RAG with Differential Privacy

Nicolas Grislain

December 5, 2024

Abstract

Retrieval-Augmented Generation (RAG) has emerged as the dominant technique to provide *Large Language Models* (LLM) with fresh and relevant context, mitigating the risk of hallucinations and improving the ovarall quality of responses in environments with fast moving knoledge bases. However, the integration of external documents into the generation process raises significant privacy concerns. Indeed, when added to a prompt, it is not possible to guarantee a reponse will not inadvertently expose confidential data, leading to potential breaches of privacy and ethical dilemmas. This paper explores a practical solution to this problem suitable to general knowledge extraction from personnal data.

Introduction

Retrieval-Augmented Generation (RAG) has become a leading approach to enhance the capabilities of Large Language Models (LLMs) by supplying them with up-to-date and pertinent information. This method is particularly valuable in environments where knowledge bases are rapidly evolving, such as news websites, social media platforms, or scientific research databases. By integrating fresh context, RAG helps mitigate the risk of "hallucinations"—instances where the model generates plausible but factually incorrect information—and significantly improves the overall quality and relevance of the responses generated by the LLM.

However, incorporating external documents into the generation process introduces substantial privacy concerns. When these documents are included in the input prompt for the LLM, there is no foolproof way to ensure that the generated response will not accidentally reveal sensitive or confidential data. This potential for inadvertent data exposure can lead to serious breaches of privacy and presents significant ethical challenges. For instance, if an LLM is used in a healthcare setting and it accidentally includes patient information from an external document in its response, it could violate patient confidentiality and legal regulations.

This paper describes a practical solution aimed at addressing these privacy concerns with *Differential Privacy* (DP). The solution is based on two pillars:

- A method to collect documents related to the question in a way that does not prevent its output to be used in a DP mechanism.
- A method to use the collected documents to prompt a LLM and produce a reponse with DP guarantees.

Related Work

A straightforward approach to *add* knowledge to an existing LLM is to continue its training with the new knowledge or *Fine Tune* (FT) it. In the case of private data, this raises challenges in the case of private data, as LLMs *memorize training data* (see (Shokri et al. 2017) or (Carlini et al. 2021)).

To mitigate this privacy risk, it is possible to *redact* sensitive content prior to the FT process, but this operation is not so reliable and requires judgement on what should be redacted. This is a difficult manual operation based on the perceived sensitivity of each field, how it can be used to re-identify an individual, especially once crossed with other publicly available data. Overall it is very easy to get wrong, leaning on the prudence side yield useless data, try to optimize utility and you may miss sensitive information.

A solution to this problem is to leverage *Differential Privacy*, a theoretical framework enabling the computation of aggregates with formal privacy garantees (See (Dwork, Roth, et al. 2014)).

The current approaches to Private LLM

A reference (Abadi et al. 2016)

(Yue et al. 2023)

Private RAG

Some solutions are based on privacy preserving synthetic data generation: (Zeng et al. 2024)

(Ponomareva et al. 2023)

(Lebensold et al. 2024)

(Lin et al. 2024)

(Xie et al. 2024)

(Tang et al. 2024)

(Wu et al. 2023)

(Hong et al. 2024)

DP-RAG

Overview

Privacy Unit Preserving Document Retrieval

Differentially Private In-Context Learning

Evaluation

Conclusion

Abadi, Martin, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. "Deep Learning with Differential Privacy." In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. CCS'16. ACM. https://doi.org/10.1145/2976749.2978318.

Carlini, Nicholas, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, et al. 2021. "Extracting Training Data from Large Language Models." In *30th USENIX Security Symposium (USENIX Security 21)*, 2633–50. USENIX Association. https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

Dwork, Cynthia, Aaron Roth, et al. 2014. "The Algorithmic Foundations of Differential Privacy." *Foundations and Trends® in Theoretical Computer Science* 9 (3–4): 211–407.

Hong, Junyuan, Jiachen T. Wang, Chenhui Zhang, Zhangheng Li, Bo Li, and Zhangyang Wang. 2024. "DP-OPT: Make Large Language Model Your Privacy-Preserving Prompt Engineer." https://arxiv.org/abs/2312.03724.

Lebensold, Jonathan, Maziar Sanjabi, Pietro Astolfi, Adriana Romero-Soriano, Kamalika Chaudhuri, Mike Rabbat, and Chuan Guo. 2024. "DP-RDM: Adapting Diffusion Models to Private Domains Without Fine-Tuning." https://arxiv.org/abs/2403.14421.

Lin, Zinan, Sivakanth Gopi, Janardhan Kulkarni, Harsha Nori, and Sergey Yekhanin. 2024. "Differentially Private Synthetic Data via Foundation Model APIs 1: Images." https://arxiv.org/abs/2305.15560.

Ponomareva, Natalia, Hussein Hazimeh, Alex Kurakin, Zheng Xu, Carson Denison, H. Brendan McMahan, Sergei Vassilvitskii, Steve Chien, and Abhradeep Guha Thakurta. 2023. "How to DP-Fy ML: A Practical Guide to Machine Learning with Differential Privacy." *Journal of Artificial Intelligence Research* 77 (July): 1113–1201. https://doi.org/10.1613/jair.1.14649.

Shokri, Reza, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. "Membership Inference Attacks Against Machine Learning Models." In *2017 IEEE Symposium on Security and Privacy (SP)*, 3–18. https://doi.org/10.1109/SP.2017.41.

Tang, Xinyu, Richard Shin, Huseyin A. Inan, Andre Manoel, Fatemehsadat Mireshghallah, Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, and Robert Sim. 2024. "Privacy-Preserving in-Context Learning with Differentially Private

Few-Shot Generation." https://arxiv.org/abs/2309.11765.

Wu, Tong, Ashwinee Panda, Jiachen T. Wang, and Prateek Mittal. 2023. "Privacy-Preserving in-Context Learning for Large Language Models." https://arxiv.org/abs/2305.01639.

Xie, Chulin, Zinan Lin, Arturs Backurs, Sivakanth Gopi, Da Yu, Huseyin A Inan, Harsha Nori, et al. 2024. "Differentially Private Synthetic Data via Foundation Model APIs 2: Text." https://arxiv.org/abs/2403.01749.

Yue, Xiang, Huseyin A. Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Hoda Shajari, Huan Sun, David Levitan, and Robert Sim. 2023. "Synthetic Text Generation with Differential Privacy: A Simple and Practical Recipe." https://arxiv.org/abs/2210.14348.

Zeng, Shenglai, Jiankun Zhang, Pengfei He, Jie Ren, Tianqi Zheng, Hanqing Lu, Han Xu, Hui Liu, Yue Xing, and Jiliang Tang. 2024. "Mitigating the Privacy Issues in Retrieval-Augmented Generation (RAG) via Pure Synthetic Data." https://arxiv.org/abs/2406.14773.