Impact of Co-occurrence on Factual Knowledge of LLMs

Findings of EMNLP 2023

Cheongwoong Kang and Jaesik Choi
KAIST

Stochastic Parrots? Intelligent Agents?

Language Model

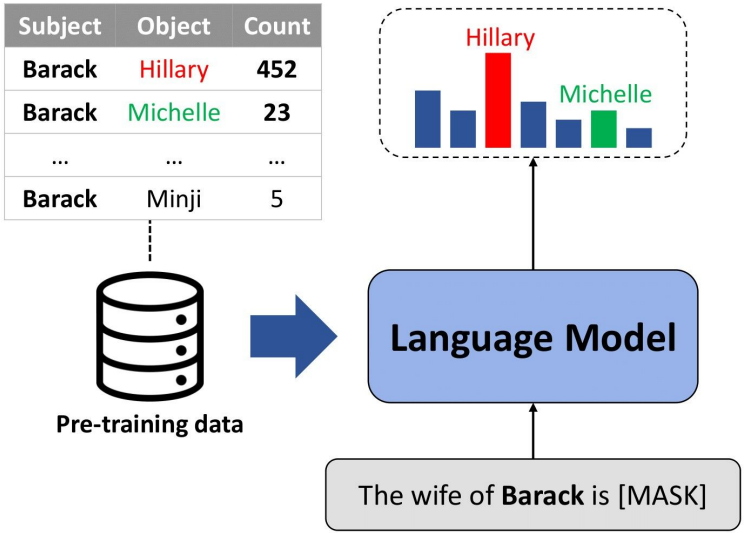
Barack Obama married Hillary

Drinking bleach can cure the common cold

New York is the capital of the U.S.

Motivation

- **Question:** Why do language models hallucinate?
- **Hypothesis:** LLMs often rely on simple co-occurrence statistics without understanding the meaning behind words, causing hallucinations.

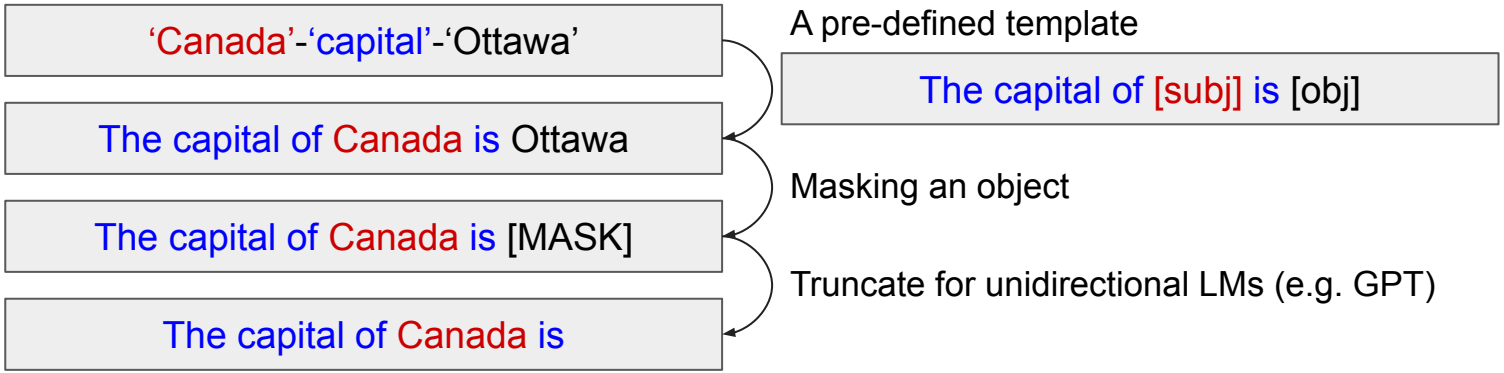


In this hypothetical example, the model **fails to answer the wife of Barack Obama by generating the most frequently co-occurring word 'Hillary'**, while the correct answer is 'Michelle.'

Factual Knowledge Probing

The LAMA Probe

- We adopt the LAMA-TREx dataset, which consists of 41 relations.
- **Facts** are represented as **subject-relation-object** triples.
- Each fact is converted to a **natural language form** based on the **templates**.
- Each sentence is converted to a **Cloze statement** by **masking an object**.

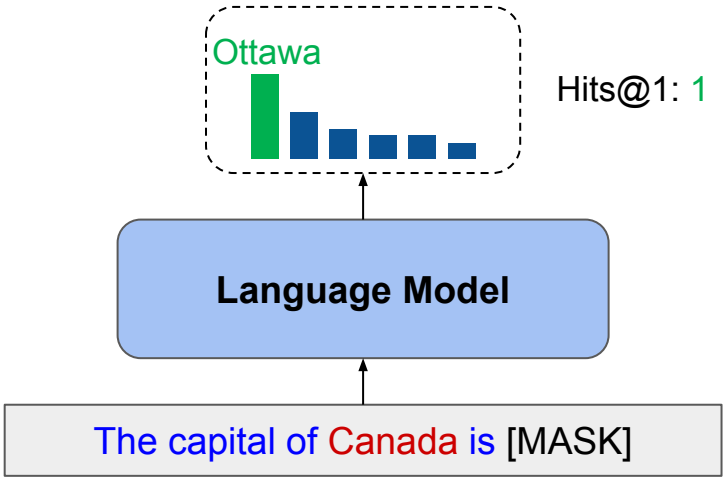


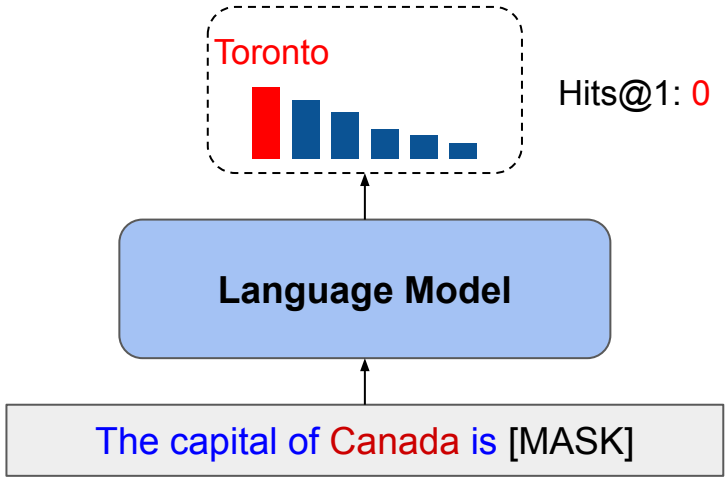
Elsahar, Hady, et al. (2018), Petroni, Fabio, et al. (2019)

Factual Knowledge Probing

Metrics

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

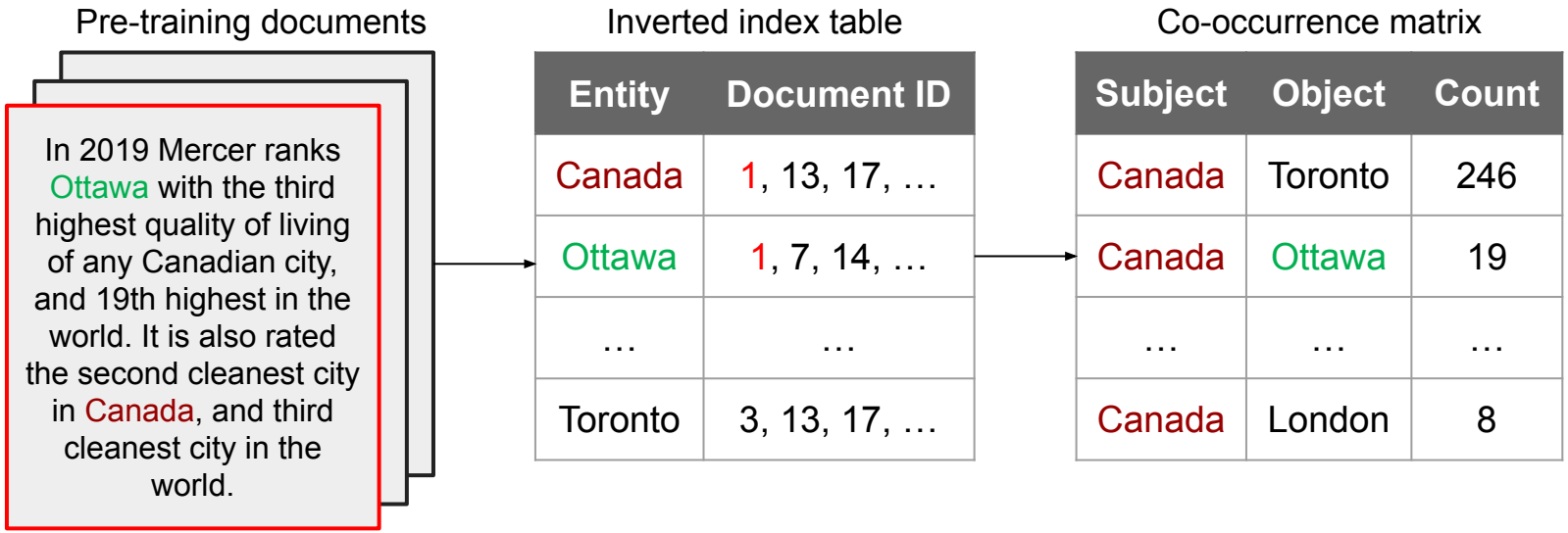




Bordes, Antoine, et al. (2011), Bordes, Antoine, et al. (2013)

Analyzing Impact of Co-occurrence Statistics

Co-occurrence Counting Pipeline



Mintz, Mike, et al. (2009)

Analyzing Impact of Co-occurrence Statistics

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 show the average hits@1 for each bin.

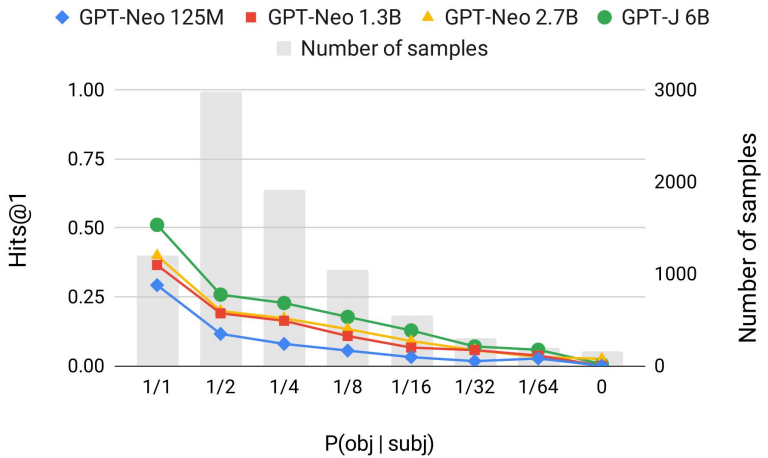
Experimental Setup

- We test open-source versions of GPT-3 with four different model sizes: **GPT-Neo 125M, 1.3B, 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.

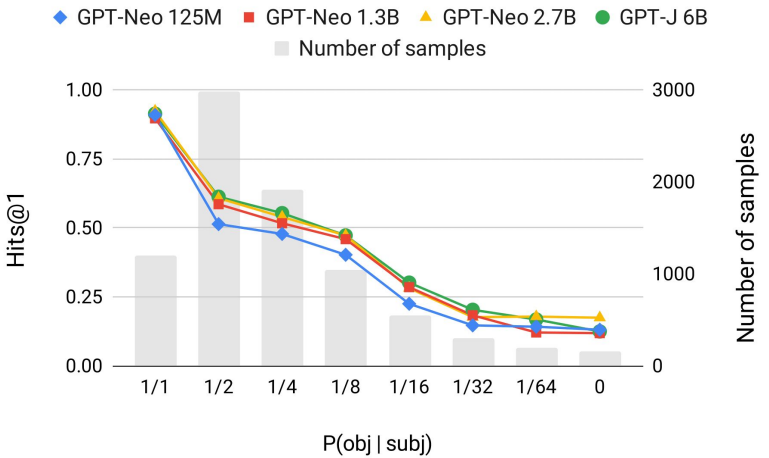
Brown, Tom, et al. (2020), Black, Sid, et al. (2021), Wang, Ben, and Aran Komatsuzaki. (2021), Wolf, Thomas, et al. (2020), Gao, Leo, et al. (2020)

Results

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



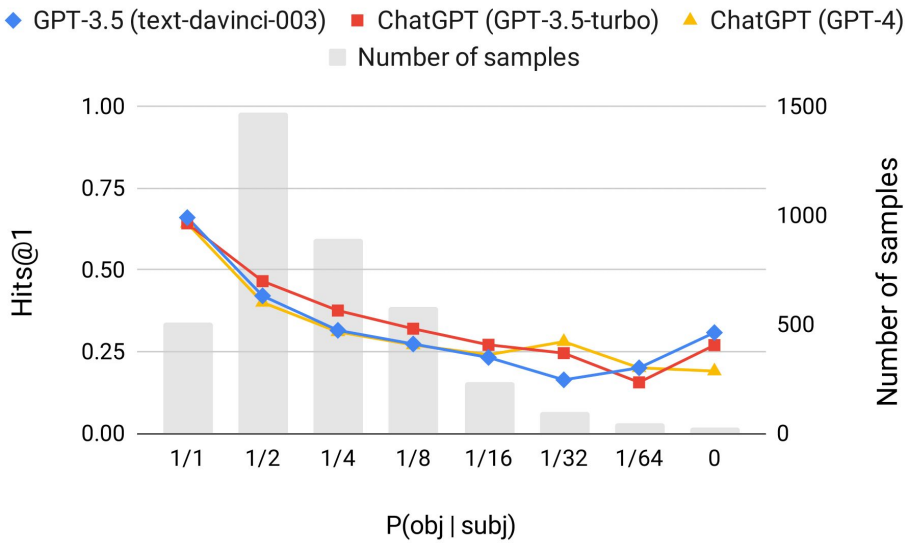
Zero-shot: We observe **a strong correlation between hits@1 and the co-occurrence**. As a result, **LLMs struggle to recall rare facts**. We observe that such correlation remains **despite scaling up model sizes**.



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

Results

- We test **larger models (GPT-3 175B and ChatGPT)** to verify that such correlation remains despite scaling up model sizes.



*Note that the pre-training data of OpenAI GPTs are not the same as the Pile, but we use the results as a proxy.

https://openai.com/blog/chatgpt

Results

- The quantitative failure analysis of GPT-J 6B, **counting how often the correct answer is overridden by a word with higher co-occurrence**. We observe that a word with higher co-occurrence is preferred over the correct answer in a total of **38%** of the failure cases. The results of different frequency bins show that the **co-occurrence bias is more problematic when recalling rare facts**.

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%

Conclusion

- We reveal that **LLMs are vulnerable to the co-occurrence bias**, defined as **preferring frequently co-occurred words over the correct answer** without understanding the meaning behind words.
- Consequently, **LLMs struggle to recall rare facts**.
- 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**.

References

- Elsahar, Hady, et al. "T-rex: A large scale alignment of natural language with knowledge base triples." *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 2018.
- Petroni, Fabio, et al. "Language Models as Knowledge Bases?." *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019.
- Bordes, Antoine, et al. "Learning structured embeddings of knowledge bases." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 25. No. 1. 2011.
- Bordes, Antoine, et al. "Translating embeddings for modeling multi-relational data." Advances in neural information processing systems 26 (2013).
- Xiong, Wenhan, et al. "Pretrained Encyclopedia: Weakly Supervised Knowledge-Pretrained Language Model." International Conference on Learning Representations. 2020.
- Ravichander, Abhilasha, et al. "On the systematicity of probing contextualized word representations: The case of hypernymy in BERT." Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics. 2020.
- Mintz, Mike, et al. "Distant supervision for relation extraction without labeled data." Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. 2009.
- Brown, Tom, et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901.
- Black, Sid, et al. "Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow." (2021).
- Wang, Ben, and Aran Komatsuzaki. "GPT-J-6B: A 6 billion parameter autoregressive language model." (2021).
- Wolf, Thomas, et al. "Transformers: State-of-the-art natural language processing." Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations. 2020.
- Gao, Leo, et al. "The pile: An 800gb dataset of diverse text for language modeling." arXiv preprint arXiv:2101.00027 (2020).

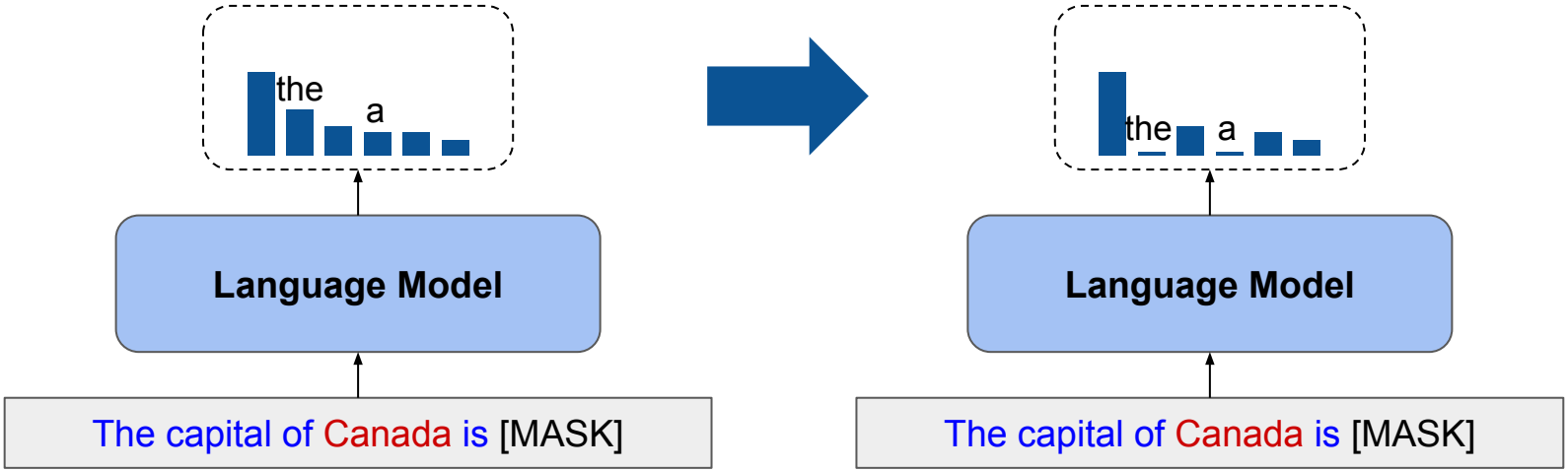
Thank You

Appendix

Factual Knowledge Probing

Restricted Output Candidates

- We restrict output candidates **as LMs are not trained to act as knowledge bases**. Specifically, we **remove stopwords** from the output candidates.



Xiong, Wenhan, et al. (2020), Ravichander, Abhilasha, et al. (2020)

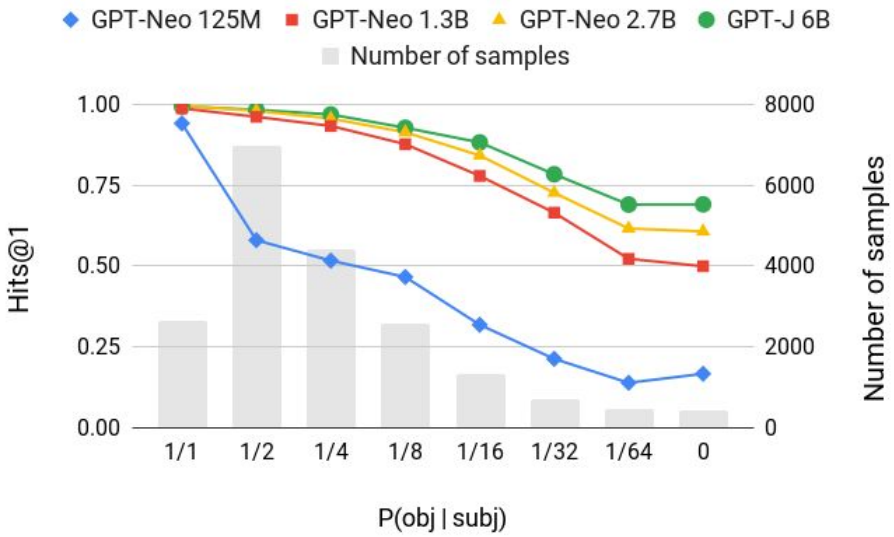
Results

- We **compare the conditional probability of predictions and gold objects.** We find that **LLMs prefer to generate words that co-occur with the subject frequently enough (** $P(\hat{obj}|subj)$ **≥ 0.26)**. In other words, **recalling rare facts is especially difficult** since **words with low co-occurrence counts are hardly generated**.

Frequency bin	$P(\hat{obj}\|subj)$	$P(obj\|subj)$
1/1	0.42±0.31	1.00±0.00
1/2	0.38±0.28	0.72±0.14
1/4	0.37±0.27	0.37±0.07
1/8	0.31±0.26	0.18±0.04
1/16	0.29±0.29	0.09±0.02
1/32	0.30±0.31	0.05±0.01
1/64	0.26±0.32	0.02±0.00
0	0.26±0.30	0.01±0.00
Total	0.35±0.29	0.46±0.32

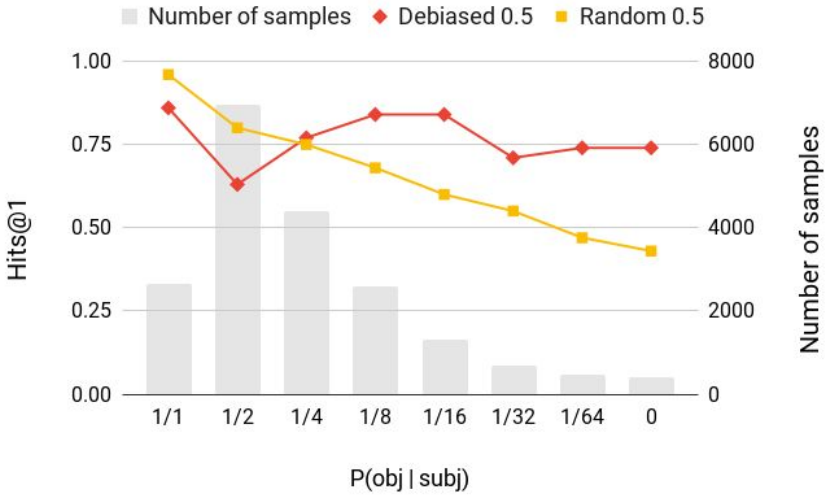
Results

- The correlation analysis of the **finetuned models on the training set**.
  - Interestingly, **LLMs struggle to learn facts that rarely appear in the pretraining** corpora **although they are explicitly given during finetuning**.

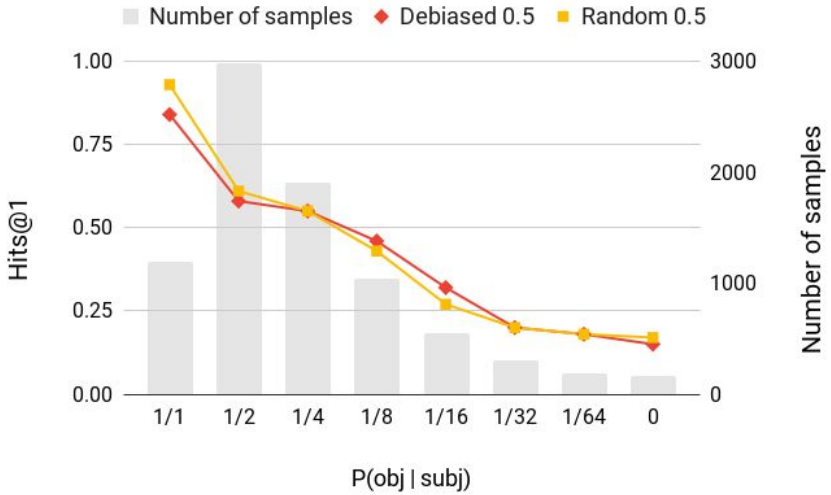


Results

- **Mitigation:** debiasing with undersampling
  - Undersampling is **effective in memorizing seen facts** with little sacrifice on frequent facts.
  - The effect is **not generalizable to unseen facts**.



Results on the **training set**



Results on the **test set**