When transformers learn "impossible" languages, what do they learn?

Ram Janarthan* **Coleman Haley*,†** **Sharon Goldwater**
University of Edinburgh
†coleman.c.haley@gmail.com

Abstract

Recent work suggests that transformer language models show a bias towards human languages over unnatural ("impossible") languages argued to be unacquirable by humans. However, this literature has largely based these claims on differences in sample efficiency and test-set perplexity, rather than on direct evaluations of the linguistic capacities that could plausibly explain non-attestation in human languages. We evaluate two theoretically motivated linking hypotheses: impossibility arising from deficiencies in *grammatical sensitivity* or *generative production*. Using GPT-2 style models trained on perturbed "impossible" variants of English, we measure sensitivity to grammaticality using BLiMP minimal pairs, finding that model performance exhibits only gradual degredation, mediated by the language's information locality. In contrast, these models exhibited pronounced failures in generation, producing substantially fewer high-quality sentences at longer lengths. Together, these results suggest generative deficiency and transmission failures as a plausible linking hypothesis between language model behaviour and non-attestation of impossible languages.

1 Introduction

Recent successes of language models (LMs) have sparked a debate over whether they can inform theories of human language acquisition. LMs appear to exhibit substantial linguistic competence, producing novel grammatical sentences (McCoy et al., 2023) and showing sensitivity to diverse linguistic phenomena (Hu and Levy, 2023; Linzen and Baroni, 2021; Wilcox et al., 2024)—leading some to suggest that general-purpose learning mechanisms may suffice to acquire language (Mahowald et al., 2024; Futrell and Mahowald, 2025). An opposing perspective instead posits that language acquisition is guided by innate, language-specific constraints that sharply restrict the space of possible human languages, accounting for both cross-linguistic universals and the systematic absence of many logically possible but unattested languages (Chomsky, 1966, 1998; Moro, 2008). Under this view, LMs, lacking language-specific biases, have been claimed to be irrelevant for explaining human language acquisition and patterns of unattested languages (Chomsky et al., 2023; Moro et al., 2023).

Responding to these claims, Kallini et al. (2024) performed an empirical study showing that transformer LMs trained on a developmentally-plausible amount of data learned English better and more quickly than modified English variants with properties argued to be impossible (henceforth "impossible languages"). Subsequently, these authors, in Kallini and Potts (2025) and Futrell and Mahowald (2025), have argued that the domain-general inductive biases present in LMs can inform linguistic theory by modelling the pattern of (un)attested human languages. These arguments have relied on differences in learning dynamics: LMs trained on impossible languages tend to converge more slowly and yield higher test-set perplexity than those trained on natural languages. This leaves the *linking hypothesis* to the non-attestation of these languages underdeveloped[1]: why should differences in perplexity or sample efficiency correspond to *exclusion* from the space of extant human languages?

In this study, we focus on two kinds of linguistic capacities as candidate linking hypotheses between LM behaviour and linguistic "impossibility," motivated by contrasting theoretical orientations in linguistics. **Sensitivity to grammatical well-formedness** plays a central role in accounts that attribute non-attestation to limits on what learners can infer from input (Chomsky, 1980, 1966), while

*These authors contributed equally to this work.

[1]Indeed, Kallini and Potts (2025) have subsequently called for the development of stronger linking hypotheses between this research program in language models and human language.

reliable **generation of well-formed sentences** is emphasized by accounts that locate impossibility in failures of transmission across speakers (Kirby et al., 2008). If models fail to acquire sensitivity to some critical aspects of a language, then impossibility might arise from unlearnability; while if models cannot generate well-formed sentences in the language, impossibility might reflect difficulties in transmitting the language.

We focus our study on the capacities acquired by transformer models trained on a cognitively plausible amount of data: either the English BabyLM corpus (Warstadt et al., 2023) or impossible variants defined by Kallini et al. (2024) as permutations of the English sentences. To evaluate *grammatical sensitivity*, we use the BLiMP minimal pair dataset (Warstadt et al., 2020), forming "impossible" variants of BLiMP by permuting the BLiMP stimuli in the same manner as the BabyLM data was permuted by Kallini et al. (2024). To evaluate *generative performance*, we leverage the fact that grammatical strings of Kallini et al.'s impossible languages can be inverted to produce grammatical strings of English. Generating sentences from each impossible language model and converting them to their English equivalents, we use an LLM to evaluate the quality of the generations.

We find models trained on impossible languages acquire substantial *passive grammatical sensitivity*, degrading gradually in performance with respect to both overall test-set perplexity and m-local entropy, a measure of information locality proposed to explain asymmetries in impossible language learning (Someya et al., 2025). This result stands in contrast to claims that human learners would fail to acquire grammatical competence from such data.[2]

In terms of *generative performance*, we find that models trained on impossible languages produce substantially fewer well-formed sentences in their languages, in a manner which does *not* strictly align with held-out test set perplexity or m-local entropy.

Together, these results suggest that language models acquire substantial (if slightly degraded) *grammatical sensitivity* to impossible languages, but tend to fail at *generative performance*, suggesting generative performance as a potential linking hypothesis between poorer LM distributional modelling of impossible languages and human non-occurrence. While these results will not resolve the debate about whether transformers are good models of human learners, they do provide a potential explanation for *why* some languages could be impossible, even for a learner with no strong language-specific inductive biases. Our study also provides a more principled methodology for studying impossibility with language models, which we encourage future studies to adopt and extend.[3]

2 Background

What kinds of languages are "impossible" is the subject of ongoing discussion among linguists, due to both the difficulty in establishing true universals of natural languages, and uncertainty about whether unattested languages could in principle be learned. Nevertheless, Kallini et al. (2024) proposed a set of candidate impossible languages by defining perturbations of the word order of a sentence applied to a natural language (English). These perturbations manipulate sentences in ways that are never observed in any human languages and have been previously hypothesized to be impossible (Moro, 2008; Mitchell and Bowers, 2020) due to their use of unattested and "unnatural" word orders. Each perturbation is applied to a tokenized sentence as described in Table 1, and all except one can be deterministically converted to English equivalents, a fact which we exploit in Section 5. This deterministic mapping also means that the true entropy of these languages is the same as the entropy of English, so a model that successfully learns both English and the impossible languages should have the same perplexity on each.

In fact, this is not what Kallini et al. (2024) found. They trained GPT-2 transformer models (Radford et al., 2019) on both the BabyLM corpus (Warstadt et al., 2023) and perturbed (impossible) versions of it, and showed that the models trained on impossible languages converged more slowly and ended up with slightly higher perplexity on held out data, suggesting more difficulty learning the impossible languages. However, results based on perplexity alone are somewhat difficult to interpret (how much worse constitutes a learning failure?), and follow-ups on other languages have shown more mixed results (Ziv et al., 2026; Yang et al., 2025). More importantly, there is no theoretically-motivated link between low perplexity and impossibility. In Section 3, we explore two

[2]For example, Chomsky (1980) outlines how human learners would be *unable* to acquire a rule based on linear ordering, as opposed to hierarchical structure.

[3]The code for this paper is available at: https://github.com/ramjanarthan/impossible-languages

Language	Abbr.	Perturbation Rule	Example Sentence
ENGLISH	E	No perturbation	Jessica stole this rabbit 's hat .
FULLREVERSE	FR	Randomly insert a special R token, and reverse the ordering of all tokens	. hat 's rabbit this R stole Jessica
PARTIALREVERSE	PR	Randomly insert a special R token, and only reverse the ordering of all tokens following it	Jessica stole R . hat 's rabbit this
LOCALSHUFFLE3	S3	Deterministically shuffle tokens within a window of size 3	this Jessica stole hat rabbit 's .
LOCALSHUFFLE5	S5	Deterministically shuffle tokens within a window of size 5	this 's rabbit Jessica stole hat .
EVENODDSHUFFLE	SEO	Reorder tokens so even-indexed tokens appear before odd-indexed tokens	Jessica this 's . stole rabbit hat
LOCALSHUFFLE10	S10	Deterministically shuffle tokens within a window of size 10	this 's rabbit . hat Jessica stole
DETERMSHUFFLE	DS	Deterministically shuffle all tokens, with shuffling seed 21	Jessica 's stole hat rabbit . this
NONDETERMSHUFFLE	NDS	Nondeterministically shuffle all tokens	this 's rabbit . hat Jessica stole

Table 1: Examples of sentences in English and impossible languages. Coloured boxes represent GPT-2 tokens. We use the same colors and abbreviations to represent languages throughout the paper. Languages in the table are ordered by increasing 4-local entropy (see Section 2).

more theoretically-elaborated hypotheses linking LM performance to impossibility.[4]

While the primary aim of this study is to test these more explicit linking hypotheses, we also secondarily study how our results relate to two prior measures of impossibility: perplexity and m-local entropy. Someya et al. (2025) proposed m-local entropy as an information-theoretic measure characterizing impossible languages. They defined m-local entropy as the next-symbol entropy given a context of size $m-1$, and estimated it for these languages using n-gram models trained on perturbed corpora very similar to those in Kallini et al. (2024). They showed a strong positive correlation between m-local entropy (strongest when $m = 4$) and the perplexity of transformer models trained on different impossible languages, concluding that transformers exhibit an information-locality bias that drives Kallini et al.'s hierarchy of impossibility.

3 Linking Hypotheses

Prior work on impossible languages on LMs has focused on learning dynamics and held-out perplexity, implicitly treating poorer compression as an explanation for, or a diagnostic of, impossibility. In this section, we argue that this move is theoretically underjustified. We propose two alternative linking hypotheses arising from opposing theoretical orientations within linguistics: deficiencies in *grammatical sensitivity* and *generative performance*, describing how they may diverge from perplexity and each other.

3.1 Grammatical sensitivity in LMs

Generative linguists have generally put substantial emphasis on the challenge of acquiring language from the limited data humans are exposed to during development (Chomsky, 1966, 1980). For example, (Chomsky, 1980) claimed that the linguistic input available to learners is insufficient to determine the correct grammatical generalizations without strong, language-specific inductive biases. On this view, impossible languages are impossible because learners cannot reliably acquire the abstract grammatical distinctions required to discriminate grammatical from ungrammatical strings, even with extensive exposure (Moro, 2008). If this claim is correct, then models trained on impossible languages should show enduring deficits in passive sensitivity to linguistic structure relative to models trained on natural languages, thereby explaining why these languages are unattested.

[4]We note that Xu et al. (2026) also evaluated models' grammatical sensitivity using minimal pairs, but they did so using typologically *implausible* (rather than impossible) languages, making their results less suitable for determining a linking hypothesis between impossibility and LM performance.

Language	Grammatical Example	Ungrammatical Example
ENGLISH	Rodney goes to **this** new mall.	Rodney goes to **these** new mall.
FULLREVERSE	. mall new **this** to goes🄁neyRod	. mall new **these** to goes🄁neyRod
LOCALSHUFFLE3	goesRodney new to **this** mall.	goesRodney new to **these** mall.
NONDETERMSHUFFLE	goesRod new. toney **this** mall	goesRod new. toney **these** mall

Table 2: Examples of a minimal pair in different (impossible) languages for the BLiMP task "Determiner Noun Agreement with Adjective".

The most widely used and successful approach to evaluating sensitivity to grammatical well-formedness in LMs is the use of minimal pair datasets like BLiMP (Warstadt et al., 2020). Specifically, minimally-differing strings, one grammatical and one ungrammatical, are compared using perplexity, with the lower perplexity taken to be the model's preferred variant. The use of minimal pairs accounts for the "noisy channel" nature of sentence probability (Hu et al., 2026; Levy, 2008), which assigns lower values to infrequent grammatical strings than to ungrammatical strings which are close to frequent grammatical strings.

This minimal-pair view of grammatical sensitivity in language models displays patterns which may diverge from the overall model loss. For example, consider the following family of Boltzmann distributions parameterized by temperature T:

$$P_T(x) = \frac{\exp(\text{logits}(x)/T)}{\sum_{x' \in V} \exp(\text{logits}(x')/T)}, \quad (1)$$

where $\text{logits}(x)$ are the logits of some neural LM for an input x. While a language model most standardly refers to the distribution where $T = 1$, in practice, generations from a LM are often drawn from a number of distributions in this family, with T controlling how conservative the generations are. These distributions have different entropies, which implies different perplexities on a test set, but varying T is a rank-preserving operation: for any x and y, $P_T(x) > P_T(y)$ implies $P_{T'}(x) > P_{T'}(y)$. That is, *the whole family of distributions* has *identical scores* on a minimal pair benchmark like BLiMP (Warstadt et al., 2020), in keeping with the intuition that samples from any of these distributions are samples from the same language model, with the same passive linguistic competence.

This example illustrates a general property of minimal-pair evaluation: it abstracts away from many sources of variation that affect overall likelihood but are orthogonal to grammatical well-formedness. Perplexity, by contrast, aggregates over all factors that influence probability mass, including lexical, frequency-based, and calibration effects. In this paper, we investigate whether differences in average perplexity reflect genuine differences in grammatical sensitivity, or are dominated by other distributional factors.

Perplexity and emergent capabilities: Recent studies of large LMs have identified so-called "emergent capabilities"—tasks that LMs are poor at below a sufficient scale. Du et al. (2024) argue that this is best understood as abilities where overall model perplexity changes a small amount, but performance on that task changes a large amount. This provides additional motivation for our study—not only is it possible that the models trained on impossible languages could have exactly the same grammatical sensitivity as measured on BLiMP tasks, but the reverse could also be the case—all models except the English model could, in principle, fail to acquire sensitivity to some key linguistic properties of their input language.

3.2 Generation and Iterated Learning

A separate group of linguists emphasizes that language is shaped by cultural transmission over generations. On this view, sometimes called the "Iterated Learning" account (Kirby et al., 2008), acquiring competence in a language is necessary, but not sufficient, for that language to persist with all its properties. The inputs to the next generation of language learners are the productions of the previous generation; accordingly, speakers must be able to reliably produce sentences in a language for it to be transmitted to the next generation. As suggested by the temperature argument in the previous section, even perfect minimal-pair discrimination does not imply the well-formedness of generations, so in Section 5 we attempt to measure generation quality

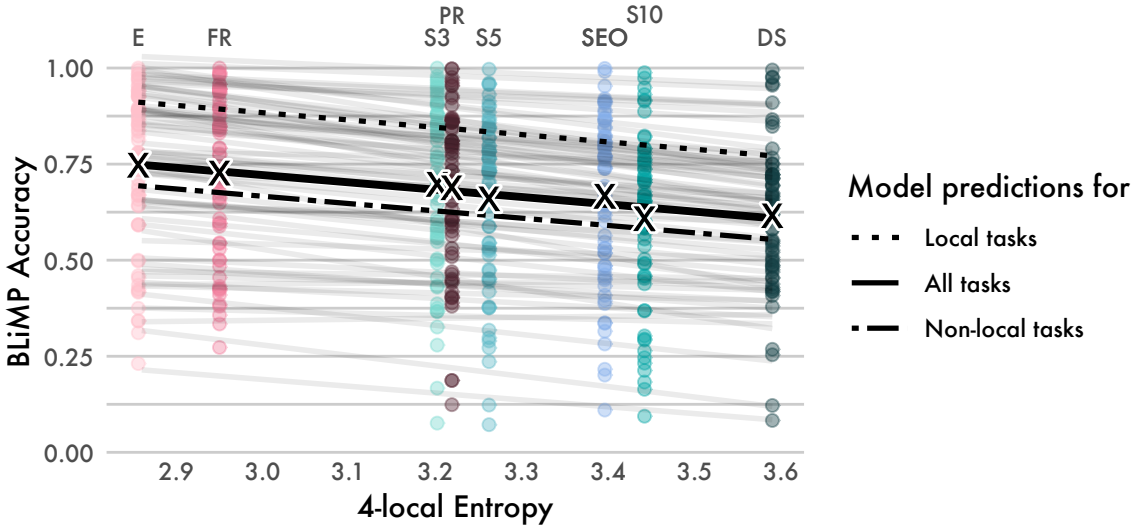

Figure 1: BLiMP task accuracy of models trained on English and impossible languages, correlated with 4-local entropy (see Appendix B for the analogous plot of model accuracy against perplexity, which fits less well). Coloured dots and grey lines show accuracy and lines of best fit for individual tasks, with Xs indicating the mean for each language (labeled at top with abbreviations from Table 1). The model indicates a modest, linear decline with increasing 4-local entropy (fitted black line). This decline is insensitive to task locality: tasks that are solved well by a 5-gram model in English exhibit almost the same rate of degradation as those that are not (dotted vs. dashed line).

explicitly using a pretrained LLM.

4 Experiment 1: Minimal Pairs

In this section, we test the grammatical sensitivity of the GPT-2 models trained by Kallini et al. (2024) on "impossible" versions of English, by evaluating them across a suite of 67 tasks that test for different grammatical phenomena (BLiMP; Warstadt et al., 2020). We constructed BLiMP datasets for each impossible language by applying the corresponding perturbation rule to each sentence.[5] Table 2 shows examples of our minimal pairs for one task.

We computed accuracy as the percentage of pairs where the log likelihood of the grammatical sentence was larger than the ungrammatical sentence. To understand how BLiMP performance degrades as the languages diverge more from human languages, we conducted two mixed effects analyses associating (1) 4-local entropy[6] or (2) perplexity with BLiMP task accuracy, using both a fixed and random slope and intercept for each task.

[5]During pre-processing, we discarded BLiMP minimal pairs that have an uneven number of tokens across the two sentences, since such pairs could become less minimally distinct in the impossible languages, due to rules which depend on linear position/number of tokens. Table 7 in Appendix D lists the final dataset sizes; most datasets keep all sentences.

[6]Someya et al. (2025) found the strongest relationship between m-local entropy and perplexity for $m = 4$, so we focus our analysis on this value of m. Calculation details for our m-local entropy and perplexity scores are in Appendix A.

4.1 Overall performance

Figure 1 shows the performance of all models on BLiMP and the fit of our main mixed effects analysis. All models perform substantially above chance on BLiMP, with average performance over all tasks (Xs in the figure) ranging from 74.7% accuracy for the English model to 61.5% for DETERMSHUFFLE. For comparison, Warstadt et al. (2020) report that GPT-2-large obtains 80.1% accuracy on BLiMP, compared to a human accuracy of 88.9% (the ceiling for meaningful performance improvements). Thus, the English model, despite having 16% as many parameters as GPT-2-large (125M vs. 775M) and being trained on 1.25% of the data (0.54GB vs. $\approx$ 40GB), still performs fairly well, comparable with the best baseline in the BabyLM challenge (OPT-125M, 75%; Warstadt et al., 2023).

Interestingly, we find that 4-local entropy captures the pattern of BLiMP scores much better than perplexity on a held-out test set (Δ AIC $= -56$). For example, PARTIALREVERSE has lower perplexity than FULLREVERSE, but lower BLiMP accuracy and higher 4-local entropy.We find a modest negative linear impact of 4-local entropy on BLiMP performance within the studied range (the solid black line in Figure 1: $\beta = -0.19, p \ll 0.001$, 95% CI: $[-0.22, -0.16]$). That is, for each bit that 4-local entropy increases (somewhat more than *double* the difference between ENGLISH and

Language	Low Perplexity (Higher Quality)	High Perplexity (Lower Quality)
ENGLISH	*"Yes, that's what I'm trying to tell you."*	*"I'll do you some fancy drawing down from me."*
EVENODDSHUFFLE	*"And what the hell are you going to do with him?"*	*"I you what wouldn said we't't I didn? ? suppose"*
LOCALSHUFFLE5	*"All right, let's just have a look at the camera."*	*"That's for all, is the? for what."*

Table 3: Examples of model generations between 11 and 15 tokens long in different languages, with perturbations inverted where applicable. The examples have perplexity scores below (left) and above (right) the 75th percentile of all English generations. As suggested by these examples, the top (worst) perplexity quartile for the English model still contains many grammatical (if nonsensical or unnatural) sentences. Generations for impossible languages that have perplexity at least as high are often ungrammatical. More examples for all languages are shown in Appendix C.

LOCALSHUFFLE5), overall BLiMP accuracy decreases by approximately 19%. For a language like LOCALSHUFFLE3 which is much closer to English than 1 bit, this translates into an accuracy of 68% compared to 74% for English.

4.2 Does task locality matter?

Some BLiMP tasks may be solvable via simple local cues; such tasks could be less impacted in impossible languages than those which require more abstract generalizations. Indeed, Chomsky's (1980) claim is precisely that rules which require structural sensitivity specifically require Universal Grammar to learn. To test whether overall BLiMP trends masked a collapse for more complex tasks, we split BLiMP tasks by their ability to be solved using local surface cues, proxied by a high accuracy score ($\geq 80\%$) by a 5-gram model in English, using the scores reported in Warstadt et al. (2020). We refer to tasks with $\geq 80\%$ accuracy as "local" tasks, and others as "non-local" tasks. Table 7 in Appendix D lists accuracy scores and locality for all tasks.

Figure 1 shows the results of including task locality as a fixed-effect in our analysis (dashed and dotted lines). The effect is significant ($p \ll 0.001$), and local tasks have about 22% higher accuracy in our analysis. However, this effect is *constant* regardless of 4-local entropy: adding an interaction between 4-local entropy and task locality does *not* result in a significantly better fit ($\chi^2(1) = 0.65, p = 0.42$). That is, non-local tasks are *not* unusually difficult to learn in impossible languages.

At the macro-level, the results outlined in this section do not align with either extreme possibility: models neither exhibit a dramatic loss of grammatical sensitivity for all impossible languages, nor are they unaffected. This could indicate that grammatical sensitivity to a language might not be a suitable linking hypothesis to linguistic impossibility. In the next section, we investigate the potential of generative performance as an alternative linking hypothesis, by analysing the quality of generations sampled from Kallini et al.'s (2024) models.

5 Experiment 2: Generative performance

Evaluating LM generation quality and naturalness without references is a challenging problem. One commonly used method is to use a higher quality model trained in the same language (usually on a larger dataset) to assess generations. To do this, we leverage the fact that these impossible languages can be deterministically reverted to English. After applying an inversion function to a model's generation, we evaluate it as valid English.[7] This tests the LM's ability to generate acceptable sentences, as well as how well it learned that particular perturbation rule. To evaluate a generation's acceptability, we computed the perplexity per token using a pretrained LLM (GPT2-large; Radford et al. (2019)).

Although perplexity is normalized by sequence length, it still tends to have a strong correlation with the number of tokens in a sentence: early tokens tend to have higher surprisal as there is less information available to predict them and more uncertainty about the text being generated. As such, we ensure to compare perplexities across models only for matched generation lengths. We further perform an analysis based on *quartiles* to evaluate the proportion of generations for each model that are of a similar quality to the English model of Kallini et al. (2024). Because the generations of the English model are fairly high-quality, this serves as a simple proxy for which of the generations of the other models are of similar quality. For instance, considering a window size of 5 tokens (e.g., sentences of lengths 6-10), we compute the 25th, 50th, and 75th percentiles of English generations in terms of perplexity. Using the same window

[7] For inversions of the reverse languages, we removed *all* instances of the R token prior to evaluation as English.

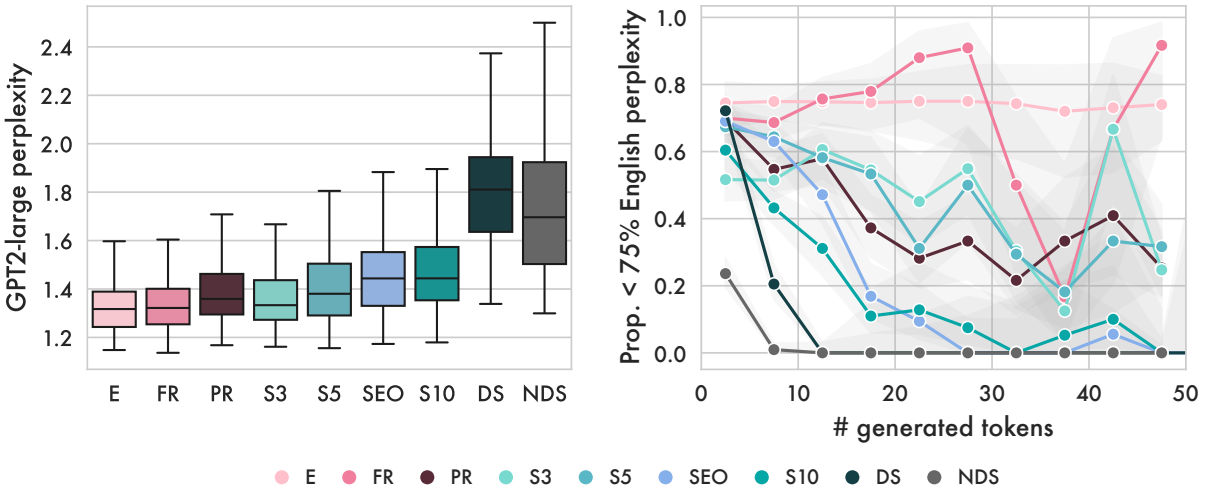

Figure 2: Evaluation of generations from the impossible language models. **Left:** the perplexity under GPT2-large of generations between 11 and 20 tokens in length after inverting the "impossible" transformations, largely corroborating previous hierarchies of impossibility. **Right:** The proportion of generations with a perplexity below 75% of English generations ("high-quality"), stratified by length. For nearly all impossible languages, substantially less than 75% of model generations fall into this perplexity range, with the proportion decreasing with length.

size, we then compare what proportion of generations in a particular impossible language fall *below* the 75th percentile/third quartile of the English model. For purposes of space and clarity, we will refer to these generations as "high-quality."[8] We generated 1000 sentences of up to 50 tokens from each model by sampling from their full trained distribution at each timestep (temperature 1). A model's beginning-of-sequence token was used as the prompt for each generation. We did not use any methods that truncate the space of considered output tokens, like top-k or top-p sampling, nor did we use beam search. Examples of generations with perplexities in different quartiles are shown in Table 3 and Appendix C.

Figure 2 summarises the main findings of our generation experiments. The left-hand plot shows the distribution of perplexities for generated sentences in each language. The ranking of languages in terms of generation perplexity is similar to that based on test-set perplexity found by Kallini et al. (2024) ($\rho = 0.93$), but is even more similar to the ranking based on 4-local entropy from Someya et al. (2025) ($\rho = 0.98$), diverging from test-set perplexity in the same places (chiefly REVERSEPARTIAL and EVENODDSHUFFLE). This provides further evidence of m-local entropy being *more* predictive of model behaviour than perplexity.

The right plot in Figure 2 reveals a richer story. All models were found to generate strings of a given length at roughly the same rate (so apparent collapses are not due to a lack of sentences of that length). All models (with the exception of NONDETERMSHUFFLE) are able to generate a substantial (50% or higher) percentage of high-quality sentences at a length of 5 tokens or less. However, the impossible language models tend to produce fewer high-quality generations as the number of generated tokens increase. While substantially outperforming NONDETERMSHUFFLE for short strings, DETERMSHUFFLE essentially never generates high-quality strings of a length greater than 10, suggesting a failure to learn the shuffling patterns for longer sequences.

The clear outlier is FULLREVERSE, which demonstrates a similar or even higher proportion of high-quality sentences compared to the base English model. The *higher* proportion may not indicate that the generations of FULLREVERSE are actually better sentences than the English model, but rather that they express less complex/more frequent messages (Hu et al., 2026), highlighting the limitations of the current evaluation approach for distinguishing between models with generally high-quality generations. Nevertheless, all other impos-

[8]We chose not to use human judgments in this study, as we wanted a scalable and inexpensive evaluation for future studies to follow; still, this is an important limitation. Notably, perplexity conflates surprise due to malformedness with surprise due to sentences that convey infrequent meanings (Hu et al., 2026); if sentences from the impossible models use more frequent meanings than those from the English model, this method could over-estimate their quality. However, we do not find this to be a major issue here in practice.

sible models have substantially fewer high-quality generations, particularly for longer sentences, suggesting substantial naturalness and acceptability issues with their generations.

6 Discussion

In this study, we asked what linguistic capacities LMs trained on impossible languages acquire, with the aim of identifying a plausible linking hypothesis between LM behaviour and impossibility. Across the two experiments, we see a divergent pattern of results. Experiment 1 found that the grammatical sentitivity of an LM trained on an impossible language depends primarily on the local entropy of the language, but this effect is very modest, with LMs still achieving substantial grammatical sensitivty. In contrast, Experiment 2 showed that sampling from the learned distributions of these models yielded substantially fewer high-quality sentences than a comparable English model, especially at longer lengths.

What implications do these divergent findings have for the impossible language debate as it connects to language models? Our minimal pair results are amenable to multiple interpretations, largely because we lack definitive evidence about the human case. Our results do not align with generativist predictions that linguistic competency from human-scale data requires the language conform to Universal Grammar; many impossible models are almost as structurally sensitive as the base English model. However, because the English model does not achieve human level grammatical comptence, a generativist could reasonably argue that the type of grammatical learning exhibited by these models does not meaningfully bear on poverty-of-the-stimulus claims. Nevertheless, the mounting evidence base of a steady linear climb towards grammatical competence at data levels close to the human scale (including the present results) increasingly challenges traditional poverty-of-the-stimulus views (Warstadt et al., 2020; Hu et al., 2024; Oh and Schuler, 2023).

At the same time, the minimal pair results do not provide compelling evidence that the languages studied here are unlearnable by a learner with similar inductive biases to these models. The modest degradation in grammatical sensitivity shown in these results predicts that either human learners differ from current LMs in critical ways, or that humans would likewise acquire substantial grammatical sensitivity in these "impossible" languages—leaving open the question of why such languages do not occur.

One possible resolution emerges from our generation results. Human language is transmitted and evolves through *Iterated Learning*, in which speakers learn language from the productions of a previous generation of learners (Kirby et al., 2008). If a language is especially difficult to generate, it would be unlikely to survive transmission through generations of learners. Such a language may not be impossible to *learn*, but to transmit. However, this linking hypothesis faces an important challenge: *unlike* human language acquisition, iterated learning of language in current LMs tends to degenerate over successive generations (LeBrun et al., 2022; Guo et al., 2024; Shumailov et al., 2024). Similar behaviour has been observed in human iterated learning experiments: even simple artificial languages degenerate over generations of human learners when there is no pressure for the language to be expressive (Kirby et al., 2008). In humans, this pressure for expressivity can be provided by the communicative function of language: when subjects in iterated learning experiments must use their acquired language in a signalling task, the artificial language changes while maintaining its expressivity (Kirby et al., 2015). While it could be possible to modify LLM training procedures to introduce an expressivity pressure (Smith et al., 2024), the issue of LM degeneration nevertheless highlights critical differences between current LM training paradigms and the dynamics of human language transmission with which a successful theory must more thoroughly contend.

Additionally, our results provide new evidence for the importance of Someya et al.'s (2025) m-local entropy in understanding transformer learning dynamics. While Someya et al. (2025) found correlations between perplexity and m-local entropy, we find that in both our experiments, 4-local entropy predicts LM behaviour better than perplexity does. This contrast is striking: perplexity is a property of the *model being evaluated*, while 4-local entropy is *solely* a property of the *training data.* Further work is needed to understand how these findings generalize across tasks, languages, and models, and to explain why m-local entropy so strongly predicts generation quality and BLiMP performance.

One limitation of this study is our focus on a single transformer architecture (GPT-2 small) and set of hyperparameters. We maintain the same models

as Kallini et al. (2024) for the purposes of comparison, but future work should ensure findings in this area are robust to variations in training data, training duration, hyperparameters, and model specifics.

Another important limitation of our findings is that they are based on a single source language: English. Prior studies extending Kallini et al. (2024)'s approach to multiple languages have yielded mixed results, with "impossible" languages sometimes achieving *lower* perplexity than real human languages (Ziv et al., 2026; Yang et al., 2025). The interpretation of these findings is unclear, because they rely on cross-linguistic comparison of absolute perplexity, which has been shown to be unreliable across languages and datasets (Poelman and de Lhoneux, 2026). Because our approach does not depend on cross-linguistic comparisons of absolute perplexity, it may be more cross-linguistically robust. Establishing this remains a critical direction for future work, as similar confounds could still arise in practice.

7 Conclusion

In the emerging literature of impossible language modelling, a persistent issue has been a lack of clarity on what model behaviour needs to explain. In this paper, we have argued that a cognitively relevant theory cannot simply provide a function for identifying if a language is "impossible," but must *also* provide a plausible linking hypothesis for why that language is unattested. Prior work had obscured the fact that models trained on impossible languages can still achieve substantial grammatical sensitivity; this fact is *not* aligned with *generative* theories of what drives language non-attestation. The picture is more promising for a transmission-oriented account, but a full account still requires substantial elaboration. Even in this simple study, though, it is clear that learning dynamics do not always align neatly with linguistic capacities of interest. We therefore urge future studies to move beyond asking how well a model fits a distribution toward asking where and why a breakdown would occur on the basis of their results.

Acknowledgements

The authors would like to thank Iona Carslaw, Nina Gregorio, Anna Kapron-King, Oli Liu, Burin Naowarat, Yen Meng, Katarzyna Pruś, and Sydelle de Souza for their feedback during the writing process. Additionally, this work has benefitted substantially from Dagstuhl Seminar 25301 "Linguistics and Language Models: What Can They Learn from Each Other?"—Coleman Haley would particularly like to thank the other members of the (Im)possible languages working group (Christopher Potts, Marie-Catherine de Marneffe, Katherine Demuth, Robert Frank, Juan Luis Gastaldi, Hagen Blix, Mark Johnson, Roger Levy, Kyle Mahowald, Mark Steedman, and Adina Williams) for the stimulating discussions that helped inspire this paper.

This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics and School of Philosophy, Psychology & Language Sciences.

References

Noam Chomsky. 1966. Explanatory Models in Linguistics. In *Studies in Logic and the Foundations of Mathematics*, volume 44, pages 528–550. Elsevier.

Noam Chomsky. 1980. Rules and representations. *Behavioral and Brain Sciences*, 3(1):1–15.

Noam Chomsky. 1998. On the Nature, Use and Acquisition of Language. In *Language and Meaning in Cognitive Science*. Routledge.

Noam Chomsky, Ian Roberts, and Jeffrey Watumull. 2023. Opinion | Noam Chomsky: The False Promise of ChatGPT. *The New York Times*.

Zhengxiao Du, Aohan Zeng, Yuxiao Dong, and Jie Tang. 2024. Understanding emergent abilities of language models from the loss perspective. In *Advances in neural information processing systems 38: Annual conference on neural information processing systems 2024 (NeurIPS 2024)*, Vancouver, BC.

Richard Futrell and Kyle Mahowald. 2025. How Linguistics Learned to Stop Worrying and Love the Language Models. *Behavioral and Brain Sciences*, pages 1–98.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2024. The Curious Decline of Linguistic Diversity: Training Language Models on Synthetic Text. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3589–3604, Mexico City, Mexico. Association for Computational Linguistics.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics.

Jennifer Hu, Ethan Gotlieb Wilcox, Siyuan Song, Kyle Mahowald, and Roger P. Levy. 2026. What Can String Probability Tell Us About Grammaticality? *Transactions of the Association for Computational Linguistics*, 14:124–146.

Michael Y. Hu, Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Ryan Cotterell, Leshem Choshen, Alex Warstadt, and Ethan Gotlieb Wilcox. 2024. Findings of the Second BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 1–21, Miami, FL, USA. 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.

Julie Kallini and Christopher Potts. 2025. Language models as tools for investigating the distinction between possible and impossible natural languages. *arXiv preprint*. ArXiv: 2512.09394 [cs.CL].

Simon Kirby, Hannah Cornish, and Kenny Smith. 2008. Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, 105(31):10681–10686.

Simon Kirby, Monica Tamariz, Hannah Cornish, and Kenny Smith. 2015. Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141:87–102.

Benjamin LeBrun, Alessandro Sordoni, and Timothy J. O'Donnell. 2022. Evaluating distributional distortion in neural language modeling. In *The tenth international conference on learning representations (ICLR 2022)*, virtual event. OpenReview.net.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Tal Linzen and Marco Baroni. 2021. Syntactic Structure from Deep Learning. *Annual Review of Linguistics*, 7:195–212.

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.

R. Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. 2023. How Much Do Language Models Copy From Their Training Data? Evaluating Linguistic Novelty in Text Generation Using RAVEN. *Transactions of the Association for Computational Linguistics*, 11:652–670.

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. 2008. *The Boundaries of Babel: The Brain and the Enigma of Impossible Languages*. The MIT Press.

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

Byung-Doh Oh and William Schuler. 2023. 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.

Wessel Poelman and Miryam de Lhoneux. 2026. Form and meaning in intrinsic multilingual evaluations. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2503–2521, Rabat, Morocco. Association for Computational Linguistics.

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.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Kenny Smith, Simon Kirby, Shangmin Guo, and Thomas L. Griffiths. 2024. AI model collapse might be prevented by studying human language transmission. *Nature*, 633(8030).

Taiga Someya, Anej Svete, Brian DuSell, Timothy J. O'Donnell, Mario Giulianelli, and Ryan Cotterell. 2025. Information Locality as an Inductive Bias for Neural Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27995–28013, Vienna, Austria. 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.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. 2024. Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*, 55(4):805–848.

Tianyang Xu, Tatsuki Kuribayashi, Yohei Oseki, Ryan Cotterell, and Alex Warstadt. 2026. Can language models learn typologically implausible languages? *Transactions of the Association for Computational Linguistics*, 14:588–611.

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.

Imry Ziv, Nur Lan, and Emmanuel Chemla. 2026. Biasless language models learn unnaturally: How LLMs fail to distinguish the possible from the impossible. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5393–5403, Rabat, Morocco. Association for Computational Linguistics.

A Details of Perplexity and 4-local Entropy

Perplexity: We computed the perplexity for each language on a sample of 7996 (1333 per subcorpus) sentences from perturbed versions of the BabyLM test dataset. We generated the perturbed versions by applying the perturbations to each line of the BabyLM test datasets.

4-local entropy: Someya et al. (2025) computed m-local entropy only for a subset of the languages we consider here, and on a different dataset, so we replicated their computations. We computed 4-local entropy on the 10 million sentence train set of BabyLM. This train set is a subset of the 100 million sentence set used by Kallini et al. (2024). We used `kenlm` Heafield (2011) to compute the 4-gram models over GPT-2 tokens, following Someya et al. (2025) in using default fallbacks for the Kneser-Ney smoothing (needed because there are no singleton unigrams after BPE tokenization). One limitation is we did not follow Kallini et al. (2024)'s sentence splitting for computing 4-local entropy, as it was very computationally expensive to run.

Language	Perplexity
ENGLISH	102.8
FULLREVERSE	123.8
PARTIALREVERSE	111.6
LOCALSHUFFLE3	164.3
LOCALSHUFFLE5	203.4
EVENODDSHUFFLE	210.1
LOCALSHUFFLE10	335.7
DETERMSHUFFLE	655.0
NONDETERMSHUFFLE	812.5

Table 4: Languages with their perplexity values

Language	m-local entropy
ENGLISH	2.856
FULLREVERSE	2.950
PARTIALREVERSE	3.219
LOCALSHUFFLE3	3.202
LOCALSHUFFLE5	3.262
EVENODDSHUFFLE	3.396
LOCALSHUFFLE10	3.442
DETERMSHUFFLE	3.590
NONDETERMSHUFFLE	4.470

Table 5: Languages with their m-local entropy value

B Perplexity and BLiMP Accuracy

See Figure 3.

C Model Generations

See Table 6.

D Data Statistics

See Table 7.

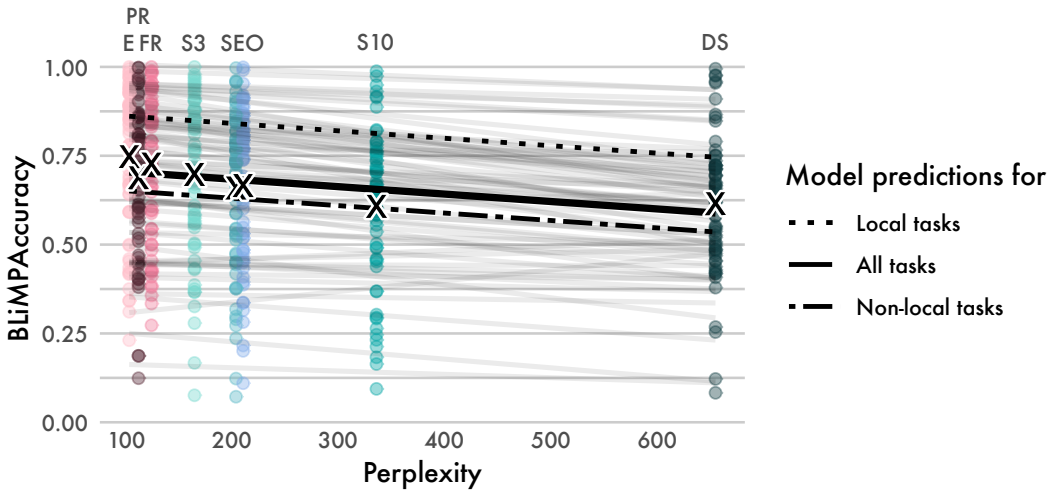

Figure 3: BLiMP task accuracy of models trained on English and impossible languages, correlated with test-set perplexity. Coloured dots and grey lines show accuracy and lines of best fit for individual tasks, with Xs indicating the mean for each language (labeled at top). The model indicates a modest, linear decline with increasing perplexity (fitted black line). This decline is insensitive to task locality: tasks that are solved well by a 5-gram model in English exhibit almost the same rate of degradation as those that are not (dotted vs. dashed line). However, we find 4-local entropy fits the data much better.

Language	**Low Perplexity (Higher Quality)**	**High Perplexity (Lower Quality)**
ENGLISH	*"Well, I really don't want to be there."*	*"And you,re a poor citizen who owns it?"*
	"I don't know what I'm gonna do this weekend, anyway."	*"He also plays for\xa0Derek and Wandaise."*
	"But I'm sorry, I didn't mean to."	*"[Tensemble.es and applause] (Gordon)"*
FULLREVERSE	*"But the way I look at it, there's no doubt about it."*	*"- Take her away! - I'm sorry.!"*
	"Oh, you're too much for me, aren't you?"	*"Watch out. - I really have to... get here."*
	"Yeah, because I haven't seen them for a long time."	*"At August 1, 18051 people lived there."*
PARTIALREVERSE	*"I don't know, I mean, that was a great idea."*	*"Alfactory (also called the Bin Malvernis A)."*
	"As a matter of fact, I've had nothing on it."	*"[sister music] [Mama chuckling]"*
	"I know what this is, but I want you to come with me."	*"[Hoor Help!RING]ONE RING]"*
LOCALSHUFFLE3	*"There are a lot of things that I want to do."*	*"No wonder it 'll have spoiled me now some day ."*
	'I don't know that you were right, but I'd rather not."	*"They were to decide that the exact sum of x "*
	"In 2016, she moved to Los Angeles, California, United States."	*"You were... in mind with him,! or not?"*
LOCALSHUFFLE5	*"You don't have to be able to get that job done, right?"*	*"That's for all, is the? for what."*
	"No, no, I believe I'm happy to say this."	*"Now,!'s on-by-door right here."*
	"All right, let's just have a look at the camera."	*"- [! in the background] - [man..."*
EVENODDSHUFFLE	*"And what the hell are you going to do with him?"*	*"M their and. friendsarks. theiroses come"*
	"In 2001, he was awarded the Nobel Prize in Chemistry."	*"Yes I,'t don we think do should. that"*
	"You really think it's going to happen, right?"	*"I you what wouldn said we't't I didn? ? suppose"*
LOCALSHUFFLE10	*'I've been on the bus at the time, it was my family."*	*"something'm you ask for I to thatI to."*
	"Now, I don't know it because I've found just about that."	*". place the most in the city ofThe is town"*
	"What were you trying to do with your work to my father?"	*"- Shut up! -!!!! (Come -.) on"*
DETERMSHUFFLE	-	*"the an good,, for way isIt same first?"*
	-	*"word a know withoutDon,'t about it man?"*
	-	*"is father like her your, then youDo?, -"*
NONDETERMSHUFFLE	-	*'of the. a isIt in university town located city"*
	-	*"..The people 6 2010 the city of population was"*
	-	*". thisBut very they a I,. when and a"*

Table 6: Qualitative comparison of model generations between 11 and 15 tokens long in all languages used in this study. The generations, after inverting perturbations where applicable, have perplexity scores below (left) and above (right) the 75% quantile of all English generations.

Table 7: Dataset statistics and local solvability scores (Continues on next page).

Phenomenon	Dataset UID	# pairs	Locally Solvable? (5-gram Score)
ANAPHOR AGREEMENT	anaphor_gender_agreement	1000	No (44)
ANAPHOR AGREEMENT	anaphor_number_agreement	1000	No (52)
ARGUMENT STRUCTURE	animate_subject_passive	723	No (70)
ARGUMENT STRUCTURE	animate_subject_trans	555	Yes (91)
ARGUMENT STRUCTURE	causative	581	No (54)
ARGUMENT STRUCTURE	drop_argument	412	No (72)
ARGUMENT STRUCTURE	inchoative	591	No (51)
ARGUMENT STRUCTURE	intransitive	456	No (68)
ARGUMENT STRUCTURE	passive_1	644	Yes (89)
ARGUMENT STRUCTURE	passive_2	607	Yes (82)
ARGUMENT STRUCTURE	transitive	550	No (71)
BINDING	principle_A_c_command	1000	No (58)
BINDING	principle_A_case_1	1000	Yes (100)
BINDING	principle_A_case_2	887	No (49)
BINDING	principle_A_domain_1	1000	Yes (95)
BINDING	principle_A_domain_2	1000	No (56)
BINDING	principle_A_domain_3	438	No (52)
BINDING	principle_A_reconstruction	1000	No (40)
CONTROL/RAISING	existential_there_object_raising	729	Yes (84)
CONTROL/RAISING	existential_there_subject_raising	850	No (77)
CONTROL/RAISING	expletive_it_object_raising	318	No (72)
CONTROL/RAISING	tough_vs_raising_1	933	No (33)
CONTROL/RAISING	tough_vs_raising_2	931	No (77)
DETERMINER-NOUN AGR.	determiner_noun_agreement_1	927	Yes (88)
DETERMINER-NOUN AGR.	determiner_noun_agreement_2	1000	Yes (86)
DETERMINER-NOUN AGR.	determiner_noun_agreement_irregular_1	703	No (53)
DETERMINER-NOUN AGR.	determiner_noun_agreement_irregular_2	1000	No (55)
DETERMINER-NOUN AGR.	determiner_noun_agreement_with_adjective_1	914	No (52)
DETERMINER-NOUN AGR.	determiner_noun_agreement_with_adj_2	1000	No (50)
DETERMINER-NOUN AGR.	determiner_noun_agreement_with_adj_irregular_1	818	No (53)
DETERMINER-NOUN AGR.	determiner_noun_agreement_with_adj_irregular_2	1000	No (55)
ELLIPSIS	ellipsis_n_bar_1	1000	No (23)
ELLIPSIS	ellipsis_n_bar_2	593	No (50)
FILLER GAP	wh_questions_object_gap	1000	No (53)
FILLER GAP	wh_questions_subject_gap	1000	Yes (82)
FILLER GAP	wh_questions_subject_gap_long_distance	1000	Yes (86)
FILLER GAP	wh_vs_that_no_gap	1000	Yes (83)
FILLER GAP	wh_vs_that_no_gap_long_distance	1000	Yes (81)
FILLER GAP	wh_vs_that_with_gap	1000	No (18)
FILLER GAP	wh_vs_that_with_gap_long_distance	1000	No (20)
IRREGULAR FORMS	irregular_past_participle_adjectives	1000	No (79)
IRREGULAR FORMS	irregular_past_participle_verbs	932	Yes (80)
ISLAND EFFECTS	adjunct_island	1000	No (48)
ISLAND EFFECTS	complex_NP_island	1000	No (50)
ISLAND EFFECTS	coordinate_structure_constraint_complex_left_branch	1000	No (32)
ISLAND EFFECTS	coordinate_structure_constraint_object_extraction	1000	No (59)
ISLAND EFFECTS	left_branch_island_echo_question	552	Yes (96)
ISLAND EFFECTS	left_branch_island_simple_question	1000	No (57)
ISLAND EFFECTS	sentential_subject_island	1000	No (61)
ISLAND EFFECTS	wh_island	1000	No (56)
NPI LICENSING	matrix_question_npi_licensor_present	634	No (1)
NPI LICENSING	npi_present_1	1000	No (47)

Table 7: Dataset statistics and local solvability scores (continued).

Phenomenon	Dataset UID	# pairs	Locally Solvable? (5-gram Score)
NPI LICENSING	npi_present_2	1000	No (47)
NPI LICENSING	only_npi_licensor_present	1000	No (57)
NPI LICENSING	only_npi_scope	762	No (30)
NPI LICENSING	sentential_negation_npi_licensor_present	1000	Yes (93)
NPI LICENSING	sentential_negation_npi_scope	1000	No (45)
QUANTIFIERS	existential_there_quantifiers_1	1000	Yes (91)
QUANTIFIERS	existential_there_quantifiers_2	1000	No (62)
QUANTIFIERS	superlative_quantifiers_1	1000	No (45)
QUANTIFIERS	superlative_quantifiers_2	1000	No (17)
SUBJECT-VERB AGR.	distractor_agreement_relational_noun	919	No (24)
SUBJECT-VERB AGR.	distractor_agreement_relative_clause	894	No (22)
SUBJECT-VERB AGR.	irregular_plural_subject_verb_agreement_1	864	No (73)
SUBJECT-VERB AGR.	irregular_plural_subject_verb_agreement_2	802	Yes (88)
SUBJECT-VERB AGR.	regular_plural_subject_verb_agreement_1	881	No (76)
SUBJECT-VERB AGR.	regular_plural_subject_verb_agreement_2	926	Yes (81)