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