# Ingest-And-Ground: Dispelling Hallucinations from Continually-Pretrained LLMs with RAG

Chenhao Fang*, Derek Larson*, Shitong Zhu*, Sophie Zeng*, Wendy Summer*, Yanqing Peng*, Yuriy Hulovatyy*, Rajeev Rao, Gabriel Forgues, Arya Pudota, Alex Goncalves, Hervé Robert
{chenhaofang,dereklarson,shitong,sophiezeng,wsummer,yanqingpeng,jura,rrrao,gforgues,arpu,alexgon,hervert}@meta.com
Meta
Menlo Park, California, USA

## Abstract

This paper presents new methods that have the potential to improve privacy process efficiency with LLM and RAG. To reduce hallucination, we continually pre-train the base LLM model with a privacy-specific knowledge base and then augment it with a semantic RAG layer. Our evaluations demonstrate that this approach enhances the model performance (as much as doubled metrics compared to out-of-box LLM) in handling privacy-related queries, by grounding responses with factual information which reduces inaccuracies.

## 1 Introduction

Robust privacy compliance [1] is achieved through a collaborative privacy review process for projects involving engineers and privacy regulation experts [2]. Although this method is effective, it is also time-consuming and labor-intensive. More specifically, engineers may possess a diverse range of knowledge regarding these privacy regulations, leading to multiple project revisions to ensure compliance. Meanwhile, reviewers must also grasp the technical details, resulting in numerous queries to thoroughly comprehend the projects.

**Goals**. Our mission is to develop tools that assist both parties in this process. However, privacy regulations are too complex for traditional rule-based or AI techniques, since they require interpretation across human language, code, and both structured and unstructured data. The advent of Large Language Models (LLMs) offers a promising solution, since they are adept at understanding different types of inputs [3], can perform logical reasoning and handle complex tasks [4], and are interactive which allow them to provide explanations for their assessments. An LLM equipped with internal privacy knowledge could help engineers and privacy moderators find answers to privacy questions more efficiently, greatly streamlining privacy reviews.

**Challenges**. The concept of "privacy risk" is dynamic and even the most advanced model can struggle with the specialized and ever-changing nature of privacy regulations and enterprise-specific data structures, leading to errors and irrelevant responses [5]. This phenomenon, known as "hallucination" [6], can significantly undermine the usability of the model. For example, in assessing data storage regulations, a LLM model experiencing hallucinations might erroneously assert compliance based on non-existent regulations or a misinterpreted data storage component. Unlike other common issues in LLM responses such as logical errors and knowledge gaps, hallucinations are particularly problematic in our context as they appear correct at first glance and are tricky for engineers and privacy reviewers to immediately recognize inaccuracies in the responses. This results in wasted time and resources as both parties need to verify the correctness of generated response.

**Solution**. To address these challenges, Retrieval-Augmented Generation (RAG) offers a promising solution [7]. RAG systems integrate information retrieval with text generation, enabling them to dynamically access the most current knowledge bases and retrieve relevant context data in real-time [8]. This approach ensures that the responses generated are not only contextually appropriate but also reflect the latest standards, while providing citations for their conclusions. In practice, effective RAG systems have been proven to significantly reduce the risk of hallucinations in different use cases [9], and we explore a similar solution to build a system called PrivacyBrain to improve the LLM-generated answers for privacy compliance assessment.

## 2 Methodology

For this reserach, we choose to continually pre-train `Llama-3.1` [10] to ingest the privacy knowledge base for enhancing the base pre-trained model on fronts of understanding and reasoning for various privacy-related tasks (e.g., identifying regulatory risks for a new product launch). While the original Llama3.1 family showed their state-of-the-art performance, including factuality assessments which measures how vulnerable the models are prone to hallucination issues [10], the additional round of pre-training can elevate them again, as the new training updates model weights without calibrating the model through any alignment process (e.g., RLHF [11]). Given so, we aim to ground hallucinations with a proper RAG system that provides factual information to the LLM, by retrieving it from a knowledge base.

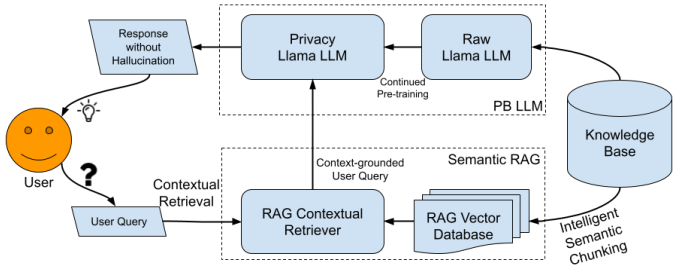


**Figure 1: System overview of PrivacyBrain: how we factually ground LLM hallucinations with RAG**

As depicted in Figure 1, PrivacyBrain includes the following main components/steps:

*Authors contributed equally

**Privacy knowledge base**. First, we need to compose the dataset that covers our desirable privacy-related documentations, to serve as the data source for both performing additional model training and providing document pieces to RAG. We follow guidance from privacy experts to construct a knowledge base of about 20,000 documents (which translates to roughly 2 million tokens) that encapsulates privacy policies and processes, as well as privacy laws and regulations for various countries and regions.

**LLM continual pre-training**. We use `Llama3.1-70b-instruct` as the base model, and perform continual pre-training over it using the collected documents in the knowledge base. We follow the general pre-training paradigm reported in [10], which utilizes Causal Token Masking as the training task.

**RAG with semantic chunking**. To reduce hallucinations, we further use the constructed knowledge base to build an intelligent RAG layer on top of the enhanced LLM. When presented with a query prompt, the RAG layer retrieves relevant documents or passages from the knowledge base and then uses both the original prompt and the retrieved information to generate a coherent and contextually enriched response. One key step of indexing documents for RAG systems is how to perform chunking across long documents, as it directly impacts the effectiveness of retrievers to locate the most relevant document pieces. Inappropriate segmentation strategies can result in chunks that either lack sufficient context or contain too much irrelevant information, thereby impairing the performance of retrieval models [12].

Traditional chunking methods, such as segmenting by sentences or paragraphs [13], typically generate snippets of uniform or similar sizes. However, they often fall short in considering text semantics, leading to sub-optimal retrieval performance. More sophisticated techniques include recursive character splitting [14], which segments texts based on a hierarchy of delimiters like paragraph breaks and spaces, respect documents' intrinsic structure better, but still compromises contextual richness. More recent algorithms are semantic-based splitting (e,g., [15]), which uses text embeddings to group text segments with similar meanings. These approaches segments documents at "semantically-natural" breaks such as sentence endings, ensuring that each chunk is contextually coherent and semantically connected. In this work, we implemented our semantic RAG module based on Meta's state-of-the-art text embedding model, Dragon-Plus [16].

**Online context grounding**. Now, with both the enhanced LLM and semantic RAG module ready, we need to put them into action in the online environment. Specifically, we feed example queries related to privacy first to the RAG, which retrieves fact-grounded and relevant document chunks as additional context to enhance the original queries, and then pass them to LLM. With the additional context that grounds the factuality of queries provided, LLM now stands much lower odds to hallucinate, as will be shown in evaluations.

## 3 Evaluations

**Data**. We construct a privacy understanding benchmark that consists of 50 example queries (e.g., "What are principles for protecting user data in general, for a social network company?"), which covers a wide range of common questions about privacy(e.g., privacy processes, privacy laws/regulations). We manually construct ground truth answer text for each question according to references, as well as a short list of keywords that capture the key points from the answer text.

**Quality measurement & metric**. Assessing quality of LLM systems, especially based on Q&A interactions, has been gaining increasing popularity, as it is intrinsically challenging. We tackle this by combining two popular and widely-adopted mechanisms: (i) *LLM-as-a-Judge rating* [17], which invokes a potent third-party LLM (GPT-4 in our case) to rate each answer generated by PrivacyBrain, with respect to the compiled ground truth. The metric for this measurement is *pass rate*. (ii) *Keyword matching*, which relies on the manually-extracted short list of key points to determine how many of them are present in the LLM response. We measure this metric by $\frac{|matched\ keywords|}{|total\ keywords|}$. We believe this combination offers a reasonable approximation to high-quality human inspection, which is otherwise costly to carry out.

**Baselines**. We choose three baselines to evaluate the performance PrivacyBrain (or PB in short) in relation to its variants, to showcase improvements by incorporating different components: (i) raw Llama3.1 with continual pre-training (PB LLM in short); (ii) raw Llama3.1 with semantic RAG; and (iii) raw Llama3.1.

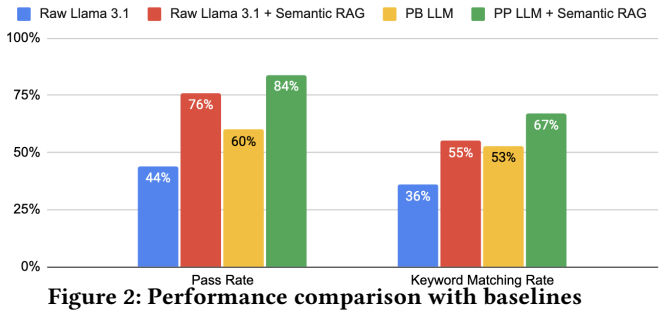


**Figure 2: Performance comparison with baselines**

**Results**. Figure 2 highlights key observations based on our evaluation set: (i) continual pre-training significantly (+16%) helps Llama LLM ingest domain knowledge that it was not initially trained with, resulting in higher rates passing both GPT4 and keyword-based quality measurements; (ii) semantic RAG visibly tames hallucinations that cause factual errors generated from the PrivacyBrain LLM, which translates to further improvements (+24% vs. PB LLM only, and 40% vs. raw Llama) over metrics in measuring the correctness of LLM responses. Collectively, we conclude that the combination of the knowledge injection via additional LLM pre-training, and the contextual RAG layer (note the additional knowledge powering these two enhancements is from the same source) strengthens the overall performance of PrivacyBrain.

## 4 Conclusion

Our preliminary results demonstrate the potential of LLM models with RAG systems in enhancing privacy compliance in enterprise settings. Future work will focus on expanding the data sources, improving the retrieval mechanisms, and further fine-tuning the system for specific sub-domains within privacy laws/regulations.

## References


[1] Meta. Protecting privacy and security. https://about.meta.com/actions/protecting-privacy-and-security/, 2024.

[2] Meta. Privacy Progress. https://about.meta.com/actions/protecting-privacy-and-security/privacy-progress/, 2024.

[3] Daye Nam, Andrew Macvean, Vincent Hellendoorn, Bogdan Vasilescu, and Brad Myers. Using an llm to help with code understanding. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, pages 1–13, 2024.

[4] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.

[5] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[6] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, 2023.

[7] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation, 2021.

[8] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[9] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*, 2021.

[10] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[11] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[12] Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context, 2023.

[13] LangChain. Split by character. https://python.langchain.com/v0.1/docs/modules/data_connection/document_transformers/character_text_splitter/, 2024.

[14] LangChain. Recursively split by character. https://python.langchain.com/v0.1/docs/modules/data_connection/document_transformers/recursive_text_splitter/, 2024.

[15] Greg Kamradt. 5 levels of text splitting. https://github.com/FullStackRetrieval-com/RetrievalTutorials/blob/main/tutorials/LevelsOfTextSplitting/5_Levels_Of_Text_Splitting.ipynb, 2024.

[16] Sheng-Chieh Lin, Akari Asai, Minghan Li, Barlas Oguz, Jimmy Lin, Yashar Mehdad, Wen-tau Yih, and Xilun Chen. How to train your dragon: Diverse augmentation towards generalizable dense retrieval. *arXiv preprint arXiv:2302.07452*, 2023.

[17] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.