Enhancing Personal Knowledge Management by Mitigating LLM Hallucinations with Retrieval-Augmented Generation

Ge Ye Ong1, a) , Mohanan Vasuky1, b) and S Prabha Kumaresan2, 3, c)

1School of Computing, INTI International College Penang, 1-Z, Lebuh Bukit Jambul, Bukit Jambul, 11900, Bayan Lepas, Pulau Pinang, Malaysia

2Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Malaysia.

3Center of Natural Language Processing (NLP), CoE for Artificial Intelligence, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Malaysia.

*c) Corresponding author: prabha.kumaresan@mmu.edu.my*

*a)marcuso1710@gmail.com  
b)vasuky.mohanan@newinti.edu.my*

**Abstract**. With the rise of generative AI, Large Language Models (LLMs) like ChatGPT have become powerful tools for knowledge acquisition, offering vast parametric knowledge in a conversational manner. However, they are prone to generating inaccurate or nonsensical content, known as hallucinations, which pose risks to knowledge integrity, especially in personal knowledge management systems where LLMs are integrated for content generation and summarization. To address this, this work explores a Retrieval-Augmented Generation (RAG) pipeline which incorporates semantic chunking and vector indexing with Snowflake's arctic-embed-m, mxbai-embed-large-v1 for reranking, and a Llama-3-8B core generation model, as a solution to mitigate hallucinations and improve LLM reliability in personal knowledge bases. Using an Iterative and Incremental Development (IID) methodology, the approach was systematically refined across multiple cycles, with evaluations conducted on NLP benchmarks such as MS MARCO and SQuAD to compare RAG with traditional LLMs in both open-book and closed-book settings. The RAG implementation achieved 87% factual accuracy, representing a 34–50% improvement over leading LLMs, while in open-domain QA, it attained a precision of 38.6%, recall of 69.6%, and an F1 score of 44.2%, significantly reducing hallucinations. These findings establish RAG as a viable approach for enhancing personal knowledge management, ensuring more accurate knowledge organization and retrieval while mitigating hallucination risks in AI driven systems.

# Introduction

The arrival of Large Language Models represents a significant leap in artificial intelligence, enabling machines to comprehend and generate human language with unprecedented fluency [1]. LLM applications like ChatGPT rapidly gained widespread adoption demonstrating proficiency in various Natural Language Processing (NLP) tasks [1], particularly excelling in question-answering (QA) [2]. Consequently, individuals are increasingly relying on LLMs for information retrieval and knowledge acquisition [3], positioning them as the future of Personal Knowledge Management Systems (PKMS). Moreover, the usage of LLMs has extended into critical fields such as healthcare [4], [5], where these high-stakes applications would require rigorous scrutiny and careful consideration.

Despite their capabilities, the reliability of LLMs is fundamentally constrained by their propensity for “hallucinations”, the generation of plausible yet factually incorrect, nonsensical, or input-deviating content [6]. Hallucinations can manifest in various forms, including contradictions with factual world knowledge, the input query, or even the model's own prior statements [6]. This intrinsic unreliability severely limits their applicability in PKMS and other domains demanding high factual integrity [7], [8].

This paper evaluates the effectiveness of a RAG system incorporating semantic chunking and reranking techniques in improving LLM factual accuracy for open-domain QA tasks to reduce occurrences of hallucinations. We quantify the reduction in hallucinations compared to several standalone LLMs using a curated dataset.

# Related Works

Transformer-based LLMs [9] have advanced NLP significantly, yet hallucination persists as a key limitation across various forms (fact-conflicting, input-conflicting, context-conflicting) [6]. There are multiple obstacles that hinder the understanding and studying of hallucinations: vast training data, versatility of LLMs, and imperceptibility of errors [6]. Due to the large amount of data needed to train LLMs, cleaning and verifying initial training data can be difficult. LLMs are also expected to perform well in many different scenarios [6], causing difficulties in mitigating hallucinations, where sometimes hallucinations pose as an advantage instead, for instance, in creative writing. The hallucinations generated by LLMs can seem so realistic to the point that it is difficult for models and even humans to detect and evaluate them.

Retrieval-Augmented Generation or RAG is a framework to enhance LLMs by incorporating chunks of information from an external knowledge base by retrieving them semantically [10]. The earliest known and most foundational RAG architecture, also known as the Naïve RAG framework involves indexing, retrieval (based on semantic similarity), and generation using the retrieved context [10]. Additional enhancements often include optimized chunking, query transformations, and reranking mechanisms which define an Advanced RAG architecture, which incorporates pre- and post-retrieval strategies [10]. Various strategies exist to optimize RAG performance. Chunking methods range from simple fixed-size splits to more sophisticated semantic or sentence-aware approaches. Pre-retrieval query transformations and post-retrieval reranking using dedicated models can further enhance the relevance of context provided to the LLM [10].

Alternative approaches to enhance LLM factuality include fine-tuning on domain-specific data [11] or leveraging models with extremely long context windows . However, fine-tuning does not easily incorporate rapidly changing information and requires significant computational resources for updates [12],while long-context models face challenges with computational cost during inference, potential performance degradation over the full context length [10] and lack the transparent explicit grounding evidence provided by RAG's retrieval step. RAG offers a flexible and computationally viable approach for improving factual consistency, particularly for knowledge-intensive tasks, and can be combined with fine-tuning for further specialization [10].

# Methodology

This study employs an experimental approach combining Iterative and Incremental Development (IID) with Evaluation-Driven Development (EDD) to build and refine RAG pipelines. EDD provides measurable progress baselines, crucial for evaluating LLM systems where output quality determines the primary success metric. IID allows flexibility to adapt components (embedding models, chunking strategies, LLMs) based on evaluation outcomes.

## ****Dataset Curation****

A curated QA dataset was created by combining and processing questions from two established benchmarks: MS MARCO [13] and SQuAD [14]. These datasets provide diverse, open-domain questions with associated contexts and answers. After preprocessing, filtering unanswerable questions, and ensuring diversity, a final set of 128 question-answer pairs was randomly sampled (64 from each source) for evaluation. The contexts associated with these pairs formed the external knowledge base for the RAG systems.

## Evaluation Strategy

This project will employ human-based evaluation, a methodology also utilised by hallucination detection papers like RAGTruth and HaluEval. Some authors believe that human evaluation remains as a robust and trustable method of hallucination evaluation. To isolate the effects of retrieval-augmented generation (RAG), we will compare responses from the RAG-enhanced LLM against those from a standalone LLM under identical conditions. Given practical limitations in manpower and budget, we evaluate a smaller sample size of 128 question/answer pairs. To maintain credibility while ensuring overall feasibility, this evaluation threshold has been previously justified in prior work [15].

## System Architectures

Three main system configurations were evaluated in this study, a non-RAG baseline, a Naive RAG system, and an Advanced RAG system (the proposed architecture). The non-RAG baseline consists of a standalone Meta Llama-3-8B-8192 model, accessed via a custom GroqCloud LlamaIndex wrapper with a temperature of 0.1 and fully utilizing its 8192-token context window. This setup establishes a baseline performance without any retrieval augmentation. A Naïve RAG pipeline was implemented using the LlamaIndex framework and compared against standalone LLMs. Input documents were processed using sentence splitting, which chunks the information by sentence delimiters after a certain token size is reached. The embedding model, Snowflake's arctic-embed-m, selected based on MTEB retrieval benchmarks will be used to index the contexts from SQuAD and MS MARCO. These embeddings were stored and queried using ChromaDB, an open-source vector database. Lastly, The core of the pipeline was the same Meta's Llama-3-8B-8192 used in the non-RAG baseline. Lastly, to improve the RAG performance further [16], a generic RAG prompt template guided the LLM to synthesize answers based on provided context. Finally, the Advanced RAG (proposed architecture in Figure 1) has a few key differences, instead of basic sentence splitting, input documents for this were processed using semantic chunking. This method leverages the Snowflake's arctic-embed-m embedding model itself for context-aware segmentation, aiming to create more semantically coherent and meaningful chunks prior to indexing. Another key addition was a post-retrieval reranking step. After the retrieval phase, Mixedbread's mxbai-embed-large-v1 model was employed as a reranker to improve context relevance by reordering the retrieved chunks, prioritizing the most pertinent information for the generation phase.

A screenshot of a computer

AI-generated content may be incorrect.

**FIGURE 1**. Architecture diagram of the proposed RAG implementation

# Results

The study finds a significant accuracy improvement with RAG. In Figure 2(a), both the Naïve RAG and the Advanced RAG (Proposed Architecture), show a substantial increase the number of "True" factual outputs compared to the non-RAG baseline, showing its effectiveness in this QA task.

In Figure 2(b), The Advanced RAG system achieved 87.5% accuracy, compared to 57.8% for the standalone Llama-3-8B baseline. This constitutes a 51.4% relative improvement and demonstrates effective hallucination mitigation by improving factual accuracy.

|  |  |
| --- | --- |
| A graph of different colored squares  Description automatically generated with medium confidence | A graph of accuracy of qa methods  Description automatically generated |
| (a) | (b) |

**FIGURE 2.** Human annotation results and accuracy score from QA task across experimented methods

Figure 3(a) presents a comparative analysis of the QA task performance between RAG architectures and conventional standalone LLMs, including both closed-source and open-source variants. Both RAG architectures significantly outperforms all tested standalone LLMs, achieving superior results compared to GPT-4o-mini (62.5%), Mixtral-8x7b (60.9%), and Gemma2-9B-it (53.9%), with a 38% relative improvement over the best performing base LLM, GPT4o-mini.

To mitigate potential biases introduced by human annotation, a token-level analysis was conducted, which has further reinforced the findings from human evaluations (refer Figure 3(b)). The RAG system consistently outperformed the baseline models across all evaluated metrics. Among conventional LLMs, Llama-3-8B achieved the highest F1-score, indicating a better balance between precision and recall compared to other standalone models (refer to Table 1).

While GPT-4o and Mixtral-8x7b exhibited strong recall metrics, although factually correct, their responses were often excessively verbose, leading to lower precision and, consequently, a reduced F1-score. This suggests that while these models retrieve relevant information effectively, their lack of conciseness negatively impacts overall performance.

|  |  |
| --- | --- |
| A graph of different colored rectangular objects  Description automatically generated | A graph of different colored bars  Description automatically generated |
| (a) | (b) |

**FIGURE 3.** Accuracy score and token-level metrics of RAG against standalone LLMs from QA task

# Discussion

**TABLE 1.** Comparison of QA method task results across all metrics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Method** | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
| Advanced RAG (Proposed Architecture) | **0.867** | **0.386** | **0.696** | **0.442** |
| Naïve RAG (SentenceSplitter + No Reranker) | 0.836 | 0.359 | 0.692 | 0.416 |
| Llama 3 8B (Baseline) | 0.578 | 0.239 | 0.396 | 0.266 |
| Gemma2-9B-it | 0.539 | 0.232 | 0.387 | 0.257 |
| GPT4o-mini | 0.625 | 0.130 | 0.438 | 0.182 |
| Mixtral-8x7b | 0.609 | 0.077 | 0.442 | 0.117 |

The integration of RAG has significantly enhanced the factual accuracy of the Llama-3-8B model, improving its performance beyond its baseline and that of other standalone LLMs of similar caliber on this open-domain QA task. The 28.9% absolute accuracy increase over the baseline highlights RAG's effectiveness in reducing hallucinations by grounding generation in retrieved context. This study finds that the proposed architecture is less likely to hallucinate, by improving factual accuracy.

The performance gain suggests the RAG enhancements such as semantic chunking likely provided more coherent contextual units than simpler methods, while the reranker effectively prioritized relevant information for the generation phase. The fact that the RAG-augmented Llama-3-8B outperformed models like GPT4o-mini and Mixtral-8x7b underscores the importance of retrieval mechanisms for knowledge-intensive QA, potentially outweighing moderate differences in base model scale or architecture for this task.

Besides the clear-cut improvements of RAG with factual accuracy. RAG also represents itself as an ethical AI framework. One of the pillars of AI Ethics is Explainability. RAG addresses this by being able to output the content retrieved into its prompt. Due to hallucinations, chatbots still require self-validation and can be difficult to use [17]. This approach improves the trustworthiness and transparency of AI, potentially increasing adoption.

Secondly, within high-fidelity domains like healthcare, Privacy is of **utmost** importance. Standard SaaS LLM solutions can risk exposing confidential HIPAA data to third parties [5]. RAG's flexibility allows for self-hosted LLMs and private knowledge bases, enabling secure deployment of applications. This approach protects sensitive information while delivering accurate, relevant answers, making it viable for current healthcare applications.

RAG's ethical focus does not end there, it also enables fairer and more inclusive AI by utilizing diverse knowledge bases relevant to specific communities. This allows for more tailored and contextually accurate AI interactions. However, deploying such tailored RAG systems effectively worldwide depends on overcoming fundamental language and communication issues, notably adapting processes like tokenization to ensure high performance across varied niche and complex language structures [18].

While our RAG implementation demonstrated superior performance, the IID process underscored that RAG systems can be “fragile,” with their effectiveness highly dependent on the careful selection of their core components. Firstly, the quality and relevance of the input knowledge base are foundational. This material forms the factual grounds for the entire pipeline, which is why our study prioritized credible sources from established academic benchmarks like MS MARCO and SQUAD. Secondly, the effectiveness of the embedding model is equally essential. As discussed, the choice of Snowflake's arctic-embed-m for our architecture was deliberate, guided by its strong performance on the MTEB Retrieval leaderboard. This indicates its capability to capture semantic similarity effectively, increasing the probability of retrieving the correct contextual chunks essential for factual grounding. Lastly, the capabilities of the core LLM significantly influence outcomes. Our QA tasks revealed that different LLMs produced varied results, clearly demonstrating that the chosen LLM plays a major role in determining the final quality and factual consistency of the RAG system's output.

Challenges in evaluating generative outputs persist within the field. Human annotation provides high quality evaluations but are costly, while token-based metrics provide objectivity but can be sensitive to verbose answers which are likely from LLM generations. By combining both approaches, this study achieves credible results. Study limitations include the dataset scale (128 pairs) and restriction to single-hop questions.

# Conclusion

This paper proves the successful implementation and evaluation of a Retrieval-Augmented Generation (RAG) pipeline designed to address the issue of hallucinations in LLMs for knowledge-intensive tasks, particularly, QA. The study found RAG significantly improved the factual accuracy of LLMs compared to their standalone counterpart and even other state-of-the-art LLMs of similar weight. Through the usage of external knowledge sources through semantic retrieval and reranking, the RAG pipeline effectively grounded the LLM's responses, leading to a substantial reduction in hallucinated content. The findings suggest that RAG-based solutions are a viable approach for improving the reliability of LLMs in PKMS. While challenges in evaluating generative models persist, the combined human and token-based analysis employed in this study provides credible evidence of RAG's effectiveness in mitigating hallucinations. This work highlights the potential of RAG to enhance the trustworthiness of LLMs for a wide range of applications

# References

1. E. Enamorado-Díaz, J.A. García-García, M.J. Escalona-Cuaresma, and D. Lizcano-Casas, “Metaverse Applications: Challenges, Limitations and Opportunities - A Systematic Literature Review,” Information and Software Technology **182**, 107701 (2025).
2. J. Yang, X. Hu, G. Xiao, and Y. Shen, “A Survey of Knowledge Enhanced Pre-trained Language Models,” ACM Trans. Asian Low-Resour. Lang. Inf. Process., 3631392 (2024).
3. M.A. Fink, A. Bischoff, C.A. Fink, M. Moll, J. Kroschke, L. Dulz, C.P. Heußel, H.-U. Kauczor, and T.F. Weber, “Potential of ChatGPT and GPT-4 for Data Mining of Free-Text CT Reports on Lung Cancer,” Radiology **308**(3), e231362 (2023).
4. J. Jayaram, Y. Kulkarni, L.V. Ganesh, P. Naveen, and E.A. Anaam, “Treatment Recommendation using BERT Personalization,” JIWE **3**(3), 41–62 (2024).
5. M.T.-T. Yong, S.-B. Ho, and C.-H. Tan, “Migraine Generative Artificial Intelligence based on Mobile Personalized Healthcare,” JIWE **4**(1), 275–291 (2025).
6. L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, “A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions,” ACM Trans. Inf. Syst. **43**(2), 1–55 (2025).
7. Pal, L.K. Umapathi, and M. Sankarasubbu, “Med-HALT: Medical Domain Hallucination Test for Large Language Models,” in *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, (Association for Computational Linguistics, Singapore, 2023), pp. 314–334.
8. M. Alavi, D. Leidner, and R. Mousavi, “Knowledge Management Perspective of Generative Artificial Intelligence (GenAI),” SSRN Journal, (2024).
9. Z. Liu, Z. Sun, S. Cheng, S. Huang, and M. Wang, “Only 5% Attention Is All You Need: Efficient Long-range Document-level Neural Machine Translation,” in *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Association for Computational Linguistics, Nusa Dua, Bali, 2023), pp. 733–743.
10. T.T. Procko, and O. Ochoa, “Graph Retrieval-Augmented Generation for Large Language Models: A Survey,” in *2024 Conference on AI, Science, Engineering, and Technology (AIxSET)*, (IEEE, Laguna Hills, CA, USA, 2024), pp. 166–169.
11. Y. Wang, M. Wang, M.A. Manzoor, F. Liu, G.N. Georgiev, R.J. Das, and P. Nakov, “Factuality of Large Language Models: A Survey,” in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, (Association for Computational Linguistics, Miami, Florida, USA, 2024), pp. 19519–19529.
12. B.C. Teixeira, C. Gao, C. Wang, A. Malhi, and D. Stein, “SA102 Using Retrieval Augmented Generation (RAG) and Agents’ Workflow for Protocol Authoring,” Value in Health **27**(12), S634 (2024).
13. Z. Yue, J. Liu, C. Zhang, C. Wang, H. Jiang, Y. Zhang, X. Tian, Z. Cen, Y. Xiao, and T. Ruan, “MA-MRC: A Multi-answer Machine Reading Comprehension Dataset,” in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, (ACM, Taipei Taiwan, 2023), pp. 2144–2148.
14. P. Rajpurkar, R. Jia, and P. Liang, “Know What You Don’t Know: Unanswerable Questions for SQuAD,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, (Association for Computational Linguistics, Melbourne, Australia, 2018), pp. 784–789.
15. L.R. Valmaggia, T.K. Bouman, and L. Schuurman, “Attention Training With Auditory Hallucinations: A Case Study,” Cognitive and Behavioral Practice **14**(2), 127–133 (2007).
16. P. Josyula, *Mastering Retrieval-Augmented Generation*, 1st ed (BPB Publications, Delhi, 2025).
17. T.-J. Goh, L.-Y. Chong, S.-C. Chong, and P.-Y. Goh, “A Campus-based Chatbot System using Natural Language Processing and Neural Network,” Journal of Informatics and Web Engineering **3**(1), 96–116 (2024).
18. Khan, K. Khan, W. Khan, S.N. Khan, and R. Haq, “Knowledge-based Word Tokenization System for Urdu,” Journal of Informatics and Web Engineering **3**(2), 86–97 (2024).