Five Queries Are Enough: Query-Efficient and Surrogate-Free Membership Inference Attacks on RAG via Entailment

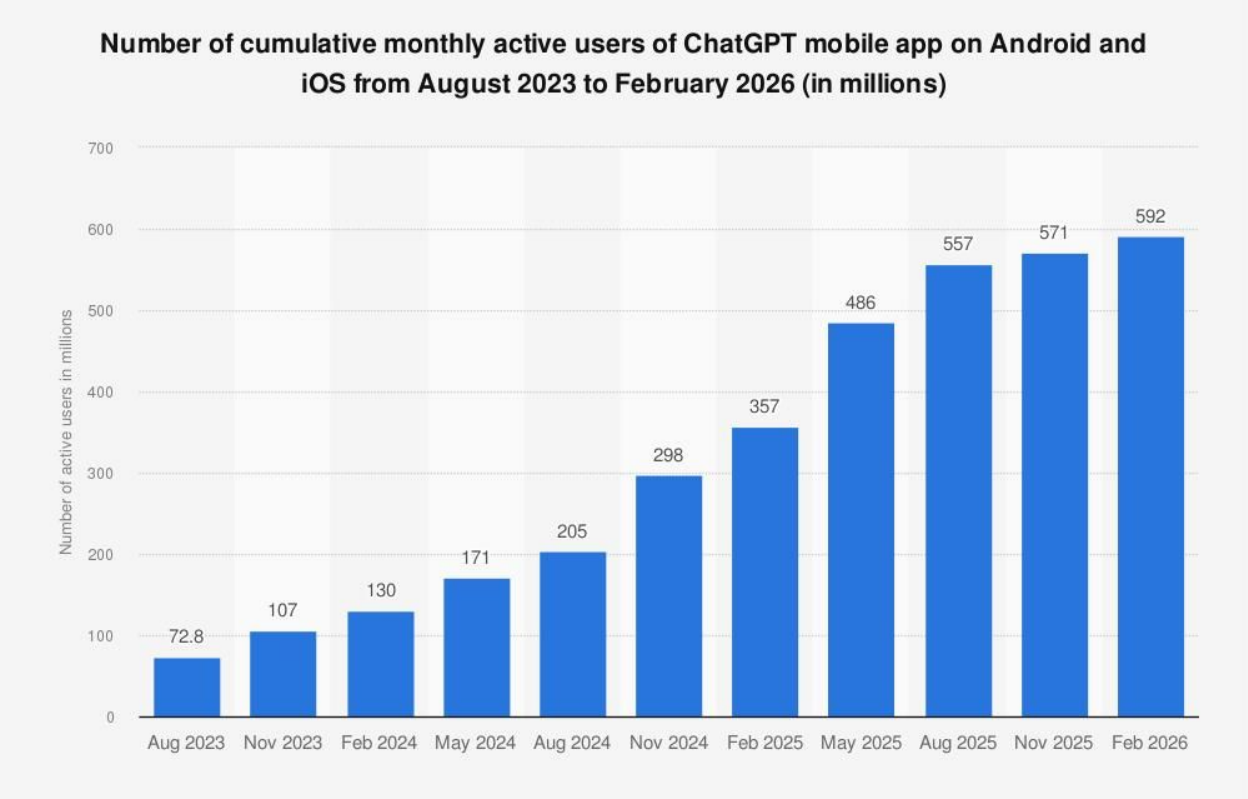
Nguyen Linh Bao Nguyen, Wanlun Ma, Viet Vo,
Alsharif Abuadbba, Minghong Fang, Jun Zhang, Yang Xiang

USENIX Security 2026



Background and Motivation

The rise of Large Language Models (LLMs)



Source: https://www.statista.com/statistics/1625099/monthly-active-users-of-chatgpt-mobile-app

ChatGPT

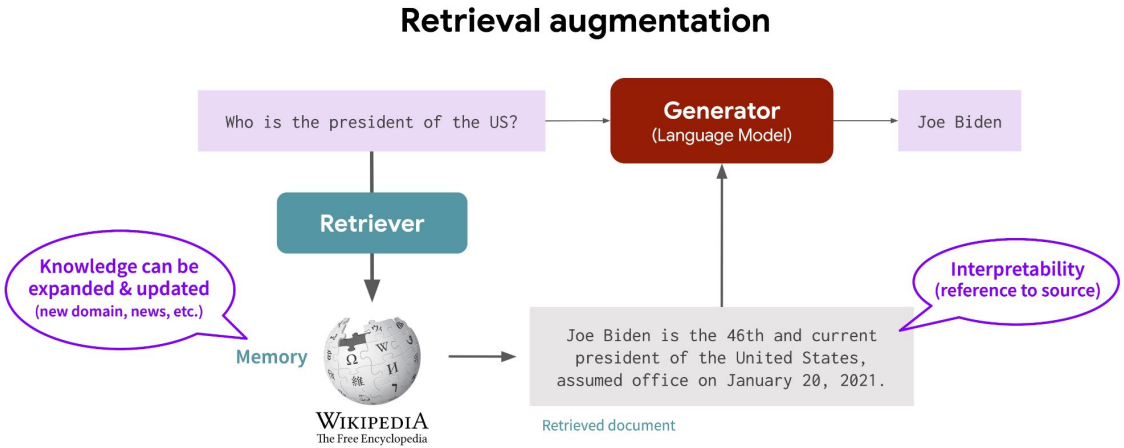
Claude

Gemini

deepseek

Retrieval-Augmented Generation (RAG)

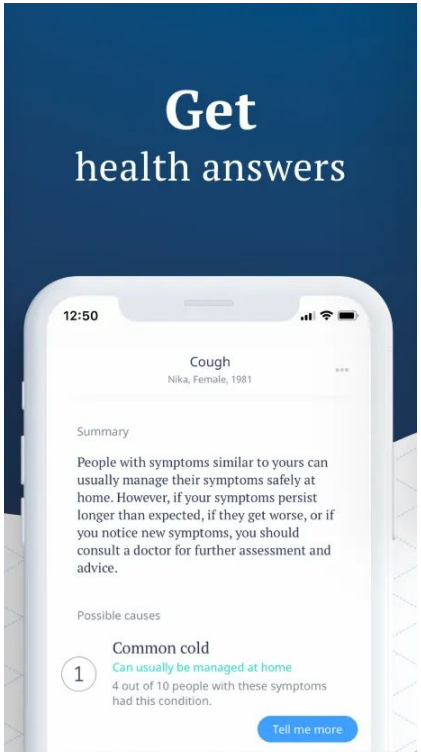
- Retriever searches for relevant information in database
- Generator produces an answer using that information.



Source: https://cs.stanford.edu/~myasu/blog/racm3/

RAG with Private Database

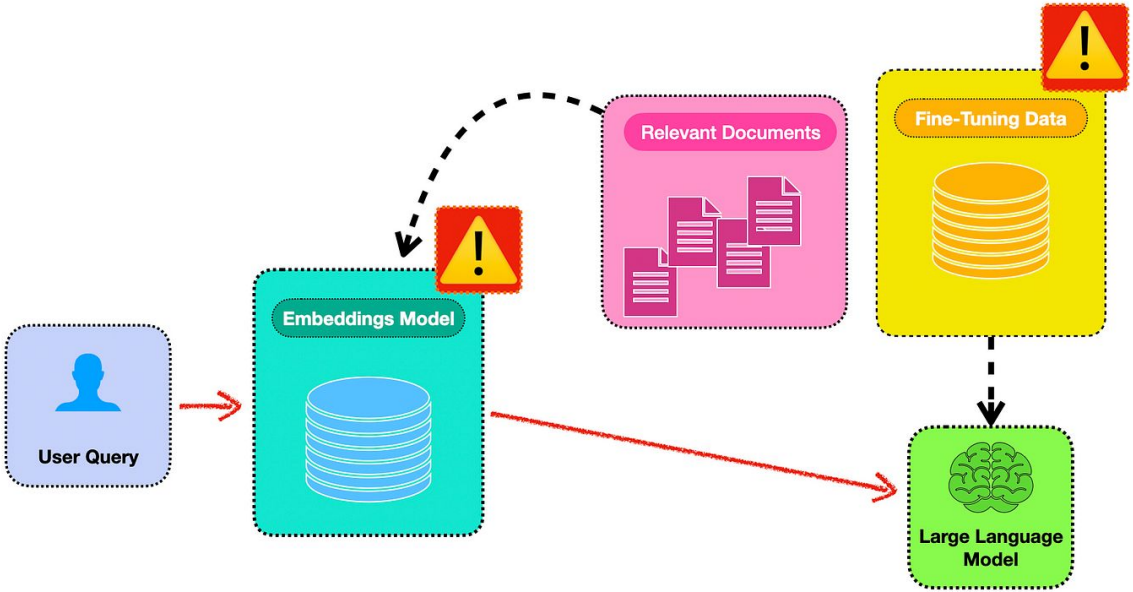
- **Ada Health**
  - Symptom assessment platform
  - Uses RAG to retrieve data from a private clinical knowledge base
- **SuperTOBi (Vodafone)**
  - Customer support chatbot
  - Uses RAG to support LLM answers with internal policy



Sources:
https://www.vodafone.com/news/newsroom/technology/meet-super-tobi-vodafone-s-new-generative-ai-virtual-assistant-now-serving-customers-in-multiple-countries
https://ada.com/

Privacy Issues in RAG

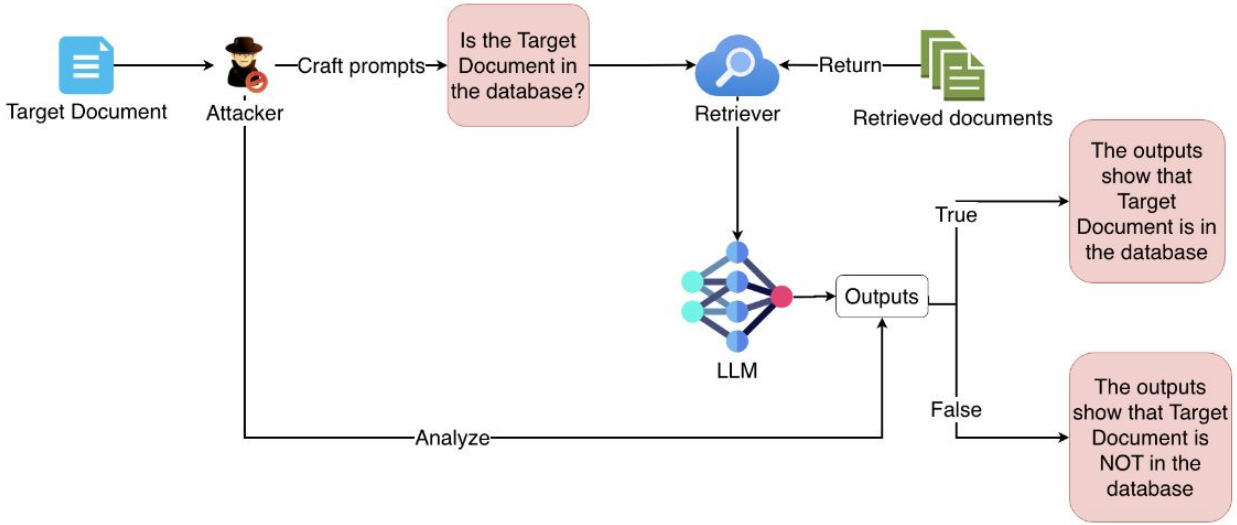
Integrating RAG into LLM generation introduces new threat surfaces



Source: https://cobusgreyling.medium.com/rag-data-privacy-attack-methods-safe-prompts-f6576a5d8962

Membership Inference Attacks in RAG

Check if specific document is present in the RAG database



Existing MIAs in RAG



Document

Recover document
(S^2-MIA [1], RAGLeak [2], Prompt-Inj [3], Good/bad [4])

Complete the following doc: [PARTIAL DOC]

Existing MIAs in RAG



Document

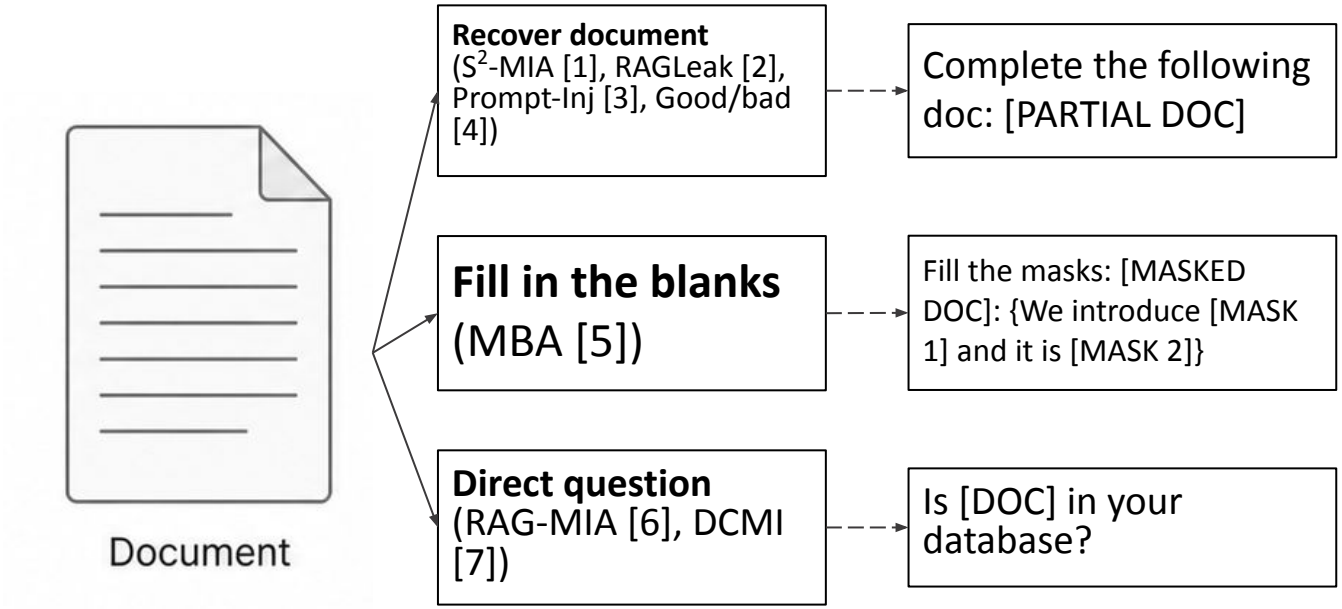
Recover document
(S^2-MIA [1], RAGLeak [2], Prompt-Inj [3], Good/bad [4])

Complete the following doc: [PARTIAL DOC]

Fill in the blanks
(MBA [5])

Fill the masks: [MASKED DOC]: {We introduce [MASK 1] and it is [MASK 2]}

Existing MIAs in RAG



Existing MIAs in RAG



Document

Recover document (S^2-MIA [1], RAGLeak [2], Prompt-Inj [3], Good/bad [4])

Complete the following doc: [PARTIAL DOC]

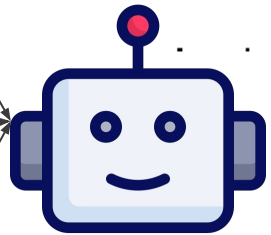
Fill in the blanks (MBA [5])

Fill the masks: [MASKED DOC]: {We introduce [MASK 1] and it is [MASK 2]}

Direct question (RAG-MIA [6], DCMI [7])

Is [DOC] in your database?

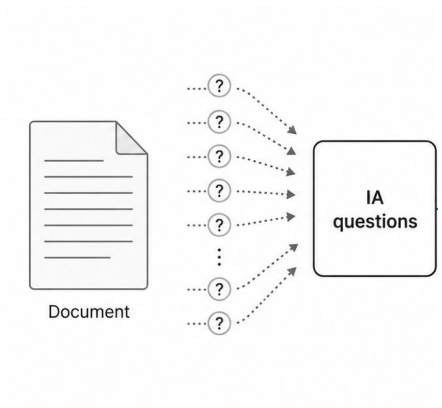
These queries look suspicious



Easily detected because of using malicious queries

Interrogation Attack (IA) [8]

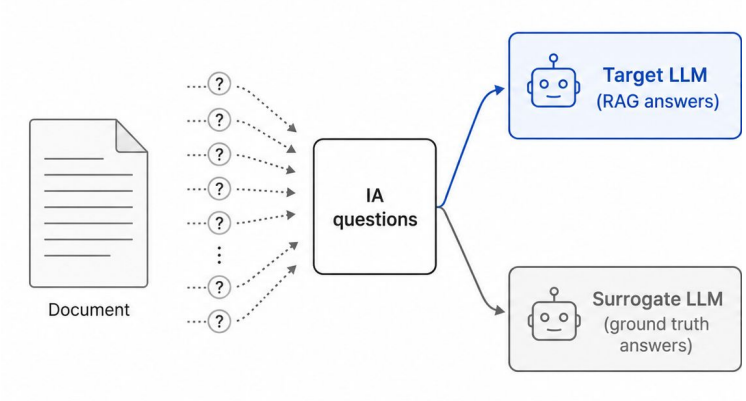
- IA crafts 30 yes/no/idk questions from the target document D. For example:
  - “In D, is experiment X’s result = Y?”



Interrogation Attack (IA) [8]

- IA crafts 30 yes/no/idk questions from the target document D. For example:
  - "In D, is experiment X's result = Y?"

- The set of questions will be sent to
  - Target LLM of RAG → RAG answers
  - Surrogate LLM → Ground truth answers

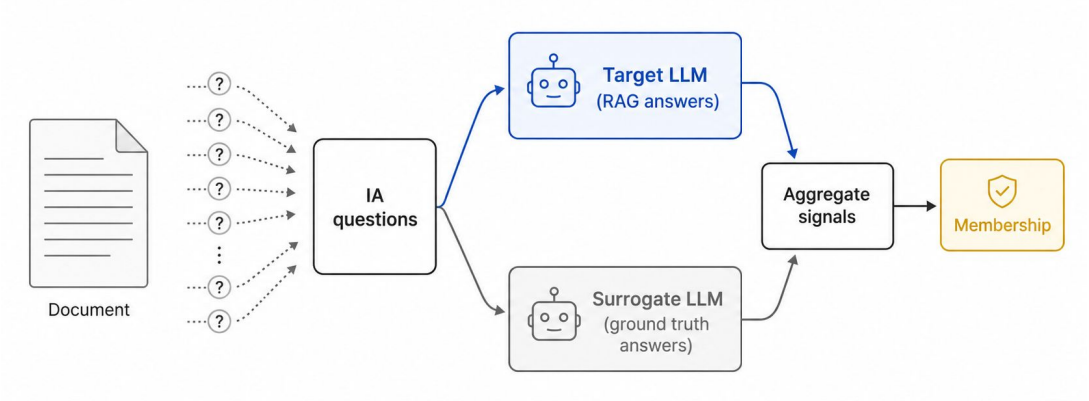


Interrogation Attack (IA) [8]

- IA crafts 30 yes/no/idk questions from the target document D. For example:
  - "In D, is experiment X's result = Y?"

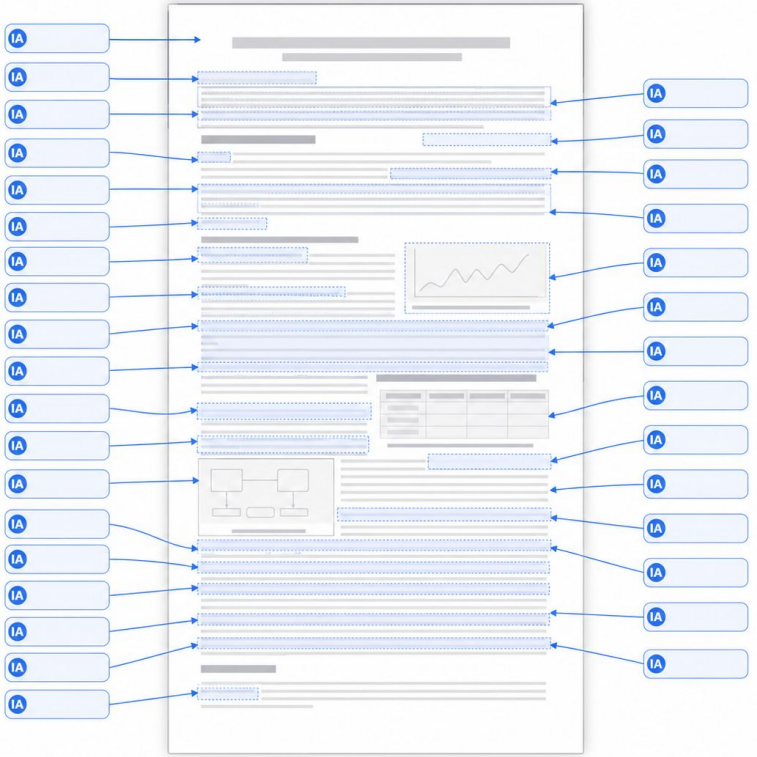
- The set of questions will be sent to
  - Target LLM of RAG → RAG answers
  - Surrogate LLM → Ground truth answers

- If RAG answers ≈ GT answers → Member



Limitations of Interrogation Attack (IA)

Each query targets small details of document
→ Need a lot queries to cover whole document



Membership Entailment Attack (MEntA)

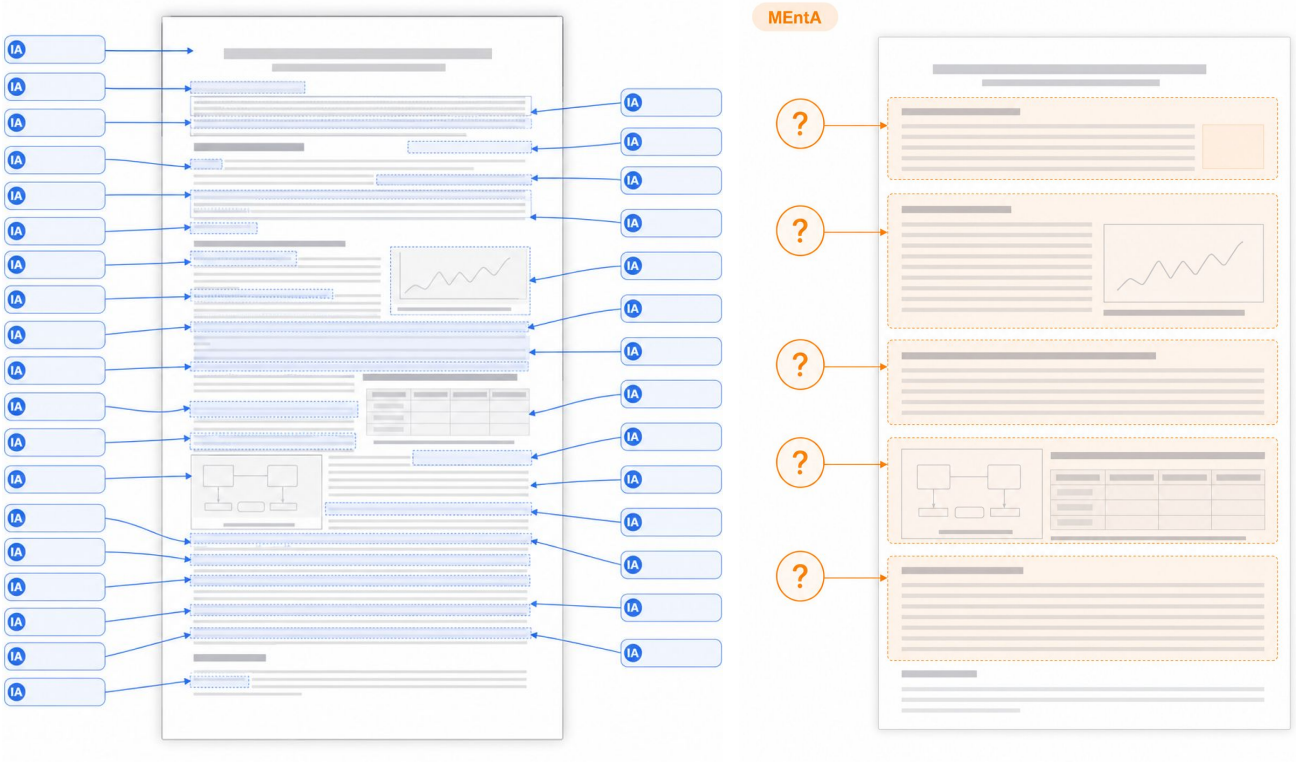


The design of MEntA (Membership Entailment Attack)

Replace **yes/no** questions (leak little info) with **open questions** (leak more info) → Fewer queries

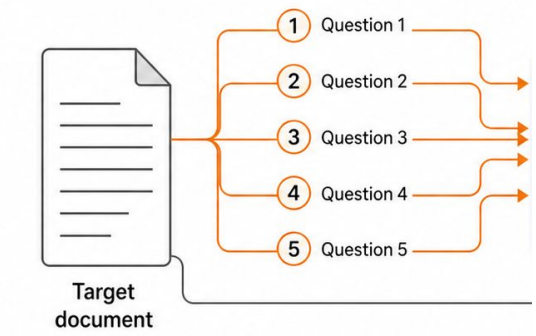
Example

- IA: “In D, is experiment X’s result = Y?” **(sentence level)**
- MEntA: “In D, explain experiment X’s result in details” **(paragraph level)**



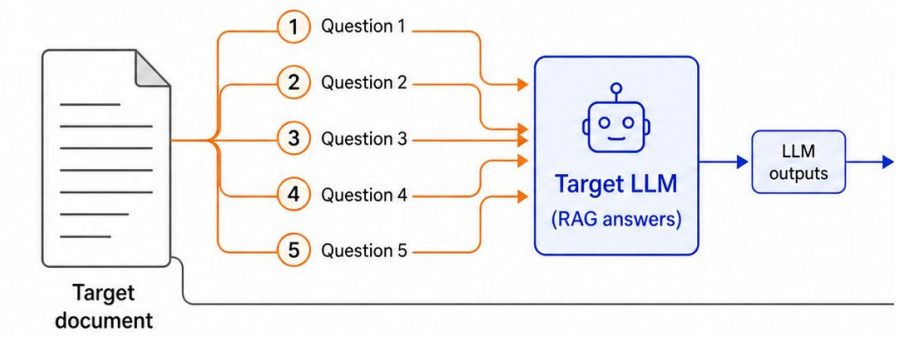
The design of MEntA (Membership Entailment Attack)

- MEntA first crafts 5 open questions from the target document D.



The design of MEntA (Membership Entailment Attack)

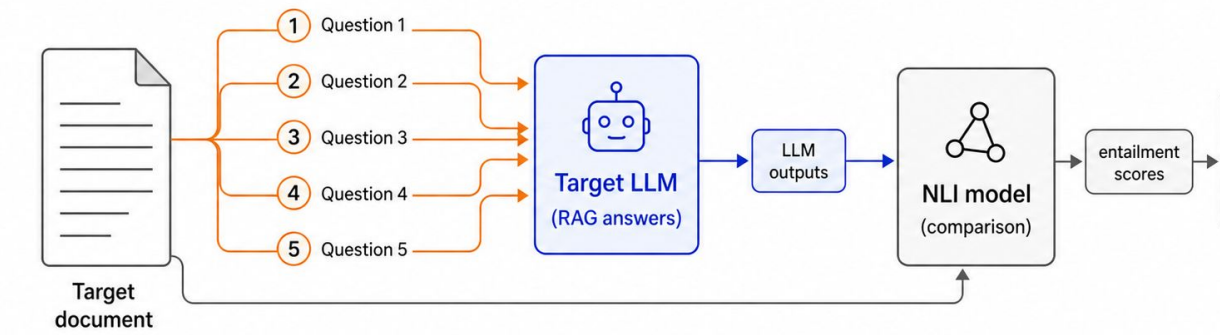
- MEntA first crafts 5 open questions from the target document D.
- The set of questions will be sent to
  - Target LLM of RAG → RAG answers



The design of MEntA (Membership Entailment Attack)

- MEntA first crafts 5 open questions from the target document D.
- The set of questions will be sent to
  - Target LLM of RAG → RAG answers
- NLI model calculates entailment scores between RAG answers and target document. For example:
  - "The result of experiment X is Y"
  - "We conduct experiment X on mice and observe Y"

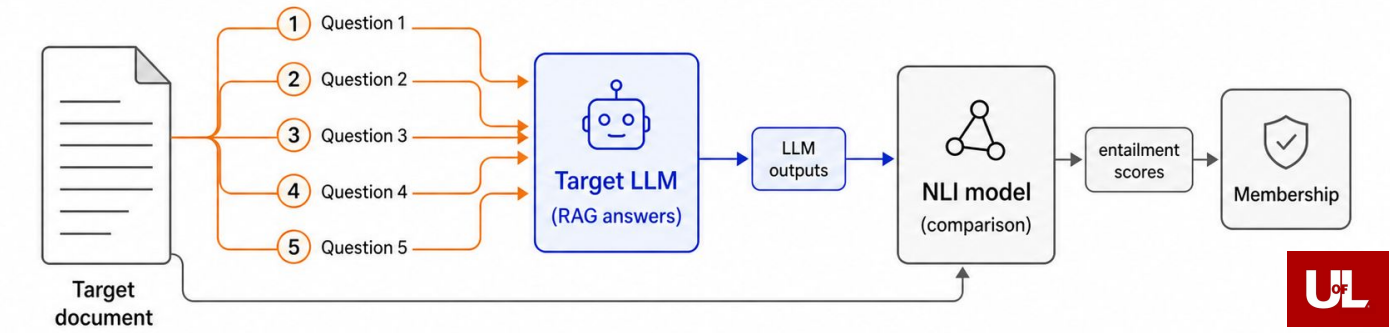
  → High entailment score



The design of MEntA (Membership Entailment Attack)

- MEntA first crafts 5 open questions from the target document D.
- The set of questions will be sent to
  - Target LLM of RAG → RAG answers
- NLI model calculates entailment scores between RAG answers and target document. For example:
  - "The result of experiment X is Y"
  - "We conduct experiment X on mice and observe Y"

  → High entailment score
- Determine membership based on entailment scores



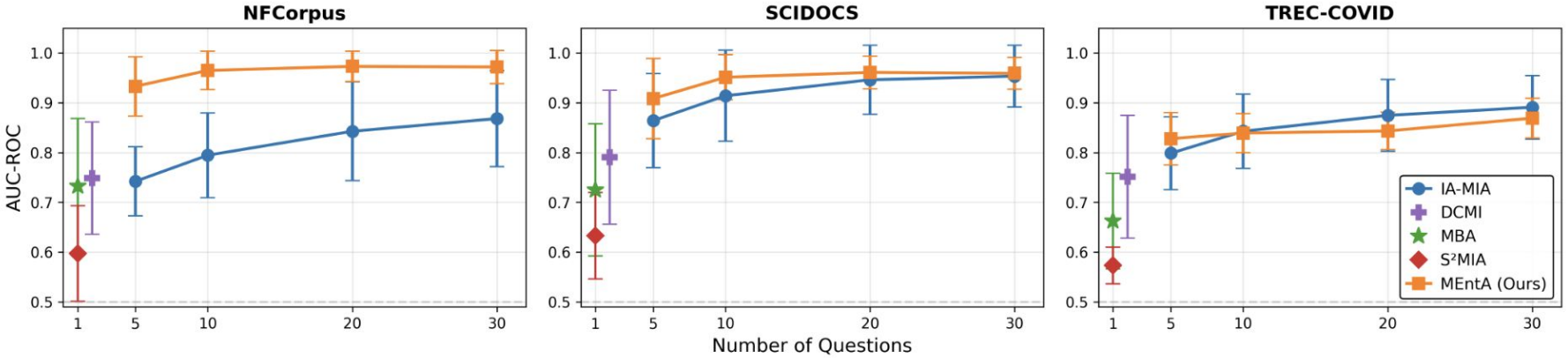
Experiments

Experimental Setup

- **Baselines**
  - S^2-MIA [1] (recover document)
  - MBA [5] (fill in the blanks)
  - DCMI [7] (direct question)
  - IA [8] (yes/no natural questions)
- **Generators**
  - Phi4-14B [9]
  - Llama3.1-8B [10]
  - CommandR-7B [11]
  - Gemma2-2B [12]
- **Datasets**: NFCorpus, SCIDOCS, TREC-COVID

Query Efficiency (AUC)

- MEntA starts with high AUC and remains high
- IA starts with low AUC and gradually increases



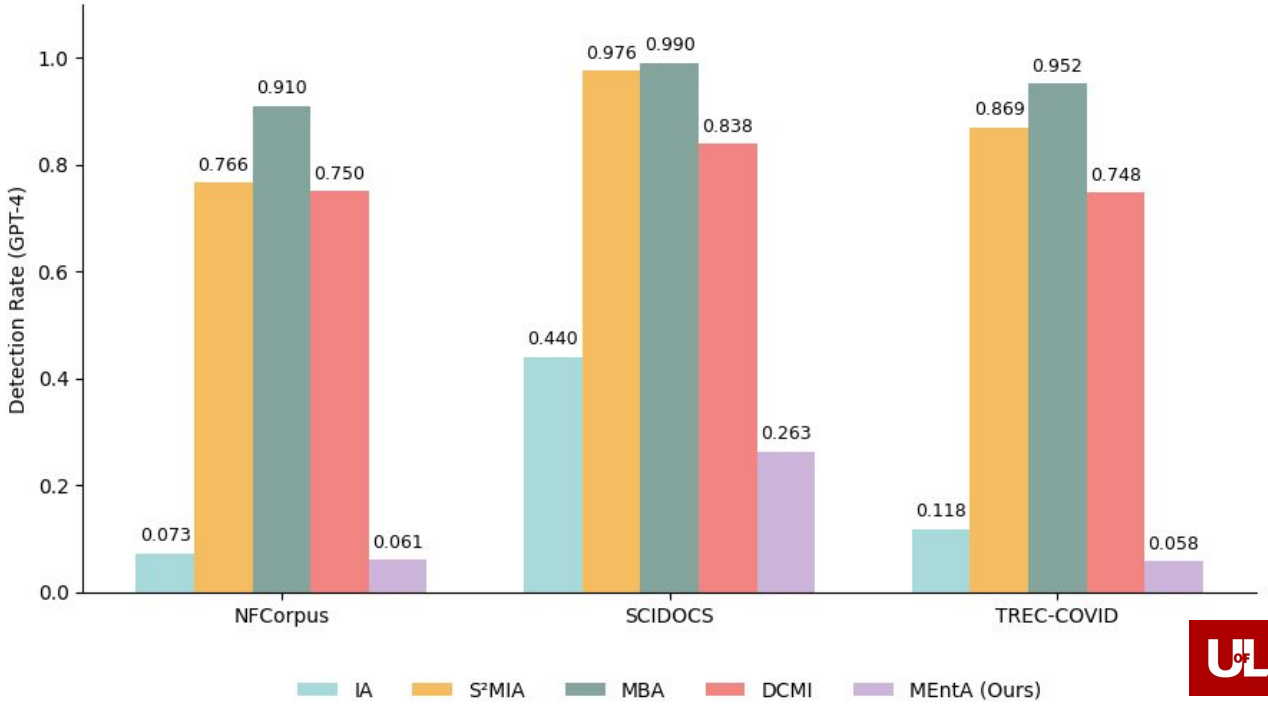
Cost Efficiency (USD)

- IA cost per attack
  - 30 outputs generated by Target Generator
  - 30 outputs generated by Shadow LLM (GPT-4o-mini)
- MEntA cost per attack
  - 5 outputs generated by Target Generator
  - 5 NLI inferences comparing 5 outputs and a target document

Generator	$/query		$/attack		$/attack ratio
	IA	MEntA	IA	MEntA	IA/MEntA
Phi4-14B	1.12e-4	**4.42e-5**	3.37e-3	**2.21e-4**	**15.24**
CommandR-7B	1.01e-4	**3.40e-5**	3.04e-3	**1.70e-4**	**17.87**
Llama3.1-8B	9.13e-5	**1.32e-5**	2.74e-3	**6.62e-5**	**41.37**
Gemma2-2B	8.56e-5	**7.89e-6**	2.57e-3	**3.95e-5**	**65.06**

Defense Robustness

Under LLM-based detection, MEntA is the most robust
(lowest detection rate among attack queries)



Defense Robustness

Under LLM-based detection, MEntA is the most robust
(lowest detection rate among attack queries)

Dataset	Attack	None	DP	Rerank	Prompt Inst.	Paraphrase
NFCorpus	S^2MIA [40]	0.711	0.655	0.669	0.850	0.883
	MBA [43]	0.874	0.803	0.876	0.824	0.805
	IA [47]	0.756	0.518	0.752	0.754	0.722
	DCMI [19]	0.828	0.828	0.832	0.773	0.843
	MEntA (Ours)	**0.989**	**0.913**	**0.975**	**0.984**	**0.977**
SCIDOCS	S^2MIA [40]	0.715	0.671	0.655	0.849	0.874
	MBA [43]	0.891	0.790	0.855	0.876	0.791
	IA [47]	0.915	0.543	0.937	0.901	0.880
	DCMI [19]	0.941	**0.943**	0.942	0.862	0.910
	MEntA (Ours)	**0.991**	0.916	**0.986**	**0.992**	**0.953**
TREC-COVID	S^2MIA [40]	0.604	0.586	0.558	0.699	0.731
	MBA [43]	0.781	0.727	0.781	0.743	0.728
	IA [47]	0.816	0.487	0.820	0.820	0.779
	DCMI [19]	0.886	**0.889**	0.873	0.827	0.846
	MEntA (Ours)	**0.893**	0.836	**0.904**	**0.908**	**0.846**

Conclusion

Conclusion

- Existing attacks are often detectable, query-intensive, or expensive, limiting real-world practicality.
- MEntA provides a practical membership inference attack requiring only 5 natural queries and no surrogate model.
- MEntA improves:
  - Query efficiency — only 5 queries.
  - Cost efficiency — up to 65× cheaper.
  - Defense robustness — remains effective against existing defenses.
- **Takeaway**: RAG systems remain vulnerable to low-cost, low-query privacy attacks, highlighting the need for stronger privacy defenses.

References

1. Yuying Li, Gaoyang Liu, Chen Wang, and Yang Yang. Generating is believing: Membership inference attacks against retrieval-augmented generation. In ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5, 2025
2. Feng, K., Zhang, G., Tian, H., Xu, H., Zhang, Y., Zhu, T., Ding, M., & Liu, B. (2025). RAGLeak: Membership inference attacks on RAG-based large language models. In W. Susilo & J. Pieprzyk (Eds.), Information Security and Privacy (pp. 147–166). Springer Nature Singapore.
3. Zhenting Qi, Hanlin Zhang, Eric Xing, Sham Kakade, and Himabindu Lakkaraju. Follow my instruction and spill the beans: Scalable data extraction from retrieval- augmented generation systems, 2024.
4. Zeng, S., Zhang, J., He, P., Liu, Y., Xing, Y., Xu, H., Ren, J., Chang, Y., Wang, S., Yin, D., & Tang, J. (2024). The good and the bad: Exploring privacy issues in retrieval-augmented generation (RAG). In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), Findings of the Association for Computational Linguistics: ACL 2024 (pp. 4505–4524). Association for Computational Linguistics.
5. Mingrui Liu, Sixiao Zhang, and Cheng Long. Mask-based membership inference attacks for retrieval-augmented generation. In Proceedings of the ACM on Web Conference 2025, WWW '25, page 2894–2907, New York, NY, USA, 2025. Association for Computing Machinery.
6. Anderson, M., Amit, G., & Goldsteen, A. (2025). Is my data in your retrieval database? Membership inference attacks against retrieval augmented generation. In Proceedings of the 11th International Conference on Information Systems Security and Privacy (pp. 474–485). SCITEPRESS - Science and Technology Publications.
7. Xinyu Gao, Xiangtao Meng, Yingkai Dong, Zheng Li, and Shanqing Guo. Dcmi: A differential calibration membership inference attack against retrieval-augmented generation. In Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security, CCS '25, page 4184–4198, New York, NY, USA, 2025. Association for Computing Machinery.
8. Ali Naseh, Yuefeng Peng, Anshuman Suri, Harsh Chaudhari, Alina Oprea, and Amir Houmansadr. Riddle me this! stealthy membership inference for retrieval-augmented generation. In Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security, CCS '25, page 1245–1259, New York, NY, USA, 2025. Association for Computing Machinery.
9. Marah Abdin et al. Phi-4 technical report, 2024.
10. Aaron Grattafiori, Abhimanyu Dubey, et al. The llama 3 herd of models, 2024.
11. Team Cohere, :, Aakanksha, Arash Ahmadian, et al. Command a: An enterprise-ready large language model, 2025.
12. Gemma Team, Morgane Riviere, et al. Gemma 2: Improving open language models at a practical size, 2024.
13. N. Jain et al., "Baseline Defenses for Adversarial Attacks Against Aligned Language Models," arXiv, 2023.
14. R. Wen, Z. Li, M. Backes, and Y. Zhang, "Membership Inference Attacks Against In-Context Learning," presented at the CCS 2024, Salt Lake City, UT, USA, 2024. [Online]. Available: https://doi.org/10.1145/3658644.3690306.
15. Y. Choi, Y. Park, J. Byun, J. Lee, and J. Park, "Safeguarding Privacy of Retrieval Data against Membership Inference Attacks: Is This Query Too Close to Home?," Suzhou, China, November 2025: Association for Computational Linguistics, in EMNLP 2025, pp. 8241–8258, doi: 10.18653/v1/2025.findings-emnlp.438. [Online]. Available: https://aclanthology.org/2025.findings-emnlp.438/
16. N. Grislain, "RAG with Differential Privacy," arXiv, 2025.
17. M. Liu, S. Zhang, and C. Long, "Mask-based Membership Inference Attacks for Retrieval-Augmented Generation," presented at the WWW 2025, Sydney NSW, Australia, 2025. [Online]. Available: https://doi.org/10.1145/3696410.3714771.













Thank you for your attention!

Nguyen Linh Bao Nguyen

PhD Student | Swinburne University of Technology

bnguyen1@swin.edu.au



