Biasless Language Models Learn Unnaturally: How LLMs Fail to Distinguish the Possible from the Impossible

Imry Ziv[1], Nur Lan[2,4], Emmanuel Chemla[2,3,4]

[1]Tel Aviv University, [2]Ecole Normale Supérieure,
[3]Earth Species Project, [4]EHESS, PSL University, CNRS

imryziv@mail.tau.ac.il, {nur.lan,emmanuel.chemla}@ens.psl.eu

Abstract

Are large language models (LLMs) sensitive to the distinction between humanly possible and impossible languages? This question was recently used in a broader debate on whether LLMs and humans share the same innate learning biases. Previous work has answered it in the positive by comparing LLM learning curves on existing language datasets and on "impossible" datasets derived from them via various perturbation functions. Using the same methodology, we examine this claim on a wider set of languages and impossible perturbations. We find that in most cases, GPT-2 learns each language and its impossible counterpart equally easily, in contrast to previous findings. We also apply a more lenient condition by testing whether GPT-2 provides any kind of separation between the whole sets of natural vs. impossible languages, based on cross-linguistic variance in metrics derived from the learning curves. Taken together, these perspectives show that GPT-2 provides no systematic separation between the possible and the impossible.

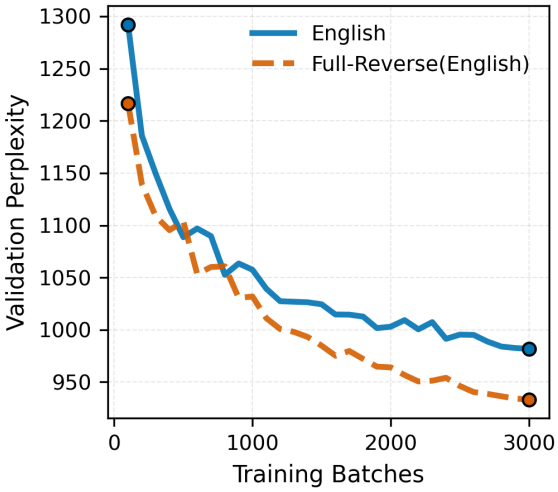

Figure 1: Learning curve of GPT-2 on a standard English dataset vs. a perturbed, impossible variant of the same language. The ease-of-learning methodology used in Kallini et al. (2024) fails to find a stable boundary between possible and impossible languages, often preferring the impossible variants as shown here. See Figure 2 for the full results.

1 Introduction

1.1 Preliminaries

A well-known thought experiment in the field of linguistics is that of Chomsky's Martian linguist (Chomsky, 2000). The idea is as follows: if a Martian linguist were to arrive on Earth, they "might reasonably conclude that there is a single human language, with differences only at the margins" (Chomsky, 2000, p. 7). Despite the seemingly vast differences between the world's languages, all languages are shaped and restricted by the learning biases that humans are biologically endowed with. In this spirit, linguists have discovered patterns which seem to hold across virtually all known natural languages, seeking to uncover what the Martian metaphor has alluded to: the restrictions that define what a human language may potentially be (Baker, 2012). One example of such a pattern is the fact that dependencies in natural language are universally stated in terms of hierarchical representations and never in terms of linear representations (Chomsky, 1965). At the same time, linguists also find that some patterns are consistently absent from all attested languages. That is the case of so-called linguistic "islands" (Ross, 1967), restriction conditions on syntactic movement which apply across virtually every known language. These varied patterns reveal a highly skewed language typology: one in which learners are drawn to very specific kinds of patterns while staying oblivious to otherwise-plausible ones.

Based on this empirical evidence, one would arrive at what we name *the generative hypothesis*: that the set of humanly attested languages is ac-

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 5393–5403

tually a small subset of all potentially observable languages, and that key aspects of linguistic typology are shaped this way because of strong learning biases humans are innately endowed with.[1] Hypothetical languages that do not adhere to these universal constraints came to be known as *impossible languages*. Some have argued that in humans, the notion of impossible languages is biologically embedded, evidenced by selective functional recruitment of different brain regions for the processing of possible languages and impossible ones (Musso et al., 2003, Moro, 2016, Moro et al., 2023).

The recent advent of large language models (LLMs) and their subsequent use as cognitive modeling tools (see Futrell and Mahowald, 2025 for an extensive review) has sparked a similar question regarding such models: are LLMs sensitive to the distinction between possible and impossible languages? Throughout this article we will refer to this issue as *sensitivity to impossibility*. As previously noticed (Chomsky et al., 2023, Moro et al., 2023, Ziv et al., 2025), a negative answer to this question may undermine LLMs as cognitive models for human linguistic cognition, since this distinction is thought to be shaped by human learning biases. On the other hand, a positive answer would provide evidence against *the generative hypothesis*.

In the following sections, we provide experimental support for a negative answer to this question. We build on the methodology from Kallini et al. (2024), which measures *sensitivity to impossibility* by comparing the validation perplexities of GPT-2 models when trained on English datasets and on artificially perturbed versions of these datasets. The perturbation functions generate linguistically impossible variants of existing datasets mainly by performing sentence-level shuffles and reversals (a common method for creating "unnatural" datasets, see Mitchell and Bowers, 2020, and Sinha et al., 2021, among others). While Kallini et al. (2024) provide a positive answer to the question of LLM *sensitivity to impossibility*, we extend their methodology to additional languages and perturbations, and show that LLMs are equally capable of learning possible languages and their perturbed impossible counterparts, in some cases showing a strong preference for the impossible variants. The cross-linguistic methodology allows us to also examine whether the LLM provides any kind of separation between the whole set of possible languages and the whole set of impossible languages. To inform this perspective, we compute basic metrics (minimal perplexity value achieved during training, area under perplexity curve) that reflect the ease-of-learning of the language in question, and then compare the variation in these metrics between possible languages with the variation between the different impossible variants of each language. Here, too, the Kallini et al. (2024) method fails to generalize cross-linguistically, as we find that each possible language patterns with its impossible variants. We conclude that if the perplexity method is indeed a good proxy for ease-of-learning by the LLM, then LLMs do not share the human learning biases that shape linguistic typology and render the perturbed languages impossible.

1.2 Impossible, Unnatural, and Unattested

We note that there has been some ambiguity in the literature regarding the terms *impossible languages*, *unnatural languages*, and *unattested languages*. As pointed out by Xu et al. (2025), there is an important difference between languages that are humanly impossible and languages that are implausible, where the latter merely violate typological tendencies; not every implausible language can meaningfully bear on cognitive issues (one can imagine an unattested version of English in which every sentence is concatenated to itself four times as an absurd example, and see also Footnote 1).

In this work, we will use the term *impossible languages* to refer to hypothetical languages that violate innate biases that are specifically linguistic. This is important since there exist certain typological tendencies that have been attributed to the workings of domain-general learning mechanisms (for example, the prevalence of harmonic word orders, see Culbertson and Newport, 2015), or to the effect of information-theoretical channel biases that guide language evolution, thus shaping the typological landscape (Clark et al., 2023, Futrell and Hahn, 2025). These are the kinds of typologically marked phenomena that have been at the center of work similar but orthogonal to ours, such as Xu et al. (2025). In contrast, the main linguistic bias our work manipulates is the reliance on recursive procedures to generate hierarchical

[1] Not all typological asymmetries stem from the learning biases — they could arise from many functional pressures, such as communicative pressures (Futrell and Hahn, 2022) or historical trajectories (Collins, 2016). We make sure to test only for asymmetries related to the human language device, see Section 1.2.

structures, which is considered at the core of human linguistic cognition (Chomsky, 1965, Hauser et al., 2002). Weaker typological tendencies do not fit into an argument regarding how learning biases in the language faculty shape the typology.

2 Related Work

2.1 The Cognitive Relevance of Large Language Models

The impressive performance of deep learning models on many linguistic benchmarks has fueled growing interest in employing artificial neural networks to model human linguistic cognition (e.g., Linzen and Baroni, 2020, Baroni, 2022, Wilcox et al., 2023, Futrell and Mahowald, 2025). This kind of work usually relies on the assumption that artificial neural networks are proxies for undisclosed theories of human linguistic cognition (Ziv et al., 2025). These proxies are relatively unbiased, in the sense that they were not built with the kinds of biases that linguistics has argued for in order to explain humans' learning processes and outcomes. If the performance of such proxies matches that of humans on key linguistic phenomena, one can argue that this weakens *the generative hypothesis* (although the claim that artificial neural networks are bias-free has been contested, see Chemla and Nefdt, 2024). This has been the argumentation in works such as Wilcox et al. (2023), Mahowald et al. (2024), and Kallini et al. (2024).

There are other objections to the use of LLMs as models of human linguistic cognition: Fox and Katzir (2024) show that LLMs fail to adhere to elementary observations in linguistics, such as the distinction between competence and performance (Yngve, 1960) and between correctness and likelihood (Chomsky, 1957). LLMs are also inconsistent with human learning and processing of certain linguistic phenomena such as syntactic islands (Lan et al., 2024). These findings imply that LLMs may fail to capture patterns that humans learn well, reducing their value as models of human cognition. Here we check whether the LLM proxy evaluated in Kallini et al. (2024) remains robust under further scrutiny on additional languages, perturbations, and statistical tests.

2.2 Large Language Models and Impossible Languages

A particularly interesting line of work within the use of LLMs as models of human linguistic cognition argues that large language models exhibit human-like sensitivity to the distinction between possible and impossible languages. Continuing a line of inquiry concerning the performance of statistical parsers on unnatural language constructions (Fong and Berwick, 2008, Fong et al., 2013), Mitchell and Bowers (2020) have shown that recurrent neural networks (RNNs, Elman, 1990) learn number agreement easily even within impossible constructions such as partially reversed sentences and randomly shuffled vocabularies. Later work (e.g., Abdou et al., 2022) examines the effect of shuffling and disrupting fixed word orders on the performance of Transformer language models (Vaswani et al., 2017) such as BERT (Devlin et al., 2018).

More recently, Kallini et al. (2024) address *sensitivity to impossibility* using an approach we adopt in the current paper: they attempt to show that LLMs' learning curves during training reflect a learning preference for English over languages that were generated from English datasets via various "unnatural" perturbations. It was shown in that work that GPT-2's learning curve (in terms of perplexity on a held-out corpus) when trained on an English dataset remains constantly below the same curve for unattested, artificially-generated variants of English. The unattested languages are generated using perturbation functions that deliberately disrupt tendencies thought to be universal — for example, by reversing or shuffling elements in ways that break constituent structure and disrupt word order. Kallini et al. (2024) took these results to suggest that asymmetries in LLMs' learning curves can account for typological asymmetries – here, the non-existence of certain languages and the prevalence of others – and that by doing so they help discard linguistic biases. Yang et al. (2025) make a similar case while extending this methodology to additional languages, and Xu et al. (2025) consider perturbations such as counterfactual word orders, that are not impossible but typologically implausible.

If Kallini et al.'s (2024) answer to the *sensitivity to impossibility* issue proves true, it would undermine *the generative hypothesis*, considering that LLMs lack the learning biases that humans come equipped with. As will become clearer in the following sections, our work adds to this debate by providing evidence to the contrary: GPT-2 fails to distinguish the possible from the impossible across many novel perturbations and languages,

both interlinguistically and intralinguistically.

2.3 The Intralinguistic and Interlinguistic Perspectives

Using LLM learning curves as proxies for *sensitivity to impossibility* raises two kinds of questions. First, are LLMs sensitive to the distinction between a language and its impossible counterpart? We dub this question the *intralingustic perspective*, and note that this is the perspective that was examined in Kallini et al. (2024), although on English only. The second question pertains to an *interlinguistic perspective*: are LLMs sensitive to the distinction between the set of all attested languages and the set of all unattested ones? One might argue that if LLM learning curves provide an adequate explanation for the typological landscape, this should be reflected in all attested languages' learning curves remaining below all unattested ones' (similarly to recent work by Yang et al., 2025). A key objection to this requirement is that language typology is shaped by factors other than the implicit biases of the learner. To account for this, we test instead a weaker criterion that controls for external factors that might affect the typology: whether the variance in learning curves between attested languages and their unattested counterparts is greater than the variance among attested languages.

3 Methodology

3.1 Impossible Perturbations

We test the two perspectives based on the methodology by Kallini et al. (2024). For each attested language (e.g., Italian) we create perturbed versions of that language which are impossible – for example, a version of Italian in which a syntactic transformation relies on linear word position (e.g., reversing a sentence from the fourth token onwards). The perturbations are listed in Table 1. Each language (either possible or impossible) then serves as a training dataset for a GPT-2 model (Radford et al., 2019).

We expand the experimental coverage in Kallini et al. (2024) in two ways. First, while Kallini et al. (2024) only trained on English (and impossible variants of it), we add eight additional languages: Danish, Finnish, French, German, Greek, Hebrew, Italian, and Russian.

Second, for each attested language we apply the perturbations given in Table 1. SHUFFLE perturbations change the word order in a sentence: SHUFFLE global shuffles each sentence deterministically based on sentence length in tokens, and SHUFFLE local switches each even-indexed token with the following odd-indexed token. The perturbation SWITCH swaps the tokens at indices 0 and 2 (SWITCH and SHUFFLE perturbations are compared to the baseline dataset NO PERTURB).

REVERSE perturbations flip the word order in a sentence: REVERSE full reverses the entire sentence while REVERSE partial reverses the order starting after a randomly inserted <rev> marker. REVERSE languages are compared to a baseline dataset REVERSE baseline where the <rev> token is inserted at the same location without additional changes, to control for the effect of additional textual material on perplexities.

HOP inserts a special marker three tokens after each verb. HOP is compared to a baseline HOP baseline where the same marker is inserted right after the verb, again to control for perplexity effects of the marker.

Note that almost all employed perturbations

Perturbation	Baseline / *Perturbed
SHUFFLE global	Colorless$_0$ green$_1$ ideas$_2$ sleep$_3$ furiously$_4$. *Sleep$_3$ ideas$_2$ colorless$_0$ furiously$_4$ green$_1$
SHUFFLE local	Colorless$_0$ green$_1$ ideas$_2$ sleep$_3$ furiously$_4$. Green$_1$ colorless$_0$ sleep$_3$ ideas$_2$ furiously$_4$.
REVERSE partial	Colorless$_0$ green$_1$ <rev>$_2$ ideas$_3$ sleep$_4$ furiously$_5$. *Colorless$_0$ green$_1$ <rev>$_2$ furiously$_5$ sleep$_4$ ideas$_3$.
REVERSE full	Colorless$_0$ green$_1$ <rev>$_2$ ideas$_3$ sleep$_4$ furiously$_5$. *Furiously$_5$ sleep$_4$ ideas$_3$ <rev>$_2$ green$_1$ colorless$_0$.
SWITCH	Colorless$_0$ green$_1$ ideas$_2$ sleep$_3$ furiously$_4$. *Ideas$_2$ green$_1$ colorless$_0$ sleep$_3$ furiously$_4$.
HOP	They$_0$ were$_1$ sleeping$_2$ **v$_3$** next$_4$ to$_5$ the$_6$ colorless$_7$ green$_8$ ideas$_9$. *They$_0$ were$_1$ sleeping$_2$ next$_3$ to$_4$ the$_5$ **v$_6$** colorless$_7$ green$_8$ ideas$_9$.

Table 1: Impossible perturbations used in our experiments. The perturbations are applied to attested languages to generate impossible variants of them.

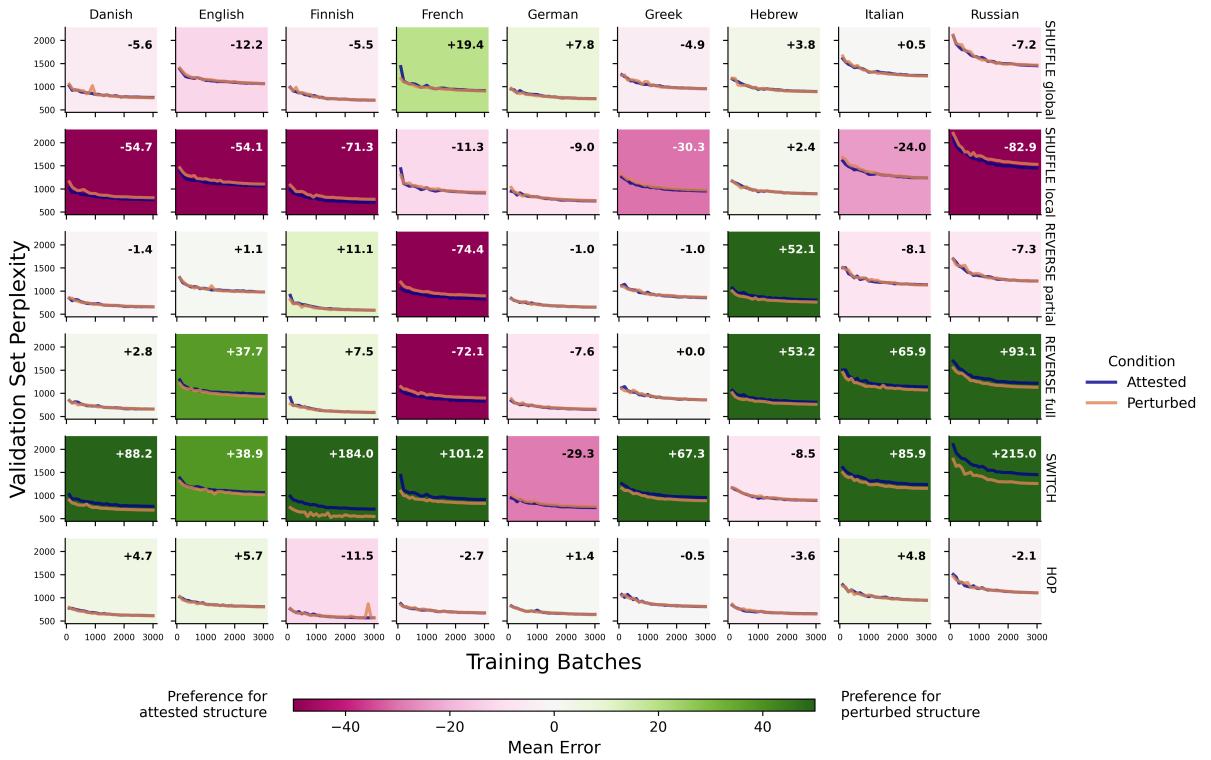

Figure 2: Learning curves for attested languages (dark blue curves) and their perturbed variants (orange). Each subplot displays the learning curves for an experiment alongside mean error values, representing the mean difference between the two learning curves. Positive values (green hues) indicate that the attested language's learning curve is, on average, above its perturbed variant, while negative values (pink hues) indicate the opposite.

disrupt the basic notion of hierarchy in natural language: The SHUFFLE, SWITCH, and REVERSE partial perturbations ignore hierarchical structure completely and break up constituent structure. While not directly at odds with hierarchical structure, the REVERSE full perturbation violates universal linguistic notions such as binding and anaphora (Chomsky, 1981), and the HOP perturbation requires a linear generalization (i.e., counting tokens sequentially) to be learned.

3.2 Baseline Dataset Creation

To create our baseline datasets, we replicate the dataset creation procedure used in Gulordava et al. (2018). We extract recent Wikipedia dumps using WikiExtractor, clean them with TreeTagger (Schmid, 1999) and tokenize them using a simple whitespace tokenizer, uniform for all languages. We then consider 90M token subsets shuffled at sentence level as our final baseline dataset. We set a uniform tokenizer vocabulary size of 50,257 per language, following work such as Arnett and Bergen (2025). Most dumps used were from January 2025, except English, Italian, Russian, and Hebrew for which we used the original datasets created by Gulordava et al. (2018). Our uniform dataset creation method ensures that the differences in perplexities stem only from inherent properties of the languages we consider, and not from unrelated variables such as non-uniform tokenization, sentence lengths, number of unknown tokens, different language registers due to different sources, etc. The languages used were chosen based on the following factors: whether they have a recent Wikipedia dump of sufficient size ($\geq$100M tokens) and whether they have a publicly available part-of-speech (POS) tagger with $\geq$90% accuracy. Alongside these considerations, our choice of languages was also informed by trying to choose languages from distinct language branches. Our choices yield seven distinct branches: North Germanic (Danish), West Germanic (English, German), Romance (French, Italian), Hellenic (Greek), Uralic-Finnic (Finnish), Semitic (Hebrew) and East Slavic (Russian).

3.3 Impossible Dataset Creation

After creating a *train*, *test*, *validation* split for each baseline dataset, we run our perturbation functions (see Table 1) on each split, sentence by sentence,

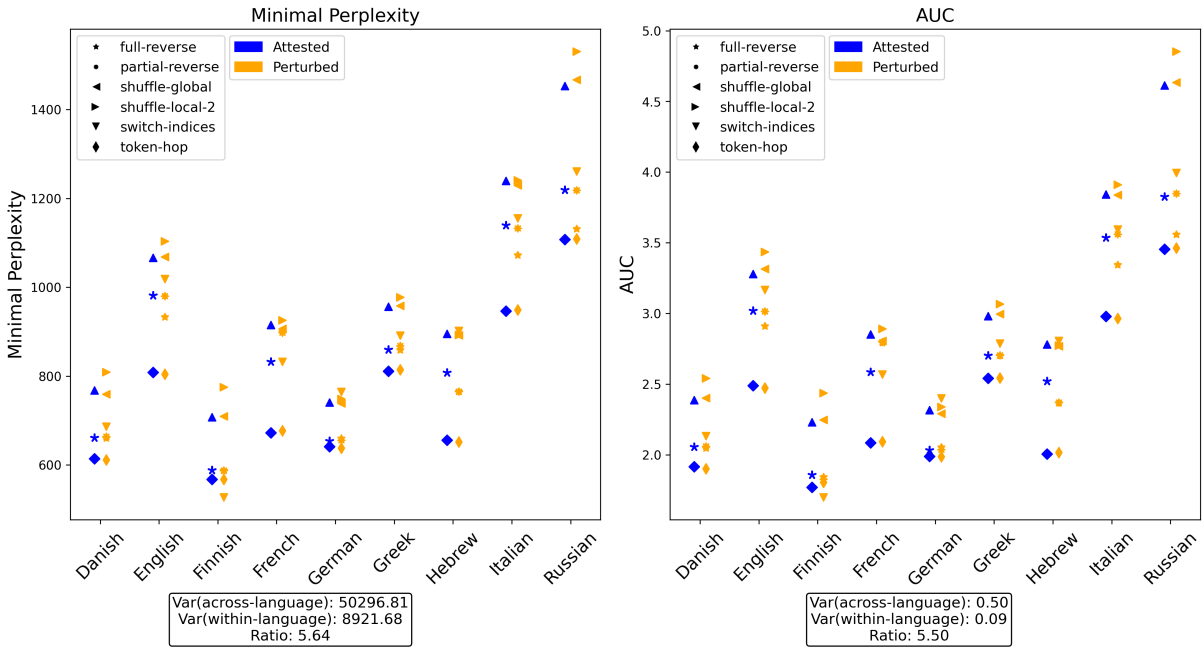

Figure 3: Cross-linguistic comparison of minimal validation perplexity values during training and of area under the curve (AUC) of the training curves from Figure 2. Languages compared to the same baseline have the same shape: triangles are compared to `NO PERTURB` (blue triangle), stars to `REVERSE baseline` (blue star) and diamonds to `HOP baseline` (blue diamond). *across-language* and *within-language* variances are denoted below each subplot. Values vary considerably more across languages (different languages across same perturbation) than within languages (different perturbations across same language). This shows that attested languages and their perturbed, impossible variants pattern together.

similarly to Kallini et al. (2024). To make sure that each sentence of each perturbed dataset has the same number of tokens as its corresponding sentence in the baseline dataset, we insert additional baseline tokens in the baseline variants of perturbations that require additional text (`REVERSE partial`, `REVERSE full`, `HOP`, see Table 1). As mentioned, this controls for the perplexity effect of the insertion of the `REVERSE`/`HOP` markers.

Our `HOP` perturbations require POS tagging, which we mostly performed using the spaCy Python library (see Table 2). For each language we choose a `HOP` marker that is a single-character token that did not previously exist in the model's vocabulary. We choose to replace Kallini et al. (2024)'s 3rd person agreement *token-hop* with POS-based marking because of the lower accuracy of morphological taggers on languages that are not English and cross-linguistic differences in agreement morphology.

3.4 Model Training

We train a GPT-2 model from scratch on the tokenized corpora using HuggingFace's Transformers' *GPT2LMHeadModel*, with a standard configuration: 12 layers, 12 attention heads, 768-dimensional embeddings, and a context size of 1024 tokens. Following Kallini et al. (2024), training uses batch size 512, a learning rate of 5×10^{-4}, and weight decay of 10^{-2}, stopping after 3,000 batches. We record validation set perplexity every 100 batches. Performance is compared by plotting learning curves for each possible/impossible pair. Training is performed using the default Trainer module from the HuggingFace *transformers* Python package. The experiments were run on Nvidia V100 and A40 GPUs. Each perturbation experiment lasted $\sim$ 1 hour, amounting to $\sim$ 81 GPU hours for all languages and perturbations.

3.5 Operationalizing *sensitivity to impossibility*

To test the *intralinguistic perspective*, we plot the perplexity curves for each attested language and its perturbed impossible counterpart. For each such pair we compute mean error (ME):

$$\frac{1}{N}\sum_{i=1}^{N}(B_i - M_i)$$

where B_i are points on the baseline/possible curve, M_i are points on the perturbed/impossible curve, and N is the number of perplexity measurement points. A positive ME is indicative of model failure in explaining the intralinguistic perspective: it means that the baseline curve is consistently above — or coincides with — the perturbed, impossible curve.

To test the *interlinguistic perspective*, we compare the minimum perplexity value achieved during training as well as the area under the curve (AUC) for all possible and impossible languages. We consider the variance in these metrics between possible languages (across-language variance) and between the different variants of each language (within-language variance).[2] If within-language variance is significantly lower than across-language variance, LLMs' learning curves cannot account for typological asymmetries, and the question of *sensitivity to impossibility* in LLMs receives a negative answer.[3]

4 Results

4.1 Intralinguistic Perspective

The learning curves for the different languages are shown in Figure 2. We plot the learning curve for each possible language alongside its impossible variants. In an overwhelming amount of languages, the learning curves coincide almost completely. In cases such as `{english, italian}-REVERSE full` and all `SWITCH` languages except Hebrew and German, the impossible language's learning curve remains below its baseline. Figure 2 also shows that mean error is often positive (green), indicating the model's preference for the impossible structure. Out of 54 test cases, 28 are in the expected direction, for which a negative mean error indicates that the possible curve is on average below its impossible counterpart. A two-sided binomial test comparing this result to chance ($P = 0.5$) yields a p-value of 0.892, indicating no statistically significant deviation from random expectation. The perturbation best explained by the model is `SHUFFLE local`, with 8 out of 9 languages in the right direction, and the worst ones are `REVERSE full` and `SWITCH`, in which the model fails to prefer the attested variant for 7 out of 9 languages. Since GPT-2 does not prefer possible languages across language pairs in a statistically significant manner, it fails to explain *the intralinguistic perspective*.

4.2 Interlinguistic Perspective

Figure 3 plots the minimal perplexity and AUC values for each of the experiments. On the interlinguistic perspective, it is apparent that possible languages and their impossible counterparts pattern together, with a distinct cluster for each attested language and all its perturbations. Russian, Italian, and all their perturbations, for example, are projected as harder than all impossible variants of all other languages. GPT-2 provides no clear-cut separation in these metrics between the set of humanly attested languages (blue data points) and the set of impossibly perturbed ones (orange data points).

Regarding the variances, *across-language* variance compares different languages under the same perturbation, capturing differences in ease-of-learning between languages. *Within-language* variance, by contrast, compares different perturbations of the same language, capturing differences between an attested language and its perturbed counterparts. If the model were to correctly separate the possible from the impossible, we would expect higher *within-language* variances: attested languages should pattern together, while impossible variants should be projected as significantly harder, thereby increasing *within-language* variance.

Our results demonstrate the opposite (see Figure 3). This is not expected if relative ease-of-learning, taken as the relative ordering of learning curves and measured by minimum perplexity and AUC, is an adequate predictor of attested typologies. GPT-2's failure to explain both perspectives suggests that it does not share the human innate biases that shape linguistic typology according to *the generative hypothesis*, and provide no evidence against strong innate biases in human linguistic cognition.

5 Discussion

The rise of LLMs, first within the field of natural language processing (NLP) and later in multiple areas of science, has offered a helpful new tool with which one can revisit questions raised

[2]This operationalization of separability is different from the linear SVM classifier method used in Yang et al. (2025).

[3]All experimental materials, source code, and results are available at https://github.com/imry-ziv/impossible-languages-acl.

by linguists since the 1950's. Recent literature (Kallini et al., 2024, Yang et al., 2025, Xu et al., 2025) has suggested that LLMs show that observations about language typology can be explained without assuming strong linguistic biases, namely, by examining the *sensitivity to impossibility* of Transformer-based language models such as GPT-2. If these models show such sensitivity, it would weaken the position that humans are born with strong innate learning biases, and that these biases are what shapes the possible-impossible frontier in linguistic typology (*the generative hypothesis*).

Here we provide an empirical examination of the claim that shows that so-called ease-of-learning, measured through perplexity curves over training, fails to tease apart attested languages and versions of them that are modified in humanly impossible ways. In most pairs of languages, the LLMs' learning curves actually indicated a dispreference (or indifference) towards existing languages when compared to impossible ones, contrasting with previous results from Kallini et al. (2024) and Yang et al. (2025). Ease-of-learning also fails to create a stable boundary between attested and unattested languages, as exhibited by the interlinguistic perspective in Figure 3. Learning curves thus seem to be an irrelevant measure for deriving language typology, at least for the question of language (non-)existence. This result holds across the variety of languages we tested. Again, this suggests that the biases present in the model architecture and/or training regimes of LLMs are quite different from those that human learners are equipped with.[4]

6 Limitations

While we expand coverage relative to previous work, our experiments rely on a limited number of perturbations and languages, which cannot capture the full diversity of the typological landscape. Future work could apply the same methodologies to a larger set of linguistic phenomena and perturbations to strengthen our conclusions.

Moreover, the perplexity curve metric used in our experiments does not suffice as a complete explanation of typological asymmetries, as there are many other factors that shape the typology other than learning asymmetries and innate biases. Future work could embed the learning component within a model of cultural evolution, in which small learning asymmetries may be amplified across generations (see Kirby et al., 2007, Niyogi and Berwick, 2009, and Brochhagen et al., 2018, among others).

Another limitation is that our experiments rely on GPT-2. This choice was made to allow for direct comparison with the methodology in Kallini et al. (2024) who relied on GPT-2 alone as well. It is possible that the results would not extend to state-of-the-art LLMs. Future work could investigate whether either conclusion holds for larger-scale models.

We also note that the current work differs from previous work on several technical aspects which might explain our different results. First, while Kallini et al. (2024) and Yang et al. (2025) used pretrained tokenizers for each language, we use a uniform whitespace tokenizer with a vocabulary size of 50,257 for all languages. This was meant to unify the different learning tasks. This might have disadvantaged the learning of more morphologically-complex languages, but we do not expect it to affect the results within same-language comparisons. Our setup also differs in hyperparameter, training set size, and training regime choices (e.g., we do not use the learning rate warmup used in Kallini et al., 2024). Admittedly these choices could tip the learning curve patterns in the opposite direction. We note however that as long as the language modeling setup is reasonable, the cognitive linking hypothesis (i.e., the ease-of-learning methodology based on learning curves) should not be so fragile as to break based on hyperparameter choices.

Acknowledgments

The authors would like to thank the TAU Computational Linguistics Lab, the MIT Computational Psycholinguistics Lab, the LINGUAE research group at ENS, and the anonymous ACL reviewers for their helpful comments and discussions. 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.

References

Mostafa Abdou, Vinit Ravishankar, Artur Kulmizev, and Anders Søgaard. 2022. Word order does mat-

[4] A similar stance contesting the Kallini et al. (2024) approach has been taken in recent work, both conceptually (Milway, 2025, Hunter, 2025) and empirically (Leivada et al., 2025).

ter and shuffled language models know it. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6907–6919, Dublin, Ireland. Association for Computational Linguistics.

Catherine Arnett and Benjamin K. Bergen. 2025. Why do language models perform worse for morphologically complex languages? In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*, pages 6607–6622, Mexico City, Mexico. Association for Computational Linguistics.

Mark Baker. 2012. Formal generative typology.

Marco Baroni. 2022. On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. In Shalom Lappin and Jürgen Bernardy, editors, *Algebraic Structures in Natural Language*, pages 1–16. CRC Press / Taylor Francis. Preprint available as arXiv:2106.08694.

Thomas Brochhagen, Michael Franke, and Robert van Rooij. 2018. Coevolution of lexical meaning and pragmatic use. *Cognitive Science*, 42(8):2757–2789.

Emmanuel Chemla and Ryan Nefdt. 2024. No such thing as a general learner: Language models and their dual optimization. Ms.

Noam Chomsky. 1957. *Syntactic Structures*. Mouton, The Hague.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA.

Noam Chomsky. 1981. *Lectures on Government and Binding*. Foris Publications, Dordrecht, Netherlands.

Noam Chomsky. 2000. *New Horizons in the Study of Language and Mind*. Cambridge University Press, Cambridge; New York. Hardback edition, with foreword by Neil Smith.

Noam Chomsky, Ian Roberts, and Jeffrey Watumull. 2023. Noam chomsky: The false promise of chatgpt. *The New York Times*. Accessed: 2025-09-10.

Thomas Hikaru Clark, Clara Meister, Tiago Pimentel, Michael Hahn, Ryan Cotterell, Richard Futrell, and Roger Levy. 2023. A cross-linguistic pressure for uniform information density in word order. *Transactions of the Association for Computational Linguistics*, 11:1048–1065.

Jeremy Collins. 2016. Some language universals are historical accidents. In Karsten Schmidtke-Bode, Natalia Levshina, Susanne Maria Michaelis, and Ilja A. Serant, editors, *Explanation in Typology: Diachronic Sources, Functional Motivations and the Nature of the Evidence*, pages 51–65. Language Science Press.

Jennifer Culbertson and Elissa L. Newport. 2015. Harmonic biases in child learners: In support of language universals. *Cognition*, 139:71–82.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211.

Sandiway Fong and Robert C. Berwick. 2008. Treebank parsing and knowledge of language: A cognitive perspective. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, page 539, Washington, DC. Cognitive Science Society. Paper No. 539.

Sandiway Fong, Igor Malioutov, Beracah Yankama, and Robert C. Berwick. 2013. Treebank parsing and knowledge of language. In Aline Villavicencio, Thierry Poibeau, Anna Korhonen, and Afra Alishahi, editors, *Cognitive Aspects of Computational Language Acquisition*, Theory and Applications of Natural Language Processing, pages 133–172. Springer.

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

Richard Futrell and Michael Hahn. 2022. Information theory as a bridge between language function and language form. *Frontiers in Communication*, 7(657725).

Richard Futrell and Michael Hahn. 2025. Linguistic structure from a bottleneck on sequential information processing. *Nature Human Behaviour*.

Richard Futrell and Kyle Mahowald. 2025. How linguistics learned to stop worrying and love the language models. *Behavioral and Brain Sciences*. Accepted manuscript.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of NAACL 2018*, pages 1195–1205.

Marc D. Hauser, Noam Chomsky, and W. Tecumseh Fitch. 2002. The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298(5598):1569–1579.

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)*, pages 14691–14714, Bangkok, Thailand. Association for Computational Linguistics.

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.

Nur Lan, Emmanuel Chemla, and Roni Katzir. 2024. Large language models and the argument from the poverty of the stimulus. *Linguistic Inquiry*, pages 1–56.

Evelina Leivada, Raquel Montero, Paolo Morosi, Natalia Moskvina, Tamara Serrano, Marcel Aguilar, and Fritz Guenther. 2025. Large language model probabilities cannot distinguish between possible and impossible language. *arXiv preprint*, arXiv:2509.15114. Submitted on 18 Sep 2025.

Tal Linzen and Marco Baroni. 2020. Syntactic structure from deep learning. *Annual Reviews of Linguistics*.

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.

Daniel Milway. 2025. On modern language models, impossible languages and anti-science. LingBuzz preprint. https://ling.auf.net/lingbuzz/008314.

Jeff Mitchell and Jeffrey Bowers. 2020. Priorless recurrent networks learn curiously. In *Proceedings of the 28th International Conference on Computational Linguistics*.

Andrea Moro. 2016. *Impossible Languages*. MIT Press, Cambridge, MA.

Andrea Moro, Matteo Greco, and Stefano F. Cappa. 2023. Large languages, impossible languages and human brains. *Cortex*, 167:82–85.

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.

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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*. Accessed: 2024-11-15.

John R. Ross. 1967. *Constraints on Variables in Syntax*. Ph.D. thesis, MIT, Cambridge, MA.

H. Schmid. 1999. *Improvements in Part-of-Speech Tagging with an Application to German*, pages 13–25. Springer Netherlands, Dordrecht.

Avi Shmidman and Shaltiel Shmidman. 2025. Restoring missing spaces in scraped Hebrew social media. In *Proceedings of the Tenth Workshop on Noisy and User-generated Text*, pages 16–25, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS 2017)*, volume 30, pages 5998–6008. Curran Associates, Inc.

Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. 2023. Using computational models to test syntactic learnability. *Linguistic Inquiry*, pages 1–44.

Tianyang Xu, Tatsuki Kuribayashi, Yohei Oseki, Ryan Cotterell, and Alex Warstadt. 2025. Can language models learn typologically implausible languages? *Preprint*, arXiv:2502.12317.

Xiulin Yang, Tatsuya Aoyama, Yuekun Yao, and Ethan Wilcox. 2025. Anything goes? a crosslinguistic study of (im)possible language learning in LMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26058–26077, Vienna, Austria. Association for Computational Linguistics.

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

Imry Ziv, Nur Lan, Emmanuel Chemla, and Roni Katzir. 2025. Large Language Models as Proxies for Theories of Human Linguistic Cognition. *Preprint*, arXiv:2502.07687.

Appendix

A License Compliance

In this study, we utilize datasets from the Colorless Green RNNs project, which are licensed under the Creative Commons AttributionNonCommercial (CC BY-NC) license. Additionally, we adopt aspects of the methodology described in Kallini et al. (2024), whose code is available under the MIT License. Both the datasets and the methodological framework are duly cited in the main text.

B Models used for POS tagging

Language	Model
Danish	spaCy da_core_news_md
English	spaCy en_core_web_sm
Finnish	spaCy fi_core_news_md
French	spaCy fr_core_news_lg
German	spaCy de_core_news_sm
Greek	spaCy el_core_news_md
Hebrew	DictaBERT (Shmidman and Shmidman, 2025)
Italian	spaCy it_core_news_sm
Russian	spaCy ru_core_news_sm

Table 2: Models used for POS tagging in each experiment.