Mission Accomplished?
Recovering Information from 'Impossible' Languages with LLMs

Anonymous ACL submission

Abstract

While there has been research into whether LLMs can learn linguistically impossible languages, it remains unclear whether they can systematically recover linguistic information from degraded input, and whether recovery difficulty reflects the degree of locality violation. We investigate whether LLMs can translate impossible languages back into possible forms and whether recovery difficulty varies by perturbation type. By fine-tuning GPT-2 models pretrained on impossible languages on three perturbation types that violate information locality to different degrees, we find that models can reconstruct grammatically well-formed output, with performance systematically modulated by perturbation nature. We further examine how training sample size and sentence length affect recovery: models trained on larger datasets show improved performance across all perturbations, though at different rates, while longer sentences provide richer contextual information that benefits perturbations with preserved structure but cannot overcome the fundamental difficulty when locality is completely disrupted. Our findings demonstrate that information locality violations create processing difficulty for LLMs, providing empirical support for information locality as a fundamental principle of efficient sequential processing rather than a categorical learnability boundary.

Anonymous Repository

1 Introduction

Human cognitive constraints determine which languages humans can learn, understand, and use (Chomsky, 1957, 1986; Jackendoff, 2002; Pinker, 1994). One goal of linguistics is to distinguish natural languages from other systems (de Saussure, 1983; Hockett, 1958; Lyons, 1968). Hence, linguists have proposed various descriptive frameworks, with some theories of universal grammar focusing on word order patterns and relationships among sentence elements, drawing on cross-linguistic typology (Greenberg, 1963; Hawkins, 1983; Dryer, 1992), formal semantics (Barwise and Cooper, 1981; Keenan and Stavi, 1986), syntactic theory (Vennemann, 1974; Primus, 2001; Hawkins, 1994), and generative grammar (Chomsky, 1981, 1995).

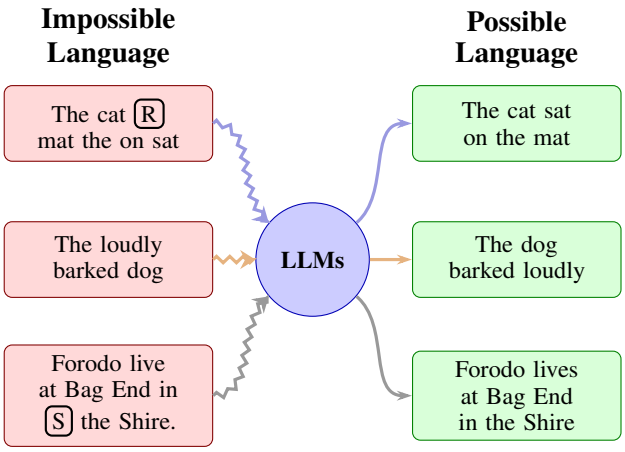


Figure 1: Translation task overview. Three separate LLMs are fine-tuned on specific perturbation types (from top to bottom PARTIALREVERSE, LOCALSHUFFLE, WORDHOP). Each model translates its corresponding impossible language input (left) into possible language output (right), recovering disrupted natural structure.

A central criticism of LLMs is that they fail to capture constraints defining human linguistic competence. Chomsky and colleagues argue that LLMs cannot distinguish between possible and impossible languages, as they can accommodate systems violating universal grammatical principles, thereby undermining their status as models of the human language faculty (Chomsky, 2023; Chomsky et al., 2023). Moro et al. (2023) argue that impossible languages exist only for humans, as neurobiological architecture constrains learnable grammars, whereas LLMs lack such constraints. Extending this view, Bolhuis et al. (2024) contend that LLMs do not model human language representation or acquisi-

tion, since their behaviour arises from statistical optimisation rather than biological mechanisms.

Although some studies argue that LLMs process impossible and possible languages similarly (Ziv et al., 2025), Kallini et al. (2024) demonstrate systematic learning differences, showing a clear bias toward natural language structures. This suggests that while LLMs can technically learn impossible languages, the principles defining human linguistic feasibility still influence their performance. To examine this possibility empirically, we focus on a well-established constraint in human language processing: *information locality*, which posits that semantically and syntactically related elements appear close together in linear order (Gibson et al., 2000; Futrell et al., 2020). The *Dependency Locality Theory* (DLT) formalizes this relation, predicting that processing difficulty increases as distance between dependent elements grows, reflecting cognitive constraints on working memory during incremental comprehension (Gibson, 1998). Cross-linguistic research supports this through dependency length minimisation (DLM), showing that natural languages systematically organise word order to keep grammatically related elements close (Hawkins, 1994; Liu, 2008; Futrell et al., 2020; Temperley, 2018). When information locality is violated through scattered dependent elements with intervening unrelated material, structures become cognitively inaccessible to humans despite intact lexical content, rendering text incomprehensible because information cannot be efficiently accessed (Levy, 2008; Hahn et al., 2020).

Some perturbations employed in this study systematically violate information locality. By scrambling word order or inserting distance-based markers, these transformations increase dependency lengths and disrupt local relationships, impeding incremental comprehension (Gibson et al., 2000; Liu, 2008). Since intelligible natural language requires related words to appear close together, translating an impossible language requires models to identify scattered syntactic/semantic dependencies and restore their characteristic proximity. Investigating whether language models can restore information locality from structurally degraded inputs provides theoretical insight into LLM architectural biases (Kallini et al., 2024; Xu et al., 2025; Tran et al., 2018; Khandelwal et al., 2018), while practically informing how systems process corrupted or noisy text. We investigate two core questions in this study: **First**, can LLMs, not subject to the same working memory limitations as humans, recover information obscured by locality violations? **Second**, and does recovery difficulty scale with violation severity?

To address this question, this study examines whether LLMs can recover linguistic information from perturbed inputs, where the perturbations amount to 'linguistically impossible' languages. Kallini et al. (2024) applied three distinct perturbation functions to the BabyLM dataset that systematically alter sentence structures, introducing linguistic inconsistencies absent in natural languages (Warstadt et al., 2023; Sinha et al., 2021). They subsequently trained separate GPT-2 models from scratch on data perturbed using these functions. We take models **pre-trained** by Kallini et al. as our starting point and fine-tune them to reconstruct valid forms (Radford et al., 2019; Ebrahimi et al., 2020). As Figure 1 illustrates, each model translates impossible language inputs into possible language outputs, recovering disrupted natural structure (one model for each impossible language).

Our experiments reveal three key findings. **First**, recovery performance varies systematically by perturbation type, directly reflecting the degree of locality violation. **Second**, dependency-based evaluation shows that models recover substantial syntactic structure even when exact reconstruction remains low, indicating that structural recovery precedes surface-level reconstruction. **Third**, longer sentences simultaneously facilitate recovery through richer contextual cues and hinder it by expanding the reordering search space. Together, these results suggest that LLMs are sensitive to information locality constraints but not fully bounded by them, exhibiting architectural biases partially aligned with human linguistic feasibility[1].

2 Background

2.1 Impossible Languages and Cognitive Constraints

Chomsky's Universal Grammar (UG) framework distinguishes possible from impossible human languages, proposing that all human languages share constrained structural principles from innate cognitive mechanisms (Chomsky, 1965). Crucially, formal inadequacy (whether grammatical formalisms describe natural language) differs from cognitive impossibility (whether humans can learn a system)

[1] All data and source code for this paper are available on GitHub

Language	Example 1	Example 2
ORIGINAL TEXT	It is nice in there	we 'd need to look at it again , would n't we
LOCALSHUFFLE	there It in is nice	we 'd need to it look again at , would n't we
PARTIALREVERSE	It is R there in nice	we 'd need R we n't would , again it at look to
WORDHOP	It be nice in there S	we 'd need to look at P it P again , would n't we P

Table 1: Example data for each perturbation function. For ease of comparison, we use a similar visualisation to Kallini et al. (2024).

(Moro, 2016; Musso et al., 2003). Impossible languages are defined by unlearnable grammars which humans cannot acquire despite exposure due to cognitive constraints, rather than failure to fit formal models. Empirical evidence suggests languages relying solely on position-based rules without hierarchical structure are impossible, conflicting with core syntactic processing mechanisms(Musso et al., 2003; Smith and Tsimpli, 1993).

Nefdt (2024) argues identifying possible grammar boundaries requires grounding in theoretical linguistic competence, not solely distributional regularities. Smith and Tsimpli (1993) showed successful acquisition of artificial languages consistent with universal principles but failure to learn one violating them, suggesting certain grammars remain unlearnable. Neurolinguistic studies show brain regions for syntactic processing selectively respond to natural language-compatible structures (Moro, 2016; Musso et al., 2003). However, empirical evidence remains remarkably limited. Smith and Tsimpli (1993) focus on a single individual and most theoretical claims lack systematic empirical testing. This gap highlights computational approaches' potential value, as LLMs offer systematic ways to test which linguistic systems are truly unlearnable, providing empirical grounding for theoretical claims about possible language boundaries.

2.2 LLMs as Models of Linguistic Competence

Against this theoretical and cognitive backdrop, artificial neural networks examined as linguistic competence models reveal important discrepancies. Hahn (2020) shows standard self-attention architectures face intrinsic limitations modelling unbounded hierarchical dependencies, challenging their adequacy as full natural language syntax models. Mitchell and Bowers (2020) demonstrate that recurrent neural networks learn highly unnatural patterns humans do not acquire, highlighting misalignment between human and neural inductive biases. While approximating complex string distributions, they are not automatically constrained by possible human languages and may generalise divergently from human cognition.

Studies of transformer capabilities show mixed results. Abdou et al. (2022) show that models trained on shuffled text retain non-trivial word order information, while Huang et al. (2023) demonstrate that lexically invariant transformer variants match standard models under long contexts. Ebrahimi et al. (2020) show that self-attention networks learn certain hierarchical languages with appropriate cues, though Delétang et al. (2023) argue standard architectures face principled challenges representing complex hierarchical dependencies without explicit structured memory.

Recent research comparing human learners and LLMs shows that both benefit from compositional, structured input (Galke et al., 2024), with transformers partially encoding hierarchical phrase structure though not fully matching generative grammar expectations (Alleman et al., 2021). Kallini et al. (2024) demonstrate that LLMs exhibit systematic learnability differences between possible and impossible languages, displaying biases toward natural patterns while struggling with highly unnatural systems. Xu et al. (2025) find that GPT-2 shows learning slowdowns aligned with typological universals when acquiring implausible counterfactual languages, though it can ultimately learn them. While LLMs can learn highly unnatural patterns, their learning is shaped by architectural and data-driven biases that only partially overlap with the cognitive constraints defining human language, motivating examination of whether LLMs can systematically recover possible languages from impossible ones by reversing perturbations violating natural linguistic structure.

3 Method

The main goal of this research is to train an LLM to translate impossible language into possible lan-

```
"Fix this text:  The cat [R] mat the on sat\nCorrected:  The cat sat on the mat"
<[-100][-100][-100][-100]...[-100][-100][-100][-100]>        <tokens>
Ignored in loss                                              Learned from
```

Figure 2: The mechanism of masking labels and calculating loss

guage. Formally, we define a perturbation function $f : A \rightarrow A'$ where A represents standard English text and A' is the perturbed version. Our objective is to train a model $\mathcal{M}$ such that $\mathcal{M}(A') := A$, effectively reversing the perturbation and recovering the original linguistic information.

A critical constraint is that $\mathcal{M}$ must not have been pre-trained on A. If $\mathcal{M}$ had prior exposure to standard English, it could leverage that knowledge rather than genuinely learning to recover linguistic structure from perturbed input, confounding our evaluation of information recovery. To satisfy this requirement, we adopt **the models pre-trained by** Kallini et al. (2024) as the starting point for our fine-tuning process, since these models were trained from scratch on impossible languages, generated by applying different perturbation functions to standard English sentences.

3.1 Data Generation

We use the *bnc_spoken* and *Gutenberg* subsets of BabyLM (Warstadt et al., 2023). These subsets allow us to analyse the impact of contrasting sentence lengths on translation performance. Applying perturbations to these subsets yields paired examples (perturbed, original) for fine-tuning. Following Kallini et al. (2024), we apply three perturbation functions (Table 1).

LOCALSHUFFLE: This perturbation randomly reorders words within a local window (size 3 in this study), disrupting the sequential arrangement while maintaining words within bounded distances.

PARTIALREVERSE: A random starting point is selected within the sentence, and an R token is placed in this position, and subsequent tokens are reversed in order. This creates a partition where the initial segment remains unchanged, the final segment is reversed, and R marks the boundary.

WORDHOP: This perturbation violates the principle that no human language requires counting words for grammatical operations (Musso et al., 2003) by adding markers (S for singular, P for plural) at fixed distances after verbs based on subject-verb agreement.

These three perturbations, selected from Kallini et al. (2024), span a spectrum of locality violation: LOCALSHUFFLE destroys local order within windows, PARTIALREVERSE disrupts order from a random midpoint, and WORDHOP preserves order entirely, adding only counting-based markers.

We created three datasets, each containing paired text (impossible, possible), by applying one perturbation strategy to each. Preprocessing included tokenization, lemmatization, POS tagging, and constituency parsing, enabling grammatical structure-dependent perturbations such as WORDHOP's verb phrase identification and marker placement. To investigate training scale effects, we use datasets of $100K$ sample pairs from *bnc_spoken*.

3.2 Model Fine-tuning

The models were fine-tuned by combining the causal language modelling paradigm of GPT-2 (Radford et al., 2019) with a masked label strategy. Each training sample consists of a pair: an impossible text and its corresponding possible version. These pairs are formatted into the following instruction-response structure:

```
Fix this text:  <impossible_text>
Corrected:  <possible_text><|endoftext|>
```

To ensure the model learns to generate corrections rather than memorising instructions, masked labelling is applied. The instruction tokens (sequence start to `Corrected`) receive label value $[-100]$, while the response tokens (correct text and `<|endoftext|>` token) retain their IDs as labels. By excluding the input from the loss calculation, the model is still presented with the full input during the forward pass, but the loss is computed only over the target (correct) tokens (Ouyang et al., 2022). The GPT-2 tokenizer was extended with special tokens {R, S, P} before fine-tuning to accommodate perturbation markers. Figure 2 displays the label masking and loss calculation.

3.3 Evaluation metrics

To assess how well the model recovers syntactic structure from perturbed input, we employ Exact Match (EM), measuring perfect reconstruction (Rajpurkar et al., 2016); and the BLEU score quantifying n-gram overlap (Papineni et al., 2002). We also consider two complementary metrics based on dependency triples:

Average Dependency Length (ADL). We compute the mean absolute positional distance between each non-root token and its syntactic head: $\text{ADL} = \frac{1}{n}\sum_{i=1}^{n} |\text{pos}(w_i) - \text{pos}(\text{head}(w_i))|$, where n is the number of non-root tokens. Scrambled text disrupts this locality, producing higher ADL.

Dependency Triple F1. From each parse, we extract triples (w_d, r, w_h) where w_d is the dependent, r is the relation label, and w_h is the head. We compute set-based precision, recall, and F1 between the triple sets in the model's prediction and the original (unperturbed) string. A triple matches only if all three elements are identical. As a baseline, we also compare the scrambled input against the actual text to quantify how many arcs survive perturbation by chance (Buchholz and Marsi, 2006; Cai and Knight, 2013).

3.4 Experimental Design

We conduct three experiments testing whether reconstruction difficulty scales with locality violation severity. We hypothesise that perturbations more severely disrupting the proximity of related elements should be harder to reverse, reflecting a gradient rather than a categorical boundary. Our perturbations systematically violate this by scattering related words across sentences, altering the normal subject-verb agreement rule. If LLMs exhibit locality constraints, translation performance should degrade as violations intensify: longer sentences increase difficulty by expanding distances between related elements and enlarging reordering search spaces, while perturbation types should vary by locality disruption severity. For full fine-tuning details see A.

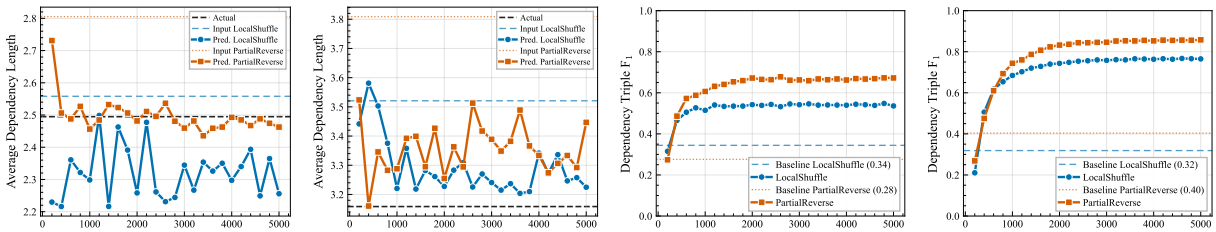
Experiment 1: Effect of Training Sample Size. This experiment investigates how training data size influences translation performance. Models were fine-tuned on the $100K$ subset of *bnc_spoken*, with learning progress evaluated at multiple checkpoints using our four metrics (§ 3.3). This multi-metric approach suits our task because impossible and possible texts share vocabulary but differ in token order, making both strict reconstruction (EM) and partial accuracy (BLEU, Triple F1), and structural locality (ADL) informative. The hypothesis is that larger datasets will improve performance across all perturbations and metrics, but with different learning rates: WORDHOP should be learned fastest due to preserved word order, while LOCALSHUFFLE and PARTIALREVERSE will require more examples to recover disrupted locality.

Dataset		Size	Average Length
bnc_spoken	train	$100K$	12.77
bnc_spoken	test	$1K$	11.72
Gutenberg	train	$100K$	40.61
Gutenberg	test	$1K$	46.78

Table 2: Average sentence lengths for the training and test datasets.

Experiment 2: Effect of Sentence Length. Building on the locality principle introduced above, this experiment tests how sentence length affects recovery difficulty. Results were compared between the *bnc_spoken* dataset and the *Gutenberg* dataset (see Table 2). The hypothesis is that longer sentences provide more training signal, potentially improving overall performance, but simultaneously increase difficulty for perturbations disrupting information locality. Specifically, LOCALSHUFFLE and PARTIALREVERSE are expected to show complex effects: benefits from additional training content may be offset by challenges from related words scattered over greater distances, which exponentially expand the space of possible reorderings and integration costs. WORDHOP is expected to benefit more straightforwardly from longer sentences since it preserves underlying word order and only adds markers at fixed distances, maintaining most local dependencies intact.

Experiment 3: Quality of Generated Text. Because the goal of this research is to translate disturbed languages and recover information, this experiment evaluates whether generated texts converge toward natural language distributions using perplexity. We measure against GPT-2 base on standard English to measure how well generated text aligns with natural language distributions (Jelinek et al., 2005). Lower perplexity value indicates greater consistency with natural language distributions (Radford et al., 2019). We hypothesise that prediction perplexity will converge toward that of the original text as training progresses, with rates



(a) *ADL - bnc_spoken* (b) *ADL - Gutenberg* (c) Triple F_1 - *bnc_spoken* (d) Triple F_1 - *Gutenberg*

Figure 3: Dependency-based evaluation across perturbation types and datasets. (a), (b) Average Dependency Length: predictions converge toward the natural baseline on both *bnc_spoken* (ADL = 2.50) and *Gutenberg* (ADL = 3.10), with higher variance for LOCALSHUFFLE on longer sentences. (c), (d) Dependency Triple F_1: both perturbation types surpass their scrambled-input baselines, with higher scores on *Gutenberg* (LOCALSHUFFLE ≈ 0.84, PARTIALREVERSE ≈ 0.81) than *bnc_spoken* (0.67 and 0.62), reflecting greater benefit from longer, well-structured contexts. For larger plots, see Figure 6

varying by perturbation type. Specifically, LOCALSHUFFLE and PARTIALREVERSE predictions are expected to show steeper perplexity reductions as these perturbations more severely disrupt natural word order, while WORDHOP predictions maintain slightly higher perplexity even after convergence due to artifacts from marker removal that do not fully restore natural linguistic flow.

4 Experimental results

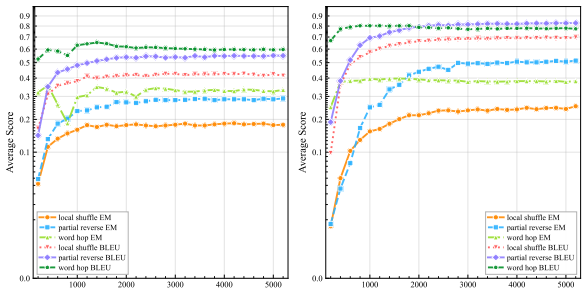
Experiment 1: Figure 4a shows that the number of fine-tuning steps critically affects performance. In the early stages of training, where the model has been exposed to fewer samples, both accuracy and output quality are notably lower. As training progresses and the model processes more data, performance improves steadily across all perturbation types, though at varying rates: WORDHOP rapidly reaches high performance (0.6 EM), reflecting the tractability of marker removal when word order is preserved, while LOCALSHUFFLE plateaus at a low ceiling (0.18 EM) even after extended training, indicating that additional fine-tuning steps cannot fully overcome the difficulty of recovering severely disrupted locality.

The dependency-based evaluation (Figures 3a and 3c) reveals that surface metrics underestimate the model's structural recovery. Despite its low exact match, LOCALSHUFFLE achieves Dependency Triple $F_1 \approx 0.55$ (Figure 3c), well above the scrambled-input baseline of 0.34, while PARTIALREVERSE reaches $F_1 \approx 0.58$ against a baseline of 0.28. This indicates that the model recovers syntactic relationships even when full string reconstruction remains elusive, and that PARTIALREVERSE's lower baseline reflects greater disruption of dependency arcs despite its higher EM. The ADL analysis (Figure 3a) shows that PARTIALREVERSE predictions converge steadily toward the actual ADL ($\approx$2.49), whereas LOCALSHUFFLE predictions fall below both the input and actual baselines, suggesting the model produces simpler-than-natural syntactic structures when facing severe locality disruption. Qualitative examples illustrating the learning progression at different checkpoints are provided in Appendix (Tables 3, 4, and 5).

Experiment 2: Table 2 shows average sentence lengths for both datasets. Comparing final performance (Figure 4) reveals substantial sentence length effects. On *Gutenberg* versus *bnc_spoken*, WORDHOP improves from 60% to 78% BLEU, PARTIALREVERSE from 55% to 82%, while LOCALSHUFFLE rises only from 18% to 25% EM (see also Appendix Figure 9).

Dependency-based evaluation reveals this gap is narrower at the structural level. On *Gutenberg*, LOCALSHUFFLE and PARTIALREVERSE achieve F_1 scores of 0.75 and 0.80, compared to 0.55 and 0.58 on *bnc_spoken* (Figures 3c and 3d). This likely reflects the nature of the source text: written language provides richer dependency cues through longer, well-formed sentences, whereas spoken data contains fragmented utterances and disfluencies that leave fewer structural signals. ADL analysis (Figures 3a and 3b) supports this: on *Gutenberg*, predictions converge toward the actual ADL ($\approx$3.16), while on *bnc_spoken*, LOCALSHUFFLE over-minimises dependency lengths below baseline, producing simpler-than-natural structures.

This indicates that WORDHOP's preserved word order and PARTIALREVERSE's partial structure enable better exploitation of longer sentences, whereas LOCALSHUFFLE's complete locality disruption presents fundamental recovery difficulties. Longer sequences increase dependency distances



(a) *bnc_spoken* (b) *Gutenberg*

Figure 4: Training curves comparing short (a) versus long (b) sentences. Longer contexts improve all perturbations, with PARTIALREVERSE showing largest gains. For larger plots, see Figure 7

and enlarge the reordering search space, consistent with prior work on long-range dependency limitations (Khandelwal et al., 2018; Hahn, 2020; Tran et al., 2018) and degraded order information under global disruption (Sinha et al., 2021; Abdou et al., 2022).

Experiment 3: Our models are fine-tuned from GPT-2(models pre-trained by Kallini et al. (2024)), we use the pre-trained (base) GPT-2 model for evaluation to measure how well the generated text aligns with general language patterns learned from large-scale pretraining data: lower perplexity indicates that the text is more consistent with natural language distributions.

Figure 5 compares perturbed language inputs with model predictions against baseline unperturbed text (perplexity = 129.34). LOCALSHUFFLE input exhibits extremely high perplexity (3902.34), which predictions reduce to 212.40 (94.6% decrease), demonstrating successful reconstruction approaching natural language distributions, though still remaining above baseline, which indicates incomplete restoration. PARTIALREVERSE shows a similar pattern: input perplexity (1500.38) reduces to 146.56 (90.2% reduction), coming closer to baseline than LOCALSHUFFLE, consistent with its higher exact match scores (0.29 vs. 0.18). WORDHOP presents a contrasting case. Input perplexity (168.11) is already near baseline since word order is preserved, yet predictions show increased perplexity (243.81, +45.0%), suggesting that while marker removal succeeds, the reconstruction introduces subtle artefacts, likely from imperfect verb agreement handling. Concrete examples at different training stages are provided in Appendix Tables 3, 4, and 5.

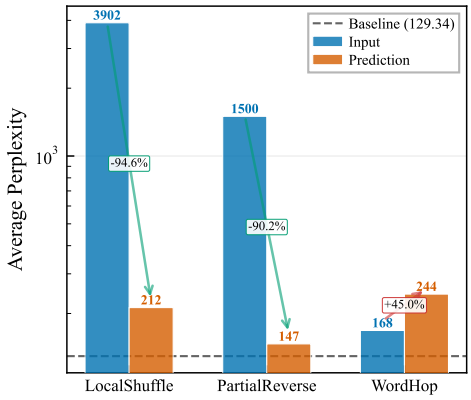


Figure 5: Perplexity of impossible inputs versus model predictions relative to baseline unperturbed text (129.34, dashed line), evaluated using pretrained GPT-2.

5 Discussion

We investigate whether LLMs can recover accessible natural linguistic structure from systematically degraded input, and whether recovery difficulty scales with the degree of locality violation. Our experiments with GPT-2, pre-trained on impossible languages by Kallini et al. (2024) and fine-tuned on three perturbation types (LOCALSHUFFLE, PARTIALREVERSE, WORDHOP), reveal important findings about neural language models' capabilities and limitations in recovering linguistically possible forms from impossible ones.

Information Locality. Our results demonstrate that model performance directly reflects the degree to which perturbations violate information locality (Gibson et al., 2000). LOCALSHUFFLE and PARTIALREVERSE, which disperse grammatically related elements and create non-local dependencies, proved most difficult, while, WORDHOP, which preserves underlying word order, achieved substantially higher performance despite introducing distance-based markers.

The Dependency Triple F_1 analysis reinforces this beyond surface-level metrics, showing that the model recovers a substantial portion of correct syntactic relationships even when full string reconstruction remains elusive. This recovery is markedly stronger on longer texts, despite the greater dependency distances introduced by perturbation. This suggests that the richer contextual signal available in longer, syntactically complete sentences outweighs the increased difficulty of the reordering task, providing the model with stronger cues for restoring locality during training. Perplexity analysis indicates that models learn generalisable structural recovery rather than

memorized mappings, with dramatic reductions for LOCALSHUFFLE (94.6%) and PARTIALREVERSE (90.2%) toward the baseline indicating successful reconstruction of locality-respecting configurations (Hahn et al., 2020; Hahn, 2020), though WORDHOP's increased perplexity reflects both marker removal artifacts and the distinct lexical distributions introduced by its design.

Performance across perturbation functions. GPT-2 performs information recovery with performance varying systematically by perturbation type. WORDHOP proved easiest (> 0.6 EM), preserving underlying word order, while LOCALSHUFFLE and PARTIALREVERSE plateaued around 0.18 and 0.29 EM, requiring reconstruction of hierarchical structures where local dependencies become non-local. This aligns with theoretical arguments about transformer limitations in capturing unbounded hierarchical dependencies (Tran et al., 2018; Delétang et al., 2023; Hahn, 2020; Khandelwal et al., 2018). Dependency-based evaluation reveals that the gap between perturbation types is narrower at the structural level than surface metrics suggest, indicating meaningful syntactic knowledge acquisition even for the hardest perturbations. The differential performance reflects each perturbation's distinct characteristics. LOCALSHUFFLE and PARTIALREVERSE maintain standard English vocabulary with reordered tokens. In contrast, WORDHOP uses lemmatized verbs and agreement markers, preserving sequential structure but altering lexical distributions.
The pattern of difficulty (WORDHOP easiest, LOCALSHUFFLE hardest) mirrors what Kallini et al. (2024) found for learning these languages, suggesting the same architectural biases constrain both acquisition and translation. Perplexity analysis confirms genuine structural recovery: prediction perplexity for LOCALSHUFFLE and PARTIALREVERSE decreased toward unperturbed text, producing sequences increasingly consistent with pretrained GPT-2 distributions (Radford et al., 2019; Holtzman et al., 2020).

Importance of sentence length. Our second experiment revealed strong sentence length effects relating to information locality and data availability. Models trained on *Gutenberg* outperform *bnc_spoken* as longer sentences provide more training content. However, longer sequences exponentially increase possible reorderings and distances between originally adjacent elements (Gibson, 1998), increasing integration costs for reconstructing relationships as processing costs grow with dependency length (Futrell et al., 2020; Gibson et al., 2000). The model must recover information over longer distances while identifying which scattered elements belong together, similar to the challenge of processing deeply nested structures (Levy, 2008). This reveals important implications for LLM inductive biases. While longer sequences provide more training signal and thereby enable better performance, they simultaneously present greater challenges for recovering non-local dependencies. Though LLMs lack human cognitive-neurobiological constraints, they exhibit architectural constraints where learning efficiency balances data availability against dependency distance, creating biases partially aligned with natural language properties and preferring local over non-local dependencies (Futrell et al., 2020; Khandelwal et al., 2018). This balance between richer signal and greater reordering difficulty mirrors the trade-off observed across perturbation types, reinforcing that locality violation severity is the central factor shaping reconstruction performance.

6 Conclusion

This study demonstrates that LLMs can translate inputs which are perturbed to represent 'linguistically impossible' sequences, into possible forms by recovering information obscured through locality violations. Performance varies systematically (WORDHOP easiest, LOCALSHUFFLE hardest), with dependency-based evaluation revealing substantial syntactic recovery even when exact reconstruction fails.

Our findings show that recovery difficulty scales systematically with locality violation severity, providing empirical support for information locality as a fundamental principle of efficient sequential processing rather than a categorical learnability boundary. LLMs and human language processing share sensitivity to information locality, though underlying mechanisms differ, with architectural biases partially aligned with cognitive constraints on human linguistic competence. This has practical applications in error correction, text normalisation, and robustness to noisy input, while raising questions about the relationship between processing constraints and learnability boundaries in both artificial and human language systems.

7 Limitations

It is important to acknowledge several limitations of this study. First, our experiments focus on a relatively small model (GPT-2) and a limited set of perturbation functions. Larger, more recent architectures with enhanced capabilities for modeling long-range dependencies might exhibit different learning patterns or overcome some of the limitations we observed.

Second, our impossible languages, while designed to violate linguistic universals, represent only a small subset of the space of possible perturbations. Future work could explore a broader range of impossible language constructions, including those that violate other aspects of linguistic structure beyond word order, and evaluate whether consistent patterns emerge in model behavior. Additionally, investigating whether models can generalize their translation abilities to novel perturbation types or combinations would provide insight into whether they learn abstract principles of information locality recovery or task-specific mappings.

The notion of "impossible language" itself requires careful interpretation. While linguistic theory identifies certain systems as impossible for human acquisition based on violations of information locality and other neurobiological and cognitive constraints, these systems remain well-defined computational problems that neural networks can, in principle, solve. Our results highlight an important distinction between cognitive impossibility, which arises from violations of information locality, and computational possibility. Recognizing this gap is essential for understanding both the potential and the limits of LLMs as models of human language and as practical language processing tools.

References

Mostafa Abdou, Vinit Ravishankar, Artur Kulmizev, and Anders Søgaard. 2022. Word order does matter (and shuffled language models know it). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*, pages 6904–6919.

Matteo Alleman, Jonathan Mamou, Miguel A. Del Rio, Hanlin Tang, Yoon Kim, and SueYeon Chung. 2021. Syntactic perturbations reveal representational correlates of hierarchical phrase structure in pretrained language models. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 263–276.

Jon Barwise and Robin Cooper. 1981. Generalized quantifiers and natural language. *Linguistics and Philosophy*, 4(2):159–219.

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

Sabine Buchholz and Erwin Marsi. 2006. Conll-x shared task on multilingual dependency parsing. In *Proceedings of the tenth conference on computational natural language learning (CoNLL-X)*, pages 149–164.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752, Sofia, Bulgaria. Association for Computational Linguistics.

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

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

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

Noam Chomsky. 1986. *Knowledge of Language: Its Nature, Origin, and Use*. Praeger, New York.

Noam Chomsky. 1995. *The Minimalist Program*. MIT Press, Cambridge, MA.

Noam Chomsky. 2023. Noam chomsky on language, left libertarianism, and progress. Podcast episode. Conversations with Tyler, Episode 182.

Noam Chomsky, Ian Roberts, and Jeffrey Watumull. 2023. The false promise of chatgpt. *The New York Times*. Guest essay, March 8, 2023.

Ferdinand de Saussure. 1983. *Course in General Linguistics*. Duckworth, London. R. Harris (Trans.). Original work published 1916.

Grégoire Delétang, Anian Ruoss, Jordi Grau-Moya, Tim Genewein, Kevin Li Wenliang, Elliot Catt, Chris Cundy, Marcus Hutter, Shane Legg, Joel Veness, and Pedro A. Ortega. 2023. Neural networks and the chomsky hierarchy. In *Proceedings of the International Conference on Learning Representations (ICLR 2023)*.

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

Javid Ebrahimi, Dhruv Gelda, and Wei Zhang. 2020. How can self-attention networks recognize dyck-n languages? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4307–4313.

Richard Futrell, Roger Levy, and Edward Gibson. 2020. Dependency locality as an explanatory principle for word order. *Language*, 96(2):372–413.

Lukas Galke, Yoav Ram, and Limor Raviv. 2024. Deep neural networks and humans both benefit from compositional language structure. *Nature Communications*, 15:10816.

Edward Gibson. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Edward Gibson et al. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*, 2000:95–126.

Joseph H. Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg, editor, *Universals of Language*, pages 73–113. MIT Press, Cambridge, MA.

Michael Hahn. 2020. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171.

Michael Hahn, Dan Jurafsky, and Richard Futrell. 2020. Universals of word order reflect optimization of grammars for efficient communication. *Proceedings of the National Academy of Sciences*, 117(5):2347–2353.

John A. Hawkins. 1983. *Word Order Universals*. Academic Press, New York.

John A. Hawkins. 1994. *A Performance Theory of Order and Constituency*. Cambridge University Press, Cambridge.

Charles F. Hockett. 1958. *A Course in Modern Linguistics*. Macmillan, New York.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Qian Huang, Eric Zelikman, Sarah Li Chen, Yuhuai Wu, Gregory Valiant, and Percy Liang. 2023. Lex-invariant language models. In *Advances in Neural Information Processing Systems*.

Ray Jackendoff. 2002. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press, Oxford.

F. Jelinek, R. L. Mercer, L. R. Bahl, and J. K. Baker. 2005. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

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 (ACL 2024)*, pages 14691–14714.

Edward L. Keenan and Jonathan Stavi. 1986. A semantic characterization of natural language determiners. *Linguistics and Philosophy*, 9(3):253–326.

Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. 2018. Sharp nearby, fuzzy far away: How neural language models use context. In *Proceedings of ACL 2018*, pages 284–294.

Roger Levy. 2008. Expectation-based syntactic comprehension. In *Proceedings of the ACL-HLT 2008*, pages 1047–1054.

Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159–191.

John Lyons. 1968. *Introduction to Theoretical Linguistics*. Cambridge University Press, Cambridge.

Jennifer J. Mitchell and Jeffrey S. Bowers. 2020. Priorless recurrent networks learn curiously. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*, pages 5002–5015.

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

Andrea Moro, Matteo Greco, and Stefano F. Cappa. 2023. Large languages, impossible languages and human brains. *Cortex*. Advance online publication.

Mariacristina Musso, Andrea Moro, Volkmar Glauche, Michel Rijntjes, Jürgen Reichenbach, Christian Büchel, and Cornelius Weiller. 2003. Broca's area and the language instinct. *Nature Neuroscience*, 6(7):774–781.

Ryan M. Nefdt. 2024. *The Philosophy of Linguistics*. Cambridge University Press, Cambridge.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Steven Pinker. 1994. *The Language Instinct*. William Morrow, New York.

Beatrice Primus. 2001. Word order typology. In Martin Haspelmath et al., editors, *Language Typology and Language Universals*, volume 2, pages 855–875. De Gruyter Mouton, Berlin.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

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

Neil V. Smith and Ianthi-Maria Tsimpli. 1993. The acquisition of possible and impossible languages by a polyglot savant. *Lingua*, 91(1):279–347.

David Temperley. 2018. Dependency-length minimization and the link between word order and cognition. *Cognitive Science*, 42(8):3039–3061.

Ke Tran, Arianna Bisazza, and Christof Monz. 2018. The importance of being recurrent for modeling hierarchical structure. In *Proceedings of EMNLP 2018*, pages 4731–4736.

Theo Vennemann. 1974. Theoretical word order studies: Results and problems. *Papiere zur Linguistik*, 7:5–25.

Alex Warstadt, Leshem Choshen, Aaron Mueller, Adina Williams, Ethan Wilcox, and Chengxu Zhuang. 2023. Call for papers – the babylm challenge: Sample-efficient pretraining on a developmentally plausible corpus.

Tianyang Xu, Tatsuki Kuribayashi, Yohei Oseki, Ryan Cotterell, and Alexander Scott Warstadt. 2025. Can language models learn typologically implausible languages? *ArXiv*, abs/2502.12317.

Imry Ziv, Nur Lan, Emmanuel Chemla, and Roni Katzir. 2025. Biasless language models learn unnaturally: How llms fail to distinguish the possible from the impossible. *arXiv preprint arXiv:2510.07178*.

A Fine-tuning Implementation Details

All models were fine-tuned using Hugging Face Transformers (v4.35.0) with PyTorch 2.1.0, employing the masked instruction-following approach described in Section 3.2.

Batch size: Effective batch size of 512 (per-device 128 with 4 gradient accumulation steps).

Optimization: AdamW optimizer with learning rate 2×10^{-5}, cosine schedule, 500 warmup steps, weight decay 0.01, and gradient clipping at norm 1.0.

Training steps: 200 steps (10K dataset), 5,000 steps (100K dataset), determined by validation loss convergence.

Precision: Mixed precision training with bfloat16 and TensorFloat-32 on NVIDIA H100 GPUs.

Regularization: Label smoothing factor of 0.1.

Checkpointing: Every 20 steps (10K) or 200 steps (100K). Best model selected by lowest evaluation loss on 10% held-out validation set.

Reproducibility: Random seed 42 for all experiments.

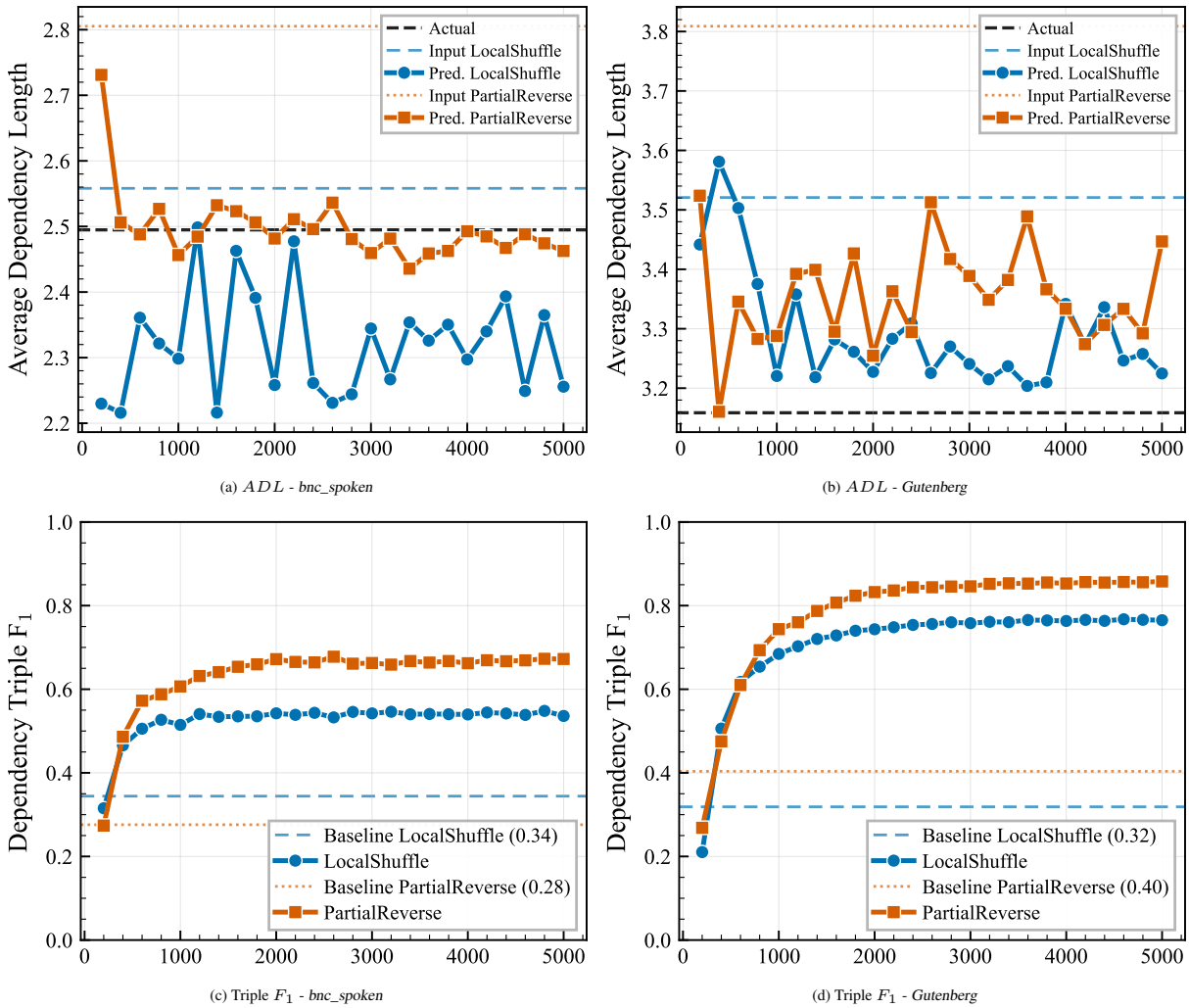
B Larger Plots

Because the images embedded in the main text are relatively small, this appendix provides enlarged versions to improve readability and highlight finer details.

C Examples

During training, we evaluated model performance at multiple checkpoints to monitor learning progress. We tracked generated text quality across checkpoints for each impossible language type. This section provides examples illustrating the improvement trajectory throughout training (Tables 3, 4, 5).

Additionally, we investigate how performance varies by input text length for the LocalShuffle perturbation. Since LocalShuffle uses a window size of 3, we classify input texts by the number of shuffling windows they contain. Specifically, we define $\text{class}(text_i) = \lceil \text{len}(text_i)/3 \rceil$, where $\text{len}(text_i)$ is the number of tokens in the text. This groups texts by how many 3-token windows were shuffled during perturbation. Figure 9 shows model improvement across these classes, revealing how performance scales with the number of shuffled segments.



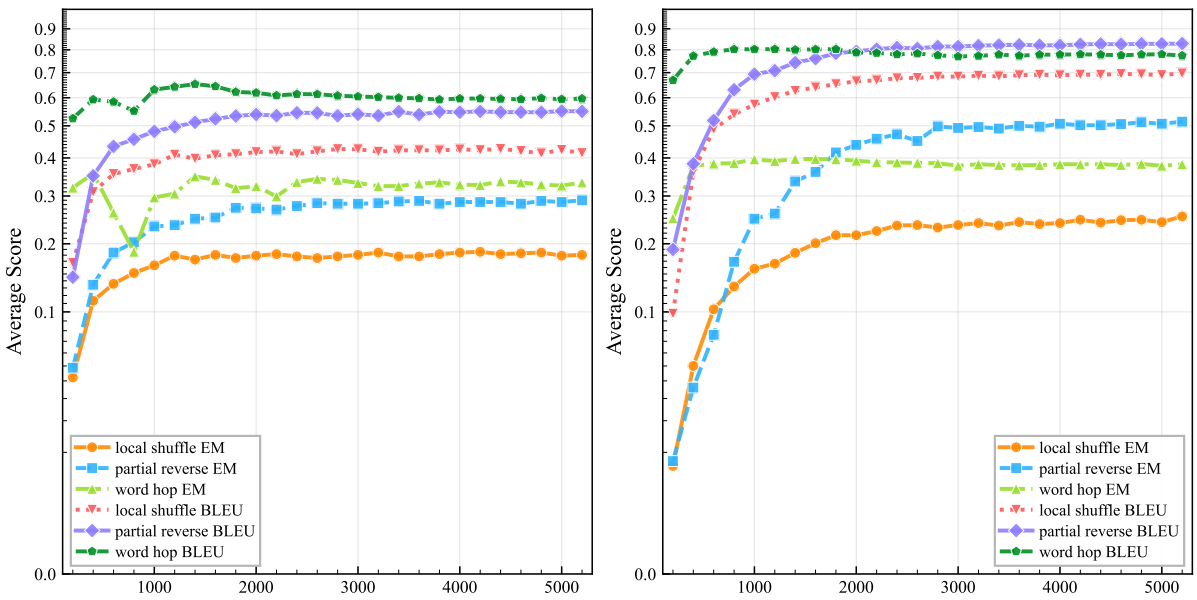
(a) ADL - *bnc_spoken*

(b) ADL - *Gutenberg*

(c) Triple F_1 - *bnc_spoken*

(d) Triple F_1 - *Gutenberg*

Figure 6: A bigger size of the plots in 3



(a) *bnc_spoken*

(b) *Gutenberg*

Figure 7: A bigger size of the plots in 4

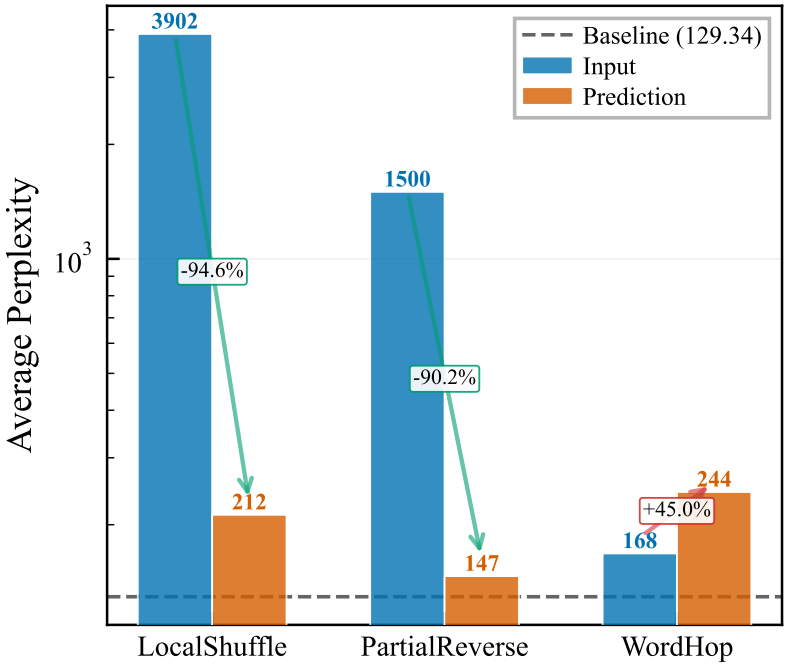


Figure 8: A bigger size of the plots in 5.

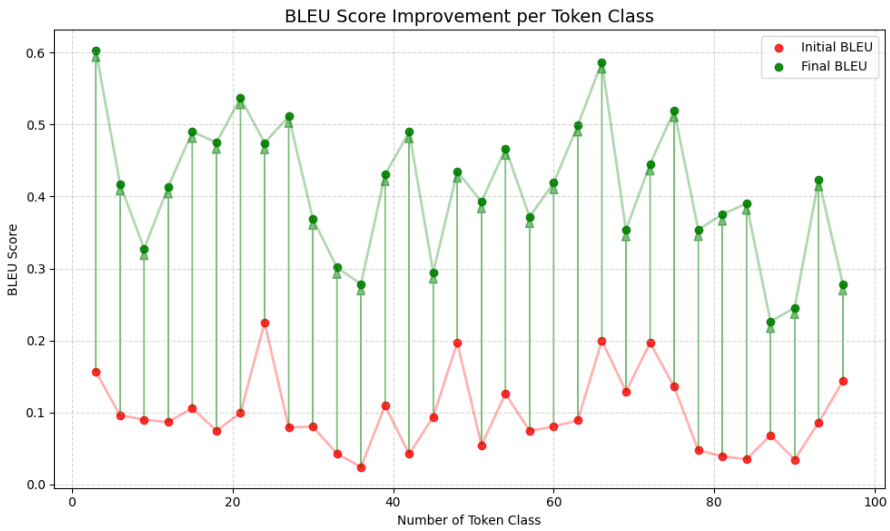


Figure 9: Improvement of the BLEU score for each class. The red line corresponds to the BLEU score before translating and for shuffled text, and the green line is for texts after translation by the final model.

Table 3: Learning process for translation LOCALSHUFFLE at different learning checkpoints in a successful example. The **red** colour determines the difference between the current text and with original text, and the **green** colour determines similarity.

Input Text	Type
Reggie was able to find out for them the place at which Fleming was putting up in Boston. Having ascertained from the clerk that he was still staying there, the next thing was to get hold of Louis Anderson.	original text
able Reggie was to find out the for them place at which putting Fleming was Boston in up ascert. Havingained the from that clerk he still was staying the there, next was thing hold to get Louis of Anderson .	LOCALSHUFFLE
"Steve had to find out what they were going to put him in the hotel, the next thing was that was that he was going to put on to Tom. Having he was staying in the Boston car, which was the next thing	Checkpoint-200
Reggie was able to find out for the them place at which Fleming was putting up in Boston. Having ascertained from the clerk that he was still staying there, the next thing was to get on Louis Anderson.	Checkpoint-600
Reggie was able to find out for them the place at which Fleming was putting up in Boston. Having ascertained from the clerk that he was still staying there, the next thing was to get to Louis Anderson of.	Checkpoint-1000
Reggie was able to find out for them the place at which Fleming was putting up in Boston. Having ascertained from the clerk that he was still staying there, the next thing was to get hold of Louis Anderson.	Checkpoint-1400

Table 4: Learning process for translation PARTIALREVERSE at different learning checkpoints in a successful example. The **red** colour determines the difference between the current text and with original text, and the **green** colour determines similarity.

Input Text	Type
Bunny and Sue, so impatient they could hardly keep still, waited. They heard the front door open. They heard their father talking. Then came a funny, squeaking, whining sound.	original text
Bunny and Sue , so impatient R **. sound whining , aking sque , funny a came Then . talking father their heard They . open door front the heard They . waited , still keep hardly could they**	PARTIALREVERSE
Bessie and Sue, so impatient and impatiently. The children were not allowed to keep them shut up. They were not allowed to keep them shut. They were not allowed to keep them shut, and they were not allowed to	Checkpoint-200
Bucky and Sue, so impatient they could hardly keep still, waited, waited, waited, waited. They heard their father talking. Then Then came a funny, squeaky voice.	Checkpoint-600
Bumper and Sue, so impatient they could hardly keep still, waited. They heard the front door open. They heard their father talking. Then came a funny, squeaking, whining sound.	Checkpoint-1000
Bunny and Sue, so impatient they could hardly keep still, waited. They heard the front door open. They heard their father talking. Then came a funny, squeaking, whining sound.	Checkpoint-1400

Table 5: Learning process for translation PARTIALREVERSE at different learning checkpoints in a successful example. The **red** colour determines the difference between the current text and with original text, and the **green** colour determines similarity.

Input Text	**Type**
"That's what I say!" came from Mart. "This is no night for a little boy to be out. It's snowing harder than ever."	original text
"That be what I say!" came [S] from Mart. "This [P] be no night for a [S] little boy to be out. It be snowing harder than ever [S] ."	WORDHOP
"That's what I say!" came [S] from Mart. "This what I say!" came [S] from Mart. "This## no night for a [S] little boy to be out. It's snowing**	Checkpoint-200
"That's what I say!" came [S] 'from Mart. "This is no night for a [S] little boy to be out. It's snowing harder than ever"	Checkpoint-600
"That's what I say!" came from Mart. "This is no night for a little boy to be out. It's snowing harder than ever."	Checkpoint-1000