Adaptive Memory Enhancement: Augmentation and Self-Check Mechanisms in Retrieval Processes (AMRAG)

Authors: Faran Zafar

Supervisor: Professor Christopher Baber

Affiliation: BSURE Program, Google DeepMind, University of Birmingham

Contents

Table of Figures

Abstract

This paper presents the framework of Adaptive Memory Retrieval Augmentation with Self-Checks, AMRAG, for enhancing the accuracy and reliability of a Retrieval-Augmented Generation, RAG system. It incorporates dynamic query refinement and web search integration for context augmentation, while at the same time diminishing hallucinations and increasing relevance by including self-check mechanisms for retrieved documents and it also proposes a system of Self-Memory Storage System which mutates with the system. This paper illustrates that the AMRAG framework outperforms the traditional RAG framework with a experiments that gauge the quality of the response generated, therefore providing a way to integrate external knowledge into large language models more robustly. This study adds up to a growing body of literature on RAGs, which has an adaptive and more reliable approach and has the potential to transform the way in which retrieval-based generation tasks are handled within different domains.

1. Introduction

Retrieval-Augmented Generation (RAG) has emerged as a powerful method for enhancing the performance of large language models (LLMs) by incorporating external knowledge into the response generation process. While RAG systems have shown remarkable improvements over traditional generation methods, they still face significant challenges, particularly in terms of retrieval accuracy and the generation of hallucinations—false or irrelevant information.

Existing RAG implementations often rely on static retrieval processes that do not adapt to the complexity or ambiguity of queries, leading to suboptimal performance. Moreover, the lack of robust self-check mechanisms within these systems further exacerbates the issue of hallucinations. To address these challenges, this paper proposes the Adaptive Memory Retrieval Augmentation with Self-Checks (AMRAG) framework. AMRAG introduces dynamic query refinement , context augmentation and self-verification processes, enabling more precise and contextually relevant retrieval while significantly reducing the incidence of hallucinations.

2. Background and Related Work

2.1 Retrieval-Augmented Generation (RAG)

Large-scale LLMs such as BERT (Devlin et al. 2019), GPT-3 (Brown et al. 2020), and T5 (Raffel et al. 2020)have revolutionized the NLP domain. Specifically, they demonstrate impressive performance on a wide range of downstream tasks by exploiting large amounts of knowledge captured in pre-training. However, one major drawback of techniques developed along this line is that they are fixed at deployment time; thus, they cannot use real-time information or perform knowledge-intensive tasks effectively.

To address this limitation, Retrieval-Augmented Generation (RAG) was introduced as a framework that combines the generative capabilities of LLMs with the precision of information retrieval systems. RAG allows models to fetch relevant documents from external sources during the generation process, thus enhancing their ability to provide accurate and up-to-date responses (Lewis et al., 2020). Despite the success of RAG systems, challenges such as retrieval inaccuracy and the generation of hallucinated content—irrelevant or incorrect information—remain prevalent (Ji et al., 2023).

2.2 Limitations of Existing RAG Systems

Self-RAG, proposed by Asai et al. (2023), introduces a self-reflection mechanism to decide when retrieval is necessary, reducing irrelevant document usage and potential hallucinations. However, its effectiveness depends heavily on the quality of initial retrievals. If the initial retrieval fails, the system may proceed with insufficient context, leading to suboptimal responses. Additionally, Self-RAG may overlook useful information by not performing multiple retrievals for complex queries.

CRAG (Corrective Retrieval-Augmented Generation) by Yan et al. (2024) enhances retrieval robustness by performing corrective web searches when initial retrievals are inadequate. Nevertheless, CRAG does not store fetched data meaning that similar queries will require repeated web searches which consume resources thus increasing latency and computational load especially in cases where identical queries are raised frequently.

An additional breakthrough is represented by RQ-RAG framework that incorporates query refinement strategies for better management of complex ambiguous queries (Chan et al., 2024). The RQ-RAG enables the model decompose and rewrite dynamic queries thereby increasing their relevance regarding retrieved documents. Nevertheless, this method improves retrieval but fails to completely deal with hallucinatory elements because it mainly targets query optimizing processes.

2.3 Hallucination Detection and Context Augmentation

In recent advancements in Retrieval-Augmented Generation (RAG), the challenge of hallucinations where a model generates incorrect or irrelevant information has gained significant attention. In most cases, traditional RAG systems find it hard to maintain accuracy of the generated content especially when documents retrieved are irrelevant to each other. Several methods have been proposed that aim at refining retrieval processes and improving generation accuracy.

One notable approach is the **Corrective Retrieval-Augmented Generation (CRAG)** framework, proposed by Yan et al. (2024). With this framework, CRAG offers a new strategy for detecting hallucinations through an easy-to-use LLM which is lighter than other hallucinogenic detectors before generation happens around it. The aim of this secondary LLM is to evaluate whether retrieved documents are adequate their use for generation so that it would spot any inaccuracies on them even if they weren't obvious at first glance. Besides this, the creators of CRAG added full-scale web searches whenever initial retrieval from stationary corpora didn't work within static values In this way, the method provides much wider and more reliable knowledge for producing accurate answers.

2.4 The Need for Adaptive Retrieval and Self-Checks

The persistence of the problems in RAG systems shows the need for further adaptive retrieval mechanisms and more robust self-check procedures. By adaptive retrieval, it means that the query is to be dynamically refined for multiple retrieval attempts to retrieve most relevant documents. This becomes very important while dealing with complex queries, which cannot be adequately addressed by a single retrieval attempt.

It means that to a great extent, the propensity for these models to yield hallucinations can be reduced by incorporating self-check mechanisms within the generation process. Self-checks include checking generated content against the retrieved documents so as to eliminate inconsistency and inaccuracy. Such an approach improves not only the reliability of output but also enhances the model's capacity to handle a wider range of queries effectively.

In this line, it proposes an Adaptive Memory Retrieval Augmentation with Self-Checks framework that puts these developments into integrating adaptive query refinement with self-verification processes. Thereby, AMRAG would be imbued with the power of adaptiveness to query complexity and the self-checking of content generated in real time, enhancing some existing key limitations of RAG systems.

4. Methodology/Framework

The AMRAG framework integrates several novel components designed to enhance the retrieval and generation processes in RAG systems. The workflow of AMRAG is illustrated in Figure 1, and its key components are described below.

4.1 Query Analysis and Decomposition

The query analysis and decomposition process is a critical component of the AMRAG framework, designed to enhance the system's ability to accurately retrieve relevant information from complex or ambiguous queries. This process begins as soon as the user inputs a query into the system.

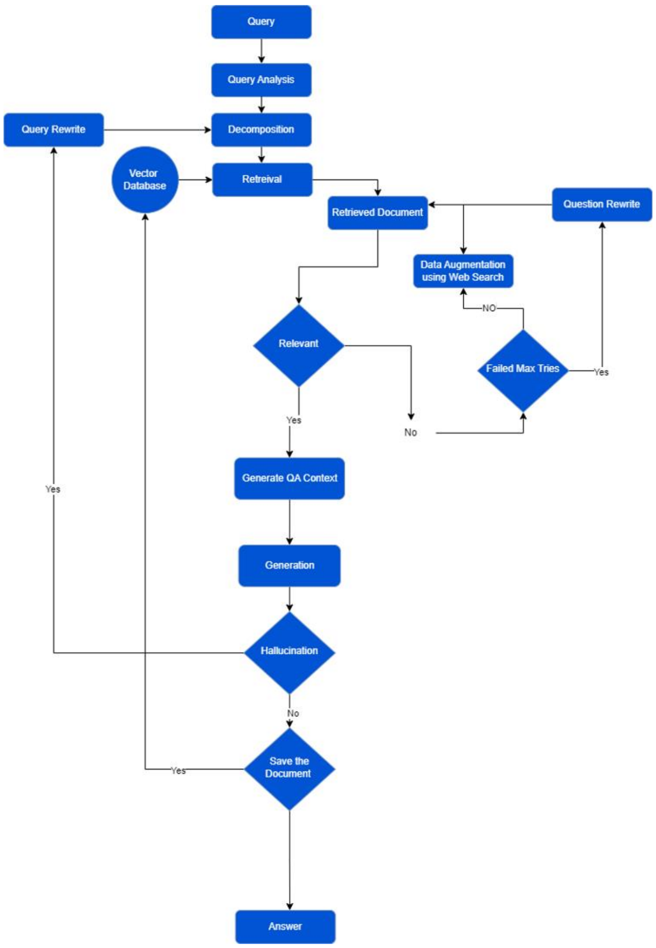


Figure 1 Showing the AMRAG Framework

4.1.1 Decomposition with GPT-4o-mini

If the initial analysis indicates that the query is complex or contains multiple facets, the AMRAG framework proceeds to decompose the query into simpler sub-queries. This decomposition is performed using GPT-4o-mini, a lightweight language model specifically for this purpose.

GPT-4o-mini takes the complex query as input and generates several sub-queries, each targeting a specific aspect of the original query. These sub-queries are designed to be more precise and narrowly focused, making it easier for the retrieval system to identify and extract relevant information from the document corpus.

For instance, this query “What are the key considerations, including benefits, challenges, and impacts on urban areas, associated with using renewable energy?” using GPT-4o-mini might break down the query into the following sub-queries:

- "What are the benefits of using renewable energy?"
- "What are the challenges of using renewable energy?"
- "How does renewable energy impact urban areas?"

By breaking down the query in this manner, AMRAG ensures that each sub-query can be treated independently during the retrieval process, leading to a more comprehensive and accurate aggregation of information in the final response.

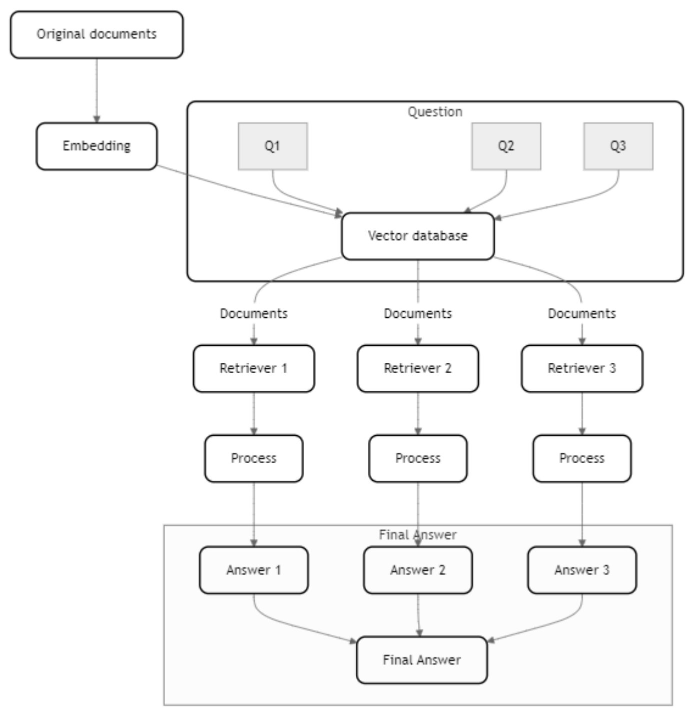


Figure 2 Showing the Decomposition Technique

4.1.2 Integration into Retrieval Process

Once the sub-queries are generated, each is processed separately through the retrieval system. The documents retrieved in response to each sub-query are then aggregated and synthesized to form a complete response to the original complex query.

This decomposition process not only improves the accuracy of the retrieved information but also helps in reducing potential hallucinations by focusing the retrieval on specific, well-defined aspects of the query. The use of GPT-4o-mini ensures that this decomposition is done efficiently, making it suitable for real-time applications within the AMRAG framework.

4.2 Retrieval and Relevance Checking

It uses a decomposed query to retrieve documents within a pre-built vector database. This is an iterative process of retrieval, where the system will back off to query rewrites in case the first attempt at retrieval does not return relevant results. In case relevant documents are still not found, **AMRAG resorts to external web searches** to increase the available data, ensuring that the most complete information will be found.

Relevant documents are then checked for retrieval. In case the retrieved document is relevant, proceed to the next step; otherwise, the query should be rewritten and the retrieval repeated. This cycle may be continued until relevant documents have been found or a maximum number of attempts reached.

4.3 Generation and Self-Check Mechanisms

First, relevant documents are identified, and then the QA context is generated from them to answer the query. The system responds with a reply based on that context. From the self-reflective approach developed by Asai et al **(**2023**)** for their SELF-RAG framework, AMRAG adds a self-check mechanism as an additional module to review the generated content critically for probable inaccuracies or hallucinations. It involves a reflection process in which the model checks its output for consistency and factual accuracy. In case of a hallucination or inconsistency, it loops back to the phase of refining the query to correct the process. Otherwise, if no issues are found, the response is finalized and delivered to the user.

4.4 Final Output

The final response, verified and refined through multiple stages of self-checking and query adjustment, is then presented to the user. This ensures that the generated answer is both accurate and contextually appropriate.

5. Results

5.1 Testing Methodology

Testing the effectiveness of the proposed Adaptive Memory Retrieval Augmentation with Self-Checks (AMRAG) framework against a traditional Retrieval-Augmented Generation system involves a thorough evaluation process. In the evaluation, we focus on producing high-quality responses pertaining to diverse natural language processing tasks from both systems.

For evaluation, we leveraged the **RAGTruth** dataset provided by Niu et al. (2023), which contains a set of questions along with their corresponding answer context passages. These have been vectorized into a vector database for fast retrieval during test time. This dataset contained question/answer pairs; we used these questions as the queries to test both the AMRAG and baseline RAG systems. This setup enabled us to evaluate the ability of the systems in retrieving relevant information and generating responses that are accurate and appropriate in context.

5.1.1 Binary Classification by a Large Language Model (LLM)

The primary evaluation technique used was presenting the responses of both simple RAG and AMRAG, then having an LLM perform binary classification. For every query, RAG and AMRAG generated a single response. Afterwards, the responses were passed through the LLM GPT-4, which had to classify which of the two responses was better.

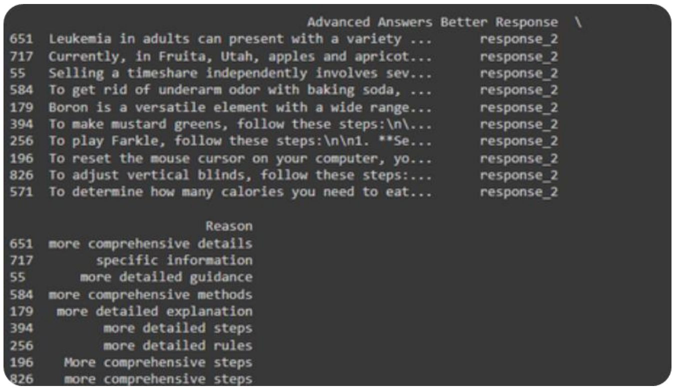
The LLM was not only asked to choose between "Response 1" from RAG and "Response 2" from AMRAG but also to justify this choice. What this did was guarantee that the classification would not be exclusively quantitative—that is, to indicate which response was better—but also qualitative, due to the fact that the reasons that the LLM gave for its choice clarified the nuances behind its decisions.

5.2 Performance Metrics

The primary metric used to assess the performance was the percentage of instances where AMRAG's response was judged to be better than the simple RAG system's response. This metric provides a direct comparison of the two systems' effectiveness.

5.3 Comparative Analysis

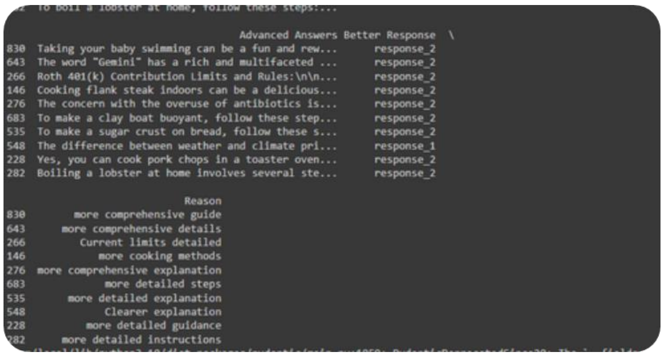
The results of the testing showed that AMRAG outperformed the simple RAG system **in 88% of the cases** of a pool of 50 questions asked each to RAG and AMRAG. Specifically, in the majority of queries, the LLM classified AMRAG's response as better due to various reasons such as the answer being more comprehensive, more detailed, clearer explanation etc.



```
                                   Advanced Answers Better Response  \
651  Leukemia in adults can present with a variety ...      response_2
717  Currently, in Fruita, Utah, apples and apricot...      response_2
55   Selling a timeshare independently involves sev...      response_2
584  To get rid of underarm odor with baking soda, ...      response_2
179  Boron is a versatile element with a wide range...      response_2
394  To make mustard greens, follow these steps:\n\...      response_2
256  To play Farkle, follow these steps:\n\n1. **Se...      response_2
196  To reset the mouse cursor on your computer, yo...      response_2
826  To adjust vertical blinds, follow these steps:...      response_2
571  To determine how many calories you need to eat...      response_2

                        Reason
651  more comprehensive details
717        specific information
55       more detailed guidance
584  more comprehensive methods
179   more detailed explanation
394         more detailed steps
256         more detailed rules
196    More comprehensive steps
826    more comprehensive steps
```

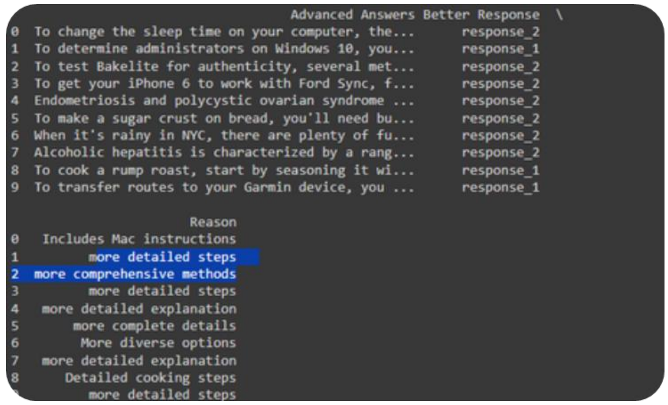
Figure 3 Batch 1 : Result of Binary Classification of Better Reponses Done by LLM



```
                                   Advanced Answers Better Response  \
830  Taking your baby swimming can be a fun and rew...      response_2
643  The word "Gemini" has a rich and multifaceted ...      response_2
266  Roth 401(k) Contribution Limits and Rules:\n\n...      response_2
146  Cooking flank steak indoors can be a delicious...      response_2
276  The concern with the overuse of antibiotics is...      response_2
683  To make a clay boat buoyant, follow these step...      response_2
535  To make a sugar crust on bread, follow these s...      response_2
548  The difference between weather and climate pri...      response_1
228  Yes, you can cook pork chops in a toaster oven...      response_2
282  Boiling a lobster at home involves several ste...      response_2

                          Reason
830       more comprehensive guide
643     more comprehensive details
266        Current limits detailed
146           more cooking methods
276  more comprehensive explanation
683            more detailed steps
535      more detailed explanation
548            Clearer explanation
228         more detailed guidance
282     more detailed instructions
```

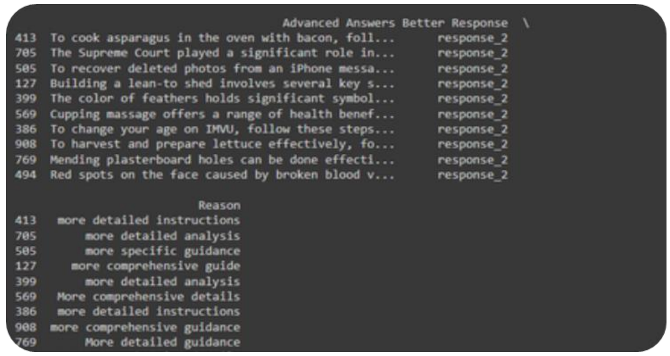
Figure 4 Batch 2 : Result of Binary Classification of Better Reponses Done by LLM



```
                                 Advanced Answers Better Response  \
0  To change the sleep time on your computer, the...      response_2
1  To determine administrators on Windows 10, you...      response_1
2  To test Bakelite for authenticity, several met...      response_2
3  To get your iPhone 6 to work with Ford Sync, f...      response_2
4  Endometriosis and polycystic ovarian syndrome ...      response_2
5  To make a sugar crust on bread, you'll need bu...      response_2
6  When it's rainy in NYC, there are plenty of fu...      response_2
7  Alcoholic hepatitis is characterized by a rang...      response_2
8  To cook a rump roast, start by seasoning it wi...      response_1
9  To transfer routes to your Garmin device, you ...      response_1

                      Reason
0   Includes Mac instructions
1         more detailed steps
2  more comprehensive methods
3         more detailed steps
4   more detailed explanation
5       more complete details
6        More diverse options
7   more detailed explanation
8      Detailed cooking steps
          more detailed steps
```

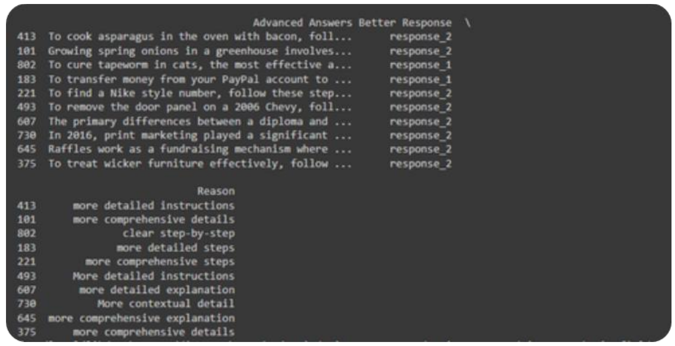
Figure 5 Batch 3 : Result of Binary Classification of Better Reponse Done by LLM



```
                                   Advanced Answers Better Response  \
413  To cook asparagus in the oven with bacon, foll...      response_2
705  The Supreme Court played a significant role in...      response_2
505  To recover deleted photos from an iPhone messa...      response_2
127  Building a lean-to shed involves several key s...      response_2
399  The color of feathers holds significant symbol...      response_2
569  Cupping massage offers a range of health benef...      response_2
386  To change your age on IMVU, follow these steps...      response_2
908  To harvest and prepare lettuce effectively, fo...      response_2
769  Mending plasterboard holes can be done effecti...      response_2
494  Red spots on the face caused by broken blood v...      response_2

                         Reason
413  more detailed instructions
705      more detailed analysis
505      more specific guidance
127    more comprehensive guide
399      more detailed analysis
569  More comprehensive details
386  more detailed instructions
908 more comprehensive guidance
769      More detailed guidance
```

Figure 6 Batch 4: Result of Binary Classification of Better Reponses Done by LLM



```
                                   Advanced Answers Better Response  \
413  To cook asparagus in the oven with bacon, foll...      response_2
101  Growing spring onions in a greenhouse involves...      response_2
802  To cure tapeworm in cats, the most effective a...      response_1
183  To transfer money from your PayPal account to ...      response_1
221  To find a Nike style number, follow these step...      response_2
493  To remove the door panel on a 2006 Chevy, foll...      response_2
607  The primary differences between a diploma and ...      response_2
730  In 2016, print marketing played a significant ...      response_2
645  Raffles work as a fundraising mechanism where ...      response_2
375  To treat wicker furniture effectively, follow ...      response_2

                          Reason
413      more detailed instructions
101      more comprehensive details
802             clear step-by-step
183            more detailed steps
221       more comprehensive steps
493      More detailed instructions
607       more detailed explanation
730          More contextual detail
645  more comprehensive explanation
375      more comprehensive details
```

Figure 7 Batch 5 : Result of Binary Classification of Better Reponses Done by LLM

5.3.1 Reasons for Superior Performance of AMRAG

The LLM's reasoning provided valuable insights into why AMRAG's responses were frequently superior. Some of the key reasons highlighted by the LLM include:

- **Better Relevance:** Because of the dynamic query refinement operated by the AMRAG system, it was able to retrieve documents more relevant to an input query than the other compared systems, hence able to respond more aligned to the user's query

- **Hallucinations Reduction:** The self-check mechanisms that were inbuilt into the AMRAG reduced drastically the increase of hallucinations. This was noted by the LLM, where a good number of the responses offered by the AMRAG were mainly factual with very minimal percentage such as speculative elements compared to the simple RAG.
- **Contextual Augmentation:** The AMRAG framework was better at dealing with questions and queries that were ambiguous given the ability to run a web search to augment that data

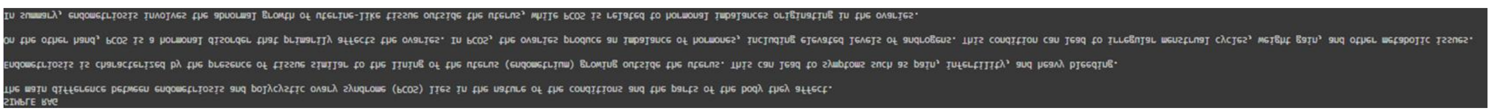
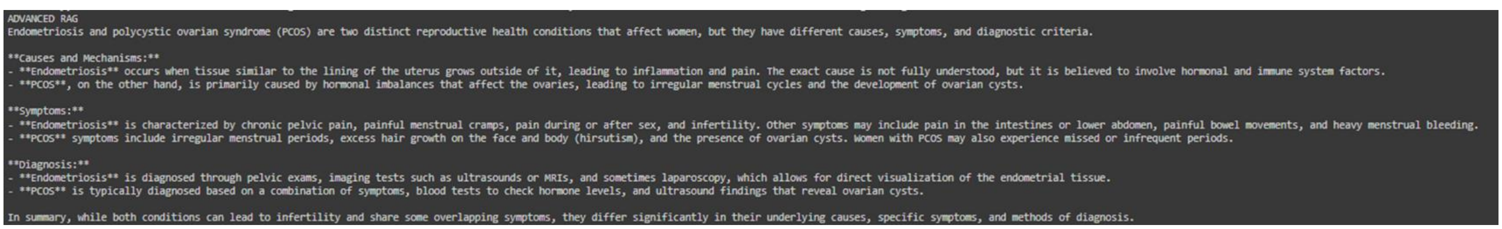


Figure 8 Showing the Response Generated by Simple RAG to the Question: What is the difference between Endometriosis and Polycystic ovarian syndrome?



ADVANCED RAG
Endometriosis and polycystic ovarian syndrome (PCOS) are two distinct reproductive health conditions that affect women, but they have different causes, symptoms, and diagnostic criteria.

Causes and Mechanisms:
- **Endometriosis** occurs when tissue similar to the lining of the uterus grows outside of it, leading to inflammation and pain. The exact cause is not fully understood, but it is believed to involve hormonal and immune system factors.
- **PCOS**, on the other hand, is primarily caused by hormonal imbalances that affect the ovaries, leading to irregular menstrual cycles and the development of ovarian cysts.

Symptoms:
- **Endometriosis** is characterized by chronic pelvic pain, painful menstrual cramps, pain during or after sex, and infertility. Other symptoms may include pain in the intestines or lower abdomen, painful bowel movements, and heavy menstrual bleeding.
- **PCOS** symptoms include irregular menstrual periods, excess hair growth on the face and body (hirsutism), and the presence of ovarian cysts. Women with PCOS may also experience missed or infrequent periods.

Diagnosis:
- **Endometriosis** is diagnosed through pelvic exams, imaging tests such as ultrasounds or MRIs, and sometimes laparoscopy, which allows for direct visualization of the endometrial tissue.
- **PCOS** is typically diagnosed based on a combination of symptoms, blood tests to check hormone levels, and ultrasound findings that reveal ovarian cysts.

In summary, while both conditions can lead to infertility and share some overlapping symptoms, they differ significantly in their underlying causes, specific symptoms, and methods of diagnosis.

Figure 9 Showing the Response Generated by AMRAG to the Question: What is the difference between Endometriosis and Polycystic ovarian syndrome?

5.5 Discussion

The results of this testing underscore the advantages of the AMRAG framework over traditional RAG systems. The combination of adaptive query refinement and self-check mechanisms not only enhances the relevance and accuracy of the responses but also contributes to a more reliable and trustworthy output.

The reasons provided by the LLM for selecting AMRAG's responses over those of the simple RAG system further validate the design choices made in developing AMRAG. By addressing the common pitfalls of traditional RAG systems—such as retrieval inaccuracies and hallucinations

6. Future Work

6.1 Enhancing Retrieval Efficiency with Self-Memory Storage System (SMSS)

The most promising area of future work will be the development and integration of a Self-Memory Storage System into the AMRAG framework. The definition of an SMSS was strongly inspired by the framework proposed by Cheng et al(2023). in their paper on retrieval-augmented text generation with self-memory, where the model generates its outputs iteratively to use them as memory during further generations. By applying this framework, SMSS would enable AMRAG to reap benefits from informative documents—also previously responding queries—to make a more efficient and accurate generation of the response over time.

6.1.1 Concept and Purpose

We envision SMSS as a learnable memory module, in which valuable information from past queries will be stored. By storing these documents, it would enable the system to avoid redundant web searches for similar future queries and thus reduce retrieval times and computational overhead considerably. This could also help assure more accurate and contextually relevant responses by using the past retrievals to inform outputs in the future

6.1.2 Anticipated Impact

It may considerably improve the retrieval efficiency and accuracy of response to similar queries, especially as encountered in a previous message. Over time, the SMSS will enable the AMRAG to develop a more personalized, dynamic memory that will boost its performance on a much wider array of NLP tasks.

6.2 Comprehensive Testing and Evaluation
Further and better understanding of the potential and limitation of the AMRAG framework calls for more extensive testing. In the future, rigorous testing across a wide range of datasets and NLP tasks is necessary to evaluate how well the framework can perform in different contexts. A number of the following elements should, to this effect, be explored:

- **Robustness Against Diverse Query Types:** This test issues a call to an evaluation of AMRAG based on its ability to process queries varying on all difficulty levels, including the intrinsically ambiguous, multifaceted, or domain-specific.
- **Effectiveness of Self-Check Mechanisms**: Evaluating how much self-check mechanisms reduce the chances of hallucinations in difficult and open-ended queries.
- **Comparison to State-of-Art Models:** Comparative studies on other retrieval-aided generation models, like CRAG and SELF-RAG, have shown that AMRAG can yield competitive performance.

6.3 Evaluation Metrics

Several metrics will be applied to give an overarching evaluation of the performance for the proposed framework of AMRAG, all of which deal with different aspects related to the functionality of a system.

- **Relevance and Accuracy**: Precision@K, Recall@K, Exact Match, and F1 Score will measure the accuracy and relevance of the recovered and generated content.
- **Robustness:** Hallucination Rate and Error Rate will help evaluate the framework's capability to come up with correct and factually accurate responses. The consistency score will be used to assess stability.
- **Efficiency:** Query latency, retrieval time, and response generation time will be crucial to measuring the applicability of this system in real-time applications and how efficient it is.
- **User Experience**: The user satisfaction score will go hand in hand with the Engagement Rate to measure the effectiveness and attractiveness of the system from the user's point of view.
- **Comparative Analysis:** Draw comparisons in performance between AMRAG and existing RAG models and frameworks using IOB and A/B testing results.

7. Conclusion

This paper introduced the Adaptive Memory Retrieval Augmentation with Self-Checks (AMRAG) framework, a novel approach to enhancing the accuracy and reliability of retrieval-augmented generation systems. By integrating adaptive query refinement and self-verification mechanisms along with context augmentation using web search, AMRAG addresses the key challenges of traditional RAG systems, such as retrieval inaccuracy and hallucinations.

The results of our experiments demonstrate that AMRAG offers significant improvements in both retrieval precision and answer relevance. Future work could explore the effect of Self-Memory Storage System (SMSS) along with the application of AMRAG to more diverse datasets and its integration with real-time, dynamic environments.

References:

1. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. and Amodei, D. (2020) ‘Language Models are Few-Shot Learners’, *arXiv preprint arXiv:2005.14165*.
2. Cheng, X., Luo, D., Chen, X., Liu, L., Zhao, D., & Yan, R. (2023). **Lift Yourself Up: Retrieval-augmented Text Generation with Self-Memory**. Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023).
3. Chan, C.-M., Xu, C., Yuan, R., Luo, H., Xue, W., Guo, Y. and Fu, J. (2024) ‘RQ-RAG: Learning to Refine Queries for

Retrieval Augmented Generation', *arXiv preprint arXiv:2404.00610*.
4. Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019) 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1, pp. 4171-4186.
5. Gao, L., Ma, X., Lin, J. and Callan, J. (2022) 'Precise Zero-Shot Dense Retrieval without Relevance Labels', *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 1706-1717.
6. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. and Madotto, A. (2023) 'Survey of Hallucination in Natural Language Generation', *ACM Computing Surveys*, 55(12), pp. 1-38.
7. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S. and Kiela, D. (2020) 'Retrieval-augmented generation for knowledge-intensive NLP tasks', *Advances in Neural Information Processing Systems*, 33, pp. 9459-9474.
8. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P. (2020) 'Exploring the limits of transfer learning with a unified text-to-text transformer', *Journal of Machine Learning Research*, 21(1), pp. 5485-5550.
9. Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2023). **SELF-RAG: Learning to Retrieve, Generate, and Critique Through Self-Reflection**. Preprint. Available at arXiv:2310.11511.
10. Niu, C., Wu, Y., Zhu, J., Xu, S., Shum, K., Zhong, R., Song, J., & Zhang, T. (2023). **RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models**.