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. It shows *differentially private token generation* is a viable approach to private RAG.

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 (Qi et al. 2024). 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 (DP-RAG) 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.

The paper describes also some empirical tests and shows that *DP-RAG* is most effective in context where enough documents give elements of response.

Related Work

In general there are 2 families of approaches to add new knowledge to an LLM. The first *Fine Tunning* (FT) and the other is *Retrieval Augmented Generation* (RAG). In both these approaches, adding privacy can be done, through simple heuristics with human validation such as *masking* or using a systematic and principle-based approach such as *Differential Privacy*.

Private Fine-Tuning

A straightforward approach to adding knowledge to an existing LLM is to continue its training with the new knowledge or to Fine Tune (FT) it. However, this raises challenges when dealing with private data, as LLMs tend to 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 (aka. *masking*), but this operation is not very reliable and requires judgment on what should be redacted. This is a difficult manual operation based on the perceived sensitivity of each field and how it can be used to re-identify an individual, especially when combined with other publicly available data. Overall, it is very easy to get wrong; leaning too much on the side of prudence can yield useless data, while trying to optimize utility may result in leaking 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 most common approache to Private LLM FT is to use Differentially-Private-Stochastic-Gradient-Descent (DP-SGD, see (Abadi et al. 2016) and (Ponomareva et al. 2023)). DP-SGD is about clipping gradients and adding them some noise while running your ordinary SGD (or standard variants such as *Adam*, etc.). This method requires the data to be organized per *privacy unit* (typically a privacy unit will be a user). Every training example should belong to one and only one privacy unit[1].

But, when new documents are frequently added to the private knowledge base, or when data is scarcen, FT may not be the best approach.

[1]Note that observations (examples) can be grouped into composite observations if one user contributes to many observations.

Private RAG

When FT is not the best approach to adding new knowledge, DP-FT cannot help with privacy. In these cases, DP can still be levereaged in different ways. A straightforward approach to DP RAG is to generate synthetic documents with differential privacy out of the private document pool and retrieve documents from tha synthetic pool instead of the private one. Another approach is to generate the LLM response in a DP way.

The approach of generating synthetic documents usable for RAG has been explored by (Zeng et al. 2024) but without DP garantees. There are three main approaches to the problem of generating DP Synthetic Data (SD):

- Fine-Tuning a pretrained generative model with DP to generate synthetic documents.
- Use some form of automated prompt tuning to generate synthetic prompts or context documents.
- And use DP aggregated generation.

Fine-Tuning a pretrained generative model with DP can be done with DP-SGD ((Abadi et al. 2016) and (Ponomareva et al. 2023)) as mentionned above. An application to synthetic text generation is described there: (Yue et al. 2023). This method is technically complex, as, DP-SGD can be challenging to implement efficiently (Bu et al. 2023).

In (Hong et al. 2024), the authors use an automated prompt tuning technique developped in (Sordoni et al. 2023) and (Zhou et al. 2023) and make it differentially private. From the evaluations presented, it seems to compare favorably to DP-FT synthetic data approaches. Similar methods, based on DP-automated prompt tuning are exposed in (Lin et al. 2024) for images and (Xie et al. 2024) for text.

A last approach to generating synthetic data is based on DP aggregation of data.(Lebensold et al. 2024) or (Wu et al. 2023) show how to aggregate images or text in their embedding space (aka. Embedding Space Aggregation). Aggregating data privately is also the approach of (Tang et al. 2024), but they do it at the token level.

This last method greatly inspired the approach described in this document, though not for SD, but to directly generate RAG output from private documents.

DP-RAG

Overview

DP-RAG is made of 2 main components:

- 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.

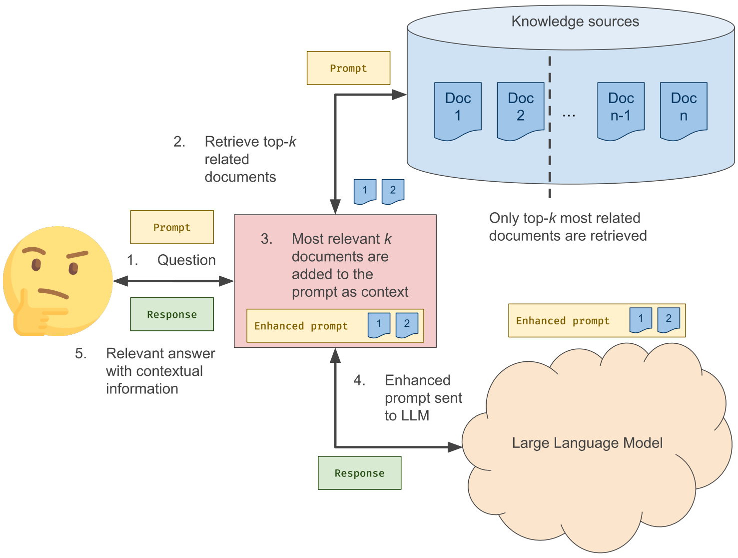

Figure 1: A broad picture of how RAG works

To understand the need for these, let's describe what RAG is usually made of, and the assumption we make for its private variant (DP-RAG).

A LLM $\mathcal{L}$ is a function, taking some text, in thr form of a sequence of tokens: $x = \langle x_0, x_1, \dots, x_{n-1} \rangle$ as input and outputing a distribution of x_n conditional on x:

$$\mathcal{L}()$$

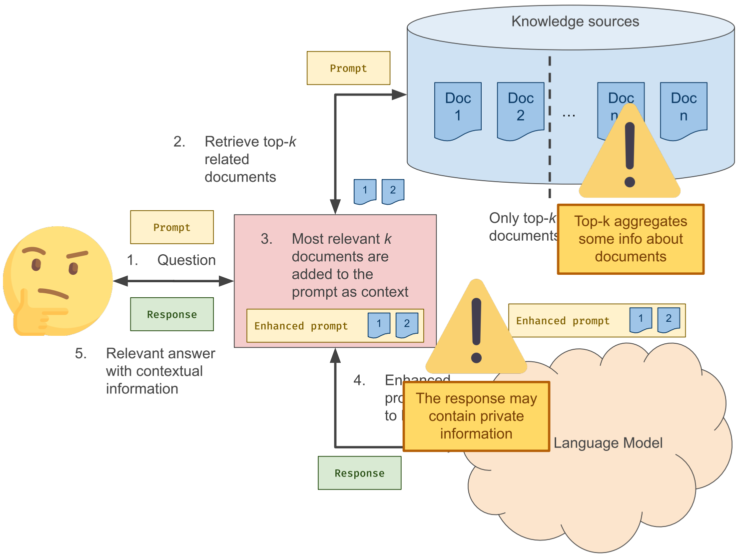

Figure 2: A broad picture of how RAG works

Privacy Unit Preserving Document Retrieval

Differentially Private In-Context Learning

(Durfee and Rogers 2019) could be used for token selection

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.

Bu, Zhiqi, Yu-Xiang Wang, Sheng Zha, and George Karypis. 2023. "Differentially Private Optimization on Large Model at Small Cost." In *International Conference on Machine Learning*, 3192–3218. PMLR.

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.

Durfee, David, and Ryan Rogers. 2019. "Practical Differentially Private Top-k Selection with Pay-What-You-Get Composition." https://arxiv.org/abs/1905.04273.

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.

Qi, Zhenting, Hanlin Zhang, Eric Xing, Sham Kakade, and Himabindu Lakkaraju. 2024. "Follow My Instruction and Spill the Beans: Scalable Data Extraction from Retrieval-Augmented Generation Systems." https://arxiv.org/abs/2402.1

7840.

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.

Sordoni, Alessandro, Xingdi Yuan, Marc-Alexandre Côté, Matheus Pereira, Adam Trischler, Ziang Xiao, Arian Hosseini, Friederike Niedtner, and Nicolas Le Roux. 2023. "Joint Prompt Optimization of Stacked LLMs Using Variational Inference." https://arxiv.org/abs/2306.12509.

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.

Zhou, Yongchao, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. "Large Language Models Are Human-Level Prompt Engineers." https://arxiv.org/abs/2211.01910.