Anything Goes? A Crosslinguistic Study of (Im)possible Language Learning in LMs

Xiulin Yang[α] **Tatsuya Aoyama**[α] **Yuekun Yao**[β] **Ethan Wilcox**[α]

[α]Georgetown University [β]Saarland University

{xy236, ta571, ethan.wilcox}@georgetown.edu ykyao@coli.uni-saarland.de

Abstract

Do language models (LMs) offer insights into human language learning? A common argument against this idea is that because their architecture and training paradigm are so vastly different from humans, LMs can learn arbitrary inputs as easily as natural languages. We test this claim by training LMs to model impossible and typologically unattested languages. Unlike previous work, which has focused exclusively on English, we conduct experiments on 12 languages from 4 language families with two newly constructed parallel corpora. Our results show that while GPT-2 small can largely distinguish attested languages from their impossible counterparts, it does not achieve perfect separation between all the attested languages and all the impossible ones. We further test whether GPT-2 small distinguishes typologically attested from unattested languages with different NP orders by manipulating word order based on Greenberg's Universal 20. We find that the model's perplexity scores do not distinguish attested vs. unattested word orders, while its performance on the generalization test does. These findings suggest that LMs exhibit some human-like inductive biases, though these biases are weaker than those found in human learners.

1 Introduction

To what extent can language models (LMs) serve as models of human language acquisition and processing? Some, such as Piantadosi (2023), argue that LMs can function as comprehensive linguistic theories, challenging traditional symbolic generative approaches. However, critics maintain that the success of LMs is largely irrelevant to human cognition due to fundamental differences in architecture and learning mechanisms (Chomsky et al., 2023; Fox and Katzir, 2024). Moreover, studies have shown that LMs fail to acquire key aspects of linguistic knowledge, suggesting that they are limited as models of human language (Fox and Katzir, 2024; Lan et al., 2024; Katzir, 2023; Dentella et al., 2024). One central argument in this debate is that LMs are highly flexible learners, capable of acquiring linguistic patterns beyond those learnable by humans, thus making the ability of LMs to learn human languages uninformative for understanding human language acquisition (Chomsky and Moro, 2022; Moro, 2023; Moro et al., 2023).

We present data favoring a more moderate stance, in line with other recent contributions (Futrell and Mahowald, 2025; Millière, 2024; Pater, 2019). Specifically, we present new empirical evidence from the study of *impossible languages* (Kallini et al., 2024) in a multilingual setting, suggesting that LMs exhibit some learning biases that align with certain aspects of human cognition. At the same time, their learning behavior is not universally human-like, suggesting that they have simultaneous biases (or a lack thereof) that diverge from human language processing.

We focus on LMs' abilities to learn different types of languages, both possible (attested or unattested) and impossible (unattested by definition). Specifically, for possible languages, we define **attested languages** as the natural languages spoken by humans (e.g., English, German, and Chinese); **unattested languages** as languages constructed on language universals and identified in typological studies as *never-occurring*. We consider **impossible languages** as those that humans cannot acquire and would never produce. Following Kallini et al. (2024), we select impossible variants as uncontroversial examples of linguistic impossibility, such as languages with shuffled or reversed word orders. To explore unattested languages, we draw from Greenberg's Universal 20 (Greenberg et al., 1963), which identifies unattested word order patterns in noun phrases (e.g., adjective-number-determiner-noun). While there is no direct evidence that such languages are unlearnable, previous studies suggest that typological feature frequencies correlate with

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 26058–26077

learnability in human learners (Culbertson et al., 2020; Gentner and Bowerman, 2009; Saffran et al., 2008).

Regarding impossible language modeling, Kallini et al. (2024) provided initial evidence that GPT-2 small can distinguish between possible and impossible variants of English, suggesting that transformer models encode human-like linguistic biases (Futrell and Mahowald, 2025). However, their study was limited to English, leaving the question of whether this finding generalizes across languages unanswered. Furthermore, their focus on impossible languages leaves the study of unattested languages largely unexplored (although see Xu et al. (2025) for recent work in this area).

This paper is organized around two main research questions: (1) **Does LMs' learning behavior distinguish between attested and impossible languages?** Specifically, (a) Within each attested language, do LMs demonstrate better learning of an attested language compared to its impossible variants? (b) Across different attested languages from multiple language families, do LMs demonstrate better learning of *all* attested languages compared to *all* impossible languages? (2) **Does LMs' learning behavior distibguish between attested and unattested languages?** Specifically, does LMs' ability to model unattested languages align with human typological biases?

Our experiments on two parallel corpora show that GPT-2 is better at language modeling attested compared to impossible languages in most settings, though this distinction weakens for certain locally shuffled variants in some languages (1a). However, the models' learning behavior does not distinguish attested from impossible languages across languages (1b). It assigns lower perplexity to unattested languages with preserved constituency and fixed word order, yet performs better on typologically attested languages in the generalization test (2). These findings suggest that LMs show certain human-like learning biases (e.g., Culbertson et al., 2020), though not full alignment.[1]

2 Related Work

2.1 Language Models & Cognitive Plausibility

Recent advances in deep learning have led to an upsurge in cognitive modeling with artificial neural networks, especially for language (e.g., Wilcox et al., 2023; Borenstein et al., 2024; Kirov and Cotterell, 2018). However, linguists remain divided on whether LMs can meaningfully inform linguistic theories. On the one hand, LMs are limited: They lack the capacity for (compositional) generalization (Yao and Koller, 2022; Kim and Linzen, 2020) and display biases inconsistent with human learning and processing of certain linguistic phenomena (de Dios-Flores et al., 2023; Davis and van Schijndel, 2020; Mitchell and Bowers, 2020). These issues suggest that, beyond functioning as sophisticated probability estimators, LMs have limited use as cognitive models (Cuskley et al., 2024; Bolhuis et al., 2024; Chomsky et al., 2023). Of particular relevance to our study is the argument that LMs can learn patterns that are difficult or even impossible for humans (Chomsky et al., 2023; Moro et al., 2023). This suggests that LMs do not share the cognitive constraints inherent to the human brain and may therefore miss patterns to which humans are naturally biased, rendering them uninformative for understanding human cognition.

On the other hand, LMs have advanced psycholinguistics by serving as highly accurate probability estimators and have already been used to test and refine theories such as Surprisal Theory (Goodkind and Bicknell, 2018; Oh and Schuler, 2023b,a; Kuribayashi et al., 2024), Uniform Information Density (Meister et al., 2021; Tsipidi et al., 2024), and other cognitive-linguistic theories and psychometrics (Pearl and Mis, 2011; Gibson et al., 2019; Kuribayashi et al., 2025). More recently, Kallini et al. (2024); Xu et al. (2025)'s experiments demonstrate that LMs can distinguish between possible and (typologically) impossible languages (Chomsky et al., 2023; Moro et al., 2023) in studies focusing on English and Japanese. These findings provide some empirical counter-evidence to the above arguments.

2.2 Multilingual Language Modeling

Whether languages vary in complexity remains a controversial topic, and linguists have taken different approaches to address this question (e.g., McWhorter, 2001, 2011; Newmeyer, 2021; Joseph and Newmeyer, 2012). While most generative linguists argue that Universal Grammar requires that all languages be equally complex, others have challenged this notion (Gil, 2008).[2]

Initial computational attempts to examine lan-

[1]Our code and data are available at `https://github.com/picol-georgetown/multilingual-LM`.

[2]See Newmeyer (2021) for a more thorough discussion.

guage complexity using LMs were limited to RNN-based architectures (Cotterell et al., 2018; Mielke et al., 2019; Johnson et al., 2021) and n-grams (Koplenig and Wolfer, 2023). These studies suggest that language complexity correlates with morphological richness and the size of speaker populations. More recently, Arnett and Bergen (2025) investigated why morphologically rich languages are harder to model. By testing monolingual LMs trained on carefully curated comparative datasets (Chang et al., 2024), they found that morphological features alone could not predict language learnability when training data size was controlled.

While valuable, previous studies often rely on non-parallel corpora, introducing inconsistencies across languages. Even with parallel corpora (Mielke et al., 2019), studies are limited by small datasets and outdated models. Our study addresses these gaps using a larger parallel corpus and modern transformer architectures.

3 Data and Implementation Details

3.1 Parallel Data Construction: OPUS12 and OPUS30

One challenge in multilingual comparisons is that texts drawn from different sources in different languages will have different amounts of information. To control for this, we construct two sentence-aligned multilingual parallel corpora to ensure that all languages in our dataset match in terms of content. This allows us to isolate the effect of how formal properties of a language might impact its learnability.

We name the two parallel corpora **OPUS12** and **OPUS30**, gathering aligned sentences from five corpora available on OPUS (Tiedemann, 2012): NLLB (Schwenk et al., 2021), TED2020 (Reimers and Gurevych, 2020), the Bible (Christodouloupoulos and Steedman, 2015), OpenSubtitles (Lison and Tiedemann, 2016), and CCAligned (El-Kishky et al., 2020). Since overlap among languages decreases as more languages are included, we decided to select a minimum of 10M words in English as a standard for our parallel corpora. 10M words also correspond to the amount of input of children's first 2 to 5 years of development (Warstadt et al., 2023).

OPUS12 is a 12-language multilingual sentence-aligned corpus[3]. There are around 10M words in the case of English. OPUS30 contains 30 languages with a smaller data size: 48K sentences with 0.7M words. While the two datasets share overlapping languages, their sentences do not overlap, making OPUS30 a suitable test set for additional language modeling experiments.

[3]The languages and their typological information are listed in Appendix C.

Data Source	OPUS12		OPUS30	
	# Sent	# Word	# Sent	# Word
NLLB	5K	0.1M	16	368
TED2020	164K	2.9M	11K	182K
Bible	40K	1M	14K	324K
OpenSubtitles	680K	4.5M	15K	60K
CCAligned	117K	1.6M	8K	111K
Overall	1M	10.1M	48K	0.7M

Table 1: Data sources of OPUS12 and OPUS30. The word counts are based on the English data.

After deduplicating and removing English sentences from non-English data split using FastText (Joulin et al., 2017), we report the statistics of our corpora in Table 1.

3.2 Validation Experiment

To ensure the reliability of our findings presented in the remainder of this paper, we replicate experiments in Kallini et al. (2024) using a scaled-down version of their original corpus (10M words). We find a perfect rank correlation between our results and theirs (Spearman's $\rho = \mathbf{1}, p < 0.001$). More information can be found in Appendix A.

3.3 Model Architecture & Training

In our experiments, following Kallini et al. (2024), we trained standard GPT-2 small models for each language and evaluated its performance based on the geometric mean perplexity over a parallel test split of 10K randomly sampled sentences. Due to limited computational resources, we trained each model using 3 random seeds instead of the 5 used in the original study, reduced the maximum training steps from 2000 to 1200 to avoid overfitting, and adjusted the warmup steps proportionally to 120.[4]

3.4 Multilingual Tokenization

Given our multilingual experiments, tokenization is crucial for fair comparison. To avoid bias toward Latin-script languages, which are overrepresented in our study, we opted against using a multilingual tokenizer with a shared vocabulary.

[4]We did not experiment with alternative warmup steps, as Kallini et al. (2024) demonstrated that changing the warmup schedule does not affect the ranking of perplexities for impossible LMs.

Group	Language	Definition
Ours	SHUFFLE_LOCAL (W=2)	The sentence is reordered with every two tokens reversed in order.
	REVERSE_FULL	Every word is reversed in order in a sentence.
K+	SHUFFLE_DETERMINISTIC (S=84)	The sentence is deterministically shuffled by length with seed 84.
	SHUFFLE_DETERMINISTIC (S=57)	The sentence is deterministically shuffled by length with seed 57.
	SHUFFLE_DETERMINISTIC (S=21)	The sentence is deterministically shuffled by length with seed 21.
	SHUFFLE_LOCAL (W=10)	The sentence is deterministically shuffled in local window size being 10.
	SHUFFLE_LOCAL (W=5)	The sentence is deterministically shuffled in local window size being 5.
	SHUFFLE_LOCAL (W=3)	The sentence is deterministically shuffled in local window size being 3.
	SHUFFLE_EVEN_ODD	The sentence is reordered with even-indexed tokens first, then odd-indexed.

Table 2: Overview of impossible languages in our Experiment1 and Experiment2. K+ languages are borrowed from Kallini et al. (2024) and the rest are new variants introduced in our experiments.

Previous monolingual experiments either set the vocabulary size of tokenizers to be the same across languages (Arnett and Bergen, 2025) or applied the formula $0.4 \times |V|$ (Koplenig et al., 2023; Mielke et al., 2019), where $|V|$ represents the number of unique word types. We conducted a series of pilot experiments on tokenization and found the latter approach unsuitable for our experimental design. Specifically, the large $|V|$ in morphologically rich languages makes it impractical to train a small model with such a large vocabulary size. Details can be found in Appendix B.

Given these considerations, we opted to use pretrained tokenizers. The rationale behind this choice is that when the tokenizer training data is sufficiently large and diverse, the resulting tokenization scheme should be equally good across languages, as long as the tokenizer algorithm and hyperparameters (e.g., vocabulary size, subword strategy) remain the same.[5] While it is difficult to say how *sufficiently large and diverse* a tokenizer training set should be for fair comparison, we consider the size of the training data for GPT-2 (Radford et al., 2019) as a reference point, as English was a high-resource language even in 2019 when the paper was published. We believe that this data size is sufficient to minimize differences that tokenization will make across languages.

One potential concern is that the BPE algorithm might not be optimized for agglutinative languages such as Turkish. However, much literature on cross-linguistic LM comparison adopts BPE tokenizers (e.g., Mielke et al., 2019; Arnett and Bergen, 2025). As an additional check, we use token counts per word (TCW; reported in Appendix E Table 5) to measure the morphological complexity of a language and report the correlation between TCW and our test-set perplexity. The results show the correlation is not significant (see Section5), suggesting that the morphological complexity of a language does not substantially impact its learnability in our experiments.

When selecting pretrained tokenizers, we use **monolingual BPE** tokenizers,[6] targeting a vocabulary size of approximately 50k, with exceptions for Romanian, Arabic, and Chinese due to limited model availability. The training data for all other languages is at least as large as the English corpus. The tokenizer details can be found in Appendix D.

4 Experiment 1: Attested vs. Impossible Languages (Intra-Language)

4.1 Impossible Languages

In this experiment, we use the deterministic shuffled languages from Kallini et al. (2024) along with two new variants (see Table 2). We include shuffled languages because (1) Kallini et al. (2024) identify them as the *most* impossible languages in their language possibility ranking, and (2) their difficulty is also indirectly supported by empirical studies showing that both adults and children exhibit a regularization bias, which can be thought of as a bias *against* shuffling (Newmeyer, 2005; Singleton and Newport, 2004).

Since all languages are deterministically shuffled, the original ones (i.e., attested ones) can be recovered from their variants through another deterministic function. If LMs function as non-human-

[5]Although tokenization quality, measured by metrics like compression (Schmidt et al., 2024) and Rényi entropy (Zouhar et al., 2023), has been linked to language modeling performance (e.g., Liang et al., 2023; Goldman et al., 2024), recent studies challenge this connection (Arnett and Bergen, 2025).

[6]However, for Chinese, we follow previous studies (Mielke et al., 2019) and use the Chinese-BERT tokenizer.

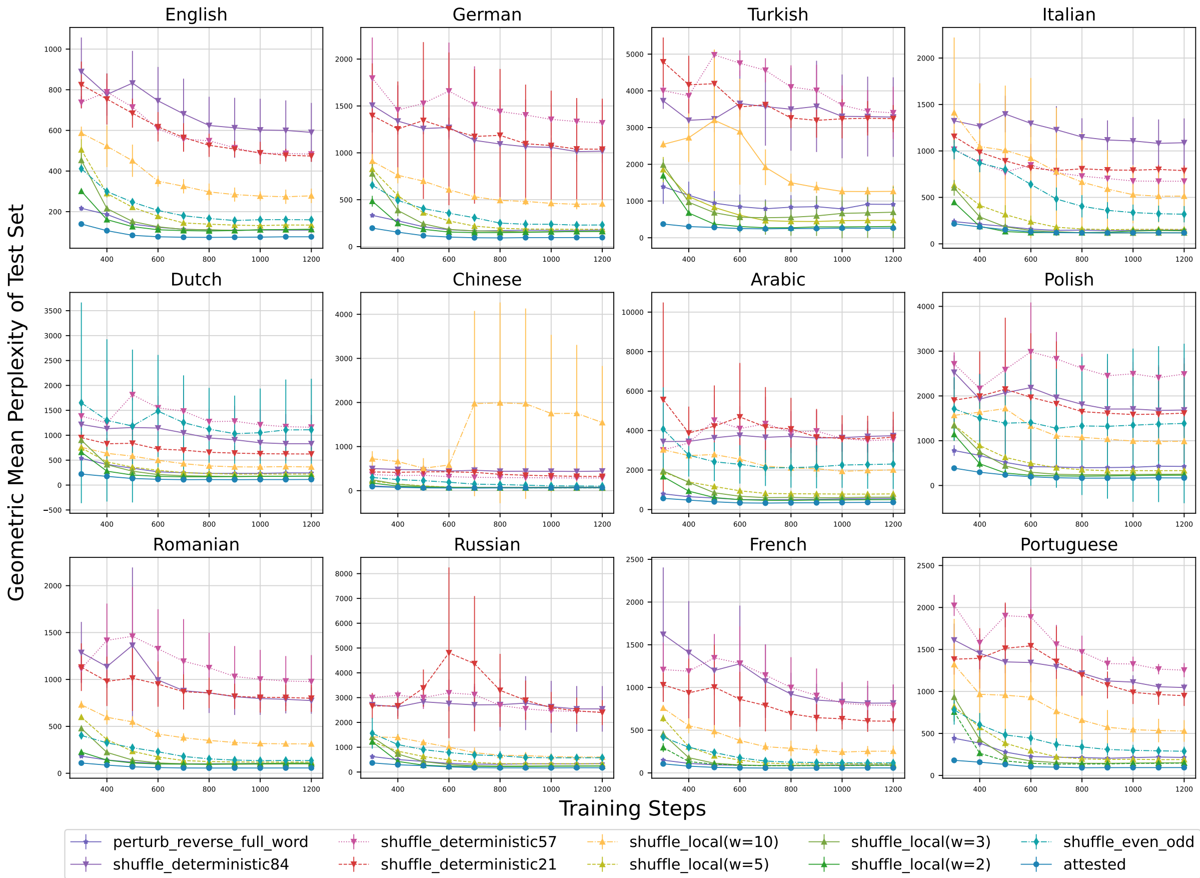

Figure 1: Attested individual Language vs. their corresponding counterparts with a 95% confidence interval over 3 random seeds tested on 10k sentences from OPUS30.

like pattern recognizers as Chomsky et al. (2023); Moro et al. (2023) argue, they should be able to learn these languages as well as attested ones.

4.2 Results & Discussion

The results of this experiment are presented in Figure 1. We note three high-level trends: First, in all languages except Italian, at every checkpoint, the attested language has a lower mean perplexity than all its impossible variants. For Italian, SHUFFLE_LOCAL (W=2) yields a slightly lower perplexity than natural Italian, though the difference is not significant (Mann-Whitney U test: $W = 63$, $p = 0.353$). Welch's t-test with Bonferroni correction across 12 checkpoints shows that for all languages, SHUFFLE_CONTROL differs significantly from other perturbations early in training, but this difference diminishes or becomes insignificant for some languages, especially French, Italian, and Portuguese.[7] Attested languages also show smaller error bars, suggesting more stable learning.

Second, smaller shuffling windows consistently yield lower perplexity. Moreover, SHUFFLE_DETERMINISTIC languages result in higher perplexity than SHUFFLE_LOCAL, likely because they shuffle based on sequence length, which autoregressive models cannot directly access. Third, as a sanity check, a Spearman's rank correlation between OPUS30 English and Kallini et al. (2024)'s results shows strong alignment (see Appendix A).

Based on these findings, we answer the first subquestion: LMs can largely distinguish each attested language from its impossible counterparts by their learning trajectories.

5 Experiment 2: Attested vs. Impossible Languages (Inter-Language)

In this experiment, we pool the results of all possible and impossible languages and investigate whether there is a separation boundary between them. If GPT-2 small can distinguish between possible and impossible languages, we expect its perplexity on the former to be lower than on the latter.

The results of different LMs are shown in Figure 2. The first thing to note is that not every lan-

[7]Details in Appendix G.

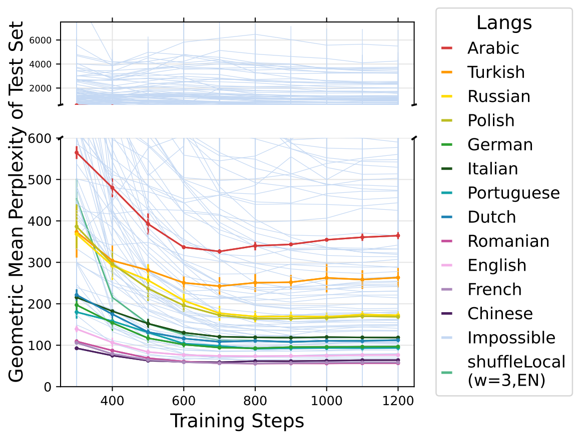

Figure 2: Attested natural languages vs. impossible languages with a 95% confidence interval over 3 random seeds. The x-axis represents the training steps, and the y-axis shows the perplexity on the test split. All the impossible languages are marked in light blue.

guage shows the same perplexity, with Arabic highest and Chinese the lowest.

We observe a moderate positive correlation between the average number of tokens per word (TCW) and perplexity of each of the last checkpoints in 11 languages (Chinese is excluded because the BERT tokenizer is a character-level tokenizer), as indicated by a Spearman's rank test ($\rho = 0.564$), but it is not significant ($p = 0.076$). This finding aligns with the observation by Arnett and Bergen (2025) that there is no significant difference in language modeling difficulty of agglutinative vs. fusional languages when the amount of information is controlled.

Turning to our main research question, although all the attested languages are distributed at the bottom of the graph, we see that some impossible languages fall between these attested languages. For example, Russian, Turkish, and Arabic all show higher perplexity than English perturbed with SHUFFLE_LOCAL (W=3). To quantify the extent GPT-2's perplexity values can separate attested from impossible languages, we train a linear SVM classifier with the perplexity value across the three random seeds of each checkpoint as features. The classifier reaches 0.75 ($sd = 0.08$) macro F1 score averaged over 10-folds cross-validation.

Based on this experiment, we answer the second sub-question posed in our paper: Although LMs tend to learn attested languages better than impossible ones, their perplexity does not distinguish all attested languages from all impossible languages.

6 Experiment 3: Attested vs. Unattested Languages

In this experiment, we investigate how well LMs can learn and generalize to **unattested languages**, languages whose structure is conceivable according to rules of grammar or morphology, but which have not been found to exist. While unattested languages are not necessarily unlearnable (e.g., Tsimpli and Smith, 1995), prior research suggests a link between typological feature frequency, cognitive biases, and language learnability (e.g., Gentner and Bowerman, 2009; Culbertson et al., 2012; Culbertson and Newport, 2015; Culbertson et al., 2020).

We focus on Greenberg's Universal 20 (Greenberg et al., 1963), which suggests that certain determiner-adjective-number-noun orders in an NP are universally unattested. Culbertson and Newport (2015, 2017); Culbertson et al. (2020) find that harmonic NP orders (i.e., ones where the dependents always all either precede or follow the head; e.g., NUM-ADJ-NOUN and NOUN-ADJ-NUM) are easier to learn than non-harmonic ones (e.g., NUM-NOUN-ADJ or ADJ-NOUN-NUM) for humans. One influential hypothesis, the Typological Prevalence Hypothesis, proposes that more common typological patterns are easier to learn (Gentner and Bowerman, 2009). Therefore, if LMs exhibit similar biases as humans, a gradient of difficulty is expected in learning different NP orders, with some unattested configurations posing greater challenges than others.

Among the 24 theoretically possible orders of adjectives, nouns, determiners, and numbers, we select five combinations, covering cases classified as FEW, MANY, and ZERO in Cinque (2005)'s typological analysis.[8] In this experiment, we only permute words within NPs. If the perplexity of these permuted languages is similar to that of attested languages, it suggests two possible reasons: (1) LMs can learn these unattested languages; (2) NPs may be small (in terms of number of tokens) with respect to the entire data size, and hence NP-internal perturbation introduces less noise compared to the entire data perturbation of the previous experiments. To rule out the latter possibility, we also construct a control condition in which words corresponding to these POS categories are randomly shuffled within

[8] Although Cinque (2005) seeks to explain why ZERO languages really are "underivable" under the minimalist program we refer to them as *unattested* to contrast them with the impossible languages of the previous section, i.e., ones that involve shuffling or reversed word order.

Langs	Attested		Example
	Typo.	Theo.	
PERTURB_NNDA	NO	NO	She enjoyed books three the fantastically interesting a lot .
PERTURB_ANND	NO	NO	She enjoyed fantastically interesting three books the a lot .
PERTURB_DANN	FEW	YES	She enjoyed the fantastically interesting books three a lot .
DPERTURB_DNAN	MANY	YES	She enjoyed the three fantastically interesting books a lot .
PERTURB_DNNA	MANY	YES	She enjoyed the books three fantastically interesting a lot .
NP_RANDOM	NO	NO	She enjoyed books fantastically three interesting the a lot .

Table 3: List of NP-perturbations with corresponding categories and examples. *Typo* refers to *typologically*-attested, while *Theo* refers to *theoretically*-attested by Cinque (2005)'s analysis.

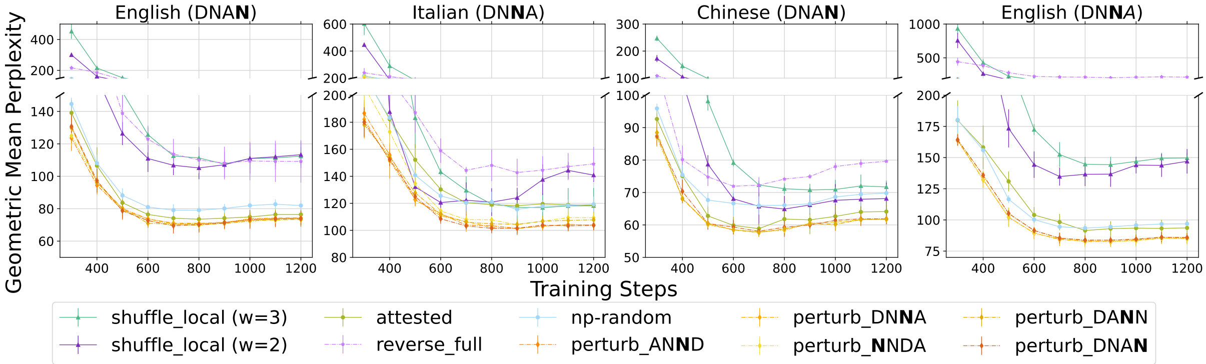

Figure 3: Attested natural languages vs. their corresponding unattested languages with a 95% confidence interval over 3 random seeds. The x-axis represents the training steps, and the y-axis shows the perplexity on the test split. Different language types are distinguished using distinct color palettes.

NPs. This language serves as a baseline, indicating the extent to which NP-internal permutations influence the learnability of a language.

Examples of perturbed NP word orders and their typological information are listed in Table 3 and their word orders are reported below:

- PERTURB_NNDA: NOUN>NUM>DET>ADJ.
- PERTURB_ANND: ADJ>NUM>NOUN>DET.
- PERTURB_DANN: DET>ADJ>NOUN>NUM.
- PERTURB_DNAN: DET>NUM>ADJ>NOUN, typical of English and Chinese.
- PERTURB_DNNA: DET>NUM>NOUN>ADJ, typical of Italian and Portuguese.
- NP_RANDOM: Random permutation of ADJ, NOUN, NUM, and DET within NPs.

Since identifying NP structures requires a constituency parser, we use Stanza (Qi et al., 2020) to parse raw text. Stanza provides constituency parsing for only Chinese, Portuguese, English, and Italian, with acceptable accuracy (>0.85)[9], so we limit our analysis to these four languages. As different parsers are trained on distinct treebanks with varying annotation guidelines, we select POS tags based on each treebank's guidelines. Details are provided in Appendix F.

Studies such as Xu et al. (2025) suggest a difference between models' perplexity and results of targeted evaluations. Therefore, we additionally conduct a targeted test to assess how well LMs trained on different perturbed languages generalize. Specifically, we propose ΔGenScore to quantify their generalization ability, measured across a test corpus of n sentences, and defined as:

$$\Delta\text{GenScore} = \text{GenScore}_{\checkmark} - \text{GenScore}_{\text{✗}} \tag{1}$$

$$\text{GenScore}_{\checkmark} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{P_{\checkmark}(s_{\checkmark,i}) > P_{\checkmark}(s_{\text{✗},i})\}$$

$$\text{GenScore}_{\text{✗}} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{P_{\text{✗}}(s_{\text{✗},i}) > P_{\text{✗}}(s_{\checkmark,i})\}$$

where GenScore$_{\checkmark}$ refers to the generalization score of a model trained on **attested (natural) languages**, while GenScore$_{\text{✗}}$ refers to the generalization score of a model trained on **unattested (perturbed) languages**. More specifically, for each test case, we form a minimal pair consisting of an original version $s_{\checkmark,i}$ and its perturbed sentence $s_{\text{✗},i}$. Let $P_{\checkmark}$ denote the probability assigned by a model trained on attested languages and $P_{\text{✗}}$ the

[9]https://stanfordnlp.github.io/stanza/constituency.html

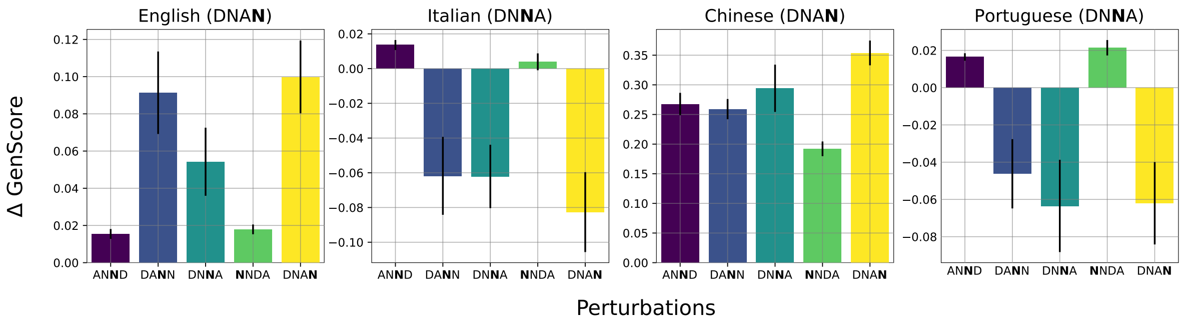

Figure 4: Mean Δ GenScore across four languages under five NP perturbations. Error bars indicate the 95% CI computed over three random seeds. Positive Δ GenScore indicates better generalization for models trained on attested languages, while negative values indicate better generalization for models trained on unattested languages.

probability assigned by a model trained on unattested languages. Then, GenScore$_{\checkmark}$ is the proportion of cases where $P_{\checkmark}(s_{\checkmark,i}) > P_{\checkmark}(s_{\text{✗},i})$, while GenScore$_{\text{✗}}$ is the proportion where $P_{\text{✗}}(s_{\text{✗},i}) > P_{\text{✗}}(s_{\checkmark,i})$. We extract attested sentences from the same test set used for perplexity evaluation but include only those with at least one perturbed NP. The minimal pair test is conducted using the last checkpoint of each language model.

If a model assigns higher probability to natural (i.e., attested) word orders *regardless of its training data*, then it would obtain a ΔGenScore of 1. Likewise, if it assigns a higher probability to unattested orders regardless of its training data, then it would have a ΔGenScore of -1. A ΔGenScore of 0 indicates that the model always assigns higher probabilities to sequences that match its training data. Therefore, we interpret positive ΔGenScore values as indicating better generalization for natural word orderings, and negative scores as indicating better generalization for perturbed orderings. We use ΔGenScore to investigate models trained on each of our natural languages, and compare them to each of our possible NP perturbations.

6.1 Results

Perplexity Our results (Figure 3, bottom subgraph) show that shuffling POS tags within NPs increases perplexity, often matching or exceeding that of attested languages. This rules out the possibility that limited perturbations do not affect model training. Surprisingly, all five NP-perturbed languages exhibit lower perplexity than their attested counterparts across all four languages, though the differences are not significant for Italian, Chinese, and Portuguese (by a Welch's t-test with Bonferroni correction).[10] No significant difference is observed between languages with attested (i.e., DANN, DNAN, and DNNA) and unattested NP orders (i.e., NNDA and ANND) either, indicating a lack of human alignment in language learning bias.

Generalization Test The results from this experiment are visualized in Figure 4 and present a mixed picture. Two observations emerge. First, models trained on NNDA and ANND, the two typologically absent orderings, consistently yield positive ΔGenScore across all languages. This indicates poorer generalization of models trained on unattested patterns than models on attested ones listed in Table 3. Second, ΔGenScore remains positive for all five NP perturbations in English and Chinese but shows mixed results for Italian and Portuguese. Since English and Chinese predominantly follow the DNAN order and Italian and Portuguese follow DNNA, this suggests models trained on DNAN orders generalize more consistently. This finding, if confirmed, supports Culbertson and Newport (2015)'s report of human biases toward harmonic languages. However, for stronger conclusions, further investigation with more typologically diverse languages and NP perturbations is needed.

Summary Experiment 3 shows that while LMs do not reflect a gradient of difficulty measured by perplexity in learning different NP orders based on typological prevalence, they may exhibit human-aligned generalization patterns for typologically unattested languages in the generalization test. The differences between perplexity and targeted evaluation results are consistent with Xu et al. (2025)'s findings, which show similar discrepancies.

[10]For English, there is no significant difference between SHUFFLE_CONTROL and DNAN, the dominant NP order in English.

6.2 Discussion of Perplexity Results

Why doesn't LM perplexity distinguish between attested and unattested languages? We propose two key factors that influence LM learning outcomes: *randomness* and *constituency structure*. By *randomness*, we refer to whether the perturbation function produces a perturbed text that can be deterministically recovered to its original form. By *constituency structure*, we mean whether the phrase structures of the original language are preserved in the perturbed version.

Regarding randomness, as LMs are simply *distributions over strings* (Borenstein et al., 2024), introducing randomness increases unpredictability of the text, thus increasing the entropy of the sequence. This explains why NP-perturbed unattested languages show lower perplexity than attested languages and NP_RANDOM variants. The reasoning is that our perturbation procedure enforces a strict ordering procedure, which may be (sometimes) violated in the original attested language. For example, although English is a DNAN language, certain constructions such as the DANN (DET-ADJ-NUM-NOUN; e.g., *a beautiful five days*) does not follow the dominant pattern. Once POS tag orders are normalized within NPs, the resulting constructions become more predictable. Therefore, all normalized NPs, including our unattested NPs, may have lower overall entropy, which could explain why they are easier to learn. In fact, the normalized DNAN, which has the same typical word order as English, shows lower perplexity than the original, unnormalized English; and the same applies to our other languages in this experiment.

Regarding constituency structure, we hypothesize that disrupting constituency weakens local dependency relations within phrase structures. This explains why in experiments 1 and 2, all LMs' perplexities for impossible languages are higher than for NP-perturbed languages, despite maintaining a deterministic order (Figure 3). Similarly, this may also explain the higher perplexity of count-based grammars in Kallini et al. (2024), where the authors insert a morphological marker a certain number of words or tokens after a host word. The count-based insertion may disrupt phrase structure integrity.[11]

In sum, this discussion points to a potential confound in our experiments: although the texts are parallel in content, languages with normalized NP structures may have lower entropy. In this case, even if LMs learn all languages equally well, lower entropy would naturally lead to lower perplexity. Future work could control for entropy across NP-perturbed languages to test whether perplexity differences persist.

7 Discussion & Conclusion

In this paper, we extend Kallini et al. (2024) to a broader multilingual context using two new parallel corpora. Our findings complement their work, suggesting that models have a preference for human-like languages, although the preference is somewhat gradient and depends on the testing setup. First, while GPT-2 small assigns lower perplexity to attested languages compared to their impossible variants, the difference is sometimes not significant, especially later in training. Second, the model does not fully separate all attested from all unattested or impossible languages, but it *does* generally learn attested languages better, achieving a separability of 0.75 between the two classes based on perplexity. In the NP word order experiments, some unattested languages exhibit lower perplexity than their attested counterparts, despite having orderings that violate Greenberg's Universal 20. However, when assessed using targeted evaluation methods, a more promising pattern emerges: GPT-2 seems to favor typologically attested, as opposed to unattested NP variants, and shows some preference for harmonic word orderings.

What to make of these results in the context of our original question–whether LMs can serve as cognitive models? While our results show that GPT-2 does not behave as we might expect from a fully human-like learner, they also demonstrate that it has a soft preference for attested over impossible languages. Skeptics have previously linked LMs to a bad theory of physics in which "anything goes."[12] In line with Kallini et al. (2024), our results demonstrate that these models do not instantiate an "anything goes" hypothesis. Rather, their incremental data-processing architectures represent a useful starting point for studying human language processing and learning. Refining models to better align with humans is possible and will likely lead to lasting insights about human cognitive architecture.

[11] We do not replicate these count-based experiments.

[12] Chomsky, quoted from an email to Gary Marcus: *"You can't go to a physics conference and say: I've got a great theory. It accounts for everything and is so simple it can be captured in two words: 'Anything goes.' All known and unknown laws of nature are accommodated, no failures. Of course, everything impossible is accommodated also."*

8 Limitations

We acknowledge that our experiments rely on GPT-2 Small, which may not generalize to larger models. This choice was made for two reasons: (1) running experiments across multiple languages is computationally expensive; (2) we aimed for comparability with Kallini et al. (2024). Future work could explore whether our findings hold for larger models or similarly sized models with different architectures. Additionally, the dataset used for training the language model is relatively small. This is a deliberate trade-off between data size and linguistic diversity. While a larger dataset might yield more robust results, our approach ensures broader typological coverage. In our experiments on unattested languages, we generated synthetic data by perturbing languages based on Universal 20. However, linguistic correlations extend beyond word order universals. For instance, Greenbergian correlations (Dryer, 1992) suggest that verb-object order often correlates with other features such as adposition-noun phrase order and determiner-noun phrase order. Future work will explore more nuanced perturbations to better capture such cross-linguistic dependencies. Lastly, the data we use has not been manually checked yet. It is possible that our parallel corpora include noise that might influence the learning results.

9 Ethics Statement

We use publicly available datasets, ensuring that no private or personally identifiable information is included. Our dataset selection prioritizes linguistic diversity while maintaining data transparency. Regarding computational resources, we use GPT-2 small trained on A-100 and V-100 GPUs. Each experiment on each language took around 10-12 hours.

10 Acknowledgments

We thank Amir Zeldes, Nathan Schneider, Yilun Zhu, Dan DeGenaro, Wesley Scivetti, and all members of PICoL and NERT for their helpful feedback and support.

References

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. AraGPT2: Pre-trained transformer for Arabic language generation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 196–207, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Catherine Arnett and Benjamin Bergen. 2025. Why do language models perform worse for morphologically complex languages? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6607–6623, Abu Dhabi, UAE. Association for Computational Linguistics.

Florbela Barreto, António Branco, Eduardo Ferreira, Amália Mendes, Maria Fernanda Bacelar do Nascimento, Filipe Nunes, and João Ricardo Silva. 2006. Open resources and tools for the shallow processing of Portuguese: The TagShare project. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Johan J Bolhuis, Stephen Crain, Sandiway Fong, and Andrea Moro. 2024. Three reasons why AI doesn't model human language. *Nature*, 627(8004):489.

Nadav Borenstein, Anej Svete, Robin Chan, Josef Valvoda, Franz Nowak, Isabelle Augenstein, Eleanor Chodroff, and Ryan Cotterell. 2024. What languages are easy to language-model? a perspective from learning probabilistic regular languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15115–15134, Bangkok, Thailand. Association for Computational Linguistics.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. 2024. Goldfish: Monolingual language models for 350 languages. *Preprint*, arXiv:2408.10441.

Noam Chomsky and Andrea Moro. 2022. *The secrets of words*. MIT Press.

Noam Chomsky, Lan Roberts, and Jeffrey Watumull. 2023. Noam Chomsky: The false promise of ChatGPT. *The New York Times*. Accessed: 2024-12-16.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the Bible in 100 languages. *Language resources and evaluation*, 49:375–395.

Guglielmo Cinque. 2005. Deriving Greenberg's Universal 20 and its exceptions. *Linguistic inquiry*, 36(3):315–332.

Ryan Cotterell, Sabrina J. Mielke, Jason Eisner, and Brian Roark. 2018. Are all languages equally hard to language-model? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 536–541, New Orleans, Louisiana. Association for Computational Linguistics.

Jennifer Culbertson, Julie Franck, Guillaume Braquet, Magda Barrera Navarro, and Inbal Arnon. 2020. A learning bias for word order harmony: Evidence from speakers of non-harmonic languages. *Cognition*, 204:104392.

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

Jennifer Culbertson and Elissa L Newport. 2017. Innovation of word order harmony across development. *Open Mind*, 1(2):91–100.

Jennifer Culbertson, Paul Smolensky, and Géraldine Legendre. 2012. Learning biases predict a word order universal. *Cognition*, 122(3):306–329.

Christine Cuskley, Rebecca Woods, and Molly Flaherty. 2024. The limitations of large language models for understanding human language and cognition. *Open Mind*, 8:1058–1083.

Forrest Davis and Marten van Schijndel. 2020. Recurrent neural network language models always learn English-like relative clause attachment. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1979–1990, Online. Association for Computational Linguistics.

Iria de Dios-Flores, Juan Garcia Amboage, and Marcos Garcia. 2023. Dependency resolution at the syntax-semantics interface: psycholinguistic and computational insights on control dependencies. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 203–222, Toronto, Canada. Association for Computational Linguistics.

Rodolfo Delmonte, Antonella Bristot, and Sara Tonelli. 2007. VIT-Venice Italian Treebank: Syntactic and quantitative features. In *Sixth International Workshop on Treebanks and Linguistic Theories*, volume 1, pages 43–54. Northern European Association for Language Technol.

Vittoria Dentella, Fritz Guenther, and Evelina Leivada. 2024. Language in vivo vs. in silico: Size matters but larger language models still do not comprehend language on a par with humans. *arXiv preprint arXiv:2404.14883*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew S Dryer. 1992. The Greenbergian word order correlations. *Language*, 68(1):81–138.

Stefan Dumitrescu. 2024. GPT-Neo Romanian 780M. Hugging Face. Available at: `https://huggingface.co/dumitrescustefan/gpt-neo-romanian-780m`.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pages 5960–5969, Online. Association for Computational Linguistics.

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

Richard Futrell and Kyle Mahowald. 2025. How linguistics learned to stop worrying and love the language models. *arXiv preprint arXiv:2501.17047*.

Dedre Gentner and Melissa Bowerman. 2009. Why some spatial semantic categories are harder to learn than others: The typological prevalence hypothesis. *Crosslinguistic approaches to the psychology of language: Research in the tradition of Dan Isaac Slobin*, 465:480.

Edward Gibson, Richard Futrell, Steven P Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. 2019. How efficiency shapes human language. *Trends in cognitive sciences*, 23(5):389–407.

David Gil. 2008. How complex are isolating languages. *Language*.

Omer Goldman, Avi Caciularu, Matan Eyal, Kris Cao, Idan Szpektor, and Reut Tsarfaty. 2024. Unpacking tokenization: Evaluating text compression and its correlation with model performance. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2274–2286, Bangkok, Thailand. Association for Computational Linguistics.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.

Joseph H Greenberg et al. 1963. Some universals of grammar with particular reference to the order of meaningful elements. *Universals of language*, 2:73–113.

Yeb Havinga. 2023. GPT Neo 1.3B pre-trained on cleaned Dutch mC4. Accessed: 2024-12-10.

iGeniusAI. 2024. Italia-9b-instruct-v0.1. `https://huggingface.co/iGeniusAI/Italia-9B-Instruct-v0.1`. Accessed: 2024-06-17.

Tamar Johnson, Kexin Gao, Kenny Smith, Hugh Rabagliati, and Jennifer Culbertson. 2021. Investigating the effects of i-complexity and e-complexity on the learnability of morphological systems. *J. Lang. Model.*, 9.

John E Joseph and Frederick J Newmeyer. 2012. 'All languages are equally complex'. *Historiographia linguistica*, 39.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

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.

Roni Katzir. 2023. Why large language models are poor theories of human linguistic cognition: A reply to Piantadosi. *Biolinguistics*, 17:1–12.

H. Toprak Kesgin, M. Kaan Yuce, Eren Dogan, M. Egemen Uzun, Atahan Uz, H. Emre Seyrek, Ahmed Zeer, and M. Fatih Amasyali. 2024. Introducing cosmosGPT: Monolingual training for Turkish language models. In *2024 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, page 1–6. IEEE.

Najoung Kim and Tal Linzen. 2020. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105, Online. Association for Computational Linguistics.

Christo Kirov and Ryan Cotterell. 2018. Recurrent neural networks in linguistic theory: Revisiting Pinker and Prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, 6:651–665.

Alexander Koplenig and Sascha Wolfer. 2023. Languages with more speakers tend to be harder to (machine-) learn. *Scientific Reports*, 13(1):18521.

Alexander Koplenig, Sascha Wolfer, and Peter Meyer. 2023. A large quantitative analysis of written language challenges the idea that all languages are equally complex. *Scientific Reports*, 13(1):15351.

Tatsuki Kuribayashi, Yohei Oseki, and Timothy Baldwin. 2024. Psychometric predictive power of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1983–2005, Mexico City, Mexico. Association for Computational Linguistics.

Tatsuki Kuribayashi, Yohei Oseki, Souhaib Ben Taieb, Kentaro Inui, and Timothy Baldwin. 2025. Large language models are human-like internally. *arXiv preprint arXiv:2502.01615*.

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.

Julien Launay, E.l. Tommasone, Baptiste Pannier, François Boniface, Amélie Chatelain, Alessandro Cappelli, Iacopo Poli, and Djamé Seddah. 2022. PAGnol: An extra-large French generative model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4275–4284, Marseille, France. European Language Resources Association.

Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabsa. 2023. XLM-V: Overcoming the vocabulary bottleneck in multilingual masked language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13142–13152, Singapore. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Ricardo Lopes, Joao Magalhaes, and David Semedo. 2024. GlórIA: A generative and open large language model for Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 441–453, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

John H McWhorter. 2001. The worlds simplest grammars are creole grammars. *Linguistic Typology*, 5(2-3):125–166.

John H McWhorter. 2011. *Linguistic simplicity and complexity: Why do languages undress?*, volume 1. Walter de Gruyter.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the Uniform Information Density hypothesis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 963–980, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sabrina J. Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2019. What kind of language is hard to language-model? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4975–4989, Florence, Italy. Association for Computational Linguistics.

Raphaël Millière. 2024. Language models as models of language. *arXiv preprint arXiv:2408.07144*.

Jeff Mitchell and Jeffrey Bowers. 2020. Priorless recurrent networks learn curiously. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5147–5158, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Andrea Moro. 2023. Embodied syntax: impossible languages and the irreducible difference between humans and machines. *Sistemi intelligenti*, 35(2):321–328.

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

Frederick J Newmeyer. 2005. *Possible and probable languages*. Oxford: Oxford.

Frederick J Newmeyer. 2021. Complexity and relative complexity in generative grammar. *Frontiers in Communication*, 6:614352.

Byung-Doh Oh and William Schuler. 2023a. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1915–1921, Singapore. Association for Computational Linguistics.

Byung-Doh Oh and William Schuler. 2023b. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Malte Ostendorff. 2023. Gpt2-xl-wechsel-german. https://huggingface.co/malteos/gpt2-xl-wechsel-german.

Joe Pater. 2019. Generative linguistics and neural networks at 60: Foundation, friction, and fusion. *Language*, 95(1):e41–e74.

Lisa Pearl and Benjamin Mis. 2011. How far can indirect evidence take us? Anaphoric one revisited. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33.

Steven T Piantadosi. 2023. Modern language models refute Chomsky's approach to language. *From fieldwork to linguistic theory: A tribute to Dan Everett*, pages 353–414.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

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.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Jenny Saffran, Marc Hauser, Rebecca Seibel, Joshua Kapfhamer, Fritz Tsao, and Fiery Cushman. 2008. Grammatical pattern learning by human infants and cotton-top tamarin monkeys. *Cognition*, 107(2):479–500.

Craig W Schmidt, Varshini Reddy, Haoran Zhang, Alec Alameddine, Omri Uzan, Yuval Pinter, and Chris Tanner. 2024. Tokenization is more than compression. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 678–702, Miami, Florida, USA. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Jenny L Singleton and Elissa L Newport. 2004. When learners surpass their models: The acquisition of American sign language from inconsistent input. *Cognitive psychology*, 49(4):370–407.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Ianthi-Maria Tsimpli and Neil Smith. 1995. *Mind of a Savant: Language, Learning and Modularity*. Blackwell.

Eleftheria Tsipidi, Franz Nowak, Ryan Cotterell, Ethan Wilcox, Mario Giulianelli, and Alex Warstadt. 2024. Surprise! Uniform Information Density isn't the whole story: Predicting surprisal contours in long-form discourse. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18820–18836, Miami, Florida, USA. Association for Computational Linguistics.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.

Ethan G. Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Michał Wojczulis and Dariusz Kłeczek. 2021. papuGaPT2 - Polish GPT2 language model.

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

Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. 2005. The Penn Chinese treebank: Phrase structure annotation of a large corpus. *Natural language engineering*, 11(2):207–238.

Yuekun Yao and Alexander Koller. 2022. Structural generalization is hard for sequence-to-sequence models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5048–5062, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dmitry Zmitrovich, Aleksandr Abramov, Andrey Kalmykov, Vitaly Kadulin, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Tatiana Shavrina, Sergei S. Markov, Vladislav Mikhailov, and Alena Fenogenova. 2024. A family of pretrained transformer language models for Russian. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 507–524, Torino, Italia. ELRA and ICCL.

Vilém Zouhar, Clara Meister, Juan Gastaldi, Li Du, Mrinmaya Sachan, and Ryan Cotterell. 2023. Tokenization and the noiseless channel. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5184–5207, Toronto, Canada. Association for Computational Linguistics.

A Experiment Results of Replicating Kallini et al. (2024)

We implement the training and evaluation following the same experiment setting from Kallini et al. (2024), only on a 10M word subset of their original data. The result is shown in Figure 5. Unlike in Kallini et al. (2024), however, we do observe that test-set perplexity does increase towards the end of training, indicating that models are overfitting on our smaller datasets. We note that we do not observe this overfitting behavior in the experiments presented in the main text, where the heldout perplexity continues to decrease (or plateau) throughout training.

We calculate Spearman's rank correlation between our results for the **shuffled* languages and those of Kallini et al. (2024) at every 200-step interval from 400 to 1,200. The Spearman's ρ is consistently **1** ($p < 0.001$), indicating perfect agreement between the rankings, showing that 10M words are sufficient enough to replicate the language modeling experiments for which Kallini et al. (2024) originally used 100M words.

In experiment 1, we also conducted a Spearman's Ranking Correlation test between the results on OPUS30 English and those from Kallini et al. (2024)'s experiments. We grouped the SHUFFLE_DETERMINISTIC languages together and observed that the ranking of our English impossible variants aligns perfectly with that reported by Kallini et al. (2024) ($\rho = 1, p = 0.0027$).

B Tokenization Pilot Experiments and Results

In our experiments, where we trained tokenizers for each language using 10M words (around 60MB data), testing vocabulary sizes ranging from 30K to 80K in increments of 10K, we observed two key findings: (1) Tokenizers trained with around 60MB data resulted in unstable language modeling outcomes, and (2) different languages require distinct optimal vocabulary sizes: morphologically richer languages tend to have a larger vocabulary size. We also observed that even when trained on the corpus with matching content, not all languages are equally learnable in terms of their perplexity. These results are shown in Figure 6. Additionally, agglutinative languages like Turkish, with their large number of unique tokens, made large vocabulary sizes impractical. For instance, Turkish has three times the number of unique words as English (467K

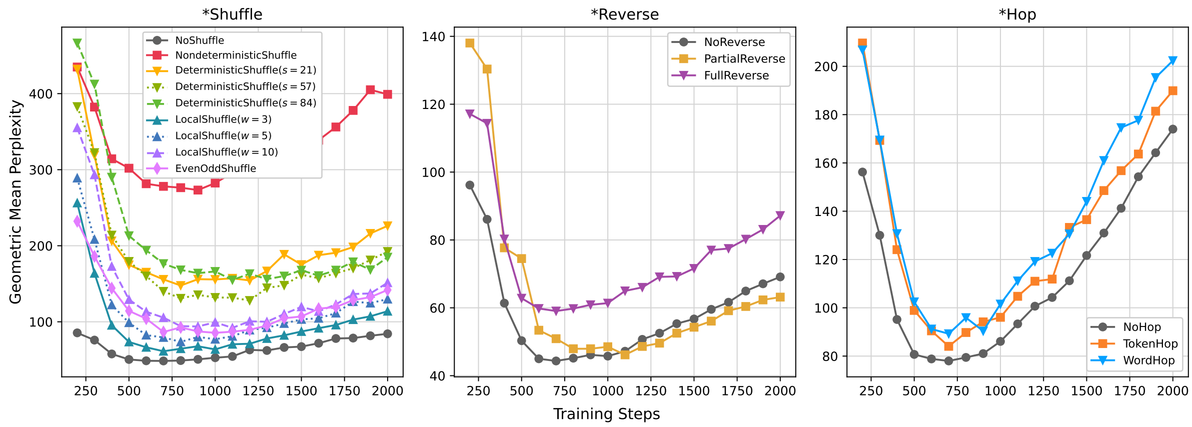

Figure 5: Replication of (Kallini et al., 2024) with 10M words from BabyLM Challenge dataset (strict-small track)

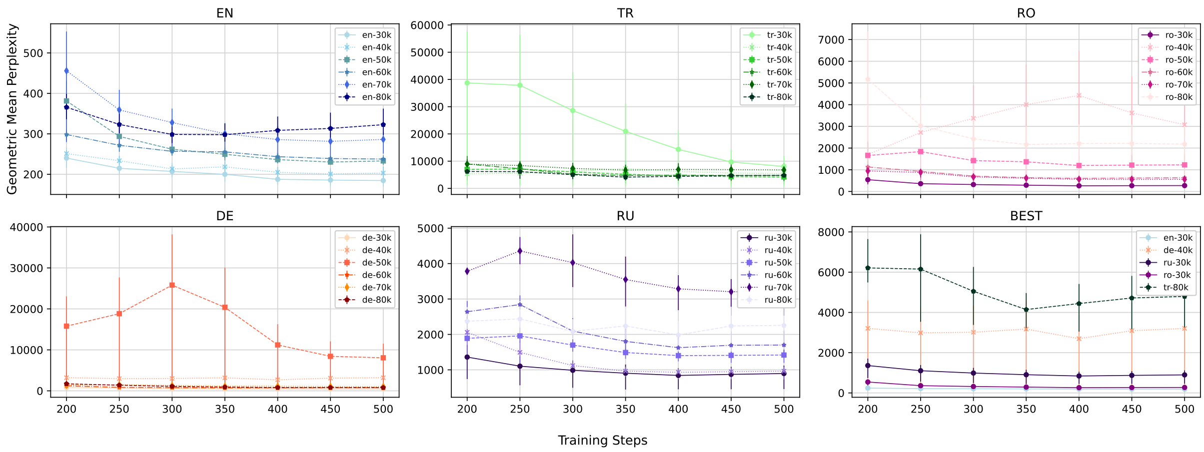

Figure 6: Perplexity results on the development set (10K sentences) for five languages (EN, TR, RO, DE, RU), trained on a 10M-sentence training set across different vocabulary sizes. Error bars represent the first and last quartiles (25% and 75%) of the results. A plot for the optimized vocabulary size (labeled 'BEST') is also included, showing high variance for TR and RU even with optimized vocabulary size.

Language	Treebank	POS-tags			
		DET	NUM	ADJ	NOUN
English	Penn Treebank (Marcus et al., 1993)	DT, PRP$, PDT, POS	QP, $, CD	RB, ADJP, JJR, JJS, JJ	NN, NNS, NNP, NNPS
Italian	VIT(Delmonte et al., 2007)	DET	NUM, SQ	ADJ, SA	NOUN, PRON, PROPN, SYM, X
Chinese	CTB 3.0(Xue et al., 2005)	DT, M, CLP, DP	CD, OD, QP	JJ, ADJP, DNP, DEC, DEG	NN, NP, NR, NT, PRP, PN, FW
Portuguese	Cintil (Barreto et al., 2006)	DET, D, DEM, POSS, POSS'	QNT, QNT', NUM, PERCENTP, PERCENTP', CARD, CARD'	ADJ, AP	N', NOUN, PRON

Table 4: POS-tag categories across languages

vs. 140K), and applying $0.4 \times |V|$ would result in a vocabulary size of 186K, which is too large for efficient language model training with the limited data available and a small model.

C Details of OPUS12 and OPUS30

The typological features of languages used in the two corpora are reported in Table 6. The licensing terms vary depending on their original sources, listed below.

LANGS	AR	TR	RU	PL	DE	IT
TCW	2.19	2.05	2.05	1.98	1.65	1.40
LANGS	PT	NL	RO	EN	FR	
TCW	1.68	1.51	1.81	1.45	1.67	

Table 5: TCW per language by each of their pretrained tokenizer

- NLLB: ODC-By
- TED2020: CC BY–NC–ND 4.0 International; for details, see the official website.
- Bible: CC0 1.0
- OpenSubtitles: GNU General Public License v3.0
- MultiCCAligned: unknown; see the official website.

D Tokenizers

Table 7 shows the details of the tokenizers we use in the experiments. When the training data for a tokenizer is unspecified, we assume it matches the training data used for the corresponding pretrained model.

E TCW & CTC

The TCW is reported in Table 5. We use it to measure the morphological richness of a language.

F POS tags of each treebank

Different constituency parsers are trained with different treebanks. We select POS-tags that are relevant to the four word classes. The detailed POS-tags for each language can be found in Table 4.

G Statistical test between impossible languages

We conducted Welch's paired t-test comparing different perturbations with shuffle_control across 12 checkpoints. The results are ordered alphabetically.

We find that for Dutch, Russian, and Turkish, the difference between SHUFFLE_CONTROL and other perturbations is always significant; by contrast, for languages including Arabic, Chinese, English, German, and Romanian, the difference becomes less significant or insignificant in the locally shuffled variants.

[1] https://huggingface.co/aubmindlab/aragpt2-base

[2] https://huggingface.co/ytu-ce-cosmos/turkish-gpt2

[3] https://huggingface.co/ai-forever/rugpt3large_based_on_gpt2

[4] https://huggingface.co/flax-community/papuGaPT2

[5] https://huggingface.co/malteos/gpt2-xl-wechsel-german

[6] https://huggingface.co/iGeniusAI/Italia-9B-Instruct-v0.1

[7] https://huggingface.co/NOVA-vision-language/GlorIA-1.3B

[8] https://huggingface.co/yhavinga/gpt-neo-125M-dutch

[9] https://huggingface.co/dumitrescustefan/gpt-neo-romanian-780m

[10] https://huggingface.co/openai-community/gpt2

[11] https://huggingface.co/lightonai/pagnol-xl

[12] https://huggingface.co/google-bert/bert-base-chinese

Language	Family	Word Order	Morphology
OPUS12			
English	Indo-European (Germanic)	SVO	Analytic
German	Indo-European (Germanic)	No dominant	Fusional
Russian	Indo-European (Slavonic)	SVO	Fusional
Romanian	Indo-European (Romance)	SVO	Fusional
Turkish	Turkic (Altaic)	SOV	Agglutinative
Dutch	Indo-European (Germanic)	No dominant	Fusional
Polish	Indo-European (Slavonic)	SVO	Fusional
Portuguese	Indo-European (Romance)	SVO	Fusional
Italian	Indo-European (Romance)	SVO	Fusional
French	Indo-European (Romance)	SVO	Fusional
Chinese	Sino-Tibetan	SVO	Analytic
Arabic	Afro-Asiatic (Semitic)	VSO	Root-based (nonconcatenative)
OPUS30			
Spanish	Indo-European (Romance)	SVO	Fusional
Czech	Indo-European (Slavonic)	SVO	Fusional
Bulgarian	Indo-European (Slavonic)	SVO	Fusional
Slovak	Indo-European (Slavonic)	SVO	Fusional
Serbian	Indo-European (Slavonic)	SVO	Fusional
Croatian	Indo-European (Slavonic)	SVO	Fusional
Ukrainian	Indo-European (Slavonic)	SVO	Fusional
Danish	Indo-European (Germanic)	SVO	Fusional
Swedish	Indo-European (Germanic)	SVO	Fusional
Greek	Indo-European (Hellenic)	No dominant	Fusional
Persian	Indo-European (Indo-Iranian)	SVO	Fusional
Lithuanian	Indo-European (Baltic)	SVO	Fusional
Vietnamese	Austroasiatic	SVO	Analytic
Hebrew	Afro-Asiatic (Semitic)	VSO	Root-based (nonconcatenative)
Hungarian	Uralic	SVO	Agglutinative
Indonesian	Austronesian	SVO	Analytic
Japanese	Japonic	SOV	Agglutinative
Korean	Koreanic	SOV	Agglutinative

Table 6: Typological features of the OPUS12 and OPUS30 corpora, with OPUS30 including 18 additional languages beyond those in OPUS12.

Language	|Vocab|	|Training|	Reference	Domain
Arabic[1]	64,000	77GB	Antoun et al. (2021)	Web Crawl, Wikipedia, News
Turkish[2]	50,257	100GB	Kesgin et al. (2024)	Web Crawl, books, news, others
Russian[3]	50,257	450GB	Zmitrovich et al. (2024)	Wikipedia, books, news, books, Web Crawl, Subtitles
Polish[4]	50,257	47GB	Wojczulis and Kłeczek (2021)	Web Crawl
German[5]	50,304	156GB	Ostendorff (2023)	Web Crawl
Italian[6]	50,176	Trillions toks	iGeniusAI (2024)	public sources, synthetic data, and domain-specific content
Portugese[7]	50,258	35B tokens	Lopes et al. (2024)	Web Crawl, News, Subtitles, EuroParl
Dutch[8]	50,257	151GB	Havinga (2023)	Web Crawl
Romanian[9]	64,000	40GB	Dumitrescu (2024)	Web Crawl, Opus, Wikipedia
English[10]	50,257	40GB	Radford et al. (2019)	Web Crawl
French[11]	50,262	130GB	Launay et al. (2022)	Web Crawl
Chinese[12]	21,128	300GB	Devlin et al. (2019)	Wikipedia

Table 7: Tokenizers, vocabulary sizes, training data sizes, references, pretrained model name, and training data domains for each language tested in our experiments.

Perturbation | Step	100	200	300	400	500	600	700	800	900	1000	1100	1200
perturb_reverse_full_word	<0.001	<0.001	0.0036	0.0484	<0.001	0.0283	0.0781	0.5293	1	0.8811	1	1
shuffle_deterministic21	1	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_deterministic57	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_deterministic84	0.7871	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_even_odd	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_local10	0.4242	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_local2	1	<0.001	<0.001	<0.001	<0.001	0.7469	1	1	1	1	1	1
shuffle_local3	1	<0.001	<0.001	<0.001	<0.001	0.0075	0.0062	0.1182	1	1	1	1
shuffle_local5	0.4024	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.3782	0.3208	1	1

Table 8: Welch's t-test comparing each perturbation with shuffle_control across 12 checkpoints for **Arabic**, with Bonferroni adjustment.

Perturbation | Step	100	200	300	400	500	600	700	800	900	1000	1100	1200
perturb_reverse_full_word	<0.001	<0.001	<0.001	1	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_deterministic21	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_deterministic57	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_deterministic84	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_even_odd	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_local10	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_local2	<0.001	<0.001	<0.001	<0.001	<0.001	0.0142	1	1	1	1	1	1
shuffle_local3	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.0012	0.0014	0.0063	0.135
shuffle_local5	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	1	1	0.6033	0.0391	0.0016

Table 9: Welch's t-test comparing each perturbation with shuffle_control across 12 checkpoints for **Chinese**, with Bonferroni adjustment.

Perturbation | Step	100	200	300	400	500	600	700	800	900	1000	1100	1200
perturb_reverse_full_word	<0.001	<0.001	0.0012	0.1697	0.0396	0.0137	0.003	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_deterministic21	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_deterministic57	1	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_deterministic84	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_even_odd	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_local10	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_local2	<0.001	<0.001	1	<0.001	0.0354	0.0277	0.0051	0.0059	0.0158	<0.001	0.0043	<0.001
shuffle_local3	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.002	0.0296	0.0111
shuffle_local5	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.0018	<0.001	0.0049	0.0026

Table 10: Welch's t-test comparing each perturbation with shuffle_control across 12 checkpoints for **Dutch**, with Bonferroni adjustment.

Perturbation | Step	100	200	300	400	500	600	700	800	900	1000	1100	1200
perturb_reverse_full_word	<0.001	<0.001	0.0188	0.0016	0.0025	0.0751	0.0028	0.684	0.938	1	0.8737	1
shuffle_deterministic21	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_deterministic57	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_deterministic84	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_even_odd	<0.001	<0.001	<0.001	0.0078	<0.001	<0.001	0.0022	0.4015	0.0068	0.0013	0.0235	0.0812
shuffle_local10	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_local2	0.0458	<0.001	<0.001	0.0055	0.0251	1	1	1	1	1	1	1
shuffle_local3	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.0269	0.0878
shuffle_local5	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.0167	0.1072

Table 11: Welch's t-test comparing each perturbation with shuffle_control across 12 checkpoints for **English**, with Bonferroni adjustment.

Perturbation | Step	100	200	300	400	500	600	700	800	900	1000	1100	1200
perturb_reverse_full_word	<0.001	0.0284	1	0.008	1	1	1	1	1	1	1	1
shuffle_deterministic21	<0.001	<0.001	<0.001	<0.001	<0.001	0.0046	0.6694	1	1	1	1	1
shuffle_deterministic57	1	<0.001	<0.001	<0.001	<0.001	0.0066	0.6324	1	1	1	1	1
shuffle_deterministic84	<0.001	0.0203	<0.001	<0.001	<0.001	<0.001	0.0402	1	1	1	1	1
shuffle_even_odd	<0.001	<0.001	1	<0.001	1	1	1	1	1	1	1	1
shuffle_local10	<0.001	<0.001	0.3162	<0.001	1	1	1	1	1	1	1	1
shuffle_local2	<0.001	<0.001	1	1	1	1	1	1	1	1	1	1
shuffle_local3	<0.001	<0.001	1	0.5046	1	1	1	1	1	1	1	1
shuffle_local5	<0.001	<0.001	1	0.0257	1	1	1	1	1	1	1	1

Table 12: Welch's t-test comparing each perturbation with shuffle_control across 12 checkpoints for **French**, with Bonferroni adjustment.

Perturbation | Step	100	200	300	400	500	600	700	800	900	1000	1100	1200
perturb_reverse_full_word	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_deterministic21	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_deterministic57	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_deterministic84	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_even_odd	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_local10	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_local2	<0.001	<0.001	<0.001	<0.001	0.034	0.0974	0.4926	0.4579	1	0.2287	1	0.4976
shuffle_local3	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.0035	0.0185	0.011	0.0675	0.14	0.2177
shuffle_local5	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.001	<0.001	<0.001	<0.001	0.0017

Table 13: Welch's t-test comparing each perturbation with shuffle_control across 12 checkpoints for **German**, with Bonferroni adjustment.

Perturbation | Step	100	200	300	400	500	600	700	800	900	1000	1100	1200
perturb_reverse_full_word	<0.001	1	1	1	1	1	1	1	1	1	1	1
shuffle_deterministic21	<0.001	<0.001	1	0.5567	1	1	1	1	1	1	1	1
shuffle_deterministic57	0.002	0.0107	0.8726	0.3123	1	1	1	1	1	1	1	1
shuffle_deterministic84	<0.001	<0.001	0.0112	0.0013	0.002	1	1	1	0.8841	1	1	1
shuffle_even_odd	<0.001	<0.001	0.1087	0.3001	0.1957	1	1	1	1	1	1	1
shuffle_local_word3	<0.001	0.0846	1	1	1	1	1	1	1	1	1	1
shuffle_local10	0.0098	<0.001	<0.001	0.4004	1	1	1	1	1	1	1	1
shuffle_local2	<0.001	1	1	0.3393	0.0606	0.1196	0.1088	0.1105	0.1625	0.1634	0.2567	0.2097
shuffle_local3	1	1	1	1	1	1	1	1	1	1	1	1
shuffle_local5	1	1	1	1	1	1	1	1	1	1	1	1

Table 14: Welch's t-test comparing each perturbation with shuffle_control across 12 checkpoints for **Italian**, with Bonferroni adjustment.

Perturbation | Step	100	200	300	400	500	600	700	800	900	1000	1100	1200
perturb_reverse_full_word	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_deterministic21	0.0018	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_deterministic57	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_deterministic84	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_even_odd	0.0049	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_local10	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_local2	0.0288	<0.001	<0.001	<0.001	<0.001	0.0166	0.0577	0.1104	0.0993	0.0446	0.0521	0.0508
shuffle_local3	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_local5	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

Table 15: Welch's t-test comparing each perturbation with shuffle_control across 12 checkpoints for **Polish**, with Bonferroni adjustment.

Perturbation | Step	100	200	300	400	500	600	700	800	900	1000	1100	1200
perturb_reverse_full_word	<0.001	<0.001	<0.001	<0.001	<0.001	0.0056	0.0028	0.0659	0.1986	1	1	1
shuffle_deterministic21	1	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.0358	0.0039	0.0778
shuffle_deterministic57	1	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	1	0.4359	1
shuffle_deterministic84	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.006	0.0012	0.0137
shuffle_even_odd	1	<0.001	<0.001	0.0569	<0.001	0.3289	1	1	1	1	1	1
shuffle_local10	1	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.1379	1	1	1	1
shuffle_local2	1	1	<0.001	1	0.1401	1	1	1	1	1	1	1
shuffle_local3	0.8382	<0.001	<0.001	0.0189	<0.001	1	1	1	1	1	1	1
shuffle_local5	1	<0.001	<0.001	0.6868	<0.001	1	1	1	1	1	1	1

Table 16: Welch's t-test comparing each perturbation with shuffle_control across 12 checkpoints for **Portuguese**, with Bonferroni adjustment.

Perturbation | Step	100	200	300	400	500	600	700	800	900	1000	1100	1200
perturb_reverse_full_word	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	1	<0.001	<0.001	0.0693	0.0291	0.0033
shuffle_deterministic21	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.6784	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_deterministic57	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	1	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_deterministic84	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.1518	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_even_odd	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	1	<0.001	<0.001	0.0059	0.011	0.1607
shuffle_local10	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	1	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_local2	<0.001	<0.001	<0.001	<0.001	<0.001	0.0311	1	0.1407	0.123	0.0092	0.0079	0.0246
shuffle_local3	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	1	<0.001	<0.001	<0.001	<0.001	0.0292
shuffle_local5	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	1	<0.001	<0.001	0.0014	<0.001	0.2879

Table 17: Welch's t-test comparing each perturbation with shuffle_control across 12 checkpoints for **Romanian**, with Bonferroni adjustment.

Perturbation | Step	100	200	300	400	500	600	700	800	900	1000	1100	1200
perturb_reverse_full_word	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_deterministic21	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_deterministic57	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_deterministic84	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_even_odd	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_local10	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_local2	<0.001	<0.001	<0.001	1	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_local3	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_local5	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

Table 18: Welch's t-test comparing each perturbation with shuffle_control across 12 checkpoints for **Russian**, with Bonferroni adjustment.

Perturbation | Step	100	200	300	400	500	600	700	800	900	1000	1100	1200
perturb_reverse_full_word	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_deterministic21	0.0551	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_deterministic57	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_deterministic84	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_even_odd	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_local10	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_local2	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.0036	<0.001	0.002	0.0107	0.0403
shuffle_local3	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_local5	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
shuffle_nondeterministic	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

Table 19: Welch's t-test comparing each perturbation with shuffle_control across 12 checkpoints for **Turkish**, with Bonferroni adjustment.