Scaling Up Membership Inference: When and How Attacks Succeed on LLMs

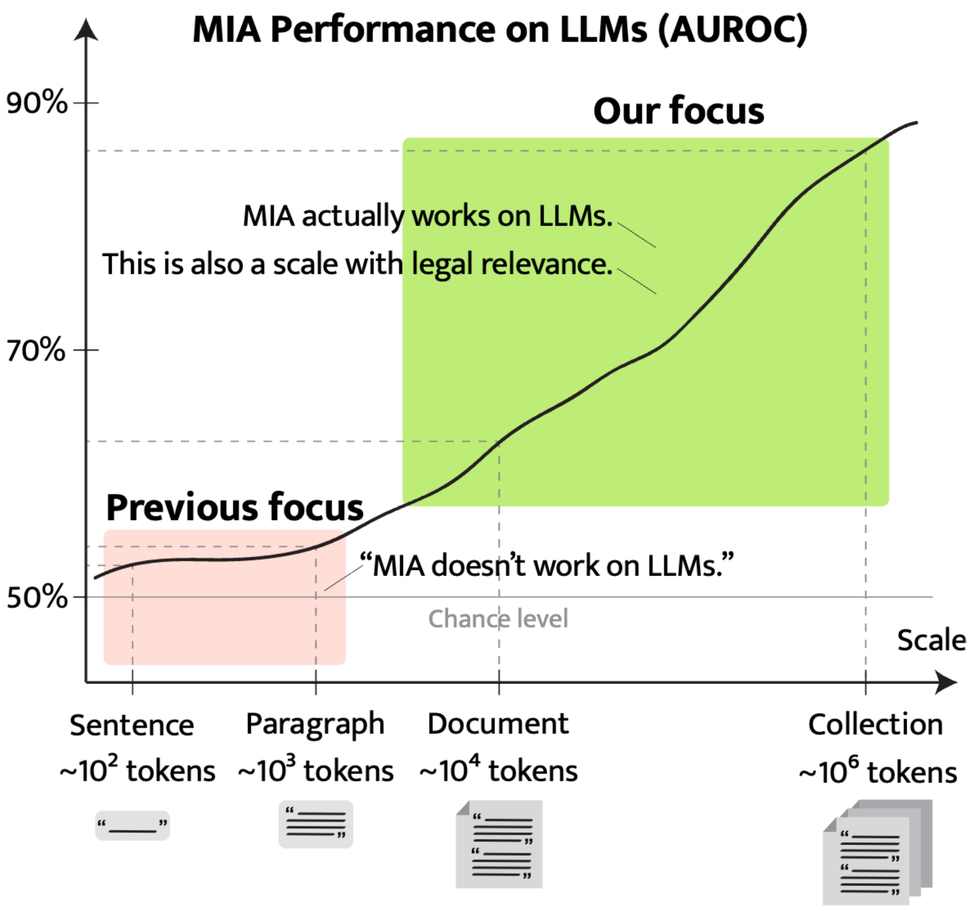
Haritz Puerto[1,2], Martin Gubri[1], Sangdoo Yun[3], Seong Joon Oh[1,4,5]

[1]Parameter Lab, [2]UKP Lab, TU Darmstadt, [3]Naver AI, [4]University of Tübingen, [5]Tübingen AI Center

haritz.puerto@tu-darmstadt.de

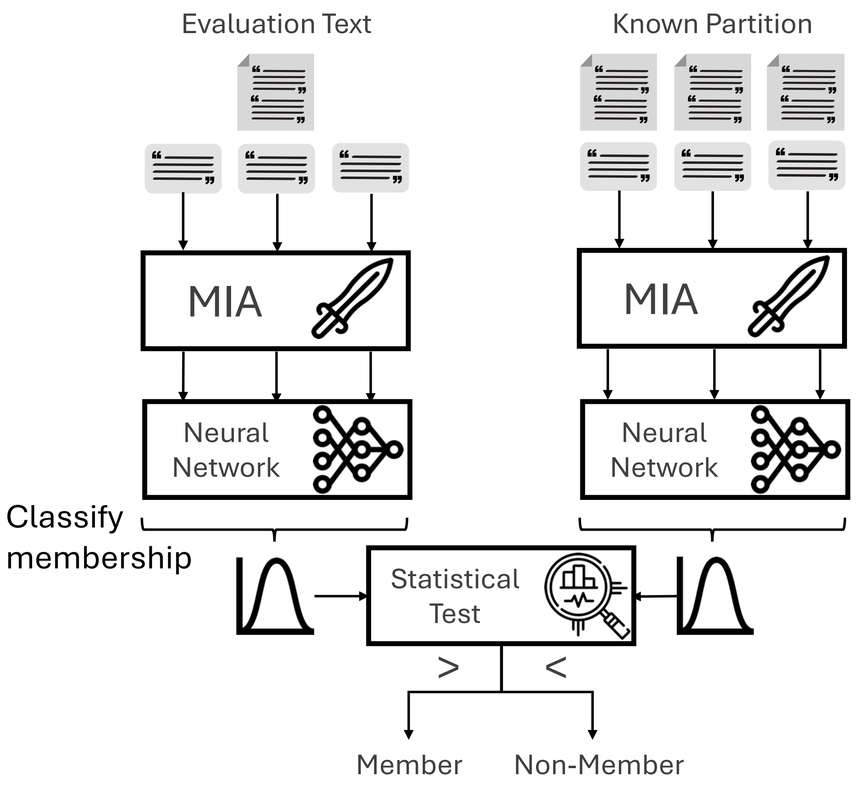
LLM training data can be detected, but we need 10k+ tokens

Motivation



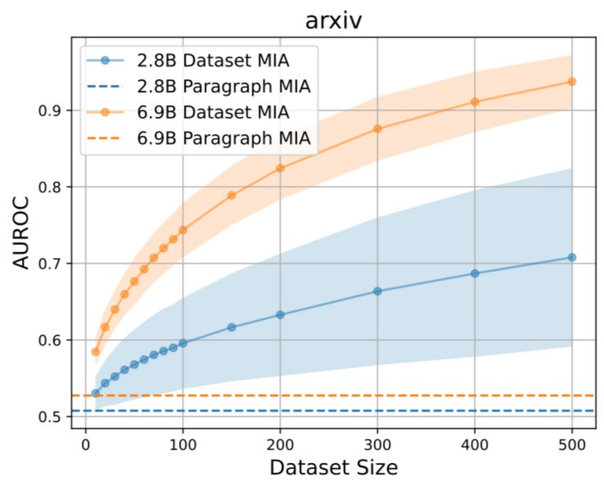
- 😡 Content creators are concerned about the use of their copyrighted data to train LLMs
- 😰 Membership Inference Attacks (MIA) are considered ineffective on LLMs for text chunks

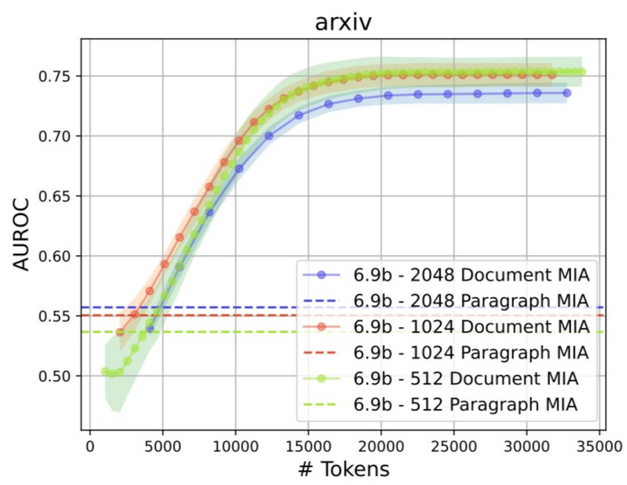
Method



- Extension of [1] to any data scale
- We compare the MIA dist. to a dist. of known members

Pretraining





Pythia + The Pile

MIA performance improves as we scale up the data

Data Scale	ArXiv	HackerNews	Wiki
Sentence	0.501±0.003	0.500±0.003	0.507±0.004
Paragraph	0.528±0.004	0.511±0.015	0.523±0.013
Document	0.697±0.060	0.513±0.040	0.560±0.011
Collection (500)	0.943±0.025	0.709±0.340	0.844±0.132

Pythia 6.9B

Training Impact

Data scale can be reduced as we increase the data recency in the training

MIA	AUROC
Sentence	0.793 ± 0.024
Collection (20)	0.993 ± 0.012

Phi 2 – FT on QA

Dataset	Scale	Pretrained	Continual Learning
ArXiv	Paragraph	0.509 ± 0.006	0.587 ± 0.009
	Document	0.523 ± 0.01	0.582 ± 0.06
	Collection	0.718 ± 0.122	1.0 ± 0.0
GitHub	Paragraph	0.494 ± 0.009	0.559 ± 0.017
	Document	0.498 ± 0.01	0.579 ± 0.014
	Collection	0.479 ± 0.069	0.885 ± 0.064
Wikipedia	Paragraph	0.534 ± 0.015	0.577 ± 0.012
	Document	0.531 ± 0.019	0.590 ± 0.015
	Collection	0.665 ± 0.169	0.997 ± 0.007

Pythia 2.8B

Conclusion & Links

- MIA can be effective on pretrained LLMs at the right scale (long documents and collections of documents)
- Fine-tuning increases the effectiveness of MIA
- MIA can be used to detect test-set contamination

Paper | Code | Data | Learn More

[1] Pratyush et al. "LLM Dataset Inference: Did you train on my dataset?." NeurIPS 2024