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.

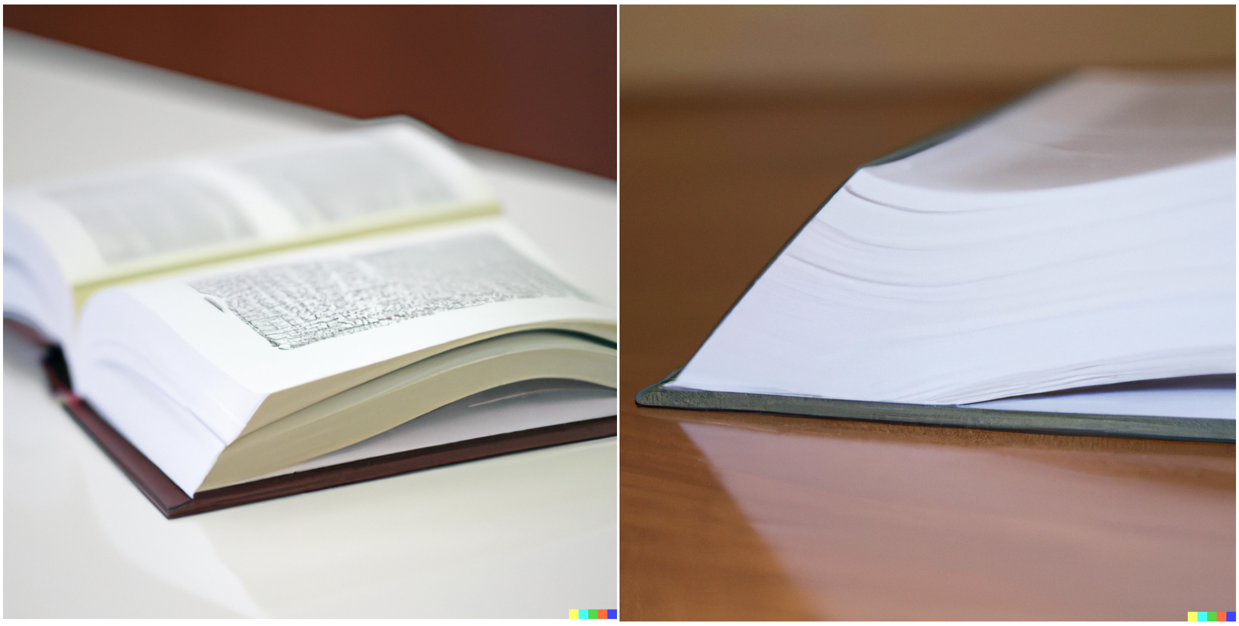

Figure 1. (Left) ***DALL-E 2 output for “A book on a table.”*** **(Right)** ***DALL-E 2 output for “A table on a book.” Representative of others in the set.***

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