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

Amirhossein Mohammadi, Laurence Frank, Albert Gatt, Ayoub Bagheri, *Utrecht University.*

Utrecht University

What Are "Impossible Languages"?

Definition

Impossile Languages are hypothetical or fictional languages that cannot exist or be fully used by humans because they violate fundamental limits of human cognition, perception, or linguistic structure.

- Possible languages: Systems humans can learn and use
- Impossible languages: Violate universal grammatical principles
- Not about formal adequacy, but cognitive learnability

Information Locality Principle

Definition

Information locality is the principle that information that is used together or accessed close together in time or space is stored or organized close together, making processing or retrieval more efficient.

- Related elements appear close together in linear order
- Dependency Locality Theory (DLT): Processing difficulty increases with distance
- Natural languages minimize dependency lengths
- Violations make text incomprehensible despite intact lexical content

Dependency Length

The book that my sister bought yesterday is on the table.

The cat that the dog that the mouse chased bit ran away.

The student who the teacher who the principal hired recommended failed.

My sister bought a book yesterday. It is on the table.

The mouse chased the dog. The dog bit the cat. The cat ran away.

The principal hired a teacher. That teacher recommended a student. The student failed.

LLMs vs Impossible Languages

Chomsky in "Conversation with Tyler" Podcast

LLMs ultimately reveal nothing about human language and thought because they cannot distinguish between possible and "impossible" languages

- Models process both systems identically without recognizing the distinction, they fail to provide insight into the specific nature of human language

LLMs vs Impossible Languages

Kallini et al. (2024)

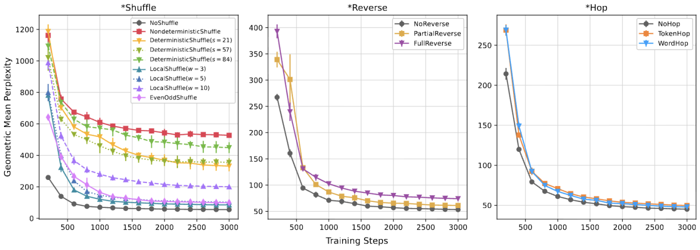


Figure: Perplexities for each impossible language model over training steps.

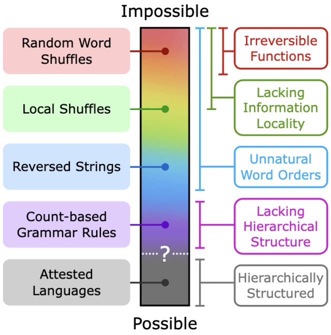


Figure: Partial impossibility continuum of languages based on complexity.

Method

Methodology

$$f : A \to A';$$
$$\mathcal{M}(A') := A$$

where f is the perturbation function, A is a possible language, A' is the corresponding impossible language with A and $\mathcal{M}$ is the LLM trained to translate A' to A.

- $\mathcal{M}$ must not be trained in Language A

How to Create Impossible Datasets?

Perturbation functions!

- **LocalShuffle:** 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 by adding markers (S for singular, P for plural) at fixed distances after verbs based on subject-verb agreement.

Perturbation Functions

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

Dataset

Two Subsets Selected from BabyLM Corpus:

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

Method

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

"Fix this text: The cat $\boxed{\text{R}}$ mat the on sat\nCorrected: The cat sat on the mat"

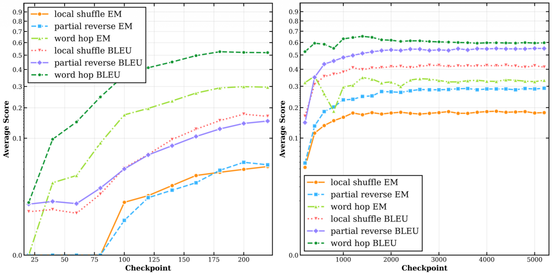
$\underbrace{\texttt{<[-100][-100][-100][-100]...[-100][-100][-100][-100]>}}_{\textcolor{red}{\texttt{Ignored in loss}}}$ $\underbrace{\texttt{<tokens>}}_{\textcolor{green}{\texttt{Learned from}}}$

Causal language model (CLM) - Instructikon Following

Experiment 1 - Effect of Training Size

Experimental Design:

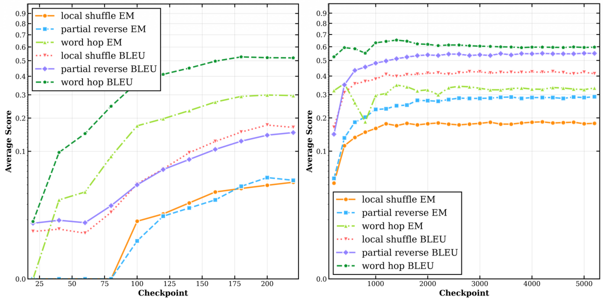
- **Training sizes:** 10K vs 100K samples (`bnc_spoken`)
- **Evaluation metrics:**
  - **Exact Match (EM):** Perfect reconstruction rate.
  - **BLEU score:** N-gram overlap measure.



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

Experiment 1 - Effect of Training Size

- Larger datasets improve performance across all perturbations
- Different learning rates by perturbation type
- More data cannot fully overcome fundamental difficulty
  - Architectural limitations, not just data limitations
  - Information locality violations create intrinsic challenges

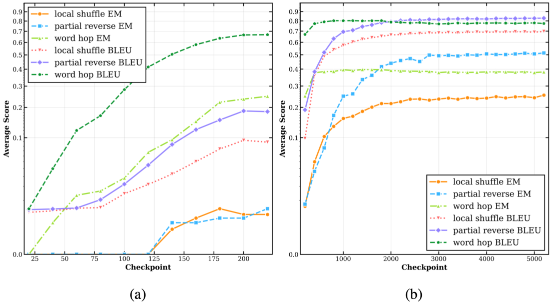


(a) $10K$ training samples (b) 100K training samples

Experiment 2 - Effect of Sentence Length

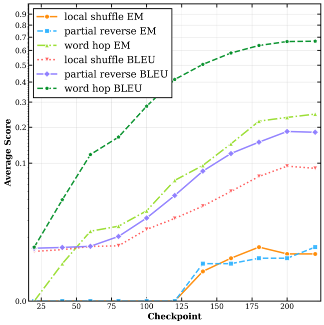
Experimental Design:

- **Compare two datasets:**
  - `bnc_spoken`: Short sentences (avg. 12 tokens)
  - `Gutenberg`: Long sentences (avg. 40 tokens)

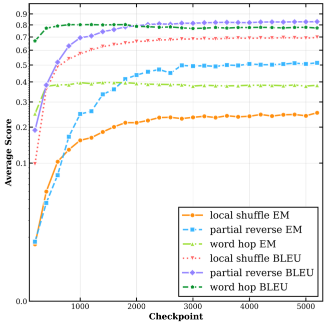


Experiment 2 - Effect of Sentence Length

- In longer sentences, perturbation functions like shuffle and reverse increase the dependency length of a text more in comparison with shorter sentences.
- Longer sentences generally improved performance because they provide a richer training signal.



(a)

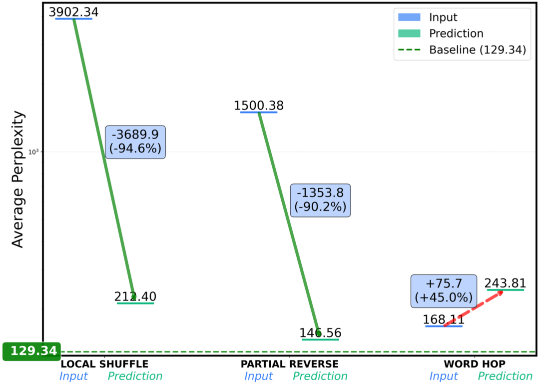


(b)

Experiment 3 - Text Quality

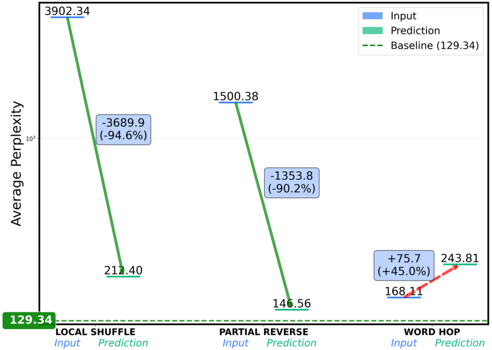
Research Question: Are models truly recovering linguistic structure or merely memorizing input-output mappings?

Experimental Design: Compare the perplexity of perturbed and translated text using normal GPT-2.



Experiment 3 - Text Quality

- For shuffling and reversal perturbations, the models successfully restored "information locality," making previously inaccessible text interpretable.
- The perplexity increase in WordHop suggests the model introduced subtle artifacts or grammatical errors (likely in verb agreement) during marker removal.



Key Finding - Information Locality Matters

- Model performance directly reflects how much a perturbation violates natural language characteristics.
- LocalShuffle and PartialReverse proved most difficult because they disperse grammatically related elements, creating non-local dependencies.
- In contrast, WordHop was easiest because it maintains the underlying word order and local dependencies.
- Perplexity analysis confirms that GPT-2 genuinely recovers underlying linguistic structure rather than just memorising mappings.

Thank you

Questions?

Utrecht University