LANGUAGE DEVELOPMENT RESEARCH

An Open Science Journal

Volume 5 | Issue 1 | September 2025
ISSN 2771-7976

About the journal

Language Development Research: An Open-Science Journal was established in 2020 to meet the field's need for a peer-reviewed journal that is committed to fully open science: LDR charges no fees for readers or authors, and mandates full sharing of materials, data and analysis code. The intended audience is all researchers and professionals with an interest in language development and related fields: first language acquisition; typical and atypical language development; the development of spoken, signed or written languages; second language learning; bi- and multilingualism; artificial language learning; adult psycholinguistics; computational modeling; communication in nonhuman animals etc. The journal is managed by its editorial board and is not owned or published by any public or private company, registered charity or nonprofit organization.

Child Language Data Exchange System

Language Development Research is the official journal of the **TalkBank system**, comprising the CHILDES, PhonBank, HomeBank, FluencyBank, Multilingualism and Clinical banks, the CLAN software (used by hundreds of researchers worldwide to analyze children's spontaneous speech data), and the Info-CHILDES mailing list, the de-facto mailing list for the field of child language development with over 1,600 subscribers.

Diamond Open Access

Language Development Research is published using the Diamond Open Access model (also known as “Platinum” or “Universal” OA). The journal does not charge users for access (e.g., subscription or download fees) or authors for publication (e.g., article processing charges).

Hosting

The **Carnegie Mellon University Library Publishing Service** (LPS) hosts the journal on a Janeway Publishing Platform with its manuscript management system (MMS) used for author submissions.

License

Language Development Research is published by TalkBank and the Carnegie Mellon University Library Publishing Service. Authors retain the copyright to their published content. This work is distributed under the terms of the **Creative Commons Attribution-Noncommercial 4.0 International license** (https://creativecommons.org/licenses/by-nc/4.0/), which permits any use, reproduction and distribution of the work for noncommercial purposes with no further permissions required provided the original work is attributed as specified under the terms of this Creative Commons license.

Peer Review and Submissions

All submissions are reviewed by a minimum of two peer reviewers, and one of our Action Editors, all well- established senior researchers, chosen to represent a wide range of theoretical and methodological expertise. Action Editors select peer reviewers based on their expertise and experience in publishing papers in the relevant topic area.

Submissions and Publication Cycle

We invite submissions that meet our criteria for rigour, without regard to the perceived novelty or importance of the findings. We publish general and special-topic articles (“Special Collections”) on a rolling basis to ensure rapid, cost-free publication for authors.

Language Development Research is published once a year, in December, with each issue containing the articles produced over the previous 12 months. Individual articles are published online as soon as they are produced. For citation purposes, articles are identified by the year of first publication and digital object identifier (DOI).

Table of Contents

Volume 5, Issue 1, September 2025

Special Issue: What Large Language Models (LLMs) Can('t) Tell Us About Child Language Acquisition

Special Issue Editor: Michael C. Frank, Stanford University, USA

Modeling the initial state of early phonetic learning in infants

Maxime Poli
CoML, ENS/PSL/EHESS/CNRS, Paris, France

Thomas Schatz
Université Aix Marseille, CNRS, LIS, Marseille, France

Emmanuel Dupoux
CoML, ENS/PSL/EHESS/CNRS, Paris, France
Meta AI Research, Paris, France

Marvin Lavechin
GIPSA-lab, Université Grenoble-Alpes, Grenoble, France

Abstract: What are the necessary conditions to acquire language? Do infants rely on simple statistical mechanisms, or do they come pre-wired with innate capabilities allowing them to learn their native language(s)? Previous modeling studies have shown that unsupervised learning algorithms could reproduce some aspects of infant phonetic learning. Despite these successes, algorithms still fail to reproduce the learning trajectories observed in infants. Here, we advocate that this failure is partly due to a wrong initial state. Contrary to infants, unsupervised learning algorithms start with little to no prior knowledge of speech sounds. In this work, we propose a modeling approach to investigate the relative contribution of innate factors and language experience in infant speech perception. Our approach allows us to investigate theories hypothesizing a more significant role of innate factors, offering new modeling opportunities for studying infant language acquisition.

Keywords: phonetic learning; language acquisition; deep learning

Corresponding author: Maxime Poli, Centre Sciences des Données, ENS, 45 rue d'Ulm, Paris, France. Email: maxime.poli@ens.psl.eu

ORCID ID: https://orcid.org/0000-0002-9377-9150

Citation: Poli, M., Schatz, T., Dupoux, E. & Lavechin, M. (2025). Modeling the initial state of early phonetic learning in infants. *Language Development Research, 5*(1), 1-34. http://doi.org/10.34842/y89t-6q31

Introduction

The ‘statistical learning hypothesis’ posits that infants learn their native languages(s) by gradually collecting statistics over their language input (Saffran & Kirkham, 2018). This is strikingly similar to how current AI’s Large Language Models (LLMs) learn: building a probabilistic model of sequences of words from the mere observation of these sequences as they occur in their language inputs[1]. How does learning in such models fare in comparison to learning in infants? First, LLMs typically learn from text, while preschool children learn from speech, which constitutes a richer, noisier, and more variable signal. Second, LLMs are trained on exceedingly large amounts of data. For instance, the recent model LLaMA was trained on 1.4T tokens, roughly 800B words (Touvron et al., 2023) while children hear only between 1M and 10M words per year (Gilkerson et al., 2017). At this rate, infants would need to live between 80,000 years and almost a million years to get the same amount of data. Therefore, current language models are outranked by children regarding robustness to input signal variability and data efficiency as already advocated in Lavechin et al. (2023) and Warstadt et al. (2023). One candidate explanation for the incredibly slow learning pace observed in LLMs is their lack of innate language capabilities. Indeed, LLMs have a relatively generic architecture that can be used to learn visual or musical patterns. In contrast, it has been claimed that language learning critically relies on evolution-supplied specialized structures unique to humans (Chomsky, 1957; Hauser et al., 2002).

Far from entering the complicated controversy about the role of innate knowledge in language and cognition, we focus in this paper on an apparently simple yet fundamental subcomponent of language: phonetics. The ability to encode the sounds of language in terms of a relatively invariant representation has been considered one of the first steps of language acquisition in infants. Quite surprisingly, preverbal infants have an excellent ability to discriminate between very subtle sound differences that sometimes escape adults. Contrary to English adult speakers, 10- to 12-month-old English-learning infants can distinguish [t̪a] from [ʈa], which is contrastive in Hindi (Werker et al., 1981). Similarly, Japanese-learning infants can discriminate [ɹa] from [la] as in ‘right’ versus ‘light’ (Kuhl et al., 2006), while Japanese adult speakers struggle to hear the difference (Best & Strange, 1992; Yamada & Tohkura, 1992). It is only when infants grow older that their perception specializes to their native language(s) (Kuhl et al., 2006; McMurray et al., 2018).

The early capacities of infants to discriminate speech sounds highlight the *initial state* of their perceptual apparatus, whereas their developmental trajectories emphasize the role of *experience* (Eimas et al., 1971; Kuhl & Iverson, 1995; Kuhl et al., 2003; Maye et al., 2002; Werker & Curtin, 2005).

[1] More precisely, they learn by predicting the conditional probability of future linguistic units — words or sub-word tokens — based on past units.

In this study, we investigate the respective contribution of initial state abilities and language experience in infant speech perception with computational modeling[2]. Our approach involves pretraining computational models of early phonetic learning to induce initial state sound discrimination capabilities. We then observe how these induced capabilities affect the learning trajectories taken by the model. Our results show that models with strong initial state capabilities better fit the observed data in 6-8 and 10-12 month-old American English and Japanese-learning infants. Our methodology allows us to explore theories positing a greater contribution of initial state factors in infant language acquisition, a theoretical space that has been largely overlooked in computational modeling until now.

Theoretical views on early phonetic learning in infants

The relative contributions of initial abilities versus language experience in phonetic learning have been subject to much debate. Aslin and Pisoni (1980) have outlined three possible theories concerning the development of speech perception in infants – see Rowland (2013) for an overview of the different theories. Those are depicted in Figure 1.

The *universal theory* hypothesizes that infants come pre-equipped with general auditory mechanisms partially shared with other species. According to this theory, newborns could initially discriminate all possible speech sound contrasts. Through exposure to speech, only sensitivity to contrasts to which the child is exposed would persist (maintenance), while sensitivity to contrasts to which the child is not exposed would decline (loss) – see Aslin et al. (2002). There exist at least two observations incompatible with the universal theory. First, infants lose sensitivity for some non-native contrasts but not all of them – see Singh et al. (2022) and Tsuji and Cristia (2014) for meta-analytic evidence. Second, infants are born capable of discriminating many sound contrasts but not all of them – e.g., see Eilers and Minifie (1975) for an example where infants fail to discriminate between [s] as in 'sing' versus [θ] as in 'thing'.

The *attunement theory*, perhaps the prevailing theory nowadays, proposes that infants come pre-equipped with language-specific mechanisms that would enable them to roughly discriminate speech sounds, although not to the same extent as adults in terms of native speech sound discrimination (Kuhl, 2004; Werker & Curtin, 2005). The attunement theory places greater importance on the role of experience by stipulating that the language(s) infants are exposed to reorganize their perceptual abilities. Through exposure to speech, infants' sensitivity to some – mostly native – contrasts would increase (facilitation), while sensitivity to some other – mostly non-native – contrasts would decline (loss). According to this theory, there may be no change in perceptual

[2] Here, we take the initial state to be the state of the perceptual system at birth. Such a system can come about through a combination of evolutionary processes (the true 'innate' components) and prenatal learning in utero. We do not attempt to distinguish these two sources of initial state abilities.

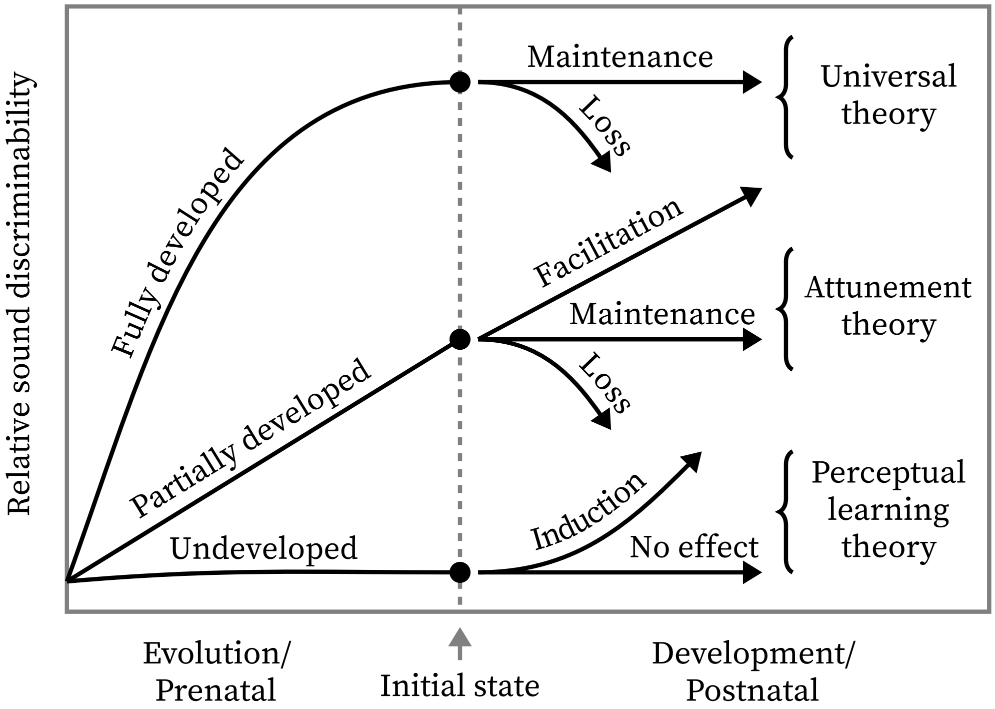

Figure 1. ***The possible effects of innate factors (Evolution/Prenatal) and language experience (Development/Postnatal) in infant speech sound perception. Adapted from Aslin and Pisoni (1980).***

abilities for some native or non-native contrasts (maintenance), which has been reported in many studies in 6- to 12-month-old infants (Best et al., 1995; Eilers & Minifie, 1975; Polka et al., 2001; Tsao et al., 2006) – see Best et al. (2016) for a review on the different speech sound discrimination trajectories observed in infants. Although there may be disagreement on the details of the implementation – e.g., the PRIMIR framework proposed by Werker and Curtin (2005) or the perceptual magnet theory proposed by Kuhl and Iverson (1995) and Kuhl et al. (2008) –, the attunement theory nicely accounts for the large array of developmental patterns observed in infants.

A major critique of both the attunement theory and the universal theory is that we may overestimate infants' capabilities to discriminate speech sounds for two reasons. First, it is common when working with infant participants to exclude those who fail to pay attention, cry, or fall asleep during the experiment. Nittrouer (2001) argues that infants may show uncooperative precisely because they cannot discriminate the stimuli presented. Consequently, excluding infants who fail to meet the criterion of the experimental procedure may result in inflated measures of discriminability. Indeed, testing 6- to 14-month-olds and 2- to 3-year-olds, Nittrouer (2001) found lower discriminability scores than typically reported in the literature – but see Aslin et al. (2002) for counterarguments. The second argument is that sound discrimination experiments use simplified stimuli in the form of prototypical sounds and cherry-picked contrasts that fail to account for the large variability of spontaneous speech encountered by infants (Pierrehumbert, 2003). Under this view, sound discrimination capabilities measured in controlled laboratory settings would not reflect the actual capabilities of infants in real-world situations (Nittrouer, 2001; Pierrehumbert, 2003; Swingley, 2009).

This brings us to the *perceptual learning theory*, which proposes a scenario where experience plays a more important role. According to this theory, there would be no need to assume innate capabilities, and infants could build the sound system of their native language(s) in a bottom-up manner from sole exposure to speech. This theory seems plausible in light of the experiments attempting to isolate learning mechanisms infants may bring to the task. For instance, Maye et al. (2002) showed that it is possible to induce different discrimination patterns in 6- and 8-month-old infants. Infants exposed to a bimodal distribution of sounds along a [ta]-[da] continuum can discriminate [ta] from [da], while those exposed to a unimodal distribution drawn from the center of the continuum cannot. The perceptual learning theory is further supported by computational modeling studies showing that it is possible to reproduce some developmental patterns in speech perceptual learning using unsupervised learning models (Lavechin et al., 2022; Räsänen et al., 2016; Schatz et al., 2021; Steels & De Boer, 2008; Vallabha et al., 2007)[3].

Current work in modeling early phonetic learning

Computational modeling studies have always been central to the debate on the relative contribution of innate factors and experience, as they shed light on what can be learned from the input signal (Ambridge & Lieven, 2011; Bates et al., 1996; Joanisse & McClelland, 2015). After all, if a model successfully reproduces the observed data in infant perceptual learning of speech sounds, do we need to posit innate factors? Despite successes in reproducing some aspects of early phonetic learning as observed in infants (Antetomaso et al., 2017; Lavechin et al., 2022; Miyazawa et al., 2010; Räsänen, 2012; Schatz et al., 2021; Steels & De Boer, 2008; Vallabha et al., 2007), we argue that computational modeling studies have thus far failed to account for the large array of infant developmental trajectories depicted in Figure 1 and reviewed in Best et al. (2016).

Let us take the example of the American English [ɹ]-[l] contrast which has received the attention of both infant development and modeling experts. In a seminal study, Kuhl et al. (2006) showed that between 6 and 8 months, Japanese- and American English-

[3] Here, our goal was to provide an overview of the main arguments supporting or challenging the different views but note that most authors do not consider these three theories to be mutually exclusive. In other words, it is unlikely that a single theory explains the development of all speech contrasts. From our perspective, the debate is not about trying to establish a single definitive theory as the absolute truth but more about where the initial state fits on the nature versus nurture continuum (vertical dashed line of Figure 1) and how this initial state influences developmental outcomes.

learning infants are capable of discriminating [ɹ] from [l] with similar performance scores. However, when tested a few months later, these same infants show markedly different perceptual patterns. By 10-12 months, American English infants show an improvement (facilitation) in their ability to discriminate the [ɹ]-[l] contrast, while Japanese infants show a decline (loss). While the effect of language exposure (higher scores for the model for whom the contrast is native) has been reproduced in numerous computational modeling studies and across different pairs of languages – e.g., Lavechin et al. (2022), Li et al. (2020), Matusevych et al. (2023), and Schatz et al. (2021) –, a closer examination of the trajectories taken by the proposed algorithms reveals notable differences with the trajectories observed in infants.

Schatz et al. (2021) used an algorithm based on a mixture of Gaussians applied to mel-frequency cepstral coefficients (MFCCs) with their first- and second-order derivatives. Their results showed that the discrimination score obtained by the Japanese model on the [ɹ]-[l] contrast increases with the quantity of speech available in the training set. In other words, for this contrast, the algorithm follows the inductive trajectory depicted in Figure 1, contrary to the loss observed in infants according to previous studies (Kuhl et al., 2006; Tsushima et al., 1994)[4].

Another example using the same algorithm from Li et al. (2020) showed a slightly different trajectory. When trained on a single speaker, the algorithm exhibits an increase (induction) on the [ɹ]-[l] contrast followed by a decrease (loss), resulting in an inverted U-shaped trajectory which, to the best of our knowledge, has not been documented in infants. Intriguingly, the same U-shaped trajectory is observed on the [w]-[j] pair (as in 'wet' versus 'yet'), which is contrastive in Japanese, and for which current theories predict either a facilitation or maintenance trajectory. This performance loss on the [w]-[j] pair, when the algorithm is trained on a large quantity of speech produced by the same speaker, may indicate that the algorithm overfits that same speaker. Lavechin et al. (2022) report the discrimination accuracy obtained by a Contrastive Predictive Coding (CPC) algorithm trained on raw speech. Although no trajectory is reported for individual contrasts, the overall discrimination accuracy averaged across all English or French contrasts also follows an inductive trajectory.

Statistical learning models, irrespective of whether they operate on handcrafted features or raw speech, are inherently rooted in the perceptual learning theory. Essentially, they begin with limited prior knowledge of speech sounds, and their performance largely tends to exhibit improvement over time. Consequently, current models fail to reproduce the large array of developmental trajectories observed in infants.

[4] In Kuhl et al.'s (2006) study, the observed decline on the [ɹ]-[l] contrast for Japanese infants was not deemed significant, contrary to Tsushima et al. (1994), where a significant decline was noted. When taken together with studies in later childhood and adulthood (e.g., Miyawaki et al. (1975)), it appears reasonable to interpret the cumulative evidence as suggestive of a decline, though additional infant experiments would be advisable.

The present study

In this study, we seek to explore the respective contribution of initial state abilities and experience on the development of speech sound discrimination capabilities. By and large, existing models of early phonetic learning implement the perceptual learning theory, where the proposed model starts with undeveloped or minimally developed discrimination capabilities (first portion of the vertical dashed line in Figure 1). Our primary contribution involves introducing a novel approach, previously used in machine learning but not yet applied to phonetic learning modeling, which consists of inducing 'innate' speech sound discrimination capabilities by pretraining our model. By controlling the initial state, we can now build computational models of early phonetic learning that posit a greater role of innate factors compared to language experience and assess which of these models better aligns with observed data in infants.

To demonstrate the relevance of our approach in modeling early phonetic learning, we simulate the learning process of American English- and Japanese-learning infants using CPC, an algorithm that learns from raw speech in an unsupervised manner already proposed in Lavechin et al. (2022, 2024) and Nguyen et al. (2020) – see Matusevych et al. (2023) for a comparison of different models. To induce 'innate' speech sound discrimination capabilities and propose models more aligned with the attunement or universal theories, we pretrain models on ambient sounds in Experiment 1, and on multilingual speech in Experiment 2. Following Schatz et al. (2021), we evaluate the model's capability to discriminate American English and Japanese contrasts using the machine ABX sound discrimination task and test whether the simulated learning trajectories align with the observed data in infants. In particular, we focus on the [ɹ]-[l] pair, which is contrastive in English but not in Japanese and for which existing data indicate a facilitation effect over the first year of life for American English-learning infants and a loss effect for Japanese-learning infants (Kuhl et al., 2006; Tsushima et al., 1994). We also analyze the performance obtained on the [w]-[j] control pair (as in 'well' versus 'yell'), contrastive in both languages, for which prevailing theories predict either a maintenance or facilitation effect over the first year of life for both American English- and Japanese-learning infants. Although fewer observations are available on the [w]-[j] contrast, see Tsushima et al. (1994) whose results are compatible with a maintenance or facilitation trajectory in Japanese-learning infants.

Experiment 1: inducing initial speech sound discrimination capabilities through pretraining on ambient sounds

In this first experiment, we ask whether it is possible to induce 'innate' speech sound discrimination capabilities in our model and how the resulting initial state affects its developmental trajectory. Following Lavechin et al. (2024), we chose a learning algorithm relying on auditory predictive coding at the core of the predictive brain hypothesis that has gained attention in the neuroscience community (Huang & Rao, 2011; Hueber et al., 2020). The algorithm learns by predicting future representations of audio based on present and past ones (see Methods).

We consider two types of models. One model starts from random initialization, which is akin to assuming little initial discrimination capabilities except those brought by the architecture which has been optimized to process human speech (see Rivière et al., 2020) and corresponds to how computational models of early phonetic learning are typically trained (e.g., see Lavechin et al., 2024; Matusevych et al., 2023; Schatz et al., 2021). This is our *no-pretraining* condition, which aligns with the perceptual learning theory. The other model follows a pre-exposure or evolutionary phase during which it is pretrained on ambient sounds (e.g., animal vocalizations, vehicles, raindrops) yielding an initial state optimized to process ambient sounds. We predicted that such a pretrained model would learn the temporal dynamics of non-speech sounds and show some initial discrimination capabilities that are not specific to any language. This is our *pretrained* condition, which aligns with attunement or universal theories.

These two types of models (no pretraining vs. pretrained) undergo an exposure phase, during which they receive the exact same language experience in the form of either Japanese or American English recordings. We then compare their learning trajectories in terms of speech sound discrimination capabilities.

Methods

Pretraining dataset

To build the dataset of ambient sounds, we started with the Animal Sound Archive (Frommolt et al., 2006; GBIF.org, 2023), which consists of 78 hours of field recordings of animals. We supplemented it with 422 hours from AudioSet (Gemmeke et al., 2017), excluding utterances annotated as human vocalizations or music and retaining only animal sounds or everyday environmental sounds. Additionally, we filtered out the remaining speech segments missed by human annotators using the model proposed in Bredin et al. (2020). Our pretraining set comprises 500 hours of ambient sounds.

Training datasets

The Japanese training set is derived from the Corpus of Spontaneous Japanese (Maekawa, 2003), and the American English corpus is made of audiobooks (Kahn et al., 2020; Kearns, 2014). For both corpora, non-speech segments were removed (Bredin et al., 2020). We then selected a subset of American English audiobooks to match the characteristics of the Japanese corpus in two aspects: the number of speakers and the duration of speech per individual speaker. Ultimately, both corpora are made of approximately 500 hours of speech data. This quantity of speech is compatible with what infants hear during their first year as current estimates vary from 60 hours (Cristia et al., 2019) to 1,000 hours (Cristia, 2023).

For each language, we built smaller datasets by partitioning them into mutually exclusive subsets of varying sizes: 1 hour, 4 hours, 20 hours, and 100 hours. In all conducted experiments, whether on Japanese or English, we trained separate models on 15 subsets for the 1-hour, 4-hour, and 20-hour splits and 5 separate models for the 100-hour split.

The learner model

We chose Contrastive Predictive Coding (CPC) as our core learning mechanism (Oord et al., 2018; Rivière et al., 2020). In the Zero Resource Speech Challenge 2021 on unsupervised representation learning, CPC achieved the best speech sound discrimination scores (Dunbar et al., 2021). This model takes as input the raw waveforms. It is designed to predict future states of a sequence from its past in an autoregressive manner. In other words, given a sequence of observations, the model aims to accurately predict the next state of the input sequence based on its past context. This predictive task is achieved through a contrastive objective, where the model learns to distinguish between positive samples — the actual future states — and a set of negative samples — sampled from other parts of the dataset — during training (see implementation details in Appendix "Contrastive Predictive Coding").

Measuring speech sound discrimination

To assess the model's ability to discriminate contrasts, we conducted the same machine ABX sound discrimination task as used by Schatz et al. (2013). This evaluation procedure presents the model with three triphone stimuli pronounced by the same speaker labeled as A, B, and X. A and X are two instances of the same triphone (e.g., 'boot'), while B differs only in the central phone while maintaining the same context (e.g., 'beet'). We compute the corresponding representations R_A, R_B, and R_X for these stimuli and calculate the pairwise distances $d(R_A, R_X)$ and $d(R_B, R_X)$, with d the angular distance. As stimuli can have different durations, we perform Dynamic Time Warping (DTW) to obtain a time alignment before computing the average angular distance along the shortest DTW path. The representations of A and X returned by the model are more similar than those of B and X if $d(R_A, R_X) < d(R_B, R_X)$, in which case the model is considered to be correct in discriminating the contrasts. The ABX accuracy is computed as the average number of times the model provides the correct answer across all possible triphones and all possible contrasts. Alternatively, the accuracy can be computed across all possible triphones containing a specific contrast (e.g., [ɹ]-[l]).

Evaluation sets

We used the same evaluation sets as in Schatz et al. (2021). These sets consist of two Japanese corpora – the left-out subset of the CSJ and the Globalphone corpus of Japanese (GPJ) (Schultz, 2002) — and two American English corpora – a subset of the Wall Street

Journal corpus (WSJ) (Paul & Baker, 1992) and the Buckeye corpus (Pitt et al., 2005). The CSJ and Buckeye corpora contain more spontaneous speech, while GPJ and WSJ are composed of read speech. The CSJ evaluation set was built from the speech of speakers absent in the training set. All four evaluation sets are made of approximately ten hours of speech along with their forced-aligned phonetic transcripts.. Across registers (read or spontaneous speech), the number of speakers, the proportion of male and female speakers, and the cumulated duration per speaker are matched.

We compute the ABX accuracy in the *native* and the *non-native* condition. In the native condition, models are evaluated on the same language they have been exposed to (e.g., the Japanese model evaluated on our two Japanese evaluation sets). In the non-native condition, models are evaluated on the language they have not been exposed to (e.g., the Japanese model evaluated on our two American English evaluation sets). When mutually exclusive training sets of the same duration are available, we consider the mean and the standard deviation of the ABX accuracy in either the native or non-native condition.

Results and discussion

We begin by measuring the ABX accuracy of both our unpretrained and pretrained learners to assess their initial speech sound discrimination capabilities. We then compare the learning trajectories displayed by our two types of learners during the language exposure phase. To gain deeper insights into the nature of our two initial states (no pretraining vs. pretrained), we visualize the separability of the representations according to phonetic categories. Finally, we reflect on how the learning trajectories exhibited by our learners on the [w]-[j] and [ɹ]-[l] contrasts fare with the data observed in infants.

Initial speech sound discrimination capabilities and developmental trajectories

Panel **a)** of Figure 2 shows the average American English and Japanese ABX accuracy obtained by our two initial states: with no pretraining (in blue) or with pretraining on ambient sounds (in orange). Our randomly initialized model, which has not been pretrained, obtains an average ABX accuracy of 62.3% (μ_{JP} = 65.2%, μ_{EN} = 59.4%). In contrast, our model pretrained on ambient sounds obtains 92.4% ABX accuracy (μ_{JP} = 93.1%, μ_{EN} = 91.8%) showing better discrimination capabilities. We interpret the surprisingly high ABX accuracy obtained by our model pretrained on non-speech sounds as evidence that learning generic representations not specific to any language is enough to discriminate most human speech sounds.

Now that our first goal – inducing initial speech sound discrimination capabilities in our model – has been achieved, we analyze the learning trajectory exhibited by our model after exposure to either American English or Japanese in panel **b)**. Here, we

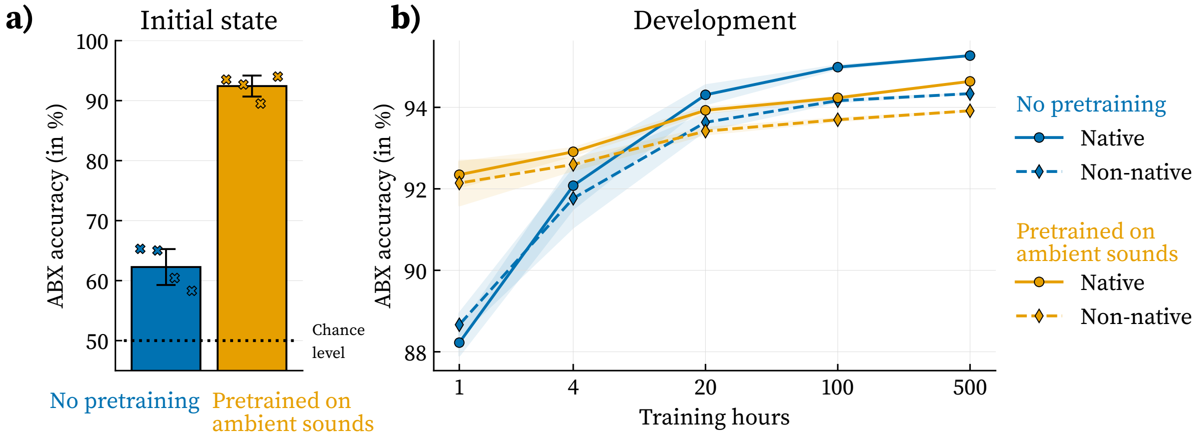

Figure 2. ***Comparison of our learner trained with no pretraining (in blue) or with pretraining on ambient sounds (in orange). Panel a) shows the average American English and Japanese ABX accuracy obtained by both types of learners before language experience (initial state). Panel b) shows the same information for native (same training and test language; dashed line) and non-native (different training and test languages; solid line) models as a function of the quantity of speech available in the training set (development). Error bars in panel a) represent +/- the standard deviation computed across our four evaluation sets. Shaded areas in panel b) represent +/- the standard deviation computed across mutually exclusive training sets whose number depends on the quantity of data.***

distinguish between native (same training and test language, solid line) and non-native (different training and test languages, dashed line) models.

Let us first focus on the trajectory exhibited by our model that has not been pretrained in blue. For low data quantities (between 1 and 4 hours), the native and non-native models obtain similar ABX accuracies, indicating that models have not yet learned language-specific representations. In other words, the American English model discriminates American English sounds as accurately as the Japanese model (and vice-versa). It is only after substantial exposure to their ‘native’ language (20 hours) that models start learning language-specific representations. Overall, we observe a positive effect of data quantity on the sound discrimination performance of our models. This is true for both the native and the non-native models. The more speech the model receives, the better it discriminates sounds. While this is expected in the native condition (e.g., exposure to Japanese makes the model better at discriminating Japanese sounds), this might be more surprising in the non-native condition. This is because there are many shared sounds across the two languages and the results reported in panel **b)** are computed across all possible contrasts – similar to what has been observed by Lavechin et al. (2022), Matusevych et al. (2023), and Schatz et al. (2021).

We now turn to the model pretrained on ambient sounds in orange. The pretrained model starts with better sound discrimination capabilities and exhibits a slower learning trajectory. Similarly to models which have not been pretrained, models pretrained on ambient sounds obtain a higher ABX accuracy with an increase in the quantity of speech. They also learn more language-specific representations as they receive more speech (the gap between the orange solid and dashed lines broadens with the number of training hours). Interestingly, after exposure to 500 hours of speech, pretrained models performed slightly worse than models that were not pretrained. Similarly, they learn representations that are less language-specific. Indeed, the relative difference in ABX error rate between native and non-native models is 16.5% in the no-pretraining condition versus 11.9% in the pretrained condition. We conducted two-way ANOVA analyses with factors nativeness and training language for each speech quantity (1h, 4h, 20h and 100h). In all settings the p-value was lower than .0001 indicating significant differences between the native and non-native models. While pretraining on ambient sounds initially steers the model in a favorable direction enabling it to discriminate speech sound contrasts effectively, this pre-exposure to non-speech sounds ends up hurting the performance of our model in processing speech sounds. Although it is hard to provide precise evidence, we hypothesize that, even after exposure to 500 hours of speech, some neurons are still dedicated to processing non-speech sounds.

Visualization of the initial sound discrimination capabilities

To better understand the initial sound discrimination capabilities induced previously through pretraining, we visualize in Figure 3 the phone representations in a two-dimensional space using the t-distributed Stochastic Neighbor Embedding (t-SNE) method – as done in de Seyssel et al. (2022) or Lavechin et al. (2022).

Panel **a)** shows the t-SNE projection of the phone representations of our two initial states: no pretraining versus pretrained on ambient sounds. Although no fine-grained separability between sonority categories was expected for the unpretrained model, we still observe some degree of separability between consonants and vowels. This aligns with the above-chance ABX accuracy of 62.3% obtained by this model (Figure 2). The model pretrained on ambient sounds show drastically different separability patterns. Here, we observe that phones are organized along a sonority continuum with a relatively good separability between the different categories, despite the model never receiving speech sounds during pretraining. Although results are only presented on our American English test sets, similar patterns are observed on our Japanese test set.

In panel **b)**, we specifically study the separability between the [ɹ]-[l] and [w]-[j] contrasts which will be the focus of the upcoming section. Our unpretrained model shows no separability for the [ɹ]-[l] or [w]-[j] contrast. However, this is not the case with our model pretrained on ambient sounds, which shows good separability for both contrasts.

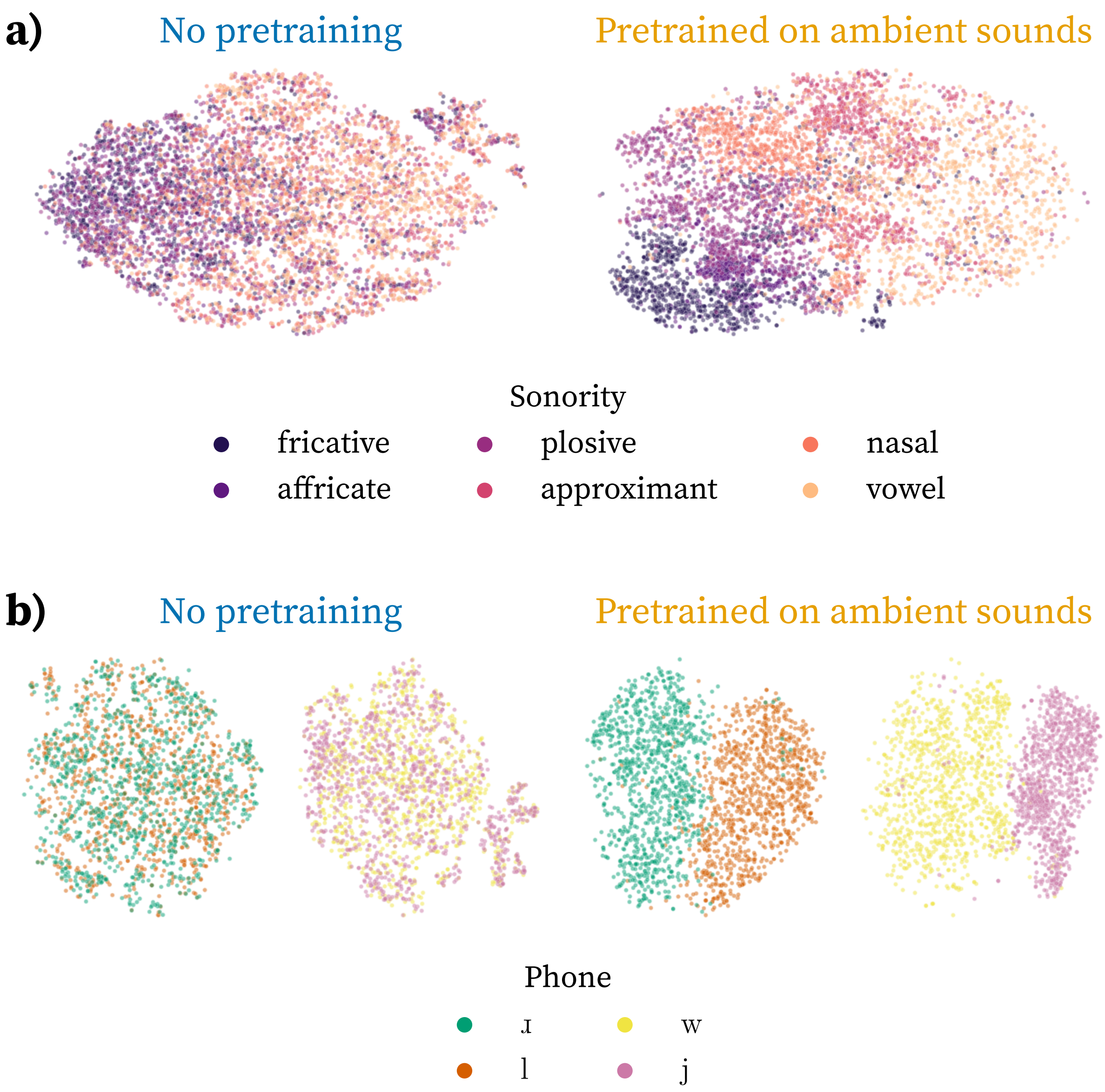

Figure 3. ***Visualization of the initial sound discrimination capabilities for our learner trained with no pretraining (in blue) or with pretraining on ambient sounds (in orange). Panel a) shows t-SNEs visualizations of the continuous representations (last layer) averaged within phones in the American English test set according to sonority. Panel b) shows the same information for the American English [ɹ]-[l] and [w]-[j] contrasts. Each point is the t-SNE projection of an individual phone's representation.***

These results demonstrate that inducing 'innate' speech sound discrimination capabilities is possible via pretraining on non-speech sounds. The first version of our model comes with no pretraining (random initialization) and shows limited initial speech sound discriminability. This version corresponds to the initial state of most computational models of early phonetic learning but does not necessarily align with dominant theories of early phonetic acquisition in infants – it implements the perceptual learning theory. A second version of our model comes with pretraining and shows relatively good speech sound discriminability – it implements the attunement or universal theory.

Now that we have two different initial states at both ends of the nature-nurture continuum, an important question arises: Which better predicts the developmental trajectory observed in infants?

Individual trajectories for the [ɹ]-[l] and [w]-[j] pairs

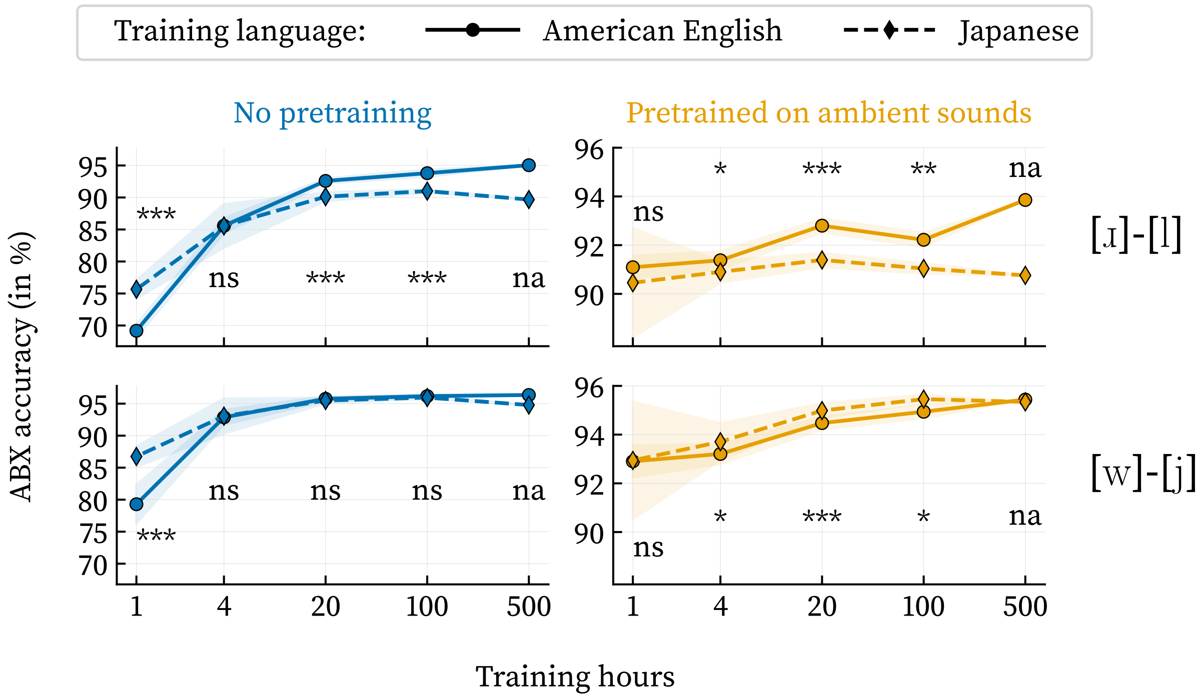

Figure 4. *Comparison of the learning trajectory exhibited by our model without pretraining (in blue) or with pretraining on ambient sounds (in orange) on the [ɹ]-[l] and [w]-[j] pairs. Models are trained on either American English (solid line) or Japanese (dashed line). [ɹ]-[l] and [w]-[j] occurrences are extracted from our American English evaluation sets. Shaded areas represent +/- the standard deviation computed across mutually exclusive training sets whose number depends on the quantity of data. Significance scores are obtained with a one-way ANOVA with factor training language (na: not applicable, ns: not significant, * p<.05, **, p<.001, *** p<.0001).*

To investigate this question, we study the learning trajectories exhibited by our models on the American English [ɹ]-[l] pair, contrastive in American English but not in Japanese, and the [w]-[j] control pair, contrastive in both languages. Figure 4 shows the ABX accuracy obtained on these contrasts for our American English model (solid line) or our Japanese model (dashed line), in the no pretraining condition (in blue) or the pretrained condition (in orange).

Let us first focus on the no pretraining condition. The American English model better discriminates the [ɹ]-[l] contrast than the non-native Japanese model. We also observe that the gap between the two models increases with the quantity of speech. On the contrary, on the [w]-[j] contrast, our native American English and our non-native Japanese models develop similar discrimination performances. These results closely replicate those of Li et al. (2020) and Schatz et al. (2021) with a different model and correspond, at least to some extent, to what has been observed in infants, namely that 10-12 month-old American English- and Japanese-learning infants show a similar discrimination performance on the [w]-[j] contrast, but American English-learning infants show better discrimination on the [ɹ]-[l] contrast.

Looking more closely at how the trajectories exhibited by our models fare with those observed in infants, we observe notable differences. While the American English model succeeds in reproducing the facilitation trajectory observed in American English infants on the [ɹ]-[l] contrast, this is not the case with the Japanese model. Indeed, our unpretrained Japanese model also follows an inductive trajectory, while Kuhl et al. (2006) and Tsushima et al. (1994) reported a loss trajectory in Japanese-learning infants between 6-8 and 10-12 months for this specific contrast. Although there is less data available on the [w]-[j] contrast, prevailing theories predict either a facilitation or a maintenance trajectory compatible with the trajectories exhibited by our unpretrained model.

We now turn to a similar analysis of the trajectories exhibited by our models pretrained on ambient sounds in orange. In this condition, our native American English model replicates the facilitation trajectory observed in American English-learning infants on the [ɹ]-[l] contrast. On this same contrast, our non-native Japanese model now exhibits a maintenance trajectory with constant performance regardless of the quantity of speech available in the training set. While this maintenance trajectory still does not perfectly match what has been observed in infants (i.e., a loss trajectory), the match is closer than in the no pretraining condition. Indeed, Kuhl et al. (2006) report a low difference in discrimination accuracy between the 6-8 month-old group and the 10-12 month-old Japanese group. Furthermore, the effect of age was found not significant for the Japanese group. Therefore, we interpret Kuhl's results as compatible with the maintenance trajectory exhibited by our Japanese model. Our interpretation of the learning trajectories exhibited by our model concerning the [w]-[j] contrast in relation to the infant literature is similar to that presented for the no pretraining condition and will not be repeated here. An interesting observation, however, is that performance on the [w]-[j] contrast still improves after 500 hours of speech, contrary to the no pretraining condition in which performance flattens after 20 hours of speech.

Experiment 2: inducing initial speech sound discrimination capabilities through multilingual pretraining

Experiment 1 showed that it is possible to induce innate speech sound discrimination capabilities by pretraining on ambient sounds. During the developmental phase, our pretrained model exposed to Japanese exhibits a maintenance trajectory on the [ɹ]-[l] contrast, more closely resembling infant behavioral data that suggest a loss trajectory (Figure 4). In the present experiment, we ask whether it is possible to induce higher initial speech sound discrimination capabilities – and perhaps obtain a loss trajectory on the [ɹ]-[l] contrast – with a different pretraining strategy: pretraining on a set of typologically diverse languages.

Methods

We use the same training sets, learner, evaluation sets, and evaluation protocol as used in Experiment 1. The only difference is that we pretrain on a multilingual corpus derived from VoxPopuli (Wang et al., 2021), a large-scale multilingual speech corpus containing recordings of European Parliament events. We remove the Germanic languages from the 23 languages present in the dataset to prevent the model trained on English from being positively biased. This procedure resulted in selecting 18 typologically diverse languages[5]. To ensure consistency with Experiment 1, our pretraining set is made of 500 hours of speech uniformly sampled across languages, resulting in approximately 28 hours per language.

Results and discussion

Initial sound discrimination capabilities and developmental trajectories

Panel **a)** of Figure 5 suggests that pretraining on multilingual is sensibly similar to pretraining on ambient sounds in terms of initial speech sound discrimination capabilities (μ_{JP} = 93.5 %, μ_{EN} = 92.1%). Contrary to our initial hypothesis, training on mulitilingual speech does not yield higher speech sound discrimination capabilities compared to pretraining on ambient sounds.

During the developmental phase, the learning trajectories obtained in the pretrained condition are similar than those obtained in Experiment 1. Two-way ANOVAs also resulted in p-values lower than .0001 for all speech quantities indicating significant differences between the native and non-native models. A notable difference compared to Experiment 1 is that pretraining on multilingual speech does not hurt the performance obtained after 500 hours of exposure, contrary to what was observed when pretraining on ambient sounds, as shown in panel **b)** of Figure 5.

[5] Bulgarian, Czech, Croatian, Estonian, Finnish, French, Greek, Hungarian, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovene and Spanish.

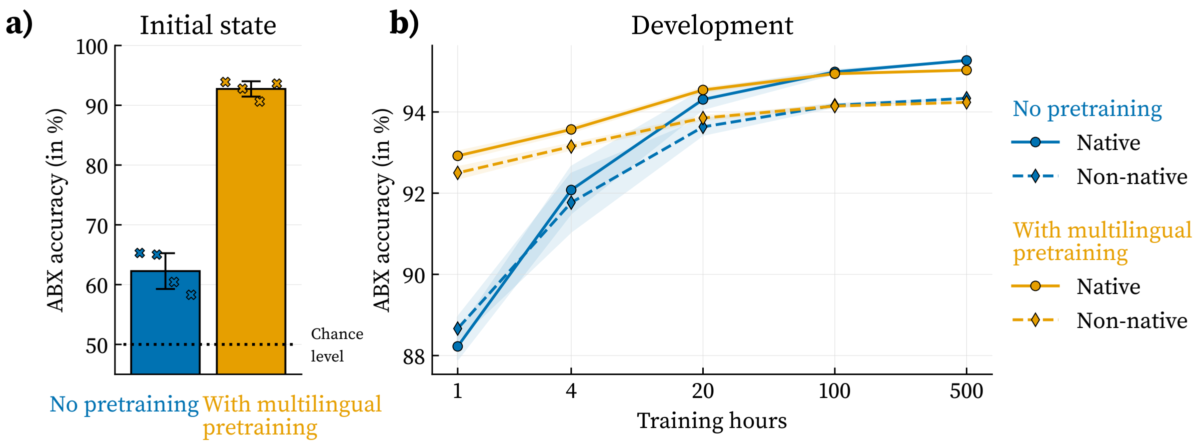

Figure 5. ***Comparison of our learner trained with no pretraining (in blue) or with multilingual pretraining (in orange) for native (same training and test language; solid line) and non-native (different training and test languages; dashed line) models as a function of the quantity of speech available in the training set (development). Shaded areas represent +/- the standard deviation computed across mutually exclusive training sets whose number depends on the quantity of data.***

Individual trajectories for the [ɹ]-[l] and [w]-[j] pairs

Again, the learning trajectories on the [ɹ]-[l] and [w]-[j] contrasts in Figure 6 are sensibly similar to those observed in Experiment 1. However, in the pretrained condition, the Japanese model seems to follow a facilitation trajectory on the [ɹ]-[l] contrast, contrary to the maintenance trajectory observed in Experiment 1. This is due to the lower ABX accuracy on the [ɹ]-[l] contrast obtained by the initial state pretrained on multilingual speech (85.8% on Buckeye, 91.7% on WSJ) compared to the initial state pretrained on ambient sounds (87.8% on Buckeye, 93.9% on WSJ).

In the context of this study, pretraining on 500 hours of multilingual speech does not seem to present any advantage as compared to pretraining on ambient sounds. Admittedly, training on larger quantities of multilingual speech – and perhaps a higher number of languages – may yield a model that starts with higher initial speech sound discrimination capabilities, as was the initial goal of this experiment.

In a concluding experiment (see Experiment 3 in Appendix), we show that our model reproduces the trajectories observed in infants: facilitation on the [ɹ]-[l] contrast and maintenance on the [w]-[j] contrast for American English-learning infants; loss on the [ɹ]-[l] contrast and maintenance on the [w]-[j] contrast for Japanese-learning infants. This is achieved through cross-lingual pretraining, where models are pre-trained on either American English or Japanese and then further trained on the language to which they have not been exposed. This protocol assumes higher non-native sound discrimination

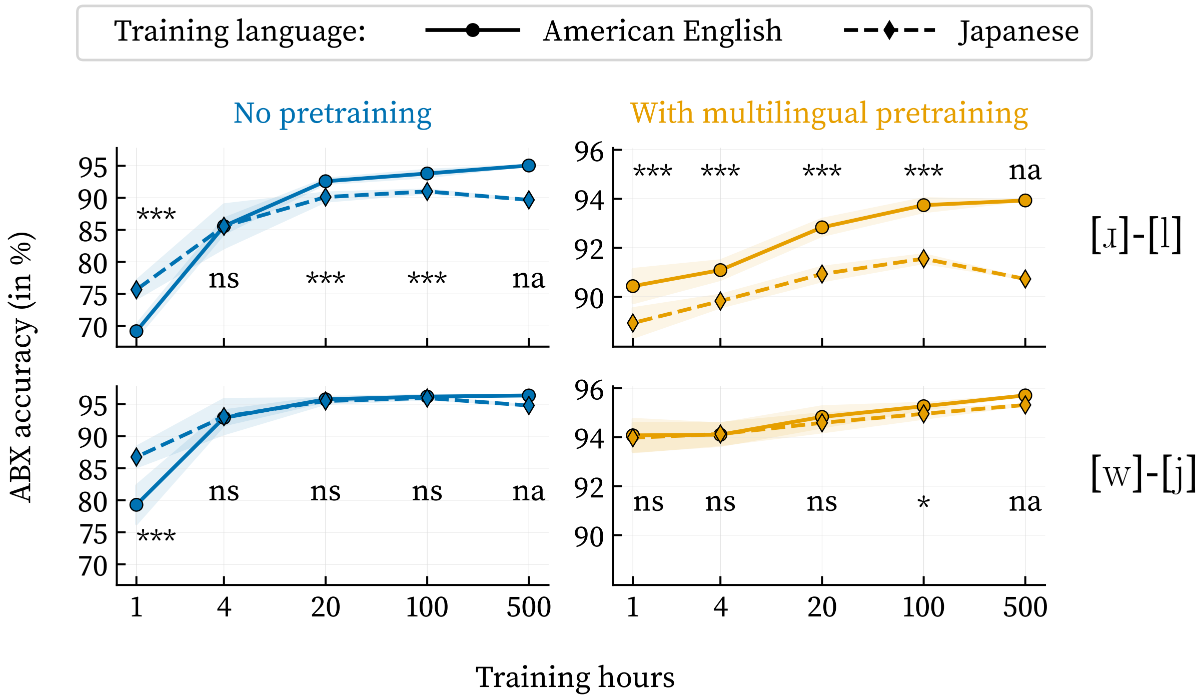

Figure 6. ***Comparison of the learning trajectory exhibited by our model without pretraining (in blue) or with multilingual pretraining (in orange) on the [ɹ]-[l] and [w]-[j] pairs. Models are trained on either American English (solid line) or Japanese (dashed line). [ɹ]-[l] and [w]-[j] occurrences are extracted from our American English evaluation sets. Shaded areas represent +/- the standard deviation computed across mutually exclusive training sets whose number depends on the quantity of data. Significance scores are obtained with a one-way ANOVA with factor training language (na: not applicable, ns: not significant, * p<.05, **, p<.001, *** p<.0001).***

capabilities than in Experiment 1 or 2. While this final experiment may not have direct relevance from an evolutionary perspective, it achieves to demonstrate that our model's performance can maintain, improve, or deteriorate depending on the interaction between innate and environmental factors.

General discussion

What is the respective contribution of innate factors and experience in child language acquisition? Without bringing indisputable evidence to the question, we proposed a novel method to build computational models of early phonetic learning that start with initial sound discrimination capabilities. Conducting two experiments, we showed that the model endowed with initial capabilities could demonstrate maintenance, facilitation, or loss trajectories, as opposed to the standard model, which learns from scratch and mostly exhibits facilitation trajectories. Here, we reflect on the implications of our findings for the existing literature on modeling infant phonetic learning. We first return to the idea of language-universal capabilities in newborns. We then propose other approaches to induce such capabilities in computational models. Finally, we reflect on how our work can be extended in a more systematic approach to the study of infant phonetic learning.

The idea of universal speech perception capabilities at the initial state is prevailing in current theories of language acquisition. In Kuhl's (2004) words, infants have an "initial universal ability to distinguish between phonetic units". Werker and Curtin (2005) write about "language-general" and "language-specific" perception. In our view, testing these theories should not only consist in collecting relevant data in infants but also in implementing them (de Seyssel et al., 2023; Dupoux, 2018). In that regard, our approach has two advantages. First, it encourages us to transform verbally-expressed ideas into implementable algorithms. Second, it offers us ways to test and compare our verbal theories.

In Experiment 1, we implemented the idea of a language-universal perceptual space by pretraining on ambient sounds. We found that the learning trajectories taken by the model during the developmental phase better fit the observed data in infants, providing evidence in favor of attunement and universal theories. In Experiment 2, we proposed a second strategy that consists of pretraining on multilingual speech data. Admittedly, one could devise different strategies – that should be equally evaluated in terms of their fit with observed data in infants. For instance, one might pretrain at a larger scale both in terms of quantity of speech data and number of languages. This could be done by training on the more than 7,000 languages being spoken worldwide[6] before comparing the learning trajectories taken by the model when trained on a single language with those observed in infants (hypothesizing rather strong initial capabilities). On the contrary, one could devise strategies to build models that assume poorer initial capabilities by training on a different source of data or by lowering the amount of data available in the pretraining set. Importantly, our goal is not to provide an explanation of the evolutionary transition from a primitive amphibian auditory system to the human auditory system. In that regard, the pretraining strategy has no other function than to hypothesize some degree of initial capabilities, offering us a rather vast ground for exploration.

In contrast to existing modeling studies (Lavechin et al., 2024; Li et al., 2020; Matusevych et al., 2023; Schatz et al., 2021), our approach goes beyond evaluating models solely based on measures of native advantage (i.e., better discrimination score for the model for whom the contrast is native). It includes assessing their fit to developmental trajectories observed in infants. This work focused on the [ɹ]-[l] pair contrastive in American

[6] https://www.ethnologue.com

English but not in Japanese and the [w]-[j] pair contrastive in both languages. However, there is available data in Zulu, Tigrinya, Taa, Nuu-Chah-Nulth, Spanish, Hindi, Czech, Nthlakampx, and Mandarin (Best et al., 2016). A more systematic approach would involve building a training set for each of these languages and studying the speech sound discrimination patterns developed by computational models. Successful models, capturing a significant proportion of the variance of the available empirical record, can then be used to obtain predictions on contrasts that have yet to be studied. These predicted trajectories can subsequently be validated or refuted through new sound discrimination experiments with infants. Alternatively, instead of focusing on data from individual studies, one could compare the learning outcomes developed by computational models against robust data from meta-analyses as proposed by Cruz Blandón et al. (2023). We strongly believe that such a systematic dialogue between experimental and modeling studies is essential to foster theory-building in psychological sciences (Frank et al., 2017).

Conclusion

Even though current AI language models have been considered as supporting empiricist views of language learning, these models offer a much larger range of theoretical options. By decomposing model training in a (potentially long) evolutionary phase and a (potentially short) developmental phase, they can implement either extreme versions of empiricism, or extreme versions of nativism, with a whole range of intermediary cases. In our work, we conducted two experiments that demonstrated the possibility of inducing initial sound discrimination capabilities in our computational model of early phonetic learning. Contrary to the randomly initialized model, the models pre-equipped with discrimination capabilities showed learning trajectories more closely resembling those observed in infants. Further research is needed to establish in a more quantitative fashion what model of the initial state would fit best the observed learning trajectories in infants.

References

Ambridge, B., & Lieven, E. V. (2011). *Child language acquisition: Contrasting theoretical approaches*. Cambridge University Press. https://doi.org/10.1017/CBO9780511975073

Antetomaso, S., Miyazawa, K., Feldman, N., Elsner, M., Hitczenko, K., & Mazuka, R. (2017). Modeling phonetic category learning from natural acoustic data. *Proceedings of the annual Boston University Conference on Language Development*. https://par.nsf.gov/biblio/10057880

Aslin, R., & Pisoni, D. (1980). Some developmental processes in speech perception. In G. H. Yeni-Komshian, J. F. Kavanagh, & C. A. Ferguson (Eds.), *Child phonology: Vol 2, Perception* (pp. 67–96). New York: Academic Press.

Aslin, R., Werker, J. F., & Morgan, J. L. (2002). Innate phonetic boundaries revisited (l). *The Journal of the Acoustical Society of America, 112*(4), 1257–1260. https://doi.org/10.1121/1.1501904

Bates, E., Elman, J., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996, October). *Rethinking innateness: A connectionist perspective on development*. The MIT Press. https://doi.org/10.7551/mitpress/5929.001.0001

Best, C. T., Goldstein, L. M., Nam, H., & Tyler, M. D. (2016). Articulating what infants attune to in native speech. *Ecological Psychology, 28*(4), 216–261. https://doi.org/10.1080/10407413.2016.1230372

Best, C. T., McRoberts, G. W., LaFleur, R., & Silver-Isenstadt, J. (1995). Divergent developmental patterns for infants' perception of two nonnative consonant contrasts. *Infant behavior and development, 18*(3), 339–350. https://doi.org/10.1016/0163-6383(95)90022-5

Best, C. T., & Strange, W. (1992). Effects of phonological and phonetic factors on cross-language perception of approximants. *Journal of phonetics, 20*(3), 305–330. https://doi.org/10.1016/S0095-4470(19)30637-0

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., & Gill, M.-P. (2020). Pyannote.audio: Neural building blocks for speaker diarization. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7124–7128. https://doi.org/10.1109/ICASSP40776.2020.9052974

Chomsky, N. (1957). Syntactic structures. Mouton de Gruyter. https://doi.org/10.1515/9783112316009

Cristia, A. (2023). A systematic review suggests marked differences in the prevalence of infant-directed vocalization across groups of populations. *Developmental Science, 26*(1), e13265. https://doi.org/10.1111/desc.13265

Cristia, A., Dupoux, E., Gurven, M., & Stieglitz, J. (2019). Child-directed speech is infrequent in a forager-farmer population: A time allocation study. *Child Development, 90*(3), 759–773. https://doi.org/10.1111/cdev.12974

Cruz Blandón, M. A., Cristia, A., & Räsänen, O. (2023). Introducing meta-analysis in the evaluation of computational models of infant language development. *Cognitive Science*, *47*(7), e13307. https://doi.org/10.1111/cogs.13307

de Seyssel, M., Lavechin, M., Adi, Y., Dupoux, E., & Wisniewski, G. (2022). Probing phoneme, language and speaker information in unsupervised speech representations. *Proc. Interspeech 2022*, 1402–1406. https://doi.org/10.21437/Interspeech.2022-373

de Seyssel, M., Lavechin, M., & Dupoux, E. (2023). Realistic and broad-scope learning simulations: first results and challenges. *Journal of Child Language*, 1–24. https://doi.org/10.1017/S0305000923000272

Dunbar, E., Bernard, M., Hamilakis, N., Nguyen, T. A., de Seyssel, M., Rozé, P., Rivière, M., Kharitonov, E., & Dupoux, E. (2021). The Zero Resource Speech Challenge 2021: Spoken language modelling. *Proc. Interspeech 2021*, 1574–1578. https://doi.org/10.21437/Interspeech.2021-1755

Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, *173*, 43–59. https://doi.org/10.1016/j.cognition.2017.11.008

Eilers, R. E., & Minifie, F. D. (1975). Fricative discrimination in early infancy. *Journal of speech and Hearing Research*, *18*(1), 158–167. https://doi.org/10.1044/jshr.1801.158

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science*, *171*(3968), 303–306. https://doi.org/10.1126/science.171.3968.303

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., Hamlin, J. K., Hannon, E. E., Kline, M., Levelt, C., et al. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, *22*(4), 421–435. https://doi.org/10.1111/infa.12182

Frommolt, K.-H., Bardeli, R., Kurth, F., & Clausen, M. (2006). The animal sound archive at the Humboldt-University of Berlin: Current activities in conservation and improving access for bioacoustic research. *Advances in Bioacoustics 2*, 139–144. https://www.ibac.info/advances-in-bioacoustics-ii#aib10

GBIF.org. (2023). GBIF occurrence download. https://doi.org/10.15468/dl.dmckt3

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., & Ritter, M. (2017). Audio set: An ontology and human-labeled dataset for audio events. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 776–780. https://doi.org/10.1109/ICASSP.2017.7952261

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Kimbrough Oller, D., Hansen, J. H., & Paul, T. D. (2017). Mapping the early language environment using all-day recordings and automated analysis. *American journal of speech-language pathology*, *26*(2), 248–265. https://doi.org/10.1044/2016_AJSLP-15-0169

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science*, *298*(5598), 1569–1579. https://doi.org/10.1126/science.298.5598.1569

Huang, Y., & Rao, R. P. (2011). Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(5), 580–593. https://doi.org/10.1002/wcs.142

Hueber, T., Tatulli, E., Girin, L., & Schwartz, J.-L. (2020). Evaluating the potential gain of auditory and audiovisual speech-predictive coding using deep learning. *Neural Computation*, *32*(3), 596–625. https://doi.org/10.1162/neco_a_01264

Joanisse, M. F., & McClelland, J. L. (2015). Connectionist perspectives on language learning, representation and processing. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*(3), 235–247. https://doi.org/10.1002/wcs.1340

Kahn, J., Rivière, M., Zheng, W., Kharitonov, E., Xu, Q., Mazaré, P., Karadayi, J., Liptchinsky, V., Collobert, R., Fuegen, C., Likhomanenko, T., Synnaeve, G., Joulin, A., Mohamed, A., & Dupoux, E. (2020). Libri-light: A benchmark for ASR with limited or no supervision. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7669–7673. https://doi.org/10.1109/ICASSP40776.2020.9052942

Kearns, J. (2014). LibriVox: Free public domain audiobooks. *Reference Reviews*, *28*(1), 7–8. https://doi.org/10.1108/RR-08-2013-0197

Kharitonov, E., Rivière, M., Synnaeve, G., Wolf, L., Mazaré, P.-E., Douze, M., & Dupoux, E. (2021). Data augmenting contrastive learning of speech representations in the time domain. *2021 IEEE Spoken Language Technology Workshop (SLT)*, 215–222. https://doi.org/10.1109/SLT48900.2021.9383605

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature reviews neuroscience*, *5*(11), 831–843. https://doi.org/10.1038/nrn1533

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 979–1000. https://doi.org/10.1098/rstb.2007.2154

Kuhl, P. K., & Iverson, P. (1995). Linguistic experience and the “perceptual magnet effect”. *Speech perception and linguistic experience: Issues in cross-language research*, 121–154.

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental science*, *9*(2), F13–F21. https://doi.org/10.1111/j.1467-7687.2006.00468.x

Kuhl, P. K., Tsao, F.-M., & Liu, H.-M. (2003). Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences*, *100*(15), 9096–9101. https://doi.org/10.1073/pnas.1532872100

Lavechin, M., De Seyssel, M., Titeux, H., Bredin, H., Wisniewski, G., Cristia, A., & Dupoux, E. (2022). Can statistical learning bootstrap early language acquisition? a modeling investigation. https://doi.org/10.31234/osf.io/rx94d

Lavechin, M., de Seyssel, M., Métais, M., Metze, F., Mohamed, A., Bredin, H., Dupoux, E., & Cristia, A. (2024). Modeling early phonetic acquisition from child-centered audio data. *Cognition*, *245*, 105734. https://doi.org/10.1016/j.cognition.2024.105734

Lavechin, M., Sy, Y., Titeux, H., Blandón, M. A. C., Räsänen, O., Bredin, H., Dupoux, E., & Cristia, A. (2023). BabySLM: language-acquisition-friendly benchmark of self-supervised spoken language models. *Proc. INTERSPEECH 2023*, 4588–4592. https://doi.org/10.21437/Interspeech.2023-978

Li, R., Schatz, T., Matusevych, Y., Goldwater, S., & Feldman, N. H. (2020). Input matters in the modeling of early phonetic learning. *Proceedings of the Annual Conference of the Cognitive Science Society*. https://par.nsf.gov/biblio/10176646

Maekawa, K. (2003). Corpus of spontaneous japanese: its design and evaluation. *Proc. ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition*, paper MMO2. https://www.isca-speech.org/archive/sspr_2003/maekawa03_sspr.html

Matusevych, Y., Schatz, T., Kamper, H., Feldman, N. H., & Goldwater, S. (2023). Infant Phonetic Learning as Perceptual Space Learning: A Crosslinguistic Evaluation of Computational Models. *Cognitive Science*, *47*(7), e13314. https://doi.org/10.1111/cogs.13314

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101–B111. https://doi.org/10.1016/s0010-0277(01)00157-3

McMurray, B., Danelz, A., Rigler, H., & Seedorff, M. (2018). Speech categorization develops slowly through adolescence. *Developmental psychology*, *54*(8), 1472. https://doi.org/10.1037/dev0000542

Miyawaki, K., Jenkins, J. J., Strange, W., Liberman, A. M., Verbrugge, R., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of japanese and english. *Perception & Psychophysics*, *18*(5), 331–340.

Miyazawa, K., Kikuchi, H., & Mazuka, R. (2010). Unsupervised learning of vowels from continuous speech based on self-organized phoneme acquisition model. *Interspeech*. https://doi.org/10.21437/Interspeech.2010-757

Nguyen, T. A., de Seyssel, M., Rozé, P., Rivière, M., Kharitonov, E., Baevski, A., Dunbar, E., & Dupoux, E. (2020). The Zero Resource Speech Benchmark 2021: Metrics and baselines for unsupervised spoken language modeling. *NeurIPS Workshop on Self-Supervised Learning for Speech and Audio Processing*. https://arxiv.org/abs/2011.11588

Nittrouer, S. (2001). Challenging the notion of innate phonetic boundaries. *The Journal of the Acoustical Society of America*, *110*(3), 1598–1605. https://doi.org/10.1121/1.1379078

Oord, A. van den, Li, Y., & Vinyals, O. (2018). *Representation Learning with Contrastive Predictive Coding*. arXiv: 1807.03748 `[cs, stat]`.

Paul, D. B., & Baker, J. M. (1992). The design for the Wall Street Journal-based CSR corpus. *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*. https://aclanthology.org/H92-1073

Pierrehumbert, J. B. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and speech*, *46*(2-3), 115–154. https://doi.org/10.1177/00238309030460020501

Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., & Raymond, W. (2005). The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. *Speech Communication*, *45*(1), 89–95. https://doi.org/10.1016/j.specom.2004.09.001

Polka, L., Colantonio, C., & Sundara, M. (2001). A cross-language comparison of/d/–/ð/perception: evidence for a new developmental pattern. *The Journal of the Acoustical Society of America*, *109*(5), 2190–2201. https://doi.org/10.1121/1.1362689

Räsänen, O. (2012). Computational modeling of phonetic and lexical learning in early language acquisition: Existing models and future directions. *Speech Communication, 54*(9), 975–997. https://doi.org/10.1016/j.specom.2012.05.001

Räsänen, O., Nagamine, T., & Mesgarani, N. (2016). Analyzing distributional learning of phonemic categories in unsupervised deep neural networks. *CogSci... Annual Conference of the Cognitive Science Society. Cognitive Science Society (US). Conference, 2016,* 1757. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5775908

Rivière, M., Joulin, A., Mazaré, P.-E., & Dupoux, E. (2020). Unsupervised pretraining transfers well across languages. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 7414–7418. https://doi.org/10.1109/ICASSP40776.2020.9054548

Rowland, C. (2013). *Understanding child language acquisition*. Routledge. https://doi.org/10.4324/9780203776025

Saffran, J. R., & Kirkham, N. Z. (2018). Infant statistical learning. *Annual Review of Psychology, 69*(1), 181–203. https://doi.org/10.1146/annurev-psych-122216-011805

Schatz, T., Feldman, N. H., Goldwater, S., Cao, X.-N., & Dupoux, E. (2021). Early phonetic learning without phonetic categories: Insights from large-scale simulations on realistic input. *Proceedings of the National Academy of Sciences, 118*(7), e2001844118. https://doi.org/10.1073/pnas.2001844118

Schatz, T., Peddinti, V., Bach, F., Jansen, A., Hermansky, H., & Dupoux, E. (2013). Evaluating speech features with the Minimal-Pair ABX task: Analysis of the classical MFC/PLP pipeline. *Proc. Interspeech 2013,* 1781–1785. https://doi.org/10.21437/Interspeech.2013-441

Schultz, T. (2002). Globalphone: a multilingual speech and text database developed at karlsruhe university. *Proc. 7th International Conference on Spoken Language Processing (ICSLP 2002),* 345–348. https://doi.org/10.21437/ICSLP.2002-151

Singh, L., Rajendra, S. J., & Mazuka, R. (2022). Diversity and representation in studies of infant perceptual narrowing. *Child Development Perspectives, 16*(4), 191–199. https://doi.org/10.1111/cdep.12468

Steels, L., & De Boer, B. (2008). Embodiment and self-organization of human categories: A case study for speech. *Body, Language and Mind, 1,* 411–430. https://doi.org/10.1515/9783110207507.3.411

Swingley, D. (2009). Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1536), 3617–3632. https://doi.org/10.1098/rstb.2009.0107

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *LLaMA: Open and efficient foundation language models*. arXiv: 2302.13971 [cs.CL].

Tsao, F.-M., Liu, H.-M., & Kuhl, P. K. (2006). Perception of native and non-native affricate-fricative contrasts: Cross-language tests on adults and infants. *The Journal of the Acoustical Society of America*, *120*(4), 2285–2294. https://doi.org/10.1121/1.2338290

Tsuji, S., & Cristia, A. (2014). Perceptual attunement in vowels: A meta-analysis. *Developmental psychobiology*, *56*(2), 179–191. https://doi.org/10.1002/dev.21179

Tsushima, T., Takizawa, O., Sasaki, M., Shiraki, S., Nishi, K., Kohno, M., Menyuk, P., & Best, C. (1994). Discrimination of english /r-l/ and /w-y/ by japanese infants at 6-12 months: language-specific developmental changes in speech perception abilities. *Proc. 3rd International Conference on Spoken Language Processing (ICSLP 1994)*, 1695–1698. https://doi.org/10.21437/ICSLP.1994-438

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*, *104*(33), 13273–13278. https://doi.org/10.1073/pnas.0705369104

Wang, C., Riviere, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J., & Dupoux, E. (2021). VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 993–1003. https://doi.org/10.18653/v1/2021.acl-long.80

Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T., & Cotterell, R. (2023). Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, 1–34. https://doi.org/10.18653/v1/2023.conll-babylm.1

Werker, J. F., & Curtin, S. (2005). PRIMIR: A developmental framework of infant speech processing. *Language learning and development*, *1*(2), 197–234. https://doi.org/10.1207/s15473341lld0102_4

Werker, J. F., Gilbert, J. H., Humphrey, K., & Tees, R. C. (1981). Developmental aspects of cross-language speech perception. *Child development*, 349–355. https://doi.org/10.2307/1129249

Yamada, R. A., & Tohkura, Y. (1992). The effects of experimental variables on the perception of american english/r/and/l/by japanese listeners. *Perception & psychophysics*, *52*, 376–392. https://doi.org/10.3758/BF03206698

Data, Code and Materials Availability Statement

The source code of all models, experimentation scripts, and data processing scripts are available at https://github.com/mxmpl/initial-phonetic-learning. This repository also contains links to download the pretraining datasets of ambient sounds, the multilingual and English training sets, model checkpoints, and comprehensive results. Audioset and the Animal Sound Archive occurrence data are made available under a CC BY 4.0 license. The Animal Sound Archive audio files are licensed under CC BY-SA 4.0 and CC BY-NC-SA 4.0. The VoxPopuli and LibriVox recordings are in the public domain. The Editor granted an exemption to materials sharing for the following datasets, on the grounds that they are subject to copyright: CSJ, GPJ, WSJ, and Buckeye.

Authorship and Contributorship Statement

M.P., T.S., E.D., M.L. designed research; M. P. performed research; M.P, M.L. wrote the manuscript with contributions from T.S. and E.D. All authors approved the final version of the manuscript and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Acknowledgements

This work was granted access to the HPC resources of IDRIS under the allocation 2023-AD011014368 made by GENCI and was supported in part by the Agence Nationale pour la Recherche (ANR-17-EURE-0017 Frontcog, ANR10-IDEX-0001-02 PSL*, ANR19-P3IA-0001 PRAIRIE 3IA Institute) and grants from CIFAR (Learning in Machines and Brains) and Meta AI Research (Research Grant). M. P. acknowledges Ph.D. funding from Agence de l'Innovation de Défense. T.S. work, carried out within the Institute of Convergence ILCB, was supported by grants from France 2030 (ANR-16-CONV-0002) and the Excellence Initiative of Aix-Marseille University (A*MIDEX).

Appendices

Contrastive predictive coding (implementation details)

Training a neural network in an unsupervised manner often requires designing a pretext task that will force the model to learn high-level representations of the input data. The pretext task in a Contrastive Predictive Coding algorithm is forward modeling. where the model is trained to predict the future states of a sequence based on its past context. During training, the model receives a positive example drawn from the near future up to 120 ms, and multiple negative examples not drawn from the near future. Given the past context of a sequence, the model has to come to reliably choose the positive sample over the negative ones.

In more technical terms, a non-linear encoder denoted as g_{enc} maps the observations x_t at time t to a latent representation $z_t = g_{\text{enc}}(x_t)$. The context-dependent representation c_t is then built by an autoregressive model, g_{ar}, which aggregates the latent representations: $c_t = g_{\text{ar}}(z_1, ..., z_t)$. Given the past context c_t, a predictor g_{pred} is asked to predict future representations z_{t+k} for $k \in \{1, ..., K\}$. The model is trained to maximize the categorical cross-entropy to correctly identify a positive future sample x_{t+k} from a set of unrelated negative samples. Formally, at step t, the loss function $\mathcal{L}_t$ for the pretext task is defined as follows:

$$\mathcal{L}_t = -\frac{1}{K}\sum_{k=1}^{K} \log\left[\frac{\exp(g_{\text{pred}}(c_t)_k^\top z_{t+k})}{\sum_{n\in\mathcal{N}_t}\exp(g_{\text{pred}}(c_t)_k^\top g_{\text{enc}}(n))}\right] \tag{1}$$

with $\mathcal{N}_t$ the set of negatives samples. The model is asked to predict up to $K = 12$ time steps in the future, equivalent to 120ms. The encoder g_{enc} comprises 5 one-dimensional convolutional layers with kernel sizes (10, 8, 4, 4, 4) and strides (5, 4, 2, 2, 2) and returns a 256-dimensional vector every 10 milliseconds. The auto-regressive model g_{ar} is a 2-layer long short-term memory network of dimension 256. The predictor g_{pred} is a single multi-head transformer layer with $K = 12$ heads, each predicting at time step $k \in \{1, ..., 12\}$. All models are trained for 100 epochs and the best epoch is selected according to the validation accuracy. For each independent dataset, 5 % of the data is used as the validation set. The other hyperparameters follow Kharitonov et al. (2021).

When training on speech, the negative samples $\mathcal{N}_t$ are drawn from within the same speaker. On the other hand, when training on ambient sounds, as there is no notion of speaker in this particular dataset, the negative samples are drawn from within the sequence. This is the only difference in the training process between the two approaches.

t-SNE visualization

To compute the t-SNE visualizations in panel **a)** of Figure 3, we first extract the audio representations of the American English Buckeye corpus. For each phone, we average the representations over time to get a single vector representation. Next, we apply the t-SNE method to reduce the 256-dimensional space into a 2-dimensional space. For the sake of clarity, only 1,000 randomly sampled phones for each category are displayed. For panel **b)** we apply the t-SNE method only on the representations of [ɹ]-[l] or [w]-[j] occurrences. Similarly, we display only 1,000 randomly sampled representations for each phonetic category.

Evaluated phonetic inventory

Table 1 shows the American English and Japanese phonetic inventory used in the ABX sound discrimination task and the t-SNE visualization.

Sonority	American English	Japanese
Fricative	f, v, θ, ð, s, z, ʃ, ʒ, h	ɸ, s, sː, z, ɕ, ɕː, ʑ, h
Affricate	ʤ, ʧ	t͡s, t͡sː, t͡ɕ, t͡ɕː
Plosive	p, b, d, t, k, g	p, pː, b, d, t, tː, k, kː, g
Approximant	w, j, ɹ, l	w, j, r
Nasal	m, n, ŋ	m, n, ɴ
Vowel	ɪ, iː, ɛ, ʌ, ɝ, æ, ɑː, ɔː, ʊ, uː, eɪ, aɪ, aʊ, ɔɪ, oʊ	ä, äː, e, eː, i, iː, o, oː, ɯ, ɯː

Table 1. ***Evaluated phonetic inventory in American English and Japanese in the International Phonetic Alphabet (IPA) standard (same as Schatz et al., 2021).***

Additional experiment: inducing initial speech sound discrimination capabilities through cross-lingual pretraining

Experiment 1 showed that it was possible to induce initial speech sound discrimination capabilities in our learner through pretraining on ambient sounds. Despite a better match between the learning trajectory exhibited by our learner and the observed data in infants, we could only simulate a maintenance trajectory on the [ɹ]-[l] contrast for the Japanese model. Experiment 2 showed that pretraining on multilingual speech did not yield higher initial speech sound discrimination capabilities than pretraining on ambient sounds.

In the present experiment, we ask whether the Japanese model can exhibit a loss trajectory on the [ɹ]-[l] contrast. We likely need to hypothesize even higher speech sound discrimination capabilities to do so. In this experiment, this is achieved through cross-lingual pretraining. Namely: we first pre-train models on either American English or

Japanese and train them on the language they have not been exposed to. This experimental protocol is akin to assuming near-perfect sound discrimination capabilities of American English contrasts by Japanese infants and near-perfect sound discrimination capabilities of Japanese contrasts by American English infants.

Arguably, such a protocol lacks ecological validity as: 1) it assumes different initial states for our American English and Japanese models; 2) it assumes near-perfect discrimination of English sounds for our Japanese model; and 3) near-perfect discrimination of Japanese sounds for our English model. However, this Experiment serves as proof that, being gifted with high enough initial sound discrimination capabilities, our Japanese model can follow a loss trajectory on the [ɹ]-[l] contrast, while maintaining a maintenance trajectory on the [w]-[j] contrast, similar to what is observed in infants (which was not shown in Experiment 1 and 2).

Methods

We use the exact same training sets, learner, evaluation sets, and evaluation protocol used in Experiment 1. The only difference is that we pretrain cross-linguistically instead of pretraining on ambient sounds. Our approach involves two distinct initial states for English and Japanese models. They are derived from the models trained on 500 hours of speech in Experiment 1. The two initial states consist of the model's weights after exposure to either 500 hours of American English or Japanese. We then train the English models starting from the Japanese weights and the Japanese models starting from the English weights.

Results and discussion

Figure 7 shows the trajectories taken by models with a cross-lingual pretraining compared to those without any pretraining. Two-way ANOVAs with factors nativeness and training language resulted in p<.0001 for each data quantity, indicating significant differences between the native and non-native models. We observe a slight negative native advantage when training on as little as 1 hour of speech. This arises from the fact that the initial state of the Japanese models, composed of weights from a model trained on 500 hours of English, performs slightly better on English than the initial state of the English models (and vice-versa). This negative native advantage reverses after exposure to 4 hours of speech.

After training on 500 hours of speech, pretrained models show identical ABX accuracies to those obtained by non-pretrained models. In other words, pre-exposure to another language does not harm or benefit the final performance obtained by the models. The cross-lingual pretraining yields different learning trajectories than those observed in Experiment 1. Here, we observe a loss trajectory for the non-native pretrained model (decreasing orange dashed line).

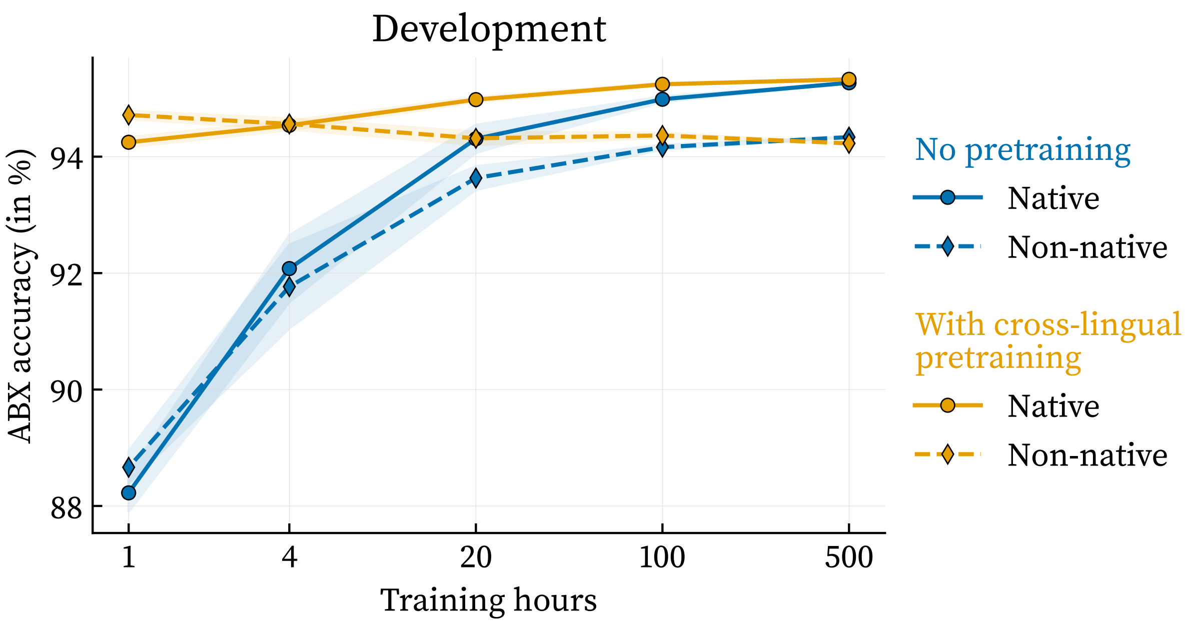

Figure 7. ***Comparison of our learner trained with no pretraining (in blue) or with cross-lingual pretraining (in orange) for native (same training and test language; solid line) and non-native (different training and test languages; dashed line) models as a function of the quantity of speech available in the training set (development). Shaded areas represent +/- the standard deviation computed across mutually exclusive training sets whose number depends on the quantity of data. Significance scores are obtained with a one-way ANOVA with factor training language (na: not applicable, ns: not significant, * p<.05, **, p<.001, *** p<.0001).***

Individual trajectories for the [ɹ]-[l] and [w]-[j] pairs

Figure 8 shows the trajectories on the [ɹ]-[l] and [w]-[j] contrasts for models that have not been pretrained (in blue) or pretrained cross-linguistically (in orange).

We will not repeat our interpretations of the trajectories taken by the unpretrained models (left column), which are the same results as those reported in Figure 4 and are left only for comparison.

In the pretrained condition, the American English model better discriminates the [ɹ]-[l] contrast than the non-native Japanese model. The American English model also successfully reproduces the facilitation trajectory observed in infants (Kuhl et al., 2006), similar to what has been observed when pretraining on ambient sounds in Experiment 1. Unlike what has been observed in Experiment 1, the Japanese model follows a loss trajectory, i.e., the performance on the [ɹ]-[l] contrast worsens as the quantity of speech increases.

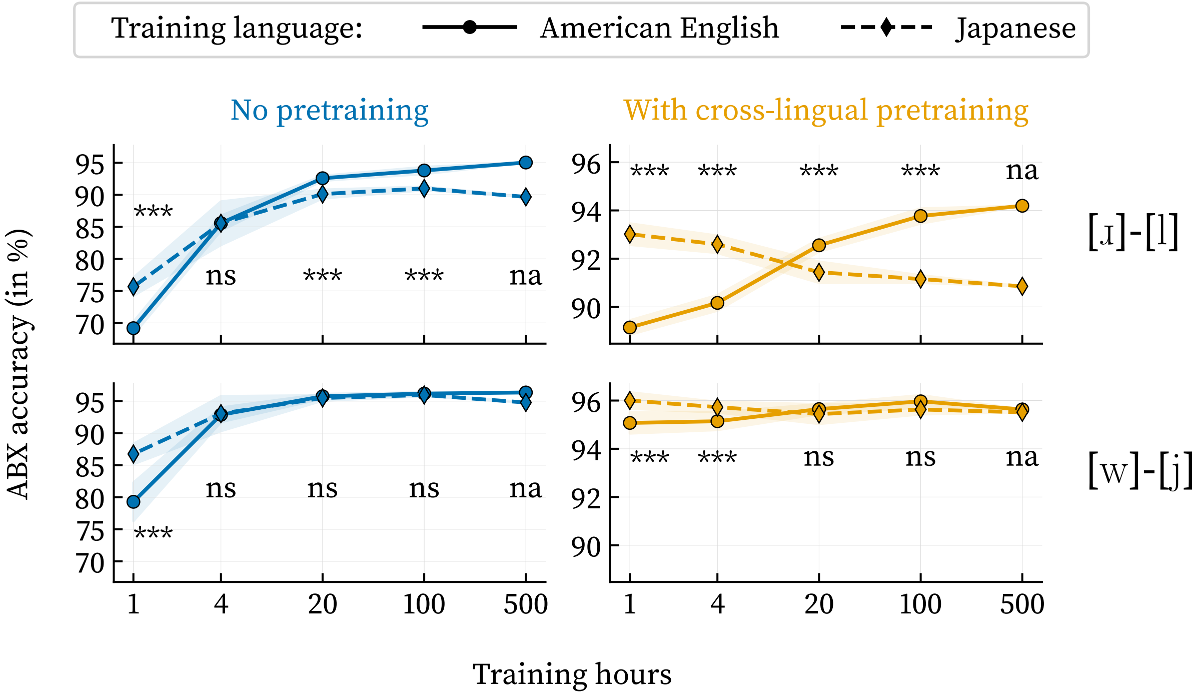

Figure 8. ***Comparison of the learning trajectory exhibited by our model without pretraining (in blue) or with cross-lingual pretraining (in orange) on the [ɹ]-[l] and [w]-[j] pairs. Models are trained on either American English (solid line) or Japanese (dashed line). [ɹ]-[l] and [w]-[j] occurrences are extracted from our American English evaluation sets. Shaded areas represent +/- the standard deviation computed across mutually exclusive training sets whose number depends on the quantity of data.***

On the [w]-[j] contrast, we now observe a maintenance trajectory instead of a facilitation trajectory in Experiment 1. In other words, after the cross-lingual pre-exposure phase, the discrimination performance obtained by our models has already converged on the [w]-[j] contrast. Performance does not benefit from further exposure to speech.

ABX sound discrimination accuracy

To enable comparisons, we provide the ABX accuracy obtained by models from Experiment 1, 2 and the additional experiment of the present study in Table 2.

Exp. #	Initial state	Training language	ABX accuracy in Japanese (CSJ / GPJ)	ABX accuracy in English (Buckeye / WSJ)
	MFCCs	–	90.8 / 91.2	87.5 / 93.4
1	No pretraining	–	65.0 / 65.3	60.4 / 58.3
1	Ambient sounds	–	92.7 / 93.5	89.5 / 94.0
2	Multilingual	–	92.8 / 93.9	90.6 / 93.6
1, 2	No pretraining	JP	96.1 / 95.5	92.0 / 94.7
1, 2	No pretraining	AE	95.2 / 95.5	93.0 / 96.4
1	Ambient sounds	JP	95.6 / 95.4	91.8 / 94.7
1	Ambient sounds	AE	94.3 / 94.9	92.1 / 95.5
2	Multilingual	JP	96.0 / 95.8	92.1 / 95.1
2	Multilingual	AE	94.6 / 95.2	92.6 / 95.8
3	Cross-lingual	JP	96.3 / 96.0	92.3 / 95.2
3	Cross-lingual	AE	94.5 / 94.9	92.7 / 96.3

Table 2. ***ABX accuracy (in %) on American English and Japanese evaluation sets.***

License

Comparing children and large language models in word sense disambiguation: Insights and challenges

Francesco Cabiddu
University College London, UK

Mitja Nikolaus
CerCo, CNRS, France

Abdellah Fourtassi
Aix-Marseille Université, France

Abstract: Understanding how children process ambiguous words is a challenge because sense disambiguation is a complex task that depends on both bottom-up and top-down cues. Here, we seek insight into this phenomenon by investigating how such a competence might arise in large distributional learners (Transformers) that purport to acquire sense representations from language input in a largely unsupervised fashion. We investigated how sense disambiguation might be achieved using model representations derived from naturalistic child-directed speech. To this end, we tested a large pool of Transformer models, varying in their pretraining input size/nature as well as the size of their parameter space. Tested across three behavioral experiments from the developmental literature, we found that these models capture some essential properties of child word sense disambiguation, although most still struggle in the more challenging tasks with contrastive cues. We discuss implications for both theories of word learning and for using Transformers to capture child language processing.

Keywords: Child Word Sense Disambiguation; Transformers; Usage-Based Learning.

Corresponding author(s): Francesco Cabiddu, UCL Psychology and Language Sciences, University College London, 26 Bedford Way, London, WC1H OAP, UK. Email: francesco.cabiddu@ucl.ac.uk.

ORCID ID(s): https://orcid.org/0000-0001-9692-4897

Citation: Cabiddu, F., Nikolaus, M., & Fourtassi, A. (2025). Comparing children and large language models in word sense disambiguation: Insights and challenges. *Language Development Research, 5*(1), 35—115. http://doi.org/10.34842/k0rg-2n19

Introduction

Large language models are deep artificial neural networks pretrained on large unlabeled datasets via self-supervised learning. These models have had a great impact in the field of Natural Language Processing (hereafter NLP) for their performance in language understanding and generation tasks (e.g., Bommasani et al., 2022). Here, we examined the plausibility of these models as distributional learners posited by usage-based approaches of language acquisition (e.g., Ambridge, 2020; Bybee, 2010). We focused on child word sense disambiguation (Cabiddu et al., 2022b; Rabagliati et al., 2013). That is, how children use sense-specific representations (e.g., band = music band, elastic band). Specifically, we examined whether the distributional learning mechanisms that allow these models to acquire linguistic knowledge at the sentence and word level could give rise to word sense disambiguation skills that children exhibit in behavioral tasks.

We tested models based on the Transformer architecture (Vaswani et al., 2017) that perform sense disambiguation using sentence context to form contextualized representations. Transformers are sensitive to both bottom-up direct word-associations (a word co-occurring frequently with another across different sentences) and top-down syntactic and semantic sentence structures (e.g., Jawahar et al., 2019; Tenney et al., 2019) on which sense disambiguation depends. Here, we refer to these high-level sentence structures as top-down cues that a usage-based learner might acquire through language experience (Alishahi & Stevenson, 2013; Bybee, 2010). These refer to any abstract knowledge that might enable an individual to generalise a certain sentence structure to novel language instances (e.g., a child knowing that "pushing a flowerpot" is more plausible than "pushing a road" even without having heard either expression before; Andreu et al., 2013). Transformers' inherent sensitivity to top-down cues allow us to apply these models to raw naturalistic language, without having to enrich the input with external, explicit resources to provide sensitivity to such structures. For example, Alishahi and Stevenson (2013) showed how a computational learner could apply familiar verbs to novel object arguments. The model they developed was provided with various pieces of knowledge, such as the positions of syntactic arguments within sentences and the semantic characteristics of each argument. From this, it was able to generalize the prototypical semantic properties that an argument of a verb should possess (i.e., verb-event structures; for instance, "The mechanic warned the driver" is more plausible than "The mechanic warned the engine"). This finding is significant because it provides in-principle evidence that a structural aspect like verb-event structures can be bootstrapped from input. However, providing the extensive knowledge presumed to be available to the learner requires several input pre-processing steps (e.g., lemmatizing the input, identifying and recoding naturalistic sentences as verb frames, tagging semantic characteristics of each argument using an external dictionary). It remains unclear whether the same results could be achieved without such pre-implemented knowledge in the model, relying only on bootstrapping verb knowledge directly from the input. Moreover, when one wants to apply a model to raw, naturalistic language, it becomes infeasible to pre-process the input for several aspects of sentence structure that the model should be sensitive to in order to perform certain tasks, such as word sense disambiguation.

Transformers have been used to form adult-like sense representations in natural language classification tasks, and the models have been tested on their ability to pick out a target sense given the sentence context (Loureiro, et al., 2021). However, such tasks may not suitably assess model developmental plausibility as they use coherent test sentences (i.e., all cues in the context unambiguously point toward one target sense). Relying on these tasks makes it difficult to differentiate whether Transformers exhibit rather adult-like or child-like performance, as both adults and children have been shown to perform well at disambiguating coherent sentences (e.g., Khanna & Boland, 2010; Rabagliati et al., 2013). Thus, the goal of the current study is to test models on contrastive tasks alongside coherent ones. Contrastive tasks put bottom-up (i.e., word associations) and top-down sentence cues in competition. They represent a more suitable test of developmental plausibility because differences exist in how children and adults behave in such tasks. In fact, in sense disambiguation children rely more on bottom-up aspects of sentence context (e.g., word associations) than adults, with less reliance on top-down cues likely due to differences in language experience or slow cognitive maturation (Khanna & Boland, 2010; Rabagliati et al., 2013).

Previous studies in NLP have computed models' representations based largely on adult language (Loureiro et al., 2021, 2022). These representations are created by using a technique that computes an average representation of a word sense given a collection of sentences (e.g., a prototypical representation of a music band). Here we apply this technique to evaluate how properties of child sense processing could be captured using sense representations formed from naturalistic *child-directed* utterances. This choice is motivated by the fact that differences in how senses are assigned to words in children and adults is likely influenced by differences in word use in child and adult environments (Meylan et al., 2021). We note that this method does not involve pre-training the models on child-directed language, although we do also include a family of models pre-trained on child-directed utterances. We show that computing child-directed sense prototypes has different benefits for capturing child performance, but we also return to its limitations in the Discussion.

We evaluated Transformers using behavioral studies that tested 4-year-old children's abilities to use bottom-up (word associations) and top-down (sentence global plausibility, verb-event structure) cues to sense disambiguation (Cabiddu et al., 2022b; Rabagliati et al., 2013). We tested a large pool of models ($N = 45$) from 14 different families. This integrative approach (see also Schrimpf et al., 2021) would allow us to study how different properties of the models may lead to different behavioral patterns, while relying on a single model could be misleading as any conclusion might be influenced by idiosyncratic aspects of this specific model (architecture, pretraining objectives, amount/type of pretraining input, etc.). Specifically, we explored how scalability in models' size (number of parameters) and pretraining data size related to sense disambiguation performance. It has been shown that increasing the number of model parameters improves models' ability to generalise, enabling them to tackle a broad spectrum of language and reasoning tasks without necessitating extensive examples during training or specific model fine-tuning (e.g., Brown et al., 2020; Chowdhery et al., 2022). Essentially, more parameters in language models means a greater capacity to store patterns and nuances from the training data. This capacity to capture a wider array of linguistic patterns may lead to improved performance in tasks such as sense disambiguation, where understanding context and subtle differences in meaning is crucial. Based on findings about word age of acquisition norms (Laverghetta Jr & Licato, 2021), we expected models with a larger number of parameters to better fit child data, also in line with NLP studies showing how increasing a model's parameter count improves its ability to track both bottom-up and top-down aspects of sentence structure (Devlin et al., 2019; Hewitt & Manning, 2019; Radford et al., 2019). Similarly, better performance and generalisation abilities can be achieved by training models on larger and more diverse datasets (e.g., Raffel et al., 2023). Training models on linguistic contexts that encompass a wide range of topics, styles, and structures increases the opportunities to abstract general schemas from the linguistic examples observed. Nevertheless, there is also evidence of small (i.e., more realistic) pretraining input being enough to align models to adult neural data and reading comprehension scores (Hosseini et al., 2022), therefore we might expect a null effect of pretraining size when attempting to capture human performance.

In summary, both model size and pre-training size are dimensions that have been linked to models' generalisation abilities. This capacity is crucial for learning top-down sentence structures that can then be generalised to new linguistic instances, which is something we focus on in our study. In the following, we first introduce evidence of child sense disambiguation. Secondly, we discuss the theoretical significance of Transformers and introduce a recent framework for evaluating models in sense disambiguation.

Child Word Sense Disambiguation

Sentence context plays a significant role in sense disambiguation (e.g., Sophia [played in / twisted] a band). Children use a similar (though lower) diversity of senses in naturalistic conversations (Meylan et al., 2021), which raises a question about which sentence properties facilitate child word disambiguation (Cabiddu et al., 2022b; Hahn et al., 2015; Khanna & Boland, 2010; Rabagliati et al., 2013). Children should access cues at different linguistic levels to successfully disambiguate senses. Here, we focused on key studies that showed that 4-year-old children could use both bottom-up and top-down disambiguation cues, although to different degrees depending on the specific cue. Table 1 shows an overview of the three experiments we consider. A general goal across experiments was to test children's sensitivity to sentence context for sense disambiguation. Further, they tested if top-down cues (global plausibility, verb-event structures) played a role beyond bottom-up word associations (when the two types of cues are in direct competition). Similarly, here we investigate if Transformers could use sentence context for word sense disambiguation like children, and if they would demonstrate comparable sensitivity to top-down cues in contrastive conditions.

Table 1. ***Behavioral experiments. Target words are shown in bold. Underlined text indicates cues to the dominant sense "elastic band", while italicized text refers to cues to subordinate "music band". The Dominant selection column indicates average dominant sense selections in children, for dominant-plausible (underlined) and subordinate-plausible conditions (italicized).***

Study	Cue type	Example	Dominant selection
(Rabagliati et al., 2013) Exp 1, Coherent cues	Prior Context	Dora [looked in her drawer / *heard some music*]. The **band** was cool	79% / *33%*
	Current Context	Dora was in her room. She [stretched / *listened to*] the **band**, which was cool.	81% / *38%*
(Rabagliati et al., 2013) Exp 2, Contrastive cues	Global Plausibility	Elmo and his class were singing songs. The teacher could play music with [anything / *anyone*], even a **band**.	39% / *21%*
(Cabiddu et al., 2022b) Contrastive cues	Verb-Event Structure	Sophia listened to some music. Then she [twisted / *played in*] a **band**.	62% / *26%*
	Verb-Lexical association	Sophia listened to some music. Then she [got / *played in*] a **band**.	60% / *26%*

The behavioral studies we considered have not only been used to test children's disambiguation skills at a certain point in development but also to examine different hypotheses on whether young children can rely on the same cues for sentence parsing as adults do, or whether there are limitations in their access to certain cues that require higher levels of linguistic analysis. One account posits that children rely solely on bottom-up cues in sentence parsing (Snedeker & Yuan, 2008), while another account emphasizes cue informativity (Trueswell & Gleitman, 2007). In the context of sense disambiguation, an informativity account would suggest that children gradually refine their estimation of the general reliability of each cue (whether bottom-up or top-down) in determining word meaning as they grow. Such gradual fine-tuning could account for the differences in how children and adults perform word sense disambiguation.

The evidence available so far supports an informativity account, showing that children rely on both top-down and bottom-up cues. However, their use of top-down cues is contingent on the strength of that cue's influence in the child's early processing. For instance, children primarily rely on bottom-up word associations instead of using top-down global plausibility at the discourse level, which is the strategy predominantly used by adults (Rabagliati et al., 2013). This likely occurs because word associations are a cue that is consistently present in children's language input, and they can use this cue from very early in development. Nonetheless, this does not imply that children cannot use top-down cues. In fact, when considering a top-down cue that children also consistently use in sentence and word processing from early in development, such as verb meaning, they indeed demonstrate the ability to rely on this cue in sense disambiguation over bottom-up word associations (Cabiddu et al., 2022b).

In all studies, children heard short stories ending with a target word and saw four pictures. Two depicted the target word's alternative senses: One frequent in child-directed speech (dominant = elastic band) and one less frequent (subordinate = music band), with a 3:1 frequency ratio. The other two pictures depicted semantic distractors (e.g., sock, sport team). After the story, children chose the picture that best matched the story's final word.

In a first experiment, Rabagliati et al. (2013) tested if children could use sentence context to disambiguate dominant and subordinate senses. Disambiguation cues were presented in a previous sentence (Prior context), or in the same sentence as the target (Current context). Example stimuli are shown in Table 1. Children showed successful disambiguation across conditions, selecting more dominant senses (above 50% chance) in dominant-plausible conditions, and more subordinate senses in subordinate-plausible conditions (i.e., less than 50% dominant selections).

However, in this experiment, children could have relied solely on bottom-up associations. For example, in *Dora was in her room. She stretched the band,* one could track the association between *stretching* and *elastic band* in naturalistic conversations without processing sentence structures (i.e., using verb-event knowledge to infer that stretchable entities are usually objects). In the second experiment from Rabagliati et al. (2013) and in the experiment from Cabiddu et al. (2022b), bottom-up and top-down cues were in competition. Stories always began with a prior context containing word associates of the target subordinate sense. As shown in Table 1, prior contexts contain the words *music* or *songs* pointing toward the subordinate *music band.* Further, in experimental conditions, stories ended with top-down cues pointing toward the opposite dominant sense *elastic band* (see underlined cues in Table 1).

In Rabagliati et al. (2013) experiment 2, experimental stories shifted global semantic plausibility toward the dominant sense. Children struggled to use global plausibility and relied heavily on word associations (39% dominant selections, below chance). In other words, children struggled to use real-world knowledge, which facilitates the comprehension of causal relations, event sequences, and social norms conveyed by the overall discourse. For example, when interpreting a sentence like *Elmo and his class were singing songs. The teacher could play music with anything, even a band* the listener would need to infer that any object could emit sound and therefore, could potentially be used as a musical instrument. In contrast, children relied mostly on bottom-up word associations (i.e., tracking co-occurrences between words) to perform shallow processing of sentence context when interpreting ambiguous words (i.e., mostly interpreting *band* as a *music group* because of its association with the words *singing, songs,* and *music*).

Still, a significant difference from a control condition emerged (21% dominant selections when the story fully supported the subordinate; see italicized cue in Table 1). This result indicated residual sensitivity to top-down global plausibility in 4-year-old children.

The study by Rabagliati et al. (2013) also highlighted the limitations of capturing adults and children's reliance on top-down cues when using a distributional computational learner that is uniquely based on tracking bottom-up word associations. They employed a bag-of-words Bayesian classifier, trained on child-directed speech, to simulate children's performance in both non-contrastive and contrastive tasks. They found that while the classifier could successfully resolve non-contrastive tasks and capture variations in child performance (experiment 1), it failed in contrastive tasks (i.e., performance at floor in experiment 2, with 0% dominant senses selected across conditions), likely due to its inability to incorporate sentence-level top-down cues in its word representations. Here, we aim to examine whether a distributional learning Transformer architecture, which has shown sensitivity to top-down sentence-level structure, could instead succeed in capturing child disambiguation performance in contrastive tasks.

Cabiddu et al. (2022b) focused on verbs. Verbs are likely to represent a particularly valid cue that young children can rely on when processing sentences and words. For example, verbs' syntactic arguments guide 3- to 5-year-old children's interpretation of ambiguous sentences (e.g., Kidd & Bavin, 2005; Snedeker & Trueswell, 2004; Yacovone et al., 2021). Further, the semantic restrictions that verbs impose on their arguments (i.e., verb-event structures) guide children's unambiguous word processing (Andreu et al., 2013; Mani et al., 2016). For example, 3-year-olds know that *pushing a flowerpot* is more plausible than *pushing a road* even if they have never heard either expression (Andreu et al., 2013).

As shown in Table 1, in a Verb-Event condition, stories ended with verbs that never co-occurred with dominant senses in naturalistic conversations (i.e., children never or rarely hear *twisting a band*, which controls for verb-object associations). However, the verbs' event structure only accepted the dominant senses (i.e., one can only twist an elastic band, not a music band), making it the only available cue.

Further, the researchers examined the effect of verb-object associations (see Verb-Lexical condition in Table 1): Verbs had a neutral verb-event structure (e.g., one could get either an elastic or music band), but often co-occurred with dominant senses in naturalistic conversations (i.e., children frequently hear *getting an elastic band*). Given the role of verb-object associations in children's word processing (Mani et al., 2016), this condition tested if children would weigh more word associations coming from a verb than the rest of the (prior) context.

Children successfully resolved dominant senses using both verb-event structures and verb-object associations, beyond bottom-up word associations from prior contexts.

Overall, results from these behavioral experiments show that children can rely on different bottom-up and top-down cues for sense disambiguation. However, it remains unclear which learning *mechanisms* might underlie these competencies. Below, we use Transformers as a scientific tool to test the extent to which purely distributional learning mechanisms account for the acquisition of word sense knowledge that is dependent on sentence context.

Word Sense Disambiguation in Transformers

Testing a usage-based learner requires an architecture that forms top-down abstractions while accounting for effects of bottom-up statistical cues in language development (e.g., Ambridge et al., 2015; McCauley & Christiansen, 2019; Saffran et al., 1996). Consider the meaning of *table* in Ambridge (2019). A fixed top-down rule defining a *table* category (e.g., has legs; used for eating; made of wood, metal, or plastic; waist height) becomes falsifiable by counterexamples (e.g., an empty barrel used as a table at a bar). A solution is to embed specific contexts in the *table* representation (Ambridge, 2020; Srinivasan & Rabagliati, 2021). Bottom-up context-dependent information allows the child to estimate the similarity between a new instance *barrel table* and previously encountered *tables*. This recursive process of estimation facilitates the emergence of a context-independent, fuzzy, and probabilistic category of *table* (i.e., a prototype). In sense disambiguation, context-dependent and context-independent representations could gradually lead to multiple sense categories for a single word (Srinivasan & Rabagliati, 2021), with clusters of instances sufficiently separated in the semantic space (e.g., an object band prototype, a music band prototype).

The way sense representations are conceptualized in these proposals of word sense acquisition aligns with the ideas proposed in accounts of word sense processing (Duffy et al., 2001; Rodd, 2020). For instance, the recent semantic-settling account (Rodd, 2020) assumes that word senses are stored in a lexical-semantic space as high-dimensional representations. Distinct senses of a word form are represented as different paths embedding a set of dimensions or features that define the mapping between the word form and each sense. During sentence parsing, a settling process guides access to specific word senses by increasing the activation of specific paths in the lexical-semantic space. This activation depends on multiple cues at the word and contextual levels, helping the system settle on one sense, from bottom-up cues (e.g., meaning expectation based on words frequently co-occurring in the sentence context) to top-down cues (e.g., real-world knowledge used for pragmatic inferences). Computational evidence supporting this processing account largely comes from adult disambiguation studies (e.g., Rodd et al., 2004). However, it is still unclear whether its predictions can extend to child processing.

The above ideas of context-dependent sense representations align with Transformers' core self-attention mechanism. For each token, these models construct distinct representations that dynamically integrate sentence context. Although children have access to referential and social cues beyond sentence context, using Transformers is useful to answer the question: *How far can a distributional learner that uniquely processes naturalistic sentence context go?*

After training, Transformers encode generalized (context-independent) knowledge. Tokens from different senses organize into separate clusters within model layers, reflecting the organization of senses in dictionaries and adult representations (Loureiro et al., 2021, 2022). In Loureiro et al. (2021), Transformers were evaluated using a nearest neighbor approach (e.g., Melamud et al., 2016; Peters et al., 2018). This uses sense-annotated corpora to create model sense prototypes by averaging the representations of a collection of tokens belonging to a specific sense (see Method). Sense prototypes are then used to evaluate the model disambiguation at test. Using this method led to a Pearson's correlation of .9 between the best model and adult annotators. This method is useful because it investigates knowledge of models that are not pretrained on disambiguation, but only on predicting a word given its context (which should be more in line with what children do). Further, compared to previous studies (Haber &

Poesio, 2020), Loureiro et al. (2021) showed that models' performance better aligned with adults' when a reference sense-annotated corpus reflected the coarse-grained knowledge that adults have (e.g., collapsing senses that adults likely do not distinguish, but that are differentiated in a dictionary). This suggests that it is possible to tailor the models' sense prototypes to a specific population. In our work, reference sentences were transcribed child-directed utterances, reflecting children's naturalistic input and containing senses known to 4-year-olds based on behavioral evidence.

Method

Models

We used 13 Transformer-based language model families with varying training tasks and input encoding mechanisms. We also included a bidirectional recurrent neural network (ELMo, Peters et al., 2018), which achieved state-of-the-art results in sense disambiguation before the introduction of Transformers (e.g., Wiedemann et al., 2019). Model descriptions can be found in Appendix S1. We also share materials and code to reproduce the study results on our GitHub page (https://doi.org/10.5281/zenodo.8200803). In various configurations within families, we varied model size (number of million parameters, *M* = 287, *range* = 8 - 1,630) and pretraining size in gigabytes of text (*M* = 103, *range* = .005 - 806). In Appendix S3, we also include results from models with randomly initialized weights, showing that performance differences were not due to architectural differences in connection patterns among units.

Model Evaluation via Nearest Neighbor

Following Loureiro et al. (2021), we extracted sense prototypes using annotated sentences (see Corpora for details) in which a word occurred in a specific sense (e.g., *elastic band* in "*when we put the rubber bands around it then we'll put your name on it so we'll know which one belongs to who*"). We extracted a model's contextualized vector for each sense occurrence, summing the last four layers. For models that work at the subword level, we first averaged representations of subword tokens for the target word. Finally, we averaged the word vectors to obtain a centroid representing the *elastic band* prototype. We repeated the process for the alternative *music band.*

In Appendix S2, we also repeat the sense prototype extraction with different random samples of sentence exemplars to provide evidence that using a Nearest Neighbor approach is not heavily dependent on the specific set of exemplar sentences we used for each target sense. This decreases the concern that results from our simulations might be related to the quality of the prototypes rather than the model representations of sense usage.

To evaluate model performance, we extracted a contextualized vector for each test sentence's target word. We used cosine similarity to compare each vector with the two prototypes representing the dominant (e.g., *elastic band*) and subordinate (e.g., *music band*) senses. The most similar prototype determined the assigned sense for the test word. We then transformed this binary measure (*Dominant* = 1, *Subordinate* = 0) into a continuous measure by computing the percentage of dominant senses assigned in a specific condition (matching the child outcome measure in Table 1).

Corpora

We took sentences for computing prototypes from ChiSense-12 (Cabiddu et al., 2022a), which contains speech directed to children up to age 4 from the English section of the CHILDES database (MacWhinney, 2000). Each sentence was tagged for occurrences of 12 ambiguous words in dominant or subordinate senses (e.g., *chicken animal*, *chicken food*). The selection of dominant and subordinate senses within the corpus drew from those used in the experiments conducted by Rabagliati et al. (2013). This approach guaranteed that the chosen senses are familiar to children, as evidenced by their performance in experimental tasks. We used 9 words, excluding homophones with different spelling (e.g., son/sun) for which no ambiguity exists as the models process orthographic input. We also tagged 4 new words to cover more items from children's experiments. The target words used were all concrete nouns: band (binding or fastening object / music group); bat (animal / sports equipment); bow (knot / weapon); button (device to control electronic operations / fastener on clothing); chicken (animal / meat); glasses (eyewear / drinking vessels); letter (alphabetical symbol / mailed communication); line (geometric line / sequence of people or things arranged one behind the other); nail (body part / metal fastener); fish (animal / meat); lamb (animal / meat); turkey (animal / meat); card (playing card / greeting card).

Details about items and annotation process are in Appendix S2. The final corpus had 15,901 sentences for 13 target words, with dominant senses appearing 69% of the time on average (3:1 dominant/subordinate ratio).

Comparing Child to Model Performance

We computed an optimal outcome measure comparing child and model performance. We examined if the models exhibited a dominant sense bias reflecting the dominant/subordinate ratio in the input. For experiment 1 in Rabagliati et al. (2013) with non-contrastive cues, we fitted a linear mixed-effects model using the percentage of dominant senses selected by each model as the outcome, and model size and pretraining size as the predictors. Model family was used as random effect intercept. The model output is reported in full in Appendix S4. Only pretraining size negatively predicted dominant selection (β = -1.53, *95% CI* = [-2.30, -.75], p < .001), but not model size (β = -1.47, *95% CI* = [-3.01, .08], p = .062). As shown in Figure 1, the models better approximated the 69% dominant sense bias as pretraining size decreased.

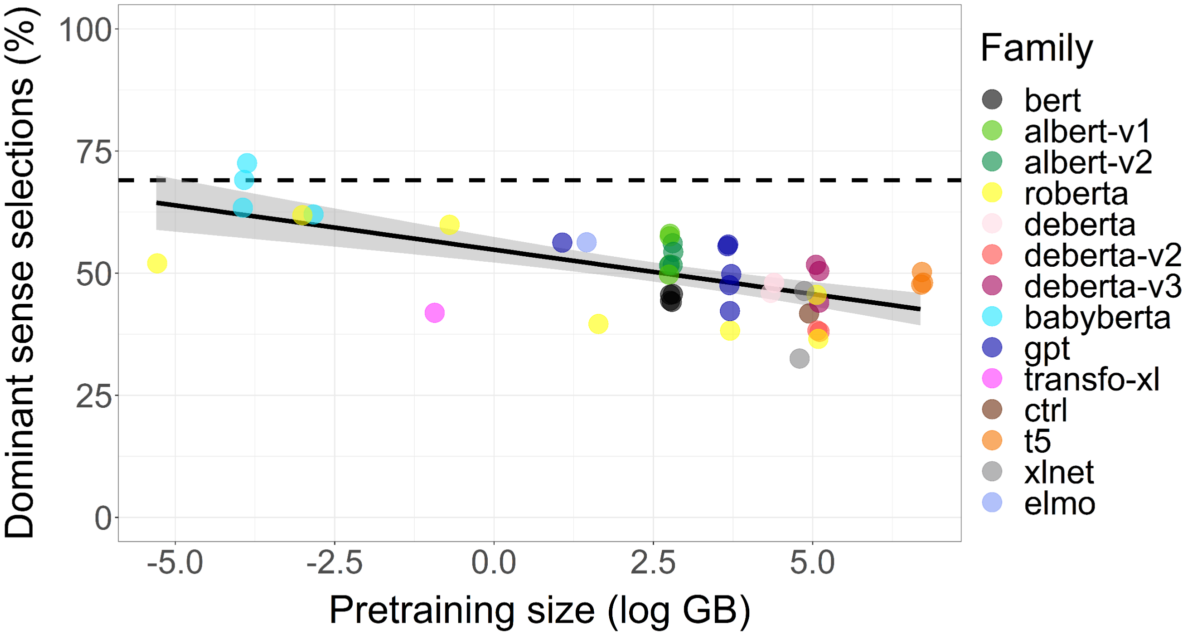

Figure 1. ***Percentage of dominant senses selected by each model in Rabagliati et al. (2013) experiment 1, by pretraining size in log GB. The dashed horizontal line indicates dominant sense prevalence in ChiSense-12.***

To further confirm that the dominant sense bias was produced by the employment of child-directed input, in Appendix S6 we also examined the models' dominant sense preference using sense prototypes computed from adult-directed speech (from utterances included in the British National Corpus; BNC Consortium, 2007). When we used adult sense prototypes, the models never approached the 69% dominant sense bias, showing equal preference for dominant and subordinate senses (50% dominant sense selections). Overall, these preliminary investigations on the effect of input speech on sense representations indicate that the use of child-directed input aligns models with children's representations of sense frequencies in naturalistic speech.

Dominant sense bias is one of the variables that can influence word disambiguation. It is an important aspect of how children disambiguate words, as well as being crucial in a model learner. However, it is not the primary focus of our examination. We aim to determine whether models are successful because they resolve the meaning of an ambiguous word using the context of the surrounding sentence, rather than from the frequency of the sense itself. For this reason, the contribution of sentence-level factors (e.g., verb information) was disentangled, for example, in Cabiddu et al. (2022b), from the contribution of word-level information (sense dominance) by statistically controlling for the latter. This was done to test whether children were indeed using verb information to resolve ambiguities. We adopted a similar approach with Transformers, by effectively separating the contribution of word-level information from that of sentence-level information. In other words, differences in dominant sense bias pose a confound: A model pretrained on a small corpus might select a similar percentage of dominant senses to children not only due to context cue sensitivity, but also because it prefers dominant senses more than a model pretrained on a large corpus. We controlled for this confound by examining the relative difference in dominant sense selections between dominant-plausible and subordinate-plausible conditions. In Appendix S5, we also include analyses that examine which models better capture children's performance when all levels (sentence-level and word-level) are considered. We return to these additional results in the Discussion.

We use relative differences in performance to control for the effect of dominant sense bias. For example, in the first experiment, children selected dominant senses (e.g., *elastic band*) in 81% of trials in the dominant-plausible condition (*She stretched the band*) and 38% in the subordinate-plausible (*She listened to the band*). For a relative difference of 81% - 38% = 43% in children, a model with 60% - 17% difference and one with 80% - 37% were considered equally similar to children. Essentially, the relative difference focused on a model's sensitivity to shifts in sentence context and compared it to children's sensitivity. The final outcome measure estimated the distance between model and children (e.g., [60% – 17%]) – [81% – 38%]), with values of 0 indicating equal sensitivity in the model and children, and values lower and higher than 0 indicating lower and higher sensitivity, respectively. Using this measure of relative distance as the outcome, we performed model comparison for each experiment between multiple nested linear mixed-effects models, which are reported in full in Appendix S4. We examined the main and interaction effects of model size and pretraining size, and employed model family as a random effect intercept in every statistical model.

Results

Rabagliati et al. (2013) - Experiment 1

Figure 2 shows models' performance by model size (2a) and pretraining size (2b). Some models reached child baseline (y = 0), while others performed worse (y < 0) or better (y > 0). The best linear mixed-effects model indicated higher context sensitivity as model size increased (β = 5.36, *95% CI* = [2.07, 8.64], p = .002) and pretraining size increased (β = 3.81, 95% CI = [2.16, 5.47], p < .001). A main effect of condition (β = -9.98, *95% CI* = [-16.18, -3.78], p = .002) showed models performing better in the current-context condition, which may not align with child performance. Although the main effect of condition was not tested in the child experiment, children's average scores might suggest similar sensitivity to prior and current context (see Table 1).

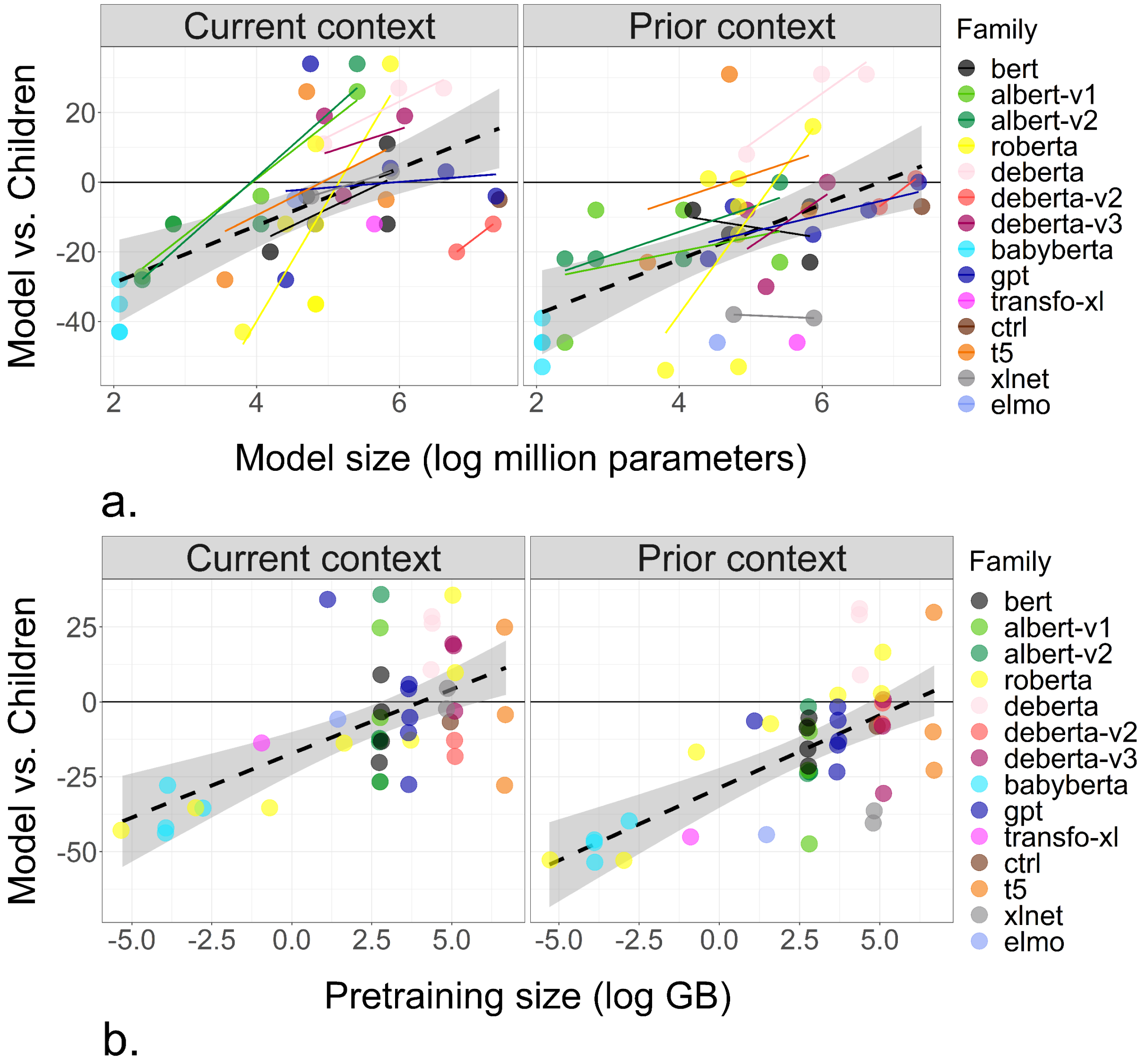

Figure 2. ***Models' relative distance from children by model size (a) and pretraining size (b), in current and prior context conditions. Model families are shown in the legend. The black horizontal line indicates child performance. The dashed regression line with 95% confidence interval shows performance across models. Colored regression lines are also shown for each model family, although only when***

examining model size as there is almost null variation in pretraining size within family. Points in panel b are jittered by 2 points in the y axis to facilitate visualization of overlapping points.

Rabagliati et al. (2013) - Experiment 2

This task used contrastive bottom-up and top-down cues, which most models seemed to struggle with: Figure 3 shows a floor effect, which led to null effects of model size ($\beta = 3.37$, *95% CI* = [-.35, 7.09], $p = .075$) and pretraining size ($\beta = 0.12$, *95% CI* = [-1.74, 1.98], $p = .895$). As confirmed in Appendix S4 (see plots showing raw dominant selection scores for each model), the floor effect led to only few models showing a difference in dominant selection between conditions. This aligns with children's residual sensitivity to top-down cues, as they displayed a difference between conditions despite low selection rates. Nevertheless, most models performed worse than children, suggesting an overall difficulty in managing contrastive cues.

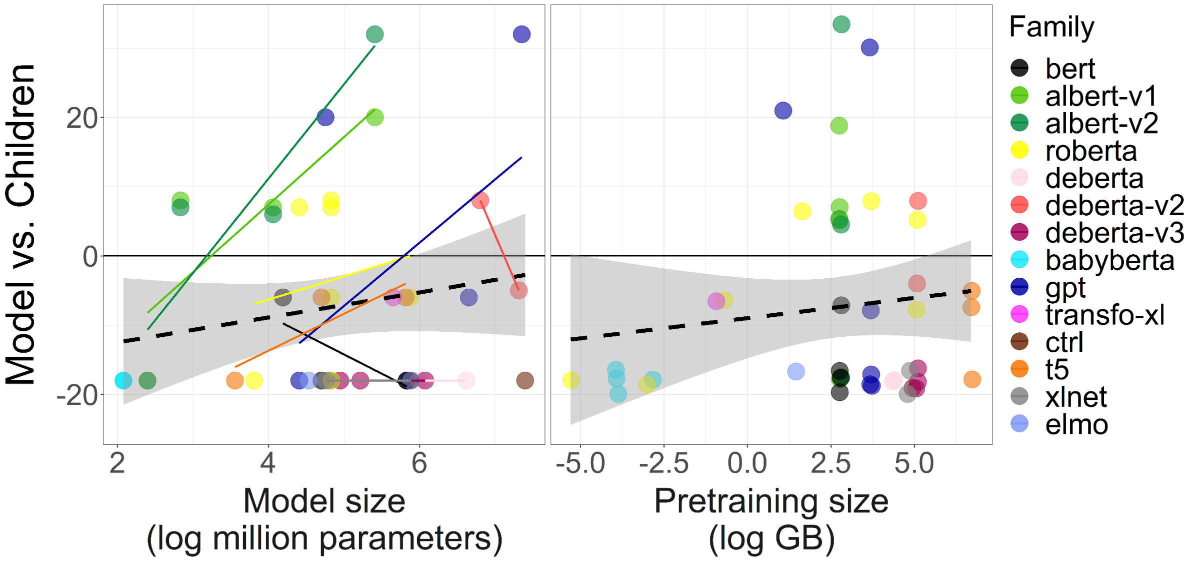

Figure 3. ***Models' relative distance from children by model size and pretraining size, in Rabagliati et al. (2013) experiment 2.***

Cabiddu et al. (2022b)

The models better handled contrastive bottom-up and top-down cues in this task, resembling the strong role of verbs in child processing. The models showed higher sensitivity to verbs with a strong event structure (Figure 4a; e.g., *She twisted a band*), with model size being positively related to models' sensitivity to verb-event cues (β = 7.57, *95% CI* = [3.48, 11.67], *p* = .001), but not pretraining size (β = -.30, *95% CI* = [-2.35, 1.74], *p* = .765). Instead, sensitivity was lower to verbs that were only lexically associated with the dominant sense (Figure 4b; e.g., *She got a band*), with no significant effects of model size (β = 1.73, *95% CI* = [-0.87, 4.34], *p* = .186) or pretraining size (β = 0.16, *95% CI* = [-1.14, 1.45], *p* = .809).

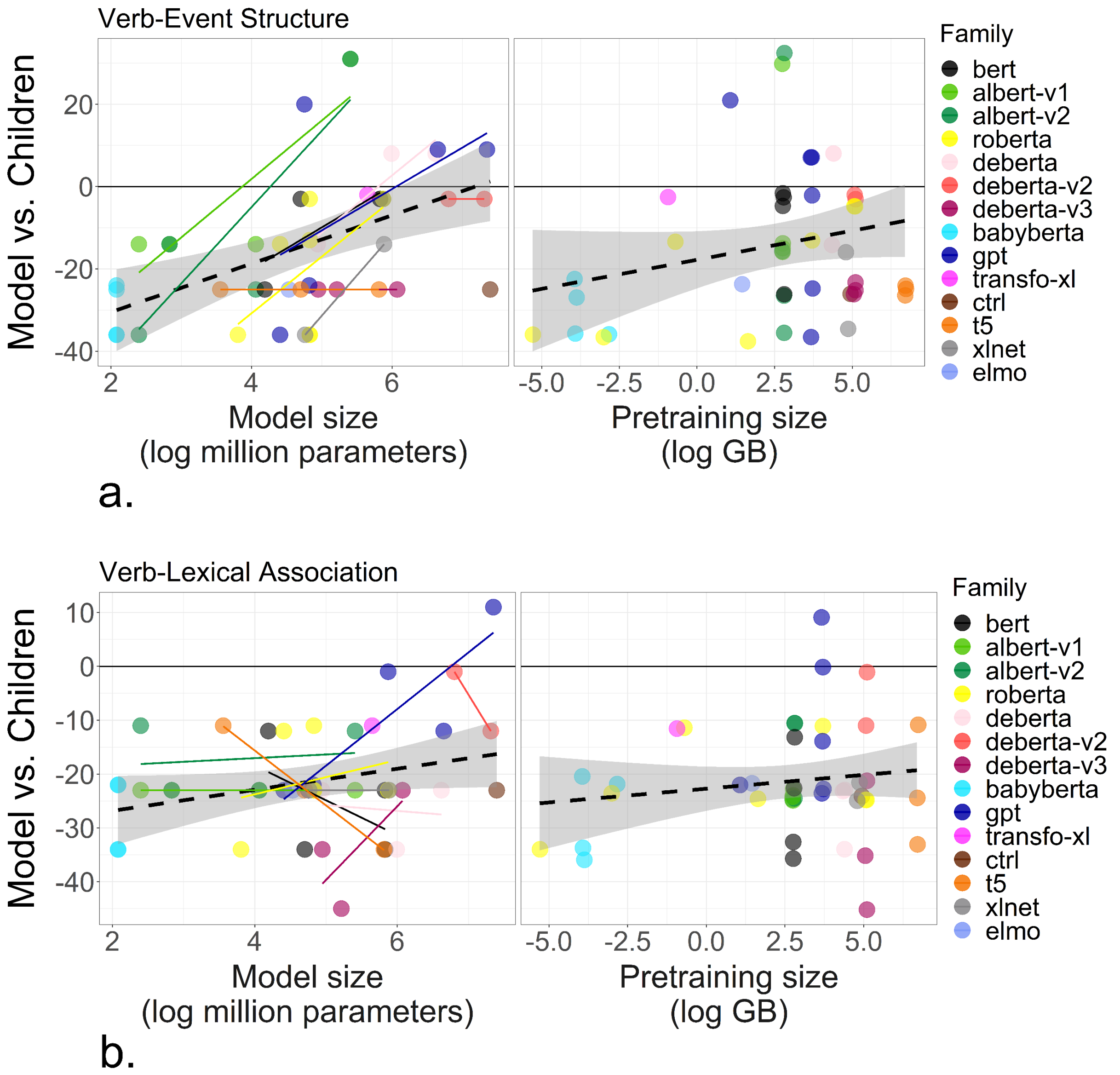

Figure 4. ***Models' relative distance from children by model and pretraining size, when examining performance at the verb-event (a) and verb-lexical conditions (b).***

Discussion

We examined the capabilities of large Transformer models in capturing child word sense disambiguation. Our results support the idea that children, like these models, might be usage-based learners who bootstrap word knowledge from the naturalistic environment (Bybee, 2010), and that child sense knowledge can, in principle, arise from probabilistic representations embedding context-dependent and context-independent information (Ambridge, 2020; Rodd, 2020; Srinivasan & Rabagliati, 2021). In line with a cue informativity account of child processing (Trueswell & Gleitman, 2007), Transformers captured the changes in word sense disambiguation performance observed across child behavioral experiments. Coherent tasks were resolved with greater ease, and performance on contrastive tasks was found to be dependent on the type of top-down cue provided (i.e., as observed in children, verbs provided a better facilitation for sense disambiguation than global plausibility).

In line with Laverghetta Jr and Licato (2021), larger models were more sensitive to both coherent (Figure 2) and contrastive cues (Figure 4a), likely because they form more precise representations based on both bottom-up and top-down aspects of sentence structure (Devlin et al., 2019; Hewitt & Manning, 2019; Radford et al., 2019).

Contrary to our prediction, models trained on larger corpora were more sensitive to coherent cues (Figure 2), while we found the predicted null effect of pretraining for contrastive cues (Figure 3 and 4). In coherent sentences, a model can rely on both word associations and top-down cues, with more pretraining likely increasing sensitivity to both. However, more pretraining might not always be as valuable for resolving *contradicting* bottom-up and top-down cues in the other conditions. Larger models might instead have an advantage in this regard.

Further, a visual inspection of models' performance at contrastive tasks (see raw plots of dominant sense selection for each model in Appendix S4) showed a stronger overall preference for subordinate senses across conditions compared to children, which might indicate models' higher sensitivity to prior context word associations (an analysis of relative differences could not highlight this, as it specifically controls for absolute differences in sense selection). In a follow-up analysis (see Appendix S5), we found evidence for this interpretation. We used an alternative outcome measure (Euclidean distance) which, compared to the relative difference, additionally looked at how close models got to $y = 0$ (Figure 2, 3, and 4) and at the exact match between models and children (i.e., difference in absolute scores): Given 81% - 38% as the children's response difference, a model performing 80% - 37% would be now closer to children than one that performs 60% - 17%. This measure might suffer from dominant sense bias (Figure 1), which we included as covariate in the statistical models to control for its effect. We replicated the positive effect of pretraining size in experiment 1 (Figure 2b), and found a *negative* effect of pretraining size in the verb-event structure condition of the third experiment (Figure 4a).

This result might indicate that smaller pretraining prevented an extreme sensitivity to word associations, allowing models to find the right balance between bottom-up and top-down cues. Interestingly, the best models in this condition received pretraining that was judged as psychologically plausible in previous studies (100 million tokens, Hosseini et al., 2022), although for an older population than ours (10-year-olds). To gain deeper insights into word association sensitivity, future work should explore how pretraining size influences the ability of large language models to track word associations and whether smaller, more realistic input can better capture children's sensitivity to these associations. Ideally, to answer this question, one would need access to the original corpora used for pretraining, which is not always possible. This would enable an understanding of precisely what types of word associations the models might have encountered during pretraining. Some recent investigations have revealed that sensitivity to word associations begins to decrease at around 1 billion tokens of input (Zhang et al., 2021). This finding might suggest the necessity to scale down to a much smaller input to avoid extreme sensitivity to bottom-up cues and to better align with child performance.

Only models with small pretraining approximated the dominant sense bias in the child input (Figure 1), and only few models (Figure 4b) showed sensitivity to verb-sense associations (e.g., *get-elastic band*), which are idiosyncrasies of the child input. One way to better align models with the child environment would be pretraining directly on child input (Hosseini et al., 2022; Warstadt & Bowman, 2022). This would also enhance the psychological plausibility of the models, which are currently pretrained on vast amounts of input, often sourced from unknown corpora and adult-directed written language. However, this task is limited by the lack of sufficiently large corpora. For example, in our study we included BabyBERTa (Huebner et al., 2021), which despite being pretrained on child input showed no sensitivity to sentence context, likely due to its small pretraining (5 million tokens). To address this gap, there is an ongoing effort within the research community to optimize model pretraining given an input limited in size, aligning more closely with human development (Warstadt et al., 2023). Model optimization also means that researchers will be able to examine and manipulate more fine-grained model dimensions than those we have considered (number of epochs, learning rate, batch size, etc.), allowing researchers to work with architectures that are likely to better approximate child learning and processing. Manipulating aspects of models' architecture will also give the opportunity to causally test their impact on the model's ability to capture child performance. For example, ablation analyses (e.g., removing parts of the model such as layers, attention heads, or specific weights) can be used to uncover necessary language knowledge within the models for successful task performance, generating hypotheses about language representations. Additionally, public release of the datasets used for training optimization will enable researchers to directly test the causal effect of input characteristics (Frank, 2023). Models can serve in controlled experiments to isolate pretraining inputs that enable effective disambiguation, offering insights into sentence-level factors that might assist children in developing word sense proficiency.

Models' performance was impaired in tasks that introduced contrastive cues (Figure 3 and 4). This suggests that this area requires further investigation, despite previous results showing that Transformers approximate adult performance in annotating word senses (Loureiro et al., 2021) or judging the semantic relatedness between word senses (Nair et al., 2020) when tested on non-contrastive sentences. Sense prototypes based on child input might have contributed to the low performance of the models in our study. In additional analyses presented in Appendix S6, we replicated all the simulations in the study using sense prototypes based on sense-tagged utterances from adult-directed speech. Specifically, we used utterances from the spoken part of the British National Corpus (BNC Consortium, 2007). We found that adult-based and child-based models produced similar percentages of correct responses in every experiment. Further, when we related models' performance to child responses, we found that child-based prototypes more closely aligned models with child performance in coherent tasks (Rabagliati et al., 2013; Study 1), but no difference was found at capturing child responses between models using child and adult sense prototypes in contrastive tasks (Rabagliati et al., 2013; Study 2; Cabiddu et al., 2022b). Overall, these supplemental results indicate that the low model performance at contrastive tasks was not due to a lack of richer linguistic cues that adult utterances might contain. The fact that the models performed poorly in tasks involving contrastive cues, even when the sense prototypes were derived from adult-directed speech, stands in contrast to the many linguistic feats of large language models (e.g., Gammelgaard et al., 2023).

Given that previous studies have not used contrastive tasks, one possibility is that such tasks might simply be difficult for models. Few models were sensitive to contrastive cues (Figure 3 and 4), indicating that at least some information about top-down structures might be captured from sentence context via distributional learning. However, overall models' performance was lower than children's. This occurs even if the task proposed to children might be more challenging than what the models faced. In fact, the models were only required to disambiguate between two alternative senses of each target word. However, other potential senses of a target word exist in dictionaries and may be known to children (e.g., for "band", not just "elastic band" and "music band", but also a “band” of bad weather). We would expect the models' ability to distinguish between word senses to deteriorate when considering a wider array of alternative senses. This is supported by Loureiro et al. (2019), which demonstrated that collapsing some of the senses in WordNet, that might not be distinguished by adults, improved Transformers' performance in word sense disambiguation.

Difficulties in approximating child knowledge could be due to the fact that children's representations of top-down structures are not only based on sentence context but also include real-world knowledge, which would need to be integrated into neural systems and could lead to abstractions more akin to human cognition (e.g., Pavlick, 2023). Specifically, while language models are capable of forming knowledge about direct word associations (bottom-up knowledge) and syntactic and semantic structures (top-down knowledge), it is crucial to acknowledge that the models' top-down generalisations about language patterns—though reflective of a form of understanding or knowledge—remain purely derived from textual patterns. For example, the models may solve experimental tasks (e.g., "Sophia listened to some music. Then, she twisted a band") by leveraging indirect associations between words—such as "twist" being associated with "bend" and "pull"—or by linking verbs to various objects (e.g., "twist" with "scarf" or "knob"), using these patterns as proxies for top-down inferences. This process enables language models to abstract semantic properties from the verbs and apply these properties to new contexts or objects that they have not explicitly encountered in their training data. The model's reliance on indirect associations to infer word meanings or predict plausible word combinations exemplifies a form of semantic generalisation. This simulates top-down processing by using the extensive network of associations encoded within their training data, thereby enabling application of these patterns to novel linguistic contexts. However, it remains an open question whether top-down abstractions based only on language patterns can approximate the generalisations that emerge from grounded representations (e.g., Pavlick, 2023, for a discussion on this topic). The challenges faced by large language models in word sense disambiguation, as highlighted in our current study, could provide valuable insights into whether grounded representations are necessary to accurately model human language processing.

For example, when modelling word acquisition trajectories, Transformers are not influenced by grounded sensorimotor, social, and cognitive factors (e.g., noun concreteness), but rely on surface features (e.g., word frequency) to a greater extent than children (Chang & Bergen, 2021). We speculate that this lack of grounded knowledge might also explain the fact that the models performed worse at disambiguating prior contexts than current contexts (Figure 2). Current contexts contained words that might appear closer to target words in naturalistic language, becoming easier to track by a distributional learner. This difficulty might not exist for children who can use their real-world knowledge for semantically-related (but distant) words (e.g., in *Dora looked in her drawer. The band was cool,* a child can infer that entities stored in a drawer are usually objects). Indeed, word acquisition trajectories can probably be better captured by neural models that process a richer multimodal signal comprising auditory features, communicative intentions, and perceptual information about word referents (e.g., Frank et al., 2009; Nikolaus & Fourtassi, 2021; Nyamapfene & Ahmad, 2007). Future work should focus on modelling child multimodal processing, currently limited by the scarcity of naturalistic multimodal corpora (e.g., Nikolaus et al., 2022).

Integrating multimodal input could also be potentially beneficial for investigating the models' performance with words varying in concreteness (e.g., concrete nouns vs more abstract verbs), which was not considered in our simulations but could be intriguing given the role of concreteness in early vocabulary learning (e.g., Braginsky et al., 2019). For instance, abstract nouns or verbs might depend more heavily on linguistic context for disambiguation, whereas concrete nouns might rely more on multimodal contexts (e.g., Sakreida et al., 2013). Highlighting this distinction could be valuable for future research, suggesting that Transformers trained on text might demonstrate superior performance with abstract words. This potential difference warrants further investigation to better understand how varying contexts influence word disambiguation across different word types.

Moreover, examining the distinction between concrete and abstract word senses could further elucidate the implications of basing model sense prototypes on child-directed or adult-directed sentences. For instance, since child-directed input often features more redundancy and a concrete vocabulary (Saxton, 2009) compared to adult-directed input, this might result in the formation of sense prototypes that better facilitate the disambiguation of concrete nouns like those used in our study. In other words, similarly to how child-directed sense prototypes may lead to a dominant sense bias typical of child-directed speech (Appendix S6), one should also find that child-directed sense prototypes lead to a bias toward concrete nouns.

Enriching models' input would allow researchers to test if acquiring multimodal knowledge suffices to capture sensitivity to top-down structures, or whether one would need to integrate domain-specific constraints in line with nativist approaches (e.g., Pinker, 1989; Thornton, 2012) or more domain-general innate biases (e.g., Perfors et al., 2011). For instance, a development of our work might involve investigating whether a purely distributional learner that can process visual object referents is able to bootstrap certain elements of sentence structure that are posited to be innate by alternative theories of language development. For example, when a word typically used as a verb (e.g., "eat") is presented in a noun context (e.g., "an eat"), 20-month-old infants more readily associate the word with a novel animal. Conversely, when a noun is strongly linked to a specific referent (e.g., "dog"), infants struggle to apply it to a different novel animal (Dautriche et al., 2018). This phenomenon indicates that employing different syntactic categories facilitates the extension of a word's meaning to encompass new referents. Given this evidence, one could examine whether a purely distributional learner, trained on input mirroring the quantity and quality available to 20-month-old infants exhibits similar facilitation from syntactic categories on word sense extension. Such empirical evidence would challenge the idea, proposed by universal grammar theories, that syntactic categories are innate rather than learned through language interaction (e.g., Valian et al., 2009).

Additionally, our method of assessing word sense disambiguation in large language models and humans could be used to evaluate approaches that view learning as an embodied and situated phenomenon. Indeed, the formation of semantic representations of words is not uniquely based on the statistics of word co-occurrences in language (the language-based distributional hypothesis). Properties of words related to the extralinguistic environment (e.g., physical properties) also play a crucial role in shaping semantic representations (the experiential hypothesis). Examining the capabilities of a distributional learner that relies exclusively on language co-occurrence statistics to capture word semantic representations can shed light on the importance of considering the real-world experiences of children. This approach can help determine how these two sources of information—linguistic and experiential—contribute independently or together to children's learning (e.g., Andrews et al., 2009). To this end, research involving language-based large language models can be expanded to also consider the combined influence of visual aspects (Lu et al., 2019; Qi et al., 2020; Sun et al., 2019; Zhuang et al., 2023).

Finally, the language that children are exposed to is often displaced, meaning caregivers frequently discuss word referents that are not present in the immediate environment (Tomasello & Kruger, 1992). Despite this, children might still leverage extralinguistic cues, such as iconicity (e.g., a caregiver mimicking the action of swinging a bat to clarify its meaning in conversation), in line with the language-as-situated hypothesis (Murgiano et al., 2021). Therefore, exploring the extent to which child semantic representations can be derived from both the linguistic and physical contexts in which children learn can reveal whether it is necessary to incorporate additional aspects of the communicative context, such as iconic cues, into our understanding of child word meaning representation.

Conclusion - What Large Language Models (LLMs) can('t) tell us about child language acquisition

We have begun to examine the capabilities and limitations of Transformer models for studying early word sense disambiguation. We have demonstrated that, as efficient distributional learners processing raw language input, large language models can be used to provide proof of principles concerning the extent to which usage-based learning can contribute to the acquisition of semantic representations at the word level. Importantly, it is this proficiency that highlights an interesting contrast: We have found that, although large language models excel at numerous language understanding and production tasks, they show significant limitations in their use of top-down cues for sense disambiguation. This results in their performance falling short compared to that of young children under certain disambiguation conditions. This finding serves as a crucial hint that an approach centered on providing more distributional linguistic cues might not be the most effective solution. Rather, it underscores the importance of either making models sensitive to additional multimodal cues or integrating specific constraints or biases into the models. This additional knowledge could potentially enable them to bridge the performance gap and align more closely to child learners.

Furthermore, in our tasks requiring the use of sentence context for word-level disambiguation, large language models have allowed us to avoid having to equip the models with syntactic and semantic knowledge at the sentence level (using external resources to pre-process the input) to ultimately perform word disambiguation. This would have required making assumptions about what knowledge the learner possesses at a certain point in development, which can come with benefits but also complications stemming from confounding effects caused by the assumptions made by the modeler.

Finally, we showed that an evaluation approach that leverages sense-annotated corpora can sensibly be used to examine the developmental plausibility of sense representations in large language models. Currently, limitations concerning model pretraining do not allow researchers to determine the impact of child language input on models' performance. However, we have seen that even the simple use of sense prototypes based on child input produced a partial alignment to child processing. This presents the prospect of combining corpus analyses of models' input with experimental simulations to elucidate the dynamics between the contribution of input characteristics and the nature of the learner's representational system.

References

Abbot-Smith, K., & Tomasello, M. (2006). *Exemplar-learning and schematization in a usage-based account of syntactic acquisition. 23*(3), 275–290. https://doi.org/10.1515/TLR.2006.011

Alishahi, A., & Stevenson, S. (2013). Gradual Acquisition of Verb Selectional Preferences in a Bayesian Model. In A. Villavicencio, T. Poibeau, A. Korhonen, & A. Alishahi (Eds.), *Cognitive Aspects of Computational Language Acquisition* (pp. 297–316). Springer. https://doi.org/10.1007/978-3-642-31863-4_11

Ambridge, B. (2019). Against stored abstractions: A radical exemplar model of language acquisition: *First Language*. https://doi.org/10.1177/0142723719869731

Ambridge, B. (2020). Abstractions made of exemplars or 'You're all right, and I've changed my mind': Response to commentators. *First Language, 40*(5–6), 640–659. https://doi.org/10.1177/0142723720949723

Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language, 42*(2), 239–273. https://doi.org/10.1017/S030500091400049X

Andreu, L., Sanz-Torrent, M., & Trueswell, J. C. (2013). Anticipatory sentence processing in children with specific language impairment: Evidence from eye movements during listening. *Applied Psycholinguistics, 34*(1), 5–44. https://doi.org/10.1017/S0142716411000592

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review, 116*(3), 463–498. https://doi.org/10.1037/a0016261

BNC Consortium. (2007). *British National Corpus, XML edition*. https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/2554

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2022). *On the Opportunities and Risks of Foundation Models* (arXiv:2108.07258). arXiv. http://arxiv.org/abs/2108.07258

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2019). Consistency and Variability in Children's Word Learning Across Languages. *Open Mind : Discoveries in Cognitive Science, 3*, 52–67. https://doi.org/10.1162/opmi_a_00026

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners* (arXiv:2005.14165). arXiv. https://doi.org/10.48550/arXiv.2005.14165

Bybee, J. (2010). *Language, Usage and Cognition*. Cambridge University Press. https://doi.org/10.1017/CBO9780511750526

Cabiddu, F., Bott, L., Jones, G., & Gambi, C. (2022a). ChiSense-12: An English Sense-Annotated Child-Directed Speech Corpus. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 5198–5205. https://aclanthology.org/2022.lrec-1.557

Cabiddu, F., Bott, L., Jones, G., & Gambi, C. (2022b). The Role of Verb-Event Structure in Children's Lexical Ambiguity Resolution. *Proceedings of the Annual Meeting of the Cognitive Science Society, 44*(44). https://escholarship.org/uc/item/9kh29212

Chang, T. A., & Bergen, B. K. (2021). *Word Acquisition in Neural Language Models* (arXiv:2110.02406). arXiv. https://doi.org/10.48550/arXiv.2110.02406

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., … Fiedel, N. (2022). *PaLM: Scaling Language Modeling with Pathways* (arXiv:2204.02311). arXiv. https://doi.org/10.48550/arXiv.2204.02311

Dautriche, I., Fibla, L., Fievet, A.-C., & Christophe, A. (2018). Learning homophones in context: Easy cases are favored in the lexicon of natural languages. *Cognitive Psychology, 104,* 83–105. https://doi.org/10.1016/j.cogpsych.2018.04.001

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Duffy, S. A., Kambe, G., & Rayner, K. (2001). The effect of prior disambiguating context on the comprehension of ambiguous words: Evidence from eye movements. In *On the consequences of meaning selection: Perspectives on resolving lexical ambiguity* (pp. 27–43). American Psychological Association. https://doi.org/10.1037/10459-002

Frank, M. C. (2023). Openly accessible LLMs can help us to understand human cognition. *Nature Human Behaviour, 7*(11), Article 11. https://doi.org/10.1038/s41562-023-01732-4

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science, 20*(5), 578–585. https://doi.org/10.1111/j.1467-9280.2009.02335.x

Gammelgaard, M. L., Christiansen, J. G., & Søgaard, A. (2023). *Large language models converge toward human-like concept organization* (arXiv:2308.15047). arXiv. https://doi.org/10.48550/arXiv.2308.15047

Haber, J., & Poesio, M. (2020). Word Sense Distance in Human Similarity Judgements and Contextualised Word Embeddings. *Proceedings of the Probability and Meaning Conference (PaM 2020),* 128–145. https://aclanthology.org/2020.pam-1.17

Hahn, N., Snedeker, J., & Rabagliati, H. (2015). Rapid Linguistic Ambiguity Resolution in Young Children with Autism Spectrum Disorder: Eye Tracking Evidence for the Limits of Weak Central Coherence. *Autism Research, 8*(6), 717–726. https://doi.org/10.1002/aur.1487

Hewitt, J., & Manning, C. D. (2019). A Structural Probe for Finding Syntax in Word Representations. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4129–4138. https://doi.org/10.18653/v1/N19-1419

Hosseini, E. A., Schrimpf, M., Zhang, Y., Bowman, S., Zaslavsky, N., & Fedorenko, E. (2022). *Artificial neural network language models align neurally and behaviorally with humans even after a developmentally realistic amount of training* (p. 2022.10.04.510681). bioRxiv. https://doi.org/10.1101/2022.10.04.510681

Huebner, P. A., Sulem, E., Cynthia, F., & Roth, D. (2021). BabyBERTa: Learning More Grammar With Small-Scale Child-Directed Language. *Proceedings of the 25th Conference on Computational Natural Language Learning*, 624–646. https://doi.org/10.18653/v1/2021.conll-1.49

Jawahar, G., Sagot, B., & Seddah, D. (2019). What Does BERT Learn about the Structure of Language? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3651–3657. https://doi.org/10.18653/v1/P19-1356

Khanna, M. M., & Boland, J. E. (2010). Children's use of language context in lexical ambiguity resolution. *Quarterly Journal of Experimental Psychology, 63*(1), 160–193. https://doi.org/10.1080/17470210902866664

Kidd, E., & Bavin, E. L. (2005). Lexical and referential cues to sentence interpretation: An investigation of children's interpretations of ambiguous sentences. *Journal of Child Language, 32*(4), 855–876. https://doi.org/10.1017/S0305000905007051

Laverghetta Jr, A., & Licato, J. (2021). Modeling Age of Acquisition Norms Using Transformer Networks. *The International FLAIRS Conference Proceedings, 34*. https://doi.org/10.32473/flairs.v34i1.128334

Loureiro, D., Jorge, A. M., & Camacho-Collados, J. (2022). LMMS Reloaded: Transformer-based Sense Embeddings for Disambiguation and Beyond. *Artificial Intelligence, 305*, 103661. https://doi.org/10.1016/j.artint.2022.103661

Loureiro, D., Rezaee, K., Pilehvar, M. T., & Camacho-Collados, J. (2021). *Analysis and Evaluation of Language Models for Word Sense Disambiguation* (arXiv:2008.11608). arXiv. https://doi.org/10.48550/arXiv.2008.11608

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *Advances in Neural*

Information Processing Systems, 32. https://proceedings.neurips.cc/paper/2019/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract.html

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk: Transcription format and programs, Vol. 1, 3rd ed* (pp. xi, 366). Lawrence Erlbaum Associates Publishers.

Mani, N., & Huettig, F. (2012). Prediction during language processing is a piece of cake—But only for skilled producers. *Journal of Experimental Psychology. Human Perception and Performance, 38*(4), 843–847. https://doi.org/10.1037/a0029284

Mani, N., Daum, M. M., & Huettig, F. (2016). "Proactive" in many ways: Developmental evidence for a dynamic pluralistic approach to prediction. *Quarterly Journal of Experimental Psychology, 69*(11), 2189–2201. https://doi.org/10.1080/17470218.2015.1111395

McCauley, S. M., & Christiansen, M. H. (2019). Language learning as language use: A cross-linguistic model of child language development. *Psychological Review, 126*, 1–51. https://doi.org/10.1037/rev0000126

Melamud, O., Goldberger, J., & Dagan, I. (2016). context2vec: Learning Generic Context Embedding with Bidirectional LSTM. *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, 51–61. https://doi.org/10.18653/v1/K16-1006

Meylan, S. C., Mankewitz, J., Floyd, S., Rabagliati, H., & Srinivasan, M. (2021). Quantifying Lexical Ambiguity in Speech To and From English-Learning Children. *Proceedings of the Annual Meeting of the Cognitive Science Society, 43*(43). https://escholarship.org/uc/item/1pq031fn

Murgiano, M., Motamedi, Y., & Vigliocco, G. (2021). Situating Language in the Real-World: The Role of Multimodal Iconicity and Indexicality. *Journal of Cognition, 4*(1), 38. https://doi.org/10.5334/joc.113

Nair, S., Srinivasan, M., & Meylan, S. (2020). Contextualized Word Embeddings Encode Aspects of Human-Like Word Sense Knowledge. In M. Zock, E. Chersoni, A. Lenci, & E. Santus (Eds.), *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon* (pp. 129–141). Association for Computational Linguistics. https://aclanthology.org/2020.cogalex-1.16

Nikolaus, M., & Fourtassi, A. (2021). Evaluating the Acquisition of Semantic Knowledge from Cross-situational Learning in Artificial Neural Networks. *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 200–210. https://doi.org/10.18653/v1/2021.cmcl-1.24

Nikolaus, M., Alishahi, A., & Chrupała, G. (2022). Learning English with Peppa Pig. *Transactions of the Association for Computational Linguistics, 10,* 922–936. https://doi.org/10.1162/tacl_a_00498

Nyamapfene, A., & Ahmad, K. (2007). A Multimodal Model of Child Language Acquisition at the One-Word Stage. *2007 International Joint Conference on Neural Networks,* 783–788. https://doi.org/10.1109/IJCNN.2007.4371057

Pavlick, E. (2023). Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 381*(2251), 20220041. https://doi.org/10.1098/rsta.2022.0041

Perfors, A., Tenenbaum, J. B., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition, 118*(3), 306–338. https://doi.org/10.1016/j.cognition.2010.11.001

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers),* 2227–2237. https://doi.org/10.18653/v1/N18-1202

Pinker, S. (1989). Learnability and Cognition. *MIT Press.* https://mitpress.mit.edu/9780262660730/learnability-and-cognition/

Qi, D., Su, L., Song, J., Cui, E., Bharti, T., & Sacheti, A. (2020). ImageBERT: Cross-modal Pre-training with Large-scale Weak-supervised Image-Text Data. *arXiv:2001.07966 [Cs].* http://arxiv.org/abs/2001.07966

Rabagliati, H., Pylkkänen, L., & Marcus, G. F. (2013). Top-down influence in young children's linguistic ambiguity resolution. *Developmental Psychology, 49,* 1076–1089. https://doi.org/10.1037/a0026918

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners.* https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2023). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer* (arXiv:1910.10683). arXiv. https://doi.org/10.48550/arXiv.1910.10683

Rodd, J. M. (2020). Settling Into Semantic Space: An Ambiguity-Focused Account of Word-Meaning Access. *Perspectives on Psychological Science, 15*(2), 411–427.

https://doi.org/10.1177/1745691619885860

Rodd, J. M., Gaskell, M. G., & Marslen-Wilson, W. D. (2004). Modelling the effects of semantic ambiguity in word recognition. *Cognitive Science, 28*(1), 89–104. https://doi.org/10.1207/s15516709cog2801_4

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science, 274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Sakreida, K., Scorolli, C., Menz, M. M., Heim, S., Borghi, A. M., & Binkofski, F. (2013). Are abstract action words embodied? An fMRI investigation at the interface between language and motor cognition. *Frontiers in Human Neuroscience, 7,* 125. https://doi.org/10.3389/fnhum.2013.00125

Saxton, M. (2009). The Inevitability of Child Directed Speech. In S. Foster-Cohen (Ed.), *Language Acquisition* (pp. 62–86). Palgrave Macmillan UK. https://doi.org/10.1057/9780230240780_4

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences, 118*(45), e2105646118. https://doi.org/10.1073/pnas.2105646118

Snedeker, J., & Trueswell, J. C. (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive Psychology, 49*(3), 238–299. https://doi.org/10.1016/j.cogpsych.2004.03.001

Snedeker, J., & Yuan, S. (2008). Effects of prosodic and lexical constraints on parsing in young children (and adults). *Journal of Memory and Language, 58*(2), 574–608. https://doi.org/10.1016/j.jml.2007.08.001

Srinivasan, M., & Rabagliati, H. (2021). The Implications of Polysemy for Theories of Word Learning. *Child Development Perspectives, 15*(3), 148–153. https://doi.org/10.1111/cdep.12411

Sun, C., Myers, A., Vondrick, C., Murphy, K., & Schmid, C. (2019). VideoBERT: A Joint Model for Video and Language Representation Learning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV),* 7463–7472. https://doi.org/10.1109/ICCV.2019.00756

Tenney, I., Das, D., & Pavlick, E. (2019). *BERT Rediscovers the Classical NLP Pipeline* (arXiv:1905.05950). arXiv. https://doi.org/10.48550/arXiv.1905.05950

Thornton, R. (2012). Studies at the interface of child language and models of language acquisition. *First Language*, *32*(1–2), 281–297. https://doi.org/10.1177/0142723711403881

Tomasello, M., & Kruger, A. C. (1992). Joint attention on actions: Acquiring verbs in ostensive and non-ostensive contexts. *Journal of Child Language*, *19*(2), 311–333. https://doi.org/10.1017/S0305000900011430

Trueswell, J. C., & Gleitman, L. R. (2007). Learning to parse and its implications for language acquisition. In M. G. Gaskell (Ed.), *The Oxford Handbook of Psycholinguistics* (p. 0). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198568971.013.0039

Valian, V., Solt, S., & Stewart, J. (2009). Abstract categories or limited-scope formulae? The case of children's determiners. *Journal of Child Language*, *36*(4), 743–778. https://doi.org/10.1017/S0305000908009082

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Warstadt, A., & Bowman, S. R. (2022). *What Artificial Neural Networks Can Tell Us About Human Language Acquisition* (arXiv:2208.07998). arXiv. https://doi.org/10.48550/arXiv.2208.07998

Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T., & Cotterell, R. (2023). Findings of the BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora. In A. Warstadt, A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, T. Linzen, & R. Cotterell (Eds.), *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning* (pp. 1–34). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.conll-babylm.1

Wiedemann, G., Remus, S., Chawla, A., & Biemann, C. (2019). *Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings* (arXiv:1909.10430). arXiv. https://doi.org/10.48550/arXiv.1909.10430

Yacovone, A., Shafto, C. L., Worek, A., & Snedeker, J. (2021). Word vs. World Knowledge: A developmental shift from bottom-up lexical cues to top-down plausibility. *Cognitive Psychology*, *131*, 101442. https://doi.org/10.1016/j.cogpsych.2021.101442

Zhang, Y., Warstadt, A., Li, X., & Bowman, S. R. (2021). When Do You Need Billions of Words of Pretraining Data? *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1112–1125. https://doi.org/10.18653/v1/2021.acl-long.90

Zhuang, C., Fedorenko, E., & Andreas, J. (2023, October 20). *Visual Grounding Helps Learn Word Meanings in Low-Data Regimes*. arXiv.Org. https://arxiv.org/abs/2310.13257v1

Data, code and materials availability statement

Raw data, simulation and analysis scripts used in the study can be found on the GitHub project repository https://doi.org/10.5281/zenodo.8200803. The ChiSense-12 corpus can be downloaded at https://gitlab.com/francescocabiddu/chisense-12. The CHILDES database is accessible at https://childes.talkbank.org/. The British National Corpus can be downloaded at https://llds.ling-phil.ox.ac.uk/llds/xmlui/handle/20.500.14106/2554.

Ethics statement

Ethics approval was not required as the study used previously-collected publicly-available data from the CHILDES database (MacWhinney, 2000).

Authorship and Contributorship Statement

Francesco Cabiddu: conceptualization; data curation; formal analysis; investigation; methodology; project administration; resources; software; writing – original draft preparation; writing – review & editing; funding acquisition. **Mitja Nikolaus:** conceptualization; formal analysis; methodology; writing – review & editing. **Abdellah Fourtassi**: conceptualization; formal analysis; methodology; writing – review & editing.

All authors approved the final version of the manuscript and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Acknowledgements

We would like to express our gratitude to Alex Warstadt for his valuable comments during the initial stages of this project, and to Abhishek Agrawal for offering helpful suggestions on the manuscript. We also extend our thanks to the anonymous reviewers for their insightful feedback. The research presented in this article has been generously supported by a Study Visit Grant from the Experimental Psychology Society (UK).

Appendix S1: Model Families

We provide a description of the model families included in the study, and details about models' configurations varying in model size and pretraining size (Table S1.1). Transformer models were downloaded using the Huggingface Transformers Python library (Wolf et al. 2020), apart from the model BabyBERTa (Huebner et al. 2021) whose pretrained weights were downloaded directly from its GitHub project page (https://github.com/phueb/BabyBERTa, October 2022). The recurrent neural model ELMo (version 3; Peters et al. 2018) was downloaded using the TensorFlow Python library (Abadi et al. 2015).

The 13 Transformer model families used were: BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019), and GTP (OpenAI GPT, Radford et al. 2018; GPT-2, Radford et al. 2019). For each of these three families we included their distilled model versions (DistilBERT, DistilRoBERTa, and DistilGPT2; Sanh et al. 2020), and the RoBERTa family also included versions pretrained on small corpora (MiniBERTa, Warstadt et al. 2020). BabyBERTa (Huebner et al. 2021); ALBERT-v1 and ALBERT-v2 (Lan et al. 2020); DeBERTa and DeBERTA-v2 (He, Gao, et al., 2021); DeBERTa-v3 (He, Liu, et al., 2021); Transformer-XL (Dai et al., 2019); CTRL (Keskar et al., 2019); T5 (Raffel et al., 2020); XLNet (Yang et al., 2020).

A first macro distinction between families concerns their unidirectional or bidirectional way of predicting a token given its context. Unidirectional Transformers (GPT, Transformer-XL, and CTRL) are trained on predicting the next token given the (previous) left sentence context. This type of training objective is in line with prediction-based approaches of children's online sentence processing (Mani & Huettig, 2012). The remaining Transformers and ELMo are instead trained on predicting tokens by taking into account both (previous) left and (following) right contexts. This type of objective is plausible because children are not only involved in predicting upcoming input when hearing speech, but they can also revise their interpretation of ambiguous words based on following cues (e.g., Qi et al., 2020). Also, in naturalistic conversations there are cases in which children would likely attend to following sentence context to disambiguate nouns (e.g., "*Look at the bat, it's flying!*").

A second macro distinction concerns how different models track the position of tokens in a text sequence. Most models track tokens' absolute positions, essentially encoding sentence word order which is required for learning syntax (e.g., distinguishing between "*The dog chased the boy*" and "*The boy chased the dog*"). Additionally, some models implement mechanisms that track both absolute and relative positions of tokens (DeBERTa, DeBERTa-v2, DeBERTa-v3) or only relative positions (Transformer-XL, T5, XLNet). Tracking relative positions means tracking the relative distance between pairs of tokens in a sequence, which translates into weighting more the words that appear closer to a target word (e.g., the contribution of "*deep*" for the vector representation of "*learning*" is higher if the two appear one next to the other, compared to when they appear in different sentences). Tracking relative positions can be considered a proxy of children's sentence local parsing (e.g., Gertner & Fisher, 2012).

BERT is a bidirectional Transformer trained on predicting tokens that are masked at random during the preprocessing of the input, with some sentences seen multiple times with the same masked tokens (i.e., static masking). It is also pretrained on predicting whether a sentence follows another in the input (next sentence prediction), with the aim of capturing relations between sentences that can be useful in Question Answering and Natural Language Inference tasks. The model is pretrained on the BookCorpus (Zhu et al., 2015) and English Wikipedia.

RoBERTa is a modification of BERT that is trained without the next sentence prediction objective, which investigations found to be not effective for improving performance in downstream tasks (e.g., Liu et al. 2019; Yang et al., 2020). It is trained by receiving larger batches of examples at every weight updating iteration. It is also trained on a larger corpus than BERT, additionally including English news articles, web content, and stories. The model also uses dynamic masking, which masks different tokens every time the same sentence is fed to the model. Its scaled-down version, MiniBERTa, is pretrained on similar input (BookCorpus and Wikipedia) but on a much smaller scale (see Table S1.1), with the configuration pretrained on the smallest corpus (1M tokens) also reduced in model size.

GPT models are unidirectional Transformers trained on a language modeling objective, namely sampling text from the input dataset and asking the model to predict the next token. OpenAI GPT was pretrained on the BookCorpus, and subsequently fine-tuned with a series of supervised language understanding tasks. GPT-2 was instead pretrained on a larger corpus of web content, with no supervised fine-tuning.

Distilled models are compressed and faster versions of the above models, based on the same architectures but with reduced number of layers. They undergo training that specifically tries to reproduce the behavior of the (parent) larger model.

BabyBERTa is a scaled-down version of RoBERTa with some key differences. It is significantly reduced in size (15x fewer parameters). It modifies the masked word prediction objective: In BERT and RoBERTa, 10% of the tokens selected for masking are left unmasked; BabyBERTa never allows unmasking. It is also pretrained on much smaller (6000x fewer tokens) and qualitatively different corpora, either separately on transcribed child-directed speech, written child-directed news articles, a small portion of Wikipedia, or a combination of the three.

ALBERT-v1 is a light version of BERT that was created with the main goal of reducing the computational costs derived from using a large number of parameters. ALBERT-v1 uses two techniques (factorization of parameters, and sharing all parameters across model layers) which significantly reduce the number of parameters without significant drops in performance in downstream tasks. Additionally, ALBERT-v1 modifies the next sentence prediction objective performing sentence order prediction instead. A key difference between the two objectives is that in next sentence prediction, the model is provided with positive examples of pairs of consecutive sentences coming from the same document, and negative examples with the second sentence of the pair swapped with one coming from a different document. The inefficiency of this task comes from the fact that negative examples contain sentences coming from different documents, which likely contain text about different topics. This results in the model being able to easily learn from negative examples by just noticing differences in word occurrences (i.e., semantically different words are used when sentences refer to different topics), focusing less on the more important aspect of discourse coherence between the two sentences. Therefore, with the new sentence order prediction objective, negative examples comprise sentence pairs coming from the same document, just swapped in order. This forces the model to focus on the coherence of one sentence following the other. This new objective significantly improved performance in downstream tasks compared to BERT. ALBERT-v1 is pretrained on the same datasets used for BERT.

ALBERT-v2 is a modification of ALBERT-v1 that improves performance at downstream tasks by using a different training regime (higher training steps and time) and by removing dropout, which is normally used to avoid that a model overfits the training dataset.

DeBERTa is a modification of RoBERTa which improves performance in downstream tasks by using mechanisms of disentangled attention and enhanced mask decoding, which essentially allow the model to integrate both absolute and relative token positions in its vector representations. DeBERTa is trained on the same corpora used for RoBERTa but excluding English news articles.

DeBERTa-v2 is an optimized version of DeBERTa, which uses a larger vocabulary, larger pretraining dataset, and larger model sizes. It shares parameters that track sentence content and relative positions to reduce model complexity. It also integrates an additional layer in the model to better learn knowledge about subword n-grams, with the aim of more precisely tracking sentence local dependences. DeBERTa-v2 is pretrained on the same RoBERTa corpora.

DeBERTa-v3 is a modification of DeBERTa-v2 that replaces the masked word prediction objective with a replaced token detection objective, which instead of randomly masking tokens during training it replaces them with plausible (but incorrect) ones. This changes the objective of the model from having to generate plausible tokens to having to discriminate between two semantically related tokens to decide which is the appropriate one in a sentence. DeBERTa-v3 is pretrained on the same RoBERTa corpora.

Transformer-XL is a unidirectional model that uses a language modeling objective as GPT. Transformer-XL introduces a recurrence mechanism in the Transformer architecture. Usually, Transformers process input in the form of text segments of a maximum length, which results in the impossibility of modelling dependencies across segments (which are treated independently). Transformer-XL uses a mechanism that recycles hidden states of previous segments and uses them as extended context for newly processed ones. Additionally, the model introduces a new mechanism that can keep track of the relative position of tokens across different segments. Recurrence and relative positional encoding allow Transformer-XL to track short-range and long-range text dependencies, which can be used to generate very long and relatively coherent articles. The model is pretrained on a small dataset of Wikipedia articles (Merity et al., 2016).

CTRL is another unidirectional Transformer that uses a language modeling objective. However, in this model the objective is modified so that the model predicts the next token of a sequence also taking into account specific codes present in the structure of the training data. These codes give information such as the specific domain of the text being processed (e.g., Wikipedia, Books), the specific style used (e.g., Horror, Science), or the specific tasks being processed (e.g., question answering, translation). These codes are extracted directly from structural components of the training data, and ultimately allow the model to better constrain its text generation process. CTRL is pretrained on a large corpus from Wikipedia, web content including news articles and Amazon reviews, translation datasets from European parliament and United Nations proceedings, and various question-answering datasets.

T5 is a bidirectional Transformer that uses an Encoder and a Decoder architecture similar to the original Transformer (Vaswani et al., 2017). It is trained on a masked prediction objective similar to BERT, representing both single (as in BERT) and sequences of tokens in the Encoder and using learned representations to generate text in the Decoder. In our study, we only used the Encoder part of the model. The model also uses a mechanism of relative positional encoding. The model is trained on the largest corpus considered in our study, which comprises scraped content from the web.

XLNet is a bidirectional Transformer that modifies the BERT training objective using a permutation modeling objective. In BERT, masked tokens within a text sequence are predicted independently from one another. In XLNet the prediction also takes into account the relations between masked tokens. Additionally, XLNet only uses a mechanism of relative positional encoding. The model is trained on the same corpora used for BERT, with the addition of various corpora of web content and news articles.

ELMo is a bidirectional recurrent neural network model. Its mechanism of recurrence allows to link current word representations to previous ones in a text sequence. This is achieved by processing input at different timesteps, and feeding the output of previous timesteps to the current one. The recurrence mechanism leads to contextualized representations that also encode information about word order. In ELMo, the input sequence is fed to the model from left to right, and again from right to left. The two output vectors are then combined to obtain a bidirectional representation. The model is trained on a corpus of News Crawl data (Chelba et al., 2014).

Table S1.1. ***Models included in the study, by pretraining size (gigabytes of text), model size (million parameters), and family type.***

Model	**Model size**	**Pretraining size**	**Family**
distilbert-base-uncased	66	16	bert
bert-base-uncased	110	16	bert
bert-large-uncased	340	16	bert
bert-large-uncased-whole-word-masking	340	16	bert
distilroberta-base	82	40	roberta
roberta-base	125	160	roberta
roberta-large	355	160	roberta
roberta-med-small-1M-2	45	0.005	roberta
roberta-base-10M-2	125	0.05	roberta
roberta-base-100M-2	125	0.5	roberta
roberta-base-1B-3	125	5	roberta
albert-base-v1	11	16	albert-v1
albert-large-v1	17	16	albert-v1
albert-xlarge-v1	58	16	albert-v1
albert-xxlarge-v1	223	16	albert-v1
albert-base-v2	11	16	albert-v2
albert-large-v2	17	16	albert-v2
albert-xlarge-v2	58	16	albert-v2
albert-xxlarge-v2	223	16	albert-v2
deberta-base	140	80	deberta
deberta-large	400	80	deberta
deberta-xlarge	750	80	deberta
deberta-v2-xlarge	900	160	deberta-v2

Table S1.1 (continued).

Model	**Model size**	**Pretraining size**	**Family**
deberta-v2-xxlarge	1500	160	deberta-v2
deberta-v3-small	141	160	deberta-v3
deberta-v3-base	184	160	deberta-v3
deberta-v3-large	434	160	deberta-v3
babyberta-ao-childes	8	0.02	babyberta
babyberta-ao-newsela	8	0.02	babyberta
babyberta-wikipedia-1	8	0.02	babyberta
babyberta-ao-childes-ao-newsela-wikipedia-1	8	0.06	babyberta
distilgpt2	82	40	gpt
openai-gpt	116	3	gpt
gpt2	124	40	gpt
gpt2-medium	355	40	gpt
gpt2-large	774	40	gpt
gpt2-xl	1558	40	gpt
transfo-xl-wt103	284	0.4	transfo-xl
ctrl	1630	140	ctrl
t5-small	35	806	t5
t5-base	110	806	t5
t5-large	335	806	t5
xlnet-base-cased	117	126	xlnet
xlnet-large-cased	360	126	xlnet
elmo	93	4.2	elmo

References (for Appendix S1)

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., … Zheng, X. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. https://www.tensorflow.org/

Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., & Robinson, T. (2014). *One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling* (arXiv:1312.3005). arXiv. https://doi.org/10.48550/arXiv.1312.3005

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2019). *Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context* (arXiv:1901.02860). arXiv. https://doi.org/10.48550/arXiv.1901.02860

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Gertner, Y., & Fisher, C. (2012). Predicted errors in children's early sentence comprehension. *Cognition*, *124*(1), 85–94. https://doi.org/10.1016/j.cognition.2012.03.010

He, P., Gao, J., & Chen, W. (2021). *DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing* (arXiv:2111.09543). arXiv. https://doi.org/10.48550/arXiv.2111.09543

He, P., Liu, X., Gao, J., & Chen, W. (2021). *DeBERTa: Decoding-enhanced BERT with Disentangled Attention* (arXiv:2006.03654). arXiv. https://doi.org/10.48550/arXiv.2006.03654

Huebner, P. A., Sulem, E., Cynthia, F., & Roth, D. (2021). BabyBERTa: Learning More Grammar With Small-Scale Child-Directed Language. *Proceedings of the 25th Conference on Computational Natural Language Learning*, 624–646. https://doi.org/10.18653/v1/2021.conll-1.49

Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., & Socher, R. (2019). *CTRL: A Conditional Transformer Language Model for Controllable Generation* (arXiv:1909.05858). arXiv. https://doi.org/10.48550/arXiv.1909.05858

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations* (arXiv:1909.11942). arXiv. https://doi.org/10.48550/arXiv.1909.11942

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (arXiv:1907.11692). arXiv. https://doi.org/10.48550/arXiv.1907.11692

Mani, N., & Huettig, F. (2012). Prediction during language processing is a piece of cake—But only for skilled producers. *Journal of Experimental Psychology: Human Perception and Performance, 38*, 843–847. https://doi.org/10.1037/a0029284

Merity, S., Xiong, C., Bradbury, J., & Socher, R. (2016). *Pointer Sentinel Mixture Models* (arXiv:1609.07843). arXiv. https://doi.org/10.48550/arXiv.1609.07843

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237. https://doi.org/10.18653/v1/N18-1202

Qi, Z., Love, J., Fisher, C., & Brown-Schmidt, S. (2020). Referential context and executive functioning influence children's resolution of syntactic ambiguity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46*(10), 1922. https://doi.org/10.1037/xlm0000886

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training*. https://gluebenchmark.com/leaderboard.
Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*. https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer* (arXiv:1910.10683). arXiv. http://arxiv.org/abs/1910.10683
Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter* (arXiv:1910.01108). arXiv. https://doi.org/10.48550/arXiv.1910.01108

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Warstadt, A., Zhang, Y., Li, H.-S., Liu, H., & Bowman, S. R. (2020). *Learning Which Features Matter: RoBERTa Acquires a Preference for Linguistic Generalizations (Eventually)* (arXiv:2010.05358). arXiv. https://doi.org/10.48550/arXiv.2010.05358

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., … Rush, A. M. (2020). *HuggingFace's Transformers: State-of-the-art Natural Language Processing* (arXiv:1910.03771). arXiv. https://doi.org/10.48550/arXiv.1910.03771

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2020). *XLNet: Generalized Autoregressive Pretraining for Language Understanding* (arXiv:1906.08237). arXiv. https://doi.org/10.48550/arXiv.1906.08237

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). *Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books* (arXiv:1506.06724). arXiv. https://doi.org/10.48550/arXiv.1506.06724

Appendix S2: Children's Target Words and Additional Annotations

In this section, we report details about the target ambiguous words used and their corresponding child-directed sentences. We used sentences from ChiSense-12 (Cabiddu et al., 2022a), a collection 53 sense-tagged corpora of American and British English child-directed speech from the CHILDES database (MacWhinney, 2000), involving 958 target children of up to 4 years of age (59 months). We selected sentences referring to 9 of the 12 ambiguous words present in the corpus, each in their dominant and subordinate sense. The remaining 3 words (flower/flour, moose/mousse, sun/son) could not be used because they had different spelling, creating no ambiguity for models' processing. Table S2.1 provides information about the number of sentences for each sense.

Some target words in the behavioral experiments were not covered by ChiSense-12. Thus, we additionally tagged all not covered words for which 40 sentences per sense were available in the same corpora used for ChiSense-12. This resulted in tagging 4 new ambiguous words (*fish* = animal/food; *lamb* = animal/food; *turkey* = animal/food; *card* = playing card/greetings card). In total, we covered 13/24 and 4/6 target words in Rabagliati et al. (2013) experiment 1 and 2 respectively, and 9/12 words from Cabiddu et al. (2022b). The sentence test items for each experiment are available in the appendices of the two original papers (Cabiddu et al., 2022b; Rabagliati et al., 2013), and in the file *test_utterances.csv* included in the R project folder of our GitHub project. The complete sets of utterances from ChiSense-12 and the new annotated words are available in the R folder of our project.

Loureiro et al. (2021) showed that a nearest neighbor approach for computing sense prototypes is stable even when drastically reducing the number of examples for each target sense. Given that we sampled a limited number of examples for each new target sense to keep the annotation work manageable (*n*=40), we verified that Loureiro's findings were supported in our case. We repeated the three modeling experiments using only the 9 target words of ChiSense-12, downsampling sentences for each sense before computing sense prototypes. The procedure was repeated 10 times, each time sampling a subset (*n* = 40) of randomly selected sentences for each sense. The results of the three experiments (Figure S2.1, S2.2, and S2.3) showed that performance remained stable even when using only 40 random sentences per sense, which justified the inclusion of the newly annotated words in our study.

Specifically, all three experiments yielded high correlations between the mean performance of each model across random samples and the performance using the full set of utterances: Experiment 1 (Rabagliati et al., 2013) $r_s = .95$; Experiment 2 (Rabagliati et al., 2013) $r_s = .95$; Experiment 1 (Cabiddu et al., 2022b) $r_s = .94$.

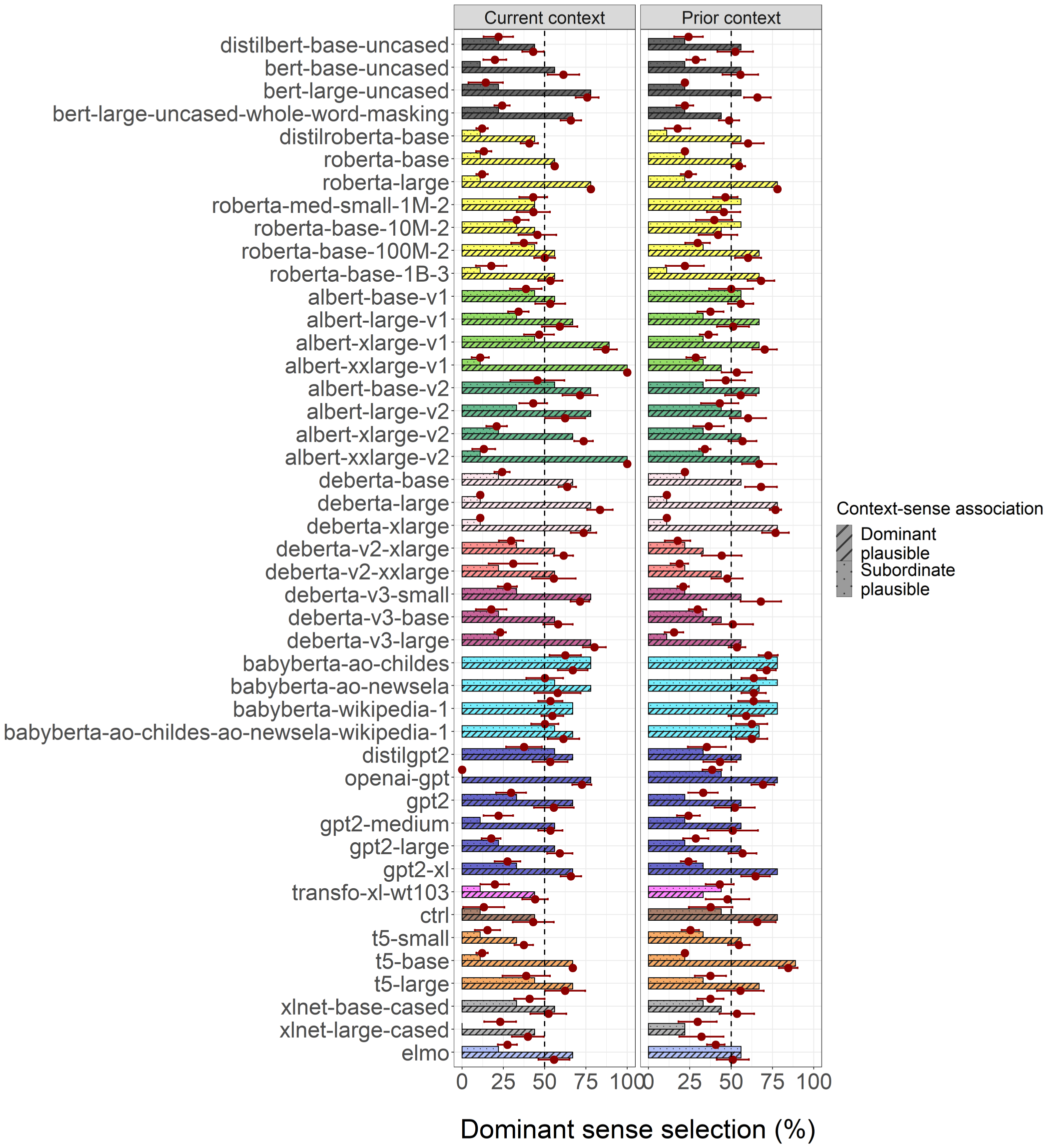

Figure S2.1. ***Percentage of dominant sense selections for experiment 1 of Rabagliati et al. (2013), in dominant-plausible and subordinate-plausible conditions (legend), when disambiguation cues were included in current or prior context (left and right panel respectively). Colored bars indicate performance of the models when the full sample of ChiSense-12 sentences is used to compute sense prototypes (Table S2.1). Red points indicate mean performance (across 10 runs) of models for which sense prototypes were computed using 40 random sentences for each sense. Error bars indicate standard deviations.***

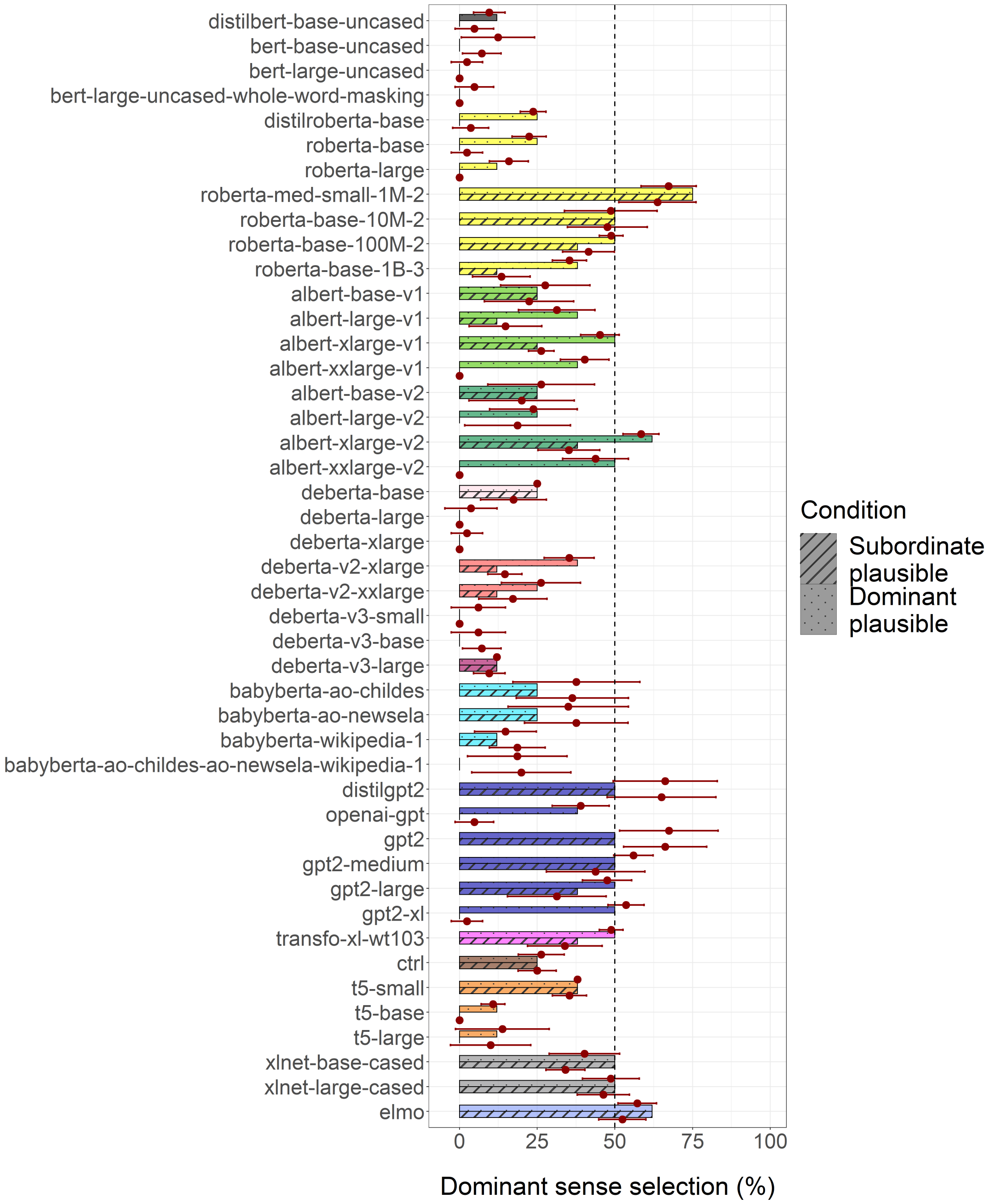

Figure S2.2. ***Percentage of dominant sense selections for experiment 2 of Rabagliati et al. (2013), in dominant-plausible and subordinate-plausible conditions (legend). The plot shows the comparison between dominant sense selection in models with prototypes computed from the full ChiSense-12 (colored bars), and models for which prototypes were computed by downsampling ChiSense-12 to 40 random sentences per sense (points and error bars indicate mean and standard deviations across 10 runs).***

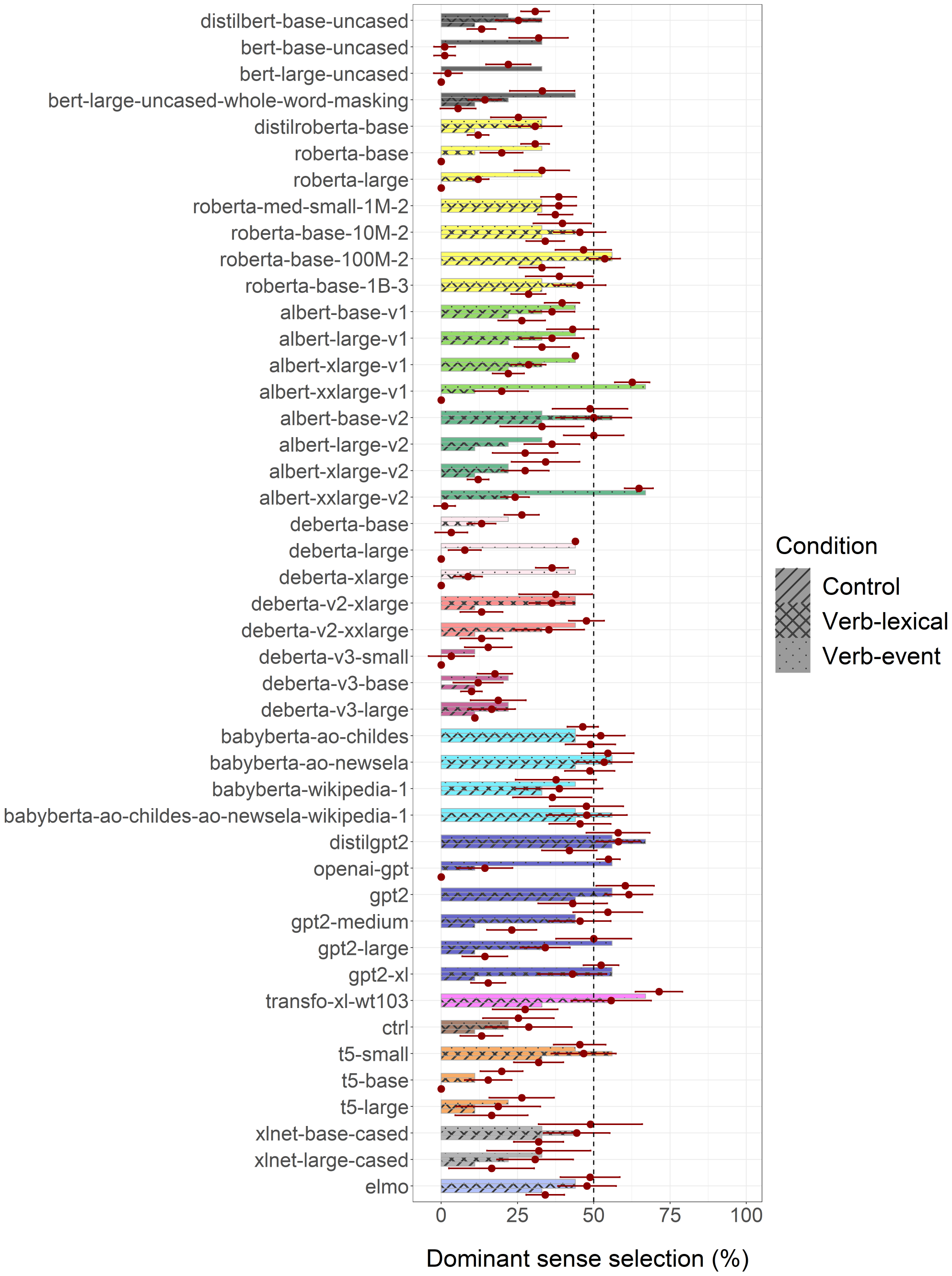

Figure S2.3. ***Percentage of dominant sense selections for Cabiddu et al. (2022b), in dominant-plausible (Verb-lexical, Verb-event) and subordinate-plausible conditions (Control). The plot shows the comparison between dominant sense selection in models with prototypes computed from the full ChiSense-12 (colored bars), and models for which prototypes were computed by downsampling ChiSense-12 to 40 random sentences per sense (points and error bars indicate mean and standard deviations across 10 runs).***

Table S2.1. ***For each target word, the table shows the raw number of utterances in which dominant (D) and subordinate (S) senses appeared, as well as the percentage of utterances in which dominant senses appeared (Dominance).***

Word (D/S)	*N* (D/S)	*Dominance*
Band (Object/Music Group)	178/58	75%
Bat (Animal/Object)	247/130	66%
Bow (Knot/Weapon)	230/27	89%
Button (Electronic/Clothing)	568/285	67%
Chicken (Animal/Food)	1463/937	61%
Glasses (Eye/Drinking)	683/620	52%
Letter (Alphabet/Mail)	1446/946	60%
Line (Geometric/Row)	471/241	66%
Nail (Finger/Tool)	460/106	81%
MEAN (*SD*)	-	69% (11%)

References (for Appendix S2)

Cabiddu, F., Bott, L., Jones, G., & Gambi, C. (2022a). ChiSense-12: An English Sense-Annotated Child-Directed Speech Corpus. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 5198–5205. https://aclanthology.org/2022.lrec-1.557

Cabiddu, F., Bott, L., Jones, G., & Gambi, C. (2022b). The Role of Verb-Event Structure in Children's Lexical Ambiguity Resolution. *Proceedings of the Annual Meeting of the Cognitive Science Society, 44*(44). https://escholarship.org/uc/item/9kh29212

Loureiro, D., Rezaee, K., Pilehvar, M. T., & Camacho-Collados, J. (2021). *Analysis and Evaluation of Language Models for Word Sense Disambiguation* (arXiv:2008.11608). arXiv. https://doi.org/10.48550/arXiv.2008.11608

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk: Transcription format and programs, Vol. 1, 3rd ed* (pp. xi, 366). Lawrence Erlbaum Associates Publishers.

Rabagliati, H., Pylkkänen, L., & Marcus, G. F. (2013). Top-down influence in young children's linguistic ambiguity resolution. *Developmental Psychology, 49*, 1076–1089. https://doi.org/10.1037/a0026918

Appendix S3: Randomly Initialized Models

We modelled the three experiments (Cabiddu et al., 2022; Rabagliati et al., 2013), running base model versions 10 times using different random initializations. For a single run, the same initialization was used to create both sense prototypes and vectors of test stimuli. None of the models showed sensitivity to sentence context across experiments (Figure S3.1, S3.2, and S3.3; i.e., same percentage of dominant sense selections across conditions), suggesting that different patterns of connections among units did not influence models' performance.

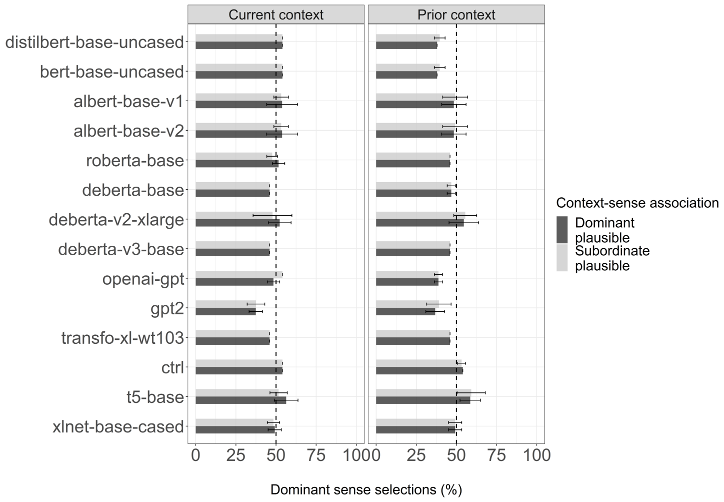

Figure S3.1. ***Mean percentage of dominant sense selections in randomly initialized models for experiment 1 of Rabagliati et al. (2013), in dominant-plausible and subordinate-plausible conditions (legend), when disambiguation cues were included in current or prior context (left and right panel respectively). Error bars indicate standard deviations over 10 model runs.***

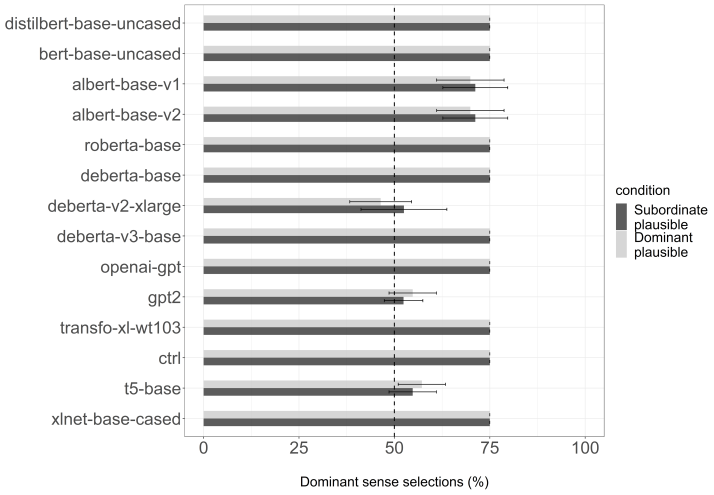

Figure S3.2. ***Mean percentage of dominant sense selections in randomly initialized models for experiment 2 of Rabagliati et al. (2013), in dominant-plausible and subordinate-plausible conditions (legend). Error bars indicate standard deviations over 10 model runs.***

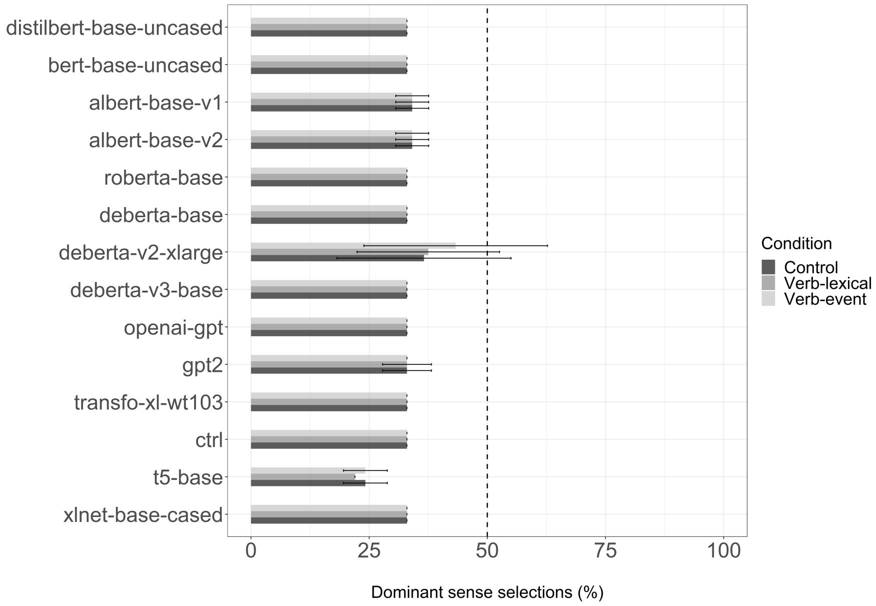

Figure S3.3. ***Mean percentage of dominant sense selections in randomly initialized models for Cabiddu et al. (2022), in dominant-plausible (Verb-lexical, Verb-event) and subordinate-plausible conditions (Control). Error bars indicate standard deviations over 10 model runs.***

References (for Appendix S3)

Cabiddu, F., Bott, L., Jones, G., & Gambi, C. (2022). The Role of Verb-Event Structure in Children's Lexical Ambiguity Resolution. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*(44). https://escholarship.org/uc/item/9kh29212

Rabagliati, H., Pylkkänen, L., & Marcus, G. F. (2013). Top-down influence in young children's linguistic ambiguity resolution. *Developmental Psychology*, *49*, 1076–1089. https://doi.org/10.1037/a0026918

Appendix S4: Relative Difference Outcome Measure

In this section, we present the results concerning the evaluation of dominance sense preference in Transformers, using child-based prototypes. This section additionally includes plots illustrating the raw performance of each model in each of the three experiments considered. Moreover, we report the output of statistical models, where the comparison between children and models' performance is made using a measure of relative difference as the outcome (see main manuscript for details about this measure).

Dominant Bias

Table S4.1. ***Model comparison between nested linear mixed-effect models via likelihood ratio test. The outcome is the percentage of dominant senses selected across conditions of Rabagliati et al. (2013) experiment 1. The predictors are log model size, log pretraining size, and their interaction. Model family was used as random effect intercept. The Null model only includes main and random effect intercepts. Subsequent models add one predictor at a time. The table shows the number of model parameters (npar), Akaike (AIC) and Bayesian (BIC) Information criterions, log-likelihood (logLik), deviance, Chi-square statistic (Chisq), degrees of freedom (Df), and p value Pr(>Chisq).***

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
Null model	3	313.52	318.94	-153.76	307.52	-	-	-
+ Model size	4	304.41	311.63	-148.20	296.41	11.11	1	**0.001**
+ Pretraining	5	293.23	302.26	-141.61	283.23	13.18	1	**0.000**
+ Interaction	6	294.46	305.30	-141.23	282.46	0.76	1	0.382

Table S4.2. ***Output of the best model selected via model comparison in Table S4.1***

	"+ Pretraining" Model Dominant sense preference		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	60.89	53.53 – 68.26	**<0.001**
Model size [log]	-1.47	-3.01 – 0.08	0.062
Pretraining size [log]	-1.53	-2.30 – -0.75	**<0.001**
Random Effects			
σ^2	25.86		
$\tau_{00\ \mathrm{family}}$	12.47		
ICC	0.33		
N_{family}	14		
Observations	45		
Marginal R^2 / Conditional R^2	0.491 / 0.657		

Rabagliati et al. (2013) – Experiment 1

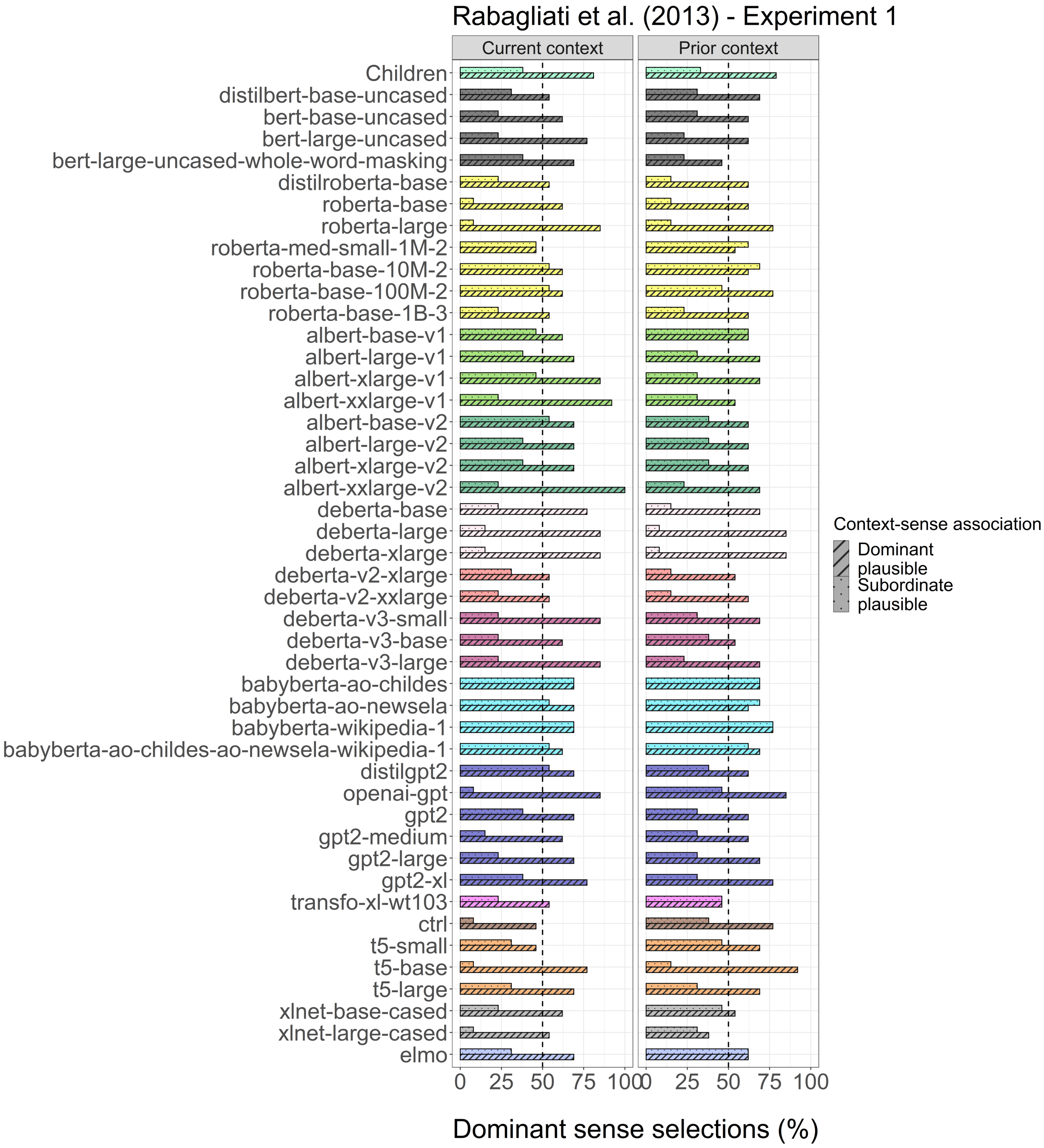

Figure S4.1. ***Percentage of dominant sense selections in models and children for experiment 1 of Rabagliati et al. (2013), in dominant-plausible and subordinate-plausible conditions (legend), when disambiguation cues were included in current or prior context (left and right panel respectively).***

Table S4.3. ***Model comparison between nested linear mixed-effect models via likelihood ratio test. The outcome is the relative difference between models and children in Rabagliati et al. (2013) experiment 1, see our main paper for more details about this outcome measure. The predictors are condition (Prior or Current context), log pretraining size, log model size , and their pairwise interactions. The random effect intercept is Model Family. The Null model includes main and random effect intercepts. Subsequent models add one predictor at a time. The table shows the number of model parameters (npar), Akaike (AIC) and Bayesian (BIC) Information criterions, log-likelihood (logLik), deviance, Chi-square statistic (Chisq), degrees of freedom (Df), and p value Pr(>Chisq).***

	npar	**AIC**	**BIC**	**logLik**	**deviance**	**Chisq**	**Df**	**Pr(>Chisq)**
Null model	3	807.62	815.12	-400.81	801.62	NA	NA	NA
+ Condition	4	803.09	813.09	-397.55	795.09	6.53	1	**0.011**
+ Pretraining	5	771.98	784.47	-380.99	761.98	33.12	1	**0.000**
+ Model size	6	764.64	779.64	-376.32	752.64	9.33	1	**0.002**
+ Pretraining*Condition	7	766.34	783.84	-376.17	752.34	0.30	1	0.584
+ Size*Condition	8	768.04	788.04	-376.02	752.04	0.30	1	0.584
+ Pretraining*Model size	9	769.64	792.14	-375.82	751.64	0.40	1	0.529

Table S4.4. ***Output of the best model selected via model comparison in Table S4.3.***

	'+ Model size' model Rabagliati et al. (2013) - Experiment 1		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-43.49	-59.94 – -27.04	**<0.001**
Model size [log]	5.36	2.07 – 8.64	**0.002**
Pretraining size [log]	3.81	2.16 – 5.47	**<0.001**
Condition [Prior context]	-9.98	-16.18 – -3.78	**0.002**
Random Effects			
σ^2	218.67		
$\tau_{00\ \text{family}}$	85.81		
ICC	0.28		
N_{family}	14		
Observations	90		
Marginal R^2 / Conditional R^2	0.510 / 0.648		

Rabagliati et al. (2013) – Experiment 2

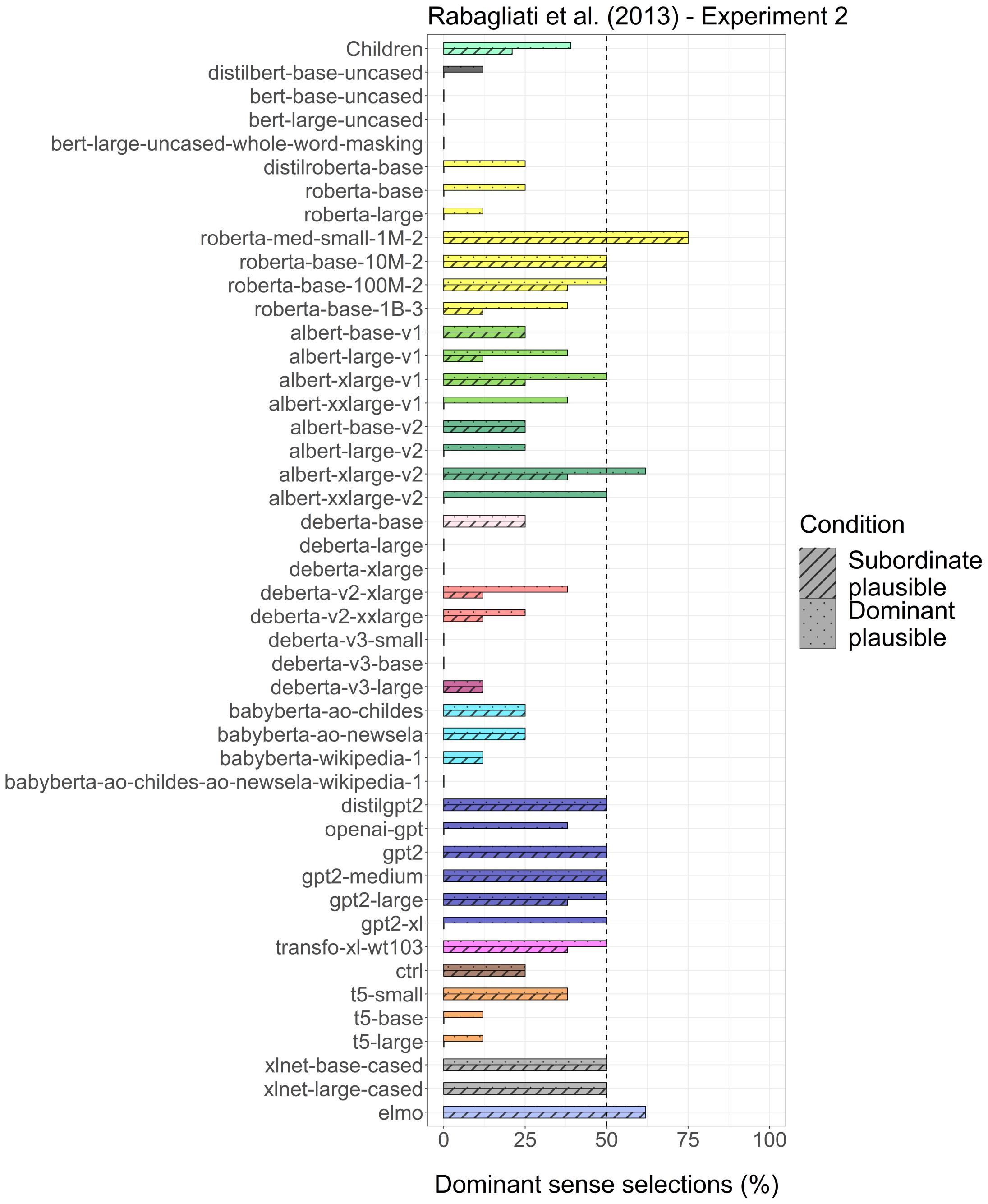

Figure S4.2. ***Percentage of dominant sense selections in models and children for experiment 2 of Rabagliati et al. (2013), in dominant-plausible and subordinate-plausible conditions (legend).***

Table S4.5. ***Model comparison between nested linear mixed-effect models via likelihood ratio test. The outcome is the relative difference between models and children in Rabagliati et al. (2013) experiment 2, see our main paper for more details about this outcome measure. The predictors are log pretraining size, log model size, and their interaction. The random effect intercept is Model Family. The Null model includes main and random effect intercepts. Subsequent models add one predictor at a time. The table shows the number of model parameters (npar), Akaike (AIC) and Bayesian (BIC) Information criterions, log-likelihood (logLik), deviance, Chi-square statistic (Chisq), degrees of freedom (Df), and p value Pr(>Chisq).***

	npar	**AIC**	**BIC**	**logLik**	**deviance**	**Chisq**	**Df**	**Pr(>Chisq)**
Null model	3	371.50	376.92	-182.75	365.50	-	-	-
+ Pretraining	4	372.70	379.93	-182.35	364.70	0.80	1	0.371
+ Model size	5	372.27	381.30	-181.13	362.27	2.44	1	0.119
+ Interaction	6	373.79	384.63	-180.89	361.79	0.48	1	0.489

Table S4.6. ***Although no model surpassed the Null model in Table S4.5, below we show the output of the model including both main effects of model size and pretraining size, to appreciate size of the estimates and variance explained.***

	'+ Model size' model Rabagliati et al. (2013) - Experiment 2		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-25.51	-43.38 – -7.63	**0.006**
Model size [log]	3.37	-0.35 – 7.09	0.075
Pretraining size [log]	0.12	-1.74 – 1.98	0.895
Random Effects			
σ^2	146.43		
$\tau_{00\ \text{family}}$	79.91		
ICC	0.35		
N_{family}	14		
Observations	45		
Marginal R^2 / Conditional R^2	0.105 / 0.421		

Cabiddu et al. (2022)

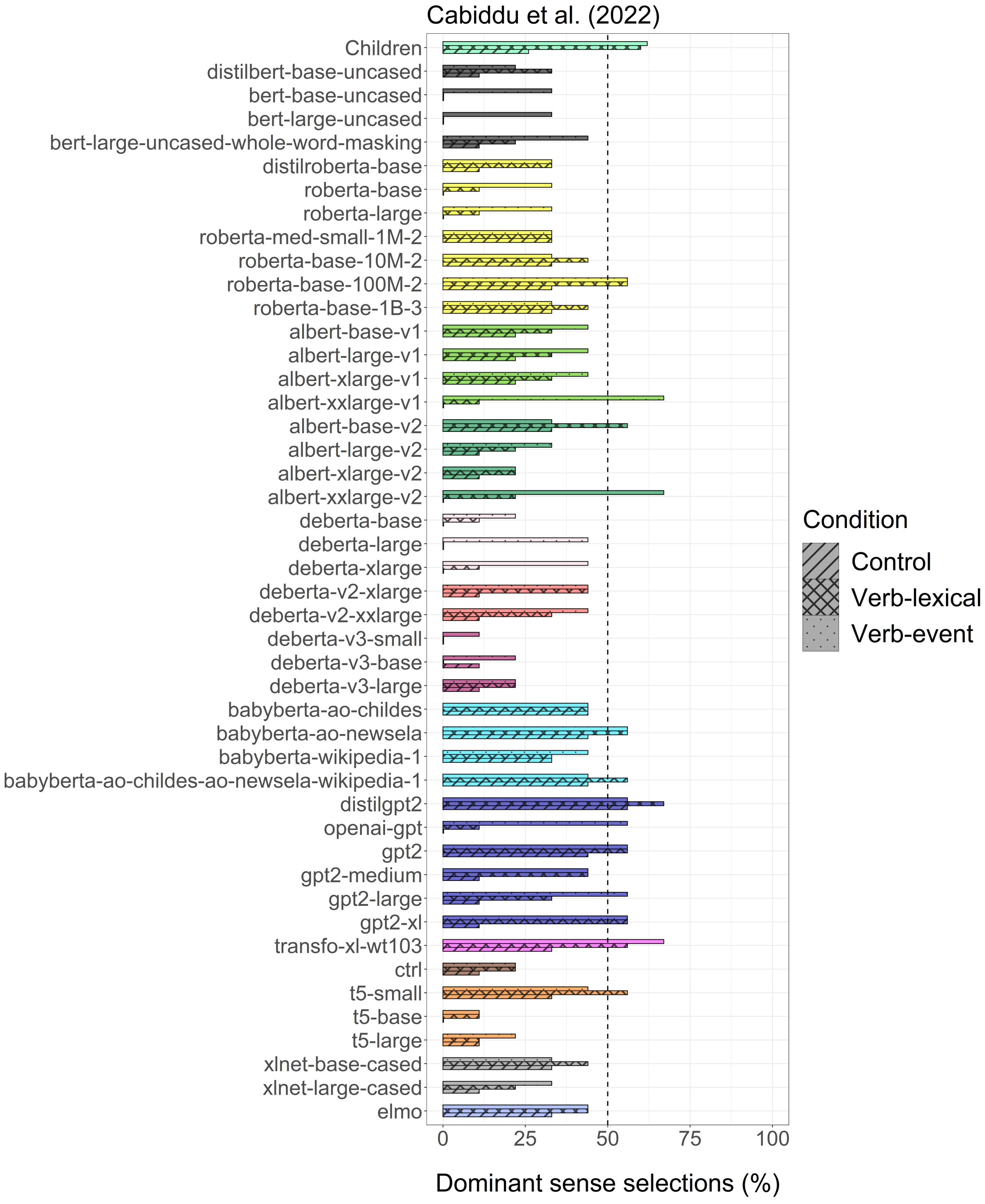

Figure S4.3. ***Percentage of dominant sense selections in models for Cabiddu et al. (2022), in dominant-plausible (Verb-lexical, Verb-event) and subordinate-plausible conditions (Control).***

Table S4.7. ***Model comparison between nested linear mixed-effect models via likelihood ratio test. The outcome is the relative difference between models and children in Cabiddu et al. (2022), when considering performance in the Verb-Event structure condition. See our main paper for more details about this outcome measure. The predictors are log pretraining size, log model size, and their interaction. The random effect intercept is Model Family. The Null model includes main and random effect intercepts. Subsequent models add one predictor at a time. The table shows the number of model parameters (npar), Akaike (AIC) and Bayesian (BIC) Information criterions, log-likelihood (logLik), deviance, Chi-square statistic (Chisq), degrees of freedom (Df), and p value Pr(>Chisq).***

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
Null model	3	390.69	396.11	-192.34	384.69	-	-	-
+ Pretraining	4	390.37	397.60	-191.19	382.37	2.32	1	0.128
+ Model size	5	381.28	390.31	-185.64	371.28	11.09	1	**0.001**
+ Interaction	6	382.77	393.61	-185.39	370.77	0.51	1	0.477

Table S4.8. ***Output of the best model selected via model comparison in Table S4.7.***

		'+ Model size' model Verb-Event Condition Cabiddu et al. (2022)	
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-50.56	-69.91 – -31.20	**<0.001**
Model size [log]	7.57	3.48 – 11.67	**0.001**
Pretraining size [log]	-0.30	-2.35 – 1.74	0.765
Random Effects			
σ^2	188.50		
$\tau_{00\ \text{family}}$	76.42		
ICC	0.29		
N_{family}	14		
Observations	45		
Marginal R^2 / Conditional R^2	0.300 / 0.502		

Table S4.9. ***Model comparison between nested linear mixed-effect models via likelihood ratio test. The outcome is the relative difference between models and children in Cabiddu et al. (2022), when considering performance in the Verb-Lexical condition. See our main paper for more details about this outcome measure. The predictors are log pretraining size, log model size, and their interaction. The random effect intercept is Model Family. The Null model includes main and random effect intercepts. Subsequent models add one predictor at a time. The table shows the number of model parameters (npar), Akaike (AIC) and Bayesian (BIC) Information criterions, log-likelihood (logLik), deviance, Chi-square statistic (Chisq), degrees of freedom (Df), and p value Pr(>Chisq).***

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
Null model	3	340.21	345.63	-167.10	334.21	-	-	-
+ Pretraining	4	341.21	348.43	-166.60	333.21	1.00	1	0.317
+ Model size	5	341.23	350.27	-165.62	331.23	1.97	1	0.160
+ Interaction	6	342.00	352.84	-165.00	330.00	1.23	1	0.267

Table S4.10. ***Although no model surpassed the Null model in Table S4.9, below we show the output of the model including both main effects of model size and pretraining size, to appreciate size of the estimates and variance explained.***

	‘+Model size’ model Verb-Lexical Condition Cabiddu et al. (2022)		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-30.39	-42.47 – -18.31	**<0.001**
Model size [log]	1.73	-0.87 – 4.34	0.186
Pretraining size [log]	0.16	-1.14 – 1.45	0.809
Random Effects			
σ^2	81.87		
$\tau_{00\ \text{family}}$	22.74		
ICC	0.22		
N_{family}	14		
Observations	45		
Marginal R^2 / Conditional R^2	0.071 / 0.273		

References (for Appendix S4)

Cabiddu, F., Bott, L., Jones, G., & Gambi, C. (2022). The Role of Verb-Event Structure in Children's Lexical Ambiguity Resolution. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*(44). https://escholarship.org/uc/item/9kh29212

Rabagliati, H., Pylkkänen, L., & Marcus, G. F. (2013). Top-down influence in young children's linguistic ambiguity resolution. *Developmental Psychology*, *49*, 1076–1089. https://doi.org/10.1037/a0026918

Appendix S5: Euclidean Distance Outcome Measure

In this section, we report results of the three experiments using an alternative outcome measure. See details about this measure in the main manuscript. In Figure S5.1, we show an example of how the measure is computed.

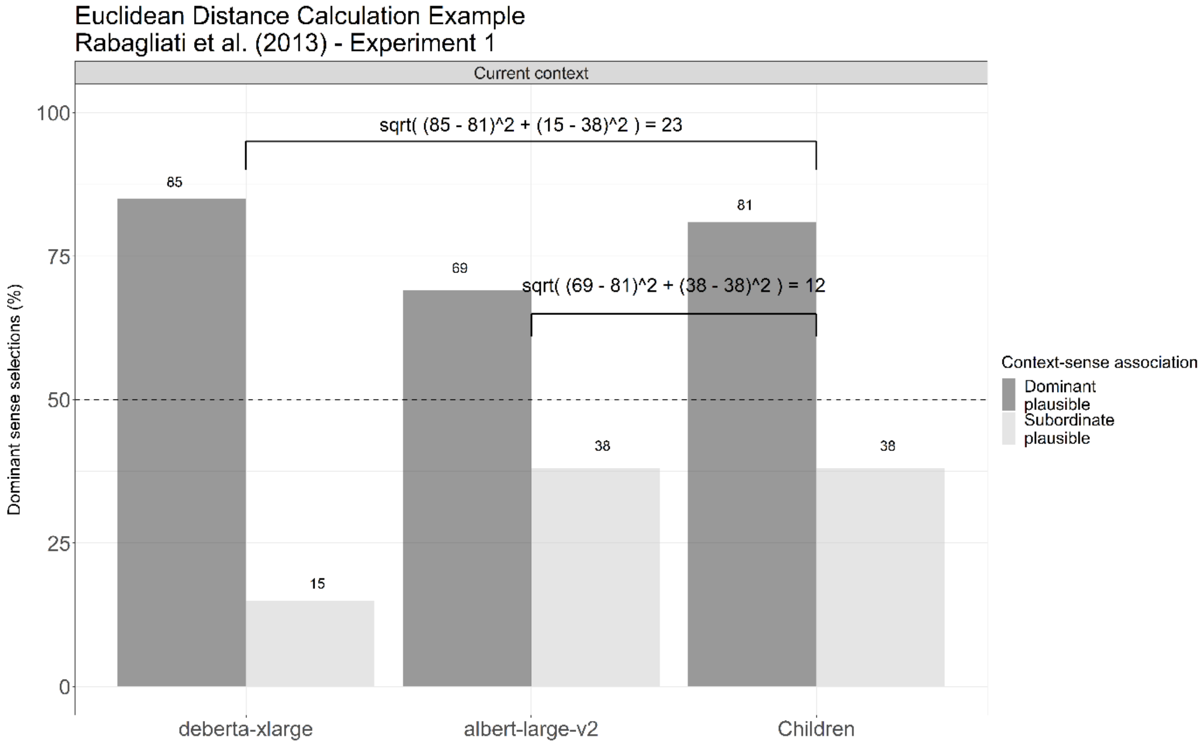

Figure S5.1. ***Example of calculation of the Euclidean Distance of deberta-xlarge and albert-large-v2 from children's scores in the Current Context condition of Rabagliati et al. (2013) experiment 1. The measure looks at the exact match between model and children.***

Rabagliati et al. (2013) – Experiment 1

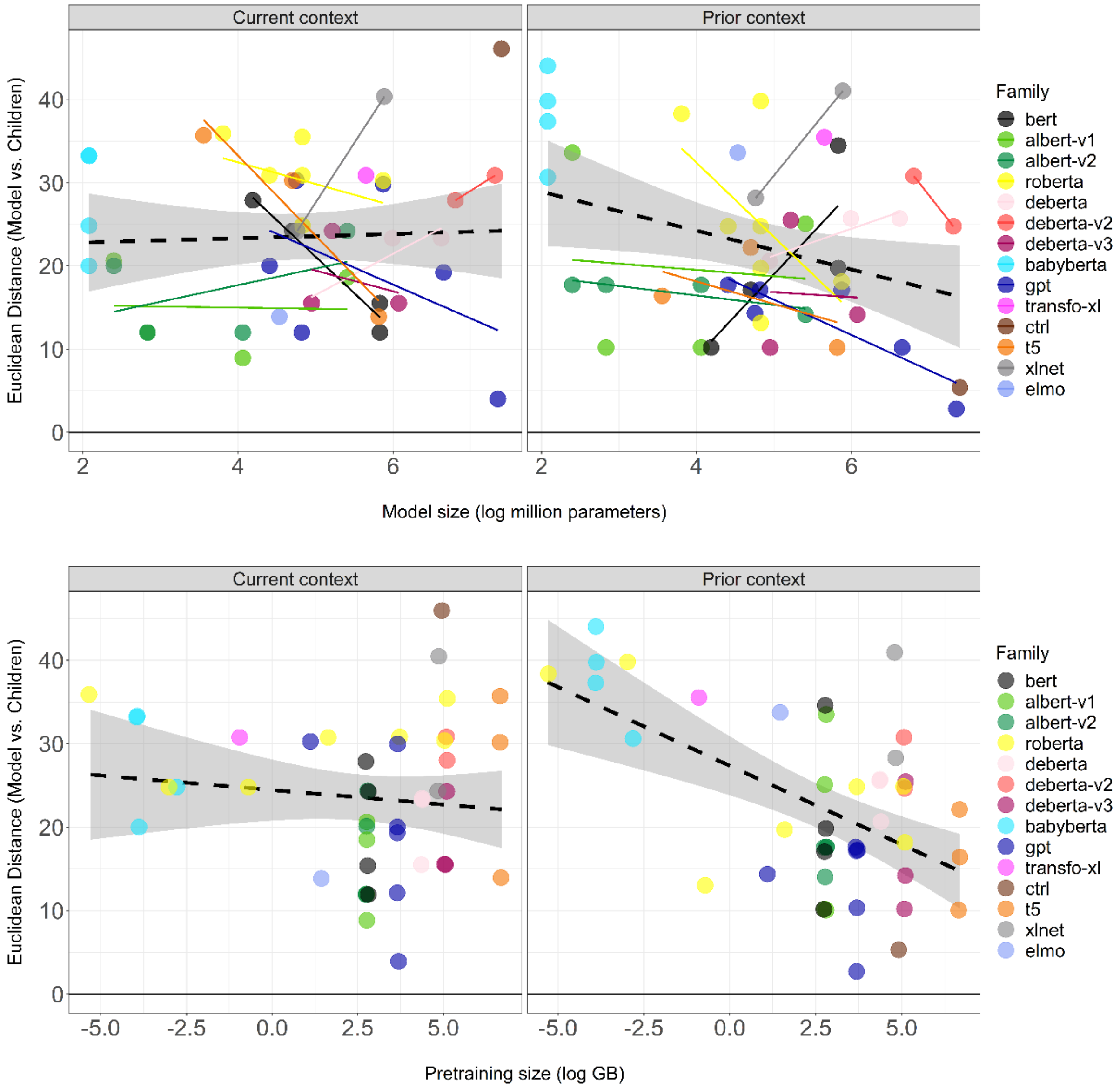

Figure S5.2. ***Models' Euclidean distance from children by model size (top row) and pretraining size (bottom row), in current and prior context conditions. Model families are shown in the legend. The black horizontal line (y=0) indicates child performance. The dashed regression line with 95% confidence interval shows performance across models. Colored regression lines are also shown for each model family, although only when examining model size as there is almost null variation in pretraining size within family.***

Table S5.1. ***Model comparison between nested linear mixed-effect models via likelihood ratio test. The outcome is the Euclidean Distance between models and children in Rabagliati et al. (2013) experiment 1. See our main paper for more details about this outcome measure. The predictors are dominant bias, condition (current, prior context), log pretraining size, log model size, and the pairwise interactions between model size, pretraining size, and condition. The random effect intercept is Model Family.***

	npar	**AIC**	**BIC**	**logLik**	**deviance**	**Chisq**	**Df**	**Pr(>Chisq)**
Null model	3	657.12	664.62	-325.56	651.12	-	-	-
+ Dominant Bias	4	658.23	668.23	-325.12	650.23	0.88	1	0.347
+ Condition	5	659.79	672.29	-324.89	649.79	0.44	1	0.505
+ Pretraining	6	646.98	661.98	-317.49	634.98	14.81	1	**0.000**
+ Model size	7	647.70	665.19	-316.85	633.70	1.28	1	0.257
+ Pretraining*Condition	8	641.30	661.30	-312.65	625.30	8.39	1	**0.004**
+ Size*Condition	9	642.39	664.89	-312.19	624.39	0.91	1	0.339
+ Pretraining*Model size	10	640.70	665.70	-310.35	620.70	3.68	1	0.055

Table S5.2. ***Output of the best model selected via model comparison in Table S5.1.***

	'Pretraining * Condition' model **Euclidean Distance** **Rabagliati et al. (2013) - experiment 1**		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	60.13	38.98 – 81.28	**<0.001**
Model size [log]	-0.95	-2.61 – 0.71	0.259
Pretraining size [log]	-1.03	-2.10 – 0.05	0.060
Condition [Prior context]	2.91	-1.31 – 7.12	0.173
Dominant bias	-0.57	-0.90 – -0.25	**0.001**
Pretraining size [log] * condition [Prior context]	-1.54	-2.60 – -0.48	**0.005**
Random Effects			
σ^2	56.96		
$\tau_{00\ \text{family}}$	13.78		
ICC	0.19		
N_{family}	14		
Observations	90		
Marginal R^2 / Conditional R^2	0.271 / 0.413		

Rabagliati et al. (2013) – Experiment 2

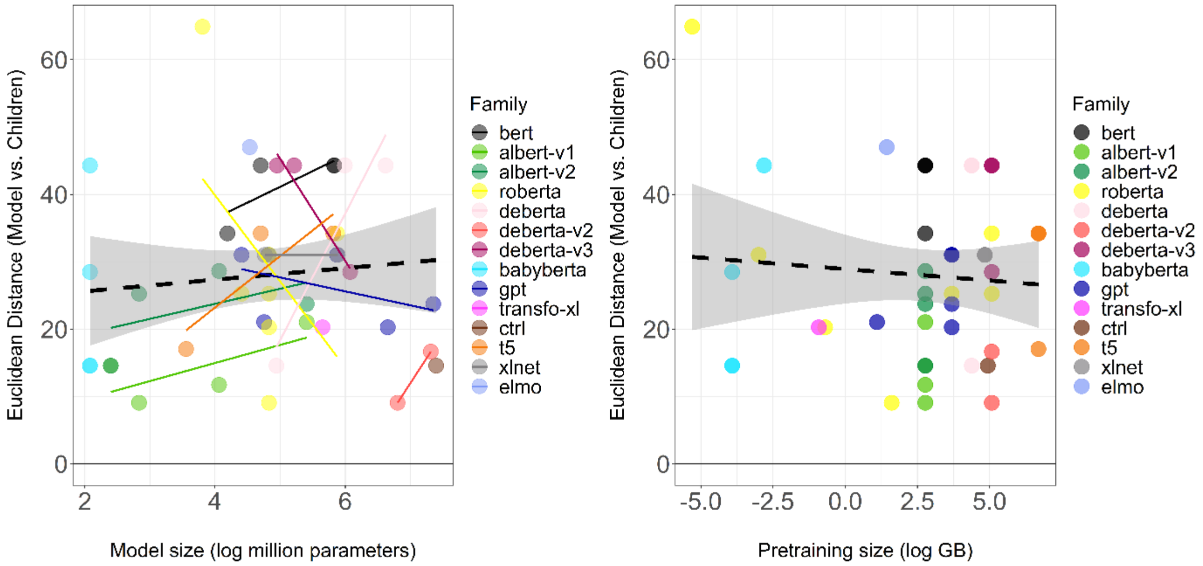

Figure S5.3. ***Models' Euclidean distance from children by model size and pretraining size in Rabagliati et al. (2013) experiment 2. Model families are shown in the legend. The black horizontal line (y=0) indicates child performance. The dashed regression line with 95% confidence interval shows performance across models. Colored regression lines are also shown for each model family, although only when examining model size as there is almost null variation in pretraining size within family.***

Table S5.3. ***Model comparison between nested linear mixed-effect models via likelihood ratio test. The outcome is the Euclidean Distance between models and children in Rabagliati et al. (2013) experiment 2. The predictors are dominant bias, log pretraining size, log model size, and the interaction between model size and pretraining size. The random effect intercept is Model Family.***

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
Null model	3	359.62	365.04	-176.81	353.62	-	-	-
+ Dominant Bias	4	361.55	368.78	-176.78	353.55	0.07	1	0.789
+ Pretraining	5	362.38	371.42	-176.19	352.38	1.17	1	0.280
+ Model size	6	363.65	374.49	-175.82	351.65	0.73	1	0.391
+ Model size * Pretraining	7	365.63	378.28	-175.82	351.63	0.02	1	0.895

Table S5.4. ***Although no model surpassed the Null model in Table S5.3, below we show the output of the model including the main effects, to appreciate size of the estimates and variance explained.***

	'+ Model size' model Euclidean Distance Rabagliati et al. (2013) - Experiment 2		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	27.14	-16.95 – 71.23	0.221
Model size [log]	1.41	-2.08 – 4.90	0.418
Pretraining size [log]	-1.26	-3.23 – 0.70	0.202
Dominant bias	-0.05	-0.73 – 0.62	0.876
Random Effects			
σ^2	120.55		
$\tau_{00\ \mathrm{family}}$	59.86		
ICC	0.33		
N family	14		
Observations	45		
Marginal R^2 / Conditional R^2	0.049 / 0.365		

Cabiddu et al. (2022)

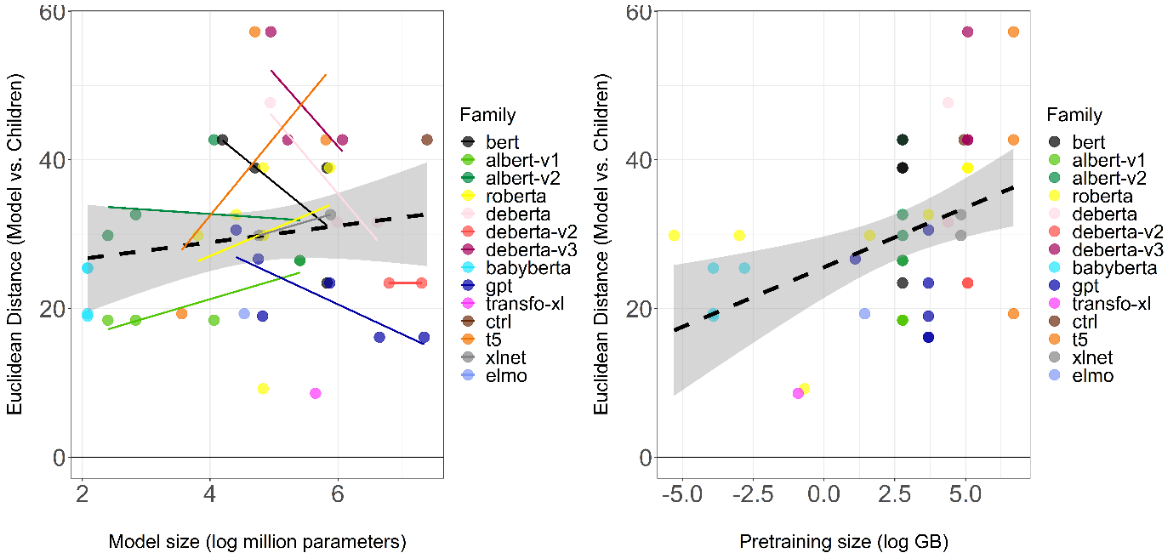

Figure S5.4. ***Models' Euclidean distance from children by model size and pretraining size, when comparing verb-event vs. control conditions in Cabiddu et al. (2022).***

Table S5.5. ***Model comparison between nested linear mixed-effect models via likelihood ratio test. The outcome is the Euclidean Distance between models and children in Cabiddu et al. (2022) when comparing Verb-Event condition to Control. The predictors are dominant bias, log pretraining size, log model size, and the interaction between model size and pretraining size. The random effect intercept is Model Family. The Null model includes main and random effect intercepts. Subsequent models add one predictor at a time. The table shows the number of model parameters (npar), Akaike (AIC) and Bayesian (BIC) Information criterions, log-likelihood (logLik), deviance, Chi-square statistic (Chisq), degrees of freedom (Df), and p value Pr(>Chisq).***

	npar	**AIC**	**BIC**	**logLik**	**deviance**	**Chisq**	**Df**	**Pr(>Chisq)**
Null model	3	344.69	350.11	-169.34	338.69	NA	NA	NA
+ Dominant Bias	4	343.94	351.17	-167.97	335.94	2.75	1	0.097
+ Pretraining	5	341.72	350.75	-165.86	331.72	4.22	1	**0.040**
+ Model size	6	343.04	353.88	-165.52	331.04	0.68	1	0.411
+ Model size * Pretraining	7	342.91	355.55	-164.45	328.91	2.14	1	0.144

Table S5.6. ***Output of the best model selected via model comparison in Table S5.5.***

		'+ Pretraining' model Euclidean Distance Verb-Event vs. Control Cabiddu et al. (2022)	
Predictors	*Estimates*	*CI*	*p*
(Intercept)	26.71	-1.86 – 55.28	0.066
Pretraining size [log]	1.54	0.01 – 3.07	**0.048**
Dominant bias	-0.02	-0.54 – 0.49	0.929
Random Effects			
σ^2	74.61		
$\tau_{00\ \text{family}}$	39.70		
ICC	0.35		
N_{family}	14		
Observations	45		
Marginal R^2 / Conditional R^2	0.167 / 0.456		

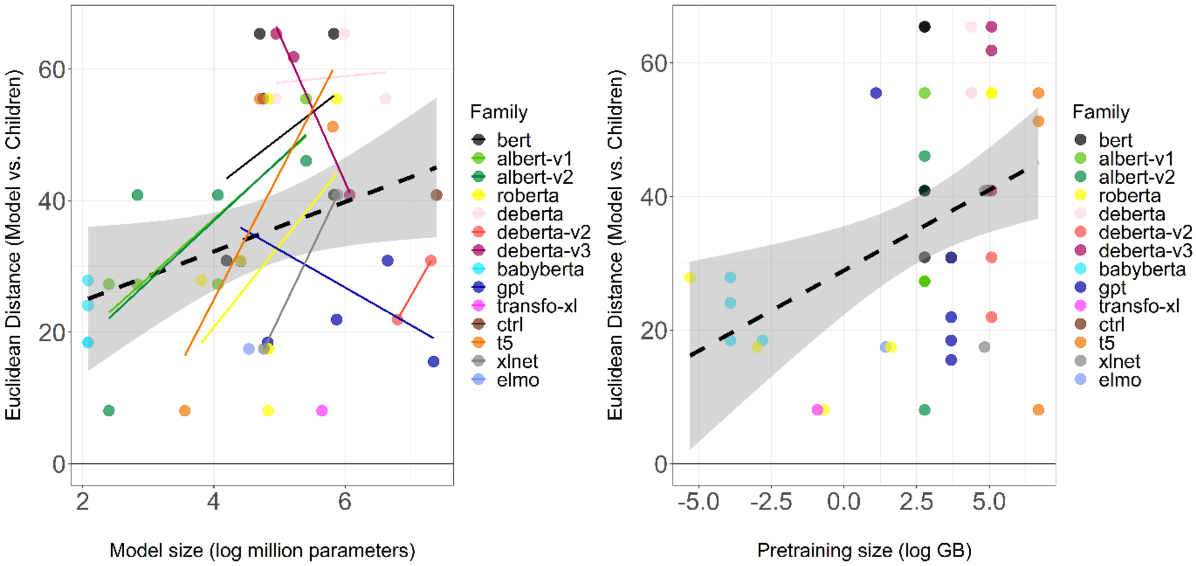

Figure S5.5. ***Models' Euclidean distance from children by model size and pretraining size, when comparing verb-lexical vs. control conditions in Cabiddu et al. (2022).***

Table S5.7. ***Model comparison between nested linear mixed-effect models via likelihood ratio test. The outcome is the Euclidean Distance between models and children in Cabiddu et al. (2022) when comparing Verb-Lexical condition to Control. The predictors are dominant bias, log pretraining size, log model size, and the interaction between model size and pretraining size. The random effect intercept is Model Family. The Null model includes main and random effect intercepts. Subsequent models add one predictor at a time. The table shows the number of model parameters (npar), Akaike (AIC) and Bayesian (BIC) Information criterions, log-likelihood (logLik), deviance, Chi-square statistic (Chisq), degrees of freedom (Df), and p value Pr(>Chisq).***

	npar	**AIC**	**BIC**	**logLik**	**deviance**	**Chisq**	**Df**	**Pr(>Chisq)**
Null model	3	390.83	396.25	-192.42	384.83	NA	NA	NA
+ Dominant Bias	4	389.72	396.95	-190.86	381.72	3.11	1	0.078
+ Pretraining	5	387.82	396.85	-188.91	377.82	3.90	1	**0.048**
+ Model size	6	389.03	399.87	-188.51	377.03	0.79	1	0.374
+ Model size * Pretraining	7	389.61	402.26	-187.81	375.61	1.41	1	0.234

Table S5.8. ***Output of the best model selected via model comparison in Table S5.7.***

	'+ Pretraining' model Euclidean Distance Verb-Lexical vs. Control Cabiddu et al. (2022b)		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	32.85	-13.44 – 79.14	0.159
Pretraining size [log]	2.28	-0.13 – 4.69	0.063
Dominant bias	-0.08	-0.91 – 0.76	0.853
Random Effects			
σ^2	242.72		
$\tau_{00\ \mathrm{family}}$	44.56		
ICC	0.16		
N_{family}	14		
Observations	45		
Marginal R^2 / Conditional R^2	0.158 / 0.289		

References (for Appendix S5)

Cabiddu, F., Bott, L., Jones, G., & Gambi, C. (2022). The Role of Verb-Event Structure in Children's Lexical Ambiguity Resolution. *Proceedings of the Annual Meeting of the Cognitive Science Society, 44*(44). https://escholarship.org/uc/item/9kh29212

Rabagliati, H., Pylkkänen, L., & Marcus, G. F. (2013). Top-down influence in young children's linguistic ambiguity resolution. *Developmental Psychology, 49*, 1076–1089. https://doi.org/10.1037/a0026918

Appendix S6: Simulations using adult-based sense prototypes

This section presents supplemental results obtained by using adult-directed speech to compute sense prototypes prior to testing the 45 Transformers in the word sense disambiguation tasks.

We initially explain how adult-directed speech was sense-tagged. Subsequently, we present plots showing the raw performance on the three experimental tasks for each adult-based Transformer.

This is followed by comparisons between adult-based models and the previously employed child-based models. For these comparisons, a preliminary examination was conducted to determine if adult-based models demonstrated superior performance than child-based models at the condition level, specifically looking at the percentage of correct responses given by a model in each experimental condition (note that this measure is independent of child performance).

Finally, we examined whether the child-based models better fit the children's data than the adult-based models. We begin by demonstrating that adult-based models did not display any dominance sense preference, thus highlighting the importance of using child-directed speech to derive child-based sense prototypes that reflect sense frequencies in the child input. We then show that child-based models better fit children's data in coherent tasks but not contrastive ones.

Sense Tagging the Spoken BNC

A question left open by previous analyses is whether the suboptimal performance of Transformers in contrastive tasks might be due to the use of sense prototypes computed from sense-tagged child-directed speech. Thus, it is possible that the models could perform better when their prototypes are based on adult-directed speech. Alternatively, the models may face difficulties with contrastive tasks for other reasons, such as a lack of real-world inference skills or multimodal data.

To build adult-based prototypes, we sense-tagged 80 utterances for each target word used in the study (40 utterances per sense). We extracted these utterances (available in our GitHub page) from adult-adult conversations present in the spoken section of the British National Corpus (BNC Consortium, 2007). One target word, *turkey*, had to be discarded because no utterances were available for one of its senses. For an additional four words, the input contained fewer than 40 utterances for one of the senses. Despite this, we used the number of utterances available and retained these target words in order to maximize the sample of items. In one case, a sense received a very low number of input utterances ($n = 3$). However, this was still retained on the basis that $n = 3$ is considered the minimum acceptable number to make sense prototypes functional in sense disambiguation (e.g., Loureiro et al., 2021). The frequencies of each tagged sense in the new adult input are displayed below in Table S6.1.

Table S6.1. ***For each target word's sense, the table displays the number of utterances tagged from the Spoken BNC.***

Target Word	Sense	n
band	music_group	40
band	object	40
bat	animal	9
bat	object	35
bow	knot	34
bow	weapon	3
button	clothing	40
button	tech	40
card	note	40
card	playing	40
chicken	animal	34
chicken	food	40
fish	animal	40
fish	food	40
glasses	drinking	40
glasses	eye	40
lamb	animal	24
lamb	food	40
letter	alphabet	40
letter	mail	40
line	geometry	40
line	order	40
nail	body_part	40
nail	object	40

Plots of Dominant Sense Selection – Raw Performance of Adult-Based Models

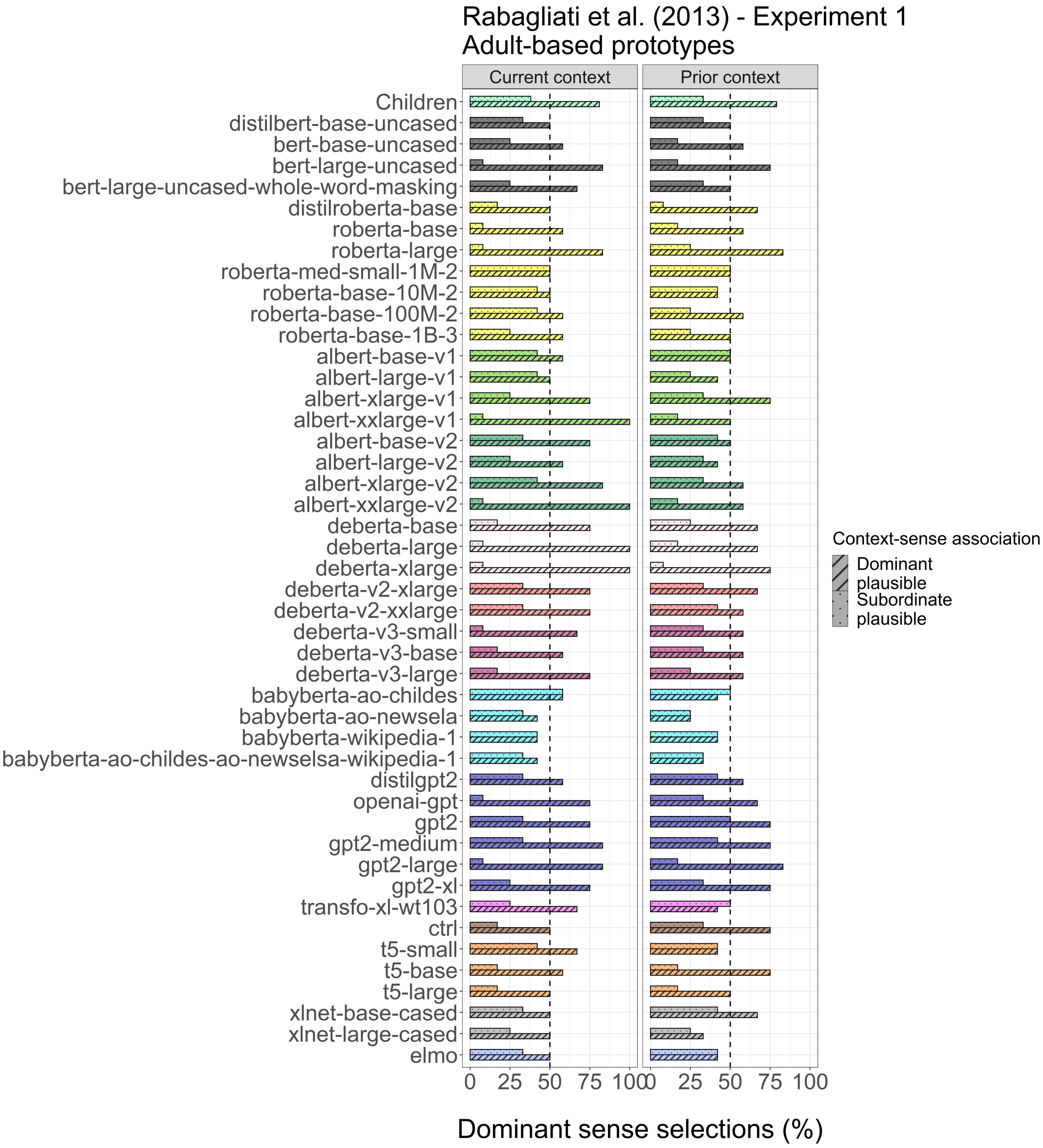

Figure S6.1. ***Percentage of dominant sense selections in adult-based models and children for experiment 1 of Rabagliati et al. (2013), in dominant-plausible and subordinate-plausible conditions (legend), when disambiguation cues were included in current or prior context (left and right panel respectively).***

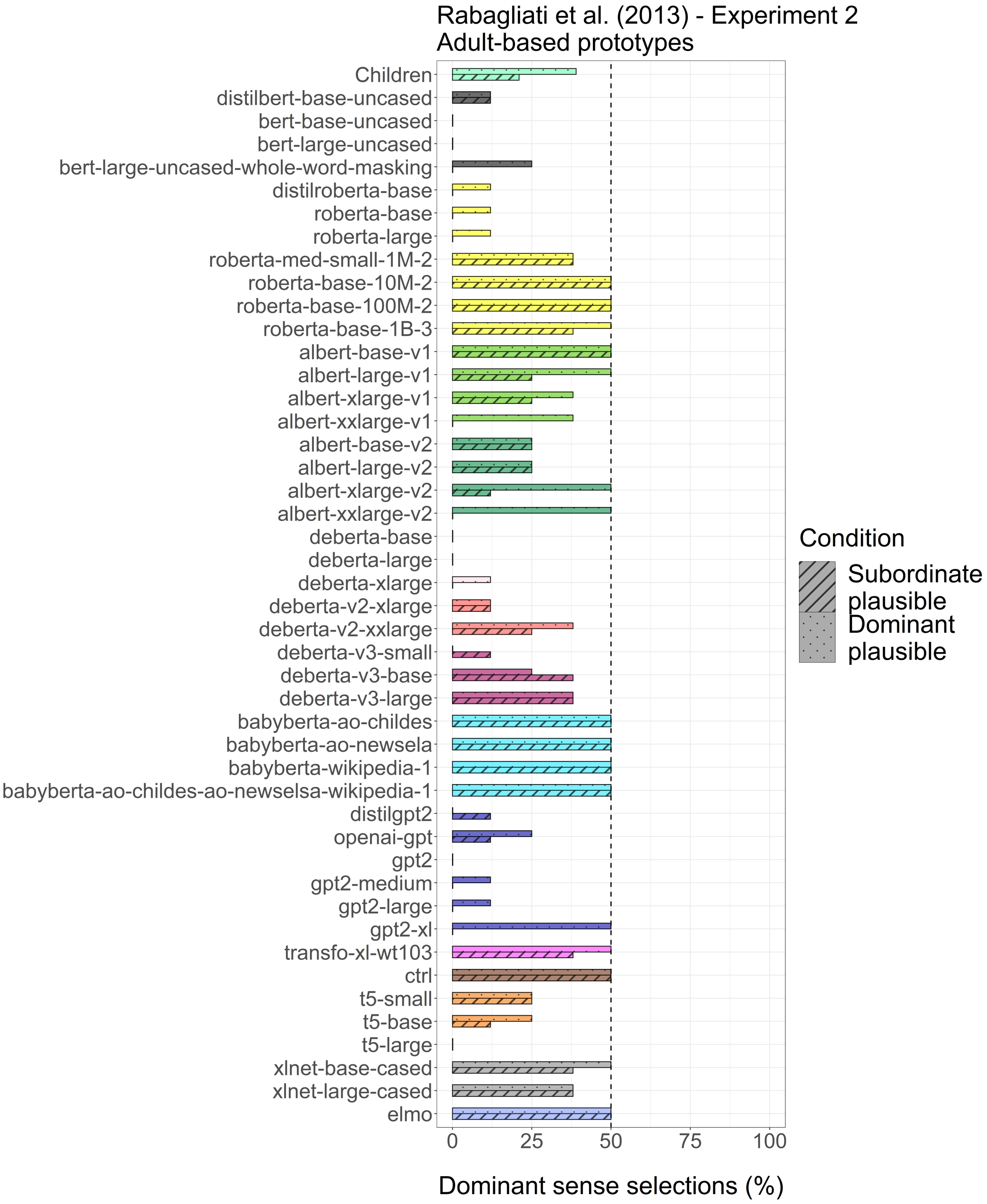

Figure S6.2. ***Percentage of dominant sense selections in adult-based models and children for experiment 2 of Rabagliati et al. (2013), in dominant-plausible and subordinate-plausible conditions (legend).***

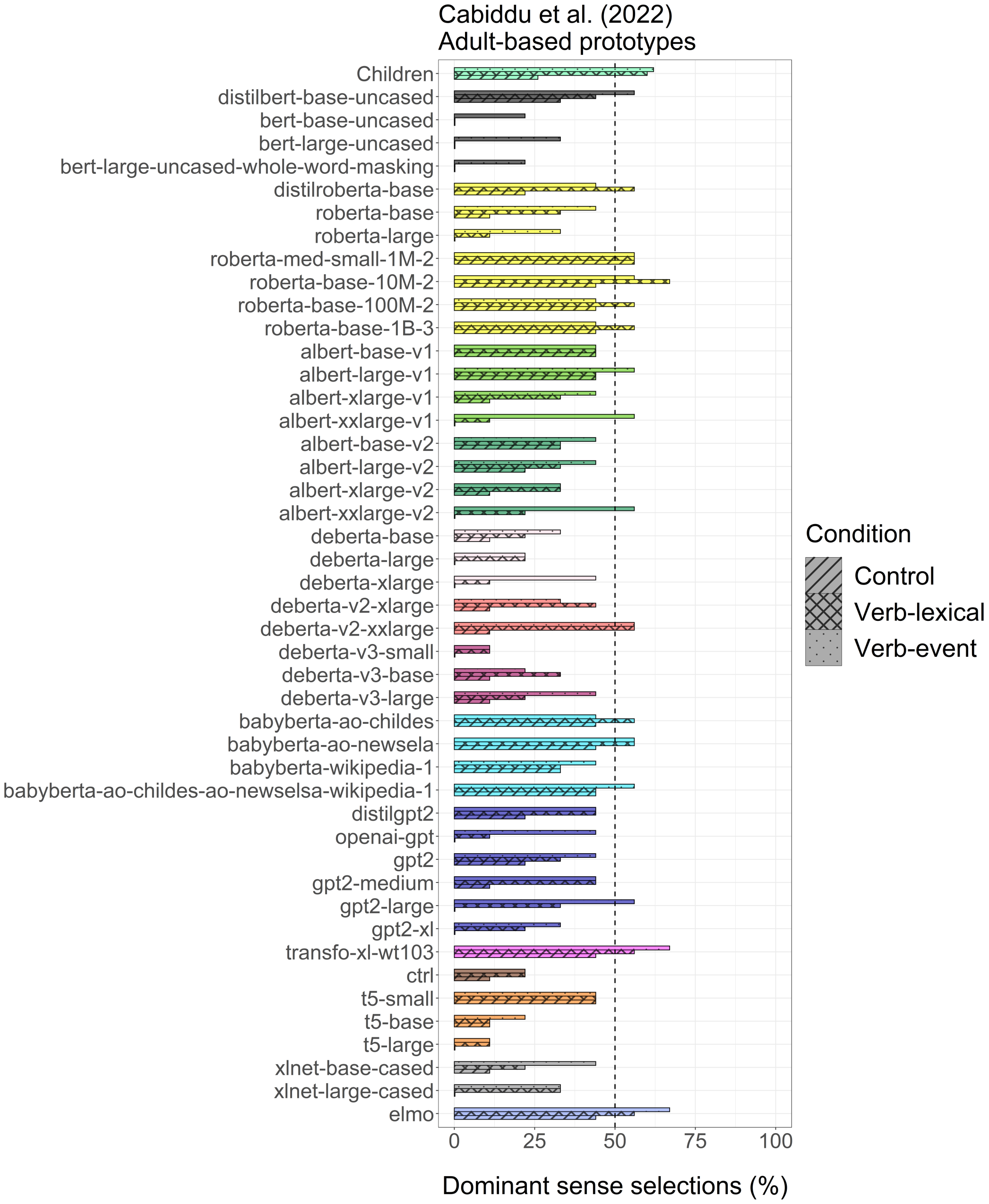

Figure S6.3. ***Percentage of dominant sense selections in adult-based models for Cabiddu et al. (2022), in dominant-plausible (Verb-lexical, Verb-event) and subordinate-plausible conditions (Control).***

Examining Correct Responses in Adult-Based and Child-Based Models

After implementing prototypes based on adult speech and rerunning the Transformers on the test stimuli, we examined the performance of the adult-based models in comparison to those child-based. As can be observed in Figure S6.4, the percentage of correct responses is remarkably similar across both age groups (adult-based models / child-based models) in each experiment considered. This reaffirms that the lower performance of Transformers in contrastive tasks, as seen in the Rabagliati Experiment 2 and Cabiddu Experiment 1, was not a consequence of deriving sense prototypes from child-directed speech.

It is important to note that this preliminary comparison does not take into account how closely the adult-based and child-based models approximate child performance. We relate the models' performance to child responses in the following section.

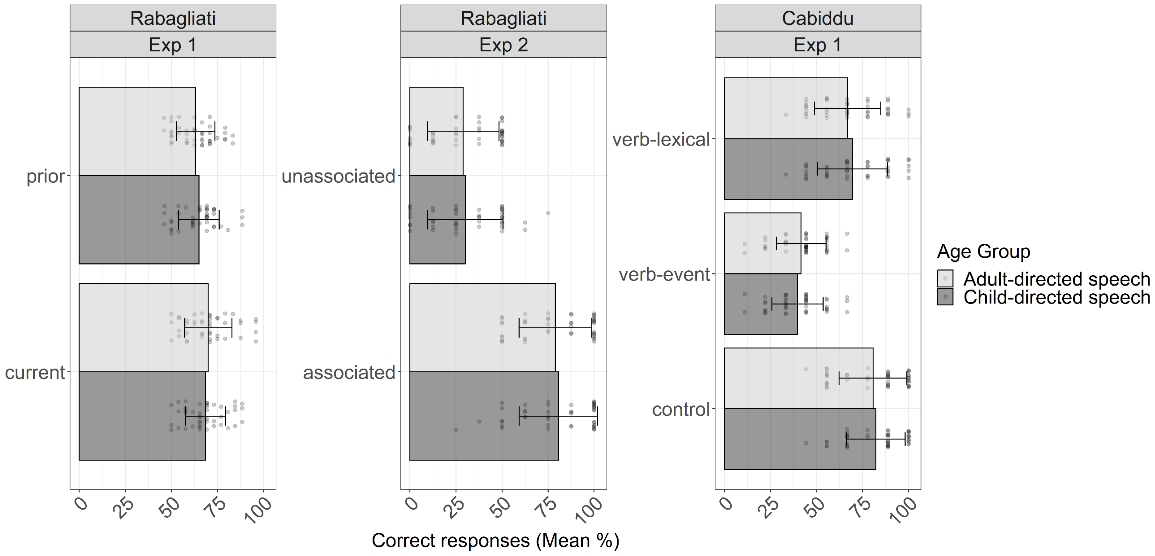

Figure S6.4. ***Mean percentage of correct responses (x-axis) in adult-based and child-based models (legend) for every condition (y-axis) in the behavioral experiments (panels). Error bars represent standard deviations around the mean percentages. Data points indicate performance for individual models in each condition.***

Relating adult-based and child-based models' performance to child responses

Dominant Sense Preference

First, we investigated whether the models based on adult speech showed any preference for dominant senses (e.g., *elastic band*) or a subordinate sense (e.g., *music band*) in the initial experiment, which used coherent sentences (Rabagliati et al., 2013; Study 1).

In the experimental study involving both adults and children (Cabiddu et al., 2022), dominance had a more pronounced effect on child performance compared to adult performance. One hypothesis suggested that the distribution of sense frequencies might not be identical in adult-directed speech as it is in child-directed speech. If this hypothesis were correct, we would anticipate that Transformers would exhibit a weak or null dominance bias when their prototypes are derived from adult input. As shown in the figure S6.5, a visual comparison between adult-based and child-based dominance preference supports this expectation: The models did not display a dominance preference when using prototypes based on adult data. Further, we found a significant difference in dominance preference between adult-based and child-based models, in interaction with both model size and pretraining size (see Table S6.2 and S6.3). This finding supports the hypothesis that the dominant bias identified in the models based on child data was likely a result of employing sense-tagged child-directed speech.

Table S6.2. ***Model comparison between nested linear mixed-effect models via likelihood ratio test. The outcome is the percentage of dominant senses selected across conditions of Rabagliati et al. (2013) experiment 1. The predictors are age group (adult-based model/child-based model), log pretraining size, log model size, and their interactions. Model family was used as random effect intercept. The Null model only includes main and random effect intercepts. Subsequent models add one predictor at a time. The table shows the number of model parameters (npar), Akaike (AIC) and Bayesian (BIC) Information criterions, log-likelihood (logLik), deviance, Chi-square statistic (Chisq), degrees of freedom (Df), and p value Pr(>Chisq).***

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
Null model	3	629.94	637.44	-311.97	623.94	NA	NA	NA
+ Age group	4	621.44	631.44	-306.72	613.44	10.50	1	**0.001**
+ Pretraining	5	613.08	625.58	-301.54	603.08	10.36	1	**0.001**
+ Model size	6	614.54	629.54	-301.27	602.54	0.54	1	0.461
+ Age group x Model size	7	590.41	607.91	-288.21	576.41	26.13	1	**0.000**
+ Age group x Pretraining	8	587.07	607.07	-285.53	571.07	5.34	1	**0.021**
+ Pretraining x Model size	9	589.05	611.55	-285.52	571.05	0.02	1	0.884
+ Age group x Pretraining x Model size	10	587.48	612.47	-283.74	567.48	3.57	1	0.059

Table S6.3. ***Output of the best model selected via model comparison in Table S6.2.***

	Dominant Sense Preference		
	'+ Age group * Pretraining' model		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	39.67	32.69 – 46.65	**<0.001**
Age group [Child-directed speech]	23.82	15.26 – 32.38	**<0.001**
Model size [log]	1.29	-0.23 – 2.80	0.096
Pretraining size [log]	-0.35	-1.10 – 0.40	0.354
Age group [Child-directed speech] × Model size [log]	-3.34	-5.28 – -1.39	**0.001**
Age group [Child-directed speech] × Pretraining size [log]	-1.08	-2.03 – -0.14	**0.025**
Random Effects			
σ^2	31.66		
$\tau_{00\ \text{family}}$	6.36		
ICC	0.17		
N $_{\text{family}}$	14		
Observations	90		
Marginal R^2 / Conditional R^2	0.428 / 0.524		

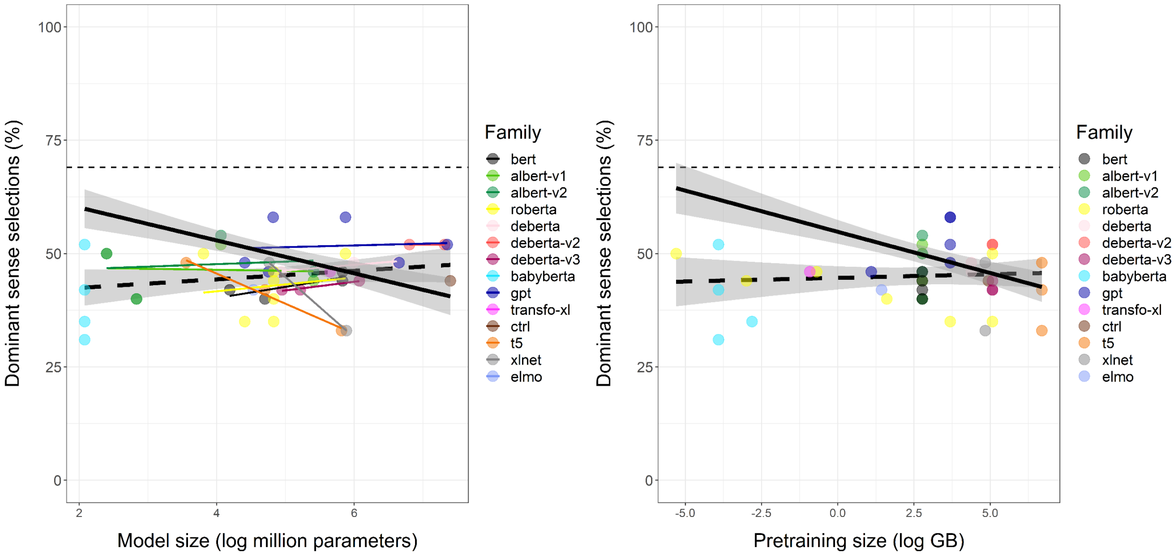

Figure S6.5. ***The percentage of dominant sense selections by adult-based models, in Rabagliati et al. (2013; Study 1), is randomly distributed around 50% and never reaches the level of child dominance bias (indicated by the dashed horizontal line). Furthermore, the selections of dominant senses do not change as a function of either the model size or the pretraining size. The solid lines display the dominant sense selection patterns in the child-based models for a visual comparison.***

Euclidean Distance Measure

In this section, we examined whether child-based models fit children's responses better than adult-based models in each of the three experiments. We used the measure of Euclidean Distance that, as presented in Appendix S5, evaluates the exact match between the model and the child.

To foresee, the only significant difference between adult-based and child-based models was found when examining performance in resolving coherent stories (Rabagliati et al., 2013; Study 1).

In Figure S6.6, we show that child-based models performed more closely to child performance than the adult models did in the first experiment. This can be observed by examining the differences between adult-based dashed regression lines and child-based solid regression lines, with child-based models' regression lines being closer to child performance ($y = 0$). The difference in Euclidean distance from children between adult-based models and child-based models was significant, as shown in table S6.4 and S6.5.

For what concerns the contrastive tasks, instead, tables S6.6 to S6.11 show non-significant differences between adult-based models and child-based models at capturing child performance.

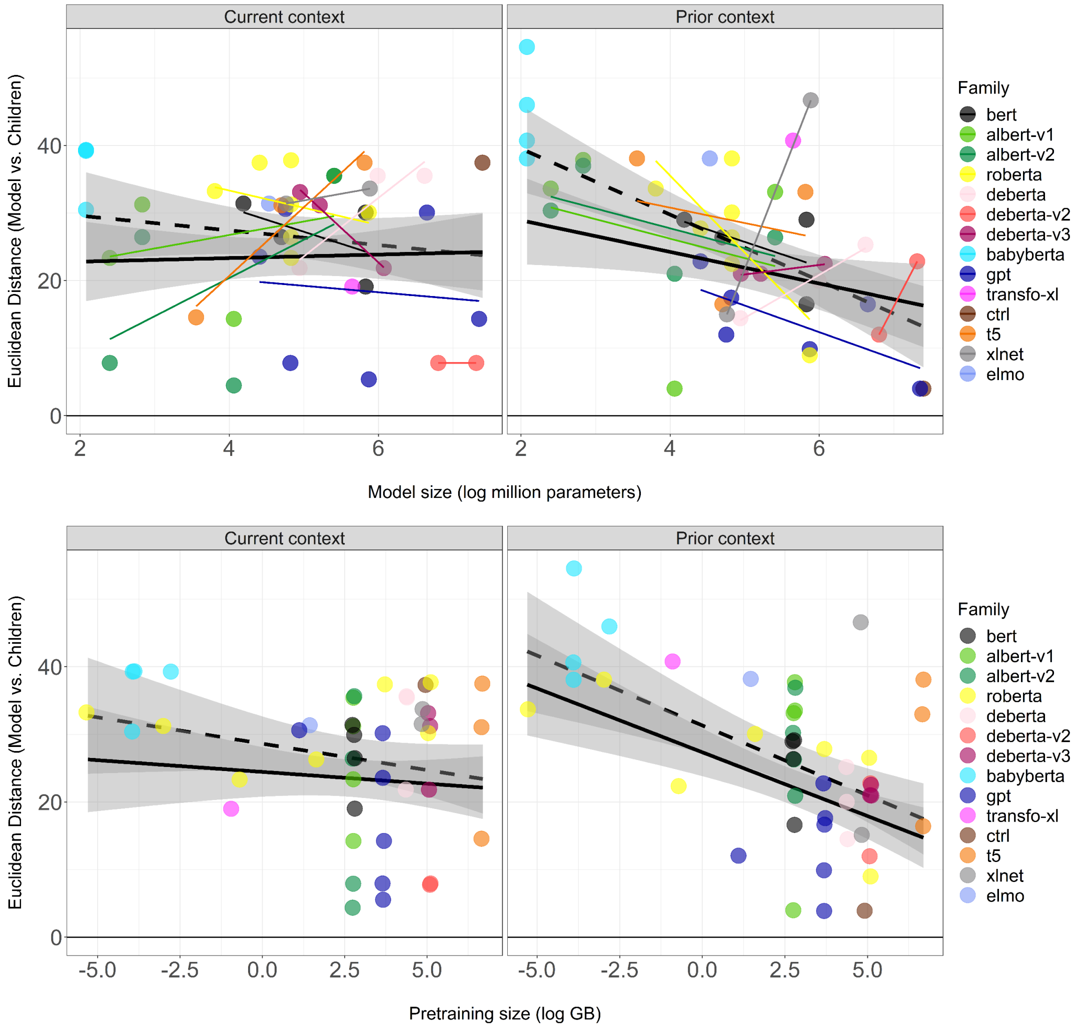

Figure S6.6. ***Models' Euclidean distance from children by model size (top row) and pretraining size (bottom row), in current and prior context conditions of Rabagliati et al. (2013), experiment 1. Model families are shown in the legend. The black horizontal line (y=0) indicates child performance. The dashed regression line with 95% confidence interval shows performance across adult-based models. The solid regression line with 95% confidence interval shows performance across child-based models. Colored regression lines are also shown for each adult-based model family, although only when examining model size as there is almost null variation in pretraining size within family. Colored data points refer to the adult-based dataset. Data points from child-based dataset are omitted for simplicity.***

Table S6.4. ***Model comparison between nested linear mixed-effect models via likelihood ratio test. The outcome is the Euclidean Distance between models and children in Rabagliati et al. (2013) experiment 1. See our main paper for more details about this outcome measure. The predictors are age group (adult-based model/child-based model), condition (current, prior context), log pretraining size, log model size, and their two-way and three-way interactions. The random effect intercept is Model Family. The Null model includes main and random effect intercepts. Subsequent models add one predictor at a time. The table shows the number of model parameters (npar), Akaike (AIC) and Bayesian (BIC) Information criterions, log-likelihood (logLik), deviance, Chi-square statistic (Chisq), degrees of freedom (Df), and p value Pr(>Chisq).***

	npar	**AIC**	**BIC**	**logLik**	**deviance**	**Chisq**	**Df**	**Pr(>Chisq)**
Null model	3	1333.28	1342.86	-663.64	1327.28	NA	NA	NA
+ Age group	4	1329.41	1342.19	-660.71	1321.41	5.87	1	**0.015**
+ Condition	5	1330.94	1346.90	-660.47	1320.94	0.47	1	0.491
+ Pretraining	6	1324.41	1343.56	-656.20	1312.41	8.53	1	**0.003**
+ Model size	7	1325.45	1347.80	-655.73	1311.45	0.95	1	0.329
+ Age group x Condition	8	1327.42	1352.97	-655.71	1311.42	0.03	1	0.862
+ Age group x Model size	9	1324.78	1353.52	-653.39	1306.78	4.64	1	**0.031**
+ Age group x Pretraining	10	1326.43	1358.36	-653.22	1306.43	0.35	1	0.556
+ Condition x Pretraining	11	1317.48	1352.60	-647.74	1295.48	10.95	1	**0.001**
+ Condition x Model size	12	1314.56	1352.88	-645.28	1290.56	4.92	1	**0.027**
+ Pretraining x Model size	13	1315.14	1356.65	-644.57	1289.14	1.42	1	0.233
+ Age group x Condition x Pretraining	14	1317.05	1361.75	-644.52	1289.05	0.09	1	0.759
+ Age group x Condition x Model size	15	1317.95	1365.85	-643.98	1287.95	1.09	1	0.296
+ Age group x Pretraining x Model size	16	1319.94	1371.02	-643.97	1287.94	0.02	1	0.891

Table S6.5. ***Output of the best model selected via model comparison in Table S6.4.***

	'Condition x Model size' model Euclidean Distance Rabagliati et al. (2013) - experiment 1		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	31.32	21.92 – 40.71	**<0.001**
Age group [Child-directed speech]	-13.22	-22.68 – -3.77	**0.006**
Condition [Prior context]	12.38	2.93 – 21.84	**0.011**
Model size [log]	-0.67	-2.67 – 1.32	0.506
Pretraining size [log]	-0.30	-1.28 – 0.69	0.554
Age group [Child-directed speech] × Condition [Prior context]	-0.46	-5.48 – 4.57	0.858
Age group [Child-directed speech] × Pretraining size [log]	-0.31	-1.31 – 0.70	0.548
Age group [Child-directed speech] × Model size [log]	2.29	0.22 – 4.37	**0.030**
Condition [Prior context] × Pretraining size [log]	-0.80	-1.81 – 0.20	0.116
Condition [Prior context] × Model size [log]	-2.29	-4.36 – -0.21	**0.031**
Random Effects			
σ^2	72.92		
$\tau_{00\ \text{family}}$	16.61		
ICC	0.19		
N $_{\text{family}}$	14		
Observations	180		
Marginal R^2 / Conditional R^2	0.195 / 0.345		

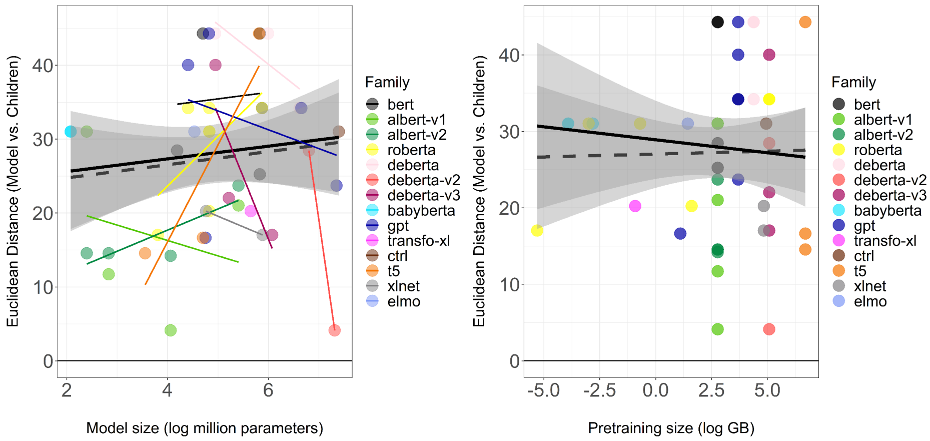

Figure S6.7. ***Models' Euclidean distance from children by model size and pretraining size in Rabagliati et al. (2013) experiment 2. Model families are shown in the legend. The black horizontal line (y=0) indicates child performance. The dashed regression line with 95% confidence interval shows performance across adult-based models. The solid regression line with 95% confidence interval shows performance across child-based models. Colored regression lines are also shown for each adult-based model family, although only when examining model size as there is almost null variation in pretraining size within family. Colored data points refer to the adult-based dataset. Data points from child-based dataset are omitted for simplicity.***

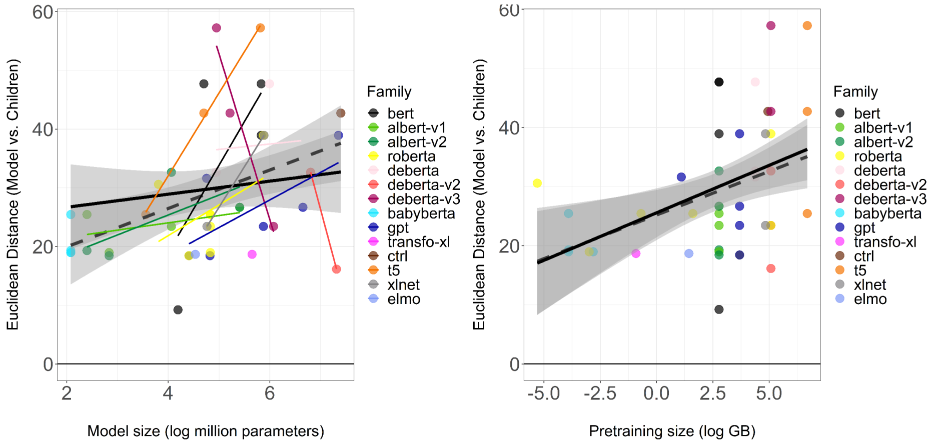

Figure S6.8. ***Models' Euclidean distance from children by model size and pretraining size in Cabiddu et al. (2022), Verb-Event condition. Model families are shown in the legend. The black horizontal line (y=0) indicates child performance. The dashed regression line with 95% confidence interval shows performance across adult-based models. The solid regression line with 95% confidence interval shows performance across child-based models. Colored regression lines are also shown for each adult-based model family, although only when examining model size as there is almost null variation in pretraining size within family. Colored data points refer to the adult-based dataset. Data points from child-based dataset are omitted for simplicity.***

Table S6.6. ***Model comparison between nested linear mixed-effect models via likelihood ratio test. The outcome is the Euclidean Distance between models and children in Rabagliati et al. (2013) experiment 2. See our main paper for more details about this outcome measure. The predictors are age group (adult-based model/child-based model), log pretraining size, log model size, and interactions. The random effect intercept is Model Family. Subsequent models add one predictor at a time. The table shows the number of model parameters (npar), Akaike (AIC) and Bayesian (BIC) Information criterions, log-likelihood (logLik), deviance, Chi-square statistic (Chisq), degrees of freedom (Df), and p value Pr(>Chisq).***

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
Null model	3	693.46	700.96	-343.73	687.46	NA	NA	NA
+ Age group	4	695.33	705.33	-343.66	687.33	0.13	1	0.716
+ Pretraining	5	697.21	709.71	-343.60	687.21	0.12	1	0.728
+ Model size	6	698.79	713.79	-343.40	686.79	0.41	1	0.520
+ Age group x Model size	7	700.79	718.29	-343.40	686.79	0.00	1	0.983
+ Age group x Pretraining	8	702.32	722.31	-343.16	686.32	0.48	1	0.490
+ Pretraining x Model size	9	703.57	726.07	-342.78	685.57	0.75	1	0.387
+ Age group x Pretraining x Model size	10	704.57	729.57	-342.29	684.57	0.99	1	0.319

Table S6.7. ***Although no model surpassed the Null model in Table S6.6, below we show the output of the model including the main effects, to appreciate size of the estimates and variance explained.***

	'+ Model size' model - Euclidean Distance Rabagliati et al. (2013) - Experiment 2		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	24.31	12.79 – 35.83	**<0.001**
Age group [Child-directed speech]	0.77	-3.51 – 5.05	0.721
Model size [log]	0.70	-1.58 – 2.99	0.542
Pretraining size [log]	-0.32	-1.47 – 0.83	0.581
Random Effects			
σ^2	104.19		
$\tau_{00\ \mathrm{family}}$	44.62		
ICC	0.30		
N_{family}	14		
Observations	90		
Marginal R^2 / Conditional R^2	0.007 / 0.305		

Table S6.8. ***Model comparison between nested linear mixed-effect models via likelihood ratio test. The outcome is the Euclidean Distance between models and children in Cabiddu et al. (2022) Verb-Event condition. See our main paper for more details about this outcome measure. The predictors are age group (adult-based model/child-based model), log pretraining size, log model size, and interactions. The random effect intercept is Model Family. The Null model includes main and random effect intercepts. Subsequent models add one predictor at a time. The table shows the number of model parameters (npar), Akaike (AIC) and Bayesian (BIC) Information criterions, log-likelihood (logLik), deviance, Chi-square statistic (Chisq), degrees of freedom (Df), and p value Pr(>Chisq).***

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
Null model	3	678.16	685.66	-336.08	672.16	NA	NA	NA
+ Age group	4	680.02	690.02	-336.01	672.02	0.14	1	0.706
+ Pretraining	5	675.80	688.30	-332.90	665.80	6.22	1	**0.013**
+ Model size	6	676.18	691.18	-332.09	664.18	1.62	1	0.203
+ Age group x Model size	7	675.29	692.79	-330.64	661.29	2.89	1	0.089
+ Age group x Pretraining	8	675.34	695.34	-329.67	659.34	1.94	1	0.163
+ Pretraining x Model size	9	673.64	696.14	-327.82	655.64	3.71	1	0.054
+ Age group x Pretraining x Model size	10	675.63	700.63	-327.81	655.63	0.01	1	0.918

Table S6.9. ***Output of the best model selected via model comparison in Table S6.8.***

	'+ Pretraining' model -Euclidean Distance Verb-Event Condition -Cabiddu et al. (2022)		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	25.80	20.83 – 30.77	**<0.001**
Age group [Child-directed speech]	0.71	-3.08 – 4.51	0.710
Pretraining size [log]	1.25	0.32 – 2.18	**0.009**
Random Effects			
σ^2	82.01		
$\tau_{00\ family}$	33.36		
ICC	0.29		
N_{family}	14		
Observations	90		
Marginal R^2 / Conditional R^2	0.110 / 0.367		

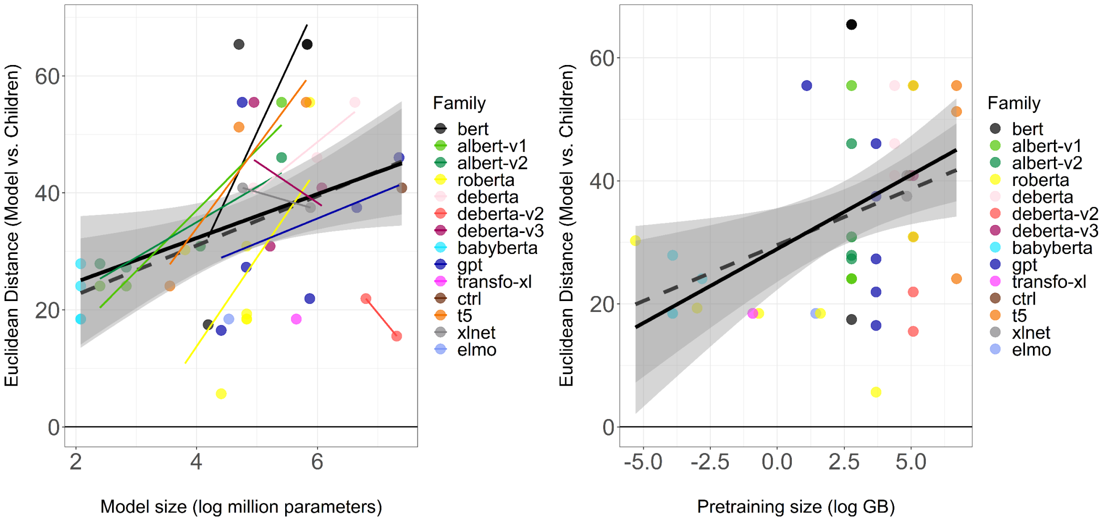

Figure S6.9. ***Models' Euclidean distance from children by model size and pretraining size in Cabiddu et al. (2022), Verb-Lexical condition. Model families are shown in the legend. The black horizontal line (y=0) indicates child performance. The dashed regression line with 95% confidence interval shows performance across adult-based models. The solid regression line with 95% confidence interval shows performance across child-based models. Colored regression lines are also shown for each adult-based model family, although only when examining model size as there is almost null variation in pretraining size within family. Colored data points refer to the adult-based dataset. Data points from child-based dataset are omitted for simplicity.***

Table S6.10. ***Model comparison between nested linear mixed-effect models via likelihood ratio test. The outcome is the Euclidean Distance between models and children in Cabiddu et al. (2022) Verb-Lexical condition. See our main paper for more details about this outcome measure. The predictors are age group (adult-based model/child-based model), log pretraining size, log model size, and interactions. The random effect intercept is Model Family. The Null model includes main and random effect intercepts. Subsequent models add one predictor at a time. The table shows the number of model parameters (npar), Akaike (AIC) and Bayesian (BIC) Information criterions, log-likelihood (logLik), deviance, Chi-square statistic (Chisq), degrees of freedom (Df), and p value Pr(>Chisq).***

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
Null model	3	755.38	762.88	-374.69	749.38	NA	NA	NA
+ Age group	4	757.29	767.29	-374.65	749.29	0.09	1	0.760
+ Pretraining	5	752.51	765.01	-371.26	742.51	6.78	1	**0.009**
+ Model size	6	746.59	761.59	-367.29	734.59	7.92	1	**0.005**
+ Age group x Model size	7	748.52	766.02	-367.26	734.52	0.06	1	0.800
+ Age group x Pretraining	8	749.62	769.62	-366.81	733.62	0.90	1	0.343
+ Pretraining x Model size	9	748.72	771.22	-365.36	730.72	2.90	1	0.088
+ Age group x Pretraining x Model size	10	750.33	775.33	-365.17	730.33	0.39	1	0.534

Table S6.11. ***Output of the best model selected via model comparison in Table S6.10.***

	'+ Model size' model -Euclidean Distance Verb-Lexical Condition Cabiddu et al. (2022)		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	7.66	-8.10 – 23.42	0.336
Age group [Child-directed speech]	0.91	-4.54 – 6.36	0.742
Model size [log]	4.80	1.76 – 7.83	**0.002**
Pretraining size [log]	0.86	-0.68 – 2.39	0.270
Random Effects			
σ^2	168.96		
$\tau_{00\ \text{family}}$	110.41		
ICC	0.40		
N $_{\text{family}}$	14		
Observations	90		
Marginal R^2 / Conditional R^2	0.214 / 0.524		

References (for Appendix S6)

BNC Consortium. (2007). *British National Corpus, XML edition.* https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/2554

Cabiddu, F., Bott, L., Jones, G., & Gambi, C. (2022). The Role of Verb-Event Structure in Children's Lexical Ambiguity Resolution. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*(44). https://escholarship.org/uc/item/9kh29212

Loureiro, D., Rezaee, K., Pilehvar, M. T., & Camacho-Collados, J. (2021). *Analysis and Evaluation of Language Models for Word Sense Disambiguation* (arXiv:2008.11608). arXiv. https://doi.org/10.48550/arXiv.2008.11608

Rabagliati, H., Pylkkänen, L., & Marcus, G. F. (2013). Top-down influence in young children's linguistic ambiguity resolution. *Developmental Psychology, 49,* 1076–1089. https://doi.org/10.1037/a0026918

License

Learning and communication pressures in neural networks: Lessons from emergent communication

Lukas Galke
Centre for Machine Learning, Department of Mathematics and Computer Science (IMADA),
University of Southern Denmark (SDU), Odense, Denmark
LEADS group, Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands

Limor Raviv
LEADS group, Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands
Centre for Social, Cognitive and Affective Neuroscience, University of Glasgow, Glasgow, UK

Abstract: Finding and facilitating commonalities between the linguistic behaviors of large language models and humans could lead to major breakthroughs in our understanding of the acquisition, processing, and evolution of language. However, most findings on human–LLM similarity can be attributed to training on human data. The field of emergent machine-to-machine communication provides an ideal testbed for discovering which pressures are neural agents naturally exposed to when learning to communicate in isolation, without any human language to start with. Here, we review three cases where mismatches between the emergent linguistic behavior of neural agents and humans were resolved thanks to introducing theoretically-motivated inductive biases. By contrasting humans, large language models, and emergent communication agents, we then identify key pressures at play for language learning and emergence: communicative success, production effort, learnability, and other psycho-/sociolinguistic factors. We discuss their implications and relevance to the field of language evolution and acquisition. By mapping out the necessary inductive biases that make agents' emergent languages more human-like, we not only shed light on the underlying principles of human cognition and communication, but also inform and improve the very use of these models as valuable scientific tools for studying language learning, processing, use, and representation more broadly.

Keywords: language acquisition; language evolution; emergent communication; large language models; learning biases; learning pressures; neural networks; neural language models; multi-agent systems

Corresponding author: Lukas Galke, Centre for Machine Learning, Department of Mathematics and Computer Science (IMADA), University of Southern Denmark (SDU), Campusvej 55, DK-5230 Odense M, Denmark. Email: galke@imada.sdu.dk

ORCID ID: https://orcid.org/0000-0001-6124-1092

Citation: Galke, L. & Raviv, L. (2024). Learning and communication pressures in neural networks: Lessons from emergent communication. *Language Development Research 5*(1), 116–143.
http://doi.org/10.34842/3vr5-5r49

Introduction

Using neural language models for language development research dates back to Elman (1993) simulating language acquisition with recurrent neural networks and conceiving the theory of "the importance of starting small". Similarly, Harris (1954)'s distributional structure has motivated word embeddings – a seminal work showing that the semantic relationship between words can be learned without supervision from text data alone (Goth, 2016; Mikolov et al., 2013). These are just some examples of where machine learning has already influenced the development and testing of linguistic theories, showcasing a thriving relationship between the two disciplines (Baroni, 2021; Contreras Kallens et al., 2023; De Seyssel et al., 2023; Dupoux, 2018). The unprecedented success of language models in recent years (Bahdanau et al., 2015; Brown et al., 2020; Devlin et al., 2019; Raffel et al., 2020; Vaswani et al., 2017) provides many opportunities to further advance our understanding of human language learning.

A growing body of work has found similarities between large language models and humans (Dasgupta et al., 2022; Schrimpf et al., 2021; Srikant et al., 2022; Webb et al., 2023; Wei et al., 2022), showing that approximate representations of the outside world can be learned from statistical patterns found in linguistic input alone (Abdou et al., 2021; B. Z. Li et al., 2021; K. Li et al., 2023; Patel & Pavlick, 2022), and manifesting the usefulness of large language models for other disciplines such as psychology (Demszky et al., 2023). However, a so far open issue is the fact that language models are exposed to different input modalities (i.e., mainly text) and have much more data available for training than humans (De Seyssel et al., 2023; Warstadt & Bowman, 2022). Resolving the discrepancy by which language models require much more data than a human child is of high interest to both cognitive science (with the goal of more representative models) and natural language processing researchers (with the goal of more efficient models). Notably, there are ongoing efforts to train language models from similar input as available to a human child, e. g., as in BabyBERTa (Huebner et al., 2021), and the BabyLM challenge[1] (Warstadt et al., 2023).

To promote a deeper understanding of how large language models may be useful for language development research, we suggest to take inspiration from the field of emergent machine-to-machine communication – where two or more neural network agents without exposure to an existing language need to engage in a communication game with the goal of successfully understanding each other (Foerster et al., 2016; Kottur et al., 2017; Lazaridou & Baroni, 2020; Lazaridou et al., 2017). Specifically, emergent communication simulations explore what happens when artificial neural networks (on which also large language models are based) need to create their own languages from scratch, i.e., without first being pre-trained on natural language corpora: do they create human-like languages by-default, or are there specific biases and constraints that

[1] https://babylm.github.io

need to be introduced in order to replicate human behavior? By attempting to simulate phenomena previously observed in humans, research on emergent communication has provided valuable insights into the processes and pressures that shape the evolution of human language, and has allowed researchers to effectively scrutinize, identify, and tease apart the relevant learning biases and conditions that underlie the communicative behaviors of artificial neural networks when they are made to communicate by themselves.

Although the setting of emergent communication is typically motivated for studying the evolution of language (see Lazaridou & Baroni, 2020; Lian et al., 2023, inter alia), language learning and language evolution are intrinsically linked: As languages are passed from generation to generation in a repeated cycle of transmission, imitation, and use, their structure is continuously shaped by the pressures and biases introduced by learners during the process of language acquisition – with such learning biases effectively shaping the evolution of languages on a longer timescale (Chater & Christiansen, 2010; Kirby et al., 2014; Smith, 2022). As such, constraints and pressures associated with learning can causally affect (and, in fact, create) the universal properties of languages, including their most fundamental structural features (Kirby, 2002, 2017; Kirby et al., 2004). As such, we believe that the field of emergent communication provides an ideal testbed for exploring the learning pressures neural networks are exposed to in the process of language learning and use, and can help shed light on (some of) the criticial inductive biases needed for replicating human linguistic behavior.

Since the theoretical usefulness of a model is dependent on its resemblance to the target entity (Zeigler et al., 2000), identifying the relevant learning pressures and biases that govern language creation in neural network models can in turn make neural language models more behaviorally plausible, and consequentially a more robust scientific tool for the language sciences. Here, we review the emergent communication literature and identify underlying learning pressures, while contrasting those with the learning pressures at play when training large language models. Thereby we shed new light on the learning dynamics of neural language models and contribute to the development of more behaviorally plausible language models for language acquisition research.

In the following, we offer a comparative perspective on humans, large language models, and deep learning agents engaging in communication games by reviewing similarities and differences in observed phenomena, discussing how mismatches in the behavior of humans and neural agents can be resolved through appropriate inductive biases, and determining the underlying learning pressures at play. We first provide a brief overview of the emergent communication literature, and then showcase initial mismatches between neural agents and humans with respect to multiple linguistic phenomena: Zipf's law of abbreviation, the benefits of compositional structure, and social factors shaping linguistic diversity (e.g., population size effects). For each of these phenomena, we describe how the initial mismatch between humans and neural network models has been resolved, and identify the underlying learning pressures giving rise to these patterns. In particular, we identify four cognitive and communicative pressures underlying both language acquisition and language evolution, and discuss whether they are inherent to the training objective (i.e., present by default given the learning environment and objective) or whether they need to be artificially incorporated into the models as inductive biases to elicit the desired outcome. We then contrast the identified pressures and biases with those present in the training of large language models, with the goal of promoting knowledge transfer between machine learning and language sciences. We conclude with concrete suggestions for future directions, aimed at developing more cognitively plausible language models for both language development and language evolution research.

Emergent communication, initial mismatches, and their resolution

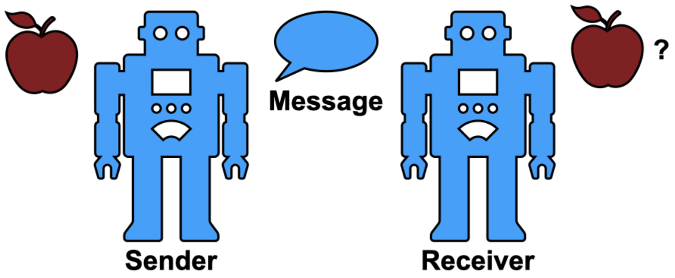

Figure 1. ***Schematics of a simple communication game. The sender sees an object and has to compose a message to describe it. The receiver only sees the message and has to discriminate the object against distractors, or fully reconstruct it.***

Computational modeling has long been used to study language evolution by simulating the process of communication and transmission between artificial agents, typically Bayesian learners (Dale & Lupyan, 2012; Gong et al., 2008; Kirby, 2002; Kirby et al., 2004; Kirby et al., 2015; Perfors & Navarro, 2014; Smith et al., 2003; Smith & Kirby, 2008; Steels, 2016). The emergence of new communication systems is similarly studied using deep neural network models (Lazaridou & Baroni, 2020), and in experimental work with human participants (Kirby et al., 2008; Raviv et al., 2019b; Selten & Warglien, 2007; Winters et al., 2015). Regardless of whether the subjects of these experiments are humans, Bayesian agents, or deep neural networks, they all share the same methodological framework, namely, sender-receiver communication games: One agent describes an input (e. g., an object or a scene), and transmits a message to another agent, that then has to guess or fully reconstruct the sender's input (see Figure 1). The agents in emergent communication experiments are typically based on deep neural networks, similar to those used in large language models.

Table 1: Observed phenomena from humans in agents from emergent communication simulations

Phenomenon in Humans	**Mismatch in Emergent Communication agents**	**Resolution**
Zipfian distribution in utterance length (frequent meanings are described by shorter utterances)	Sender agents exploit the full channel capacity because longer messages are easier to distinguish by receiver agents.	Introducing a penalty on long utterances (simulating "laziness") restores the Zipfian distribution on utterance length.
Compositional structure reliably emerges during communication and cultural transmission, and is beneficial for language learning and generalization	Inconsistent emergence of compositional structure in neural agents, and seemingly no advantage of more compositional protocols for generalization	Periodically resetting agents' parameters (simulating generational turnover) gives rise to compositional protocols, which are easier to learn for neural network agents
Population size affects the emergence of compositional structure (larger communities create more systematic languages)	Larger populations of neural agents do not create more compositional protocols	Introducing population heterogeneity (simulating individual differences) or production-comprehension symmetry (simulating role alternation in language use) leads to larger populations creating more systematic protocols

In a typical communication game, the sender acts as a conditioned-generation model, taking a target input (for example, an image or a set of attribute values) and produces a message consisting of multiple symbols. The symbols of the message are generated one by one without any pre-defined vocabulary. The generated message is then transmitted to the receiver. The receiver is trained to infer the sender's input based on the message, by selecting the correct object among distractors or by fully reconstructing it.

Emergent communication models start with randomly-initialized parameters, without any pre-defined list of words or look-up table. Thus, the messages start out as random, and only over the course of training and interaction do the models develop a communication protocol. In fact, it is the central assumptions of *emergent* communication that the agents are not seeded with some initial language or communication protocol, but that they develop the communication system on their own during interaction. Thus, agents start from scratch and are guided primarily by communicative success. Yet, there is room for inductive biases, i. e., additional biases that are imposed on the learning system to promote desired behaviours (Mitchell, 1980). While cognitive biases in biological learning systems occur naturally, inductive biases in machine learning are artificially introduced to guide the learning dynamics. For a profound overview of the emergent communication literature, we refer to recent review and survey papers by Lazaridou and Baroni (2020), Galke et al. (2022), and Brandizzi (2023).

Notably, methods from the field of emergent communication and from the closely related field of reinforcement learning (see Kosoy et al., 2020; Kosoy et al., 2022, inter alia) have already been used for language development research (see Ohmer et al., 2020; Portelance et al., 2021, inter alia), for example, to study the emergence of a mutual exclusivity bias with pragmatic agents.

While emergent communication simulations hold a great potential for advancing our understanding of how languages emerge, we can only expect insights gained with deep neural networks to inform language evolution research if the resulting languages actually show the same properties as natural languages (Galke et al., 2022). Consequently, most emergent communication simulations try to compare the properties of their emerging communication protocols to the properties found in natural languages (Havrylov & Titov, 2017; Kottur et al., 2017; Lazaridou et al., 2017). By following this approach, the field has unveiled substantial differences between humans and machines in how they learn to communicate and what kinds of languages they develop.

Crucially, although the emergent languages of neural networks initially did not exhibit many of the linguistic properties typically associated with human languages, most of these differences could be reconciled by adding adequate inductive biases, such as laziness and impatience – which, when introduced, recovered the effects found in humans. Notably, some linguistic phenomena such as the word-order/case-marking trade-off seem to occur in communicating neural networks without specific inductive biases (Lian et al., 2023). Below we review selected properties of human languages in which initial mismatches between humans and neural network agents were resolved and discuss the inductive biases that were necessary for their recovery. Table 1 provides an overview of the three phenomena and their occurrence in neural simulations.

Zipfian distribution in utterance length

Perhaps the most illustrative example of mismatches between the languages developed by humans and machines was the initial absence of Zipf's law of abbreviation in machine learning simulations. According to Zipf's law of abbreviation, the relationship between word frequency and word length follows a power law distribution, such that more frequent words are typically shorter while less frequent words are typically longer (Newman, 2005; Zipf, 1949). Zipf (1949) suggested that this effect is caused by the principle of least effort, i.e., since frequent words are produced often, and shorter words are easier to produce. Critically, Zipf's law has important implications for language evolution (Kanwal et al., 2017) and language acquisition (Ellis & Collins, 2009), with active restructuring of lexicon towards more efficient communication (Gibson et al., 2019).

Initial findings in emergent communication showed that Zipf's Law of Abbreviation is absent from the languages developed by neural agents, which was dubbed as 'anti-efficient coding' (Chaabouni et al., 2019). This was because neural senders were not under any pressure to communicate efficiently or to reduce effort. In fact, longer messages were easier for the receiver agent to process because they allowed for more opportunities to differentiate between meanings: for a 1-symbol utterance, the sender can select only 1 item from the alphabet of size k, but for a n-symbol utterance, the sender can produce k^n different combinations. The more distinct utterances are from another, the easier it is for the receiver to distinguish the target meaning from other possible meanings. Thus, longer utterances are advantageous for conveying the meaning correctly – especially when there is no penalty for utterance length.

The mismatch with human language was resolved by adjusting the optimization objective in a direction that made sender agents "lazy" (i.e., longer messages were penalized) and receiver agents "impatient" (i.e., receivers tried to infer the meaning as early as possible in a sequential read) (Rita et al., 2020). This inductive bias, which aims at mimicking real human behavior during language production and comprehension, has recovered Zipf's Law of Abbreviation in emergent communication simulations – showing that when such biases for efficiency are introduced, communication protocols developed by neural agents do show a similar frequency–length relationship as found in natural languages.

The emergence of compositional structure and its benefits for learning and generalization

Compositional structure is considered a hallmark feature of human language (Hockett, 1960; Szabó, 2022): there is a systematic mapping between linguistic forms (e.g., words, morphemes) and their meanings (e.g., concepts, grammatical categories), such that the meaning of a complex expression can be typically derived from the meanings its constituent parts. For example, the meaning of the phrase "small cats" is directly derived from the meanings of the words "small", "cat", and the marker "-s" (denoting plurality). The presence of such compositional structure underlies the infinite expressive and productive power of human languages, allowing us to describe new meanings in a way that is transparent and understandable to other speakers (Kirby, 2002; Zuidema, 2002).

In experiments simulating the evolution of languages in the lab using sender-receiver communication games, the need to communicate over a growing number of different items or in an open-ended meaning space leads to the emergence of compositional languages (Nölle et al., 2018; Raviv et al., 2019a). Crucially, the degree of compositional structure in linguistic input then predicts adults' learning and generalization accuracy, such that, compared to languages with little to no compositionality, languages with more compositional structure are learned better and faster and result in better (i.e., more transparent and systematic) generalizations to new meanings, which are also shared across different individuals who never interacted before (Raviv et al., 2021). Thus, the evolution of more compositional and systematic linguistic structure allows for more productive generalization and facilitates communication and convergence between strangers.

The learning advantage of more compositional structure for adult participants is also echoed in numerous iterated learning studies, which have shown that artificial languages become more compositional and consequently easier to learn over the course of cross-generational transmission (Beckner et al., 2017; Carr et al., 2017; Kirby et al., 2008; Kirby et al., 2014).

Testing the limits of our imagination, neural networks seemed to generalize well even without compositional communication protocols (Chaabouni et al., 2020; Lazaridou et al., 2018). Specifically, Chaabouni et al. (2020) found that, after many repetitions of an emergent communication experiment, all compositional languages generalized well, but so did non-compositional languages. This finding spurred numerous follow-up studies that aimed at improving the learning dynamics through inductive biases or by making the communication game more difficult (more complex stimuli, larger alphabet, longer messages, more agents) to successfully promote the emergence of compositional structure (Chaabouni et al., 2022; Rita, Tallec, et al., 2022). However, the lack of correlation between the degree of compositional structure – as measured by topographic similarity (Brighton & Kirby, 2006) – and generalization performance had remained.

The most reliable way to promote the emergence of compositional languages is periodically resetting the parameters of the neural network agents (Chaabouni et al., 2022; F. Li & Bowling, 2019; Zhou et al., 2022), similar to Kirby et al. (2014)'s iterated learning paradigm – leading to the hypothesis that compositional languages have a learnability advantage (Chaabouni et al., 2020; Chaabouni et al., 2022; Guo et al., 2019; F. Li & Bowling, 2019). However, these attempts did not directly test language learnability in a purely supervised fashion.

Recently, Conklin and Smith (2022) have re-analyzed the setting of Chaabouni et al. (2020) and found that, in fact, the lack of correlation between compositionality and generalization performance in the original simulation was caused by a fallacy of the topographic similarity metric that had been used to measure compositionality. For instance, homonyms (different forms for same meaning) obscure compositionality under the topographic similarity measure. When taking this variation into account, compositional structure does reliably emerge and is beneficial for generalization. In other words, it is probably the case that there was not really a mismatch between humans and neural agents in the first place.

Supporting this view, Galke et al. (2023) have replicated a large-scale language learning study originally conducted with human participants (Raviv et al., 2021) with deep neural networks and have confirmed the advantage of compositional structure for learning and generalization in neural networks. The results showed similar pattern across three learning systems – humans, small-scale recurrent neural networks trained from scratch, and the large pre-trained language model GPT-3 – with compositional structure being advantageous for all types of learners. Specifically, the results showed that neural networks benefit from more structured linguistic input, and that their productions become increasingly more similar to human productions when trained on more structured languages. This structure bias can be found in the networks' learning trajectories and their generalization behavior, mimicking previous findings with humans: although all languages can eventually be learned, languages with a higher degree of compositional structure were led to better and more human-like generalization to new, unseen items.

Population size effects

Socio-demographic factors such as population size have long been assumed to be important determinants of language evolution and variation (Lupyan & Dale, 2010; Nettle, 2012; Wray & Grace, 2007). Supporting this idea, global cross-linguistic studies report that bigger communities tend to have languages with more regular and transparent structures (Lupyan & Dale, 2010). Similarly, in experimental work, larger groups of interacting participants generally develop languages with more systematic (i.e., compositional) grammars (Raviv et al., 2019b). These findings are typically attributed to compressibility pressures arising during communication: remembering partner-specific variants becomes increasingly more challenging as group size increases and shared history decreases, which lead larger groups to prefer easier-to-learn-and-generalize variants and thus converge on more transparent and systematic languages.

Tieleman et al. (2019) has investigated populations of autoencoders. Autoencoders are neural network models composed of an encoder module and a decoder module that learn to "good" representations (the code) by reconstructing their own input. Now Tieleman et al. (2019) have decoupled encoder and decoders and exchanged them throughout training – while communicating in a continuous channel. There, larger communities produced representations with less idiosyncrasies and lead to better convergence among different agents. While a promising starting point, the communication was modeled as exchanging continuous vectors and training the encoder decoder modules together, as if they were one model. This is arguably natural communication paradigm for neural networks because it is optimized in the same way as the communication between layers in a single neural network. However, this continuous channel stands in contrast with the discrete nature of human communication (Hockett, 1960). Most other approaches in emergent communication, however, do consider a discrete channel (Galke et al., 2022).

While Chaabouni et al. (2022) argued that it is necessary to scale up emergent communication experiments in different aspects including population size in order to better align neural emergent communication with human language evolution, they have not found a consistent advantage of population size in generalization and ease-of-learning (in contrast with (Tieleman et al., 2019)). Similarly, Rita, Strub, et al. (2022) found that language properties are not enhanced by population size alone.

While emergent communication in populations of agents has been investigated earlier (Fitzgerald, 2019; Graesser et al., 2019; Lowe et al., 2019, e.g.), the effect of population size on structure with groups of more than two agents has only recently been analyzed (Chaabouni et al., 2022; Michel et al., 2023; Rita, Strub, et al., 2022). Out of these, two studies aimed to recover the group size effect in populations of neural network agents by introducing population heterogeneity (Rita, Strub, et al., 2022) and manipulating sender-receiver ties (Michel et al., 2023).The first study by Rita, Strub, et al. (2022) modeled population heterogeneity by giving each agent a different random learning rate While previous simulations used populations of identical agents, Rita et al. modeled population heterogeneity by giving each agent a different random learning rate. Results showed that in this scenario, group size effects could be partially recovered. Notably, the authors found that it is important to give sender agents having (much) higher learning rates than receivers.

Secondly, while most emergent communication simulations keep senders and receivers distinct (i.e., agents that produce never comprehend, and vice versa), there is also work that emphasizes linking production and comprehension components within the agents (e.g., by sharing some of the model parameters) (Graesser et al., 2019; Portelance et al., 2021). Galke et al. (2022) argue that this naturalistic property of alternating between sending and receiving (i.e., engaging in both production and comprehension in typical language use) may be a crucial ingredient to ensure more linguistically plausible learning dynamics – and could lead to recovering the group size effect. Subsequently, Michel et al. (2023) have introduced sender-receiver ties via gradient blocking, such that a sender and a receiver together form a single agent and each receiver is only optimized for its corresponding sender. This change indeed led to a recovery of the group size effect, with larger population of agents creating more compositional protocols. Another promising approach is to have agents model other agents' knowledge, allowing them to communicate differently with different agents - something that has been implied to underlie group size effects in humans (Lutzenberger et al., 2021; Meir et al., 2012; Mudd et al., 2020; Thompson et al., 2020). While such "theory of mind" is generally absent from emergent communication simulations in populations, the ability to infer other agents' beliefs has been successfully implemented in various reinforcement learning setups, e.g., (Filos et al., 2021; Ohmer et al., 2020).

Underlying learning pressures and inductive biases

In general, there are two types of learning biases and pressures. First, some biases and pressures seem to be present naturally, or universally, across all different learning systems investigated here, including deep learning agents. An example for this is the structure-bias, i.e., the learnability and generalization advantage of more compositional communication protocols (Galke et al., 2023) (see above). This structure-advantage seems to be present for both humans and neural networks, even without specific inductive biases. In contrast, some biases need to be artificially introduced in order to recover the effects found in humans. These include, for example, adding a length-penalty for senders, which effectively makes agents "lazy". In the above examples, we demonstrated the flexibility and adaptive nature of neural simulations and how they can be tweaked to replicate human behavioral patterns. While many features associated with natural languages were initially absent from such simulations, these mismatches have been fully or partially resolved by introducing theory-driven and human-inspired cognitive biases and learning pressures to the learning system – and these inductive biases have consequentially led to better alignment between neural agents and humans. Below, we outline on a more fine-grained level what pressures are relevant for language learning and evolution in neural networks, contrasting them with the pressures to which current large language models are exposed, and to what extent incorporating the pressures may promote the relevance of large language models for developmental research. Table 2 provides an overview of the comparison of learning pressures in emergent communication agents and large language models. Notably,

Table 2: Pressures derived from emergent communication simulations and their operationalization in neural agents and large language models

Derived Pressure	**Emergent Communication Agents**	**Large Language Models**
Pressure for successful communication	The main training objective in communication games	Absent in pre-training and fine-tuning. Only introduced when learning from human preferences in RLHF.
Pressure for learnability	Can be artificially introduced through parameter reset and iterated learning	Neural networks underlying large language models have a tendency to find the simplest solution first
Pressure to reduce production effort	Can be artificially introducing, e.g., through a penalty term for long messages	Production length is learned from LLM's training data and human feedback in RLHF.
Memory constraints	Absent because the high capacity of neural agents is sufficient to memorize even unstructured mappings	Huge capacity due to extremely high amount of parameters, yet "working memory" for in-context learning is limited by context window (how many tokens the models can process at a time)
Production-comprehension symmetry	Can be artificially introduced by linking sender and receiver modules	By design – LLMs employ the same neural network modules and parameters for comprehension and production
Modeling other agents' internal states	Can be modeled explicitly, e.g., for pragmatic reasoning	In the RLHF training stage, a reward model is trained and consulted to estimate human preferences.

this is not an exhaustive list – it focuses on the specific pressures that underlie the phenomena described above, but do not consider many other important aspects that govern natural language learning, such as grounding, a noisy environment, multi-modal communication, or referential and iconic signs.

Pressure for successful communication

In order to achieve successful communication, language users need distinguish between a variety of meanings. This expressivity pressure is hypothesized to underlie human language evolution, and serves as a "counter pressure" for simplicity/compressibility (i.e., the idea that languages should be as simple and as learnable as possible) (Kirby et al., 2015). The pressure for communicative success, e. g., to accurately reconstruct the meaning of referents from a message during interaction, is the most straight-forward pressure found in collaborative communication games (and, arguably, in real-world interaction). In emergent communication with deep neural networks, this pressure is encoded right in the optimization objective of the neural networks.

In contrast, for large language models such as GPT-3.5, the main objective during pre-training is not communication success. The standard language modeling objective used during pre-training of large language models instead optimizes for utterance completion (i. e., learning to predict words from their context). While this language modeling objective leads to tremendous success regarding language competence other emergent abilities (Devlin et al., 2019; Wei et al., 2022), it is clearly a different training objective than optimizing for communicative success, as in emergent communication simulations. After large-scale pre-training, large language models are fine-tuned using small datasets of human-generated pairs of instructions and their corresponding responses, usually with the same training objective as in pre-training. In other words, the models are made to learn from interactions by completing utterances from human-generated interactions, but not by interacting themselves. Only during the last stage of training, the models are trained via Reinforcement Learning from Human Feedback (RLHF), where a reward model estimates human preferences based on human ratings of different machine-generated responses (Ouyang et al., 2022; Schulman et al., 2017). Only in this final RLHF training stage of LLMs, the models are optimized for successful communication. Yet, this stage is important to turn base models into chat assistants that engage in conversations with humans (OpenAI, 2023; Ouyang et al., 2022).

In general, while emergent communication simulations are tuned for communicative success by design, this is in fact an extra step in large language models after pre-training on utterance completion. Thus, the learning paradigms of fine-tuning and subsequent learning from human feedback are worth further exploration for the goal of having language models being more representative of human behavior. For instance, a recent study has showcased that fine-tuning large language models on data from psychological tests turns them into useful cognitive models (Binz & Schulz, 2023).

Pressure to reduce production effort

Humans constantly strive to reduce effort during interaction (Gibson et al., 2019). For instance, this is demonstrated by our tendency to shorten or erode highly frequent words (Kanwal et al., 2017; Zipf, 1949). However, the pressure to communicate with least effort is absent in neural networks, and is usually not reflected in their training objective. In other words, it simply does not cost more "effort" for a neural network to generate a longer message. By introducing a bias for more efficient communication, Rita et al. (2020) have shown that typical human behavior can be recovered. Since language models similarly don't have an 'innate' pressure to reduce effort, it may be worth considering integrating such a pressure for efficient communication into these models for the sake of mimicking human behavior with respect to language development. However, one needs to strike a balance, as imposing a least-effort bias could also lead to communication failure in emergent communication scenarios (Lian et al., 2021), calling for further investigation of how a least-effort bias is best incorporated.

In large language models, there is no pressure to reduce production effort: LLMs are trained on next-token production over large corpora of text data, which is being piped through the model in a batched fashion to maximize throughput (see for instance Brown et al., 2020; Touvron et al., 2023, inter alia). Thus, the main driver for production length is simply the utterance length in data, and the placement of specific separator tokens, e.g., at the end of each unit of consecutive text during training. Moreover, the RLHF stage of training large language models (Ouyang et al., 2022; Schulman et al., 2017), which is supposed to align LLMs with human preferences, even promotes the generation of longer utterances, as they are deemed to be more "helpful" by (instructed) human annotators (Singhal et al., n.d.).

At inference time, when the LLM is prompted to generate text, a hard cut-off on the number of tokens or a soft length penalty may be introduced – the details of these techniques, however, are often not publicly available. Regardless, the training procedure itself does usually not include a length penalty, which needs to be taken into account when planning to use large language models for language development research.

Pressure for learnability

Based on our review, a pressure for learnability (or continual re-learning) also governs the development of communication protocols between neural network agents. That is, agents should prioritize communication protocols (or single variants) that are easier to learn, and such protocols should in turn boost performance. This learnability pressure is strongly connected to the fact that languages must be transmitted, learned, and used by multiple individuals, often from limited input and with limited exposure time (Smith et al., 2003). Yet, there is a subtle difference to strict transmission chains of iterated learning, as it is sufficient with neural networks to reset only some of the agents (F. Li & Bowling, 2019), or only parts of a single agent (Zhou et al., 2022). In numerous different settings, it has been shown that learnability pressures are crucial for compositional structure to emerge (Chaabouni et al., 2022; F. Li & Bowling, 2019; Zhou et al., 2022).

This also suggests that under repeated learning, either in Iterated Learning with human participants or with parameter reset in neural networks, weak learning biases can get amplified in the process of cross-generational transmission (Reali & Griffiths, 2009). But what are these learning biases exactly? How can they be operationalized? And how do they actually translate into language learning in the real-world? For example, do these biases differ between children and adults, or between different levels of linguistic analyses (e.g., vocabulary vs. syntax)? At the moment, these are still open questions. However, they highlight the need to seriously consider the meaning and implications of different modeling choices when simulating language acquisition using language models and deep neural networks.

As for large language models, Chen et al. (2024) have made relevant findings by analyzing the learning dynamics: language models pick up grammar as the simplest explanation for the data very early on during training (structure onset), and only shortly thereafter, general linguistic capabilties arise. In addition, when suppressing grammar as a possible way to explain the data, the models learn other strategies, but do not go back to grammar when the constraint is removed later in training.

This finding connects well with more general findings of simplicity bias in neural networks (Geirhos et al., 2020). In addition, it also connects with the findings of emergent communication in emphasizing that re-learning (e. g., through parameter reset) is important for compositional structure to emerge (F. Li & Bowling, 2019). Our hypothesis is that, if there was no pressure for re-learning, then agents would fall for the earliest successful strategy and do not consider alternatives – stressing the importance of the learnability pressure.

Memory constraints

Human language learning is governed by cognitive constraints such as a limited memory capacity. These, in turn, affect processes of language evolution and promote greater convergence to a common language within a community: once groups become too big, it becomes hard to maintain unique communication protocols with different partners (i.e., idiolects) (Wray & Grace, 2007).

Such constraints have been shown to underlie patterns of cross-linguistic diversity, whereby larger populations develop more structured and less variable languages (Raviv et al., 2019b). Yet, neural networks have virtually no memory constraints because they are commonly heavily over-parametrized. Due to this over-parametrization, neural networks have no problem to keep a large number of different partner-specific variants in their memory, and have little need to converge on a single shared language. However, simply reducing the number of model parameters to the theoretical minimum is not feasible either, as explored in emergent communication by Resnick et al. (2020). This is because over-parametrization is, in fact, a critical ingredient for the success of deep neural networks (Arora et al., 2019; Cybenko, 1989; Nakkiran et al., 2021; Zhong et al., 2017). But given the importance of such memory constraints for human language learning and evolution, it may be worth considering how such pressures can nonetheless be mimicked or introduced as inductive biases when employing deep neural networks as models for language development research.

While large language models have even higher model capacity with billions of learnable parameters, there is an interesing conceptual connection with working memory: As the model parameters are not updated at inference time (when the model is prompted with a specific input), the model can only base its generation on what is available in the prompt, which is limited by the LLMs' context window of how many tokens can be processed at a time. Although also these context windows grow larger and larger with the development of new models (OpenAI, 2023), it allows researchers to explicitly control what information is available to the model at a specific point in time.

Production-comprehension symmetry

In addition, in naturalistic settings with proficient language users, every person capable of producing a language is also capable of understanding it (Hockett, 1960) – a property that was typically absent from emergent communication simulations (Galke et al., 2022). Indeed, introducing an inherent connection between production and comprehension in neural networks has led to an increase in the desirable properties of emergent languages (Michel et al., 2023). Interestingly, comprehension and production are intrinsically linked in autoregressive large language models as the same model parameters are used for processing and for generation (Radford et al., 2019). Such results again underscore the importance of keeping seemingly basic psycholinguistic features in mind when using large language models and neural networks as models for human language learning and use.

Modeling other agents' internal states

Furthermore, another intriguing direction is to explicitly model other agents' internal states. For instance, Ohmer et al. (2020) integrates pragmatic reasoning into the agents, leading to accelerated learning – an effect that is even stronger with Zipfian input distributions compared to uniform input distributions. Explicitly modeling other agents internal states and social learning has been shown to be successful in other reinforcement learning scenarios, where agents can cooperate or compete about resources (Filos et al., 2021; Ndousse et al., 2021). Interestingly, these ideas of explicitly modeling the internal state of the interlocutor are already present in the final training stage of large language models, when optimizing for human preferences via RLHF (Ouyang et al., 2022; Schulman et al., 2017): the common procedure is to learn a specific reward model that estimates human preferences on new data, which is then be employed for steering the generations of the language model in a particular direction – here the reward model is specifically designed to estimate to what extent humans would prefer one generation over the other, which is closely resembles the idea of modeling other agents' (or humans') internal states.

Discussion

Several important mismatches between humans and neural agents with respect to language emergence can be explained by the absence of key cognitive and communicative pressures, such as memory constraints and production-comprehension symmetry, which drive language evolution. Here, we demonstrated how including these factors in neural agents can resolve said mismatches, and lead to more accurate simulations that mimic the settings and pressures operating during human language learning and use – and consequentially resulting in emergent neural communication protocols that are more linguistically plausible. Notably, additional psycho- and sociolinguistic factors may affect language evolution and learning, and might also play a role in explaining further discrepancies in behavioral patterns across learning systems.

In the current paper we presented a number of initial mismatches between humans and agents engaging in communication games – and demonstrated how they could be resolved through inductive biases. So far, there is no unified approach that consolidates all of the resolutions mentioned above. We deem this a promising direction of future work – e. g., merging the techniques of population heterogeneity, laziness and impatience, and sender-receiver ties, which have so far only been evaluated independently.

As exemplified by recent work, it is promising to keep up and nourish the knowledge exchange between researchers working on human languages and those working on computational simulations of language, e. g., via theory diffusion from language studies into machine learning and vice versa. A famous example is cultural evolution (Tomasello, 2008) and the iterated learning paradigm (Kirby et al., 2008; Kirby et al., 2014), which sparked the idea of iteratively training neural networks while resetting some of the networks' parameters (Frankle & Carbin, 2018; F. Li & Bowling, 2019; Nikishin et al., 2022; Zhou et al., 2022). This idea has, for instance, advanced our understanding of neural networks (their reliance on sparse sub-networks) and led to favorable learning dynamics that cause better and more systematic generalization beyond the training distribution. Similarly, the discrete and compositional structure of natural languages inspired researchers to incorporate discrete representations into neural network architectures in order to advance the models' generalization performance and continual learning capabilities (Liu et al., 2021; Träuble et al., 2023).

In conclusion, The emergent communication literature provided the opportunity to assist in developing linguistic theories in the spirit of Elman (1993), while, conversely, reflecting on how phenomena and biases known from humans may ultimately enhance neural networks, as in lifelong and open-world learning, which is still a major open problem in machine learning. For making use of large language models in language development research, we consider it a promising direction for future work to take inspiration from the emergent communication literature, and see which inductive biases (such as the ones sketched here) have helped to recover patterns from human language learning. Concretely, this would entail ingesting a training objective for communicative success earlier in language model training, and integrating a pressure to keep utterances as short as possible. Integrating these biases into large language models may very well lead to more cognitively plausible models for gaining new insights on how children acquire their first language.

References

Abdou, M., Kulmizev, A., Hershcovich, D., Frank, S., Pavlick, E., & Søgaard, A. (2021). Can language models encode perceptual structure without grounding? a case study in color. *Proceedings of the 25th Conference on Computational Natural Language Learning*, 109–132. https://doi.org/10.18653/v1/2021.conll-1.9

Arora, S., Du, S., Hu, W., Li, Z., & Wang, R. (2019). Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks. *Proceedings of the 36th International Conference on Machine Learning*, 322–332.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. *Proceedings of ICLR*.

Baroni, M. (2021). On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. *arXiv preprint, abs/2106.08694*. https://arxiv.org/abs/2106.08694

Beckner, C., Pierrehumbert, J. B., & Hay, J. (2017). The emergence of linguistic structure in an online iterated learning task. *Journal of Language Evolution*, *2*(2), 160–176.

Binz, M., & Schulz, E. (2023). Turning large language models into cognitive models. *arXiv:2306.03917*.

Brandizzi, N. (2023). Towards More Human-like AI Communication: A Review of Emergent Communication Research. *arXiv:2308.02541*.

Brighton, H., & Kirby, S. (2006). Understanding linguistic evolution by visualizing the emergence of topographic mappings. *Artificial Life*, *12*(2), 229–242. https://doi.org/10.1162/artl.2006.12.2.229

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems 33*.

Carr, J. W., Smith, K., Cornish, H., & Kirby, S. (2017). The cultural evolution of structured languages in an open-ended, continuous world. *Cognitive science*, *41*(4), 892–923.

Chaabouni, R., Kharitonov, E., Bouchacourt, D., Dupoux, E., & Baroni, M. (2020). Compositionality and generalization in emergent languages. *ACL*, 4427–4442.

Chaabouni, R., Kharitonov, E., Dupoux, E., & Baroni, M. (2019). Anti-efficient encoding in emergent communication. *NeurIPS*, 6290–6300.

Chaabouni, R., Strub, F., Altché, F., Tarassov, E., Tallec, C., Davoodi, E., Mathewson, K. W., Tieleman, O., Lazaridou, A., & Piot, B. (2022). Emergent communication at scale. *ICLR*. https://openreview.net/forum?id=AUGBfDIV9rL

Chater, N., & Christiansen, M. H. (2010). Language Acquisition Meets Language Evolution. *Cognitive Science*, *34*(7), 1131–1157. https://doi.org/10.1111/j.1551-6709.2009.01049.x

Chen, A., Shwartz-Ziv, R., Cho, K., Leavitt, M. L., & Saphra, N. (2024). Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs. *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=MO5PiKHELW

Conklin, H., & Smith, K. (2022). Compositionality with Variation Reliably Emerges in Neural Networks. *The Eleventh International Conference on Learning Representations*.

Contreras Kallens, P., Kristensen-McLachlan, R. D., & Christiansen, M. H. (2023). Large Language Models Demonstrate the Potential of Statistical Learning in Language. *Cognitive Science*, *47*(3), e13256. https://doi.org/10.1111/cogs.13256

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Math. Control. Signals Syst.*, *2*(4), 303–314. https://doi.org/10.1007/BF02551274

Dale, R., & Lupyan, G. (2012). Understanding the origins of morphological diversity: The linguistic niche hypothesis. *Advances in complex systems, 15*(03n04), 1150017.

Dasgupta, I., Lampinen, A. K., Chan, S. C. Y., Creswell, A., Kumaran, D., McClelland, J. L., & Hill, F. (2022). Language models show human-like content effects on reasoning. *arXiv:2207.07051.*

De Seyssel, M., Lavechin, M., & Dupoux, E. (2023). Realistic and broad-scope learning simulations: First results and challenges. *Journal of Child Language*, 1–24. https://doi.org/10.1017/S0305000923000272

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., JonesMitchell, N., Ong, D. C., Dweck, C. S., Gross, J. J., & Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*. https://doi.org/10.1038/s44159-023-00241-5

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Dupoux, E. (2018). Cognitive Science in the era of Artificial Intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition, 173*, 43–59. https://doi.org/10.1016/j.cognition.2017.11.008

Ellis, N., & Collins, L. (2009). Input and Second Language Acquisition: The Roles of Frequency, Form, and Function Introduction to the Special Issue. *The Modern Language Journal, 93*(3), 329–335. https://doi.org/10.1111/j.1540-4781.2009.00893.x

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition, 48*(1), 71–99. https://doi.org/10.1016/0010-0277(93)90058-4

Filos, A., Lyle, C., Gal, Y., Levine, S., Jaques, N., & Farquhar, G. (2021). Psiphi-learning: Reinforcement learning with demonstrations using successor features and inverse temporal difference learning. *International Conference on Machine Learning*, 3305–3317.

Fitzgerald, N. (2019). To populate is to regulate. *EmeCom workshop at NeurIPS*.

Foerster, J., Assael, I. A., de Freitas, N., & Whiteson, S. (2016). Learning to Communicate with Deep Multi-Agent Reinforcement Learning. *Advances in Neural Information Processing Systems, 29*.

Frankle, J., & Carbin, M. (2018). The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. *International Conference on Learning Representations*.

Galke, L., Ram, Y., & Raviv, L. (2022). Emergent communication for understanding human language evolution: What's missing? *Emergent Communication Workshop at ICLR 2022*. https://openreview.net/forum?id=rqUGZQ-0XZ5

Galke, L., Ram, Y., & Raviv, L. (2023). What makes a language easy to deep-learn? *arXiv:2302.12239*.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, *2*(11), 665–673.

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How Efficiency Shapes Human Language. *Trends in Cognitive Sciences*, *23*(5), 389–407. https://doi.org/10.1016/j.tics.2019.02.003

Gong, T., Minett, J. W., & Wang, W. S.-Y. (2008). Exploring social structure effect on language evolution based on a computational model. *Connection Science*, *20*(2-3), 135–153.

Goth, G. (2016). Deep or shallow, NLP is breaking out. *Communications of the ACM*, *59*(3), 13–16. https://doi.org/10.1145/2874915

Graesser, L., Cho, K., & Kiela, D. (2019). Emergent linguistic phenomena in multi-agent communication games. *EMNLP/IJCNLP (1)*, 3698–3708.

Guo, S., Ren, Y., Havrylov, S., Frank, S., Titov, I., & Smith, K. (2019). The emergence of compositional languages for numeric concepts through iterated learning in neural agents. *arXiv preprint, abs/1910.05291*.

Harris, Z. S. (1954). Distributional Structure. *WORD*, *10*(2-3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Havrylov, S., & Titov, I. (2017). Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. *NeurIPS*, 2149–2159.

Hockett, C. F. (1960). The origin of speech. *Scientific American*, *203*(3), 88–97.

Huebner, P. A., Sulem, E., Cynthia, F., & Roth, D. (2021). BabyBERTa: Learning more grammar with small-scale child-directed language. *Proceedings of the 25th Conference on*

Computational Natural Language Learning, 624–646. https://doi.org/10.18653/v1/2021.conll-1.49

Kanwal, J., Smith, K., Culbertson, J., & Kirby, S. (2017). Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. *Cognition, 165,* 45–52.

Kirby, S. (2002). Learning, bottlenecks and the evolution of recursive syntax.

Kirby, S. (2017). Culture and biology in the origins of linguistic structure. *Psychonomic bulletin & review, 24*(1), 118–137.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences, 105*(31), 10681–10686.

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current opinion in neurobiology, 28,* 108–114.

Kirby, S., Smith, K., & Brighton, H. (2004). From ug to universals: Linguistic adaptation through iterated learning. *Studies in Language, 28*(3), 587–607.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition, 141,* 87–102.

Kosoy, E., Collins, J., Chan, D. M., Huang, S., Pathak, D., Agrawal, P., Canny, J., Gopnik, A., & Hamrick, J. B. (2020). Exploring exploration: Comparing children with rl agents in unified environments. *Bridging AI and Cognitive Science workshop at ICLR*.

Kosoy, E., Liu, A., Collins, J. L., Chan, D., Hamrick, J. B., Ke, N. R., Huang, S., Kaufmann, B., Canny, J., & Gopnik, A. (2022). Learning causal overhypotheses through exploration in children and computational models. In B. Schölkopf, C. Uhler, & K. Zhang (Eds.), *Proceedings of the first conference on causal learning and reasoning* (pp. 390–406). PMLR. https://proceedings.mlr.press/v177/kosoy22a.html

Kottur, S., Moura, J., Lee, S., & Batra, D. (2017). Natural language does not emerge 'naturally' in multi-agent dialog. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2962–2967. https://doi.org/10.18653/v1/D17-1321

Lazaridou, A., & Baroni, M. (2020). Emergent multi-agent communication in the deep learning era. *arXiv preprint, abs/2006.02419*.

Lazaridou, A., Hermann, K. M., Tuyls, K., & Clark, S. (2018). Emergence of linguistic communication from referential games with symbolic and pixel input. *ICLR*.

Lazaridou, A., Peysakhovich, A., & Baroni, M. (2017). Multi-agent cooperation and the emergence of (natural) language. *ICLR*.

Li, B. Z., Nye, M., & Andreas, J. (2021). Implicit representations of meaning in neural language models. *Proc. of ACL*, 1813–1827. https://doi.org/10.18653/v1/2021.acl-long.143

Li, F., & Bowling, M. (2019). Ease-of-teaching and language structure from emergent communication. *NeurIPS*, 15825–15835.

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., & Wattenberg, M. (2023). Emergent world representations: Exploring a sequence model trained on a synthetic task. *Proc. of ICLR*. https://openreview.net/forum?id=DeG07_TcZvT

Lian, Y., Bisazza, A., & Verhoef, T. (2021). The Effect of Efficient Messaging and Input Variability on Neural-Agent Iterated Language Learning. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 10121–10129. https://doi.org/10.18653/v1/2021.emnlp-main.794

Lian, Y., Bisazza, A., & Verhoef, T. (2023). Communication Drives the Emergence of Language Universals in Neural Agents: Evidence from the Word-order/Case-marking Trade-off. *Transactions of the Association for Computational Linguistics, 11*, 1033–1047.

Liu, D., Lamb, A. M., Kawaguchi, K., ALIAS PARTH GOYAL, A. G., Sun, C., Mozer, M. C., & Bengio, Y. (2021). Discrete-Valued Neural Communication. *Advances in Neural Information Processing Systems, 34*, 2109–2121.

Lowe, R., Gupta, A., Foerster, J., Kiela, D., & Pineau, J. (2019). Learning to learn to communicate. *Proceedings of the 1st Adaptive & Multitask Learning Workshop*.

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PloS one*, *5*(1), e8559.

Lutzenberger, H., De Vos, C., Crasborn, O., & Fikkert, P. (2021). Formal variation in the kata kolok lexicon. *Glossa: a journal of general linguistics, 6*.

Meir, I., Israel, A., Sandler, W., Padden, C. A., & Aronoff, M. (2012). The influence of community on language structure: Evidence from two young sign languages. *Linguistic Variation, 12*(2), 247–291.

Michel, P., Rita, M., Mathewson, K. W., Tieleman, O., & Lazaridou, A. (2023). Revisiting Populations in multi-agent Communication. *The Eleventh International Conference on Learning Representations*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems 26*, 3111–3119.

Mitchell, T. M. (1980). The Need for Biases in Learning Generalizations.

Mudd, K., De Vos, C., & De Boer, B. (2020). An agent-based model of sign language persistence informed by real-world data. *Language Dynamics and Change*, *10*(2), 158–187.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., & Sutskever, I. (2021). Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, *2021*(12), 124003.

Ndousse, K. K., Eck, D., Levine, S., & Jaques, N. (2021). Emergent social learning via multi-agent reinforcement learning. *International conference on machine learning*, 7991–8004.

Nettle, D. (2012). Social scale and structural complexity in human languages. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1597), 1829–1836.

Newman, M. E. (2005). Power laws, pareto distributions and zipf's law. *Contemporary physics*, *46*(5), 323–351.

Nikishin, E., Schwarzer, M., D'Oro, P., Bacon, P.-L., & Courville, A. (2022). The Primacy Bias in Deep Reinforcement Learning. *Proceedings of the 39th International Conference on Machine Learning*, 16828–16847.

Nölle, J., Staib, M., Fusaroli, R., & Tylén, K. (2018). The emergence of systematicity: How environmental and communicative factors shape a novel communication system. *Cognition*, *181*, 93–104. https://doi.org/10.1016/j.cognition.2018.08.014

Ohmer, X., König, P., & Franke, M. (2020). Reinforcement of semantic representations in pragmatic agents leads to the emergence of a mutual exclusivity bias. *CogSci*.

OpenAI. (2023). GPT-4 Technical Report. *arXiv:2303.08774*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. (2022).

Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Patel, R., & Pavlick, E. (2022). Mapping language models to grounded conceptual spaces. *Proc. of ICLR*. https://openreview.net/forum?id=gJcEM8sxHK

Perfors, A., & Navarro, D. J. (2014). Language evolution can be shaped by the structure of the world. *Cognitive science*, *38*(4), 775–793.

Portelance, E., Frank, M. C., Jurafsky, D., Sordoni, A., & Laroche, R. (2021). The emergence of the shape bias results from communicative efficiency. *Proceedings of the 25th Conference on Computational Natural Language Learning*, 607–623.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*, 140:1–140:67.

Raviv, L., de Heer Kloots, M., & Meyer, A. (2021). What makes a language easy to learn? a preregistered study on how systematic structure and community size affect language learnability. *Cognition*, *210*, 104620.

Raviv, L., Meyer, A., & Lev-Ari, S. (2019a). Compositional structure can emerge without generational transmission. *Cognition*, *182*, 151–164.

Raviv, L., Meyer, A., & Lev-Ari, S. (2019b). Larger communities create more systematic languages. *Proceedings of the Royal Society B*, *286*(1907), 20191262.

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, *111*(3), 317–328. https://doi.org/10.1016/j.cognition.2009.02.012

Resnick, C., Gupta, A., Foerster, J. N., Dai, A. M., & Cho, K. (2020). Capacity, bandwidth, and compositionality in emergent language learning. *AAMAS*, 1125–1133.

Rita, M., Chaabouni, R., & Dupoux, E. (2020). "lazimpa": Lazy and impatient neural agents learn to communicate efficiently. *CoNLL*, 335–343.

Rita, M., Strub, F., Grill, J.-B., Pietquin, O., & Dupoux, E. (2022). On the role of population heterogeneity in emergent communication. *ICLR*. https://openreview.net/forum?id=5Qkd7-bZfI

Rita, M., Tallec, C., Michel, P., Grill, J.-B., Pietquin, O., Dupoux, E., & Strub, F. (2022). Emergent Communication: Generalization and Overfitting in Lewis Games. *Advances in Neural Information Processing Systems*.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences, 118*(45), e2105646118. https://doi.org/10.1073/pnas.2105646118

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Selten, R., & Warglien, M. (2007). The emergence of simple languages in an experimental coordination game. *Proceedings of the National Academy of Sciences, 104*(18), 7361–7366.

Singhal, P., Goyal, T., Xu, J., & Durrett, G. (n.d.). A long way to go: Investigating length correlations in RLHF [to appear in the Conference on Language Modeling 2024]. *arXiv:2310.03716*.

Smith, K. (2022). How language learning and language use create linguistic structure. *Current Directions in Psychological Science, 31*(2), 177–186.

Smith, K., Brighton, H., & Kirby, S. (2003). Complex systems in language evolution: The cultural emergence of compositional structure. *Advances in complex systems, 6*(04), 537–558.

Smith, K., & Kirby, S. (2008). Cultural evolution: Implications for understanding the human language faculty and its evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences, 363*(1509), 3591–3603.

Srikant, S., Lipkin, B., Ivanova, A. A., Fedorenko, E., & O'Reilly, U.-M. (2022). Convergent representations of computer programs in human and artificial neural networks. *Advances in Neural Information Processing Systems*.

Steels, L. (2016). Agent-based models for the emergence and evolution of grammar. *Philosophical Transactions of the Royal Society B: Biological Sciences, 371*(1701), 20150447.

Szabó, Z. G. (2022). Compositionality. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2022). Metaphysics Research Lab, Stanford University.

Thompson, B., Raviv, L., & Kirby, S. (2020). Complexity can be maintained in small populations: A model of lexical variability in emerging sign languages.

Tieleman, O., Lazaridou, A., Mourad, S., Blundell, C., & Precup, D. (2019). Shaping representations through communication: Community size effect in artificial learning systems. *arXiv:1912.06208*.

Tomasello, M. (2008). *Origins of Human Communication*. MIT Press.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Träuble, F., Goyal, A., Rahaman, N., Mozer, M. C., Kawaguchi, K., Bengio, Y., & Schölkopf, B. (2023). Discrete Key-Value Bottleneck. *Proceedings of the 40th International Conference on Machine Learning*, 34431–34455.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems 30*, 6000–6010.

Warstadt, A., & Bowman, S. R. (2022). What Artificial Neural Networks Can Tell Us About Human Language Acquisition. *arXiv:2208.07998*.

Warstadt, A., Choshen, L., Mueller, A., Williams, A., Wilcox, E., & Zhuang, C. (2023). Call for papers – the BabyLM challenge: Sample-efficient pretraining on a developmentally plausible corpus. *arXiv:2301.11796*.

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 1–16. https://doi.org/10.1038/s41562-023-01659-w

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent abilities of large language models. *Transactions of Machine Learning Research, 2022*. https://openreview.net/forum?id=yzkSU5zdwD

Winters, J., Kirby, S., & Smith, K. (2015). Languages adapt to their contextual niche. *Language and Cognition*, *7*(3), 415–449.

Wray, A., & Grace, G. W. (2007). The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua*, *117*(3), 543–578.

Zeigler, B. P., Praehofer, H., & Kim, T. G. (2000). *Theory of modeling and simulation*. Academic press.

Zhong, K., Song, Z., Jain, P., Bartlett, P. L., & Dhillon, I. S. (2017). Recovery Guarantees for One-hidden-layer Neural Networks. *Proceedings of the 34th International Conference on Machine Learning*, 4140–4149.

Zhou, H., Vani, A., Larochelle, H., & Courville, A. (2022). Fortuitous Forgetting in Connectionist Networks. *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*.

Zuidema, W. (2002). How the poverty of the stimulus solves the poverty of the stimulus. *Advances in neural information processing systems, 15*.

Data, Code and Materials Availability Statement

This review paper does not introduce any new data, code, or materials.

Authorship and Contributorship Statement

LG conceptualized the idea, reviewed the literature and wrote the paper. LR conceptualized the idea and helped write the paper.

Acknowledgements

We thank Mitja Nikolaus and Mathieu Rita for insightful comments and discussions. We thank Eva Portelance and Michael C Frank for their valuable comments on an initial version of the manuscript.

License

Whither developmental psycholinguistics?

Victor Gomes
University of Pennsylvania, USA

Abstract: Large Language Models (LLMs; e.g., GPT-n) have attracted the attention of psycholinguists who see a potential for solutions to ancient problems in them. This paper argues that, thus far, LLMs have not, in fact, suggested any new solutions, but instead just appear to by virtue of their sheer size and “double” opaqueness (both as models and as products). In the realm of cross-situational word learning, LLMs run into the same issues that long-discussed "global models" do in accounting for the rapidity and low-resourced nature of language acquisition. In the realm of meaning, they run into largely the same issues as the long-established conceptual theories they are often compared to. In neither case do they appear to represent a true resolution to known issues, and as such broadly encouraging the use of LLMs in developmental psycholinguistics is a gamble. This paper then argues that LLMs come with a range of immediate costs (to privacy, labor, and the climate) and so encouraging their use is not simply a low-risk gamble. These costs should be kept in mind when deciding whether to conduct any research with LLMs, whether it is to prove that they have some capacity or lack it. One way of keeping these costs in mind is to learn about them and talk about them with each other, rather than deciding that ethical questions are solely under the purview of some other discipline(s).

Keywords: LLMs, word learning, concepts, compositionality, AI ethics

Corresponding author(s): Victor Gomes, Department of Psychology, Harvard University, 33 Kirkland Street, Cambridge, MA, USA. Email: vgomes@fas.harvard.edu.

ORCID ID(s): https://orcid.org/0000-0002-8251-080X

Citation: Gomes, V. (2025). Whither developmental psycholinguistics? *Language Development Research, 5*(1), 144—187. http://doi.org/10.34842/gomesllm

Introduction

This passion of our kind
For the process of finding out
Is a fact one can hardly doubt,
But I would rejoice in it more
If I knew more clearly what
We wanted the knowledge for…

- W.H. Auden, 1962

Like any story that is old enough, it was only a matter of time until the connectionism/symbolism debate was deemed fit for rebooting. However, this time, the performance criticism seems wholly irrelevant, as these transformer-based Large Language Models (LLMs) are capable of generating grammatical and seemingly relevant sentences. Since these debates are complicated by the many potential points of disagreement (e.g., the Whorfian question) that can pop up while discussing any specific topic (e.g., conceptual structure), it is important to be clear that this paper by no means aims at exhaustively reviewing all questions that LLMs have been argued to be relevant to in the philosophy of language. While various positions on LLMs and learning and meaning may sometimes cluster, they do not neatly separate into two stable camps. Though I believe these models also fail to move the needle in debates in other areas of developmental psycholinguistics, this paper will not discuss whether LLMs strike down poverty of the stimulus arguments or whether they prove anyone wrong or right (Kodner, Payne, & Heins, 2023; Katzir, 2023; Rawski & Baumont, 2023; Milway, 2023) or whether their mechanisms are biologically plausible as connectionist naming conventions have at times suggested (Yang & Wang, 2020). While these questions are relevant to word-learning researchers, they would greatly extend the length of the present paper. Instead, this paper will focus on meaning and learning the meanings of words because it is my focus and because most people were not excited about GPT because it could produce grammatical sentences.[1] No, it is because GPT-n's outputs go beyond grammaticality to seem relevant and, as a result, seem "meaningful" to users. This has led many to argue that such models exhibit some kind of understanding (see Mitchell & Krakauer, 2023 for a review of such claims), and its outputs are therefore meaningful much like ours. So, this paper will ask: how do LLMs come to represent words, and what do they represent about them? Do humans do things similarly, and therefore, could LLMs provide insight into how children learn words? Have any long-standing problems truly been settled by LLMs? Often, word learning researchers break things down into two broad questions: 1) how words are paired with meanings, and 2) how those meanings are structured. The former is a question of Cross-Situational Word Learning (CSWL), and the latter is one of conceptual content. I aim to argue that, in both cases, 1) LLMs do not present radically new theoretical approaches as they 2) still seem to possess the same issues as those

[1] Impressive as this may be to us experts.

previous, similar approaches when inspected closely and therefore 3) do not resolve any long-known issues.

This paper will argue that these performance improvements, while very impressive, are still data-dependent. Like their ancestors, LLMs struggle to generalize beyond the datasets they are trained on. However, proving this has become much more challenging. This is not just because they are trained on immense amounts of data but also because current LLMs are products sold by businesses that have deemed details about training data proprietary to maintain a competitive advantage (e.g., see OpenAI, 2023a). Ultimately, this paper aims to caution developmental psycholinguists against using extant LLMs, which are first and foremost products. This is especially true if one lacks clear motivation and plans for interpretation (i.e., not just "to see what it can do" and publish it). Importantly, this is to say that I am *not* cautioning against the exploration of emerging modeling approaches and architectures (like transformers) to see if they address any of of the problems reviewed in **Too Holistic** and **Too Global** *unless* doing so requires paying the same costs to privacy, labor, and the climate that are outlined in **Too Costly**.

Roadmap

I will begin by clarifying what is being critiqued and what is not, then introduce the notion of compositionality and explain why it is still relevant today. Based on a range of tasks with compositional systems (primarily math and language), LLMs still struggle to generalize beyond their experience in a manner comparable to human beings and this is to be expected given where they may fit into these debates. In the case of semantics, I will first argue that LLMs as cognitive theories of language are too holistic an approach to meaning (a Conceptual Role Theory (CRT) to semantics; Piantadosi & Hill, 2022; Block, 1986; Block, 2016; Fodor & Lepore, 1992), and, in the case of early CSWL, that they are too global an algorithm of CWSL (Stevens et al., 2017). In neither case should any critique presented be read as suggesting that there is no space for connectionist or interactive approaches in CWSL or semantics and that the answer to such questions will necessarily be purely symbolic, modular, domain-specific innate, etc. Instead, the primary goal of the critiques presented is to situate LLMs within long-standing debates and note potential limitations associated with the approaches LLMs have been likened to. This allows us to then ask whether LLMs have resolved any of the existing issues of the approaches they have been likened to. In other words, if LLMs instantiate (or are otherwise similar to) global CSWL models or CRTs, do they resolve standing critiques that have been made of those approaches? Would LLMs do well in the sorts of situations that these frameworks have historically struggled to accommodate? As a consequence, the focus will be on explaining the critiques rather than giving an exhaustive review of both early CSWL and compositionality, as successfully reviewing both sides of both debates would require much more than a single article. Suffice it to say there is much debate on both sides in the realm of how central compositionality is to language (see chapters in Calvo & Symons, 2014 for arguments that compositionality is irrelevant to contemporary researchers) and how many word learning mechanisms there are, how much they vary across individuals, tasks, development, and so on (Roembke et al., 2023). The critical point is not that the critiqued positions are uncontroversially wrong and some others are uncontroversially correct - but rather that such controversies remain despite the development and success of LLMs and are likely to remain.

The paper will also additionally spend some time on the issues LLMs pose to interpretation (**Practical Meta-Theoretical Concerns**), which further limit their potential to resolve any existing controversial debates straightforwardly. That is what has remained the same, but what has changed is the social and legal context surrounding the production of these models. To that end, the paper will end with a brief but critical discussion of how these models are produced and governed solely by an industry that operates with little oversight. This section will discuss the consequences of the fact that their development and continued maintenance require immense amounts of infrastructure that is mainly made invisible to end-users (Birhane, 2020; McQuillan, 2022). Bringing all of that together, I plan on arguing that Large Language Models (LLMs) are too much: Too global, too holistic, and yet still not systematic enough (Fodor & Pylyshyn, 1988; Fodor & Lepore, 1992), and as such they fail to settle any long-standing debates decisively. I will finish by arguing that the current social context should make us think twice about integrating these models into how we do science and that the costs of using these models should be seriously considered before employing.

How LLMs M Ls

LLMs are not just large; they are also (at least so far) transformer-based architectures. The attention mechanism which transformers implement introduced two primary advantages over previous approaches (e.g., RNN, LSTM models): 1) transformers can conduct some computations in parallel, and 2) they have a better “memory.” Their increased efficiency due to (1) allows these models to be trained on larger datasets more quickly. As for (2), this is because transformers are not as limited as prior models have been in their ability to access previous states of the model (e.g., facts about the state of parameters x sentences ago). Both (1) and (2) are thanks to features of the attention mechanism. Before discussing attention, however, it is important to note that transformer-based architectures also inherit familiar features from past connectionist approaches. Weights are still randomly set at the start of training (though now, there are additional weights since there are more components). LLMs still tokenize words into subword tokens to approach something more like a morphemic representation (e.g., *birdhouse* > *bird, house*[2]). They even regularly include multi-layer perceptrons (Radford et al., 2019; Vaswani et al., 2017; Brown et al., 2020; Linzen & Baroni, 2021).

The primary difference from previous models is therefore the presence of a decoder or encoder, which implements similar but distinct attention mechanisms. At its most basic, attention allows a model to consider a string ("The rats the cat chased hate themselves.") and for each word (*rats*), identify the other words that are likely related (*The, hate, themselves*) without being as heavily biased by recency (e.g., by being biased towards *cat* in guessing the number agreement for *hate* simply because it came later than *rat*; Galassi et al., 2021). Attention allows for the model to track more information about each token than previous approaches and pass this more detailed information onto other layers (e.g., to a feedforward neural network). Most transformer-based models employ layers of various attention "heads." As a consequence of these features of transformers, aspects of training[3] have also changed. Unlike previous models, which were solely tasked with "predicting the next word," some transformer-based architectures could be more aptly said to "predict the missing word." Because these models use positional encoding, "predicting the missing word" allows them to use more than just the preceding tokens in translating a text. This is accomplished by randomly masking a certain percentage of words and asking the model to guess the missing word using context "from the future" (e.g., "the best lack all conviction" might become "the MASK lack all conviction" rather than guessing what would come after *the*). Some transformer-based models, especially those tasked with machine translation, employ attention in two kinds of layers: an encoder and a decoder. However, encoder-decoder models require more computation (you have to train an encoder) and more annotated data (paired sentences in source vs. target language). While the ability to conduct masked training is a clear benefit from an engineering perspective, it is not clear whether this is a motivated model of human language learning (i.e., accurate to the time course of early cross-situational word learning). But, more practically for this paper, many of the widespread LLMs today do not use decoder-encoder architectures, often opting for just a decoder (Fu et al., 2023). In the case of decoders, "future" information (that is, words one has not yet encountered) is negatively weighted, so it does not meaningfully affect the output. As such, whether a model can be said to "predict the next word" or "to predict the missing word(s)" depends on the model and cannot be generalized to a claim about how all LLMs are trained.

[2] The following conventions will be adopted: italics will be used for mentions of words, caps lock will be used as a shorthand for concepts (meanings), such that I would say *pink* means PINK. Double quotes will be used for sentences, whether spoken by another or not. Furthermore, examples will always use English words for ease of reading, even though LLMs operate at a subword level.

[3] *Training* will be used interchangeably with *pre-training* when discussing LLMs, except for in particular cases where questions about continued training arise (e.g., in **Too Costly** when considering environment costs).

As the success of LLMs is often credited to the development of transformers and the attention mechanism, a critique of current LLMs may, therefore, also seem to be a critique of transformers, but that is not the goal of the present paper. Transformers, like n-grams or Bayesian approaches, may be an interesting and useful addition to the modeler's toolkit when investigating particular questions. Based on features shared by things currently called LLMs, as well as the ethical questions discussed, I will cautiously suggest that the present critique primarily applies to 1) transformer-based architectures 2) with an immense number of parameters that are 3) trained on an immense amount of data, and that likely 4) have no specialized subsystems which bake in rules.[4] While it is possible that the issues LLMs face are or may become relevant to other models that do not perfectly satisfy those conditions (e.g., hybrid approaches, yet-to-come approaches that are not transformer-based but meet 2-4), that will require more specific details about the model in question.[5]

Too Holistic

We are already[6] a bit into GPT-4 (Achiam et al., 2023), and like any good reboot, the stakes have increased. The audience demands that much more than just the local hamlet is in danger, and so the claim is that we are seeing "sparks of artificial general intelligence" (Bubeck et al., 2023). The new model can seemingly write code and, perhaps most shockingly, is capable of Theory of Mind. Now, of course, there are some practical caveats we should attend to: descriptions of theory of mind tasks and others are very likely present in its training set (in code, Narayanan & Kapoor, 2023; and in logic, Liu et al., 2023), passing any task is not proof of some capacity without auxiliary assumptions (Guest & Martin, 2022), greater care should be taken in applying "rich" psychological terms to AI (Shevlin & Halina, 2019), and so on. However, momentarily running with the claim that GPT-4 may be able to reason about minds, it is bewildering in light of all these social and general task-based competencies that it struggles so much with mathematical and logical reasoning (related issues hold for earlier

[4] This is because, for example, GPT-4 performing well with arithmetic prompts when given access to the Wolfram Alpha plugin likely says less about GPT-4 than Wolfram Alpha, and at the very least complicates the question of which to credit.

[5] I ask that the reader keep in mind that LLM is a marketing term referring primarily to size rather than a term with clearly defined formal or cognitive commitments (Portelance & Jasbi, 2023). Most current LLMs are mostly transformer-based, but that does not guarantee this word will always be used to describe only transformer-based models. It does not even guarantee that future transformer-based models in this vein are guaranteed to be called LLMs, for example, if the term were to become skunked. This means it is difficult, if not outright impossible, to provide any truly in-principle critique of LLMs, as it seems unlikely the LLM refers to a principled category (e.g., as opposed to n-gram).

[6] At the time this was originally written and submitted.

models; see Lake & Murphy, 2021). Dziri et al. (2024) found that both chatGPT and GPT-4 achieve 55% and 59% accuracy on multiplication problems that involve two three-digit numbers (e.g., 123 times 456). For context, Adults typically performing near ceiling on comparable tasks (Miller et al., 1984; Geary et al., 1993; LeFevre et al., 1996). Work published since the submission of this article has found that even newer models display stark fragility in mathematical reasoning task, with accuracy varying both when information critical to the problem (i.e., number) as well as superficial information (i.e., name) are changed (Mirzadeh et al., 2024).[7]

Failures on these mathematical tasks should concern those hoping for a semantic theory, as it suggests that LLMs do not systematically understand the tokens underlying these digits - what else could explain the effect of linear order? Indeed, Dziri et al. (2024) suggest that such tasks are accomplished through linearized subgraph matching, rather than compositionally (i.e., by combining symbols according to rules to create/understand novel descriptions in a systematic manner, but see next section for extended discussion of compositionality). Regardless of how they try to do it, if an LLM were able to capture compositionality, then it should certainly be able to do simple arithmetic on unfamiliar sequences, at the very least to the same extent that people do based on their limited experience with infinity. Currently, they do not, and present research suggests that this may be an issue that scale cannot resolve but may rather serve primarily to obfuscate. LLMs struggle with logical reasoning (Liu et al., 2023; Arkoudas, 2023) and coding (Narayanan & Kapoor, 2023) when tested on benchmarks outside of the training set. Training models on more and more data may create an illusion of competency, as it reduces the chance that users (both academics and non-) will encounter failures in compositionality in typical use. Some may respond that people are not all equally great and regular at math/logic either; they may struggle when multiplying large numbers or interpreting a sentence with multiple negations. Is this because their representational systems are non-compositional? No, what makes people vary in math performance (aside from access to math education) probably has little to do with their syntactic and semantic representations of the rules of arithmetic. Instead, it is easily explained by performance factors (e.g., misremembering/forgetting, limited memory, being tired, being in alternate states of experience). The reason we struggle with larger numbers likely has more to do with the fact that as more operations need to be completed, there is more opportunity for a host of errors to occur rather than not having observed the multiplication of enough, e.g., 5-digit numbers before. Or alternatively, humans may exhibit errors as the result of testing different strategies. Indeed, some have pointed out that many of the errors exhibited in children learning arithmetic are "rational" errors – that is, applying a rule incorrectly (e.g., always subtracting the smaller digit from the larger (e.g., 202-133 =131 rather than 69 because 2-1=1 3-0=3 and 3-2=1; see VanLehn, 1990; Ben-Zeev,

[7] As suggested by earlier findings on the effect of irrelevant information on LLM mathematical performance (Shi et al., 2023).

2012). However, as mentioned earlier, these errors are eventually overcome as adults near ceiling (Miller et al., 1984; Geary et al., 1993; LeFevre et al., 1996). It is not shocking that LLMs do better at things in the training set, nor perhaps things within a certain distance from it (were there a straightforward way to quantify that for such open-ended tasks). The trouble is that the productivity of language means we may never approximate its systematicity solely by gathering more and more data or adding more parameters. These issues are clearer (and more down to Earth) when considering image-from-text models that incorporate LLMs into their architecture, like DALL-E 2 (Ramesh et al., 2022) and 3 (Betker et al., 2023), Stable Diffusion (Rombach et al., 2022) as well as multi-modal models like GPT-4 (Achiam et al., 2023) and Gemini (Team et al, 2023). The issues faced by image generation models perhaps more clearly demonstrate the ways these approaches struggle with composition, and it additionally allows us to consider whether adding more modalities resolves long-standing similar questions in the philosophy of language and concept literature. Before relating these issues to known criticisms of theories LLMs have been likened to, it will help to briefly discuss why these issues with simple compositional systems might suggest that LLMs are not learning in a manner that meaningfully generalizes from the training data and that their impressive performance may be largely due to their sheer coverage and the amount of information it can store.

Compositional systems assume regularity to represent discrete combinatorial infinity (i.e., no largest number, no longest sentence). This makes it easy for researchers to generate data for training and testing by controlling for features that are irrelevant to some given formalism. For example, linear order does not matter in summation as it is commutative; therefore, a system trained on a single-digit addition dataset in which the larger number comes last (e.g., 1+2, 2+3, 4+5) and performs at chance when the larger number comes first (2+1, 3+2, 5+4) can not reasonably be said to have generalized the rules of addition, when approaching higher digits that are likelier to be outside the training data, the rules of arithmetic fall apart for LLMs. If one learns to add in general, one should learn that it applies regularly beyond the training set - even if an advantage on familiar items remains. However, a drastic difference in performance between training and test suggests that a given model has not converged on the rules of the compositional system but is instead being swayed by other information. Though mostly linguistic examples will be used, this is also not to imply that compositional rules are all that is required to explain all verbal behavior - indeed, linguistic performance is uncontroversially shaped by a range of factors that are very unlikely to be compositional as spelled out (e.g., frequency effects). Any exhaustive account of verbal behavior will have to, at the very least, make some space for non-compositional mechanisms. And, though there is debate about the compositionality of language, there are those who feel compositionality is an important part of understanding how languages work (e.g., see Quilty-Dunn et al., 2023 and responses for many appeals to compositionality in contemporary literature). But, regardless of one’s beliefs about the extent of compositionality in language, compositionality is an especially useful guide in the present moment because it allows us to set a standard for successfully learning a rule. Such a standard would likely be less necessary were there more transparency about the data these models are trained on, as it would, therefore, be widely possible to determine how similar a new set of stimuli is to the training data (though theoretical questions about the proper similarity metrics would still remain). Now that we have reviewed some data suggesting that LLMs[8] struggle with basic compositional systems like math and logic, we will now discuss compositionality more closely and how it has caused issues for conceptual frameworks of the past before relating this to LLMs.

Representational Theory of Mind & Compositionality

Interest in word learning often comes along with an interest in what it means to know a word. Not just how it relates to some form (e.g., a morpheme) or even purely distributional facts, but rather its meaning. What do words map onto? What are they like? Since questions about meaning and concepts are so intertwined with other fundamental questions in psychology, there is little consensus about the particulars. This is why talk about concepts is so prone to desiderata-listing, or what one would want a theory of concepts to do in the first place. An important one is that a theory of concepts is compatible with RTM, or the belief that propositional attitudes (e.g., wanting, believing, knowing) are relationships between individuals and mental representations (Fodor, 1975). To be fair, such ideas were old and fairly nontendentitious within psychology, but before Fodor, no one had thought to acronymize the name. If you add in the idea that the mind is like a computer, you get the Computational Theory of Mind (CTM), which says that the internal states of RTM are (classically syntactically) structured symbols. Under such a view, thinking involves combining and transforming symbols, and though LLMs are not classical, they still involve structural transformations. RTM is a "non-negotiable" because, without RTM, there cannot be any real psychological laws; they instead must ultimately reduce completely and directly to terms of more basic sciences (e.g., to neuroscientific laws, but potentially ultimately physical laws; Churchland, 1986). Such an extreme approach may slice questions too thin (as will be discussed in **Double Opaque**) and complicate discussion about the most relevant rules. For example, while studying what has been used as currency helps in understanding the histories of economies and markets, attempting to provide translations of economic generalizations into physical descriptions of items and their transfer may result in missing the forest for the trees (Fodor, 1980). CTM is "non-negotiable" because it is our "best available theory of mental processes" – that is, computers are our best working models of a physical system that is capable of representation that can be discussed at a meaningful level (Fodor, 1985). In linguistics, both are considered deeply related to the compositionality of language. To say that

[8] This may indeed be a case where issues with LLMs straightforwardly translate to current approaches focusing on transformers.

language is compositional is to say that whatever some sentence means is going to be a function of its constituents plus the rules of syntax (Frege, 1892; Fodor and Lepore, 1992).

1. Monroe married Luis.
2. Luis married Monroe.

Systematicity requires that if you are the type of thinker that can think the thought expressed by (1), you are also necessarily the type of thinker that can think the thought expressed by (2). Assuming you know other words, you are also the type of thinker who can think other thoughts involving *Monroe, Luis,* and *married*. In other words, you also get productivity, or the idea that theres no upper limit on the longest sentence you could generate, assuming one's syntax allows for recursion. Compositionality thus guarantees systematicity and productivity respectively (making infinite use of finite means as per von Humboldt (1836) qua Chomsky (2014)), which has been useful to both linguists and non- when thinking about language (Fodor & Lepore, 1992). While that may explain the meanings of sentences, that does not seem to tell us much about the constituents of sentences and how they get their meanings. However, keeping the constraints of CTM (due to compositionality) in mind will help in considering the following ideas about meaning, as the main issue will be that they struggle to allow for compositionality. We will discuss how this relates to one theory of concepts, Conceptual Role Theories,[9] and LLMs and how adding more modalities is unlikely to solve this problem. But before we continue, we will first consider one notable theory, the Classical Theory of Concepts, to demonstrate some of the difficulties with definitions and the consequences this has had for conceptual theorizing since.

Definition and its discontents

One popular and eloquent conceptual theory, the Classical Theory of Concepts, often associated with Locke (1850) and other British empiricists, is that the meanings of words allow us to pick things out in the world because they have a sensory (or perceptual) basis. This, along with a compositional system, should explain the productivity (or open-endedness) of language. Sensation provides a foundation for a compositional system to act on; this allows mental states to interact with the world causally. Thus, a color concept, like ORANGE$_{\text{color}}$, can be defined by the sensations triggered during labeling contexts, cones responding to light with a wavelength of 585 and 620 nanometers. In this example, the meaning of *orange*$_{\textit{color}}$ is the range of sensations that can cause ORANGE$_{\text{color}}$ thoughts.

[9] This is a theory with many aliases: Conceptual/Inferential/Functional Role Semantics, Procedural Semantics. Problems with analyticity aside, assume they are all synonymous with CRT in this case (Block, 1998).

These primitive concepts can then be used to modify the features of some other concept selectively; for example, ORANGE$_{\text{color}}$ FRUIT modifies the thought FRUIT (whatever they might be) so that any related color sensations are now ORANGE$_{\text{color}}$ rather than something else. In this example, the meaning of *orange*$_{\text{fruit}}$ may be a complex concept (ORANGE$_{\text{color}}$ FRUIT) rather than a primitive one. Complex concepts may then, in turn, be combined with other primitive and complex concepts, like KENNEL FOR ORANGE$_{\text{color}}$ DOGS. The Classical Theory of Concepts is eloquent because it not only accounts for reference, and hence more "synthetic truths" (truths by virtue of experience), but also maintains "analytic truths" (truths by virtue of meaning). So, not only can KENNEL pick out kennels in the world and be used to consider facts about them ("This is a kennel," "Julian left Charlie at his favorite kennel.") but also distinguish those facts from other beliefs that are central and necessary for the concept ("Kennels are shelters," "Kennels are for dogs," etc.). Similarly, it explains why kennels in the world reliably cause KENNEL thoughts but only sometimes lead to CHARLIE thoughts.

Though the Classical Theory is an elegant way of accounting for the referential and truth-preserving aspects of meaning, nothing gold can stay. Briefly put, its demise resulted from an inability to unite these two aspects of meaning in a non-circular way. The Classical Theory posits that a statement may be true for one of two reasons: due to the nature of the terms themselves and rules of syntax (analytic) or because they say something true about the world (synthetic). For example, you do not need to look to the world to determine whether someone being a bachelor makes them an unmarried man, but you do need to check it to determine whether some given individual is a bachelor (e.g., by asking them or others whether they are married). It is, therefore, compositional under this view: UNMARRIED MAN composes into BACHELOR, which can then be decomposed back into UNMARRIED MAN. Setting aside the difficulty this approach has in defining abstract words like "virtue," the biggest problem seemed to be that no one's ever found a good definition in general (Berkeley, 1881). It is unclear how you get to JUSTICE from RED and TINNY, but it is also unclear how you get to seemingly simpler, more concrete meanings like CHAIR. More recently, Quine (1951) argued that this is because the analytic/synthetic divide is circular: analyticity rests on an assumption of synonymy between a term and its definition such that they are interchangeable (e.g., BACHELOR could be subbed in with UNMARRIED MAN in any sentence and it remains true, and vice-versa). However, determining whether terms are synonymous requires a notion of necessity that distinguishes accidental coextension from the required extensions. For example, in "Necessarily all and only creatures with a heart are creatures with kidneys," both *creatures* have the same extension (because all known creatures with hearts have kidneys), but no one would argue this is an analytic fact (unlike "Necessarily all and only bachelors are unmarried men."). To Quine, this meant there was a vicious circularity in the distinction: analyticity requires synonymy, and synonymy requires interchangeability of terms without a change in meaning, but how is it determined if terms are interchangeable? If synonymy is determined by looking at our experiences in the world, then it cannot be the basis for analyticity at the pain of circularity. Though Quine's focus was primarily on scientists (or rather science) rather than word learners, similar concerns bear on concepts and therefore heavily influenced that debate.

With analyticity gone, so with it goes a straightforward distinction between matters of meaning and matters of experience. With the issue being the inadequacy of physical features for definition and also definition itself, one potential approach is to 1) allow for some sort of internal states (rather than purely sensational ones) and 2) relax definition to something more graded. Conceptual Role Theories (CRTs) explores these possibilities. The goal of this section is not to say that CRT is wrong but that if LLMs are like them, they leave the same issues unresolved (as in the last section), and this is evident in their performance on a range of tasks involving composition. The next section begins by defining CRTs before discussing why they have been argued to struggle to account for compositionality.

Conceptual Role Theories

Talking about CRTs requires casting a wide net, though, unlike LLMs, CRTs are much more precisely defined. Generally speaking, CRTs broadly agree that meaning is functional and that what constitutes the character of a mental state is the role it has in interacting with other mental states. This can be restated psycholinguistically as the idea that the meaning of a word is its role in a language, or as it is often put, that "meaning depends on role in a conceptual scheme" (Harman, 1999). For example, we make an inference when we go from the statement that "p" ("Grass is green.") and a separate statement that "q" ("They paint the grass.") to the statement "p and q" ("Grass is green and they paint the grass."). Natural language analogs to logical operators, like *and*, are go-to examples of non-referential meaning, and their role in a sentence is what defines them (Block, 1998). CRTs often extend this idea to all words. Block (1986) famously used the example of high-school physics, in which the meaning of words like force, acceleration, and mass are interdefinable (f=ma) within a conceptual scheme (physics) rather than translated into known words outside this system.[10] It is because it treats meaning as relational in this way that some have analogized it to LLMs since they learn (probabilistic) relations between tokens (Hill & Piantadosi, 2022; Pavlick, 2022). Importantly, this has been used to argue that referential abilities are not needed since reference is not necessary in CRT approaches. However, that is not entirely true. CRTs also often make room for other systems that are innate (e.g., core cognition like object or magnitude; Carey, 2009) or that ground reference (Block, 1998). These dual-role theories are popular, even amongst those who conceive of the CRT-relations between roles, like those of tokens in LLMs, as being probabilistic (Field, 1977). Notably, CRTs differ wildly in how they cash out interactions with the

[10] I have yet to see anyone mention it but this always struck me as bad pedagogy.

world, so we will leave that aside for now.

The basic issue with CRTs is that if some relations are seen as more central than others, some old problems discussed in the previous section are reintroduced (Fodor & Lepore, 1992). For example, while you can reliably infer something about Caio's age from "Caio is 28 years old," you can also reliably infer something about Caio's weight (>10 pounds). While we could say that the former is tied to the symbol and the latter is tied to the symbol plus auxiliary beliefs (e.g., 28-year-olds are adults for humans, and adults weigh more than ten pounds), drawing such a line is hard (example adapted from Fodor, 1984), and requires reintroducing a version (albeit fuzzier) of the analytic/synthetic split, but it is not clear how that resolves the circularity in question (Quine, 1960). Unfortunately, unless one can provide an answer to how the lines are drawn, that means inferential roles are not compositional. Consider the idea that I enjoy the flavor of ARTIFICIAL STRAWBERRY, and therefore, one of the inferences licensed by this fact is simply "Victor likes artificial strawberry." However, until college, I also happened to hate strawberries and was not typically big on artificial flavors either, so neither of its constituents would have licensed the inference "Victor likes it." Why not? Or consider the opposite scenario, wherein I like houses and boats but find houseboats vulgar and offensive. In this case, an inference is licensed by both constituents but not the complex concept they enter into, so where does it go? In both cases, the inferences that can be licensed are not inherited from the utterances' constituents. In the former, the inference is not present in constituents; in the latter, it is not composed of its constituents. That is because the inferential roles of both *artificial strawberry* and *houseboat* depend not just on the inferential role of their constituents but on what you believe about them. In other words, those inferences are synthetic rather than analytic, and, of course, it is important to separate the two (if one leans into the divide) to explain why it is people can think the same thing by "dog" despite likely differing in the synthetic inferences they would entertain about them (e.g., "I'm a dog person," "Labradoodles are not real dogs," and so on). In this sense, CRTs run into similar sorts of issues as prototype theories (Connolly et al., 2007), which is just to say that neither are compositional, though there are good reasons to think that our concepts are (Fodor & Lepore, 1992). LLMs struggle to learn simple compositional rules for similar reasons: there are so many possible associations between strings of digits in their training data, and it is not guaranteed that LLMs will land on the set of associations most strongly related to the compositional rules of arithmetic. The total context-sensitivity of tokens also likely complicates learning how to handle compositional systems, as how likely four is to follow three should have no effect at all on arithmetic, and the same holds for the variables and operators in logic.

How Do LLMs Fit in?

Before continuing, it is important to note that LLMs are composed of connectionist submodels, but this does not necessarily commit it to a particular conceptual framework (or, broadly speaking, cognitive framework; see Portelance & Jasbi, 2023). This is doubly true if one considers connectionist models as implementational rather than computational (i.e., in the way neurons implement the mind; Fodor & Pylyshyn 1988). Therefore, when claiming LLMs “have” a particular sort of semantics, this could be read as either a claim about them being capable of instantiating such a semantics (i.e., as a brain might) or as a claim about them being equivalent to a theory of semantics (i.e., moreso a claim about the mind; Blank, 2023). I am skeptical that LLMs have CRT-like semantics in both senses or, at the very least, that little is gained by such an analogy presently. However, it appears the motivation for such claims seems to be that some consider LLMs to be plausible models of cognition (rather than simply implementational), but they cannot refer to the world (though see Mandelkern & Linzen, 2023), so from this basis, some critics (sensibly) argued that their representations of meaning are prima facie unlike ours (Bender & Koller, 2020). Fortunately, CRTs allow for aspects of meaning that are non-referential, so perhaps CRTs are what LLMs have (Hill & Piantadosi, 2022; Pavlick, 2022). I have yet to encounter a more robust argument for this analogy, but there is already a systematic review of why connectionist models are problematic models of cognition, and I am assuming it is in the common ground (Fodor & Pylyshyn, 1988; for a reply see Smolensky, 1991 and Smolensky, 1988, and for a reply to those replies see Fodor & McLaughlin, 1990). This paper will therefore focus on the potential that LLMs are implementations of CRTs.[11]

If we try to consider LLMs as CRTs, the first issue we run into is that the idea of a conceptual role seems to presuppose a mapping to conceptual structures (Leivada et al., 2023). Indeed, as we will briefly touch on later, many two-factor theories assume that there are conceptual systems, like those of perception (e.g., analog magnitude and parallel individuation) or others that are part of core cognition (see Carey, 2009 for review). LLMs do not have any systems like those, but they can represent tokens and their related embeddings, so for now, we will assume they may have something like a conceptual role (in that it is representational and causal) even if they have significantly fewer types of conceptual roles or they are fundamentally unlike any of ours. If we try to translate LLMs into words familiar to the word learning literature (as will be discussed in **Too Global**), then an LLM’s hypothesis for the meaning of a token is its relationship to other tokens. This means that at the end of pre-training, the hope is that there is a pretty good hypothesis for relationships between tokens. The conceptual role in question here is the role a token has in predicting the next token because that is what it contributes to a sentence that contains it. Because of its mechanics, a token’s meaning is a consequence of how likely it is to carry information about another token or how likely it is to occur in the context of other tokens (while keeping its position in the string in mind). As such, though CRTs are not necessarily

[11] Naturally, these arguments will share the mouthfeel of critiques of connectionist models because CRTs and connectionist models both run the risk of holism, but there are clear divergences (e.g., CRTs obviously cannot serve as an implementation level theory).

committed to a predictive processing account generally, LLMs instantiating a CRT run into similar issues. That is, their representations do not seem to be compositional in the way concepts are because they are neither systematic nor productive. This is because an LLM's resulting hypotheses, despite being very complex, remain closely tied to their training (Lake & Murphy, 2021; Dziri et al., 2024), as the issues around the analytic/synthetic distinction should have prepared us for.

If the meaning of *cat* is what *cat* contributes to a prediction and it is related to a host of other words, then what *cat* means changes based on the current context (even if wholly irrelevant). Were it to change too drastically based on the current context, that presents an issue to systematicity. This is an issue because it means that *cat* would likely mean slightly different things in "The cat chased the rat" and "the rat chased the cat." If the difference were purely syntactic (subject vs object), or even homophony, that would be completely fine, but it is likely to vary in more ways. For example, the *cat* in the former may activate "things cat chase" more than it activates "things that chase cats," and ceteris paribus the *cat* in the latter. It is again important to note this is not to claim that there are no non-compositional mechanisms that can contribute to inferential processes more generally, merely that it is still common today to take seriously the notion that there is some sort of compositional component at play in language (see Quilty Dunn et al., 2023 and responses). This issue in systematicity, as we will see, leads to limitations in productivity, so we will now turn to empirical data showing that LLMs struggle with this, though it is important to keep general issues with benchmarks in mind (Narayanan & Kapoor, 2023).

More Modalities May Run into Similar Issues

A familiar argument we have discussed is that 1) maybe LLMs can represent things like we do, they just need to be more grounded, and 2) maybe LLMs do not represent things like we do *because* they are not grounded. In the case of the former, a solution may be sought in a two-factor version of CRT and, in the latter, in State-Space Semantics. The problem with the former is that the causal connection is still difficult to cash out in two-factor theories, and the problem with the latter is that it is rooted in similarity rather than conceptual role (Churchland, 1986, see Fodor & Lepore, 1999 for a reply). Though I disagree with these approaches personally, I am categorically not trying to suggest in any way that these approaches to meanings are psychologically or philosophically worthless or uncontroversially wrong. It just feels relevant that they also struggle with composition too. This is because the issue at hand is not simply with the format of the data (text vs. image) but rather the global and holistic nature of these approaches. Composition simply does not seem to fall out of solely trying to determine what is likely to happen next or what is similar to what. I will not speak more on two-factor CRT because there are many versions on offer, and many of those that interest developmental psychologists make recourse to some innate cognitive structures (e.g., see Carey, 2009 for an example of perceptual systems), which is not helpful in this regard since LLMs lack those. Instead, there will be a brief discussion on State-Space Semantics, its issues, and how they seem to arise in DALL-E 2 (among others).

The primary issue with state-space semantics is that similarity is not much less holistic than predicting the next word. This is because similarity hinges on what primitives one assumes (Goodman, 1965) and therefore in the absence of such commitments anything can be deemed similar to anything else (Goodman, 1972), which introduces the risk that any observation can support any hypothesis. Beyond that, it reintroduces the problems of the Classical Theory, but in a continuous rather than definitional manner - simply replacing identity with similarity. This results in a similarity space, with words getting their meaning by virtue of their position in this space. Instead of, e.g., GREEN being defined as BLUE PLUS YELLOW, GREEN is simply like BLUE and like YELLOW in a way that places it near both. Importantly, this similarity space is often assumed to be sensorimotor/perceptual. In the case of color, the dimensions would indicate the coding frequency for the reflectance of different wavelengths (Churchland, 1986). However, it is not clear how these dimensions are individuated and, therefore, which dimensions are innate. Furthermore, since such approaches are typically probabilistic, they introduce additional questions about how meanings in such approaches compose (for discussions, see Armstrong et al., 1983; Fodor & Lepore, 1996; though see Smith & Osherson, 1984 for a response to this line of criticism). But, like the Classical Theory, they ultimately run into the same issue: there are no good definitions. Setting these issues aside, we will now consider generative models that can produce images and how they run into the same sorts of issues we have been considering.

An Image Is Worth a Thousand Captions

Given that our minds seem to display the systematicity we are after, it is hard for us to imagine what sort of thinker could think “John loves Mary” but would be incapable of thinking “Mary loves John." I propose that such a mind, in the cleanest case, could not distinguish between the two descriptions, whether that means believing only one interpretation holds regardless of the linear order or believing that both interpretations always hold. Each sentence may be considered holophrastic (johnlovesmary and marylovesjohn are different words; importantly, with no internal structure), or arguments may be ignored, and features may be blended. That seems to be exactly what DALL-E 2 struggles with (Fig. 1). DALL-E 2 (Ramash et al., 2022) is another transformer-based system produced by OpenAI, but instead of predicting continuations of text, it generates an image that the text provided by the user is likely to be a caption of.

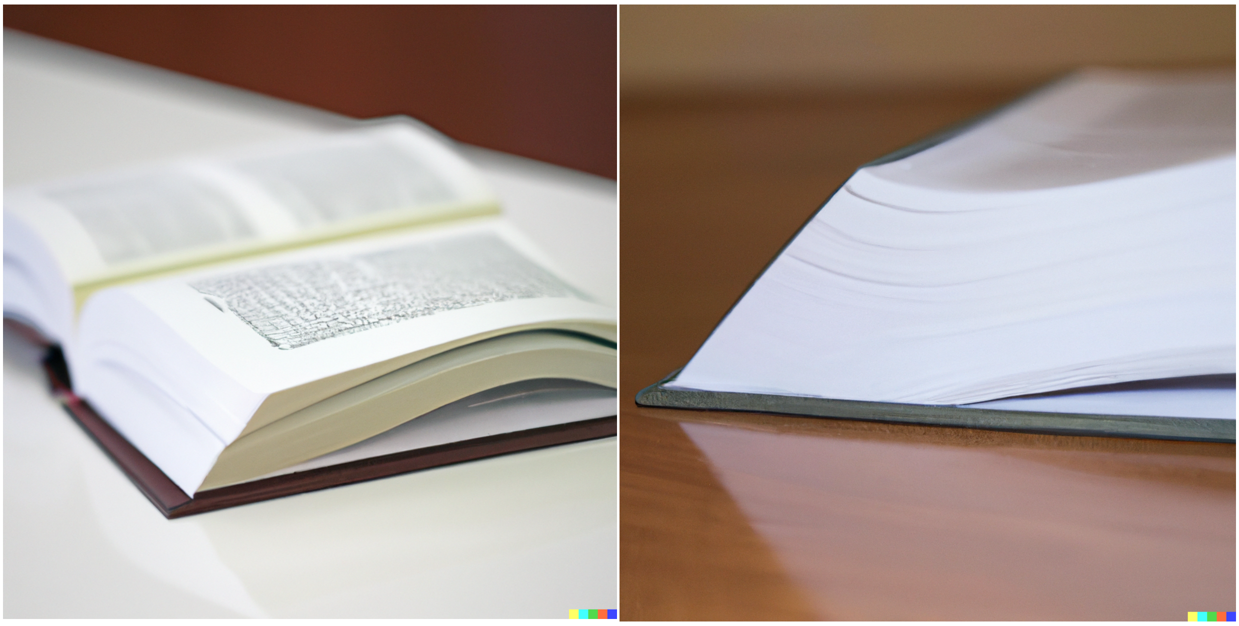

Figure 1. (Left) ***DALL-E 2 output for “A book on a table.”*** **(Right)** ***DALL-E 2 output for “A table on a book.” Representative of others in the set.***

The text encoder in CLIP is based on GPT-2 (Ramash et al., 2022; Radford et al., 2019), so that is the only point in the model at which it tracks the position of tokens within the input. The text embedding produced by it at this stage is then fed into a model, which is tasked with outputting an image embedding based on the text embedding, with the image embedding finally getting passed to the decoder to guide image generation. Each step includes a transformer model, but these last two steps also include diffusion models, which operate by reducing noise from an image towards some signal (e.g., the caption text, going from an image of static to an image of a cat through successive denoising; Ramash et al., 2022). Because of this, as Conwell & Ullman (2022) point out, information about the text’s position or even number may be outweighed by any of the steps beyond the initial encoding. This means that though image outputs can give insight into different aspects of meaning, which may be difficult to probe with text alone, they may not make full use of the information the model initially has about the text.

Conwell & Ullman (2022) investigated DALL-E 2’s ability to generate images based on relational prompts (e.g., “the book is on the table”). They generated 75 prompts by randomly sampling from a set of 15 relations (8 physical, seven agents) and 12 entities (6 animate, 6 inanimate) and used each to prompt D2. Online participants were then given a prompt and 3x6 array and asked to select the images that matched. On average, they found that participants were in 22% agreement with D2 across all relations, with 17% agreement on physical relations and 28% on agentic ones. Agreement varied greatly between relations, with only three relations significantly above 25% chance ("touching," "helping," and "kicking"). Though they did not provide an analysis of the generated images (other than participant response), the example images indicate a range of potential problems: producing a novel object, missing an item, and producing similar images for different relations/prompts. Most importantly, of course, these are the sorts of mistakes we would not expect people to make. Their drawings are likely to be worse or so abstract as to be difficult to understand what is what, but it is unlikely that upon hearing "draw me a cylinder on a cup," a person (or child) would regularly forget to draw the cup. While other work with more stimuli and newer models suggest a modest improvement in depicting spatial relations (around 45% based on human ratings; Huang et al., 2023), this work used a different approach where participants were given image-text pairs and asked to judge their appropriateness rather than presenting an array of images and asking participants to select ones which depicted the prompt. Additionally, no breakdown of performance by particular spatial relation (e.g., "touching" vs. "on") spatial relation type (e.g., agentic vs. physical) was provided by the authors. It is thus unclear whether this improvement is systematic, or similarly displays the sort of fragility observed by Conwell & Ullman (2022).

Leivada et al. (2023) tested D2 on a range of tests related to grammatical compositionality. The ones of particular relevance to the current discussion are failures in Word Order & Thematic Role (e.g., like in Fig. 1, not distinguishing between "the dog is chasing the man" and "the man is chasing the dog") and coordination (e.g., "a man is drinking water and a woman is drinking orange juice" showed both drinking one or the other). Similarly, Rassin et al. (2022) demonstrated that DALL-E 2 regularly violates what they refer to as "resource sensitivity," or the constraint that each symbol is given a different role. Though a symbol may be ambiguous ("bat" the animal vs. the instrument), when it is used in a sentence, it cannot denote various entities at once (e.g., to refer to both an animal and an instrument in the environment). An interesting example demonstrating the "leakage" of one token's set of relationships to another is what the authors refer to as "second-order stimuli." For example, their prompt of "a bird at a construction site" yielded a normal bird and no (construction) crane, but "a tall, long-legged bird at a construction site" did, along with its homophonous bird (crane). In this case, presumably, "tall, long-legged" activated $\text{CRANE}_{\text{bird}}$ while "construction site" biased it towards $\text{CRANE}_{\text{construction,}}$ so both of *crane*'s senses become involved in the embedding. This kind of error is harder to explain solely due to "not getting syntax" because the context (construction site) supports further inferences. Regardless, even if these issues are only due to not representing the syntactic structure of the sentences (or knowing anything about English syntax), the systematicity of words is deeply related to syntax, so that is to be expected.

These issues are likely to be true for other modalities, too, as well as future image-based systems, assuming they use similar approaches. As for text models, adding more training data and more parameters will make it harder to tell what they struggle with because it increases its coverage. However, it is little consolation to the cognitive psychologist that adding more and more of the world into the training set makes it harder to notice the limitations of these models. The question is whether that is how we do it. The fact that such models struggle with compositionality would be exciting if that were not already expected. With questions about the nature of concepts in word learning reviewed, we can now turn to questions about the word learning mechanisms themselves - that is, how are meaning hypotheses tested and updated across experiences?

Too Global

Most acquisitionists agree that to learn words children must be able to track them across exposures and use information from different experiences to motivate a meaning hypothesis (Yu & Smith, 2007; Fazly et al., 2010; Siskind, 1996; Trueswell et al., 2013; Stevens et al., 2017). This is largely uncontroversial because 1) we often use words in the absence of a referent, and 2) natural language, as well as experiences involving it, are rife with ambiguity (Quine, 1960; Medina et al., 2011). Much of the research in this area concerns itself with ostensive labeling (Gleitman & Trueswell, 2020; Wojcik et al., 2022) and thus involves hearing nonsense words (“dax”) paired with images or video of possible referents (Yu & Smith, 2007; Trueswell et al., 2013; Woodard et al., 2016). Text-only LLMs do not have access to referential information, being limited only to text, so it may seem like anything developed within a reference-based paradigm is wholly irrelevant, but multi-modal models are more common, like GPT-4 (Achiam et al., 2023) and Gemini (Team et al., 2023) as previously mentioned. Additionally, a critical debate in the area of early CSWL concerns how much information is stored between word-learning contexts and brought to bear on new exposure, which bears relevance to text-only and multi-modal models. In this regard, the token modeling process of LLMs resembles “global” algorithms, and of the potential issues with this class of approaches, LLMs 1) solve none and 2) run into the same issues (Stevens et al., 2017; Yang, 2020).

Global and Local Learners

A popular class of cross-situational word learning theories depends on scaling to solve the problems of ambiguity and absence. These global models propose that learners aggregate possible referents across situations for a particular word as well as across the lexicon generally (Yu, 2008). As such, global models make use of all previous word experiences, as Yu (2008) puts it, to “maximize the likelihood function of observing the whole data set.” On the one hand, global approaches rely on the very reasonable assumption that one can learn more from more information. However, there are two issues with this class of models: they do not explain trial-by-trial behavior in word learning experiments and (relatedly) do not account for the sort of insight learning evident from experiments on “fast mapping” (Carey & Bartlett, 1978). Storing more information and conducting more computations is more costly, and there is little evidence that young children remember much from a word learning context beyond their best guess (Trueswell et al., 2013; Woodard et al., 2016). If all information from previous experiences is stored, then in the absence of the "best guess" (e.g., CAT for the word *cat*), young learners should be more likely than not to pick other referents that tended to be present when a label was uttered. Experimental evidence with kids suggests this is not what happens; instead, they revert to chance as though they had no memory of the other referents that were present during labeling. Not only that but making incorrect guesses that are similar to the correct guess does not seem to increase accuracy (LaTourrette et al., 2022). Though much fast mapping research was done in the lab and was therefore certainly far less ambiguous than naturalistic learning contexts, children exhibit stark similar insight learning patterns in referentially ambiguous contexts (Woodard et al., 2016). They do not gradually approach understanding a word's meaning; instead, it seems more like they are guessing until they get more evidence for a guess, resulting in an "Eureka!" moment (Woodard et al., 2016; Medina et al., 2011). Findings along these lines have been used to argue for an alternative, more local approach to cross-situational word learning.

Local word-learning algorithms assume that learners resolve ambiguity as they encounter it and store only their best guess. Such algorithms do not rely on scaling, and, in fact, at one extreme, such a model may only have memory of 1 hypothesis (Trueswell et al., 2013). This 1-hypothesis model posits that upon hearing a new label ("div"), the young a learner proposes one hypothesis (e.g., a bottle) based on a host of ambiguity-resolution mechanisms and stores the label alongside their guess. Upon encountering the same novel label, they retrieve their hypothesis and check if it is the best guess in this context as well. If it is, they have learned the word. If not, then they propose a new one and continue the process until they successfully confirm a hypothesis for that word. An unfortunate consequence of such a drastically limited memory is that a learner could get stuck in a vicious circle of bat (animals) and bat (baseball instrument) and never learn that *bat* can mean either (Stevens et al., 2017). Later models in this vein have increased the memory to allow for multiple hypotheses to be tracked while retaining the stipulation that only one hypothesis is made per exposure (Soh & Yang, 2021; Yue et al., 2023). With the addition of reinforcement learning, homophones can be accounted for (Stevens et al., 2017).

Nonreferential, Yet Global

As mentioned at the start, LLMs do not track referents (though, see **More modalities may run into similar issues** for discussion). A host of arguments as to why they do not understand language center on it not being able to refer to things in the world (Bender & Koller, 2020; Pavlick, 2023). But, LLMs do make hypotheses about the relationship(s) between the tokens, which is all that is required for the present analogy to hold. The "meaning hypothesis" for LLMs is that tokens are related in the way that the present parameters assume them to be. With each new experience (a new string), they update a host of parameters to shift this hypothesis to one that can better accommodate the most recent bit of evidence. The only wrinkle is that in contemporary LLMs, there are various systems of associations being learned (e.g., masked transformer, multi-layer perceptron). However, none of these changes push any of these models towards more local approaches, as there is no limit placed on how much is tracked across exposures. As such, it does not meaningfully resolve any of the issues in the existing debate in the CSWL literature (e.g., that children have limited memory, difficulties accounting for insight learning). To my knowledge, no one has ever argued that the issue with global word learning models is that they cannot perform well in various tasks if given enough memory, a large amount of data, and so on, nor that such a model could not seem like it works under many circumstances. The argument has been over how to accommodate particular facts about infants (memory and amount of exposure) with experimental evidence (showing non-gradual learning patterns). If that is true, LLMs do not add any more to this debate than existing global approaches already have. But maybe these primarily scale-based approaches in early CSWL or semantics could resolve issues anyway, given enough data and parameters and fine-tuning. We will now turn to additional issues posed to interpretation that likely affect the potential usefulness of LLMs to cognitive scientists.

Practical Meta-theoretical Concerns

A common retort to any assertion that LLMs are just predicting the next word is that perhaps it is possible an LLM *can* create a world model. Indeed, a host of overgeneralizations have been made by suggesting that good performance at a task means it may be doing something human-like (Bubeck et al., 2023; though see Guest & Martin, 2022; van Rooij et al., 2023). However, we need more reason to think this sort of modeling could construct seemingly specialized modeling systems that correspond to those that we use to reason about the world. The only thing in its favor is that it could happen. And, while it could, the problem with conceivability arguments is that so could a lot of things. It could also *not* happen. Most importantly, this is a wholly unfalsifiable line of argument. No one can prove the limitations of the next model because the next model never actually arrives. Like tomorrow, the next model is forever out of reach today. Technology advances so quickly that it is certainly easy to worry that one may be proven wrong in a few months, but being proven wrong is the name of the game in science. If one formulates a hypothesis such that it can never be falsified (scaling could fix this, scaling can construct new conceptual abilities, world models, etc.), then it is difficult to have a productive theoretical argument. There is little support for this line of argument other than arguing from uncertainty and previous error. As Fodor (1999) put it, “If the best you can say for your research strategy is ‘you can never tell, it might pan out,’ you probably ought to have your research strategy looked at.” We will now consider how the opacity of these models practically limits their usefulness in research and presents further challenges to interpretation.

Doubly Opaque

LLMs are doubly opaque. As mentioned, it is not clear what LLMs learn without rigorous testing (Lake & Murphy, 2022; Dziri et al., 2024; Guest & Martin, 2022), but from a few such tests, it does appear they are more data-dependent than advertised. Unfortunately, the fact that most LLMs are products adds another layer of opacity, as aspects of the training set and even model and pre-training become “proprietary” and kept private due to the competitiveness of the landscape and the safety implications of LLMs (OpenAI, 2023). We will first discuss how being a product adds an extra layer of opacity before touching on how their black box nature complicates interpretation to begin with. It is important to note that the additional layer of corporate opacity is not just an isolated incident involving GPT-4. Some of the other big names in LLMs, Bard (running on Palm 2 (Anil et al., 2023); though also true of some earlier models, e.g., Thoppilan, 2022) and LLaMMa 2 (Touvron et al., 2023), have followed suit.[12] Given the work cited in sections above, making it harder to access the data it is LLMs are dependent on is a practical issue researchers studying LLM performance have to face. As a consequence, researchers are often forced to rely on indirect methods or assumptions about what is in the training data (e.g., GPT-3 was pre-trained on text up to 2021). Even so, this discussion is all the more complicated by the fact that as subscription and usage-based products, these LLMs are additionally updated to ensure better service. For example, Bard was recently updated with “implicit code execution” so it can develop code to respond to prompts (e.g., about math, see Krawczyk & Subramanya, 2023). As an opt-in feature, chatGPT optionally offers plugins that make up for its issues in reasoning and mathematics, like Code Interpreter, which can implement Python code to respond to a prompt (Lu, 2023) and a Wolfram API for e.g., solving equations (Wolfram, 2023). So, when we ask Bard to do something, we do not know whether it is responding by virtue of its ‘pure’ LLMs or by virtue of additional API calls, and the same is possibly true more generally if any details concerning the architecture are kept classified due to the competitiveness of the landscape. We also know some models like chatGPT are updated, e.g., with "improved factuality and mathematical capabilities” (Natalie, 2023). These updates may be based on end-user interactions with chatGPT (Schade, 2023), or they may be motivated by analysis of interaction data (OpenAI, 2023b). As such, it is also possible that, with “glitches” going viral (like how many *n*’s are in *mayonnaise*), the model may receive more data from users about the topic, or the models could even be fine-tuned in response to these issues. In either case, the users (often academics) are effectively doing quality control for multi-billion dollar companies by continuously probing these models for glitches or errors.

These issues discussed above relate to a more general problem: how should an LLM’s

[12] The case of LLaMMa is especially odd considering Meta is attempting to position it as “open” (Touvron et al., 2023; for issues surrounding openness see Liesenfeld et al, 2023).

failures be interpreted? That is exactly my concern with using these models as baselines or comparisons for human participants: How do we interpret failure? Could it be failing for one of the reasons mentioned in the previous paragraphs? Not enough data/parameters? Or is it for more fundamental reasons? Sadly, the problem does not disappear when approached from the other direction: how do we interpret success? The data dependence of these models complicates attempts to falsify it. For any failure to match human performance, one can always claim it is because it was not trained on enough data or the right sort (multimodal, speech rather than text, etc.). Or, even if the approach is correct, the particular instantiation of it may not be. This is because a consequence of code (as opposed to theory) is that one must make various commitments that may be fundamentally unrelated to the theory in question. The precise mechanism of tokenization (or segmentation) may not be relevant to understanding word learning, but it can affect performance on a range of tasks (Rust et al., 2021). Of course, segmentation and word learning must interface, and of course, research in either can benefit from considering the other. But the current approach suggests either starting from the bottom and handling these earlier stages first or committing to not just a specific theory (e.g., of word learning or segmentation) but to a set of theories about other processes (which may themselves be contested) involved in completing a general task. The result is that cognitive theories that could be falsified in principle by any LLM are at perhaps too fine a grain to inform psychological theory development. This is perhaps a broader problem of code-as-theory approaches, but it is especially salient given the complexity of LLMs.

There are some things LLMs need to do that may be separable from others in some learning contexts, like using Byte-Pair Encoding or how a model determines relationships between tokens in a string. It is hard to decide on which component to credit with success or failure in a task. In the case of an agreed-upon failure, what is falsified is too specific. Anyone who has played 20 Questions can immediately recognize the issue with this approach, and as Allen Newell (1973) famously stated, “You can’t play 20 questions with nature and win.” Unlike 20 Questions, however, even in the case of an agreed-upon success, much more experimentation is required for any of the big questions. Going from “If the model does what people do, then the model correlates with human behavior and/or neuroimaging data” and “The model correlates with human behavior” being true to “Therefore it does what people do” requires affirming the consequent, which is not a valid chain of inference (Guest & Martin, 2022). In a sense, we are then back in the same situation we are already in with people – minus the ability to introspect. For example, if one considers LLMs (or some distillation/summary of them) a grammar, it is, at best, a descriptively adequate theory for the dataset, but the goal of linguistics is to reach an explanatorily adequate theory (Dresher & Hornstein, 1976). Indeed, recent work has even argued that creating human-like AI is computationally intractable and provided a formal proof to that end (van Rooij et al., 2023). It is unclear how an LLM could explain why the language it describes is the one it ends up with; it just ends up with it. Finding another black box does not feel like much cause for celebration. This is a different scenario from better-understood models, say n-grams or even Bayesian approaches. Instead, the effectiveness of transformers is still something that is being worked out by computer scientists, like a lot of deep learning currently. This concludes the section arguing that these models 1) do not move the needle on existing debates about meaning and 2) are difficult to interpret for a host of reasons. Because of this, getting any insight from it is a high-risk gamble. We will now consider the cost of making this gamble.

Too Costly

The "costs" to such a gamble are ethical/moral in addition to literal. My argument will not be that LLMs are wrong in the abstract but in the particular. As academics, however, our focus is on the abstract, which can result in particular costs of doing business being elided and normalized. In essence, these costs run the genuine risk of being forgotten as costs. This is doubly true, given how invisible the infrastructure that supports current models is. This abstraction is a crucial feature of exploitation, but exploitation is not the only concern as we will see. For the sake of space, the arguments listed are not intended to be comprehensive (though see Weidinger et al., 2021, for a more thorough review). The focus here will be on 1) privacy concerns, 2) labor concerns, and 3) climate concerns. What unites these is the sheer data hunger of these models. Paired with the previous arguments, I feel they suggest the best course of action is to exhibit caution in using these models and to be willing to justify their use on a case-by-case basis rather than as a broad programmatic change in how we do research. At the very least, the data hunger suggests the importance of developing algorithms for machine translation, among other things, that do not require the construction of more and more "dark Satanic mills" (Blake, 1808) with massive cooling bills in an age of climate, labor, and privacy anxiety such as ours. Before we discuss those issues, we will briefly consider whether it is possible (in the particular, not the abstract) to construct an LLM (rather than an RSLM) for our purposes that can avoid these ethical costs.

The scale of processing power and the amount of data necessary complicate the development of LLMs within an academic context. Given the amount of data used by current models (GPT-3 had 499 billion tokens, approximately 374 billion words (Brown et al., 2022); LLaMMa 2 had 2 trillion tokens, ~1.75 trillion words), constructing a dataset of similar size would be especially costly if it had to be audited for identifiable information, copyrighted text, or hate speech. Multi-modal datasets introduce even more problems surrounding informed consent (Prabhu & Birhane, 2020; Birhane et al., 2023). Given the present focus on language acquisition, if developing a massive corpus of child-directed speech is a priority, then that introduces further obstacles: greater scrutiny under IRB due to collecting data from vulnerable populations since there is likely very limited child-directed text available online (unless transcribed from audio/video), time taken to transcribe and annotate, and the typical complications of developmental work (recruiting parents, scheduling, child comfort/fussiness). For context, adding together all the words in CHILDES' English, North America corpora (MacWhinney, 2000) put together have 13 million non-child words and 2.5 million child words (calculated in summer 2022). The oldest corpus dates to 1973 (Brown, 1973), which means that since then, roughly 260,000 child-directed words a year have been added to CHILDES(certainly not uniformly, of course). At that rate, it would take a thousand years to get about enough data for a child-directed speech corpus for GPT-3. And, of course, the bottleneck is not just technological. Ensuring a diverse dataset requires that parents from a range of communities feel comfortable trusting scientists into their homes and with their child's data, so this approach risks either further erasing the linguistic experiences of marginalized groups or encouraging thinking of such groups extractively. Given the variability of environments, flexibility will be required on the part of recorders, transcribers, and annotators such that automated approaches may not help. Though such products and services will likely not be usable regardless because of unclear privacy and use policies since the data will be of vulnerable populations in their home. So, it will also be costly, especially if we ensure that the recorders, transcribers, and annotators are paid fairly for their time. These are the practical issues surrounding the construction of a more ethical LLM for academic purposes. This is not to say these issues are insurmountable, nor in any way meant to discourage the construction of high-quality datasets encompassing a diverse range of speakers, languages, and communities. But merely to highlight that it is critical the field does not engage in "plug and chug" thinking and attempt to match the speed and scale of current LLMs dataset construction, lest we risk merely changing who is doing the exploitation and extraction rather than creating a more beneficent solution. But, regardless of whether an academic LLM is likely to be developed, currently, LLM research[13] has consisted primarily of probing models developed and often hosted by large corporations. The present critiques therefore hold only until an alternative is developed that resolves these issues.

For example, one promising area of research in developmental psycholinguistics involves training statistical models on more "human-sized data." Though these would not necessarily qualify as LLMs given the significantly more modest size of the datasets they are trained on, RSLM may be more apt, as noted in the introduction. RSLMs are certainly a welcome direction as they stand to minimize data hunger, which can exacerbate or cause many of the issues that will be discussed. For example, the recent BabyLM challenge included multiple tracks with different limitations on training data, with the Strict-small track limited to a ten million-word corpus and the Strict track to a 100 million-word corpus. Similarly, an earlier RSLM, BabyBERTA, made modifications to RoBERTA (Liu et al., 2021) and limited the training dataset to only 5 million words (Huebner et al., 2021). Additionally, Vong et al. (2023) recently made waves for training a CLIP-based model on paired audio-video data from a

[13] And **not** RSLM or general LM research.

corpus including transcribed text (37,500 utterances) and video (600,000 frames) from a single child (6-25 months, 61 hours of recording). RSLMs are beyond the scope of the present paper, which attempts to focus on clear-cut cases of LLMs, though naturally, some of the potential issues noted for LLMs may be relevant to RSLMs. A proper survey of RSLMs would be able to go into far greater detail for each model, as RSLM papers provide much more information about the model and training data. One problem that uncontroversially remains for both LLMs and RSLMs, however, is how exactly success is determined as discussed in **Double Opaque**. Currently, the benchmark approach is commonly employed to measure success, but such an approach is entirely dependent on the quality of available benchmarks. If a benchmark were to contain confounds that a much more limited model could take advantage of, then this might suggest that generalizing from success on such benchmarks is limited. Indeed, Martínez et al. (2023) found success on BLiMP (Warstadt et al., 2022; used in BabyLM (Warstadt et al., 2023)) and Zorro (used to test BabyBERTa; Huebner et al., 2021) using a 5-gram model. The authors of this paper suggest the LI-Adger (Sprouse & Almeida, 2012) dataset as a better benchmark with fewer linear confounds, but it is important to keep in mind that as theories develop, we may need to critique and develop benchmarks to accommodate new confounds we may discover. This comment is certainly not intended to discourage continued attempts to do more with less, nor is the present paper aiming its critiques squarely at such approaches, but it is worth keeping in mind that these practical limitations (i.e., there is no uncontroversial benchmark) will likely remain. This paper does not seek to critique such approaches outright, as they are capable of reducing the data-hunger of LLMs, which is likely a central cause of many of the risks with the development of LLMs that will be discussed in the next section.

Privacy

One of the primary benefits of transformers is parallelization, which makes transformer-based architectures faster at processing the same amount of data as earlier models. This, in turn, motivates the construction of larger datasets for training, with the hope that this will lead to more increases in performance. But these larger datasets do not come from nowhere. Scraping publicly available data is a pre-existing issue, and it alone already introduces ethical issues surrounding attribution, existing bias, and consent more generally (Prabhu & Birhane, 2020). This is because the “move fast and break things” mentality does not allow for time to ask individuals whether their data could be used and instead puts the onus on others to opt out. However, LLMs’ continuous demands for more data may mean that soon, even all the publicly available text on the internet will not cut it anymore (Villalobos et al., 2022). This means if scaling continues to be seen as the answer, other sources will have to be considered. This is especially concerning considering two of the major players in LLMs handle large amounts of text for their users: Meta via Facebook and Instagram and Alphabet via Gmail. Though these companies state their current models do not use their users' data (Jackson, 2023), they may reach a point where they have to to stay competitive (or may need to purchase it from others). It may sound unlikely, but some companies have already begun changing their policies. Zoom recently updated its privacy policy to state that information from its users' calls may be used to train a machine-learning model (Ivanovs, 2023), and Twitter has similarly updated its Terms of Service to suggest they can do the same despite previously allowing users to opt-out (Maruf, 2024). Setting those concerns aside, there is a fundamental issue posed by the internet that has consequences for the data gathered: it is not all nice. This means datasets can and do include graphic, and even illegal, text and imagery, which can affect training and, unchecked, reproduce existing biases (Prabhu & Birhane, 2020). Both these issues suggest a necessity for auditing or developing compensatory corrective systems, however, and this leads to the second cost: labor.

Automation and Labor

There are two labor issues: one has to do with the initial dataset, and the other has to do with the creation of further datasets. In the case of the former, privacy issues relate directly to labor and attribution issues. Academic texts are often publicly available, but like other publicly available texts, this comes with certain conditions – primarily that the article will be credited (typically through citation). Image-generating transformers highlight this issue in a more straightforward manner, as artists who had been putting their art online did so under the expectation that their art will not be used for commercial purposes (e.g., a logo for your lab). However, these image transformers are 1) used in many ways by end-users who may want to monetize the outputs of their prompts, and 2) primarily effective thanks to the vast amount of art produced and put online by humans and, therefore, would perform a significantly worse if they did not make use of that data. This means many artists see these models as profiting by providing a service that is built upon their work (in the aggregate) as well as facilitating and even obfuscating plagiarism. Importantly, obfuscating plagiarism becomes an even bigger issue when generative AI is marketed as a replacement for artists and graphic designers. In such cases, artists can often worry their work is being stolen to train their replacement.

The second set of issues falls under the umbrella of "data enrichment" labor (Partnership on Open AI, 2023). This refers to labor intended to improve the performance of these models by annotating or creating new data and often takes the form of annotating data for potential harms or explicating tasks (like coding) in English. In both cases, US companies run the risk of contributing to ongoing "algorithmic colonization" by suppressing the development of local products abroad while keeping individuals dependent on the West for these kinds of products and infrastructure (Birhane, 2020). One type of data enrichment involves paying individuals to read, watch, or look at a lot of content, much of which is likely to be highly graphic in various ways (e.g., sexually, racially, physically, and so on) to flag whether it violates any laws (e.g., hate speech) or is otherwise undesirable in a model (e.g., violent imagery). Or, in the case of OpenAI, rather than hiring individuals to do this work, it is instead off handed to a contractor (like Sama) and outsourced to Kenya, where labor is significantly cheaper (about $1.46 and $3.74 per hour). To save money, these workers were, of course, not provided support or access to any counseling services (Perrigo, 2023; Rowe, 2023). Other kinds of data enrichment labor are also subject to subcontracting. For example, some data enrichment tasks aim to improve performance of a model in particular areas (e.g., code, reasoning) and therefore requires creating datasets in which reasoning is often made explicit or otherwise described in English. OpenAI notably used such annotators in their push to provide code generation through GPT (Albergotti & Matsakas, 2023). Though these issues are, of course, exacerbated when outsourcing to contractors in the global south, this growing form of labor is likely to be subcontracted in the US as well. While pay often starts significantly higher (in the case of OpenAI, $15 per hour), no benefits are provided (Ingram, 2023), employment is often precarious, and can be, in the case of data-enrichment jobs for Bard, high-pressure and fast-paced (Chowdhury, 2023), with subcontracted employees having little to no say in their working conditions (De Vynck, 2023). Domestically and abroad, LLMs engage in and encourage bad labor practices to attain the level of scale necessary for the performance they would like to advertise. We will now turn to the final cost we will cover: the environmental costs.

Climate Concerns

LLM companies have, in some cases, decided to abide by best practices and disclose their estimates of their emissions, but it is important to note that it is difficult to compare estimates without knowing more details about how they were reached (Dodge et al., 2022; Patterson et al., 2021). LLaMMa 2 reported an estimate of 539 tCO2 consumed during training (Touvron et al., 2023), while external researchers have estimated GPT's to be 552 tCO2. Regardless, the numbers do look quite high, and that is because the data hunger naturally translates into many computations being performed over a long period. For context, the lifetime carbon footprint of a mid-size car (120,000 miles) is 63 tCO2 (Center for Sustainable Systems, 2018, Strubell, 2019). The average American drove 13,489 miles per year in 2021 (Hardesty, 2023), which means where a car may take nine years, an LLM takes less than one to emit almost nine times the carbon. To get a holistic view of the current and potential of LLMs, It is important to keep in mind that not only are there various companies developing them, but many of these companies have developed more than one. So far, this article has mentioned five different models (chatGPT, GPT-3, GPT-4, LLaMMa 2, and PaLM (Bard)), the oldest of which came out in 2020. It is important to note this estimate is just for pre-training; they do not account for continued running costs (responding to prompts) and the various updates that may occur along the way. In the case of the former, it is essential to keep in mind that these models are not simply “looking up” values in a database but, rather, are crunching statistics. Practically, this means it is difficult to determine what the carbon costs are of a single study with LLMs. However, even if running carbon emissions were transparently available, the question of whether and how to count the carbon emissions during pre-training in these studies would likely remain, especially as research employing LLMs like GPT-4 to make outsized claims about its ("cognitive") abilities may serve to boost their use and perceptions of legitimacy which may, in turn, contribute to the pre-training of future models. The running costs are especially relevant if LLMs become a part of daily life as their regular use may quickly add up. For example, Microsoft and Alphabet have announced their interest in integrating LLMs into online searches (Reid, 2023; Mehdi, 2023). In 2009, a single Google search was estimated to be 0.2g of carbon (Hözle, 2009); though there may have been gains in computational efficiency since then, adding LLMs to the process may jeopardize these gains. Since this paper was submitted, sustainability reports have revealed that Alphabet's carbon emissions have gone up by 48% since 2019 (Milmo, 2024) while Microsoft's have gone up by 30% from 2020 to 2023 (Hodgson, 2024). These increases will make it significantly harder for both companies to meet their goals of reaching net-zero emissions by 2030.

It is important to note that many of these companies use carbon offsets or may otherwise use other strategies or algorithms to optimize energy efficiency (though Bard's footprint is unknown, Alphabet is known to use various methods to manage their data center's energy usage; Google, 2023). However, there are limitations to strategies that do not seek to reduce energy usage but instead to either optimize or offset continued usage. For example, it is unclear whether offsets do what they promise to do, at least in the immediate timescale. Offsetting can include paying non-profits to plant trees or distribute energy-efficient gear in the global south. While these approaches may be great, they are unlikely to offset the carbon in the short run. This is because it can take decades before a tree offsets the carbon promised by such providers (Fairs, 2021), or because the returns are not as effective as possible since energy consumption in the global north outweighs that in the south; for example, per capita carbon emissions are 40 times higher in the United States than Kenya (Energy Use Per Person, 2023). Furthermore, some argue that many of the funds that go towards carbon offsetting go towards projects that would have been carried out regardless, thereby resulting in misallocated resources (Calel et al., 2021). Of course, the biggest concern is that carbon offsetting does not reduce emissions in the first place (Forster, 2022). Given the limitations of current offsetting approaches, and the urgency of the climate crisis, reducing the use of carbon has the highest impact. Finally, it is important to note that carbon emissions are not the whole story as far as climate costs are concerned. Since data centers are constantly computing, they generate heat and therefore require cooling. This requires water, so it is also important to consider the water extracted from various ecosystems, many of them fairly dry to begin with (Sattiraju, 2020). It is difficult to estimate how much water is used to train and maintain an LLM. But, this means that a more holistic view of environmental costs includes not only the carbon offset during pre-training but also a currently-hard-to-estimate estimate offset for continued use and fine-tuning in addition to the water used for cooling, especially if LLMs continue to be integrated into everyday products like online search.

Conclusion

What can LLMs tell us about long-standing debates in word learning? My argument thus far can be summarized as follows: little more than we could gain from reading the existing literature. Some may prefer querying an LLM to running an experiment, constructing their own models, or reading philosophy, and while it is of course not necessarily impossible some LLM experiment could produce an interesting finding , such work is different from theory development. The present issue with LLMs is that it is not clear how to characterize them, given their novelty and size. My point, however, is not that LLMs must be like some particular existing theory but rather that when considering existing debates and the questions they raised, LLMs run into the same issues most theories in these spaces have run into. They have yet to resolve them despite hyperbolic claims to the contrary. At best, they sidestep what makes these questions interesting, and at worst, they ignore psychological plausibility and existing empirical findings. While NLP researchers are certainly free to decide whether or not to shape their models based on psychological principles (Lake & Murphy, 2021), we developmental psycholinguists have no such freedom.

This special issue asks whether LLMs can tell us anything. Most LLM discourse seems to take this form: what can LLMs do, and what problems could they solve? Joseph Weizenbaum, one of the however many fathers of AI at this point, said the following in an interview (ben Aaron, 1985) when asked what the role of computers in education should be:

> "The questioning should start the other way -- it should perhaps start with the question of what education is supposed to accomplish in the first place. Then perhaps [one should] state some priorities -- it should accomplish this, it should do that, it should do the other thing. Then one might ask, in terms of what it's supposed to do, what are the priorities? What are the most urgent problems? And once one has identified the urgent problems, then one can perhaps say, 'Here is a problem for which the computer seems to be well-suited.' I think that's the way it has to begin."

As far as I have seen, no one has articulated why LLMs as such (i.e., GPT-4, Gemini, etc.)[14] are uniquely well-suited to the task of conducting word learning research in

[14] This is not a critique, as stated repeatedly throughout the paper, of approaches like those in BabyLM and BabyBERTa. As a reminder, this is because these approaches immediately fail criteria 3 (trained

light of the clear problems they pose to interpretation (noted in **Doubly Opaque**), and the potential costs (noted in **Too Costly**). It is true that they could in the sense that the future is unknowable, and LLMs certainly *are* mysterious, much like the brain, and yes, they seem impressive. All of this could generate inspiration, ideas, or publications, but I have yet to see a coordinated plan that takes the interpretative challenges reviewed in **Double Opaque** seriously. The costs, in my opinion, are especially marked given the high-risk nature of the decision to integrate proprietary LLMs into the field broadly and uncritically. This is not a free lunch, and if we are not pleased with the consequences of taking this bet, we will still have to pay for it. It is a very live possibility that LLMs teaches us little about language acquisition, and that we have contributed much more to the erosion of privacy as an individual right, ongoing social and financial inequality, climate change, and even more (e.g., amplifying prejudice, misinformation, security concerns (Weidinger et al., 2021)) in the process.

The past would suggest that we refrain from playing with shiny new toys even if it seems like they can do absolutely anything. However, if you feel you must, please deliberate over it and ensure it is worth it for that particular case. Consider whether there are means of conducting the study without using LLMs (e.g., maybe a homegrown RSLM would work, or an even simpler model). Stay up to date with best practices in NLP and consider how they may apply to work in our own field (e.g., perhaps working towards using standardized model cards for RSLMs as is done for LLMs (Mitchell et al., 2019)). Considering these points may mean honestly asking oneself whether a potential paper speaks to big questions or is just provocative and easily preparable. This may require reading and determining what is under debate now as well as historically and asking whether LLMs completing some task truly tell us anything. If it fails, can it tell us more than it failed? If it appears to succeed, will we allow it to confirm our biases rather than conducting further tests and refining our benchmarks? As in conducting any study, it is critical to approach big claims carefully. Using the best work in developmental psychology may serve as a good guide – that is, ensuring other possible strategies for completing a task are not available before providing strong interpretations based on success (Martínez et al., 2023; Frank, 2023). Finally, it is critical as a field that we become open to critique over our decisions. The discipline cannot move forward if discussing questions of value, cost, and ethics is considered rude, irrelevant, or an attack. We, as cognitive scientists, must be open to more than just discussions about what LLMs can(‘t) teach us about word learning. We need frank and honest conversations on whether we should, which means being able to consider the costs listed above as well as others. Yes, this may be difficult, and yes, it may be emotional, but given the costs, those moments of personal discomfort are likely well worth sitting with. Deeply deliberating beforehand about whether to use

on an immense amount of data) in the definition given in **How LLMs M Ls.** These approaches also attempt to use fewer parameters, and so are relatively better along criteria 2 than LLMs. It is harder to determine how many parameters is “too much,” though, relative to words or speech.

such an LLM also better prepares one to receive and respond to critiques of the sort provided here. Finally, it is imperative that in pursuing any work using LLMs, cognitive scientists take care not to 1) do free quality control for major corporations and 2) launder the reputation of their products by suggesting they are human-like and therefore further contributing to the hype cycle. The former can be done by ensuring any paper has a point beyond the simple “LLMs can/’t do X.” The latter can be done by 1) ensuring hyperbolic claims are not made about LLM capacities within the scientific community or to the press (Shevlin & Halina, 2019) and 2) including some of the costs as limitations of the methods and approach. While learning from the past is an individual decision at the end of the day, it stands to benefit us all.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Albergotti, R., & Matsakis, L. (2023). OpenAI has hired an army of contractors to make basic coding obsolete | Semafor. *Semafor*. https://www.semafor.com/article/01/27/2023/openai-has-hired-an-army-of-contractors-to-make-basic-coding-obsolete

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., ... & Wu, Y. (2023). Palm 2 technical report. *arXiv preprint arXiv:2305.10403.*

Arkoudas, K. (2023). *GPT-4 Can’t Reason* (arXiv:2308.03762). arXiv. http://arxiv.org/abs/2308.03762

Armstrong, S. L., Gleitman, L. R., & Gleitman, H. (1983). What some concepts might not be. *Cognition*, 13(3), 263-308.

Auden, W.H. (1962). After Reading A Child's Guide To Modern Physics. *The New Yorker*. New York, NY.

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463

ben-Aaron, D. (1985). Weizenbaum examines computers and society. The Tech. https://web.archive.org/web/20210311142401/http://tech.mit.edu/V105/N16/weisen.16n.html

Ben-Zeev, T. (2012). When erroneous mathematical thinking is just as “correct”: The oxymoron of rational errors. In *The nature of mathematical thinking* (pp. 55-79). Routledge.

Berkeley, G. (1881). A treatise concerning the principles of human knowledge. JB Lippincott & Company.

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., ... & Ramesh, A. (2023). Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf, 2*(3), 8.

Birhane, A. (2020). Algorithmic colonization of Africa. *SCRIPTed, 17,* 389.

Birhane, A., Prabhu, V., Han, S., & Boddeti, V. N. (2023). *On Hate Scaling Laws For Data-Swamps* (arXiv:2306.13141). arXiv. http://arxiv.org/abs/2306.13141

Blake, W. (1808). *Milton.*

Blank, I. A. (2023). What are large language models supposed to model?. *Trends in Cognitive Sciences.*

Block, N. (1986). Advertisement for a Semantics for Psychology. *Midwest Studies in Philosophy, 10,* 615–678. https://doi.org/10.1111/j.1475-4975.1987.tb00558.x

Block, N. (2016). Semantics, conceptual role. In *Routledge Encyclopedia of Philosophy* (1st ed.). Routledge. https://doi.org/10.4324/9780415249126-W037-1

Brown, R. (1973). *A first language: The early stages.* Cambridge, MA: Harvard University Press.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). *Language Models are Few-Shot Learners* (arXiv:2005.14165). arXiv. http://arxiv.org/abs/2005.14165

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). *Sparks of Artificial General Intelligence: Early experiments with GPT-4* (arXiv:2303.12712). arXiv. http://arxiv.org/abs/2303.12712

Calel, R., Colmer, J., Dechezleprêtre, A., & Glachant, M. (2021). Do carbon offsets offset carbon?.

Calvo, P., & Symons, J. (Eds.). (2014). The architecture of cognition: Rethinking Fodor and Pylyshyn's systematicity challenge. MIT Press.

Carey, S. (2009). The Origin of Concepts. Oxford University Press.

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. Papers and Reports on Child Language Development, 15, 17-29.

Center for Sustainable Systems (2018). Carbon Footprint Factsheet. *University of Michigan*. Pub. No. CSS09-05.https://web.archive.org/web/20190531184229/http://css.umich.edu/sites/default/files/Carbon_Footprint_Factsheet_CSS09-05_e2018_0.pdf

Chomsky, N. (2014). *Aspects of the Theory of Syntax* (No. 11). MIT press.

Chowdhury, H. (2023, July 13). Google's ChatGPT rival is trained by workers who are under pressure to audit AI answers in as little as 3 minutes, documents show. *Business Insider*. https://www.businessinsider.com/googles-bard-ai-chatgpt-trained-under-pressure-workers-2023-7?op=1

Churchland, P. M. (1986). Some reductive strategies in cognitive neurobiology. *Mind, 95*(379), 279-309.

Connolly, A. C., Fodor, J. A., Gleitman, L. R., & Gleitman, H. (2007). Why stereotypes don't even make good defaults. *Cognition, 103*(1), 1-22.

Conwell, C., & Ullman, T. (2022). *Testing Relational Understanding in Text-Guided Image Generation* (arXiv:2208.00005). arXiv. http://arxiv.org/abs/2208.00005

De Vynck, G. (2023, June 15). They helped train Google's AI. Then they got fired after speaking out. *Washington Post*. https://www.washingtonpost.com/technology/2023/06/14/google-ai-bard-raters-chatbot-accuracy/

Dodge, J., Prewitt, T., Tachet des Combes, R., Odmark, E., Schwartz, R., Strubell, E., ... & Buchanan, W. (2022, June). Measuring the carbon intensity of ai in cloud instances. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency* (pp. 1877-1894).

Dresher, B. E., & Hornstein, N. (1976). On some supposed contributions of artificial intelligence to the scientific study of language. *Cognition, 4*(4), 321–398. https://doi.org/10.1016/0010-0277(76)90015-9

Dziri, N., Lu, X., Sclar, M., Li, X. L., Jiang, L., Lin, B. Y., ... & Choi, Y. (2024). Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems*, 36.

Energy use per person. (n.d.-c). *Our World in Data*. Retrieved August 15, 2023, from https://ourworldindata.org/grapher/per-capita-energy-use

Fairs, M. (2021). Planting trees "doesn't make any sense" in the fight against climate change due to permanence concerns, say experts. *dezeen.* https://www.dezeen.com/2021/07/05/carbon-climate-change-trees-afforestation/

Fazly, A., Alishahi, A., & Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. *Cognitive Science, 34*(6), 1017-1063.

Field, H. H. (1977). Logic, meaning, and conceptual role. *The Journal of Philosophy, 74*(7), 379-409.

Fodor, J. A. (1975). *The language of thought*. Harvard University Press.

Fodor, J. A. (1980). Special sciences, or the disunity of science as a working hypothesis. In *The language and thought series* (pp. 120-133). Harvard University Press.

Fodor, J. A. (1984). Semantics, Wisconsin style. *Synthese, 59(3),* 231-250.

Fodor, J. A. (1985). Fodor's guide to mental representation: The intelligent auntie's vade-mecum. *Mind, 94*(373), 76-100.

Fodor, J. (1999). Diary: why the brain? *London Review of books.* https://www.lrb.co.uk/the-paper/v21/n19/jerry-fodor/diary

Fodor, J. A., & LePore, E. (1992). *Holism: A shopper's guide*. Blackwell.

Fodor, J., & Lepore, E. (1996). The red herring and the pet fish: Why concepts still can't be prototypes. Cognition, 58(2), 253-270.

Fodor, J., & Lepore, E. (1999). All at Sea in Semantic Space: Churchland on Meaning Similarity. *The Journal of Philosophy,* 24.

Fodor, J., & McLaughlin, B. P. (1991). Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work (pp. 331-354). Springer Netherlands.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28*(1–2), 3–71. https://doi.org/10.1016/0010-0277(88)90031-5

Forster, P. (2022). *Here's how to fix carbon offsetting to make it effective*. World

Economic Forum. https://www.weforum.org/agenda/2022/11/fix-carbon-offsetting-environment-emissions-climate-change

Frank, M. C. (2023). Baby steps in evaluating the capacities of large language models. *Nature Reviews Psychology, 2*(8), 451-452.

Frege, G. (1892) "Über sinn und bedeutung." *Zeitschrift für Philosophie und philosophische Kritik* 100, 25-50.

Fu, Z., Lam, W., Yu, Q., So, A. M. C., Hu, S., Liu, Z., & Collier, N. (2023). Decoder-only or encoder-decoder? interpreting language model as a regularized encoder-decoder. *arXiv preprint arXiv:2304.04052*.

Geary, D. C., Frensch, P. A., & Wiley, J. G. (1993). Simple and complex mental subtraction: strategy choice and speed-of-processing differences in younger and older adults. *Psychology and aging, 8*(2), 242.

Gleitman, L. R., & Trueswell, J. C. (2020). Easy words: Reference resolution in a malevolent referent world. *Topics in cognitive science, 12*(1), 22-47.

Goodman, Nelson. (1965). The new riddle of induction. In Nelson Goodman (ed.), Fact, *Fiction, and Forecast*, 59-83. Cambridge, MA: Harvard University Press.

Goodman, N. (1972). *Seven Strictures on Similarity*. In N. Goodman (Ed.), *Problems and projects*. New York: Bobbs-Merrill.

Google. (2023). 2023 Environmental Report. *Google Sustainability*. https://sustainability.google/reports/google-2023-environmental-report/

Guest, O., & Martin, A. E. (2023). On Logical Inference over Brains, Behaviour, and Artificial Neural Networks. *Computational Brain & Behavior*, *6*(2), 213–227. https://doi.org/10.1007/s42113-022-00166-x

Harman, G. (1999). *Reasoning, meaning, and mind*. OUP Oxford.

Hodgson, C. (2024, May 15) "Microsoft's emissions jump almost 30% as it races to meet AI demand." *Financial Times.* https://www.ft.com/content/61bd45d9-2c0f-479a-8b24-605d5e72f1ab.

Hözle, U. (2009). *Powering a Google search*. Official Google Blog. https://googleblog.blogspot.com/2009/01/powering-google-search.html

Huang, K., Sun, K., Xie, E., Li, Z., & Liu, X. (2023). T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 78723-78747.

Huebner, P. A., Sulem, E., Cynthia, F., & Roth, D. (2021, November). BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th conference on computational natural language learning* (pp. 624-646).

Ingram, D. (2023, May 6). *The AI revolution is powered by these contractors making $15 an hour*. NBC News. https://www.nbcnews.com/tech/innovation/openai-chatgpt-ai-jobs-contractors-talk-shadow-workforce-powers-rcna81892

Ivanovs, A. (2023). Zoom's updated terms of service permit training AI on user content without Opt-Out. *Stack Diary*. https://stackdiary.com/zoom-terms-now-allow-training-ai-on-user-content-with-no-opt-out/

Jackson, S. (2023, March 22). Google's new Bard chatbot told an AI expert it was trained using Gmail data. The company says that's inaccurate and Bard "will make mistakes." *Business Insider*. https://www.businessinsider.com/google-denies-bard-claim-it-was-trained-using- gmail-data-2023-3?op=1

Katzir, R. (2023). Why large language models are poor theories of human linguistic cognition. A reply to Piantadosi (2023). *Manuscript. Tel Aviv University. url: https://lingbuzz. net/lingbuzz/007190*.

Kodner, J., Payne, S., & Heinz, J. (2023). *Why Linguistics Will Thrive in the 21st Century: A Reply to Piantadosi (2023)* (arXiv:2308.03228). arXiv. http://arxiv.org/abs/2308.03228

Krawczyk, J. & Subramanya, A. (2023). Bard is getting better at logic and reasoning. *Google*. https://blog.google/technology/ai/bard-improved-reasoning-google-sheets-export/

Lake, B. M., & Murphy, G. L. (2021). Word meaning in minds and machines. *Psychological Review*. http://arxiv.org/abs/2008.01766

LaTourrette, A. S., Yang, C., & Trueswell, J. (2022). When close isn't enough: Semantic similarity does not facilitate cross-situational word-learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44, No. 44).

LeFevre, J. A., Sadesky, G. S., & Bisanz, J. (1996). Selection of procedures in mental addition: Reassessing the problem size effect in adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(1), 216.

Leivada, E., Murphy, E., & Marcus, G. (2023). DALL· E 2 fails to reliably capture common syntactic processes. *Social Sciences & Humanities Open, 8*(1), 100648.

Liesenfeld, A., Lopez, A., & Dingemanse, M. (2023). Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators. *Proceedings of the 5th International Conference on Conversational User Interfaces,* 1–6. https://doi.org/10.1145/3571884.3604316

Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics, 7*(1), 195-212.

Liu, H., Ning, R., Teng, Z., Liu, J., Zhou, Q., & Zhang, Y. (2023). Evaluating the logical reasoning ability of chatgpt and gpt-4. *arXiv preprint arXiv:2304.03439*.

Liu, L. Z., Wang, Y., Kasai, J., Hajishirzi, H., & Smith, N. A. (2021). Probing across time: What does RoBERTa know and when?. *arXiv preprint arXiv:2104.07885*.

Locke, J. (1850). An essay concerning human understanding. And a treatise on the conduct of the understanding. Philadelphia: Troutman & Hayes.

Lu, Y. (2023, July 11). What to know about ChatGPT's new Code Interpreter feature. *The New York Times.* https://www.nytimes.com/2023/07/11/technology/what-to-know-chatgpt-code-interpreter.html

MacWhinney, B. (2000). *The CHILDES project: The database* (Vol. 2). Psychology Press.

Martínez, H. J. V., Heuser, A. L., Yang, C., & Kodner, J. (2023). Evaluating neural language models as cognitive models of language acquisition. *arXiv preprint arXiv:2310.20093.*

Maruf, R. (2024, Oct. 21) "X Changed Its Terms of Service to Let Its AI Train on Everyone's Posts. Now Users Are up in Arms." *CNN.* www.cnn.com/2024/10/21/tech/x-twitter-terms-of-service/index.html.

McQuillan, D. (2022). Resisting AI: an anti-fascist approach to artificial intelligence. Policy Press.

Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences, 108*(22), 9014-9019.

Mehdi, Y. (2023, May 16). *Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web - The Official Microsoft Blog*. The Official Microsoft Blog.

https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/

Miller, K., Perlmutter, M., & Keating, D. (1984). Cognitive arithmetic: comparison of operations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(1), 46.

Milmo, D. (2024, July 2) “Google’s emissions climb nearly 50% in five years due to AI energy demand.” *The Guardian.* https://www.theguardian.com/technology/article/2024/jul/02/google-ai-emissions.

Milway, D. (2023.). A Response to Piantadosi (2023).

Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., & Farajtabar, M. (2024). Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. arXiv preprint arXiv:2410.05229.

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI’s large language models. *Proceedings of the National Academy of Sciences, 120*(13), e2215907120.
Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. In Proceedings of the conference on fairness, accountability, and transparency (pp. 220-229).

Narayanan, A., & Kapoor, S. (2023). GPT-4 and professional benchmarks: the wrong answer to the wrong question. *Medium.* https://www.aisnakeoil.com/p/gpt-4-and-professional-benchmarks

Natalie. (2023). *ChatGPT — Release Notes | OpenAI Help Center.* https://help.openai.com/en/articles/6825453-chatgpt-release-notes
Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium.

OpenAI. (2023a). *GPT-4 Technical Report* (arXiv:2303.08774). arXiv. http://arxiv.org/abs/2303.08774

OpenAI. (2023b). *Privacy policy*. Retrieved August 14, 2023, from https://openai.com/policies/privacy-policy

Partnership on AI. (2023). Improving Conditions for Data Enrichment Workers. *PAI.* https://partnershiponai.org/responsible-sourcing-library/

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). *Carbon Emissions and Large Neural Network Training*.

Pavlick, E. (2023). Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 381*(2251), 20220041. https://doi.org/10.1098/rsta.2022.0041

Perrigo, B. (2023, January 18). Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic. *Time*. https://time.com/6247678/openai-chatgpt-kenya-workers/

Piantadosi, S. T., & Hill, F. (2022). *Meaning without reference in large language models* (arXiv:2208.02957). arXiv. http://arxiv.org/abs/2208.02957

Portelance, E., & Jasbi, M. (2024). The roles of neural networks in language acquisition. *Language and Linguistics Compass, 18*(6), e70001.

Prabhu, V. U., & Birhane, A. (2020). *Large image datasets: A pyrrhic win for computer vision?* (arXiv:2006.16923). arXiv. http://arxiv.org/abs/2006.16923

Quilty-Dunn, J., Porot, N., & Mandelbaum, E. (2022). The best game in town: The re-emergence of the language of thought hypothesis across the cognitive sciences. *Behavioral and Brain Sciences*, 1-55.

Quine, W. V. (1951). Main Trends in Recent Philosophy: Two Dogmas of Empiricism. *The Philosophical Review, 60*(1), 20. https://doi.org/10.2307/2181906

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog, 1*(8), 9.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). *Hierarchical Text-Conditional Image Generation with CLIP Latents* (arXiv:2204.06125). arXiv. http://arxiv.org/abs/2204.06125

Rassin, R., Ravfogel, S., & Goldberg, Y. (2022). *DALLE-2 is Seeing Double: Flaws in Word-to-Concept Mapping in Text2Image Models* (arXiv:2210.10606). arXiv. http://arxiv.org/abs/2210.10606

Rawski, J., & Baumont, L. (2023). Modern Language Models Refute Nothing.

Reid, E. (2023, May 10). Supercharging Search with generative AI. *Google*. https://blog.google/products/search/generative-ai-search/

Roembke, T. C., Simonetti, M. E., Koch, I., & Philipp, A. M. (2023). What have we learned from 15 years of research on cross-situational word learning? A focused review. *Frontiers in Psychology, 14.*

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10684-10695).

Rowe, N. (2023). 'It's destroyed me completely': Kenyan moderators decry toll of training of AI models. *The Guardian.* https://www.theguardian.com/technology/2023/aug/02/ai-chatbot- training-human-toll-content-moderator-meta-openai

Rust, P., Pfeiffer, J., Vulić, I., Ruder, S., & Gurevych, I. (2021). How Good is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3118–3135. https://doi.org/10.18653/v1/2021.acl-long.243

Sattiraju, N. (2020, April 2). The secret cost of Google's data centers: billions of gallons of water to cool servers. *Time.* https://time.com/5814276/google-data-centers-water/

Schade, M. (2023). *How your data is used to improve model performance | OpenAI Help Center.* https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance

Shevlin, H., & Halina, M. (2019). Apply rich psychological terms in AI with care. Nature Machine Intelligence, 1(4), 165-167.

Shi, F., Chen, X., Misra, K., Scales, N., Dohan, D., Chi, E. H., ... & Zhou, D. (2023, July). Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning* (pp. 31210-31227). PMLR.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition, 61*(1–2), 39–91. https://doi.org/10.1016/S0010-0277(96)00728-7

Smith, E. E., & Osherson, D. N. (1984). Conceptual combination with prototype concepts. *Cognitive science, 8*(4), 337-361.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and brain sciences, 11*(1), 1-23.

Smolensky, P. (1991). The constituent structure of connectionist mental states: A reply to Fodor and Pylyshyn. *The Southern Journal of Philosophy*, *26*(S1), 137–161. https://doi.org/10.1111/j.2041-6962.1988.tb00470.x

Soh, C., & Yang, C. (2021). Memory constraints on cross situational word learning. In Proceedings of the annual meeting of the cognitive science society (Vol. 43, No. 43).

Sprouse, J., & Almeida, D. (2012). Assessing the reliability of textbook data in syntax: Adger's Core Syntax1. *Journal of Linguistics*, *48*(3), 609-652.

Stevens, J. S., Gleitman, L. R., Trueswell, J. C., & Yang, C. (2017). The Pursuit of Word Meanings. *Cognitive Science*, *41*, 638–676. https://doi.org/10.1111/cogs.12416

Strubell, E., Ganesh, A., & McCallum, A. (2019). *Energy and Policy Considerations for Deep Learning in NLP* (arXiv:1906.02243). arXiv. http://arxiv.org/abs/1906.02243

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J. B., Yu, J., ... & Ahn, J. (2023). Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805.

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y., Lee, H., Zheng, H. S., Ghafouri, A., Menegali, M., Huang, Y., Krikun, M., Lepikhin, D., Qin, J., ... Le, Q. (2022). *LaMDA: Language Models for Dialog Applications* (arXiv:2201.08239). arXiv. http://arxiv.org/abs/2201.08239

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology*, *66*(1), 126–156. https://doi.org/10.1016/j.cogpsych.2012.10.001

van Rooij, I., Guest, O., Adolfi, F. G., De Haan, R., Kolokolova, A., & Rich, P. (2023). *Reclaiming AI as a theoretical tool for cognitive science* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/4cbuv

VanLehn, K. (1990). Mind bugs: The origins of procedural misconceptions. MIT press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (arXiv:1706.03762). arXiv.

http://arxiv.org/abs/1706.03762

Villalobos, P., Sevilla, J., Heim, L., Besiroglu, T., Hobbhahn, M., & Ho, A. (2022). *Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning* (arXiv:2211.04325). arXiv. http://arxiv.org/abs/2211.04325

von Humboldt, W. (1836). Über die Verschiedenheit des menschlichen Sprachbaues und ihren Einfluss auf die geistige Entwickelung des Menschengeschlechts. Dümmler.

Vong, W. K., & Lake, B. M. (2022). Cross-Situational Word Learning With Multimodal Neural Networks. *Cognitive science, 46*(4), e13122.

Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., ... & Cotterell, R. (2023). Findings of the BabyLM Challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S. F., & Bowman, S. R. (2020). BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics,* 8, 377-392.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., … Gabriel, I. (2021). *Ethical and social risks of harm from Language Models* (arXiv:2112.04359). arXiv. http://arxiv.org/abs/2112.04359

Wojcik, E. H., Zettersten, M., & Benitez, V. L. (2022). The map trap: Why and how word learning research should move beyond mapping. *WIREs Cognitive Science, 13*(4). https://doi.org/10.1002/wcs.1596

Wolfram, S. (2023, March 23). ChatGPT Gets Its "Wolfram Superpowers"!. *Stephen Wolfram Writings*. https://writings.stephenwolfram.com/2023/03/chatgpt-gets-its-wolfram-superpowers/

Woodard, K., Gleitman, L. R., & Trueswell, J. C. (2016). Two- and Three-Year-Olds Track a Single Meaning During Word Learning: Evidence for Propose-but-Verify. *Language Learning and Development, 12*(3), 252–261. https://doi.org/10.1080/15475441.2016.1140581

Yang, C. (2020). How to Make the Most out of Very Little. *Topics in Cognitive Science, 12*(1), 136–152. https://doi.org/10.1111/tops.12415

Yang, G. R., & Wang, X.-J. (2020). Artificial Neural Networks for Neuroscientists: A Primer. *Neuron, 107*(6), 1048–1070. https://doi.org/10.1016/j.neuron.2020.09.005

Yu, C. (2008). A Statistical Associative Account of Vocabulary Growth in Early Word Learning. *Language Learning and Development, 4*(1), 32–62. https://doi.org/10.1080/15475440701739353

Yu, C., & Smith, L. B. (2007). Rapid Word Learning Under Uncertainty via Cross-Situational Statistics. *Psychological Science, 18*(5), 414–420. https://doi.org/10.1111/j.1467-9280.2007.01915.x

Yue, C. S., LaTourrette, A. S., Yang, C., & Trueswell, J. (2023). Memory as a computational constraint in cross-situational word learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 45, No. 45).

Acknowledgements

I would like to thank the journalists who investigate how Generative AI is impacting us in the present. I would also like to thank Iffy Book for graciously hosting a range of public discussions on the ethical questions discussed in this article.

License

Models of human learning should capture the multimodal complexity and communicative goals of the natural learning environment

Jessica E. Kosie
Arizona State University

Mira L. Nencheva
Stanford University

Justin A. Jungé
Casey Lew-Williams
Princeton University

Abstract: Children do not learn language from language alone. Instead, children learn from social interactions with multidimensional communicative cues that occur dynamically across timescales. A wealth of research using in-lab experiments and brief audio recordings has made progress in explaining early cognitive and communicative development, but these approaches are limited in their ability to capture the rich diversity of children's early experience. Large language models represent a powerful approach for understanding how language can be learned from massive amounts of textual (and in some cases visual) data, but they have near-zero access to the actual, lived complexities of children's everyday input. We assert the need for more descriptive research that densely samples the natural dynamics of children's everyday communicative environments in order to grasp the long-standing mystery of how young children learn, including their language development. With the right multimodal data and a greater focus on active participation in a social environment, researchers will be able to go beyond large language models to build developmentally grounded efficient communication models that truly take into account the dimensionality of children's diverse perceptual and social environments.

Keywords: language learning; large language models; multimodal communication; natural input

Corresponding author(s): Jessica Kosie, School of Social and Behavioral Sciences, Arizona State University, Phoenix, AZ. Email: jessica.kosie@asu.edu

ORCID ID(s): Jessica Kosie: https://orcid.org/0000-0002-2390-0963; Mira Nencheva: https://orcid.org/0000-0003-4854-4608; Justin Jungé: https://orcid.org/0009-0006-8347-3146; Casey Lew-Williams: https://orcid.org/0000-0002-8781-4458

Citation: Kosie, J. E., Nencheva, M. L., Jungé, J. A., & Lew-Williams, C. (2025). Models of human learning should capture the multimodal complexity and communicative goals of the natural learning environment. *Language Development Research, 5*(1), 188–220. https://doi.org/10.34842/ldr2025-786

Introduction

With the rapid development of large language models (LLMs), many developmental researchers have begun to see their potential for furthering knowledge of how children learn language. To address the question posed in this special issue: “What can(‘t) Large Language Models (LLMs) tell us about child language acquisition?”, we highlight the ways in which LLMs differ from child language learners and how these differences impact the inferences that can be made from LLMs about how children learn language.[1] Our hope is that researchers across fields – including developmental science, computer science, linguistics, cognitive science, and artificial intelligence – will consider and address these differences as they develop LLMs and compare them to human learners.

One notable contrast between LLMs and human language learners is the amount of input required for learning. For example, Frank (2023) estimates that to “acquire language,” LLMs require 4-5 orders of magnitude more language data than human children. How children learn – given this relative dearth of input – is likely due to two key differences between these two systems: the content of the learning input and the learning goal.

Recent efforts to compare language models to natural child learning illustrate the importance of going beyond simple prediction of the next word to incorporate features of learners’ natural input and experience, finding that models that incorporate non-speech signals and inductive biases are key to linking language models to language development. For example, Vong and colleagues (2024) demonstrated that a model trained on correlated visual and linguistic data streams – naturalistic video and audio data acquired from a head-mounted camera that a child wore regularly from 6 to 25 months – was able to acquire word-referent mappings and generalize object labels to new referents. While this is an important advance in understanding how infants learn from their combined visual and auditory input, language learning is much more complex than word-object mapping alone (e.g., Wojcik et al., 2022). In another study, Lavechin and colleagues (2024) investigated perceptual attunement in infants (i.e., the process through which infants become experts at discriminating the sounds of their native language while losing this ability for sounds not in their language) by applying a prediction algorithm to clean audiobook data and ecologically valid longform recordings of children's speech input. They found that, while perceptual attunement was present in the clean data, it only emerged in the naturalistic data when the

[1] While we acknowledge that there is a large literature on computational modelling outside of LLMs, our focus here is on features of LLMs specifically and not computational modelling in general.

algorithm incorporated language learners' inductive biases (e.g., a speech preference). These results provide important insight about the role of infants' preference and expectations in influencing their ability to learn from natural input. As a result, the authors argue for the importance of model input that reflects learners' actual experience, because failing to account for features of real-world, everyday experience leads to inaccurate conclusions about the complexity of the learning problem and how human language learners succeed in the face of such a challenge.

The goal of the first widely popular LLMs was to accurately predict words and simulate human language given what was gathered from analyzing large bodies of existing text (Blank, 2023). In contrast, while learning to predict the next word is helpful for child language learning, the goal of human children is not simply to learn language. Instead, the ultimate goal of human children is to become active, integrated members of their social environment (e.g., Casillas, 2023) who can process and respond to input as it changes across multiple timescales, adapting to in-the-moment communicative demands. While learning language is in service of this goal, becoming an active member of the social environment involves much more than language alone.

Regarding the question of what LLMs can('t) tell us about child language acquisition, we argue that LLMs have limited ability to provide insight into child language acquisition until we can better account for the true complexities of children's everyday communicative input. Further, as of now, existing knowledge of the natural input to the human language learning system is incomplete. While we suggest that *large language models (LLMs)* are limited in what they can tell us about how children learn, the development and refinement of what we are calling *"efficient communication models" (*or *ECMs)* may get us closer to approximating how humans approach true, multimodal learning challenges.

What is the input to large language models? What *can* they do and what are they not designed to do?

Using prediction-based processes, LLMs are designed and trained for a wide variety of uses, including conversation and customer support, linguistic analysis (e.g., semantic & sentiment), evaluation and feedback (e.g., automated grading and comments), debugging and optimizing code, and many others (e.g., Demszky et al., 2023). To date, none of the well-known models are intended to mirror or model the specific natural language learning trajectory of human children. Criticizing LLMs for being poor models of human language learning would be a bit like criticizing helicopters for being poor models of bald eagles. Nevertheless, LLMs are a new class of entity exhibiting advanced linguistic competence, and as such, they offer both an opportunity to explore principles of language and learning (Futrell & Mahowald, 2025; Piantadosi, 2023), and a collection of computational methods and tools that could potentially be modified and rearranged in order to produce future viable models of natural human language learning (see Orhan et al., 2020; Vong et al., 2024).

For language-only LLMs (contrasted with multimodal models currently available, and discussed more below), tokens are units of meaning: individual words, or words broken into components (e.g., ambidextrous → ambi & dexterous), or phrases combined into a single unit (e.g., hit the hay → hit-the-hay). Tokens are converted to vectors in a high-dimensional space (e.g., 300 dimensions; small-to-large, dark-to-light, good-to-bad, inanimate-to-animate, etc). These dimensions are discovered from statistics of natural language; they can be non-linear and their endpoints do not necessarily correspond to human-interpretable words or familiar concepts. Positions in the high-dimensional vector space correspond to word meanings, and a sentence can be thought of as a path through the space. One goal of a language model is to take a given path through space and predict its future trajectory – to take a sentence or paragraph and predict what words will likely come next. The process of training LLMs leads them to encode the transitional probabilities between larger and larger units of meaning (strings of tokens) in order to make increasingly accurate predictions. The predictions themselves then become the prize as automatically generated text, which can be bootstrapped as input into another round of prediction, iteratively generating more and more complex and sophisticated units of meaning as conversations, essays, entire books, and more.

While the first several generations of large language models were trained only on tokenized text inputs (e.g., LlaMA2, Touvron et al., 2023), in the past couple of years (and in the time since the first draft of this article), popular “multimodal” models have been released that operate over several types of information: text, audio, images, and video (e.g., Gemini Team et al., 2025; Berkovich et al., 2025) and interface with robotics (Gemini Robotics Team et al., 2025; Koubaa, Ammar, & Boulila, 2025).

Predict-next-word is a fair (admittedly approximate) description of the goal when training language-only LLMs; newer “multimodal” models might be described as token-context-inference. Some tokens are words and others are features, objects, and events in a visual scene or video. These models operate over tokens in a substantially higher-dimension vector space inclusive of visual content – made possible by sophisticated pre-processing in machine vision, and other technical achievements. A sentence of word tokens is a trajectory through vector space and has a visual counterpart that is a trajectory through another region of this same larger vector space in a region corresponding to visual features, objects, and events. The context window is the number of tokens "actively" considered when predicting the next token. LLaMa2 released in 2023 had a context window of 4,096 tokens (Touvron et al., 2023). A version of LLaMa4 released in 2025 has a potential context window of 10 million tokens (Berkovich et al., 2025). Prediction is one form of inference, and training procedures increasingly involve more types of inference, e.g., fill-in-the-blank showing the first and last sentence with the middle sentence missing. Covering part of an image and inferring what is missing is a visual counterpart to this fill-in-the-blank structure. Starting from an image and generating a verbal description of the image (or vice versa) is also a process of inference.

An open-source, natively multimodal LLM, LLaMa4, released in the spring of 2025 (Llama Team, 2025), has specifications that can be used to illustrate the input, goals, and output of multimodal LLMs. The largest version of LLaMa4 has 2 trillion parameters (288 billion active parameters), and is trained on 40 trillion multimodal tokens – which is not a psychologically plausible amount of information to process, comprehend, and remember during the first decade of human life (it would take around 110,000 years for a human to read this much at a rate of 750 tokens per minute). Human brains have around 100 billion neurons, each with an average of 1000 connections, although this statistic hides great variability. Depending on the accounting methods, LLMs and human brains can hypothetically be described as similarly complex, or the human brain could be considered to exhibit a few orders of magnitude more or less complexity than current LLMs (e.g., for comparison to LLM parameters, should we count all neurons, only neocortical neurons, only brain areas involved in communication? Do we count individual neurons, individual synapses, or individual modifiable proteins or other molecules at each synapse?). Human children are exposed to millions of words each year, but these words are richly embedded in relevant multimodal interactions, social environments, and spatiotemporal contexts, and it is another open-ended accounting task to determine how many LLM-input tokens might correspond to a minute or year of multimodal stimuli presented to a child. As the transformer architecture is used increasingly to support multimodal models (Gemini Team et al., 2023; Jiang et al., 2025), new opportunities will arise for using ecologically valid datasets to train models that communicate.

What is the input to human learners? What are the goals?

The ultimate goal of children's communicative development, of which language is one integral part, is to become functional members of their social environments (e.g., Casillas, 2023). Next-word prediction (a primary process underlying LLMs) is an important part of communicative development, but children go beyond this by communicating about complex meanings, mental states, beliefs, and goals with others in their community. Further, unlike the learning process of LLMs, children's learning is shaped by the moment-to-moment pressure to successfully communicate with their caregivers throughout development (McMurray, 2016).

The input to young learners reflects these complex goals. Child-directed input is multimodal in a quite different sense from multimodal LLMs. Input is deeply multidimensional, incorporating a diverse set of communicative cues. Further, this multidimensional input is highly variable over time and across individuals, communities, and cultures (Bergelson, Amatuni, et al., 2019; Bergelson, Casillas, et al., 2019; Casillas et al., 2020; Holler & Levinson, 2019; Kosie & Lew-Williams, 2024a; Piazza et al., 2021; Ryskin & Fang, 2021; Schatz et al., 2022; Suarez-Rivera et al., 2022; Yu & Smith, 2012). There is no "one-size-fits-all" characterization of human input, and any model of learning (language learning included) needs to account for and/or be robust to this massive variation. Even so, findings in the field of developmental psychology often emphasize consistency rather than variability across individuals and models of human learning frequently focus on averages (e.g., the average age of acquisition for a given word; Kachergis et al., 2022). In order for LLMs to provide insight into human learning, they must account for the fact that, even in the face of this extreme variability, nearly all children around the world learn spoken or signed language. In what follows, we provide an overview of the complexity of infants' everyday experience by briefly highlighting some examples of the multidimensionality of communicative input, describing ways in which it is adapted to infants and children, and identifying sources of variation in this input.

Speech

In many cultures around the world, caregivers modify their speech during interactions with infants (e.g., Cox et al., 2022; Ferguson, 1964; Fernald et al., 1989; Hilton et al., 2022; Kuhl et al., 1997; Piazza et al., 2017; Snow & Ferguson, 1977). These modifications – frequently referred to as "motherese" or "infant-directed speech" (IDS) – include higher and more variable pitch, shorter utterances, increased repetition, and simplified vocabulary. Modifications to IDS appear to support infants' learning by increasing their attention to speech input, enhancing their discrimination of speech sounds, and helping them to segment words out of continuous speech (e.g., Cooper & Aslin, 1990; Fernald, 1985; Golinkoff et al., 2015; Graf Estes & Hurley, 2013; Ma et al., 2011; ManyBabies Consortium, 2020; Soderstrom, 2007; Trainor & Desjardins, 2002). However, the overall amount of IDS that infants encounter varies across cultures (Casillas et al., 2020; Cristia et al., 2019; Ochs & Schieffelin, 1984; Shneidman & Goldin-Meadow, 2012) and, even within a single culture, there is variation in both the amount and "quality" of IDS (Kosie & Lew-Williams, 2024a; Outters et al., 2020). Variation in infants' experience of infant-directed speech also impacts their preference for this speech register. For example, infants who experience more IDS in their everyday input show a stronger IDS preference (Outters et al., 2020). Further, caregivers tailor their use of IDS to their infants' ages and abilities. While the overall pitch of caregivers' speech (a primary feature of IDS) is high when they are interacting with younger infants, it becomes more adult-like as children get older and produce more mature vocalizations (e.g., two-word utterances; Amano et al., 2006; Cox et al., 2022). Additionally, caregivers modify their speech as children learn new words. Roy and colleagues (2009) demonstrated, using recordings of the speech directed to a single child from 9 to 24 months of age, that the mean length of utterances surrounding a word decreases until the child produces that word and begins to increase afterwards. Similarly, Schwab and colleagues (2018) showed that fathers repeat words less frequently as children's language skill increases. But caregivers modify IDS from moment to moment as well, simplifying their speech in response to infants' babbling, providing more contingent responses to more mature vocalizations, and increasing pitch when infants provide positive feedback (Elmlinger et al., 2019; Gros-Louis et al., 2006; Smith & Trainor, 2008). Thus, in addition to changes in the language (words) that infants encounter, extra-linguistic features (e.g., pitch and utterance length) vary over time as well. In sum, even the "speech" input to infants is more than speech alone, is tailored in ways that impact attention and learning, and varies across and within individual infants.

Action

As caregivers talk about objects, they frequently act on these objects as well (Karmazyn-Raz & Smith, 2022; Meyer et al., 2011; Schatz et al., 2022, 2022; Suanda et al., 2016). Like speech, infant-directed actions are modified in a variety of ways (including more enthusiasm, repetition, simplification, larger range of motion, and being performed close to the infant; Brand et al., 2002) and these modifications appear to enhance both infants' attention to actions and exploration of associated objects (Brand & Shallcross, 2008; Koterba & Iverson, 2009; Meyer et al., 2022; Williamson & Brand, 2014). Beyond enhancing attention and exploration, caregivers' use of infant-directed action has been linked to infants' language learning. Specifically, caregivers' use of object motion in synchrony with vowel sounds and words helps infants map labels to objects (e.g., Gogate & Bahrick, 1998; Matatyaho & Gogate, 2008). Additionally, in order to learn about actions and their associated labels, infants must be able to segment individual action units out of a continuously unfolding stream of activity (e.g., to learn what "waving goodbye" is, they must be able to find that particular action unit within all of the motor activity that occurs before and after the hand waving; Friend & Pace, 2011; Golinkoff & Hirsh-Pasek, 2008; Levine et al., 2019). Caregivers' modifications to infant-directed action seem to support this ability - infants more readily identify the boundaries of action segments when those actions are demonstrated using infant-directed modifications (versus demonstrations that are "adult-directed"; Kosie et al., 2022). The extent to which caregivers modify infant-directed action varies as well. For example, Fukuyama and colleagues (2015) demonstrated that, when infants had the motor skills necessary to perform an action, but were not yet actually performing the action themselves, caregivers increased the variability of their movements (a feature of infant-directed action) relative to cases in which the infant already demonstrated proficiency in the action or did not yet have the motor skill necessary to perform the task. Thus, it seems that caregivers may tailor their actions to their infants' abilities, leading to variation in action input across time and across infants.

Gesture

Gesture, too, is a common feature of everyday caregiver-infant interactions (e.g., Goldin-Meadow, Susan, 2005; Kosie & Lew-Williams, 2024a; Rowe et al., 2008; Schmidt, C. L., 1996; Vigliocco et al., 2019). Like speech and action, caregivers modify gestures when interacting with infants versus adults. Gestures directed to infants are much simpler than the gestures that occur in adult-adult interaction and primarily involve use of deictic gestures, like pointing (e.g., Iverson et al., 1999; Murphy & Messer, 1977). In interactions with infants, versus adults, gestures are more likely to be redundant with information contained in speech, reinforcing the message rather than providing new information (Iverson et al., 1999; Özçalişkan & Goldin-Meadow, 2005). This gesture-speech redundancy appears to support infants' word learning in "typically developing" children as well as those with language difficulties (Booth et al., 2008; Hollich et al., 2023; Matatyaho & Gogate, 2008; S. Vogt & Kauschke, 2017). In the longer term, caregivers' use of gesture is positively predictive of infants' gesture use which, in turn, is linked to their language development (Iverson et al., 2008; Rowe et al., 2008; Rowe & Goldin-Meadow, 2009). However, caregivers' use of gesture varies for multiple reasons. For example, caregivers modify and adapt their use of gesture as infants' object knowledge and lexical mapping abilities grow over time (e.g., using more frequent synchrony between words and object motion with younger infants; Dimitrova & Moro, 2013; Gogate et al., 2000). Both the type and frequency of caregivers' gesture use, as well as relations to infants' communicative development, also varies across cultures (e.g., Tamis-LeMonda et al., 2012; P. Vogt et al., 2020) and children growing up in more gesture-rich cultures, like Italy, develop larger and more diverse gesture repertoires (Iverson et al., 2008).

Emotion

Caregivers also frequently change their facial movements and tone of voice to convey emotion. When caregivers address infants, they use exaggerated facial displays of emotion, sometimes called "emotionese" (Brand et al., 2002; Kosie & Lew-Williams, 2024a; Wu et al., 2021), and a happy vocal tone (Fernald, 1992; Fernald et al., 1989; Kitamura & Burnham, 2003; Panneton et al., 2023; Singh et al., 2002; Trainor et al., 2000). Researchers are just beginning to characterize the kinds of emotion displays that infants observe in their natural environments. For instance, Ogren et al. (2023) found that despite researchers' overwhelming focus on canonical facial displays (like furrowing brows for anger or pouting for sadness), infants rarely see facial configurations that match these patterns in real-world settings. This highlights the importance of descriptive data-driven research on this topic in order to understand how emotional information co-occurs with speech. Presenting emotional information concurrently with other communicative cues has several benefits. First, emotional displays can enhance infants' attention and engagement. For instance, infants prefer emotionally charged vs. neutral speech (Kitamura & Burnham, 1998; Panneton et al., 2006; Singh et al., 2002), actions (Zieber et al., 2014) and faces (LaBarbera et al., 1976; Reider et al., 2022). Second, emotions provide useful context that can help children construct complex meanings (Nencheva et al., 2023; Wu et al., 2021). Although we still have a very limited understanding of how affective displays interact with other communicative cues, there is some evidence that vocal emotion may benefit aspects of children's language development, such as recognizing words embedded in a speech stream (Singh, 2008). As is the case with other cues surrounding communication, emotion displays also vary across individuals (Kosie & Lew-Williams, 2024a) and cultures (Tsai, 2017) both in quantity (e.g., the extent to which caregivers display their emotions), as well as quality (the specific emotional expressions caregivers use).

Touch

Touch is yet another modality that caregivers systematically use when communicating with infants (e.g., Anisfeld et al., 1990; Feldman et al., 2010; Ferber et al., 2008; Franco et al., 1996; Hertenstein, 2002; Jean et al., 2009; Stack & Arnold, 1998; Stack & Muir, 1990). From birth, contact with caregivers has numerous benefits for infants, including regulating infants' stress response and increasing positive affect (Feldman et al., 2002, 2010, 2014; Stack & Muir, 1992) and caregivers use different types of touch to elicit specific behaviors from their infants (e.g., Hertenstein, 2002; Jean & Stack, 2009; Stack & LePage, 1996). Caregivers also use speech and touch cues in tandem to enhance communication with infants; their use of speech and touch are frequently aligned during natural interactions with infants and, when these cues are used together, caregiver speech is more exaggerated (i.e., "infant-directed") and touches are longer (Abu-Zhaya et al., 2017). Other research demonstrates that caregivers' simultaneous use of speech and touch supports infants' learning of auditory patterns (Lew-Williams et al., 2019), speech segmentation (Seidl et al., 2015), and word mapping (Tincoff et al., 2019). However, caregivers' use of touch adapts to infants' changing behaviors and evolves over time (e.g., Ferber et al., 2008; Jean et al., 2009). The type of touch that caregivers use also varies across cultures (Franco et al., 1996; Lowe et al., 2016) and caregivers align speech and touch even more frequently with children who are deaf and hard of hearing (Abu-Zhaya et al., 2019).

Communication is multimodal

Though we have just described each of these dimensions of communication separately, they do not occur in isolation. In fact, our own recent work shows that nearly 60% of the speech that infants hear overlaps with one or more non-speech communicative cue(s) (Kosie & Lew-Williams, 2024a), and there is strong evidence that multimodality like this enhances infants' learning. A substantial body of experimental work on intersensory redundancy (Bahrick & Lickliter, 2000) has demonstrated that exposure to multimodal cues helps to direct infants' attention to relevant features of input and supports infants' discrimination of qualities including tempo, rhythm, and affect (e.g., Bahrick et al., 2002, 2004; Flom & Bahrick, 2007). These effects have been validated in descriptive, naturalistic research as well. Play bouts in which mothers simultaneously touch and talk about objects are longer than unimodal bouts and are more likely to hold infants' attention (Schatz et al., 2022; Suarez-Rivera et al., 2019; Suarez-Rivera et al., 2022). In addition to supporting infants' attention and discrimination, multimodal input assists young infants' learning of abstract rules (Frank et al., 2009) and toddler's learning of novel words (Booth et al., 2008). Specifically, Booth and colleagues (2008) found that greater redundancy among communicative cues (including speech, gaze, pointing, touch, and object manipulation) during exposure to a novel word promoted toddlers' learning of that word. Thus, the multimodality in everyday communication appears to benefit the infant learner beyond speech or language alone.

Depicting – which occurs frequently during everyday communication – involves the use of multiple cues across modalities to create a physical scene that serves to represent, or *depict,* another scene that a person intends to communicate about (Clark, 2016). For example, if someone is talking about the antics of their naughty cat Rex, they might point to an object on the table, dramatically wave their hand in a gesture indicating that an object was knocked off of the table, and make a "whooshing" sound. Together, these components generate a scene that the interlocutor can easily visualize in a way that is richer and more precise than if the producer had simply said "my cat knocked the object off of the table." In addition to evidence that multimodality supports attention and learning, it also enhances communication more broadly through mechanisms like depicting.

One potential way to conceptualize these multimodal cues is as units of information that facilitate the interpretation of the message being communicated. However, it is not clear how to conceive of the amount of information gained by each component of a multimodal event, and it is unlikely that they all contribute equally (i.e., the total information gained by a multimodal communicative event is likely not simply the sum of its parts). Somewhat analogous to video where consecutive frames often contain redundancy (Jiang et al., 2025), multimodal input can exhibit varying degrees of cross-modal correlation and unique information. This leaves open an exciting avenue for future computational work that seeks to understand how cues are combined to generate or enhance communicative meanings. Overall, multimodality is a central component of communication that supports efficiency in processing and learning and should be accounted for in any model of early learning. As multimodal AI models advance, it is possible and plausible that they will provide more insight into development than large language models alone.

Additional influences on infants' experience and processing of communicative input

Although the cues we have discussed – speech, action, gesture, emotion, and touch – underscore the extensive multidimensionality of infants' natural input, this is not an exhaustive list of the ways that humans communicate. For example, eye gaze, proximity, and response contingency are all involved in natural communicative interactions and can be modified or tailored in ways that influence learning (e.g., Brooks & Meltzoff, 2005; Goldstein & Schwade, 2008; Salo et al., 2021). The set of communicative cues in infants' everyday learning environment spans numerous modalities and varies both across and within infants.

Beyond just the cues that occur, infants' experience of communication happens within a system that is constantly changing (see Thelen & Smith, 1994 for a review). Factors including infants' internal states and features of the environment vary at multiple timescales and influence the way that communicative input is encountered and processed (Mani & Ackermann, 2018; Outters et al., 2023; Pomper & Saffran, 2019). As one example, recent evidence suggests that the presence or absence of highly salient familiar objects may influence infants' word learning. Pomper and Saffran (2019) demonstrated that infants were slower and less accurate in looking to a novel object and learning its name when it was presented alongside a highly salient familiar item.

When the familiar item was of low salience, infants readily fixated on the novel object and learned its name, suggesting that something as simple as the identity of surrounding objects shapes infants' processing of communicative input. Infants' developmental milestones influence their natural input as well. In addition to changing infants' view of the world (e.g., Kretch et al., 2014), the manner of infants' locomotion – crawling versus walking – elicits different types of verbal feedback from caregivers. Thus, infants' language input changes as they acquire a new skill in a seemingly unrelated domain (i.e., motor development; Karasik et al., 2014).

Within infants' constantly changing experience, a variety of linguistic and non-linguistic contexts provide stable and predictable cues to support early learning. While everyday activities in the home (e.g., mealtime, playtime, book sharing) are one commonly recognized type of non-linguistic context in which infant learning occurs (e.g., Kosie & Lew-Williams, 2024b; Tamis-LeMonda et al., 2019) there is no clear-cut definition for what does and does not count as "context". Emotional states, spatial locations, social and political systems, communities and neighborhoods, and cultural values and beliefs are all examples of how context arises in infants' everyday experiences (Custode & Tamis-LeMonda, 2020; Outters et al., 2023; Rowe & Weisleder, 2020; Roy et al., 2015; Wu et al., 2021). Context influences infants' experience in multiple ways: certain words are likely to occur in specific locations within the home (e.g., "bubbles" in the bathroom at bathtime or "bye" next to the front door; Custode & Tamis-LeMonda, 2020; Roy et al., 2015) and caregivers' use of multimodal cues tends to be similar from day to day within an activity context but not across different contexts (Kosie & Lew-Williams, 2024b). The consistency that arises from contexts, broadly defined, may provide a source of predictability in infants' otherwise changing environment that can be supportive of early learning (e.g., Benitez & Smith, 2012; Roy et al., 2015; Vlach & Sandhofer, 2011).

Finally, infants and caregivers co-construct the learning environment. A bursting literature now exists that characterizes infants as active learners who contribute meaningfully to their own learning (e.g., Begus et al., 2014; Elmlinger et al., 2023; Gureckis & Markant, 2012; Kuchirko et al., 2018; Slone et al., 2019; L. B. Smith et al., 2018; Zettersten & Saffran, 2021). By examining turn-taking and leader-follower dynamics across modalities, we stand to gain a deeper understanding of how caregivers and infants jointly shape the features of infants' everyday experience.

When all of these factors are taken into account, it becomes clear that it is not possible to characterize everyday input in a way that applies to all infants, or even to an individual child, as their input and processing of that input is changing from month to month, day to day, and even moment to moment. Any model of human language learning that does not take into account the complex richness of communicative experience would be deeply limited in its utility for understanding human language development. While there has been progress in diversifying the input to LLMs beyond language alone, more careful descriptive and computational work is needed to understand the varied and changing nature of input across development and how this input influences learning in the real world.

How might we conceptualize developmentally grounded *efficient communication models*?

In order to develop efficient communication models that map onto human language development, we need to learn more about the nature of young children's communicative environments. In particular, developmental scientists will need to devote time, effort, and resources to the collection of audiovisual corpora that capture children's lives. The ideal datasets will have four key features.

First, they will need to harness multimodal communicative behaviors, including speech, action, gesture, emotion, touch, and more (e.g., Kosie & Lew-Williams, 2024a). This will make it possible to explore the dynamics of eye gaze, physical proximity, body pose, and interactions with objects and events, all of which are among the many components of successful communication. The potential of this approach cannot be overstated, as the field will go far beyond industry-generated approaches that scrape textual data from the internet. As an example: Documenting how well-timed instances of words can be reinforced with gestures or emotional displays, all within the context of social routines like mealtimes, will be far more useful to the development of plausible models compared to streams of decontextualized unimodal text. Further, input that is tailored to the learner's current knowledge and abilities may scaffold learning better than input that is randomly structured over time.

Second, it will be important to follow the same children over developmental time, from birth onward (e.g., Long et al., 2024; Sullivan et al., 2021; Vong et al., 2024). This will make it possible to pinpoint how children make incremental gains in learning, with trial-and-error behaviors that are inherent in children's physical, communicative, and social lives. While scientists have carried out excellent experimental work on infant cognition and sociality, experiments inherently treat development as discontinuous. An embracing of *continuity,* spanning milliseconds and years, will be needed to create comprehensive models.

Third, rather than focusing on the child alone, or the child and one parent (as is typical in developmental research), corpora should be representative of children's rich social environment. The presence or absence of caregivers, siblings, friends, and members of the wider social network can substantially change the nature of children's communicative input and impact their language development (e.g., Bulgarelli & Bergelson, 2024; Kosie et al., 2022; Okocha et al., 2024). Further, children's language development is driven by the desire to connect with and be understood by others (Bloom, 2013). A model that reflects human-like communicative development would include such social goals and would be trained in a contingent communicative environment (with human or artificial agents). Examining the multifaceted influences of a child's social connections – as they change from moment to moment and over longer periods of time – will allow us to better approximate how children achieve the goal of becoming an active member of their social environment.

Finally, scientists will need to prioritize variability across contexts, cultures, and communities (Kline et al., 2018; Singh et al., 2023). By capturing the lives of children and families from diverse communities, we will be able to frontload the idea that there are many pathways toward outcomes that matter in context. We will be able to understand the true variation in early language learning, as opposed to attempting to create one model that learns like the average infant. This approach will yield 'large' amounts of data, but critically, these data correspond to a developmentally plausible amount of data, enabling us to learn how infant brains and bodies – situated in diverse social environments – make efficient gains in learning.

Recordings of everyday lives will be only the first step. Beyond this stage, scientists spanning many fields will need to collaborate on the development of tools that provide accurate, automated annotation of behaviors of interest (e.g., Weng et al., 2022), as comprehensive hand coding will be impossible given the volume of datasets coming to our field in the next decade or two. Although many annotation programs currently exist – spanning domains such as language, emotion, visual object perception, gestures, bodily movements, proximity, or their combination – few have achieved accuracy on par with human coders. This is because real life does not fit into the neat categories put forth by the last half-century of psychological research. For example, basic emotion categories do not map onto the real emotion experiences or displays in children's lives (Ogren et al., 2023); and speech does not arrive to the child's brain in a noise-free, single-stream, grammatically coherent way, but instead comes from a noisy kin network with constant restarts and imperfections. Further, most of these tools have been trained on adult-adult interactions and are not tuned to the specifics of infant-directed or infant-generated communicative signals. To make the challenge even harder, infants change a lot over time, and no individual tool will be able to keep up. Computer scientists will need to engage with psychologists, neuroscientists, and linguists to achieve higher accuracy with automated annotation.

These suggestions may appear contradictory to our statement that we need to develop *efficient* communication models, as including all of this information seems like it would actually make LLMs *less* efficient. However, this may be an example of how "efficiency" means different things for a human versus a machine. While it is currently a computational challenge for LLMs to simultaneously integrate multiple streams of data across modalities, this integration may require substantially less effort for humans. For example, it has been demonstrated that adults process multimodal communication (i.e., speech and gesture combined) faster than unimodal communication (i.e., speech alone; see Holler & Levinson, 2019 for a review).

To first approximation, an ECM – benchmarked to human communication learning – is one that can take the same quantity and quality of data input as a child receives over a relevant developmental window (e.g., birth to age 5) and then communicate as effectively as a (median) child of that age. With such a benchmark established, efficiency gains can be operationalized by restricting the data input to less than this quantity and achieving similar results – thus achieving and quantifying (in the hypothetical future) super-human efficiency in the acquisition of communication. However, assessing the models' communicative ability should go beyond simply predicting language and may include, for example, accomplishing more complex social goals within the context of the child's everyday environment. While instructions for actually building such a model are beyond the scope of the current paper (and of the current authors), it seems likely that more interactive training would be required, where a model would not simply receive language and multimodal input, but actually interact with humans or other machines.

With multimodal, longitudinal, densely sampled, contextually grounded, and culturally diverse datasets at our disposal, and with validated tools for automated annotation of natural behaviors, we will be positioned to take models to the next level, far beyond existing LLMs. This will herald an era of understanding how machines can be genuinely intelligent, with reciprocal implications for understanding the nature of children's early learning. GEMINI (Gemini Team et al., 2023), as just one example, has made incredible progress toward incorporating more dimensions of multimodality into their model (specifically, image, audio, video, and text). Even so, fully comprehensive datasets that capture the diversity of natural human communication will take decades to do right. In the meantime, continued incremental progress in this endeavor will generate new insights into the dynamic experiences that support children's learning as well as catalyze advances in AI.

Conclusion

To return to the question posed in this special issue: "What can('t) LLMs tell us about child language acquisition?" we suggest that LLMs do provide insights into potential mechanisms that support language learning, but substantial work remains for illuminating how children actually learn language from their natural input. For example, the success of LLMs demonstrates that large text corpora (even in the absence of multimodal and social information) contain a lot of information that enables a model to produce and respond effectively to language. The success of current LLMs additionally underscores the power of prediction as a mechanism of language learning. However, just because LLMs can learn language from their restricted textual input, it cannot be inferred that this is how infants learn language via their everyday input.

The everyday communicative environment of infants and young children is incredibly rich and varied, while the primary source of input to LLMs is textual (and sometimes visual) corpora. Focusing on only one or just a few dimensions of input (like language alone or language and objects) vastly reduces the richness of experience, and if we attempt to understand human learning from this simplistic picture of input, we only learn about what infants *can* do under restricted and unusual circumstances. If we want to know what infants actually *do* do, and avoid making inaccurate conclusions about how infants deal with the true complexity of the language learning problem (e.g., Lavechin et al., 2024), we need to understand the full complexity of the multimodal, contingent, dynamic input with which they are actively engaged and how this input supports them in becoming integrated members of their social environment. While advances in artificial intelligence – as of 2025 – are making progress in integrating across particular modalities (Gemini Team et al., 2023; Orhan et al., 2020; Vong et al., 2024), they will not be able to tell us much about how human infants and children learn until they can be immersed in real-world environments and adopt the communicative goals of young learners.

References

Abu-Zhaya, R., Kondaurova, M. V., Houston, D., & Seidl, A. (2019). Vocal and tactile input to children who are deaf or hard of hearing. *Journal of Speech, Language, and Hearing Research, 62*(7), 2372–2385. https://doi.org/10.1044/2019_JSLHR-L-18-0185

Abu-Zhaya, R., Seidl, A., & Cristia, A. (2017). Multimodal infant-directed communication: how caregivers combine tactile and linguistic cues. *Journal of Child Language, 44*(5), 1088–1116. https://doi.org/10.1017/S0305000916000416

Amano, S., Nakatani, T., & Kondo, T. (2006). Fundamental frequency of infants' and parents' utterances in longitudinal recordings. *The Journal of the Acoustical Society of America, 119*(3), 1636–1647. https://doi.org/10.1121/1.2161443

Anisfeld, E., Casper, V., Nozyce, M., & Cunningham, N. (1990). Does infant carrying promote attachment? An experimental study of the effects of increased physical contact on the development of attachment. *Child Development, 61*(5), 1617-1627. https://doi.org/10.2307/1130769

Bahrick, L. E., Flom, R., & Lickliter, R. (2002). Intersensory redundancy facilitates discrimination of tempo in 3-month-old infants. *Developmental Psychobiology, 41*(4), 352–363. https://doi.org/10.1002/dev.10049

Bahrick, L. E., & Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental Psychology*, *36*(2), 190–201. https://doi.org/10.1037/0012-1649.36.2.190

Bahrick, L. E., Lickliter, R., & Flom, R. (2004). Intersensory redundancy guides the development of selective attention, perception, and cognition in infancy. *Current Directions in Psychological Science, 13*(3), 99–102. https://doi.org/10.1111/j.0963-7214.2004.00283.

Begus, K., Gliga, T., & Southgate, V. (2014). Infants learn what they want to learn: responding to infant pointing leads to superior learning. *PLoS ONE, 9*(10), e108817. https://doi.org/10.1371/journal.pone.0108817

Benitez, V. L., & Smith, L. B. (2012). Predictable locations aid early object name learning. *Cognition, 125*(3), 339–352. https://doi.org/10.1016/j.cognition.2012.08.006
Bergelson, E., Amatuni, A., Dailey, S., Koorathota, S., & Tor, S. (2019). Day by day, hour by hour: Naturalistic language input to infants. *Developmental Science*, *22*(1), e12715. https://doi.org/10.1111/desc.12715

Bergelson, E., Casillas, M., Soderstrom, M., Seidl, A., Warlaumont, A. S., & Amatuni, A. (2019). What do North American babies hear? A large-scale cross-corpus analysis. *Developmental Science*, *22*(1). https://doi.org/10.1111/desc.12724

Bercovich, A., Levy, I., Golan, I., Dabbah, M., El-Yaniv, R., Puny, O., ... & Chung, E. (2025). Llama-nemotron: Efficient reasoning models. *arXiv preprint: https://arxiv.org/pdf/2505.00949*

Blank, I. A. (2023). What are large language models supposed to model? *Trends in Cognitive Sciences, 27*(11), 987–989. https://doi.org/10.1016/j.tics.2023.08.006

Bloom, L. (2013). Language acquisition and the power of expression. In *Language and communication* (pp. 95-113). Psychology Press.

Booth, A. E., McGregor, K. K., & Rohlfing, K. J. (2008). Socio-pragmatics and attention: contributions to gesturally guided word learning in toddlers. *Language Learning and Development, 4*(3), 179–202. https://doi.org/10.1080/15475440802143091

Brand, R. J., Baldwin, D. A., & Ashburn, L. A. (2002). Evidence for 'motionese': Modifications in mothers' infant-directed action. *Developmental Science, 5*(1), 72–83. https://doi.org/10.1111/1467-7687.00211

Brand, R. J., & Shallcross, W. L. (2008). Infants prefer motionese to adult-directed action. *Developmental Science, 11*(6), 853–861. https://doi.org/10.1111/j.1467-7687.2008.00734.x

Brooks, R., & Meltzoff, A. N. (2005). The development of gaze following and its relation to language. *Developmental Science, 8*(6), 535–543. https://doi.org/10.1111/j.1467-7687.2005.00445.x

Bulgarelli, F., & Bergelson, E. (2024). Linking acoustic variability in the infants' input to their early word production. *Developmental Science, 27*(6), e13545. https://doi.org/10.1111/desc.13545

Casillas, M. (2023). Learning language in vivo. *Child Development Perspectives, 17*(1), 10–17. https://doi.org/10.1111/cdep.12469

Casillas, M., Brown, P., & Levinson, S. C. (2020). Early language experience in a Tseltal Mayan village. *Child Development, 91*(5), 1819–1835. https://doi.org/10.1111/cdev.13349

Clark, H. H. (2016). Depicting as a method of communication. *Psychological Review, 123*(3), 324–347. https://doi.org/10.1037/rev0000026

Cooper, R., & Aslin, R. (1990). Preference for infant-directed speech in the first month after birth. *Child Development, 61*(5), 1584–1595. https://doi.org/10.2307/1130766

Cox, C., Bergmann, C., Fowler, E., Keren-Portnoy, T., Roepstorff, A., Bryant, G., & Fusaroli, R. (2022). A systematic review and Bayesian meta-analysis of the acoustic features of infant-directed speech. *Nature Human Behaviour, 7*(1), 114–133. https://doi.org/10.1038/s41562-022-01452-1

Cristia, A., Dupoux, E., Gurven, M., & Stieglitz, J. (2019). Child-directed speech is infrequent in a forager-farmer population: A time allocation study. *Child Development, 90*(3), 759–773. https://doi.org/10.1111/cdev.12974

Custode, S. A., & Tamis-LeMonda, C. (2020). Cracking the code: Social and contextual cues to language input in the home environment. *Infancy, 25*(6), 809–826. https://doi.org/10.1111/infa.12361

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., JonesMitchell, N., Ong, D. C., Dweck, C. S., Gross, J. J., & Pennebaker, J. W. (2023). Using Large Language Models in Psychology. *Nature Reviews Psychology, 2*(11), 688–701. https://doi.org/10.1038/s44159-023-00241-5

Dimitrova, N., & Moro, C. (2013). Common ground on object use associates with caregivers' Gesturese. *Infant Behavior and Development, 36*(4), 618–626. https://doi.org/10.1016/j.infbeh.2013.06.006

Elmlinger, S. L., Goldstein, M. H., & Casillas, M. (2023). Immature vocalizations simplify the speech of Tseltal Mayan and U.S. caregivers. *Topics in Cognitive Science, 15*(2), 315–328. https://doi.org/10.1111/tops.12632

Elmlinger, S. L., Schwade, J. A., & Goldstein, M. H. (2019). The ecology of prelinguistic vocal learning: Parents simplify the structure of their speech in response to babbling. *Journal of Child Language, 46*(5), 998–1011. https://doi.org/10.1017/S0305000919000291

Feldman, R., Eidelman, A. I., Sirota, L., & Weller, A. (2002). Comparison of skin-to-skin (kangaroo) and traditional care: Parenting outcomes and preterm infant development. *Pediatrics, 110*(1), 16–26. https://doi.org/10.1542/peds.110.1.16

Feldman, R., Rosenthal, Z., & Eidelman, A. I. (2014). Maternal-preterm skin-to-skin contact enhances child physiologic organization and cognitive control across the first 10 years of life. *Biological Psychiatry, 75*(1), 56–64.

https://doi.org/10.1016/j.biopsych.2013.08.012

Feldman, R., Singer, M., & Zagoory, O. (2010). Touch attenuates infants' physiological reactivity to stress. *Developmental Science, 13*(2), 271–278. https://doi.org/10.1111/j.1467-7687.2009.00890.x

Ferber, S. G., Feldman, R., & Makhoul, I. R. (2008). The development of maternal touch across the first year of life. *Early Human Development, 84*(6), 363–370. https://doi.org/10.1016/j.earlhumdev.2007.09.019

Ferguson, C. A. (1964). Baby talk in six languages. *American Anthropologist, 66*(6_PART2), 103–114. https://doi.org/10.1525/aa.1964.66.suppl_3.02a00060

Fernald, A. (1985). Four-month-old infants prefer to listen to motherese. *Infant Behavior and Development, 8*(2), 181–195. https://doi.org/10.1016/S0163-6383(85)80005-9

Fernald, A. (1992). Meaningful melodies in mothers' speech to infants. In *Nonverbal vocal communication: Comparative and developmental approaches* (pp. 262–282). Cambridge University Press.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language, 16*(3), 477–501. https://doi.org/10.1017/S0305000900010679

Flom, R., & Bahrick, L. E. (2007). The development of infant discrimination of affect in multimodal and unimodal stimulation: The role of intersensory redundancy. *Developmental Psychology, 43*(1), 238–252. https://doi.org/10.1037/0012-1649.43.1.238

Franco, F., Fogel, A., Messinger, D. S., & Frazier, C. A. (1996). Cultural differences in physical contact between Hispanic and Anglo mother–infant dyads living in the United States. *Early Development and Parenting, 5*(3), 119–127. https://doi.org/10.1002/(SICI)1099-0917(199609)5:3<119::AID-EDP123>3.0.CO;2-Y

Frank, M. C. (2023). Bridging the data gap between children and Large Language Models. *Trends in Cognitive Sciences, 27*(11), 990–992. https://doi.org/10.1016/j.tics.2023.08.007

Frank, M. C., Slemmer, J. A., Marcus, G. F., & Johnson, S. P. (2009). Information from multiple modalities helps 5-month-olds learn abstract rules. *Developmental*

Science, 12(4), 504–509. https://doi.org/10.1111/j.1467-7687.2008.00794.x

Friend, M., & Pace, A. (2011). Beyond event segmentation: Spatial- and social-cognitive processes in verb-to-action mapping. *Developmental Psychology, 47*(3), 867–876. https://doi.org/10.1037/a0021107

Fukuyama, H., Qin, S., Kanakogi, Y., Nagai, Y., Asada, M., & Myowa-Yamakoshi, M. (2015). Infant's action skill dynamically modulates parental action demonstration in the dyadic interaction. *Developmental Science, 18*(6), 1006–1013. https://doi.org/10.1111/desc.12270

Futrell, R., & Mahowald, K. (2025). How linguistics learned to stop worrying and love the language models. *arXiv preprint: https://arxiv.org/abs/2501.17047*

Gemini Team, Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., Silver, D., Petrov, S., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., … Vinyals, O. (2023). *Gemini: A family of highly capable multimodal models* (arXiv:2312.11805). *arXiv preprint: http://arxiv.org/abs/2312.11805*

Gemini Team, Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., … & Shan, Z. (2025). Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint: https://arxiv.org/abs/2507.06261*

Gemini Robotics Team, Abeyruwan, S., Ainslie, J., Alayrac, J. B., Arenas, M. G., Armstrong, T., … & Zhou, Y. (2025). Gemini robotics: Bringing AI into the physical world. *arXiv preprint: https://arxiv.org/abs/2503.20020*

Gogate, L. J., & Bahrick, L. E. (1998). Intersensory redundancy facilitates learning of arbitrary relations between vowel sounds and objects in seven-month-old infants. *Journal of Experimental Child Psychology, 69*(2), 133–149. https://doi.org/10.1006/jecp.1998.2438

Gogate, L. J., Bahrick, L. E., & Watson, J. D. (2000). A study of multimodal motherese: The role of temporal synchrony between verbal labels and gestures. *Child Development, 71*(4), 878–894. https://doi.org/10.1111/1467-8624.00197

Goldin-Meadow, Susan. (2005). *Hearing gesture: How our hands help us think*. Harvard University Press. https://doi.org/10.2307/j.ctv1w9m9ds

Goldstein, M. H., & Schwade, J. A. (2008). Social feedback to infants' babbling facilitates rapid phonological learning. *Psychological Science, 19*(5), 515–523. https://doi.org/10.1111/j.1467-9280.2008.02117.x

Golinkoff, R. M., Can, D. D., Soderstrom, M., & Hirsh-Pasek, K. (2015). (Baby)talk to me: The social context of infant-directed speech and its effects on early language acquisition. *Current Directions in Psychological Science, 24*(5), 339–344. https://doi.org/10.1177/0963721415595345

Golinkoff, R. M., & Hirsh-Pasek, K. (2008). How toddlers begin to learn verbs. *Trends in Cognitive Sciences, 12*(10), 397–403. https://doi.org/10.1016/j.tics.2008.07.003

Graf Estes, K., & Hurley, K. (2013). Infant-directed prosody helps infants map sounds to meanings. *Infancy, 18*(5), 797–824. https://doi.org/10.1111/infa.12006

Gros-Louis, J., West, M. J., Goldstein, M. H., & King, A. P. (2006). Mothers provide differential feedback to infants' prelinguistic sounds. *International Journal of Behavioral Development, 30*(6), 509–516. https://doi.org/10.1177/0165025406071914

Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science, 7*(5), 464–481. https://doi.org/10.1177/1745691612454304

Hertenstein, M. J. (2002). Touch: Its communicative functions in infancy. *Human Development, 45*(2), 70–94. https://doi.org/10.1159/000048154

Hilton, C. B., Moser, C. J., Bertolo, M., Lee-Rubin, H., Amir, D., Bainbridge, C. M., Simson, J., Knox, D., Glowacki, L., Alemu, E., Galbarczyk, A., Jasienska, G., Ross, C. T., Neff, M. B., Martin, A., Cirelli, L. K., Trehub, S. E., Song, J., Kim, M., … Mehr, S. A. (2022). Acoustic regularities in infant-directed speech and song across cultures. *Nature Human Behaviour, 6*(11), 1545–1556. https://doi.org/10.1038/s41562-022-01410-x

Holler, J., & Levinson, S. C. (2019). Multimodal language processing in human communication. *Trends in Cognitive Sciences, 23*(8), 639–652. https://doi.org/10.1016/j.tics.2019.05.006

Hollich, G. J., Hirsh-Pasek, K., & Golinkoff, R. M. (2023). *Breaking the language barrier: An emergentist coalition model for the origins of word learning.*

Iverson, J. M., Capirci, O., Longobardi, E., & Caselli, M. C. (1999). Gesturing in mother-child interactions. *Cognitive Development, 14*, 57–75. https://doi.org/10.1016/S0885-2014(99)80018-5

Iverson, J. M., Capirci, O., Volterra, V., & Goldin-Meadow, S. (2008). Learning to talk in a gesture-rich world: Early communication in Italian vs. American children. *First Language, 28*(2), 164–181. https://doi.org/10.1177/0142723707087736

Jean, A. D. L., & Stack, D. M. (2009). Functions of maternal touch and infants' affect during face-to-face interactions: New directions for the still-face. *Infant Behavior and Development, 32*(1), 123–128. https://doi.org/10.1016/j.infbeh.2008.09.008

Jean, A. D. L., Stack, D. M., & Fogel, A. (2009). A longitudinal investigation of maternal touching across the first 6 months of life: Age and context effects. *Infant Behavior and Development, 32*(3), 344–349. https://doi.org/10.1016/j.infbeh.2009.04.005

Jiang, J., Li, X., Liu, Z., Li, M., Chen, G., Li, Z., ... & Byeon, W. (2025). Token-efficient long video understanding for multimodal LLMs. *arXiv preprint: https://arxiv.org/pdf/2503.04130*

Kachergis, G., Marchman, V. A., & Frank, M. C. (2022). Toward a "Standard Model" of early language learning. *Current Directions in Psychological Science, 31*(1), 20–27. https://doi.org/10.1177/09637214211057836

Karasik, L. B., Tamis-LeMonda, C. S., & Adolph, K. E. (2014). Crawling and walking infants elicit different verbal responses from mothers. *Developmental Science, 17*(3), 388–395. https://doi.org/10.1111/desc.12129

Karmazyn-Raz, H., & Smith, L. B. (2022). Discourse with few words: Coherence statistics, parent-infant actions on objects, and object names. *Language Acquisition*, 1–19. https://doi.org/10.1080/10489223.2022.2054342

Kitamura, C., & Burnham, D. (1998). Acoustic and affective qualities of IDS in English. *5th International Conference on Spoken Language Processing (ICSLP 1998)*, paper 0909-0. https://doi.org/10.21437/ICSLP.1998-371

Kitamura, C., & Burnham, D. (2003). Pitch and communicative intent in mother's speech: Adjustments for age and sex in the first year. *Infancy, 4*(1), 85–110. https://doi.org/10.1207/S15327078IN0401_5

Kline, M. A., Shamsudheen, R., & Broesch, T. (2018). Variation is the universal: Making cultural evolution work in developmental psychology. *Philosophical Transactions of the Royal Society B: Biological Sciences, 373*(1743), 20170059. https://doi.org/10.1098/rstb.2017.0059

Kosie, J. E., Tsui, R. K. Y., Martinez, T., Sander, A., Fibla, L., Potter, C., Byers-Heinlein, K., & Lew-Williams, C. (2022). *Children's exposure to language switching in bilingual homes across two communities.* Talk presented at the Workshop on Infant Language Development (WILD). San Sebastian, Spain.

Kosie, J. E., Bala, A., & Baldwin, D. (2022). *Pupillometry sheds light on how caregivers scaffold infants' learning* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/zx4ek

Kosie, J. E., & Lew-Williams, C. (2024a). Infant-directed communication: Examining the many dimensions of everyday caregiver-infant interactions. *Developmental Science, 27*(5), e13515. https://doi.org/10.1111/desc.13515

Kosie, J. E., & Lew-Williams, C. (2024b). *Everyday caregiver-infant communication is shaped by activity context* [Talk]. International Congress of Infant Studies, Glasgow, Scotland.

Koterba, E. A., & Iverson, J. M. (2009). Investigating motionese: The effect of infant-directed action on infants' attention and object exploration. *Infant Behavior and Development, 32*(4), 437–444. https://doi.org/10.1016/j.infbeh.2009.07.003

Koubaa, A., Ammar, A., & Boulila, W. (2025). Next-generation human-robot interaction with ChatGPT and robot operating system. *Software: Practice and Experience, 55*(2), 355-382. https://doi.org/10.1002/spe.3377

Kretch, K. S., Franchak, J. M., & Adolph, K. E. (2014). Crawling and walking infants see the world differently. *Child Development, 85*(4), 1503–1518. https://doi.org/10.1111/cdev.12206

Kuchirko, Y., Tafuro, L., & Tamis LeMonda, C. S. (2018). Becoming a communicative partner: Infant contingent responsiveness to maternal language and gestures. *Infancy, 23*(4), 558–576. https://doi.org/10.1111/infa.12222

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., Stolyarova, E. I., Sundberg, U., & Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science, 277*(5326), 684–686. https://doi.org/10.1126/science.277.5326.684

LaBarbera, J. D., Izard, C. E., Vietze, P., & Parisi, S. A. (1976). Four- and six-month-old infants' visual responses to joy, anger, and neutral expressions. *Child Development, 47*(2), 535. https://doi.org/10.2307/1128816

Llama Team, Meta (2025, April 5) The Llama 4 herd: The beginning of a new era of natively multimodal Ai innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/

Lavechin, M., de Seyssel, M., Métais, M., Metze, F., Mohamed, A., Bredin, H., ... & Cristia, A. (2024). Modeling early phonetic acquisition from child-centered audio data. *Cognition, 245,* 105734. https://doi.org/10.1016/j.cognition.2024.105734

Levine, D., Buchsbaum, D., Hirsh-Pasek, K., & Golinkoff, R. M. (2019). Finding events in a continuous world: A developmental account. *Developmental Psychobiology, 61*(3), 376–389. https://doi.org/10.1002/dev.21804

Lew-Williams, C., Ferguson, B., Abu-Zhaya, R., & Seidl, A. (2019). Social touch interacts with infants' learning of auditory patterns. *Developmental Cognitive Neuroscience, 35,* 66–74. https://doi.org/10.1016/j.dcn.2017.09.006

Long, B., Xiang, V., Stojanov, S., Sparks, R. Z., Yin, Z., Keene, G. E., ... & Frank, M. C. (2024). The BabyView dataset: High-resolution egocentric videos of infants' and young children's everyday experiences. *arXiv preprint: https://arxiv.org/abs/2406.10447*.

Lowe, J. R., Coulombe, P., Moss, N. C., Rieger, R. E., Aragón, C., MacLean, P. C., Caprihan, A., Phillips, J. P., & Handal, A. J. (2016). Maternal touch and infant affect in the still face paradigm: A cross-cultural examination. *Infant Behavior and Development, 44,* 110–120. https://doi.org/10.1016/j.infbeh.2016.06.009

Ma, W., Golinkoff, R. M., Houston, D. M., & Hirsh-Pasek, K. (2011). Word learning in infant- and adult-directed speech. *Language Learning and Development, 7*(3), 185–201. https://doi.org/10.1080/15475441.2011.579839

Mani, N., & Ackermann, L. (2018). Why do children learn the words they do? *Child Development Perspectives, 12*(4), 253–257. https://doi.org/10.1111/cdep.12295

ManyBabies Consortium. (2020). Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Advances in Methods and Practices in Psychological Science*, *3*(1), 24–52. https://doi.org/10.1177/2515245919900809

Matatyaho, D. J., & Gogate, L. J. (2008). Type of maternal object motion during synchronous naming predicts preverbal infants' learning of word-object relations. *Infancy*, *13*(2), 172–184. https://doi.org/10.1080/15250000701795655

McMurray, B. (2016). Language at three timescales: The role of real-time processes in language development and evolution. *Topics in Cognitive Science*, *8*(2), 393–407. https://doi.org/10.1111/tops.12201

Meyer, M., Hard, B., Brand, R. J., McGarvey, M., & Baldwin, D. A. (2011). Acoustic packaging: Maternal speech and action synchrony. *IEEE Transactions on Autonomous Mental Development*, *3*(2), 154–162. https://doi.org/10.1109/TAMD.2010.2103941

Meyer, M., van Schaik, J. E., Poli, F., & Hunnius, S. (2022). how infant-directed actions enhance infants' attention, learning, and exploration: Evidence from EEG and computational modeling. *Developmental Science*, *26*(1). https://doi.org/10.1111/desc.13259

Murphy, C. M., & Messer, D. J. (1977). Mothers, infants and pointing: A study of a gesture. In *Studies in mother-infant interaction*. Academic Press.

Nencheva, M. L., Tamir, D. I., & Lew-Williams, C. (2023). Caregiver speech predicts the emergence of children's emotion vocabulary. *Child Development*, *94*(3), 585–602. https://doi.org/10.1111/cdev.13897

Ochs, E., & Schieffelin, B. B. (1984). Language acquisition and socialization: Three developmental stories. In R. A. Shweder & R. A. LeVine (Eds.) *Culture theory: Essays on mind, self, and emotion* (pp. 276–320). Cambridge University Press.

Ogren, M., Leotti, L., Hoemann, K., Oakes, L., Feldman Barrett, L., & LoBue, V. (2023). *What do they see and hear? 6-month-olds' natural emotional input from faces and language*. [Talk] Biennial Meeting of the Society for Research in Child Development, Salt Lake City, UT.

Okocha, A., Burke, N., & Lew-Williams, C. (2024). Infants and toddlers in the United States with more close relationships have larger vocabularies. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001609

Orhan, A. E., Gupta, V. V., & Lake, B. M. (2020). *Self-supervised learning through the eyes of a child*. 34th Conference on Neural Information Processing Systems, Vancover, Canada.

Outters, V., Hepach, R., Behne, T., & Mani, N. (2023). Children's affective involvement in early word learning. *Scientific Reports, 13*(1), 7351. https://doi.org/10.1038/s41598-023-34049-3

Outters, V., Schreiner, M. S., Behne, T., & Mani, N. (2020). Maternal input and infants' response to infant-directed speech. *Infancy, 25*(4), 478–499. https://doi.org/10.1111/infa.12334

Özçalişkan, Ş., & Goldin-Meadow, S. (2005). Do parents lead their children by the hand? *Journal of Child Language, 32*(3), 481–505. https://doi.org/10.1017/S0305000905007002

Panneton, R., Cristia, A., Taylor, C., & Moon, C. (2023). Positive valence contributes to hyperarticulation in maternal speech to infants and puppies. *Journal of Child Language,* 1–11. https://doi.org/10.1017/S0305000923000296

Panneton, R., Kitamura, C., Mattock, K., & Burnham, D. (2006). Slow speech enhances younger but not older infants' perception of vocal emotion. *Research in Human Development, 3*(1), 7–19. https://doi.org/10.1207/s15427617rhd0301_2

Piantadosi, S. T. (2023). Modern language models refute Chomsky's approach to language. In E. Gibson & M. Poliak (Eds.), *From fieldwork to linguistic theory: A tribute to Dan Everett* (pp. 353-414). Language Science Press.

Piazza, E. A., Iordan, M. C., & Lew-Williams, C. (2017). Mothers consistently alter their unique vocal fingerprints when communicating with infants. *Current Biology, 27*(20), 3162-3167.e3. https://doi.org/10.1016/j.cub.2017.08.074

Piazza, E. A., Nencheva, M. L., & Lew-Williams, C. (2021). The development of communication across timescales. *Current Directions in Psychological Science, 30*(6), 459–467. https://doi.org/10.1177/09637214211037665

Pomper, R., & Saffran, J. R. (2019). Familiar object salience affects novel word learning. *Child Development, 90*(2). https://doi.org/10.1111/cdev.13053

Reider, L. B., Bierstedt, L., Burris, J. L., Vallorani, A., Gunther, K. E., Buss, K. A., Pérez-Edgar, K., Field, A. P., & LoBue, V. (2022). Developmental patterns of affective attention across the first 2 years of life. *Child Development, 93*(6). https://doi.org/10.1111/cdev.13831

Rowe, M. L., & Goldin-Meadow, S. (2009). Differences in early gesture explain SES disparities in child vocabulary size at school entry. *Science, 323*(5916), 951–953. https://doi.org/10.1126/science.1167025

Rowe, M. L., Özçalışkan, Ş., & Goldin-Meadow, S. (2008). Learning words by hand: gesture's role in predicting vocabulary development. *First Language, 28*(2), 182–199. https://doi.org/10.1177/0142723707088310

Rowe, M. L., & Weisleder, A. (2020). Language development in context. *Annual Reviews of Developmental Psychology, 2*(1), 201-223. https://doi.org/10.1146/annurev-devpsych-042220-121816

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences, 112*(41), 12663–12668. https://doi.org/10.1073/pnas.1419773112

Roy, B. C., Frank, M. C., & Roy, D. (2009). Exploring word learning in a high-density longitudinal corpus. *Thirty-First Annual Conference of the Cognitive Science Society*.

Ryskin, R., & Fang, X. (2021). The many timescales of context in language processing. In *Psychology of Learning and Motivation* (Vol. 75, pp. 201–243). Elsevier. https://linkinghub.elsevier.com/retrieve/pii/S0079742121000244

Salo, V. C., Pannuto, P., Hedgecock, W., Biri, A., Russo, D. A., Piersiak, H. A., & Humphreys, K. L. (2021). Measuring naturalistic proximity as a window into caregiver–child interaction patterns. *Behavior Research Methods, 54*(4), 1580–1594. https://doi.org/10.3758/s13428-021-01681-8

Schatz, J. L., Suarez-Rivera, C., Kaplan, B. E., & Tamis-LeMonda, C. S. (2022). Infants' object interactions are long and complex during everyday joint engagement. *Developmental Science, 25*(4). https://doi.org/10.1111/desc.13239

Schmidt, C. L. (1996). *Scrutinizing reference: How gesture and speech are coordinated in mother-child interaction. 23*(2), 279–305. https://doi.org/10.1017/S0305000900008801

Schwab, J. F., Rowe, M. L., Cabrera, N., & Lew-Williams, C. (2018). Fathers' repetition of words is coupled with children's vocabularies. *Journal of Experimental Child Psychology, 166*, 437–450. https://doi.org/10.1016/j.jecp.2017.09.012

Seidl, A., Tincoff, R., Baker, C., & Cristia, A. (2015). Why the body comes first: effects of experimenter touch on infants' word finding. *Developmental Science, 18*(1), 155–164. https://doi.org/10.1111/desc.12182

Shneidman, L. A., & Goldin-Meadow, S. (2012). Language input and acquisition in a Mayan village: How important is directed speech? *Developmental Science, 15*(5), 659–673. https://doi.org/10.1111/j.1467-7687.2012.01168.x

Singh, L. (2008). Influences of high and low variability on infant word recognition. *Cognition, 106*(2), 833–870. https://doi.org/10.1016/j.cognition.2007.05.002

Singh, L., Cristia, A., Karasik, L. B., Rajendra, S. J., & Oakes, L. M. (2023). Diversity and representation in infant research: Barriers and bridges toward a globalized science of infant development. *Infancy, 28*(4), 708–737. https://doi.org/10.1111/infa.12545

Singh, L., Morgan, J. L., & Best, C. T. (2002). Infants' listening preferences: Baby talk or happy talk? *Infancy, 3*(3), 365–394. https://doi.org/10.1207/S15327078IN0303_5

Slone, L. K., Smith, L. B., & Yu, C. (2019). Self-generated variability in object images predicts vocabulary growth. *Developmental Science, 22*(6), e12816. https://doi.org/10.1111/desc.12816

Smith, L. B., Jayaraman, S., Clerkin, E., & Yu, C. (2018). The developing infant creates a curriculum for statistical learning. *Trends in Cognitive Sciences, 22*(4), 325–336. https://doi.org/10.1016/j.tics.2018.02.004

Smith, N. A., & Trainor, L. J. (2008). Infant-directed speech is modulated by infant feedback. *Infancy, 13*(4), 410–420. https://doi.org/10.1080/15250000802188719

Snow, C. E., & Ferguson, C. A. (1977). *Talking to children*. Cambridge University Press.

Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review, 27*(4), 501–532. https://doi.org/10.1016/j.dr.2007.06.002

Stack, D., & LePage, D. E. (1996). Infants' sensitivity to manipulations of maternal touch during face-to-face interactions. *Social Development, 5*(1), 41–55. https://doi.org/10.1111/j.1467-9507.1996.tb00071.x

Stack, D. M., & Arnold, S. L. (1998). Changes in mothers' touch and hand gestures influence infant behavior during face-to-face interchanges. *Infant Behavior and Development, 21*(3), 451–468. https://doi.org/10.1016/S0163-6383(98)90019-4

Stack, D., & Muir, D. W. (1990). Tactile stimulation as a component of social interchange: new interpretations for the still-face effect. *British Journal of Developmental Psychology, 8*(2), 131–145. https://doi.org/10.1111/j.2044-835X.1990.tb00828.x

Stack, D., & Muir, D. W. (1992). Adult tactile stimulation during face-to-face interactions modulates five-month-olds' affect and attention. *Child Development, 63*(6), 1509–1525. https://doi.org/10.2307/1131572

Suanda, S. H., Smith, L. B., & Yu, C. (2016). The multisensory nature of verbal discourse in parent–toddler interactions. *Developmental Neuropsychology, 41*(5–8), 324–341. https://doi.org/10.1080/87565641.2016.1256403

Suarez-Rivera, C., Schatz, J. L., Herzberg, O., & Tamis-LeMonda, C. S. (2022a). Joint engagement in the home environment is frequent, multimodal, timely, and structured. *Infancy, 27*(2), 232–254. https://doi.org/10.1111/infa.12446

Suarez-Rivera, C., Smith, L. B., & Yu, C. (2019). Multimodal Parent Behaviors Within Joint Attention Support Sustained Attention in Infants. *Developmental Psychology, 55*(1), 96–109. https://doi.org/10.1037/dev0000628

Sullivan, J., Mei, M., Perfors, A., Wojcik, E., & Frank, M. C. (2021). SAYCam: A large, longitudinal audiovisual dataset recorded from the infant's perspective. *Open Mind, 5*, 20–29. https://doi.org/10.1162/opmi_a_00039

Tamis-LeMonda, C. S., Custode, S., Kuchirko, Y., Escobar, K., & Lo, T. (2019). routine language: Speech directed to infants during home activities. *Child Development, 90*(6), 2135–2152. https://doi.org/10.1111/cdev.13089

Tamis-LeMonda, C. S., Song, L., Leavell, A. S., Kahana-Kalman, R., & Yoshikawa, H. (2012). Ethnic differences in mother–infant language and gestural communications are associated with specific skills in infants. *Developmental Science, 15*(3), 384–397.

https://doi.org/10.1111/j.1467-7687.2012.01136.x

Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action*. MIT Press. https://doi.org/10.7551/mitpress/2524.001.0001

Tincoff, R., Seidl, A., Buckley, L., Wojcik, C., & Cristia, A. (2019). Feeling the way to words: parents' speech and touch cues highlight word-to-world mappings of body parts. *Language Learning and Development, 15*(2), 103–125. https://doi.org/10.1080/15475441.2018.1533472

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *LLaMA: Open and efficient Foundation Language Models* (arXiv:2302.13971). arXiv. http://arxiv.org/abs/2302.13971

Trainor, L. J., Austin, C. M., & Desjardins, R. N. (2000). Is infant-directed speech prosody a result of the vocal expression of emotion? *Psychological Science, 11*(3), 188–195. https://doi.org/10.1111/1467-9280.00240

Trainor, L. J., & Desjardins, R. N. (2002). Pitch characteristics of infant-directed speech affect infants' ability to discriminate vowels. *Psychonomic Bulletin & Review, 9*(2), 335–340. https://doi.org/10.3758/BF03196290

Tsai, J. L. (2017). Ideal affect in daily life: Implications for affective experience, health, and social behavior. *Current Opinion in Psychology, 17*, 118–128. https://doi.org/10.1016/j.copsyc.2017.07.004

Vigliocco, G., Motamedi, Y., Murgiano, M., Wonnacott, E., Marshall, C., Milán-Maillo, I., & Perniss, P. (2019). *Onomatopoeia, gestures, actions and words: How do caregivers use multimodal cues in their communication to children?* [Preprint]. PsyArXiv. https://osf.io/v263k

Vlach, H. A., & Sandhofer, C. M. (2011). Developmental differences in children's context-dependent word learning. *Journal of Experimental Child Psychology, 108*(2), 394–401. https://doi.org/10.1016/j.jecp.2010.09.011

Vogt, P., Mastin, J., Masson-Carro, I., & De Jong, C. (2020). *Multimodal interactions among infants in three radically different learning environments* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/xfkag

Vogt, S., & Kauschke, C. (2017). Observing iconic gestures enhances word learning in typically developing children and children with specific language impairment. *Journal of Child Language, 44*(6), 1458–1484. https://doi.org/10.1017/S0305000916000647

Vong, W. K., Wang, W., Orhan, A. E., & Lake, B. M. (2024). Grounded language acquisition through the eyes and ears of a single child. *Science, 383*(6682), 504–511. https://doi.org/10.1126/science.adi1374

Weng, Z., Wang, K.-C., Kanazawa, A., & Yeung, S. (2022). Domain adaptive 3D pose augmentation for in-the-wild human mesh recovery. *2022 International Conference on 3D Vision (3DV)*, 261–270. https://doi.org/10.1109/3DV57658.2022.00038

Williamson, R. A., & Brand, R. J. (2014). Child-directed action promotes 2-year-olds' imitation. *Journal of Experimental Child Psychology, 118*, 119–126. https://doi.org/10.1016/j.jecp.2013.08.005

Wojcik, E. H., Zettersten, M., & Benitez, V. L. (2022). The map trap: Why and how word learning research should move beyond mapping. *Wiley Interdisciplinary Reviews: Cognitive Science, 13*(4), e1596. https://doi.org/10.1002/wcs.1596

Wu, Y., Schulz, L. E., Frank, M. C., & Gweon, H. (2021). Emotion as information in early social learning. *Current Directions in Psychological Science, 30*(6), 468–475. https://doi.org/10.1177/09637214211040779

Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition, 125*(2), 244–262. https://doi.org/10.1016/j.cognition.2012.06.016

Zettersten, M., & Saffran, J. R. (2021). Sampling to learn words: Adults and children sample words that reduce referential ambiguity. *Developmental Science, 24*(3), e13064. https://doi.org/10.1111/desc.13064

Zieber, N., Kangas, A., Hock, A., & Bhatt, R. S. (2014). Infants' perception of emotion from body movements. *Child Development, 85*(2), 675–684. https://doi.org/10.1111/cdev.12134

Data, Code, and Materials Availability Statement

This review paper does not involve any new data, code, or materials.

Authorship and Contributorship Statement

The manuscript was led by **Jessica E. Kosie**, with all authors (**Jessica E. Kosie, Mira L. Nencheva, Justin A. Jungé, and Casey Lew-Williams**) contributing to conceptualization, writing, and revision. All authors approved the final version of the manuscript prior to submission.

License