Did the Neurons Read your Book? Document-level Membership Inference for Large Language Models

Matthieu Meeus, Shubham Jain, Marek Rei, Yves-Alexandre de Montjoye

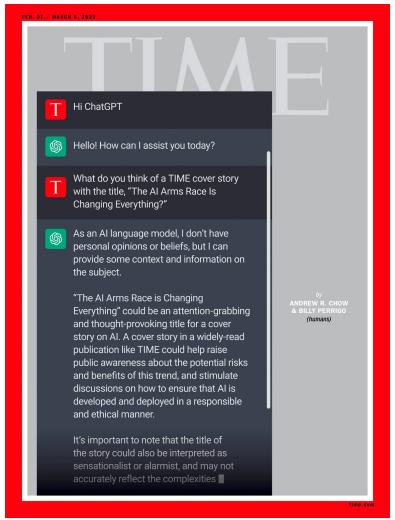
33rd USENIX Security Symposium (USENIX Security 2024) - August 15, Philadelphia

Imperial College London

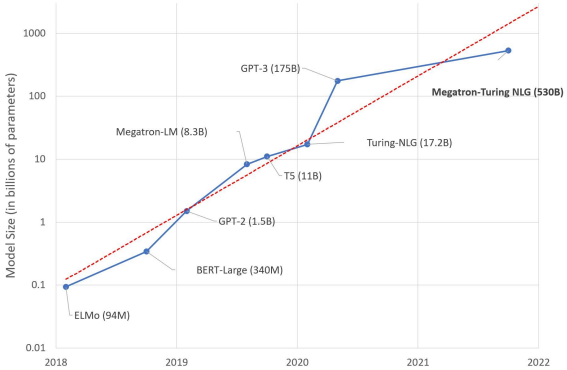
COMPUTATIONAL PRIVACY GROUP

Large Language Models (LLMs) are ubiquitous

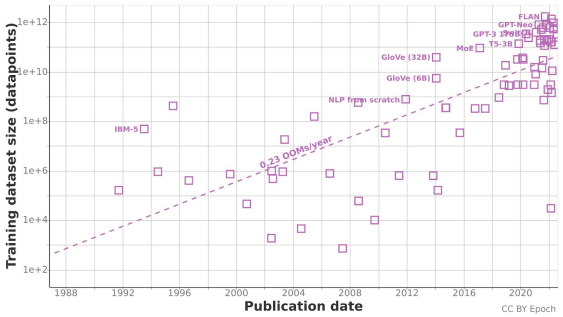
LLMs are **increasingly embedded** into our daily lives



Their capabilities come from ever **growing models** (GPT-4 has 1.76 trillion parameters)...



… trained on ever **growing datasets** of human generated text (LLaMA-3 is trained on 15 trillion tokens)

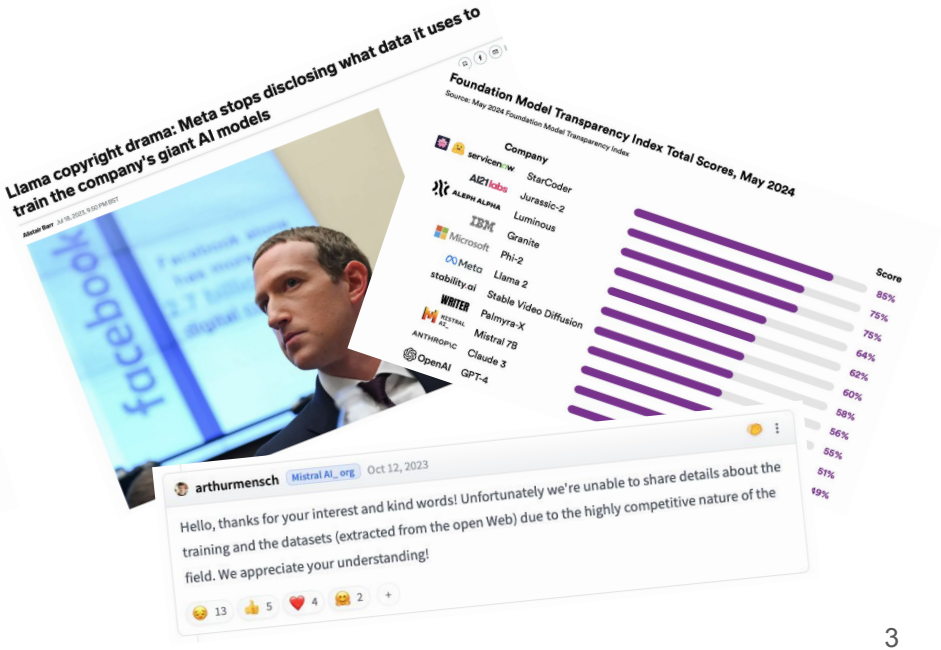


Growing demand for transparency in the pretraining dataset

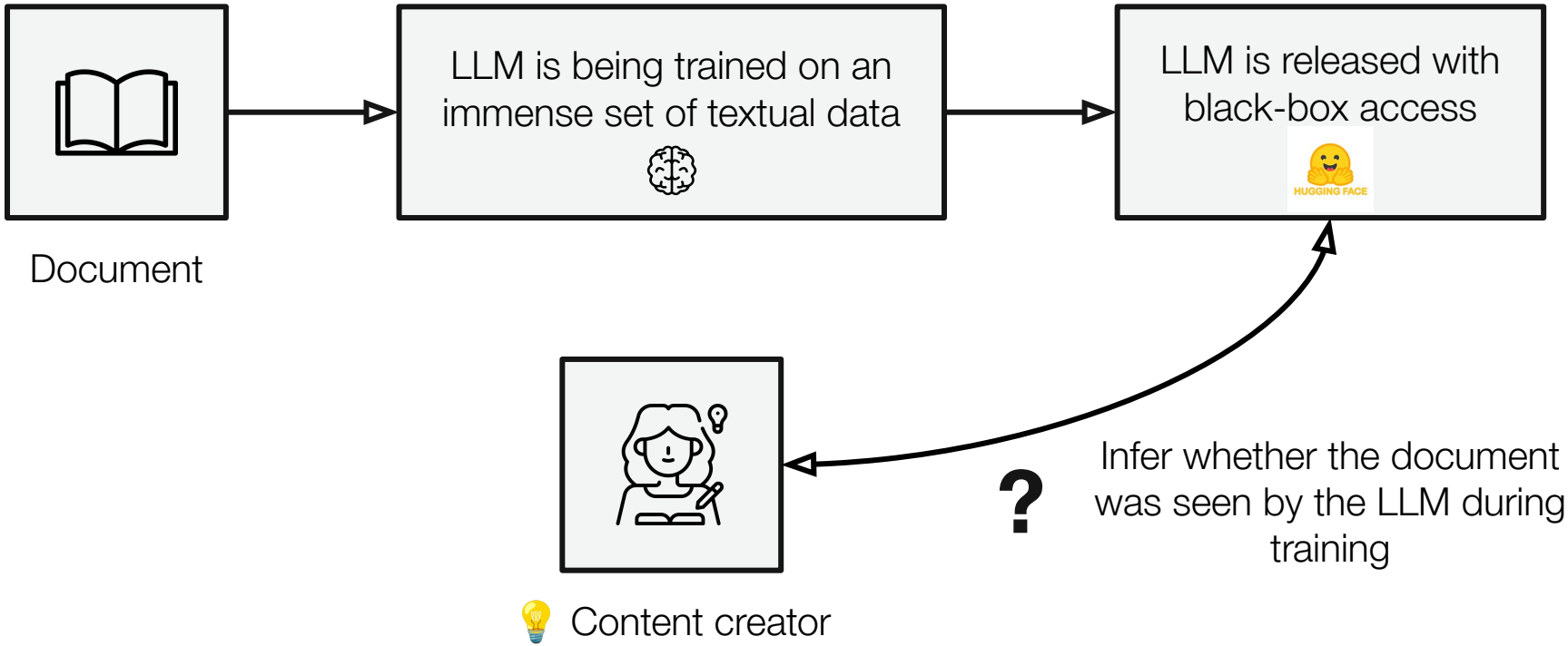
Questions are being raised about what these models actually learn from, e.g.

- Could they propagate bias, misinformation?
- Are they trained on copyrighted content?
- Was there any data contamination?

LLM developers are increasingly reluctant to disclose details on their training data.



We introduce document-level membership inference for LLMs



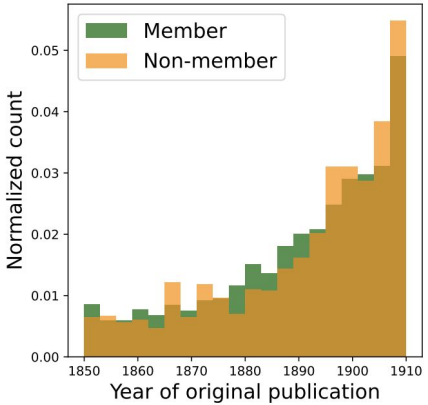
Methodology and results*

*We elaborate on the **work done in October ‘23**. Since then, the field has moved on, with **concerns** raised on document-level membership inference in practice. *More on that in a bit.*

Step 1: collect *member* and *non-member* documents

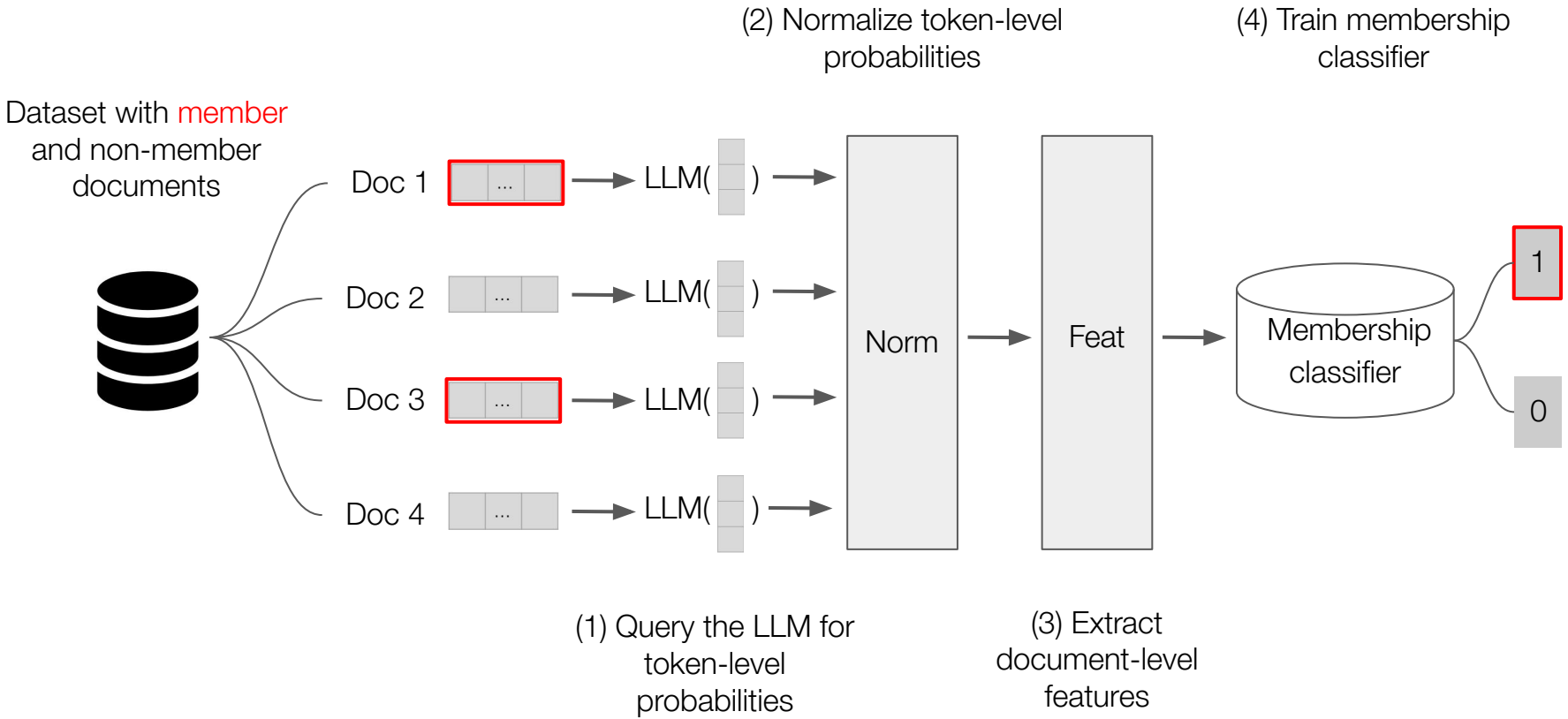
- LLMs are trained on immense amounts of text, typically scraped from the internet (e.g. Wikipedia, Reddit).
- But also from more high-quality sources such as **books** (Project Gutenberg, Books3) or **academic papers from ArXiv**.

- Collect documents that were and were not used for training:
  a. ***Members****:* documents sampled from **datasets often used to train LLMs**.
  b. ***Non-members:*** documents made available on the same sources **after the model release date**, and thus likely not used to train the LLM.
     - Books: every day books are being added to Project Gutenberg, which can easily be scraped.
     - Academic papers (LaTeX): every day novel research is made available on ArXiv.



Control for original publication year for books

Step 2: build a document-level membership classifier



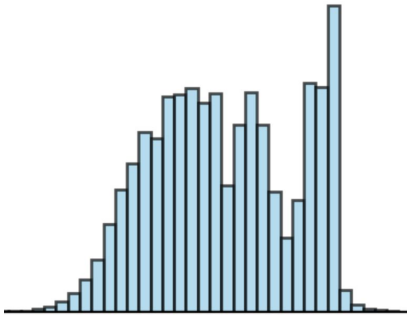
Step 2: build a document-level membership classifier

(1) Querying the LLM: We query the LLM for predicted probability of the true token appearing in the document.

- We run through the entire book with a certain context length to retrieve a predicted probability for each token in the book (100k+ values).

…"It's true, indeed, Smerdyakov is accused only by the prisoner, his two brothers, and Madame Svyetlov. But there are others who accuse him: there are vague rumors of a question, of a suspicion, an obscure report, a feeling of expectation…

LLM(rumors|context) = 0.09



All normalized probabilities in a document

(2) Normalization: The predicted probability might not carry meaningful information about membership, as the LLM can just be good at generalization <> memorization.

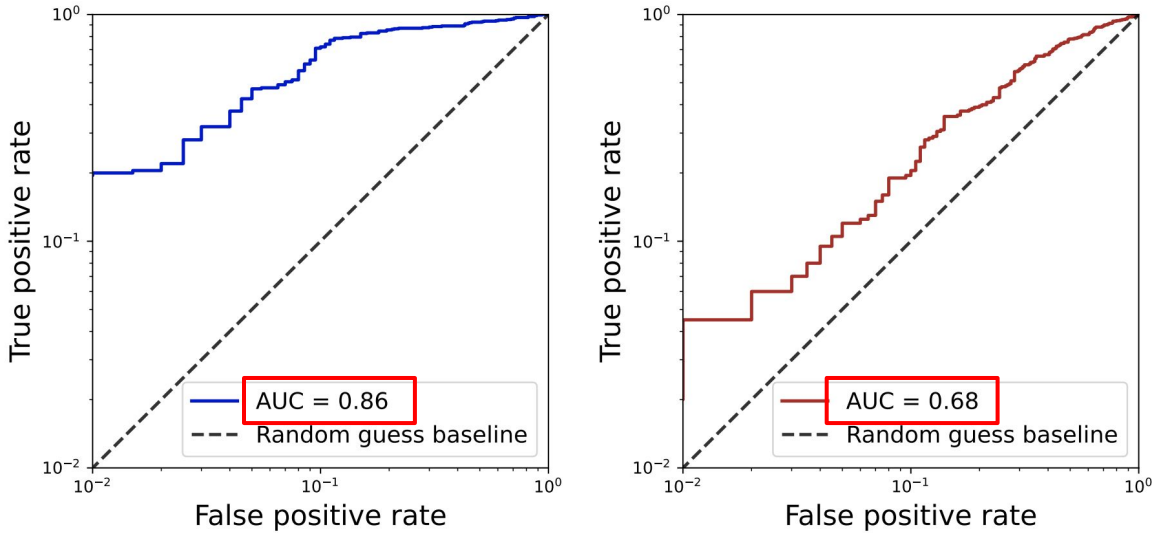
- We **normalize the token-level probability** by how frequently the token appears in a reference dataset.

(3) Feature aggregation: compute the normalized count within the bin of a histogram to capture the entire distribution.

(4) Membership classifier: feed these features into a binary random forest classifier to predict membership.

Results: membership AUC reaches 0.86 for books

- We apply our methodology to OpenLLaMA 7B - an open reproduction of LLaMA, to be in full control of which data has been used for training.
- AUC for binary membership for books (left) and ArXiv papers (right) in the best setup.
  - Reaching **high values** for both setups!



The field has moved on since then

Other work has further studied post-hoc MIAs for LLMs…

- Concurrent work[1] (ICLR 2024) has proposed pretraining data detection for LLMs.
  - Also relying on post-hoc collection of non-member data, with Wikipedia articles.
  - Also reaching a membership AUC of 0.8+.
- The same dataset of members and non-members has since then been widely used to evaluate membership inference for LLMs[2].

… but has also raised concerns.

- Concerns have been raised that this post-hoc collection of non-members constitutes a **distribution shift** between members and non-members[3].
- This would make the resulting **AUC not attributable to the memorization** of the target LLM.

1. Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. arXiv preprint arXiv:2310.16789, 2023. October 25, 2023
2. Zhang, J., Sun, J., Yeats, E., Ouyang, Y., Kuo, M., Zhang, J., Yang, H. and Li, H., 2024. Min-K%++: Improved Baseline for Detecting Pre-Training Data from Large Language Models. arXiv preprint arXiv:2404.02936. April 3, 2024
3. Duan, M., Suri, A., Mireshghallah, N., Min, S., Shi, W., Zettlemoyer, L., ... & Hajishirzi, H. (2024). Do membership inference attacks work on large language models?. *arXiv preprint arXiv:2402.07841*.

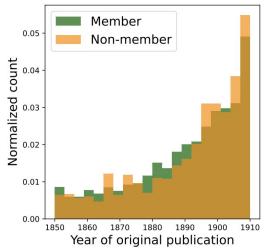
Distribution shift between members and non-members

- Concurrent works have proposed **model-less baselines**[1,2]: a classifier to distinguish between members and non-members by just looking at the documents and not at the model (e.g. bag of words).
- The **resulting AUC** shows how this dataset indeed suffers from a strong distribution shift, ***rendering it impossible to attribute the previously reported AUC to LLM memorization.***

Dataset	AUC
Project Gutenberg (full)	0.970
ArXiv	0.720
WikiMIA	0.987

- Sometimes this distribution shift **can be very subtle**.
  - Recall that we control for publication year for member and non-member books?
  - The most predictive words in the bag of word classifier include the use of politically incorrect language, which has changed with the upload date.

Most predictive words
{ colored,cæsar,n****,mediæval }



1. Meeus, M., Jain, S., Rei, M., & de Montjoye, Y. A. (2024). Inherent Challenges of Post-Hoc Membership Inference for Large Language Models. *arXiv preprint arXiv:2406.17975*.
2. Das, D., Zhang, J., & Tramèr, F. (2024). Blind Baselines Beat Membership Inference Attacks for Foundation Models. *arXiv preprint arXiv:2406.16201*.

What are the alternatives?

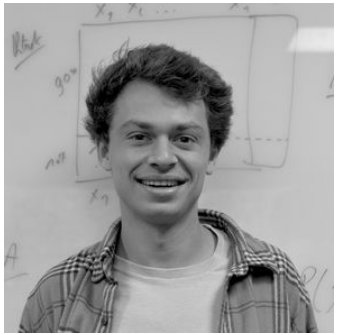
- Some publicly available models (e.g. Pythia suite) provide a **randomized split of train and test data.**
  - This provides a great setup to develop and evaluate new MIAs.
  - But also comes with its challenges
    - Research has found a **significant overlap** between train/test documents, making it harder for MIAs to succeed[1].
    - Such a held-out test split is **not available** for recent, real-world LLMs.
- Sample member and non-members **as closely as possible to the training data cutoff date.**
  - We instantiate this for ArXiv papers published just one month apart and confirm that the distribution shift is largely mitigated (bag of words classifier AUC of 0.52).
  - Our document-level membership inference methodology performs only marginally better than a random guess, with an AUC of 0.54[2].
- Other approaches: **controlled injection of highly unique sequences** in the pretraining dataset, e.g. copyright traps (ICML 2024)[3,4].

1. Duan, M., Suri, A., Mireshghallah, N., Min, S., Shi, W., Zettlemoyer, L., ... & Hajishirzi, H. (2024). Do membership inference attacks work on large language models?. *arXiv preprint arXiv:2402.07841*.
2. Meeus, M., Jain, S., Rei, M., & de Montjoye, Y. A. (2024). Inherent Challenges of Post-Hoc Membership Inference for Large Language Models. *arXiv preprint arXiv:2406.17975*.
3. Meeus, M., Shilov, I., Faysse, M., & de Montjoye, Y. A. Copyright Traps for Large Language Models. In *Forty-first International Conference on Machine Learning*.
4. Wei, J. T. Z., Wang, R. Y., & Jia, R. (2024). Proving membership in LLM pretraining data via data watermarks. *arXiv preprint arXiv:2402.10892*.

Conclusion

- Questions are being raised on what data LLMs are trained on…
- …while model developers become reluctant to disclose details on their training dataset.
- We propose the task of document-level membership inference for LLMs and a methodology to do so.
- Implementing this in practice comes with **inherent challenges.** Still to be determined if
  - (i) it works in a clean setup
  - (ii) if a clean setup can be achieved for real-world LLMs.
- In any case, we emphasize the need of a **clean and controlled setup to develop and evaluate MIAs against LLMs.**

Thank you for your attention! Happy to discuss any further questions.



@matthieu_meeus

@shubhamjain0594

@MarekRei

@yvesalexandre

Main paper

Follow-up paper elaborating on the challenges

Code

Imperial College London

COMPUTATIONAL PRIVACY GROUP