Impact of Co-occurrence on Factual Knowledge of Large Language Models

Cheongwoong Kang, Jaesik Choi

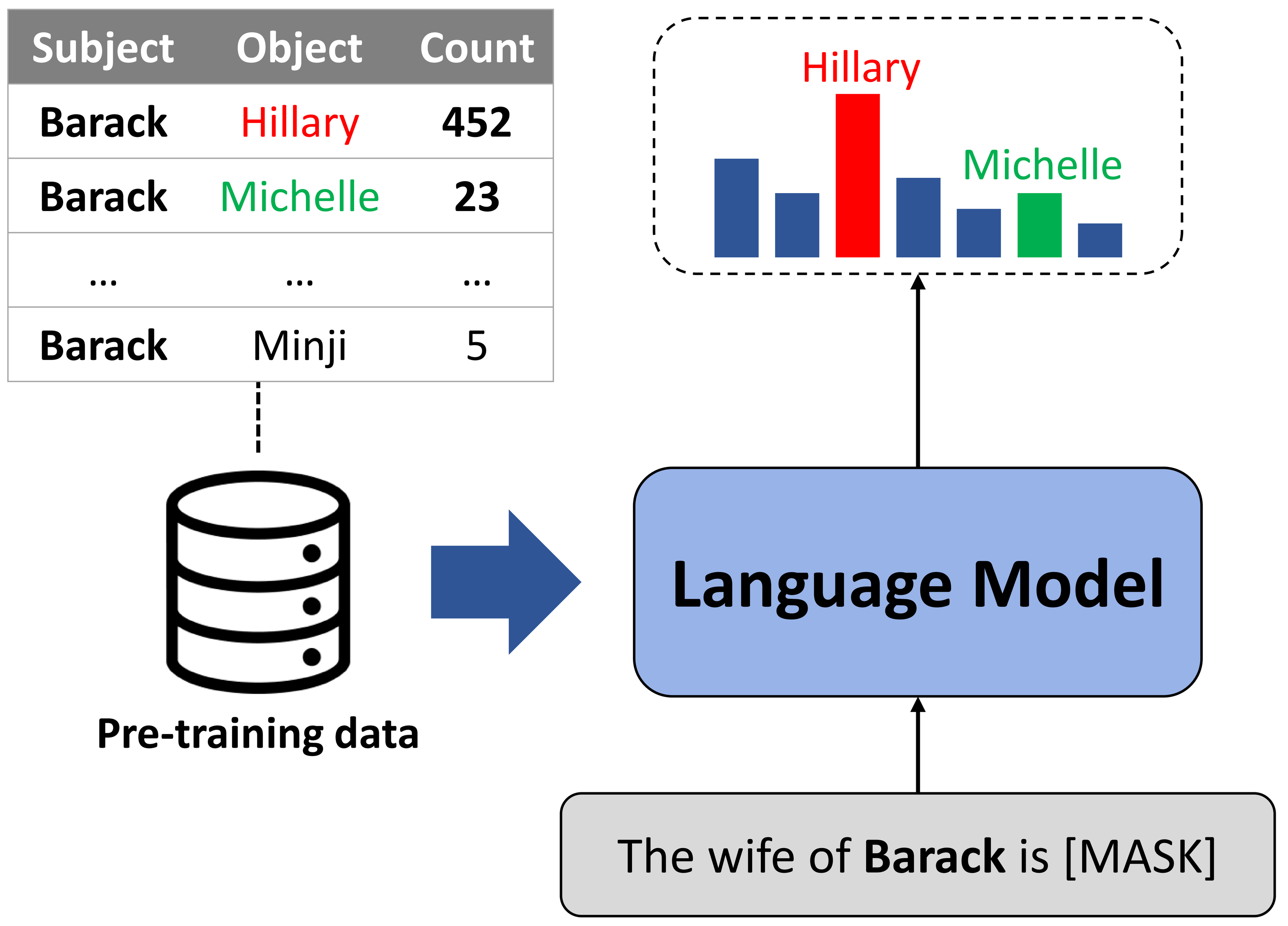
KAIST

EMNLP 2023

Project GitHub

Stochastic Parrots 🦜 or Intelligent Agents 🧑‍🎓?

Hypothesis: Large language models (LLMs) often rely on simple co-occurrence statistics without understanding the meaning behind words, causing hallucinations.

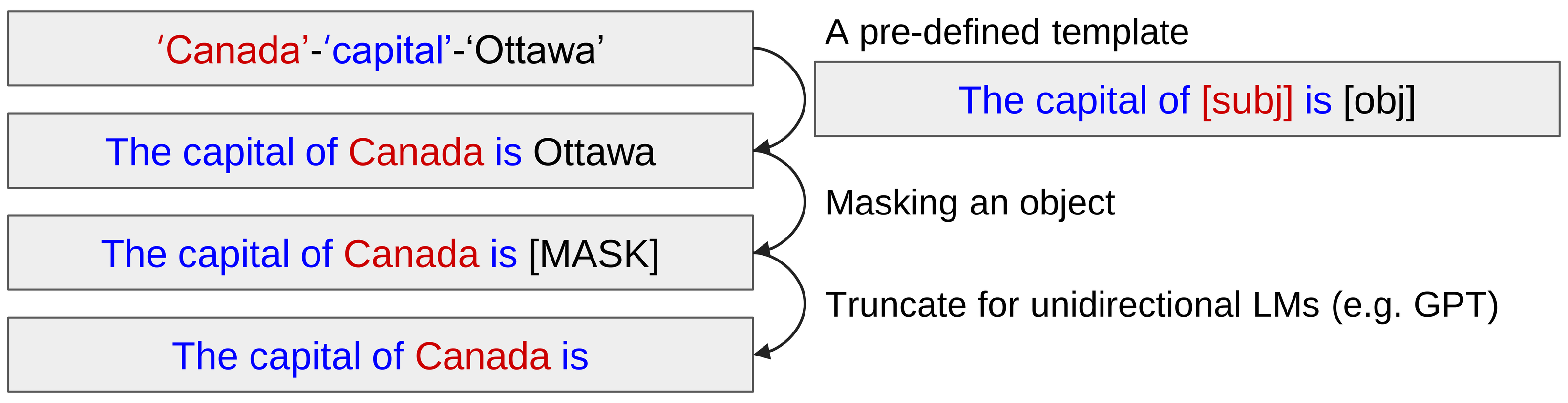


In the hypothetical example, the model **fails to answer the question about the wife of Barack Obama by generating the most frequently co-occurring word 'Hillary'**, while the correct answer is 'Michelle.'

Factual Knowledge Probing

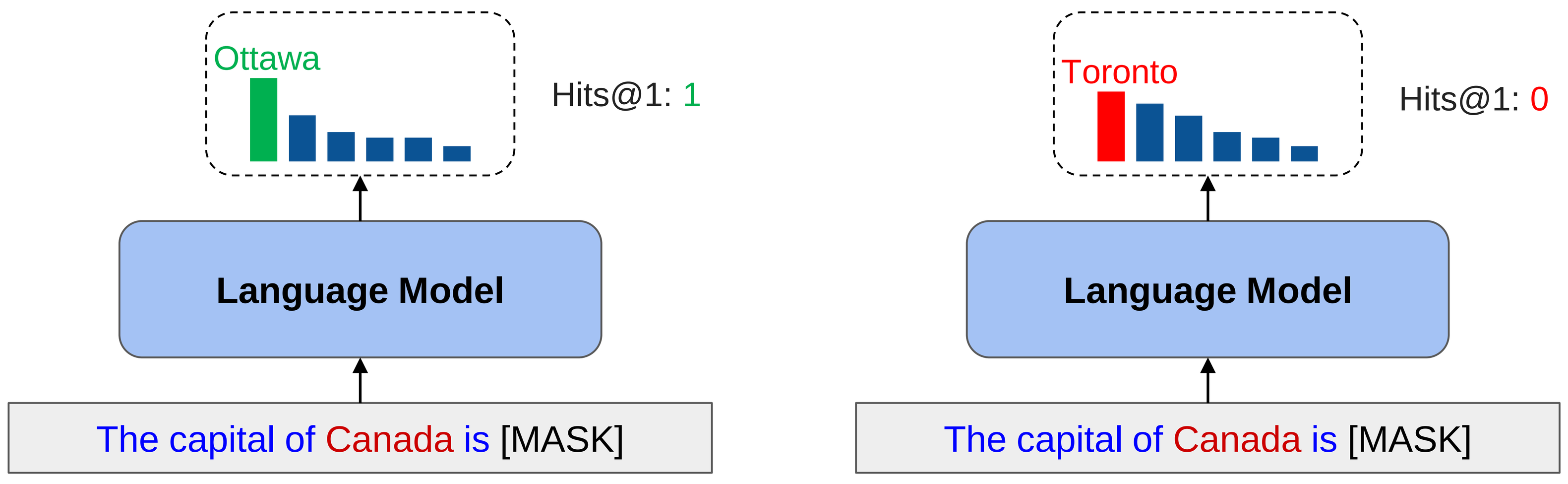
The LAMA Probe

- We adopt the LAMA-TREx dataset, which consists of 41 relations.



Metrics

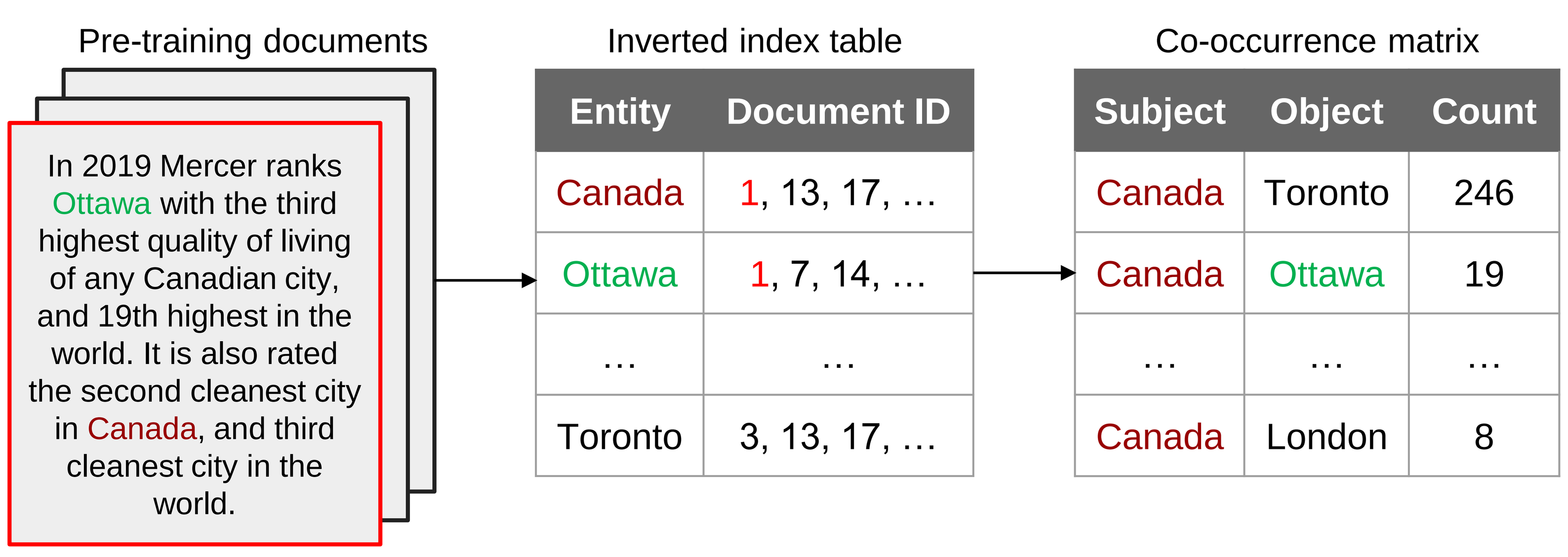
- **Hits@1:** hits@1 is 1 if the correct answer is ranked top-1, otherwise 0.



Analyzing Impact of Co-occurrence Statistics

Co-occurrence Statistics

- We consider the **subject-object co-occurrence of the pre-training dataset.**



Correlation Analysis

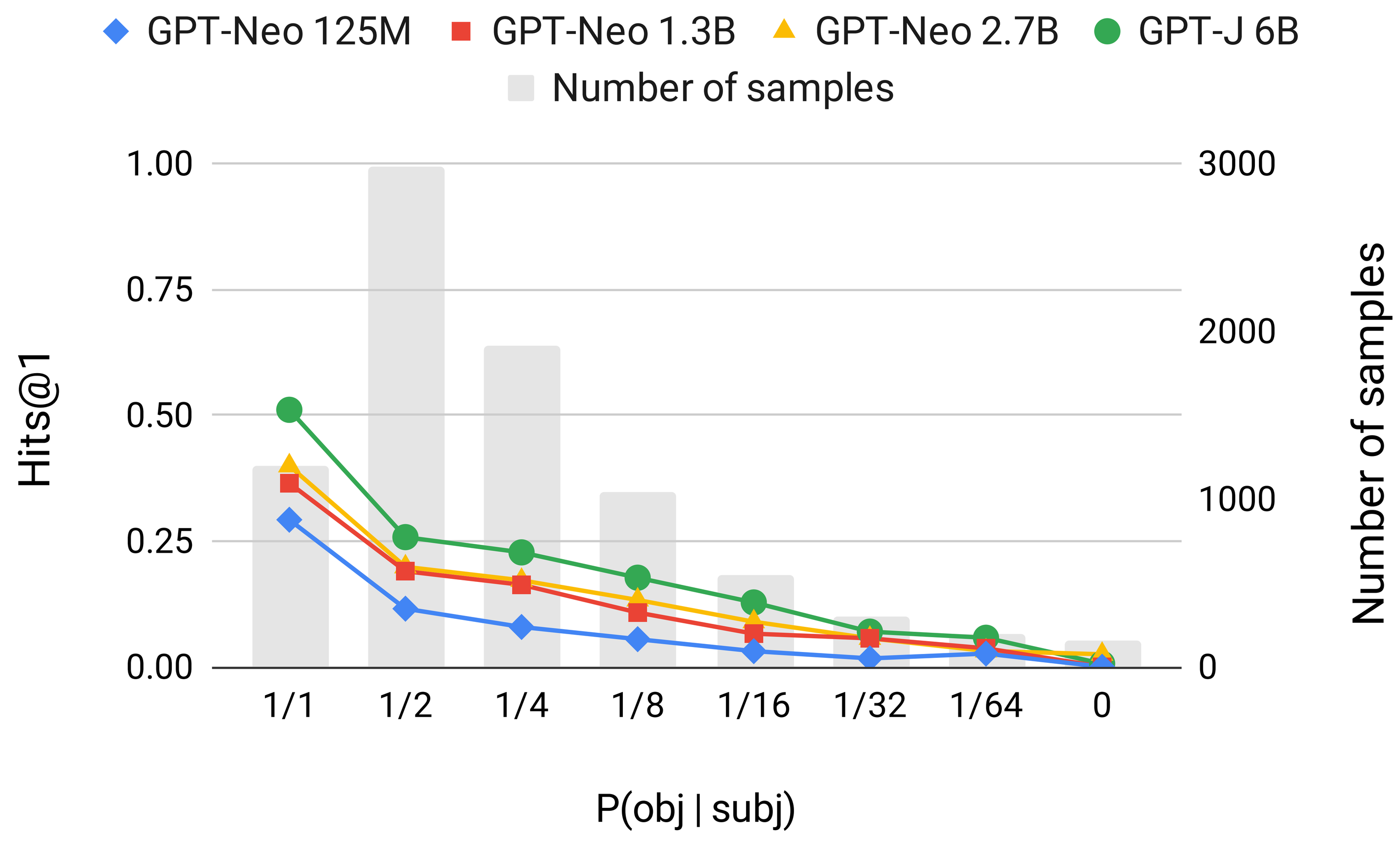
- We **plot hits@1 of the target LLMs against the conditional probability** of the gold object given a subject. Here, we divide the samples into multiple frequency (conditional probability) bins and report the average hits@1 for each bin.

Experimental Setup

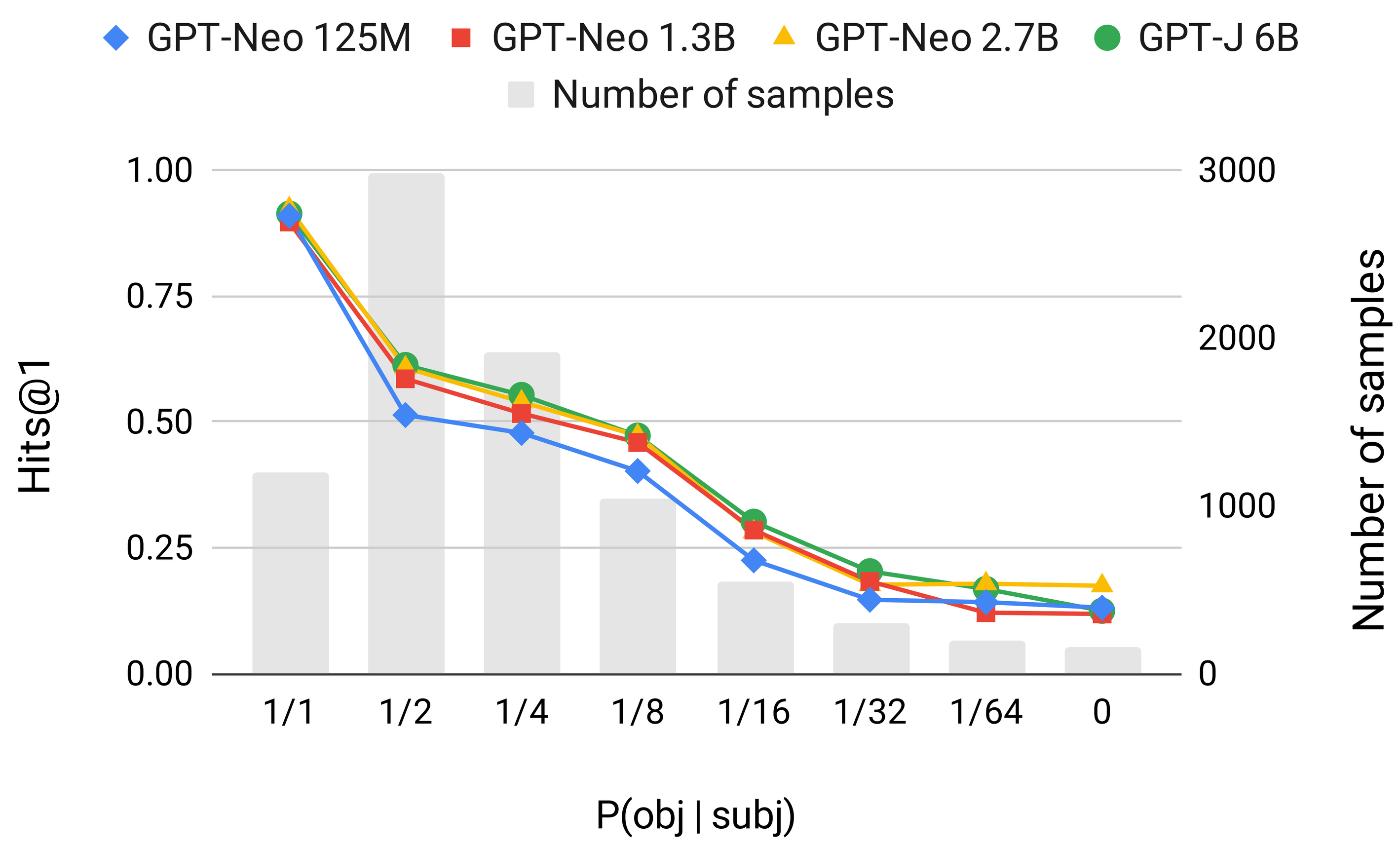
- We test open-source versions of GPT-3 with four different model sizes: **GPT-Neo 125M, GPT-Neo 1.3B, GPT-Neo 2.7B and GPT-J 6B**, which are publicly available on Huggingface's transformers.
- These models are pre-trained on **the Pile**, which is a publicly available dataset that consists of 800GB of high-quality texts from 22 different sources.

Results

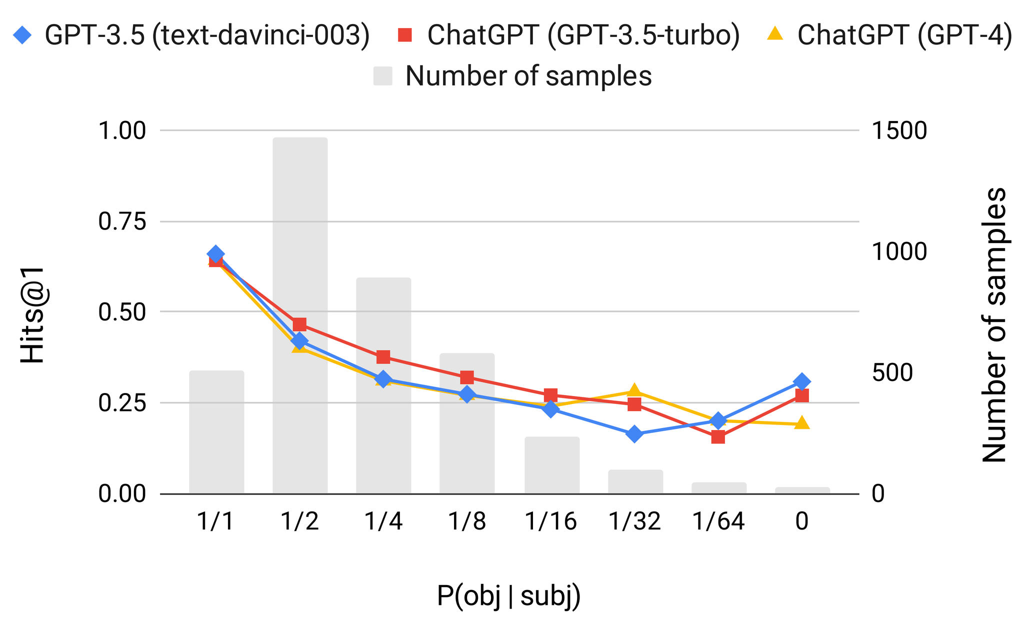
The correlation between co-occurrence and factual knowledge probing accuracy: We plot hits@1 against $P_{pretrain}(obj|subj)$ on the test set.



Zero-shot: We observe **a strong correlation between hits@1 and the co-occurrence count**. As a result, **LLMs struggle to recall rare facts**. We observe that such correlation remains **despite scaling up model sizes.**



Finetuned: We observe that the correlation remains **despite finetuning.**



Frequency bin	Ratio
1/1	0%
1/2	15%
1/4	42%
1/8	56%
1/16	70%
1/32	78%
1/64	85%
0	95%
Total	38%

Left: We test **larger models (GPT-3 175B and ChatGPT)** to verify that such correlation remains despite scaling up model sizes. Right: **The correct answer is overridden by a word with higher co-occurrence counts** in a total of **38%** of the failure cases of GPT-J 6B. The ratio is **much higher when recalling rare facts.**

Takeaways

- Our results reveal that **LLMs are vulnerable to the co-occurrence bias**, defined as **preferring frequently co-occurred words over the correct answer.**
- Consequently, **LLMs struggle to recall facts whose subject and object rarely co-occur** in the pre-training dataset.
- Co-occurrence bias remains **despite scaling up model sizes or finetuning.**
- Therefore, we **suggest further investigation on mitigating co-occurrence bias to ensure the reliability of language models** by preventing potential harms.

GitHub: https://github.com/cheongwoong/impact_of_cooccurrence

Email: cw.kang@kaist.ac.kr