

Large language models as proxies for theories of human linguistic cognition

Imry Ziv¹, Nur Lan^{1,2}, Emmanuel Chemla^{2,3}, and Roni Katzir¹

¹ Tel Aviv University

² École Normale Supérieure

³ Earth Species Project

ABSTRACT

We consider the possible role of current large language models (LLMs) in the study of human linguistic cognition. We focus on the use of such models as proxies for theories of cognition that are relatively linguistically-neutral in their representations and learning but differ from current LLMs in key ways. We illustrate this potential use of LLMs as proxies for theories of cognition in the context of two kinds of questions: (a) whether the target theory accounts for the acquisition of a given pattern from a given corpus; and (b) whether the target theory makes a given typologically-attested pattern easier to acquire than another, typologically-unattested pattern. For each of the two questions we show, building on recent literature, how current LLMs can potentially be of help, but we note that at present this help is quite limited.

Keywords:
learning, poverty
of the stimulus,
linguistic
typology, large
language models

INTRODUCTION

1

The question of whether current large language models (LLMs) might be relevant scientifically for the study of human linguistic cognition (HLC) is often framed in terms of a perceived threat that LLMs pose to

generative linguistics. On the one hand, some researchers have suggested that these models discredit prevailing views within linguistics (Piantadosi, 2024), while others, such as Chomsky *et al.* (2023), have suggested that LLMs can teach us nothing about HLC.

Here we chart a more collaborative path forward for the integration of LLMs into the study of HLC. While the idea that current LLMs are themselves good theories of HLC is untenable (as reviewed below), we note that LLMs are a potentially useful tool that can inform the evaluation of theories of HLC. In this capacity, LLMs can serve as proxies for explicit theories of HLC that are relatively unbiased with respect to linguistic representations and learning but different from current LLMs in important ways. We take it that much of the recent literature on the potential cognitive relevance of LLMs takes current models to support relatively linguistically-neutral theories of this kind. To the extent that this use of LLMs as proxies is justified, they can assist in comparing such relatively unbiased theories of HLC with more traditional theories of HLC coming out of theoretical linguistics, where the combination of representations and learning is often highly biased in favour of linguistic patterns.

We start in Section 2 by briefly reviewing some considerations pertaining to the role of LLMs in studying HLC. We start with a reminder that the study of HLC, like the rest of empirical science, proceeds through inference to the best explanation, with explicit theories competing in light of known observations. We review the recent idea that LLMs themselves can serve as better explanations than generative theories and then summarize some of the reasons noted in the literature to reject this idea. We then consider the idea mentioned above that the best explanation is that HLC is quite different from current LLMs but still involves representations and learning that are relatively linguistically neutral. It is the potential of LLMs to serve as proxies for such a relatively linguistically-neutral theory that will preoccupy us in the remainder of the paper, and in particular the question of whether evidence from LLMs can bear on the question of whether the linguistically-neutral theory is a better explanation than generative approaches. In Section 3 we ask – building on Wilcox *et al.* 2023 and Lan *et al.* 2024b, among others – whether evidence from LLMs suggests that a relatively linguistically-neutral theory of HLC might succeed in acquiring various linguistic patterns from a developmentally-

realistic corpus. In Section 4 we ask – building on Kallini *et al.* 2024 and others – whether evidence from LLMs suggests that a relatively linguistically-neutral theory of HLC makes attested languages easier to acquire than various unattested variants of those languages. In both cases, we will see that the evidence from LLMs gives no reason to think that a linguistically-neutral theory of HLC could account for the data. At least as far as the models and phenomena considered here are concerned, then, current LLMs contribute little to the ability of such linguistically-neutral theories to challenge generative theories. But moving the discussion further is likely to require much more detail about the relevant linguistically-neutral theories than what the literature supporting such theories currently offer, and we hope that the present paper serves to encourage those who are optimistic about such theories to provide the missing details.

BACKGROUND: LLMS, HLC, AND THE INFERENCE TO THE BEST EXPLANATION

2

Being part of empirical science, the study of HLC proceeds through inference to the best explanation: explicit theories are evaluated in light of the observations, and those theories that provide the best explanations of the data are preferred to the other contenders. The theories coming out of generative linguistics over the past 70 years or so are routinely evaluated in this way. These theories are often strongly linguistically biased, in the sense that they make it relatively easy to represent and learn knowledge of the kind found across languages and hard or impossible to represent or learn many other patterns. Let us use H_1 to refer to some concrete instantiation of the generative approach. We will return to the biases embodied by the architecture and representations of H_1 in the following sections.

More recently, Piantadosi (2024) has suggested that current LLMs should be treated as explicit theories of HLC that should be taken seriously as part of this evaluation. Following Fox and Katzir 2024 we

refer to this new approach to HLC as the *LLM Theory*, and we discuss it in Section 2.1, through a concrete instantiation we will call H_2 . Perhaps in light of the glaring inadequacies of the LLM Theory, some recent literature, including Wilcox *et al.* 2023 and Mahowald *et al.* 2024, has stopped short of explicitly endorsing the LLM Theory while still taking the performance of current LLMs to challenge central claims within theoretical linguistics. It is conceivable that this more recent literature is still meant to support the LLM Theory, in which case the objections to that theory, reviewed below, still apply. But it is also possible that at least some of these works aim at a more nuanced view on the role of LLMs in studying HLC, one in which the performance of current models can be taken to challenge generative views and support a theory of HLC H_3 that is not as strongly linguistically-biased as H_1 but is also very different from current models themselves, and therefore different from H_2 . Our goal in this paper is to articulate this more nuanced view and to make a preliminary attempt to evaluate it in light of current evidence. We also hope to encourage adherents of this view to make explicit the target H_3 they have in mind.

The idea, as we understand it, is the following. If one wishes to challenge claims for strong linguistic biases, as in H_1 mentioned above, and to defend a more linguistically-neutral alternative H_3 , one would need to address the arguments that theoretical linguistics has provided for H_1 from acquisition and from the typology, among other things. But doing so directly can be difficult. For one thing, H_3 might make it hard to test acquisition predictions based on developmentally-realistic corpora. It can also be hard to evaluate the ability of H_3 to predict typological patterns. It is at this point that current LLMs can be helpful. While these models might be quite different from H_3 , they too lack the strong linguistic biases that theoretical linguistics has argued for, so their successful performance on tasks related to learning and typology might lend support to the possibility that H_3 will succeed on these tasks despite not having the strong linguistic biases that theoretical linguistics has argued for. In this way, current LLMs can serve as proxies for H_3 . We refer to this idea as the *Proxy View*:

- (1) The Proxy View of LLMs in HLC: current LLMs are not themselves good theories of HLC, but their performance is indicative of the success of some other theory of HLC, presumably

more linguistically-neutral than the generative approach, and can inform the inference to the best explanation.

Since the literature that promotes the Proxy View generally does not make its supported theory explicit, our ability to evaluate it at present is limited. We will therefore focus on methodological considerations that pertain to the evaluation of the target theory once it is provided and show how LLMs might bear on this evaluation. We show that, at least at present and with respect to the phenomena considered here, LLMs lend no support to the Proxy View.

The LLM Theory and its inadequacy

2.1

As mentioned, Piantadosi (2024) argues for the LLM Theory of HLC. That is, he suggests that current LLMs, taken as real theories of HLC, are a better explanation of the known observations than anything coming out of the generative tradition. This claim is incorrect, for reasons discussed in detail in the literature and summarized briefly here. Following Katzir 2023 and Fox and Katzir 2024 we structure our brief discussion of the failure of the LLM Theory around three key issues: (a) competence vs. performance, (b) correctness vs. probability, and (c) learning and representations.¹

Competence vs. performance. Humans struggle with center embedding even at very modest depths of embedding, as in “The mouse that the cat that the dog chased bit died”. A naive account might attribute this to the grammar, part of the competence of speakers: perhaps human grammars rule out center embedding beyond a certain (low) threshold. A much better explanation, going back to Yngve 1960 and Miller and Chomsky 1963, attributes the difficulty with center embedding not to the competence of speakers but rather to their performance, typically by attributing it to limitations of their working memory. Among other things, the account in terms of performance explains why center embedding is hard, as seen in reaction times and various errors, as well as from introspection, rather than being just unacceptable. Unacceptability due to the grammar alone,

¹ See Kodner *et al.* 2023, Moro *et al.* 2023, Rawski and Baumont 2023, and others for further critical discussion of the LLM Theory.

as in “John ate apple the”, does not share this property; it is simply judged to be bad. The account in terms of the distinction between competence and performance also explains why the unacceptability of center embedding is gradient, growing with each level of embedding, and why it can vary with certain factors (including with the use of pronouns instead of referring nouns in the most embedded position). Again, the account in terms of competence alone has no explanation for these facts.

Differently from common generative proposals, the design and behaviour of current LLMs does not suggest a meaningful distinction between competence and performance of the kind discovered in humans: there is no reason to think that more (or less) time and working memory would make them change their behaviour on center embedding or agreement. This leaves current LLMs with no good explanation for the increasing difficulty of sentences with increasing levels of center embedding or for the other observations that have been taken to support the distinction between competence and performance.

Correctness vs. likelihood. Humans recognize some sentences as correct but unlikely (one example again is multiple center embedding as in the example above), and they recognize some other sentences as incorrect but likely (for example, so-called agreement attraction, as in the incorrect “The keys to the cabinet is on the table”, where *is* appears in the singular despite a plural subject *keys*, presumably because of the linear proximity of the singular *cabinet*). The significance of the distinction between correctness and likelihood for linguistics was pointed out by Chomsky (1957) and discussed further by Berwick (2018), Sprouse *et al.* (2018), and others.

LLMs provide straightforward information regarding their representation of likelihood: given a prefix, they provide a probability distribution over the next token. But they do not, at present, provide information about correctness as distinct from likelihood, and it remains unclear whether they even have suitable representations to support this distinction.²

²A potentially important area of research concerns understanding the inner workings of LLMs, particularly through the study of mechanistic interpretation. See Lakretz *et al.* 2019, Ravfogel *et al.* 2021, Rai *et al.* 2024 and López-Otal *et al.* 2025, among others. Such investigations have the potential of making the inner

We note that the literature often compares the likelihood assessments of LLMs to acceptability judgments of speakers. Such, for example, is the methodology of works that evaluate the preferences of LLMs within minimal pairs, including Linzen *et al.* 2016, Bernardy and Lappin 2017, Chowdhury and Zamparelli 2018, Gulordava *et al.* 2018, Kuncoro *et al.* 2018, Marvin and Linzen 2018, Wilcox *et al.* 2018, 2019, 2023, Bhattacharya and van Schijndel 2020, Chaves 2020, Warstadt *et al.* 2020, Huebner *et al.* 2021, Ozaki *et al.* 2022, Yedetore *et al.* 2023, and Lan *et al.* 2024b. This methodology can be informative for particular purposes and under particular conditions, and in Section 3 we adopt it ourselves, but this should not be taken to imply that the relevant likelihood assessments correspond to correctness. As mentioned, the two notions are distinct.³

Learning and representations. The inductive leaps that humans make, sometimes based on very little data, suggest specific representations and biases. This is the essence of so-called *arguments from the poverty of the stimulus* (APSs). A particularly well-known APS, originally outlined by Chomsky (1971) and discussed in detail in much later work (see Berwick *et al.* 2011 for a more recent assessment), concerns hierarchical structure, often stated in terms of constituency.⁴

workings of LLMs less opaque, and it is conceivable that they will eventually uncover a notion of correctness within LLMs.

³ A different method that might appear to offer insight into a notion of correctness or grammaticality in LLMs is prompting, often through interfaces such as ChatGPT. This roughly corresponds to comparing the distribution over ‘yes’ and ‘no’ given a prefix such as “Is the following grammatically correct?” + <target-sentence>. (The precise evaluation depends on the particular prompt that is used and can vary further depending on the specifics of the relevant chat interface.) This is often a convenient way to anecdotally illustrate a general point, by making an LLM appear to engage in reporting its acceptability assessment of the target sentence. While some work, including Dentella *et al.* 2023 and Qiu *et al.* 2025, uses prompting on a larger scale, aiming at making general points rather than just illustration, we find this use hard to interpret. In any event, we know of no particular reason to think that prompting provides access to any internal notion of correctness within the LLM (if such a notion even exists).

⁴ As a reviewer notes, this APS concerns several other syntactic notions in addition to hierarchical structure. For presentational purposes we focus the current discussion on hierarchical structure.

This is often illustrated with the formation of yes-no questions, as in (2). The generalization that English-learning children reach is that such questions are formed by fronting the auxiliary that is structurally highest, as in (2a). They do not seem to seriously consider an alternative generalization that says that the fronted auxiliary is the one that is first linearly (which would yield the incorrect (2b)), even though the data that children are typically exposed to has been argued to be unhelpful in making the choice.

- (2) [The boy who is singing] can dance
- a. Can [the boy who is singing] __ dance?
 - b. * Is [the boy who __ singing] can dance?

Generative theories of HLC typically derive this inductive leap by assuming that children are born with a combination of representations and learning biases that favour dependencies that rely on hierarchical structure over dependencies that rely on linear order (in some approaches by preventing the latter from being represented in the first place). Current LLMs, on the other hand, seem to provide no basis for this inductive leap. In fact, recent work by Lan *et al.* (2022, 2024a) casts doubt on the ability of networks trained by current methods to even acquire the basic hierarchical notions, let alone to explain the preference for dependencies that rely on these notions over dependencies that rely on linear order. Not surprisingly, then, Yedetore *et al.* (2023) find evidence suggesting that at least some neural networks fail to acquire the correct hierarchically-based generalization from their training data. We will return to APSs in Section 3, where we will consider the possible role of LLMs in reasoning about the inductive leaps predicted by non-LLM theories of HLC.

Evidence about the representations and learning biases of humans also comes from the highly skewed distribution of linguistic patterns cross-linguistically. We can again consider hierarchical structure, following Fox and Katzir 2024, though we emphasize that a very wide range of cross-linguistic patterns have been discussed in the linguistic literature. We just mentioned that biases that refer to hierarchical structure are part of the best explanation of the competence of English speakers, noting evidence from the formation of yes-no questions. Notions of hierarchical structure are central to the explanation of many

other aspects of English, including center-embedding and agreement mentioned earlier, wh-movement discussed in Section 3 below, the computation of meaning, and more. But even if one ignores the APSs for hierarchical structure, it seems clear that hierarchical structure is not just an accident of actual English: it is part of the best explanation of almost all languages that have been studied closely, which suggests that humans are born with representations and learning biases that favour hierarchical structure. Again, the representations and learning biases of current LLMs might not even be able to support the acquisition of the relevant notions of hierarchical structure, let alone explain the cross-linguistic prevalence of these notions. We will return to cross-linguistic evidence in Section 4, where we will consider the possible role of LLMs in reasoning about the typological patterns predicted by non-LLM theories of HLC.

LLMs as proxies for good learners

2.2

But if the LLM Theory is discredited, does this mean that current LLMs are irrelevant to the scientific study of HLC? As mentioned above, the Proxy View explains why the answer is no: in principle current LLMs can serve as proxies for theories of a third kind, theories that involve relatively linguistically-neutral representations and learning (and are different in this regard from typical generative theories) but that are not vulnerable to the most immediate arguments against the LLM Theory. Table 1 schematically compares the three kinds of theories of

Table 1: Comparison of the three theories outlined above on key issues, following Fox and Katzir 2024

Criterion	Generative linguistics	LLM theory	Proxy view
Linguistic bias	Strong	Weak	Intermediate
Competence vs. performance	✓	×	✓
Grammaticality $\neq$ likelihood	✓	×	✓
Explanation for learning asymmetries	✓	×	✓
Concrete instantiation	H_1	H_2	H_3

HLC that we discuss here: a concrete generative theory, H_1 ; a concrete instantiation of the LLM Theory, H_2 ; and a concrete, relatively linguistically-neutral, H_3 , possibly very different from current LLMs, for which these LLMs might serve as proxies. The Proxy View, as stated in (1) above, uses current LLMs as tools for evaluating H_3 , in comparison with H_1 and as part of the inference to the best explanation in light of the existing observations and arguments.

At this point we should mention again an immediate obstacle to evaluating the Proxy View: while recent literature, including Wilcox *et al.* 2023 and Mahowald *et al.* 2024, suggests that the performance of LLMs supports linguistically-neutral theories of HLC, this literature does not provide much by way of specifics. Until such specifics of H_3 are provided, and until H_3 is shown to be a realistic candidate for the best explanation in light of the existing literature, expressions of hope about H_3 remain just that.

Still, we can outline some methodological considerations regarding the possible role of current LLMs as proxies for H_3 once the specifics of H_3 are made sufficiently explicit. Here we consider two such uses of LLMs as proxies for H_3 . Following Lan *et al.* 2024b we use LLMs as tools for reasoning about whether H_3 predicts a given inductive leap based on a given amount of linguistic exposure. And following Kallini *et al.* 2024 we use LLMs as tools for reasoning about whether H_3 derives a learning asymmetry that might be part of an explanation of a typological asymmetry.

3 LARGE LANGUAGE MODELS AND ALIGNMENT WITH THE STIMULUS

Consider then an H_3 that is neutral with respect to some linguistic property *PROP* for which generative theories such as H_1 include a strong linguistic bias. A potential argument against H_3 and in favour of H_1 might take the form of an APS: if humans show knowledge of *PROP* after exposure to a corpus that seems too impoverished to support the learning of *PROP* by a linguistically-neutral learner, this will appear to support the strongly biased H_1 . The problem is that it is generally very hard to determine what a given learner might acquire from a given

corpus, even for a theory such as H_3 for which both the representations and the learning mechanism are explicitly provided.

LLMs might be able to help. We can train them on a suitable corpus and then probe their knowledge. Note that this is not entirely straightforward, for various reasons discussed in Lan *et al.* 2024b. Most immediately, while we might be able to directly inspect a grammar within a theory of HLC such as H_3 and ask whether it has knowledge of *PROP*, we cannot do the same with current LLMs, whose inner workings are extremely complex and remain mostly opaque, as discussed in Section 2.1. Instead of inspection, we might turn to behavioural measures such as acceptability judgments in minimal pairs (see Section 2.1 above), where one member satisfies *PROP* and is acceptable and a minimal variant violates it and is unacceptable. An immediate concern with this methodology is that, while the theory of HLC under consideration might offer such judgments, current LLMs provide only likelihood assessments, through their distribution over the next token given a prefix. Likelihood assessments are very different from acceptability judgments, as mentioned above. However, if we focus on areas where acceptability and likelihood are reasonably well aligned, we might be justified in looking at whether the LLM assigns a higher probability to the *PROP*-satisfying member of the pair than to the *PROP*-violating one (a methodology that has been used in Marvin and Linzen 2018, Hu *et al.* 2020 and Wilcox *et al.* 2019, among many other works; see Section 2.1). If this probabilistic preference is sufficiently strong, we might consider the fact that the LLM has acquired this preference as reason to think that the learner under our target theory of HLC H_3 would acquire the actual *PROP*. This might be so even if the probabilistic preference of the LLM reflects a flawed approximation of *PROP* rather than knowledge of *PROP* in any meaningful sense.

Following Wilcox *et al.* and Lan *et al.*, our success criterion considers the model successful if and only if it assigns a higher probability to the grammatical member of the minimal pair. Specifically, we evaluate probabilities on a critical region token whose preceding context is the same in both the grammatical and the ungrammatical sentence (see Table 3, where critical regions are indicated with underscores). Note that this is an extremely lenient notion of success: the model is considered successful even if the preference for the grammatical member of the minimal pair is extremely small, and even if ungrammatical

sentences are considered more likely than some (less related) grammatical sentences.

3.1

Methods

Let us illustrate this kind of reasoning with concrete examples. To do so, we use an empirical setup that consists of eight models of different architectures and training schemes, ranging from the rough equivalent of several months of human linguistic input to hundreds of thousands of years, with multiple models in between, as summarized in Table 2.⁵ Human equivalents are based on the assumption by Hart and Risley (1995) that children are exposed to around 30,000 words per day, while roughly equating words with model tokens.

Table 2:
Training data
size of the eight
language models
considered here,
and the human
linguistic
experience
equivalent to
these data sizes

Model	Train dataset size	Human equivalent
CHILDES LSTM, Transformer	8.6 million tokens	10 months
BabyLM 10M	10 million tokens	1 year
Wikipedia Transformer	90 million tokens	8 years
BabyLM 100M	100 million tokens	9 years
BERT Base Uncased	≈ 3.5 billion tokens	320 years
GPT-2	≈ 8 billion tokens	730 years
LLaMA-3.2-3b	≈ 9 trillion tokens	821,250 years

The models were tested on three linguistic phenomena using the very lenient success criterion discussed above (see Table 3). The models that are most directly relevant to assessing H_3 with respect to

⁵We use two models from Yedetore et al. 2023, an LSTM and a Transformer, that were trained on the CHILDES corpus of child-directed speech; a Transformer we trained on a subset of English Wikipedia, for which we used one of the large Transformer architectures used in Yedetore et al.; the *base-strict* and *base-strict-small* versions of the BabyLM LLMs from Warstadt et al. 2023, trained on the 100M and 10M BabyLM datasets respectively; the *BERT base uncased* version of BERT (Devlin et al. 2019), trained on the BooksCorpus and English Wikipedia datasets; OpenAI’s GPT-2 (Radford et al. 2019); and Meta’s *LLaMA 3.2-3b* model (released September 2024, Grattafiori et al., 2024). See Appendix A for further technical detail. All experimental material can be found at <https://osf.io/5zh6q/>.

Phenomenon	Example (Grammatical/*Ungrammatical)
Across-the-board movement (ATB)	Which boy did you say that Kim hated and that Mary loved yesterday? * Which boy did you say that Kim hated and that Mary loved <u>Ann</u> yesterday?
Parasitic gaps (PG)	I know who John's talking to is going to annoy <u>soon</u> . * I know who John's talking to is going to annoy <u>you</u> soon. * Who did you say that <u>loves</u> Sue?
That-trace effects (TTE)	Who did you say that <u>Sue</u> loves? Who did you say <u>loves</u> Sue? Who did you say <u>Sue</u> loves?

Table 3:
Examples of linguistic phenomena with grammatical and ungrammatical sentences. Model probabilities are compared on the underscored critical regions

any given APS are the smaller ones, which are trained on inputs that roughly match the number of words that a child might be exposed to. If these models prefer the grammatical option over the ungrammatical one in the minimal pairs of the test set for a given phenomenon, this can support the idea that H_3 would actually acquire the relevant piece of linguistic knowledge from a developmentally-realistic corpus.⁶ Conversely, lack of such a preference can weaken the plausibility that H_3 would succeed in acquiring the relevant knowledge from a similar corpus.⁷

Two of the models (BERT and GPT-2) were trained on the equivalent of hundreds of years of linguistic input. This is clearly beyond what a human child has been exposed to by the time they show evidence of the relevant linguistic knowledge. We included these models

⁶ Recall that we do not ask whether the LLMs actually acquire the linguistic knowledge under consideration, nor do we expect them to.

⁷ As should be clear, even if there is evidence that H_3 could acquire the relevant knowledge from a developmentally-realistic corpus there might still be good reasons to adopt an H_1 that is strongly biased in favour of this knowledge. Among other things, H_1 might provide a better explanation than H_3 for the cross-linguistic pattern. We believe that this is the case for the phenomena discussed in the present section and refer the reader to the literature cited below.

to allow proponents of the Proxy View to use LLMs as broadly adequate proxies for H_3 that just happen to be extremely data inefficient.

We included the much larger LLaMA mainly to see whether current architectures are sensible proxies for H_3 for the phenomena at hand. If they cannot show even a slight preference for the correct element in minimal pairs after training on the equivalent of hundreds of thousands of years of linguistic input, then perhaps they are incapable of approximating the pattern under consideration, which would suggest that they are not very good proxies for H_3 with respect to this pattern.

3.2

Test case: across-the-board movement and parasitic gaps

Above we mentioned displacement phenomena, illustrating with the position of the auxiliary in yes-no questions in English. Displacement phenomena have featured in recent discussions of the role of LLMs in evaluating APSs, both for the placement of the auxiliary (see, e.g., Yedetore *et al.* 2023) and for wh-movement (see Chowdhury and Zamparelli 2018, Bhattacharya and van Schijndel 2020, Chaves 2020, Warstadt *et al.* 2020, Ozaki *et al.* 2022, Wilcox *et al.* 2023, and Lan *et al.* 2024b). The latter will concern us in the current two subsections. Here is a simple illustration:

- (3) [Which book] did you say that Mary read ___ last week?

The common analysis of wh-movement in generative theories involves a constituent that is attached in one position and is then re-attached higher up in the structure. In (3), the constituent [*which book*] is attached within the embedded clause as a sister to *read* and is then re-attached at or near the root of the matrix clause, which among other things leads to it being pronounced in the beginning of the sentence.

Some work within generative linguistics (Pearl and Sprouse 2013, Phillips 2013) suggests that a relatively linguistically-neutral theory cannot account for the full knowledge of wh-movement given the kind of data that children are exposed to, an instance of the APS. This is not a consensus view, however, and some recent work such as Wilcox *et al.* 2023 suggests that LLMs undermine this APS and that these models show that a linguistically-neutral theory would, in fact, acquire a

full knowledge of wh-movement from a developmentally-realistic corpus. Following Lan *et al.* let us show how LLMs (used as proxies for an undisclosed linguistically-neutral theory of HLC) might bear on the APS under consideration and why they currently do nothing to undermine it.

Before proceeding, let us reiterate that it is hard to tell whether LLMs are good proxies for the relatively linguistically-neutral H_3 that seems to be envisioned in works such as Wilcox *et al.* 2023 in the absence of at least some specifics about that theory. But if they are, and if we focus on cases where correctness and likelihood are reasonably well-aligned, we could reason that if the LLM overwhelmingly assigns much higher probability to the correct member of each minimal pair, this indicates that there was sufficient information in the training data for the target linguistically-neutral theory H_3 to acquire the relevant knowledge. And if the LLM does not perform as well, this suggests that the target theory H_3 might also struggle. The former possibility weakens the APS, and the latter strengthens it.

Following Lan *et al.* (2024b) we focus on a family of wh-movement configurations where there is complex interaction between two gaps, in a way that seemingly violates island constraints on syntactic movement.⁸ The first such phenomenon is *across-the-board (ATB) movement* in coordinate structures. While extraction from a single coordinate is ungrammatical, as in (4a) and (4b), extraction can become grammatical when it takes place from both coordinates, as in (4c):

- (4) a. * [Which book] did you say [that Kim wrote __ last year] and [that Mary read a new novel yesterday]?

⁸ See Ross 1967, Williams 1977, 1990, Engdahl 1983, Haik 1985, Munn 1992, Postal 1993, Fox 2000, Nissenbaum 2000, Hornstein and Nunes 2002, and Sportiche 2024, among many others, for many important observations about the relevant phenomena that we will not be able to discuss in the present paper. Our simplified treatment here suffices to illustrate the methodological point about evaluating the Proxy View, and since our conclusions below are negative even for the simplified picture, we see no reason to think that current models will successfully approximate the complexities of these phenomena discussed in the literature. Any attempt to make a positive claim about these phenomena would of course need to address these complexities.

- b. * [Which book] did you say [that Kim wrote a new novel last year] and [that Mary read __ yesterday]?
- c. [Which book] did you say [that Kim wrote __ last year] and [that Mary read __ yesterday]?

The second nuance of wh-movement we test is when extraction from an island is made possible by the presence of another gap downstream, a phenomenon known as a *parasitic gap* (PG):

- (5) a. * I know who [John's talking to __] is going to annoy you soon.
- b. I know who [John's talking to __] is going to annoy __ soon.

We test the eight models from Table 2 on their approximation of across-the-board movement and parasitic gaps using the minimal pair method. We compare sentences that both have an upstream filler – like *which book* in (4), or *who* in (5) – and differ in whether they are gapped accordingly.⁹ As shown in Figure 1, all models other than LLaMA prefer the ungrammatical over the grammatical members of the minimal pairs in the vast majority of cases. This failure suggests that the linguistically-neutral H_3 for which the LLMs are taken to be proxies would not acquire wh-movement from a developmentally-realistic corpus.

3.3

Test case: that-trace effects

Another property of wh-movement in English is that when the moved wh-phrase is a subject, the clause from which it moves cannot have an overt complementizer such as *that*:

- (6) a. * Who did you say that __ loves Sue? _(+that, +trace)
- b. Who did you say that Sue loves __ ? _(+that, -trace)

⁹The paradigm sentences in Lan et al. 2024b were generated by separate templates for ATB and PG. Here we unify the templates for both phenomena so that they use the same lexical choices where relevant (e.g., proper names), and by adding more lexical choices. From each template, represented by a context-free grammar, we sampled 10,000 sentence pairs and ran them through each model to get the relevant probability values (see Appendix B).

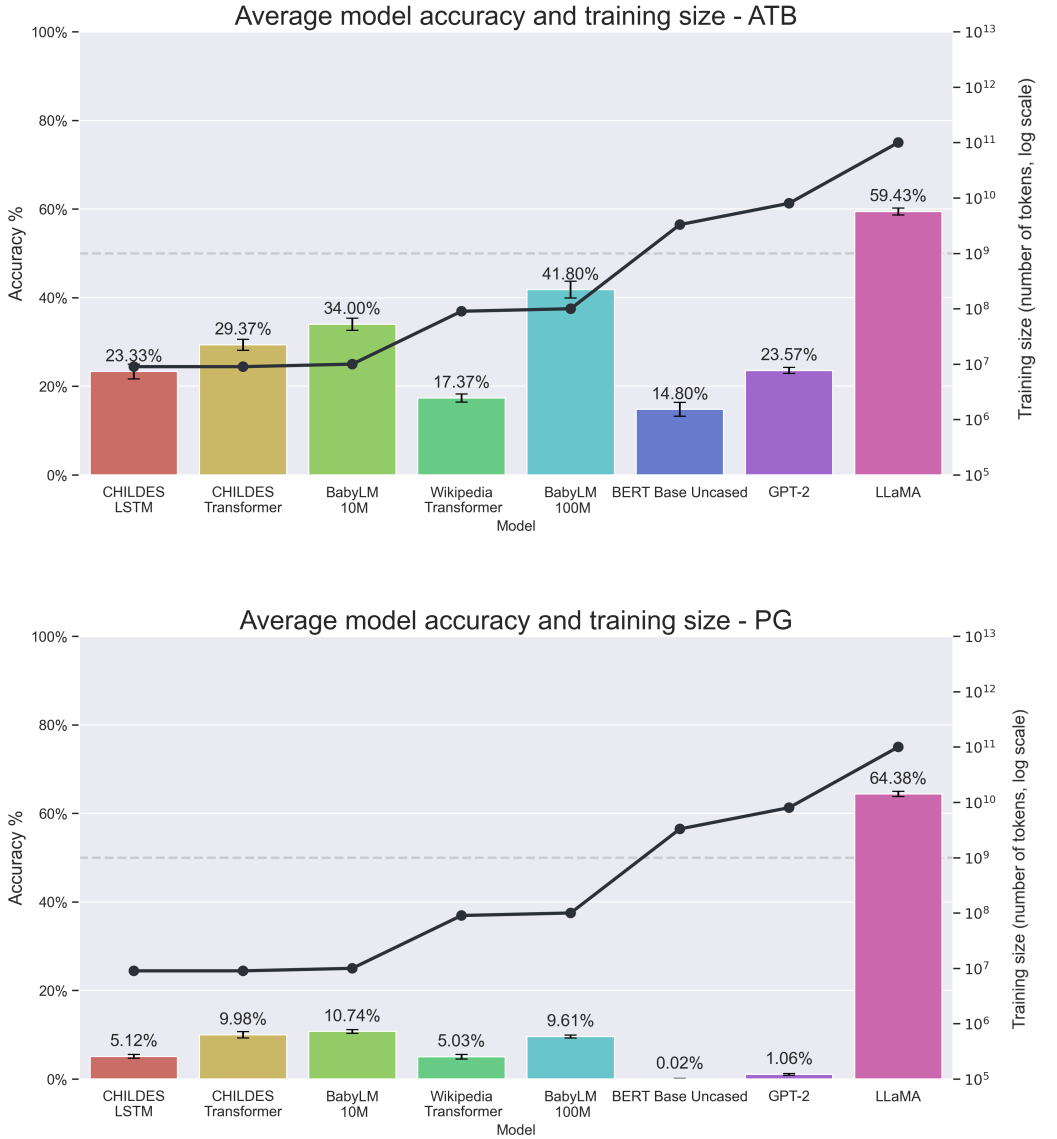


Figure 1: Model accuracy on ATB and PG datasets averaged over five experiment seeds. Accuracy is measured as the proportion of cases where the model assigns a higher probability to the grammatical sentence continuation. The dark plot represents model training sizes (on a logarithmic scale, right)

- c. Who did you say loves Sue?_(-that, +trace)
- d. Who did you say Sue loves ?_(-that, -trace)

This phenomenon, known as *that*-trace effects (TTE), has been the topic of much discussion in the linguistic literature.¹⁰ Building on corpus data from Pearl and Sprouse 2013, Phillips (2013) points out that extraction of any kind from clauses with *that* is vanishingly rare in adult-directed corpora and even rarer in child-directed speech: out of 11,308 analyzed utterances from child-directed speech, there were only two instances of non-subject extraction with *that* and zero instances of subject extraction with *that*, contrasting with 159 instances of object extraction without *that*, and 13 instances of subject extraction without *that*. As Phillips (2013) notes, these distributional observations strongly suggest an APS for the knowledge behind TTE. Still, one might try to defend an H_3 that is not biased in favour of TTE and in which knowledge of TTE is acquired from the input data, perhaps pieced together from several indirect sources of evidence. As with the ATB and PG cases above, we use LLMs as proxies for such a linguistically-neutral H_3 : if these models adequately approximate TTE, this can be taken to suggest that H_3 might succeed in learning TTE, thus weakening the current APS.

We test the same eight LLMs as we did for ATB and PG (see Table 2). The basic TTE paradigm in (6) suggests a condition that, through the minimal pair method, reflects an adequate approximation of TTE: given an overt complementizer upstream, the model should prefer an overt subject (*-trace*) to a gap (*+trace*) in the embedded clause. This preference should translate into probabilities so that $P(\text{loves} \mid \text{Who did you say that})$ in (6a) should be smaller than $P(\text{Sue} \mid \text{Who did you say that})$ in (6b). We thus refer to this condition as $P(+trace \mid +that) < P(-trace \mid +that)$.

Results for the $P(+trace \mid +that) < P(-trace \mid +that)$ criterion are presented in Figure 2.

¹⁰ See Perlmutter 1968, Bresnan 1972, Chomsky and Lasnik 1977, and much subsequent work. See Pesetsky 2017 for a more recent overview and discussion. As with ATB movement and PG, the present discussion is a simplification, and we refer the reader to the literature for observations and analysis that would be important for any attempt to make positive claims about LLM performance and its potential support for H_3 .

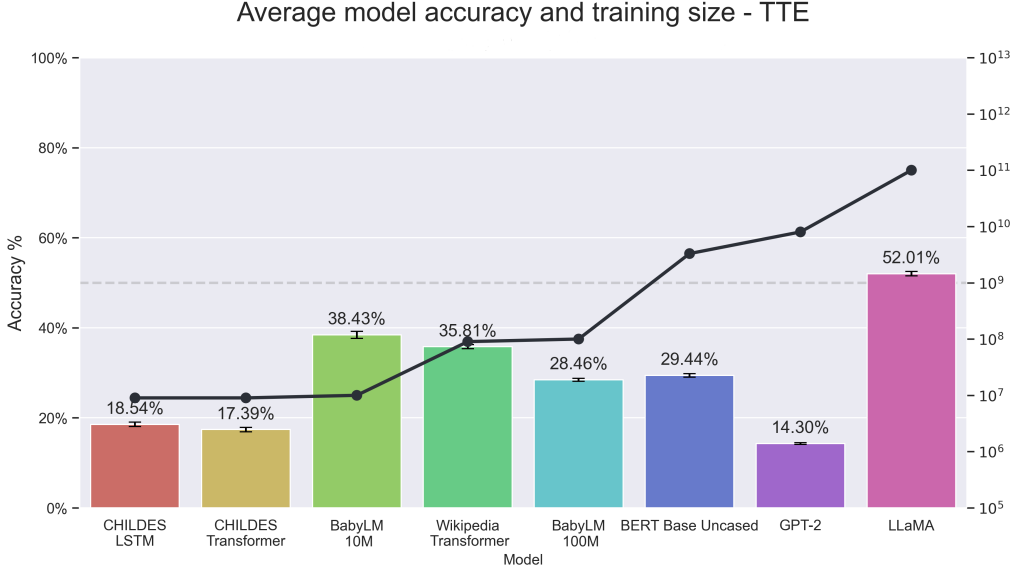


Figure 2: Model accuracy for the $P(+trace \mid +that) < P(-trace \mid +that)$ criterion over samples of 10,000 sentence pairs from TTE test set. Results are averaged over five seeds. The dark plot represents model training sizes (on a logarithmic scale, right)

All models except LLaMA overwhelmingly prefer the ungrammatical element to the grammatical one in the minimal pairs of the test. This is very similar to what we saw for ATB and PG above. It is thus hard for us to see how these results can be used to support the ability of H_3 to acquire knowledge of TTE from a developmentally-realistic corpus.¹¹

¹¹ As a reviewer notes, our results are in tension with the apparent ability of more recent versions of ChatGPT to handle TTE (see Pesetsky 2024). In the absence of sufficient detail by OpenAI about its training process we cannot pinpoint the source of this tension, but we note that the training corpora for current versions of GPT are many orders of magnitude bigger than those used for our models, with the possible exception of LLaMA. As the performance of LLaMA reported here suggests, training on sufficiently large corpora (the equivalent of hundreds of thousands of years of linguistic exposure) may suffice for above chance performance by current models. Moreover, it is conceivable that the pretraining of these current versions includes academic papers that explicitly discuss TTE. It is also possible that at least some of the apparent awareness of these versions to TTE comes from the human-guided stages that follow pretraining.

LARGE LANGUAGE MODELS AND CROSS-LINGUISTIC EVIDENCE

Consider again the comparison of a strongly linguistically-biased generative theory H_1 and a more linguistically-neutral alternative H_3 . Suppose that the strong linguistic biases of the generative theory H_1 can be shown to help derive a cross-linguistic pattern, perhaps by making it easier to represent and learn the languages that adhere to the pattern than those that do not. In the relatively linguistically-neutral theory H_3 it can be much harder to see directly whether the attested patterns are predicted to be easier to learn than the unattested ones. One might try to evaluate this matter empirically, by simulating learning of languages of the attested and the unattested kind given the linguistically-neutral theory, but this can be hard to do. Consequently, it can be hard to see whether the theory can explain the cross-linguistic pattern under consideration and whether it can compete with the generative theory. Again, however, if an LLM is a good proxy for H_3 , the problem is mitigated to some extent. One can examine the ease with which the LLM learns to approximate both languages that follow the pattern under consideration and languages that violate it. If the LLM approximates the former more easily than the latter, this can suggest that the target theory might be able to account for the cross-linguistic pattern (subject to various further assumptions about how learning affects the typology). If, on the other hand, some of the pattern-violating languages are easier to learn than some of the pattern-satisfying ones, the linguistically-neutral theory would need further mechanisms to derive the typology. We outline a preliminary investigation of this kind, building on the setup of Kallini *et al.* (2024). We discuss limitations of this approach in Section 4.6.

4.1 *Ease of learning by LLM and typological asymmetries*

Kallini *et al.* (2024) compare GPT-2 when trained on an English corpus and when trained on corpora that are derived from the original corpus through various perturbations, each perturbed corpus corresponding to an unattested kind of language. Kallini *et al.* find that

the original English corpus was easier for GPT-2 than the derived corpora of languages of unattested kinds on a very particular notion of ease of learning, defined through test set perplexities through time. In doing so, Kallini *et al.* can be taken to make use of LLMs as proxies for a linguistically-neutral H_3 in the context of explaining various cross-linguistic skews. These skews might initially be taken to support an H_1 that is strongly linguistically biased against languages corresponding to the perturbations under consideration. The performance of GPT-2, however, might suggest that a more linguistically-neutral H_3 is also compatible with the typology. On this view, languages corresponding to the perturbations might be unattested because of general, non-linguistic biases that are shared by H_3 and GPT-2.¹²

In the following subsections we extend the setup of Kallini *et al.* 2024 by adding to their examination of English a few other languages for which the evaluation is simple given current resources. We use corpora for these languages as the basis for the perturbations used by Kallini *et al.* (2024) and for one additional perturbation, listed in Table 4. For each perturbation, we use a baseline and a perturbed version for each of the four languages – English, Italian, Russian, and

Perturbation	Attested / Perturbed
<i>partial-reverse</i> (Section 4.2)	Colourless ₀ green ₁ <rev> ₂ ideas ₃ sleep ₄ furiously ₅ . Colourless ₀ green ₁ <rev> ₂ furiously ₅ sleep ₄ ideas ₃ .
<i>full-reverse</i> (Section 4.3)	Colourless ₀ green ₁ <rev> ₂ ideas ₃ sleep ₄ furiously ₅ . Furiously ₅ sleep ₄ ideas ₃ <rev> ₂ green ₁ colourless ₀ .
<i>switch-indices</i> (Section 4.4)	Colourless ₀ green ₁ ideas ₂ sleep ₃ furiously ₄ . Ideas ₂ green ₁ colourless ₀ sleep ₃ furiously ₄ .
<i>token-hop</i> (Section 4.5)	They ₀ were ₁ sleeping ₂ v ₃ next ₄ to ₅ the ₆ colourless ₇ green ₈ ideas ₉ . They ₀ were ₁ sleeping ₂ next ₃ to ₄ the ₅ v ₆ colourless ₇ green ₈ ideas ₉ .

Table 4:
Perturbation test cases.
Ease of learning is
evaluated for the attested
(top, green) and perturbed
(bottom, orange) versions
of English, Italian, Hebrew
and Russian

¹²Kallini *et al.* (2024) contextualize their work differently, relating it to the notion of impossible languages (Moro, 2016; Musso *et al.*, 2003). We will not attempt to determine whether the present work is in line with their goals.

Hebrew, based on datasets from Gulordava *et al.* 2018. We then train a transformer LLM based on the architecture and training scheme from Yedetore *et al.* 2023 and consider the validation set perplexities as training progresses.¹³

As outlined above, if the setup in Kallini *et al.* 2024 is intended to support H_3 , then the behaviour of GPT-2 with respect to their perplexity metric should show that H_3 can be part of a successful account of the typology and that the strong linguistic biases of H_1 are not necessary. A relatively straightforward argument would be if ease of learning directly matched the typological divide *across languages*, with languages of the attested kind being uniformly easier to learn than languages of the unattested kind. But ease of learning may well be just one out of several interacting factors that give rise to the typology; other plausible factors include communicative efficiency, sociolinguistic pressures, and historical accident. Kallini *et al.*'s (2024) setup does not incorporate such factors explicitly, but one can perhaps consider them as part of the background. Even when such factors are taken into consideration, however, a setup such as Kallini *et al.*'s (2024) would presumably only be useful if it showed that ease of learning by the relatively linguistically-neutral H_3 can play some role in an account of the typology. Perhaps, rather minimally, the Kallini *et al.* approach could succeed *within languages*, in that the linguistically-unnatural perturbations of each language would be harder to learn than the attested baseline.

As we show, LLM performance does not provide this kind of support to H_3 . Ease of learning in the sense of Kallini *et al.* (2024) does not match the general typological divide *across languages*, and even *within languages*, the perturbed version is sometimes easier to learn than the attested baseline. From the perspective of the Proxy View, this means that the relatively linguistically-neutral H_3 fails to receive support from the performance of the proxy LLM: if H_3 can be part of an explanation of the typological pattern, this is not shown by the current results.

¹³ The model was trained for 48 hours and validation perplexity was computed every 200 training batches. The 48 hours of training amounted to roughly 1.5 epochs per dataset, with batch size 10, such that each epoch amounted to an average of 181,552 batches. See Appendix B for further details.

We demonstrate the use of ease of learning in our first two test cases, in which we consider reversed constructions. In our first test case of this sort – a cross-linguistic reproduction of the *partial-reverse* perturbation from Kallini *et al.* 2024 – each sentence is reversed starting from a randomly chosen index in the input sentence, in which a special marker token `<rev>` is inserted:

- (7) a. Baseline: Colourless₀ green₁ `<rev>`₂ ideas₃ sleep₄
 furiously₅.
 b. Perturbed: Colourless₀ green₁ `<rev>`₂ furiously₅ sleep₄
 ideas₃.

Reversing linguistic data is generally considered an effective way to generate unnatural data, since it disrupts the language’s consistent word orders (Mitchell and Bowers, 2020), and in the case of *partial-reverse*, also unnaturally entangles dependencies involving constituents that cross the `<rev>` token boundary. The target, relatively linguistically-neutral theory H_3 in this case might be a version of a generative theory H_1 in which there is a parameter for reverse. Any grammar that is statable within H_3 now comes in two variants. If the parameter is set to 0, the grammar works as in the original generative theory. But if the parameter is set to 1, every derivation ends with reversing the output starting from a random point. This may or may not be the target theory that Kallini *et al.* have in mind; other target theories are of course imaginable, and Kallini *et al.* do not comment explicitly on this. We note that, like Kallini *et al.*’s other perturbations, reversal seems quite far away from the typological questions that generative linguistics have usually concerned themselves with. Here we stay with Kallini *et al.*’s perturbations and setup simply in order to illustrate the relevant methodological considerations, but a proper exploration of the Proxy View would need to look at those cross-linguistic patterns that have informed linguistic research.

We find that validation perplexities follow an attested-unattested divide *within languages*: the *partial-reverse* perturbed versions of En-

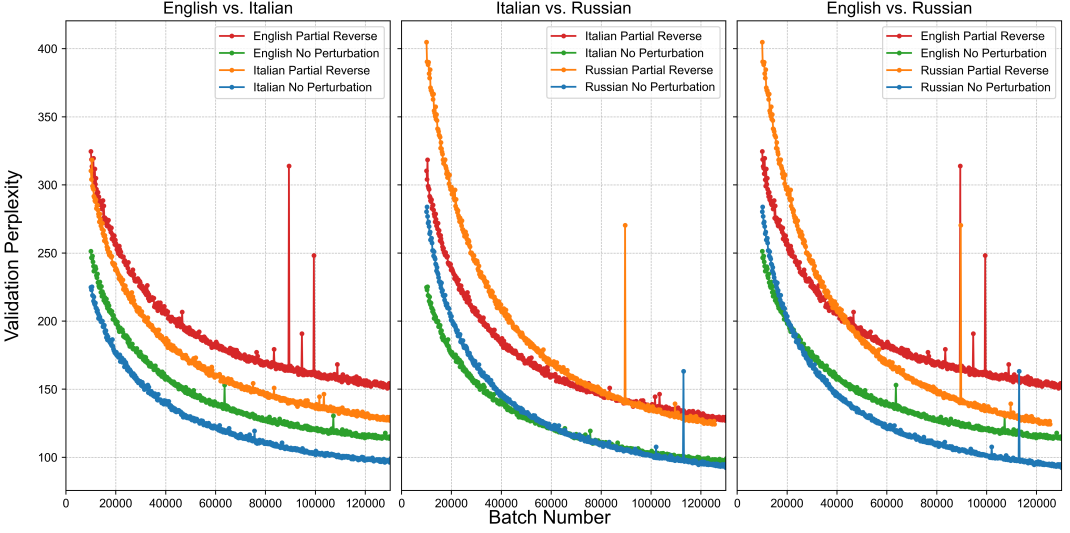


Figure 3: Validation perplexity during training for English, Italian, and Russian and their *partial-reverse* perturbations. The results indicate that the learning curves for attested languages are consistently below those of their *partial-reverse* counterparts

glish, Italian and Russian are harder to learn than their attested counterparts (see Figure 3).¹⁴

This, as discussed above, could be part of an argument that linguistic biases are not needed to rule out partial-reverse languages. Consider an H_3 that is a serious contender as a theory of HLC, addressing the various arguments regarding HLC in the literature. Suppose that H_3 is relatively linguistically neutral in the sense of allowing not just the languages sanctioned by a standard generative H_1 but also their partial-reverse versions. An argument in favour of H_1 and against H_3 might be that these partial-reverse variants are in fact unattested. Our results, following Kallini *et al.* 2024, can be taken as a response to this objection: using GPT-2 as a proxy for H_3 , there is preliminary rea-

¹⁴ The *partial-reverse* languages are compared to a modified baseline of English in which the marker `<rev>` is inserted randomly without reversing, to control for the effect of additional textual material in the sentence. Since we use the same seed for both perturbations, the markers are inserted in the same indices (see (7)).

son to think that attested languages might be slightly easier to learn than their partial-reverse variants, which might be part of an explanation of why the latter are unattested. This, in turn, could allow us to keep H_3 as a contender.

Other perturbations paint a different picture, as we now show.

Test case: *full-reverse*

4.3

We now consider the perturbation *full-reverse* from Kallini *et al.* 2024, where each sentence is reversed in its entirety and a special marker token is randomly inserted:

- (8) a. Baseline: Colourless₀ green₁ <rev>₂ ideas₃ sleep₄ furiously₅.
b. Perturbed: Furiously₅ sleep₄ ideas₃ <rev>₂ green₁ colourless₀.

We find that *across languages*, the perturbed versions of Italian, Russian and Hebrew are projected as easier than attested English (see Figure 4), even though the perturbation is not considered humanly

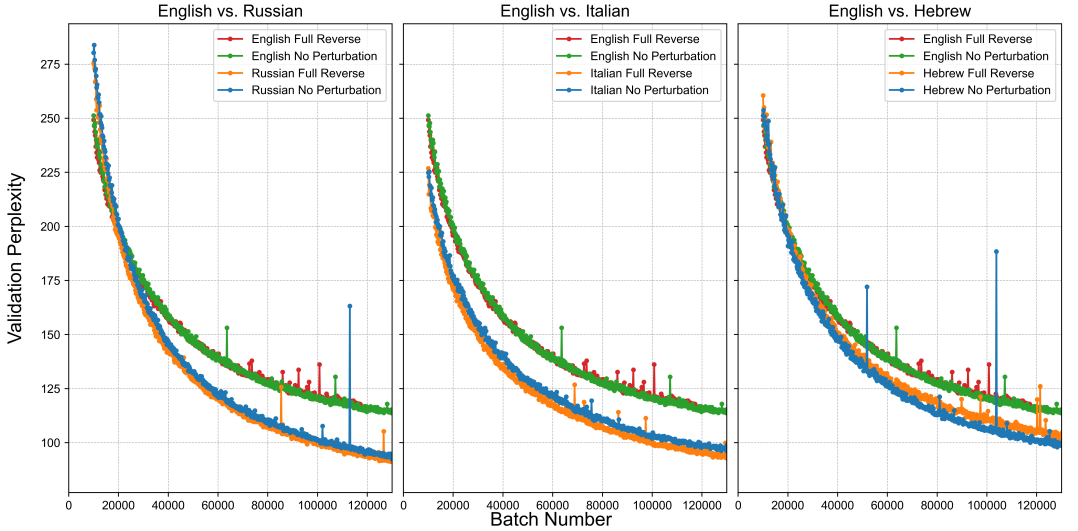


Figure 4: Validation perplexity during training for attested (baseline) and *full-reverse* versions of English, Russian, Italian, and Hebrew

possible. *Within languages*, Italian is projected as harder than its perturbed version. In this case, ease of learning according to Kallini *et al.*'s (2024) perplexity metric does not provide support for the idea that the target theory can explain away the typological pattern through asymmetries of ease of learning.

4.4

Test case: switch-indices

We turn to another typological asymmetry that the Proxy View does not at present explain. Consider *switch-indices*, where the tokens at index 0 and index 2 in every sentence are switched:

- (9) a. Baseline: Colourless₀ green₁ ideas₂ sleep₃ furiously₄.
 b. Perturbed: Ideas₂ green₁ colourless₀ sleep₃ furiously₄.

One can again imagine various target theories that allow for languages like English as well as their switch-index variants. And again the question would be whether such a target theory can account for the typological observation that such switch-index variants are unattested.

If the target theory can account for the typological asymmetry, this is again not detected through the perplexity curve of the LLM proxy: *within languages*, each attested language is wrongly projected as harder than its perturbed version (see Figure 5).

4.5

Test case: token-hop, no-hop

We look at one final typological asymmetry, based on the *token-hop* perturbation employed by Kallini *et al.* (2024). For each of English, Italian, and Russian, we create a perturbed variant by inserting a new marker token three tokens after each verb:¹⁵

- (10) a. no-hop (baseline): They₀ were₁ sleeping₂ v₃ next₄ to₅ the₆ colourless₇ green₈ ideas₉.
 b. token-hop (perturbed): They₀ were₁ sleeping₂ next₃ to₄ the₅ v₆ colourless₇ green₈ ideas₉.

¹⁵New marker tokens were randomly generated as single-character tokens that do not already exist in the model's vocabulary (English: "v", Italian: "v", Russian: "v").

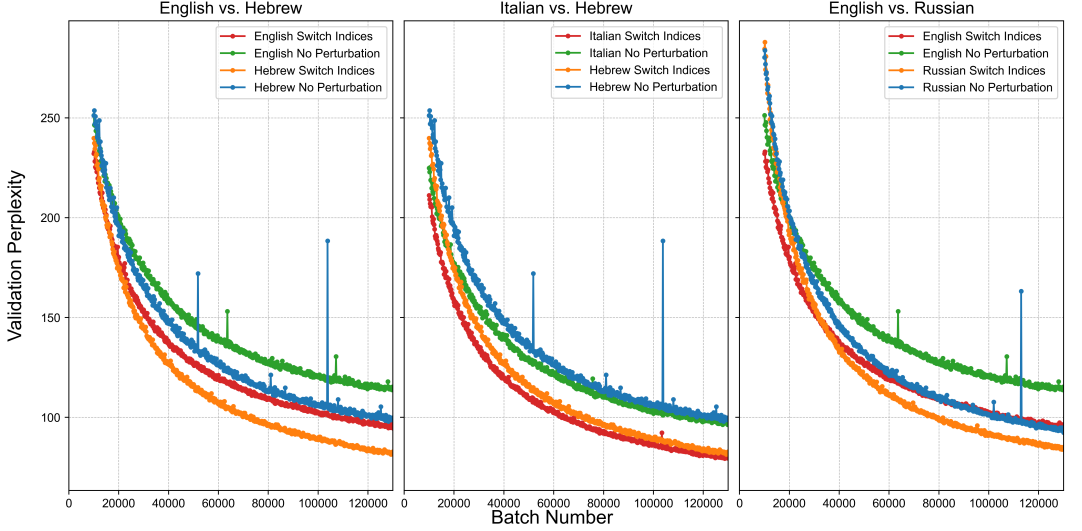


Figure 5: Validation perplexity during training for the attested (baseline) and *switch-indices* versions of English, Italian, Hebrew and Russian

Since additional textual material is inserted, in a way that might confound perplexities over time, we compare the perplexities achieved by our model on *token-hop* datasets to *no-hop* datasets, in which the marker token is inserted exactly after each verb. This controls for the effect of the marker token itself, such that difference in perplexities can be attributed solely to the effect of the token-counting generalization. The *no-hop* version adheres to attested generalizations, as it can be thought of as a post-verbal clitic that marks for parts of speech. Learning *token-hop* on the other hand crucially requires counting tokens, an ability that is particularly non-humanlike, because rules that require counting linear positions in a string violate the strictly hierarchical nature of linguistic generalizations across languages. Although *within languages* the distance of the token from its associated verb has an effect on perplexity, *across languages* we find that the *no-hop* version of English is still harder to learn than *token-hop* Italian and Russian (see Figure 6). Unperturbed English datasets are consistently projected as harder than *token-hop* perturbed versions of Italian and Russian, despite a clear typological skew against counting rules. The same point is made with the validation perplexities of *token-hop* Italian and *no-hop* Russian patterning very similarly.

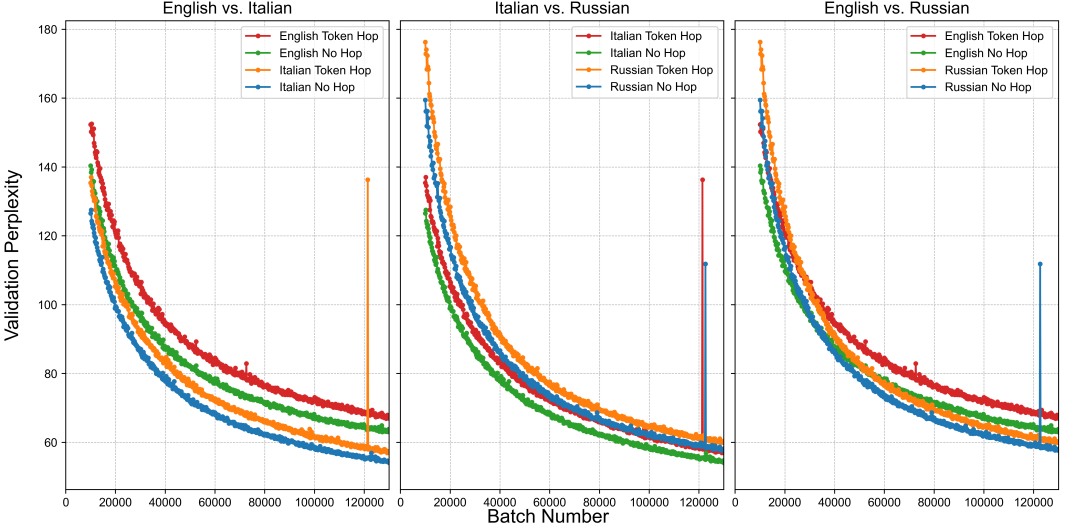


Figure 6: Validation perplexity during training for the *no-hop* (baseline) and *token-hop* versions of English, Italian, and Russian

4.6

Limitations of the typology experiment

In this section, we have considered ease of learning by an LLM as a proxy for a linguistically-neutral theory that can explain typological patterns. It should be emphasized that the inferences from this argument are particularly weak. On the one hand, even if the linguistically-neutral theory H_3 derives a learning asymmetry between pattern-satisfying and pattern-defying languages, this asymmetry might not suffice to account for a robust typological pattern. On the other hand, even if the linguistically-neutral theory H_3 does not derive a learning asymmetry (as was the case with respect to several perturbations above), it might still be able to account for it in terms of other factors, such as communicative pressure. In order to overcome this, one could try to situate the learning component within a model of cultural evolution, which allows for the amplification of small asymmetries over generations (see Kirby *et al.* 2007, Niyogi and Berwick 2009, and Brochhagen *et al.* 2018, among others) and can potentially balance multiple interacting pressures. We believe that such a model is the proper context for the use of ease of learning by LLMs to evaluate typological asymmetries, but following Kallini *et al.*, whose setup we

use here, we leave such a model for future work and only assess currently available models. All we can conclude at present is that, with the possible exception of partial reverse, Kallini *et al.*'s (2024) setup does not currently provide support to the undisclosed H_3 on the Proxy View.¹⁶

CONCLUSION

5

LLMs have been presented as a challenge to more traditional approaches to the study of HLC, and in particular to generative linguistics. A blunt version of this idea is the LLM Theory, promoted by Piantadosi (2024), which maintains that current LLMs are themselves good theories of HLC. Current LLMs can, of course, be viewed as theories of HLC, but this treatment does these models no favours, as the literature was quick to note.

A more cautious approach is the Proxy View, possibly the position of Wilcox *et al.* (2023) and Mahowald *et al.* (2024), which maintains that current LLMs are good proxies – as far as various behavioural measures are concerned – for good theories of HLC that are more linguistically-neutral than generative theories. No commitment to shared essentials between LLMs and the relevant good theories is implied. The Proxy View is of course perfectly coherent, but in contrast to the LLM Theory, it is currently too vague about the theories of HLC that it champions to be of much direct use within empirical science. Empirical science is based on inference to the best explanation and proceeds through competition between reasonably well-understood theories; saying just that one's favourite theory learns roughly as well as an LLM is insufficient for this purpose. One would like to see enough detail at least for a proper comparison of the relevant theory to generative linguistics' theories in light of the discoveries and considerations in the literature of the past 70 years or so. More than anything else, the present paper is a plea for supporters of the Proxy View to provide this kind of detail and evaluation for the theory they have in mind.

¹⁶ For further discussion of Kallini *et al.* 2024 and its limitations see Hunter 2025.

Still, already at this stage we can outline an empirical evaluation of the Proxy View on various kinds of test cases, and here we focused on two: alignment with the stimulus and cross-linguistic variation. The results were not encouraging for the Proxy View. The LLMs failed to even approximate key pieces of knowledge that humans have, even when the models were trained on much larger corpora than children are exposed to. The LLMs also had an easier time approximating various languages of typologically-unattested kinds than actual human languages, at least with respect to a success criterion used in Kallini *et al.* 2024. We conclude that, at least as far as the models and phenomena discussed here are concerned, current LLMs do nothing to suggest changes to linguistic theory.

ACKNOWLEDGEMENTS

We thank Danny Fox, Fred Landman, Roger Levy, Eyal Marco, and audiences at Leibniz ZAS, MIT, and TAU, as well as the reviewers and editors at JLM. RK has been supported by ISF grant #1083/23 and by the Alexander von Humboldt Foundation. This project was provided with computer and storage resources by GENCI at IDRIS thanks to the grant AD011013783R2 on the supercomputer Jean Zay’s V100 and CSL partitions.

APPENDIX: EXPERIMENT DETAILS

A

General Notes

A.1

In all APS experiments, probabilities are evaluated on the critical region in boldface (as in the example sentences in Tables 6, 7, 8). For models listed as unidirectional, the probability value takes into consideration the left context of the evaluated token, whereas for bidirectional models, the probability value factors in both the left and the right context. The bidirectional models we consider were pretrained with masked language modelling (*fill-mask*).

In all generated sentences, the critical region unambiguously disambiguates the sentence as an instance of the phenomenon under consideration (there is no other grammatical reading). To promote lexical variation in target regions, possible target tokens were chosen at random from a POS-tagged English Wikipedia dump. We filtered out potential target tokens that were not part of the vocabulary of one of the models considered in our setup, which are detailed in Table 5.

Table 5: Details of the models introduced in Table 2 and used in the experiments in Section 3

Model Name	Training Size	Directionality	Source
CHILDES LSTM	8.6M	Uni	Yedetore <i>et al.</i> 2023
CHILDES Transformer	8.6M	Uni	Yedetore <i>et al.</i> 2023
BabyLM 10M (<i>strict-small</i>)	10M	Bi	Warstadt <i>et al.</i> 2023
Wikipedia Transformer	90M	Uni	Trained by authors
BabyLM 100M	100M	Bi	Warstadt <i>et al.</i> 2023
BERT Base Uncased	≈3.5B	Bi	Devlin <i>et al.</i> 2019
GPT-2	≈8B	Uni	Radford <i>et al.</i> 2019
LLaMA 3.2-3b	≈9T	Uni	Grattafiori <i>et al.</i> 2024

A.2

ATB and PG

For the ATB and PG experiments, we replicate the experimental setup from Lan *et al.* 2024b with a wider variety of test sentences to increase robustness. We also unify the templates for both phenomena so that they use the same lexical choices where relevant (e.g., proper names). From each template, represented by a context-free grammar, we generated tuples of sentences that correspond to the (+*filler*, +*gap*), (+*filler*, -*gap*) conditions (see CFGs and example sentences in Tables 6 and 7). We then evaluated the success criterion

Table 6:
Excerpt from the
context-free
grammar used
to generate
Across-the-Board
sentences for the
experiments
in Section 3.
Critical regions
are indicated
in boldface.
Spillover regions
relevant only
for bidirectional
models are
indicated
in italics

ATB Grammar
$S \rightarrow \langle \text{PREAMBLE} \rangle \langle F \rangle \langle \pm G \rangle \langle \text{SPILLOVER} \rangle$ $\langle \text{PREAMBLE} \rangle \rightarrow I \text{ know } \langle WH \rangle \langle \text{EMBEDDER} \rangle$ $\langle WH \rangle \rightarrow \text{'which boys' 'which girls'}$ $\langle \text{EMBEDDER} \rangle \rightarrow \text{'you think' 'we believe'}$ $\langle +F \rangle \rightarrow \text{that } \langle \text{NAME1} \rangle \langle \text{VP1} \rangle \langle \text{ADV1} \rangle$ $\langle +G \rangle \rightarrow \langle \text{LINK} \rangle \langle \text{NAME2} \rangle \langle \text{VP2} \rangle \langle \text{ADV2} \rangle$ $\langle -G \rangle \rightarrow \langle \text{LINK} \rangle \langle \text{NAME2} \rangle \langle \text{VP2} \rangle \langle \text{OBJ} \rangle \langle \text{ADV2} \rangle$ $\langle \text{NAME1} \rangle \rightarrow \text{'Bob' 'John'}$ $\langle \text{NAME2} \rangle \rightarrow \text{'Mary' 'Jennifer'}$ $\langle \text{LINK} \rangle \rightarrow \text{'and that'}$ $\langle \text{ADV1} \rangle \rightarrow \text{'shortly' 'eventually'}$ $\langle \text{ADV2} \rangle \rightarrow \text{'soon' 'today' 'now'}$ $\langle \text{VP1} \rangle \rightarrow \text{'will meet' 'will see'}$ $\langle \text{VP2} \rangle \rightarrow \text{'will hug' 'will slap' 'will kiss'}$ $\langle \text{OBJ} \rangle \rightarrow \text{'you' 'us' 'Kim'}$ $\langle \text{SPILLOVER} \rangle \rightarrow \text{'or some other time' 'or next week'}$... $\Rightarrow I \text{ know } \underline{\text{which boys}} \text{ you think that John will meet shortly and that Mary will hug } \textit{\textbf{soon or some other time.}}_{(+\text{filler}, +\text{gap})}$ $\Rightarrow *I \text{ know } \underline{\text{which boys}} \text{ you think that John will meet shortly and that Mary will hug you } \textit{\textbf{soon or some other time.}}_{(+\text{filler}, -\text{gap})}$

PG Grammar
$S \rightarrow \langle \text{PREAMBLE} \rangle \langle \pm F \rangle \langle \pm G \rangle \langle \text{SPILLOVER} \rangle$ $\langle \text{PREAMBLE} \rangle \rightarrow I \text{ know}$ $\langle +F \rangle \rightarrow \text{who} \langle \text{NAME1} \rangle \langle \text{GEN} \rangle \langle \text{NP} \rangle$ $\langle -F \rangle \rightarrow \text{that} \langle \text{NAME1} \rangle \langle \text{GEN} \rangle \langle \text{NP} \rangle \langle \text{NAME2} \rangle$ $\langle +G \rangle \rightarrow \langle \text{LINK} \rangle \langle \text{V} \rangle \langle \text{ADV} \rangle$ $\langle -G \rangle \rightarrow \langle \text{LINK} \rangle \langle \text{V} \rangle \langle \text{OBJ} \rangle \langle \text{ADV} \rangle$ $\langle \text{GEN} \rangle \rightarrow 's$ $\langle \text{NP} \rangle \rightarrow \langle \text{NP_SIMPLE} \rangle \mid \langle \text{NP_COMPLEX} \rangle$ $\langle \text{NP_SIMPLE} \rangle \rightarrow \langle \text{GERUND} \rangle$ $\langle \text{NP_COMPLEX} \rangle \rightarrow \langle \text{N_EMBEDDED} \rangle \text{ 'to' } \langle \text{V_EMBEDDED} \rangle$ $\langle \text{LINK} \rangle \rightarrow \text{'is about to' } \mid \text{'is likely to' } \mid \text{'is going to' } \mid \text{'is expected to'}$ $\langle \text{V} \rangle \rightarrow \text{'bother' } \mid \text{'annoy' } \mid \text{'disturb'}$ $\langle \text{OBJ} \rangle \rightarrow \text{'you' } \mid \text{'us' } \mid \text{'Kim'}$ $\langle \text{GERUND} \rangle \rightarrow \text{'talking to' } \mid \text{'dancing with' } \mid \text{'playing with'}$ $\langle \text{N_EMBEDDED} \rangle \rightarrow \text{'decision' } \mid \text{'intent' } \mid \text{'effort' } \mid \text{'attempt' } \mid \text{'failure'}$ $\langle \text{V_EMBEDDED} \rangle \rightarrow \text{'talk to' } \mid \text{'call' } \mid \text{'meet' } \mid \text{'dance with' } \mid \text{'play with'}$ $\langle \text{ADV} \rangle \rightarrow \text{'soon' } \mid \text{'eventually'}$ $\langle \text{SPILLOVER} \rangle \rightarrow \text{'or some other time' } \mid \text{'or next week'}$... $\Rightarrow I \text{ know } \underline{\text{who}} \text{ John's talking to is going to annoy } \textit{\text{soon or some other time.}}_{(+\text{filler}, +\text{gap})}$ $\Rightarrow * I \text{ know } \underline{\text{who}} \text{ John's talking to is going to annoy } \textit{\text{you soon or some other time.}}_{(+\text{filler}, -\text{gap})}$

Table 7:
Excerpt from the context-free grammar used to generate Parasitic Gap paradigms for the experiments in Section 3. Critical regions are indicated in boldface. Spillover regions relevant only for bidirectional models are indicated in italics

$P(+\text{filler}, -\text{gap}) < P(+\text{filler}, +\text{gap})$ on a sample of 10,000 tuples for each phenomenon. Underlined words alternate according to the $\pm\text{filler}$ condition; words marked in boldface in Tables 6 and 7 denote the position where the $\pm\text{gap}$ condition becomes evident and probability is measured. We evaluate the performance of all eight models from Table 5, including bidirectional ones. To render the Lan *et al.* setup viable for testing bidirectional sentences we added a spillover region at the end of every sentence, following a reviewer’s suggestion.

A.3

TTE

To create test sets, we generate 1,178,496 sentence pairs, out of which we randomly sample 10,000 pairs per experiment. Paradigms are varied lexically in several ways: we first create paradigms that differ structurally from each other, to control for the possibility that our results are an artifact of the specific sentence structure introduced in the example sentences in Table 8. We also introduce lexical variation to all non-target elements, such as the main subject, the embedding verb, the wh-element and the verb inside the wh-clause. Most importantly, lexical choices for target verbs and nouns are varied using 30 target nouns and 30 target verbs. We perform five inference experiments on five such samples and consider average accuracy over all experiments. Pairs are evaluated with the success criterion introduced in Section 3.3.

Table 8: Excerpt from the context-free grammar used to generate That-Trace-Effect paradigms for the experiments in Section 3

TTE Grammar

$S \rightarrow \langle \text{PREAMBLE} \rangle \langle F \rangle \langle \text{EMBEDDING_VERB} \rangle \langle \text{COMP} \rangle \langle \text{TARGET_SUBJ} \rangle \langle \text{TARGET_VERB} \rangle$
 $\langle \text{PREAMBLE} \rangle \rightarrow \text{'He knows' | 'The boy asked' | 'The girl knew' ...}$
 $\langle F \rangle \rightarrow \text{who | } \varepsilon$
 $\langle \text{EMBEDDING_VERB} \rangle \rightarrow \text{'think' | 'believe' | 'say'}$
 $\langle \text{COMP} \rangle \rightarrow \text{that | } \varepsilon$
 $\langle \text{TARGET_SUBJ} \rangle \rightarrow \text{'people' | 'children' | 'parents' | 'scientists' | 'kids' ...}$
 $\langle \text{TARGET_VERB} \rangle \rightarrow \text{'are' | 'have' | 'ate' | 'got' | 'saw' | 'met' ...}$
 ...

$\Rightarrow$ * He knows who you think that **have** children $(+that, +trace)$
 $\Rightarrow$ He knows who you think that **children** have $(+that, -trace)$
 $\Rightarrow$ He knows who you think **have** children $(-that, +trace)$
 $\Rightarrow$ He knows who you think **children** have $(-that, -trace)$

APPENDIX: DATASETS AND MODEL BUILDING

B

Dataset Creation

B.1

Our baseline datasets are based on Wikipedia dumps in English, Italian, Hebrew and Russian from Gulordava *et al.* 2018, who extracted 90M token subsets for each language and switched all tokens that do not belong to the top 50K most frequent tokens into *<unk>* tokens (see Gulordava *et al.* 2018). For each baseline dataset we create perturbed *train*, *test* and *validation* datasets (see Table 4), similarly to Kallini *et al.* 2024, such that each perturbed dataset has the same number of tokens as its corresponding baseline dataset.

The *no-hop* and *token-hop* perturbations require POS tagging, which we performed using the spaCy Python library. Since the POS tagging accuracy is low for Hebrew, we did not perform *no-hop* and *token-hop* experiments on it. New marker tokens were randomly generated as single-character tokens that do not already exist in the model’s vocabulary (English: “v”, Italian: “v”, Russian: “v”).

We used the following spaCy NLP pipelines to perform POS tagging and identify verbs:

1. “en_core_web_sm”, accuracy of POS tagger: 0.97
2. “it_core_news_sm”, accuracy of POS tagger: 0.97
3. “ru_core_news_sm”, accuracy of POS tagger: 0.99

We chose to replace the 3rd person agreement *token-hop* used in Kallini *et al.* 2024 with a POS marker because of the lower accuracy of morphological taggers on languages that are not English and cross-linguistic differences in verb morphology.

Model Training and Evaluation

B.2

We use the optimal Transformer architecture chosen in a hyperparameter search conducted by Yedetore *et al.* (2023), which has four layers, a hidden and embedding size of 800, a batch size of 10, a dropout

rate of 0.2 and a learning rate of 5. Every 200 batches, we evaluate the average perplexity per token of the current trained model on a held out validation set. We set a 48-hour training threshold, during which each model processed a different number of batches. For consistency in comparison graphs, we use the smallest batch count achieved and compare all models up to that point. Table 9 details the number of training batches achieved by the Transformer model on each perturbed dataset.

Table 9:
Number
of training
batches achieved
in 48 hours
of training
on our perturbed
datasets
from Section 4

Language	Perturbation	Size of validation set (batches)	Training batches
English	<i>no-perturb</i>	138430	137600
	<i>full-reverse</i>	143518	117400
	<i>partial-reverse</i>	143517	131800
	<i>switch-indices</i>	166116	205516
	<i>no-hop</i>	181502	189102
	<i>token-hop</i>	181502	142200
Italian	<i>no-perturb</i>	138589	137200
	<i>full-reverse</i>	143840	154200
	<i>partial-reverse</i>	143506	152000
	<i>switch-indices</i>	166307	213507
	<i>no-hop</i>	181552	146400
	<i>token-hop</i>	181552	191752
Russian	<i>no-perturb</i>	138602	136200
	<i>full-reverse</i>	145993	153600
	<i>partial-reverse</i>	145984	125400
	<i>switch-indices</i>	166323	213323
	<i>no-hop</i>	178771	148800
	<i>token-hop</i>	178771	192571
Hebrew	<i>no-perturb</i>	137773	138000
	<i>full-reverse</i>	142884	154800
	<i>partial-reverse</i>	171460	203860
	<i>switch-indices</i>	165328	210128
	<i>no-hop</i>	–	–
	<i>token-hop</i>	–	–

REFERENCES

- Jean-Philippe BERNARDY and Shalom LAPPIN (2017), Using Deep Neural Networks to Learn Syntactic Agreement, *Linguistic Issues in Language Technology*, 15(2), <https://aclanthology.org/2017.lilt-15.3/>.
- Robert C. BERWICK (2018), Revolutionary new ideas appear infrequently, in Norbert HORNSTEIN, Howard LASNIK, Pritty PATEL-GROSZ, and Charles YANG, editors, *Syntactic Structures after 60 Years: The Impact of the Chomskyan Revolution in Linguistics*, pp. 177–194, Walter de Gruyter.
- Robert C. BERWICK, Paul PIETROSKI, Beracah YANKAMA, and Noam CHOMSKY (2011), Poverty of the stimulus revisited, *Cognitive Science*, 35(7):1207–1242.
- Debasmita BHATTACHARYA and Marten VAN SCHIJNDEL (2020), Filler-gaps that neural networks fail to generalize, in Raquel FERNÁNDEZ and Tal LINZEN, editors, *Proceedings of the 24th Conference on Computational Natural Language Learning*, pp. 486–495, Association for Computational Linguistics.
- Joan BRESNAN (1972), *Theory of complementation in English syntax*, Ph.D. thesis, Massachusetts Institute of Technology.
- Thomas BROCHHAGEN, Michael FRANKE, and Robert VAN ROOIJ (2018), Coevolution of lexical meaning and pragmatic use, *Cognitive Science*, 42(8):2757–2789.
- Rui P. CHAVES (2020), What don’t RNN language models learn about filler-gap dependencies?, *Proceedings of the Society for Computation in Linguistics*, 3(1):20–30.
- Noam CHOMSKY (1957), *Syntactic structures*, Mouton.
- Noam CHOMSKY (1971), *Problems of knowledge and freedom: The Russell lectures*, Pantheon Books.
- Noam CHOMSKY and Howard LASNIK (1977), Filters and control, *Linguistic Inquiry*, 8:425–504.
- Noam CHOMSKY, Ian ROBERTS, and Jeffrey WATUMULL (2023), The false promise of ChatGPT, *The New York Times*, <https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html>, accessed: 2025-09-10.
- Shammur Absar CHOWDHURY and Roberto ZAMPARELLI (2018), RNN simulations of grammaticality judgments on long-distance dependencies, in Emily M. BENDER, Leon DERCZYNSKI, and Pierre ISABELLE, editors, *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 133–144, Association for Computational Linguistics.
- Vittoria DENTELLA, Fritz GÜNTHER, and Evelina LEIVADA (2023), Systematic testing of three language models reveals low language accuracy, absence of response stability, and a yes-response bias, *Proceedings of the National Academy of Sciences*, 120(51):e2309583120.

Jacob DEVLIN, Ming-Wei CHANG, Kenton LEE, and Kristina TOUTANOVA (2019), BERT: Pre-training of deep bidirectional transformers for language understanding, in Jill BURSTEIN, Christy DORAN, and Tamar SOLORIO, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Association for Computational Linguistics, doi:10.18653/v1/N19-1423, <https://aclanthology.org/N19-1423/>.

Elisabet ENGDAHL (1983), Parasitic gaps, *Linguistics and philosophy*, pp. 5–34.

Danny FOX (2000), *Economy and semantic interpretation*, MIT Press.

Danny FOX and Roni KATZIR (2024), Large language models and theoretical linguistics, *Theoretical Linguistics*, 50(1–2):71–76.

Aaron GRATTAFIORI, Abhimanyu DUBEY, Abhinav JAUHRI, Abhinav PANDEY, Abhishek KADIAN, Ahmad AL-DAHLE, Aiesha LETMAN, Akhil MATHUR, Alan SCHELTEN, Alex VAUGHAN, et al. (2024), The llama 3 herd of models, *arXiv preprint arXiv:2407.21783*.

Kristina GULORDAVA, Piotr BOJANOWSKI, Edouard GRAVE, Tal LINZEN, and Marco BARONI (2018), Colorless green recurrent networks dream hierarchically, in Marilyn WALKER, Heng JI, and Amanda STENT, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1195–1205, Association for Computational Linguistics, doi:10.18653/v1/N18-1108, <https://aclanthology.org/N18-1108/>.

Isabelle HAÏK (1985), *The syntax of operators*, Ph.D. thesis, Massachusetts Institute of Technology.

Betty HART and Todd R. RISLEY (1995), *Meaningful differences in the everyday experience of young American children*, Paul H. Brookes.

Norbert HORNSTEIN and Jairo NUNES (2002), On asymmetries between parasitic gap and across-the-board constructions, *Syntax*, 5(1):26–54.

Jennifer HU, Jon GAUTHIER, Peng QIAN, Ethan WILCOX, and Roger LEVY (2020), A systematic assessment of syntactic generalization in neural language models, in Dan JURAFSKY, Joyce CHAI, Natalie SCHLUTER, and Joel TETREAULT, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1725–1744, Association for Computational Linguistics, Online, doi:10.18653/v1/2020.acl-main.158.

Philip A. HUEBNER, Elior SULEM, Fisher CYNTHIA, and Dan ROTH (2021), BabyBERTa: Learning more grammar with small-scale child-directed language, in Arianna BISAZZA and Omri ABEND, editors, *Proceedings of the 25th Conference on Computational Natural Language Learning*, pp. 624–646, Association for Computational Linguistics, Online.

Tim HUNTER (2025), Kallini et al. (2024) do not compare impossible languages with constituency-based ones, *Computational Linguistics*, 51(2):641–650.

Julie KALLINI, Isabel PAPADIMITRIOU, Richard FUTRELL, Kyle MAHOWALD, and Christopher POTTS (2024), Mission: Impossible language models, in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14691–14714, Association for Computational Linguistics, doi:10.18653/v1/2024.acl-long.787, <https://aclanthology.org/2024.acl-long.787/>.

Roni KATZIR (2023), Why large language models are poor theories of human linguistic cognition: A reply to Piantadosi, *Biolinguistics*, 17(e13153).

Simon KIRBY, Mike DOWMAN, and Thomas L. GRIFFITHS (2007), Innateness and culture in the evolution of language, *Proceedings of the National Academy of Sciences*, 104(12):5241–5245.

Jordan KODNER, Sarah PAYNE, and Jeffrey HEINZ (2023), Why linguistics will thrive in the 21st century: A reply to Piantadosi (2023), arXiv preprint arXiv:2308.03228.

Adhiguna KUNCORO, Chris DYER, John HALE, Dani YOGATAMA, Stephen CLARK, and Phil BLUNSOM (2018), LSTMs can learn syntax-sensitive dependencies well, but modeling structure makes them better, in Iryna GUREVYCH and Yusuke MIYAO, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1426–1436, Association for Computational Linguistics, doi:10.18653/v1/P18-1132, <https://aclanthology.org/P18-1132/>.

Yair LAKRETZ, German KRUSZEWSKI, Theo DESBORDES, Dieuwke HUPKES, Stanislas DEHAENE, and Marco BARONI (2019), The emergence of number and syntax units in LSTM language models, in Jill BURSTEIN, Christy DORAN, and Tamar SOLORIO, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 11–20, Association for Computational Linguistics, doi:10.18653/v1/N19-1002.

Nur LAN, Emmanuel CHEMLA, and Roni KATZIR (2024a), Bridging the empirical-theoretical gap in neural network formal language learning using minimum description length, in Lun-Wei KU, Andre MARTINS, and Vivek SRIKUMAR, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13198–13210, Association for Computational Linguistics.

Nur LAN, Emmanuel CHEMLA, and Roni KATZIR (2024b), Large language models and the argument from the poverty of the stimulus, *Linguistic Inquiry*, pp. 1–28, ISSN 0024-3892, doi:10.1162/ling_a_00533, https://doi.org/10.1162/ling_a_00533.

- Nur LAN, Michal GEYER, Emmanuel CHEMLA, and Roni KATZIR (2022), Minimum description length recurrent neural networks, *Transactions of the Association for Computational Linguistics*, 10:785–799.
- Tal LINZEN, Emmanuel DUPOUX, and Yoav GOLDBERG (2016), Assessing the ability of LSTMs to learn syntax-sensitive dependencies, *Transactions of the Association for Computational Linguistics*, 4:521–535.
- Miguel LÓPEZ-OTAL, Jorge GRACIA, Jordi BERNAD, Carlos BOBED, Lucía PITARCH-BALLESTEROS, and Emma ANGLÉS-HERRERO (2025), Linguistic interpretability of transformer-based language models: a systematic review, <https://arxiv.org/abs/2504.08001>.
- Kyle MAHOWALD, Anna A. IVANOVA, Idan A. BLANK, Nancy KANWISHER, Joshua B. TENENBAUM, and Evelina FEDORENKO (2024), Dissociating language and thought in large language models, *Trends in Cognitive Sciences*, 28(6):517–540.
- Rebecca MARVIN and Tal LINZEN (2018), Targeted syntactic evaluation of language models, in Ellen RILOFF, David CHIANG, Julia HOCKENMAIER, and Jun’ichi TSUJII, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1192–1202, Association for Computational Linguistics.
- George MILLER and Noam CHOMSKY (1963), Finitary models of language users, in R. Duncan LUCE, Robert R. BUSH, and Eugene GALANTER, editors, *Handbook of Mathematical Psychology*, volume 2, pp. 419–491, Wiley.
- Jeff MITCHELL and Jeffrey BOWERS (2020), Priorless recurrent networks learn curiously, in Donia SCOTT, Nuria BEL, and Chengqing ZONG, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 5147–5158, International Committee on Computational Linguistics, doi:10.18653/v1/2020.coling-main.451, <https://aclanthology.org/2020.coling-main.451/>.
- Andrea MORO (2016), *Impossible languages*, MIT Press, Cambridge, MA, ISBN 9780262034890.
- Andrea MORO, Matteo GRECO, and Stefano F. CAPPÀ (2023), Large languages, impossible languages and human brains, *Cortex*, 167:82–85, doi:10.1016/j.cortex.2023.07.003.
- Alan MUNN (1992), A null operator analysis of ATB gaps, *The Linguistic Review*, 9(1):1–26.
- Mariacristina MUSSO, Andrea MORO, Volkmar GLAUCHE, Michel RIJNTJES, Juergen REICHENBACH, Christian BÜCHEL, and Cornelius WEILLER (2003), Broca’s area and the language instinct, *Nature Neuroscience*, 6(7):774–781, doi:10.1038/nn1077.
- Jon NISSENBAUM (2000), *Investigations of covert phrase movement*, Ph.D. thesis, Massachusetts Institute of Technology.

- Partha NIYOGI and Robert C. BERWICK (2009), The proper treatment of language acquisition and change in a population setting, *Proceedings of the National Academy of Sciences*, 106:10124–10129.
- Satoru OZAKI, Dan YUROVSKY, and Lori LEVIN (2022), How well do LSTM language models learn filler-gap dependencies?, *Proceedings of the Society for Computation in Linguistics*, 5(1):76–88.
- Lisa PEARL and Jon SPROUSE (2013), Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem, *Language Acquisition*, 20(1):23–68.
- David M. PERLMUTTER (1968), *Deep and surface structure constraints in syntax*, Ph.D. thesis, Massachusetts Institute of Technology.
- David PESETSKY (2017), *Complementizer-trace effects*, pp. 1–34, John Wiley & Sons, Ltd, ISBN 9781118358733, doi:10.1002/9781118358733.wbsyncom108, <https://onlinelibrary.wiley.com/doi/10.1002/9781118358733.wbsyncom108>.
- David PESETSKY (2024), Is there an LLM challenge for linguistics? (or is there a linguistics challenge for LLMs?), slides available at ling.auf.net/lingbuzz/008494.
- Colin PHILLIPS (2013), On the nature of island constraints II: Language learning and innateness, in Jon SPROUSE and Norbert HORNSTEIN, editors, *Experimental Syntax and Island Effects*, pp. 132–157, Cambridge University Press.
- Steven T. PIANTADOSI (2024), Modern language models refute Chomsky’s approach to language, in Edward GIBSON and Moshe POLIAK, editors, *From Fieldwork to Linguistic Theory: A tribute to Dan Everett (Empirically Oriented Theoretical Morphology and Syntax 15)*, pp. 353–414, Berlin: Language Science Press, <https://lingbuzz.net/lingbuzz/007180>.
- Paul M. POSTAL (1993), Parasitic gaps and the across-the-board phenomenon, *Linguistic Inquiry*, 24(4):735–754.
- Zhuang QIU, Xufeng DUAN, and Zhenguang G. CAI (2025), Grammaticality representation in ChatGPT as compared to linguists and laypeople, *Humanities and Social Sciences Communications*, 12(1):617.
- Alec RADFORD, Jeffrey WU, Rewon CHILD, David LUAN, Dario AMODEI, Ilya SUTSKEVER, *et al.* (2019), Language models are unsupervised multitask learners, *OpenAI blog*, 1(8):9.
- Daking RAI, Yuchen ZHOU, Shi FENG, Abulhair SAPAROV, and Zhiting YAO (2024), A practical review of mechanistic interpretability for transformer-based language models, *arXiv preprint arXiv:2407.02646*.
- Shauli RAVFOGEL, Grusha PRASAD, Tal LINZEN, and Yoav GOLDBERG (2021), Counterfactual interventions reveal the causal effect of relative clause

representations on agreement prediction, in Arianna BISAZZA and Omri ABEND, editors, *Proceedings of the 25th Conference on Computational Natural Language Learning*, pp. 194–209, Association for Computational Linguistics, Online, doi:10.18653/v1/2021.conll-1.15.

Jon RAWSKI and Lucie BAUMONT (2023), Modern Language Models Refute Nothing, <https://ling.auf.net/lingbuzz/007203>.

John R. ROSS (1967), *Constraints on variables in syntax*, Ph.D. thesis, Massachusetts Institute of Technology.

Dominique SPORTICHE (2024), The Coordinate Structure Constraint: not a constraint on movement, ms., available at ling.auf.net/lingbuzz/007934.

Jon SPROUSE, Beracah YANKAMA, Sagar INDURKHYA, Sandiway FONG, and Robert C. BERWICK (2018), Colorless green ideas do sleep furiously: gradient acceptability and the nature of the grammar, *The Linguistic Review*, 35(3):575–599.

Alex WARSTADT, Aaron MUELLER, Leshem CHOSHEN, Ethan WILCOX, Chengxu ZHUANG, Juan CIRO, Rafael MOSQUERA, Bhargavi PARANJABE, Adina WILLIAMS, Tal LINZEN, and Ryan COTTERELL, editors (2023), *Proceedings of the babylm challenge at the 27th conference on computational natural language learning*, Association for Computational Linguistics, Singapore.

Alex WARSTADT, Alicia PARRISH, Haokun LIU, Anhad MOHANANEY, Wei PENG, Sheng-Fu WANG, and Samuel R. BOWMAN (2020), BLIMP: The benchmark of linguistic minimal pairs for English, *Transactions of the Association for Computational Linguistics*, 8:377–392.

Ethan WILCOX, Roger LEVY, and Richard FUTRELL (2019), What Syntactic Structures block Dependencies in RNN Language Models?, *Proceedings of the Annual Meeting of the Cognitive Science Society*, 41(0), <https://escholarship.org/uc/item/7zr089k7>.

Ethan WILCOX, Roger LEVY, Takashi MORITA, and Richard FUTRELL (2018), What do RNN language models learn about filler–gap dependencies?, in Tal LINZEN, Grzegorz CHRUPAŁA, and Afra ALISHAHI, editors, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 211–221, Association for Computational Linguistics, doi:10.18653/v1/W18-5423, <https://aclanthology.org/W18-5423/>.

Ethan Gotlieb WILCOX, Richard FUTRELL, and Roger LEVY (2023), Using computational models to test syntactic learnability, *Linguistic Inquiry*, pp. 1–44, doi:10.1162/ling_a_00491.

Edwin WILLIAMS (1990), The ATB theory of parasitic gaps, *The Linguistic Review*, 6:265–279.

Edwin S. WILLIAMS (1977), Across-the-board application of rules, *Linguistic Inquiry*, 8(2):419–423.

Aditya YEDETÖRE, Tal LINZEN, Robert FRANK, and R. Thomas MCCOY (2023), How poor is the stimulus? evaluating hierarchical generalization in neural networks trained on child-directed speech, in Anna ROGERS, Jordan BOYD-GRABER, and Naoaki OKAZAKI, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9370–9393, Association for Computational Linguistics, Toronto, Canada.

Victor H. YNGVE (1960), A model and an hypothesis for language structure, *Proceedings of the American Philosophical Society*, 104(5):444–466.

Imry Ziv

Ⓘ 0009-0006-8691-0217
imryziv@mail.tau.ac.il

Department of Linguistics
Tel Aviv University

Nur Lan

Ⓘ 0000-0003-0712-4236
nur.lan@ens.psl.eu

Laboratoire de Sciences Cognitives et
Psycholinguistique
Ecole Normale Supérieure

Emmanuel Chemla

Ⓘ 0000-0002-8423-5880
emmanuel.chemla@ens.psl.eu

Ecole Normale Supérieure, EHESS,
PSL University, CNRS,
Earth Species Project

Roni Katzir

Ⓘ 0000-0002-0241-1896
rkatzir@tauex.tau.ac.il

Department of Linguistics and Sagol
School of Neuroscience, Tel Aviv
University

Imry Ziv, Nur Lan, Emmanuel Chemla, and Roni Katzir, Large language models as proxies for theories of human linguistic cognition, *Journal of Language Modelling*, (Early Access)

This work is licensed under the *Creative Commons Attribution 4.0 Public License*.
ⒸⒻ  <http://creativecommons.org/licenses/by/4.0/>