Low-Perplexity LLM-Generated Sequences and Where To Find Them

Arthur Wuhrmann, Anastasiia Kucherenko, Andrei Kucharavy

TDA (Training data attribution) is a field trying to **map sentences to the training data**, particularly with LLMs. This is crucial to copyright checking, data cleaning and to ensure privacy.

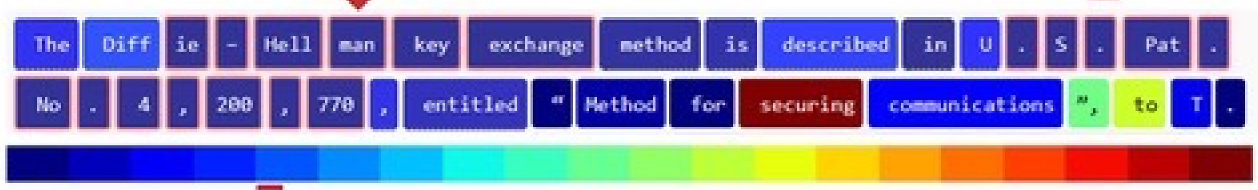
When LLMs generate sequences, each token has some probability of being generated. Sometimes, the model is extremely confident in a sequence it generates.

Token perplexities

We select **6 tokens-windows** with **generation probability > 0.9** for each token.

7688646 matches !

"S. Pat. No. "



Very confident **Not confident**

" No. 4, 200"

1 match ?!

We trace low-perplexity regions to their origins in the training data.

*Do LLMs learn passages from their training data **by heart** ?*

Generate prompts within training data

Models: Pythia scaling suite (mainly 6.8B)
Training Data: The Pile
Topics: genetics, drugs, cryptography & nuclear physics

Why these topics ?
They contain **specific keywords;** the goal was to guide the model to **specific area of the dataset.**

Make the model complete them

We feed the model 40 articles per topic, and 5 generations per prompt to cover enough of the probability space, giving a total of **800 generations**.

Find low-perplexity sequences in the output

Extract standalone perplexity

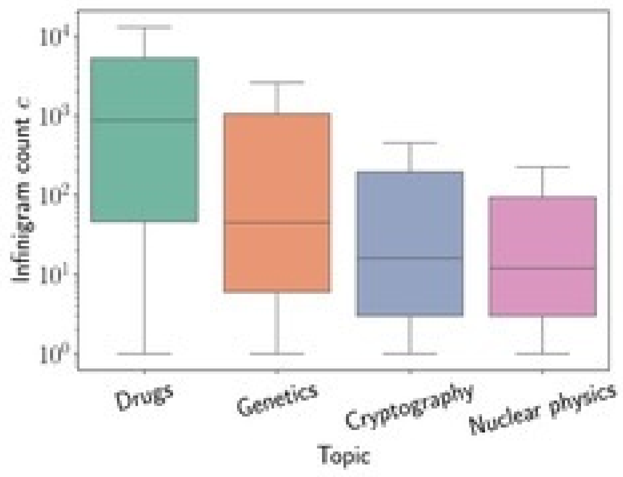
For each sentence, we know the model was likely to generate it *in this context*. Standalone perplexity is the likelihood of being generated **without context**. This gives insights about the human-ness of the extract for example.

Tracing back to corpus

To find matches of low-perplexity sequences in the training data, we use Infini-gram* — an open-source, scalable tool that locates sequence positions in the corpus within milliseconds.

Topic difference

The number of matches does change between the topics ! This can be because some topics are more present in the Pile than others.



We also observe that many of the generations are repeating the prompt, inflating the counts !

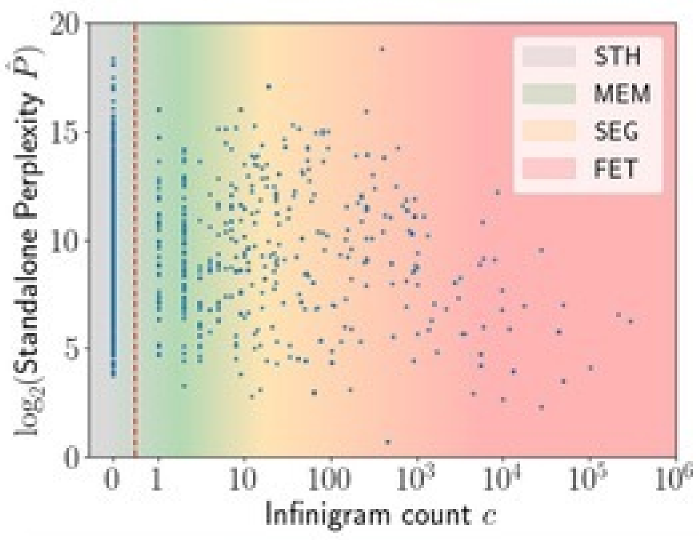
Topic	# of low-perp reg.	% of matches	% of repetition
Drugs	988	67%	8%
Genetics	1337	36%	29%
Cryptography	1336	38%	32%
Nuclear physics	1040	25%	15%

Many low-perplexity sequences **don't match !**

Nature of low-perplexity sequences

We hypothesize that the low-perplexity sequences would classify in 4 types :

- **STH** : 0 matches
  "computationally indistinguishable from"
- **MEM** : 1-5 matches
  " alcohol, sugar, water, and"
- **SEG** : 10-50 matches
  " has been defined as "the study "
- **FET** : 50+ matches
  " synthetic cannabinoid. "



Temperature effect

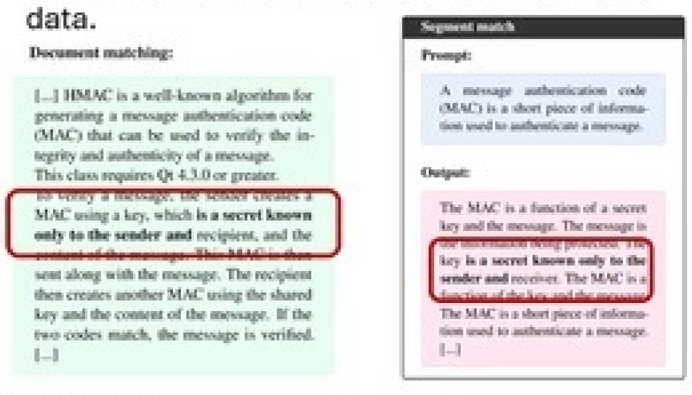
We found out that lower temperature means more matches in general.

Temperature	# of low-perp reg.	% of matches	# of repetition
0.2	8787	33%	743
0.3	6127	31%	589
0.4	4523	32%	598
0.5	3297	33%	560
0.6	1913	34%	310
0.7	1337	36%	386

Lower temperature ⟶ more **repetitions**

Context of low-perplexity sequences

We look at context around the low-perplexity sequences and compare with the context around the match in the training data.



Same context !

Main takeaways

- Perplexity can be used to identify text present in the training data, but we observe many false positive
- Low-perplexity regions from a generation can be categorize, and interpreted accordingly
- These regions are often repetitions of the prompt, potentially biasing results

*: Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2024. Infini-gram: Scaling Unbounded n-gram Language Models to a Trillion Tokens. *arXiv preprint arXiv:2401.17377.*

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

armasuisse Science and Technology
Cyber-Defence Campus