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

Anonymous ACL submission

Abstract

Understanding whether large language models operate under constraints comparable to human linguistic cognition remains a central question in AI and cognitive science. While there has been some research into whether LLMs can learn linguistically possible and impossible languages, it is less clear whether they can systematically recover linguistic structure and meaning from systematically degraded input. In this work, we investigate whether LLMs can translate and recover impossible languages back into possible forms and whether there is any difference between different types of impossible languages for recovery. By fine-tuning GPT-2 on several perturbation types, we find that models can reconstruct grammatically well-formed output, with performance systematically modulated by the nature of the perturbation. Models trained on longer sentences benefit from richer training contexts, although longer sequences also increase the difficulty of resolving non-local dependencies. Overall, our findings indicate that LLMs display a preference for local over distant dependencies, yet can still overcome structural violations that render input unintelligible, revealing a partial alignment between neural architectural constraints and human linguistic biases.

`https://anonymous.4open.science/r/impossible_translation-9C4D`

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).

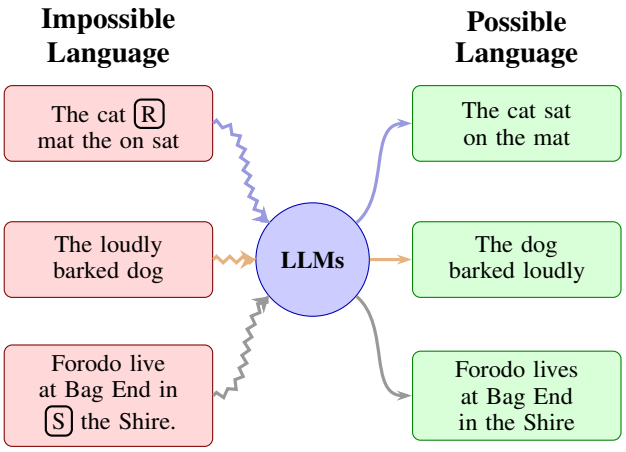


Figure 1: Translation task overview. Three separate LLMs are fine-tuned on specific perturbation types (LocalShuffle, PartialReverse, WordHop). Each model translates its corresponding impossible language input (left) into possible language output (right), recovering disrupted natural structure.

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

While some 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.

A fundamental principle underlying human language is *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, 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 natural languages systematically organise word order to keep grammatically related elements close together. (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).

This study examines whether LLMs can translate linguistically impossible languages into possible forms by recovering information rendered inaccessible through structural violations. We applied perturbation functions to the BabyLM dataset that systematically altered sentence structures, introducing linguistic inconsistencies absent in natural languages (Warstadt et al., 2023; Sinha et al., 2021). We then fine-tuned three separate GPT-2 models on these transformed datasets 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).

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, enabling incremental comprehension (Gibson et al., 2000; Liu, 2008). Since intelligible natural language requires semantically related words to appear close together, translating an impossible language requires models to identify scattered semantic dependencies and restore characteristic proximity. This raises a critical question: can LLMs, not subject to the same working memory limitations as humans, recover information made inaccessible by locality violations?

Investigating whether neural language models can translate these structurally degraded inputs back into forms respecting information locality provides insight into LLM architectural biases and their relationship to cognitive constraints defining possible human languages (Kallini et al., 2024; Xu et al., 2025; Tran et al., 2018; Khandelwal et al., 2018). Beyond theoretical implications for understanding neural architectures and human linguistic constraints, this information recovery capability has practical applications in processing corrupted text, restoring degraded documents, and handling noisy linguistic input. This study provides insights into limits and capabilities of LLMs in language learning, demonstrating their potential to interpret and transform linguistically impossible inputs into coherent outputs, contributing to understanding machine learning applications in linguistics, cognitive science, and artificial intelligence[1].

2 Background and Previous Work

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 structure) differs from cognitive impossibility (whether humans can learn a linguistic system) (Moro, 2016; Musso et al., 2003). Impossible languages are defined by unlearnable grammars which humans cannot acquire despite exposure due to cognitive constaints, rather than failure to fit formal grammatical models. Empirical evidence suggests languages relying solely on position-based rules without hierarchical structure are impossible, conflicting with core syntactic processing mechanisms and fundamentally concerning learnability rather than formal expressiveness (Musso et al., 2003; Smith and Tsimpli, 1993).

Nefdt (2024) argues identifying possible grammar boundaries requires grounding in theoretical

[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).

linguistic competence, not solely distributional regularities. Smith and Tsimpli (1993)'s work on the polyglot savant Christopher showed successful acquisition of an artificial language consistent with universal principles but failure to learn one violating them, suggesting certain grammars remain unlearnable even for exceptionally linguistically talented individuals. Neurolinguistic studies show brain regions for syntactic processing selectively respond to natural language-compatible structures, supporting that impossible languages lack neurobiological realisation (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 human language boundaries.

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 language. 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

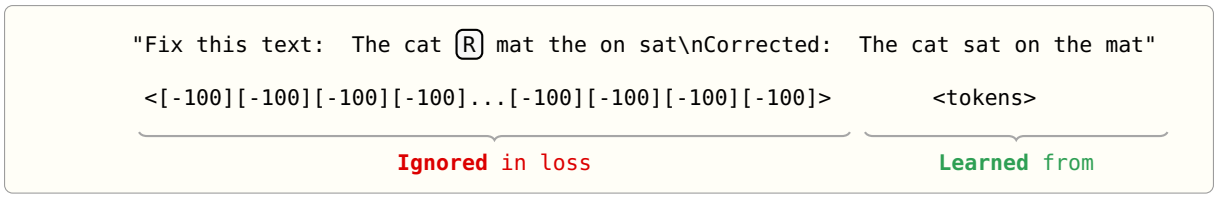


Figure 2: The mechanism of masking labels and calculating loss

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. Therefore, $\mathcal{M}$ must be pre-trained on an impossible language and then fine-tuned for the translation task. 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

The BabyLM (Warstadt et al., 2023) dataset was used for these experiments, specifically selecting the *bnc_spoken* and *Gutenberg* subsets to analyse the impact of contrasting sentence lengths on translation performance. Applying perturbations to these subsets yields paired examples (perturbed, original) for training.

Following Kallini et al. (2024), we apply three perturbation functions that systematically violate natural characteristics of language.

LocalShuffle: This perturbation randomly reorders words within a local window, 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 reversed, and the 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.

Table 1 presents examples of each function.

We created three datasets, each containing paired texts (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 created datasets of $10K$ and $100K$ sample pairs from *bnc_spoken*.

3.2 Model Fine-tuning

The models were fine-tuned by combining the causal language modelling (CLM) 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 focuses on producing correct output rather than memorising instructions, masked labelling is applied. The instruction portion (from sequence start to the word `Corrected`) receives label value $[-100]$, while the response portion (correct text and `<|endoftext|>` token) retains its tokens as labels. Only these tokens contribute to the cross-entropy loss, enabling the model to learn accurate corrections conditioned on the instruction. Figure 2 displays the label masking and loss calculation.

3.3 Experimental Design

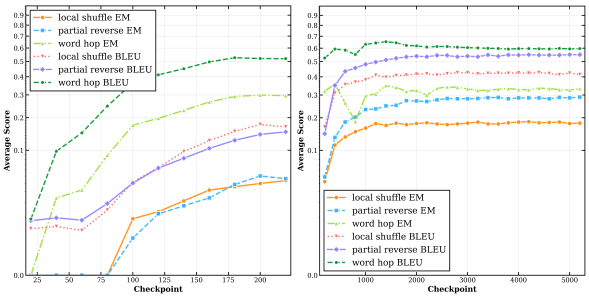
We conduct three experiments testing whether LLMs exhibit locality constraints when translating impossible languages. We hypothesise that translation performance degrades as perturbations intensify locality violations. Our perturbations systematically violate this by scattering related words across sentences, 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.

In all experiments, models were trained on `NVIDIA H100 96GB` GPUs. We train with a batch size of 512. Training the 100K dataset requires approximately 1 GPU hour per model.

Experiment 1: Effect of Training Sample Size. This experiment investigates how training data size influences translation performance. Models were fine-tuned on $10K$ and $100K$ subsets of *bnc_spoken*, with learning progress evaluated at multiple checkpoints using Exact Match (EM), measuring perfect reconstruction (Rajpurkar et al., 2016), and BLEU score quantifying n-gram overlap (Papineni et al., 2002). These metrics suit the task because impossible and possible texts share vocabulary but differ in token order, making both perfect reconstruction (EM) and partial accuracy (BLEU) informative. The hypothesis is that larger datasets will improve performance across all perturbations, 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.

Experiment 2: Effect of Sentence Length. Building on the locality principle introduced above, this experiment directly tests how sentence length affects the difficulty of recovering grammatical structure. Results were compared between the *bnc_spoken* dataset (shorter sentences, avg. 12 tokens) and the *Gutenberg* dataset (longer sentences, avg. 40 tokens). 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.



(a) $10K$ training samples (b) 100K training samples

Figure 3: Training progress on *bnc_spoken* with (a) $10K$ and (b) $100K$ samples. WORDHOP learns fastest, reaching ∼60% BLEU and ∼30% EM on $100K$ data, while LOCALSHUFFLE and PARTIALREVERSE converge more slowly at lower performance levels. For larger plots, see Figure 6.

Experiment 3: Quality of Generated Text. This experiment evaluates the extent to which generated texts are similar to human-written text using perplexity, computed under pretrained GPT-2 to measure how well generated text aligns with natural language distributions (Jelinek et al., 2005). A lower perplexity value indicates that a text is more consistent 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 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 potential artifacts from marker removal that do not fully restore natural linguistic flow.

4 Experimental results

Experiment 1: Figure 3 shows that training size critically affects performance, with $100K$ samples substantially outperforming $10K$ across all perturbations. However, improvement rates vary: WORDHOP's high performance (0.6 EM) reflects the tractability of marker removal when word order is preserved, while LOCALSHUFFLE's low ceiling (0.18 EM) even with $100K$ samples indicates that additional data cannot fully overcome the difficulty of recovering severely disrupted locality, suggesting fundamental architectural limitations in reconstructing scrambled dependencies. Qualitative examples illustrating the learning progression at different checkpoints are provided in Appendix

Dataset		Size	Average Length
bnc_spoken	train	$10K$	12.54
bnc_spoken	train	$100K$	12.77
bnc_spoken	test	$1K$	11.72
Gutenberg	train	$10K$	40.56
Gutenberg	train	$100K$	40.61
Gutenberg	test	$1K$	46.78

Table 2: Average sentence lengths for the training and test datasets at $10K$ and $100K$ sample sizes for training and $1K$ for testing.

(Tables 3, 4, and 5).

Experiment 2: Table 2 shows average sentence lengths for both datasets. Comparing final performance at 100K samples (Figures 4 and 3) reveals substantial sentence length effects that vary by perturbation type. On *Gutenberg* (avg. 40 tokens), WORDHOP achieves approximately 78% BLEU compared to 60% on *bnc_spoken* (avg. 12 tokens), representing an 18 percentage point improvement. PARTIALREVERSE shows even larger gains, improving from 55% BLEU on short sentences to 82% on long sentences (27 percentage points). In contrast, LOCALSHUFFLE demonstrates minimal improvement, rising from 18% EM to only 25% EM (7 percentage points), remaining challenging regardless of sentence length. Analysis of performance improvement by input text length class is shown in Appendix Figure 9.

This differential response indicates that WORDHOP's preserved word order and PARTIALREVERSE's partial structure enable better exploitation of additional training content in longer sentences, whereas LOCALSHUFFLE's complete locality disruption presents fundamental recovery difficulties that additional context cannot overcome. Figure 4 further confirms that while model performance improves on longer texts, learning progress is notably slower than on shorter sentences.

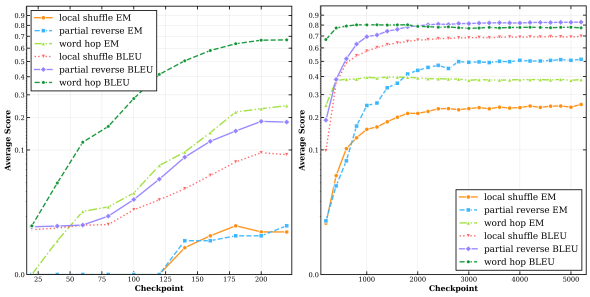
For perturbations that directly disrupt intra-sentential dependencies, sentence length becomes a crucial factor: as sentences grow longer, dependency distances increase and the perturbation affects a larger portion of the sequence, thereby enlarging the effective search space of possible reorderings. This is consistent with prior work showing that neural language models rely more heavily on nearby context and struggle with long-range dependencies and hierarchical structure, especially when architectural capacity does not scale with input length (Khandelwal et al., 2018; Hahn, 2020; Tran et al., 2018). It also aligns with findings that models trained on shuffled or perturbed word orders retain only partial and noisy order information, with performance degrading as structures become more globally disrupted (Sinha et al., 2021; Abdou et al., 2022).

Experiment 3: Although our model is fine-tuned from GPT-2, 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 impossible language inputs with model predictions against baseline unperturbed text (perplexity = 129.34). Results reveal striking differences across perturbation types. LOCALSHUFFLE input exhibits extremely high perplexity (3902.34), which model predictions reduce significantly to 212.40, representing a 94.6% decrease. This substantial reduction demonstrates successful reconstruction approaching natural language distributions, though predictions remain above baseline, indicating incomplete restoration. PARTIALREVERSE shows similar patterns with less extreme values. Input perplexity (1500.38) reduces to 146.56, a 90.2% reduction. Predictions come closer to baseline than LOCALSHUFFLE (146.56 vs. 212.40), suggesting more successful structural recovery, aligning with higher exact match scores (0.29 vs. 0.18). WORDHOP presents a contrasting pattern. Input perplexity (168.11) is already near baseline, as this perturbation preserves underlying word order. However, predictions show increased perplexity (243.81), a 45.0% increase. While the model successfully removes inserted markers, the reconstruction introduces subtle artefacts deviating from standard distributions, likely from imperfect verb agreement handling. Concrete examples of text quality at different training stages are provided in Appendix Tables 3, 4, and 5.

5 Discussion

Recent research demonstrates that LLMs exhibit disparities in their capacity to acquire impossible languages, attributed to underlying inductive biases (Kallini et al., 2024; Xu et al., 2025). Building on this foundation, we adopt a reverse approach, examining whether large language models can recover accessible linguistic structure from systemat-



(a) (b)

Figure 4: Training progress on *Gutenberg* with (a) $10K$ and (b) $100K$ samples. Longer sentences yield higher performance than *bnc_spoken*. For larger plots, see Figure 7.

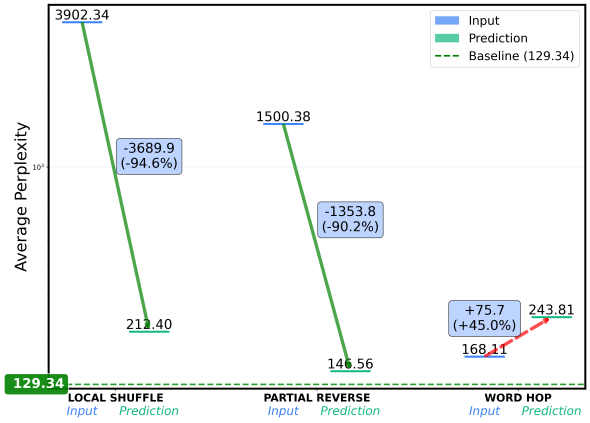


Figure 5: Perplexity of impossible inputs (blue) versus model predictions (green) relative to baseline unperturbed text (129.34, dashed line), evaluated using pretrained GPT-2. For larger plots, see Figure 8.

ically degraded input. Specifically, we investigate whether models are influenced by universal principles constraining human linguistic feasibility and whether their ability to recover information is affected by the locality of the perturbation. We argue that this serves as a test of the Information Locality hypothesis, which holds that natural languages tend to group related elements closer together. 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 across sentences and create non-local dependencies, proved most difficult (0.18 and 0.29 EM respectively), requiring models to reconstruct hierarchical structures from scrambled input. Conversely, WORDHOP, which preserves underlying word order and maintains most local dependencies, achieved substantially higher performance (>0.6 EM) despite introducing distance-based markers. Crucially, perplexity analysis reveals that models genuinely recover underlying structure rather than memorizing mappings. The dramatic perplexity reductions for LOCALSHUFFLE (94.6%) and PARTIALREVERSE (90.2%) toward the baseline indicate successful reconstruction of locality-respecting configurations, where scattered elements are reordered into interpretable structures restoring local dependencies (Hahn et al., 2020; Hahn, 2020). This convergence demonstrates that translation effectively recovers information rendered inaccessible by locality violations, though WORDHOP's increased perplexity suggests artifacts from imperfect marker removal.

Performance across perturbation functions Our results demonstrate GPT-2 can perform information recovery, with performance varying systematically by perturbation type. WORDHOP proved easiest, achieving exact match scores above 0.6 on $100K$ samples, as it preserves underlying word order despite introducing distance-based markers. In contrast, LOCALSHUFFLE and PARTIALREVERSE plateaued around 0.18 and 0.29 EM respectively, requiring reconstruction of hierarchical syntactic structures from scrambled input where local dependencies become non-local. This difficulty aligns with theoretical arguments about transformer architectural limitations in capturing unbounded hierarchical dependencies (Tran et al., 2018; Delétang et al., 2023; Hahn, 2020; Khandelwal et al., 2018). Our work examines the reverse task: whether models pre-trained on impossible languages can recover possible language structure through translation. The pattern of difficulty we observe in translation—WORDHOP easiest, LOCALSHUFFLE hardest—mirrors the pattern Kallini et al. (2024) found for learning these languages, suggesting the same architectural biases that make certain impossible languages harder to acquire also make them harder to translate. Perplexity analysis confirms models genuinely recover underlying structure rather than memorizing mappings. For LOCALSHUFFLE and PARTIALREVERSE, prediction

perplexity decreased toward original unperturbed text, with fine-tuned models producing sequences increasingly consistent with pre-trained GPT-2 distributions (Radford et al., 2019; Holtzman et al., 2020), demonstrating effective recovery of information rendered inaccessible by locality violations.

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 expand 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).

6 Conclusion

This study demonstrates that LLMs can translate linguistically impossible inputs into possible forms by recovering information obscured through information locality violations. Performance varies systematically (WORDHOP easiest, LOCALSHUFFLE hardest), with longer sentences providing more training content while increasing recovery difficulty. While LLMs are sensitive to locality constraints, they overcome intractable violations.

Our findings reveal incomplete alignment between LLMs and human linguistic competence. Models show sensitivity to information locality yet operate with distinct inductive biases, indicating that putative constraints on human learnability do not necessarily impose bounds on learnability by neural networks. The same architectural properties constrain both acquisition and translation, revealing biases partially aligned with cognitive constraints.

This raises intriguing questions about child language acquisition: Could similar approaches help explain how children extract linguistic structure from limited and imperfect input (Chomsky, 1986; Pinker, 1994)? More speculatively, these methods might help analyse extraterrestrial communication systems violating human universals, with models distinguishing incomprehensible from unfamiliar aspects. Systematic investigation of impossible languages through computational models reveals boundaries of linguistic possibility beyond human cognition.

This study's approach could be applied to some real-world problems, such as recovering disrupted information locality, enabling error correction, text normalisation, and robustness to noisy input. Future work should investigate scaling to larger architectures, generalisation to novel perturbations, and cross-linguistic extension, contributing to understanding the computational basis of language and boundaries defining linguistic possibility.

7 Limitations and Future Directions

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, we evaluate performance primarily through exact match, BLEU scores, and perplexity, which capture surface-level accuracy but may not fully reflect whether the model has genuinely learned the underlying principles of information locality restoration or merely exploited distributional shortcuts.

Third, 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.

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. Lexinvariant 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.

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, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI technical report.

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*.

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

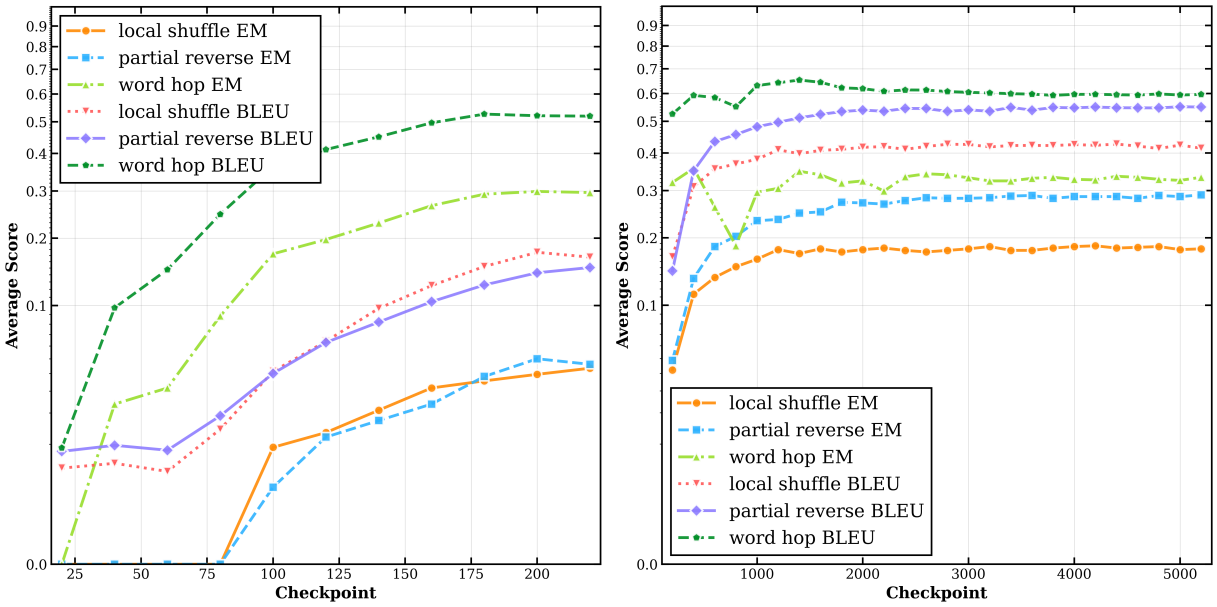
A Larger Plots

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

B 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.

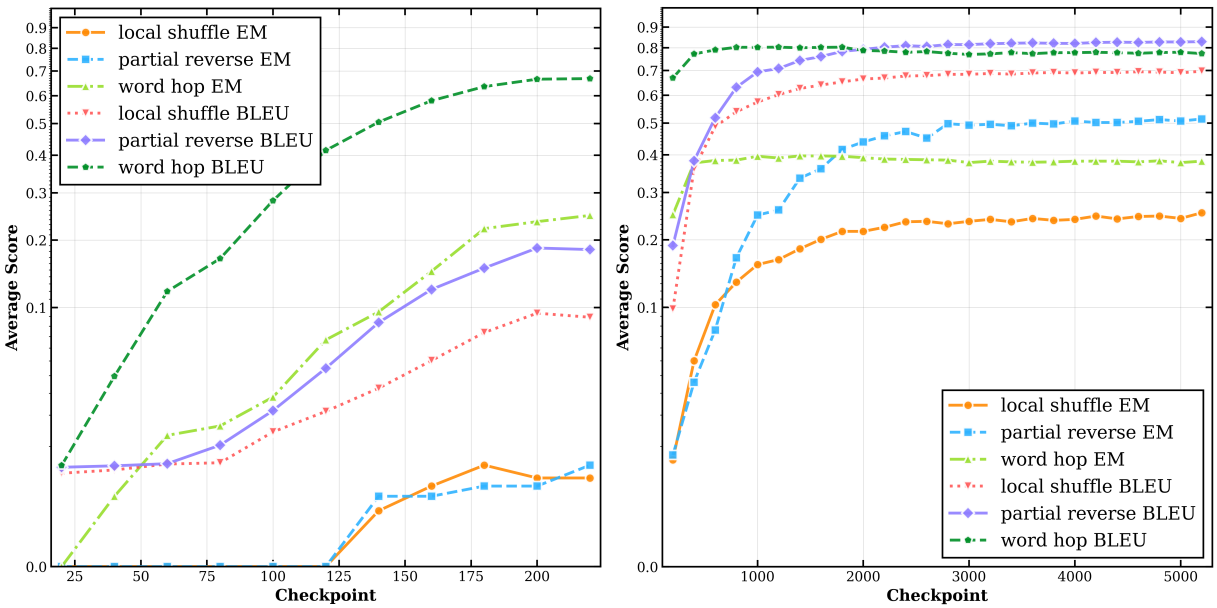
Additionally, we investigate how performance varies by input text length for the LOCALSHUFFLE perturbation. Since LOCALSHUFFLE uses a window size of 3, we classify input texts by the number of shuffling windows they contain. Specifically, we define $\text{class}(text_i) = \lceil \text{len}(text_i)/3 \rceil$, 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) $10K$ training samples

(b) 100K training samples

Figure 6: Training progress on *bnc_spoken* with (a) $10K$ and (b) $100K$ samples. WORDHOP learns fastest, reaching ∼60% BLEU and ∼30% EM on $100K$ data, while LOCALSHUFFLE and PARTIALREVERSE converge more slowly at lower performance levels.



(a)

(b)

Figure 7: Training progress on *Gutenberg* with (a) $10K$ and (b) $100K$ samples. Longer sentences yield higher performance than *bnc_spoken*

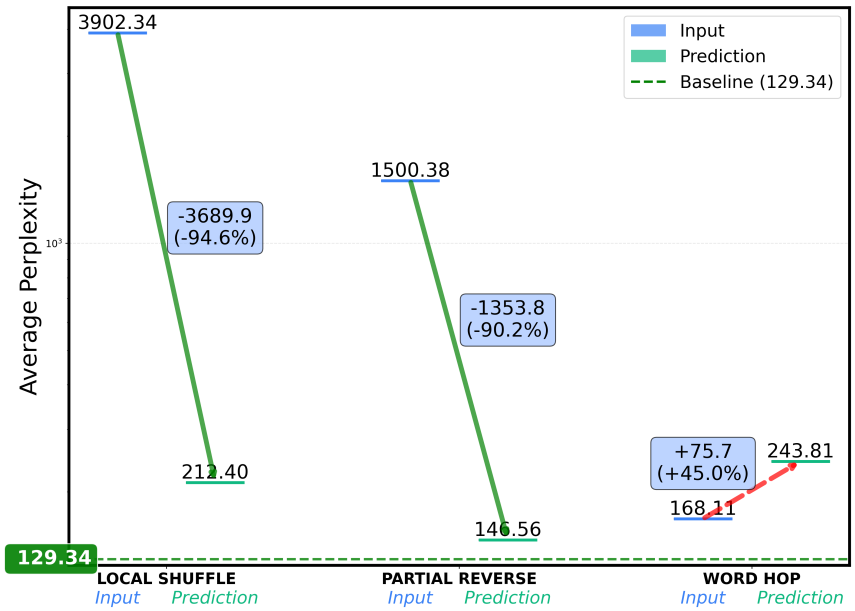


Figure 8: Perplexity of impossible inputs (blue) versus model predictions (green) relative to baseline unperturbed text (129.34, dashed line), evaluated using pretrained GPT-2.

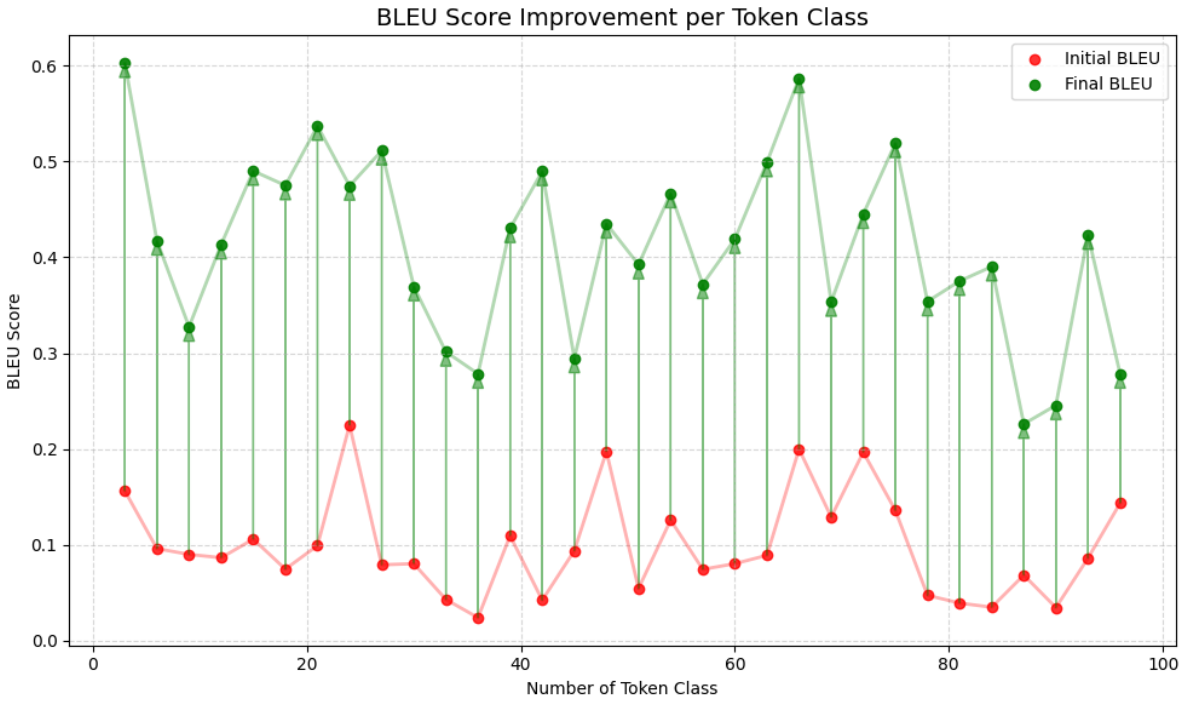


Figure 9: Improvement of the BLEU score for each class. The red line corresponds to the BLEU score before translating and for shuffled text, and the green line is for texts after translation by the final model.

Table 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