Universität Trier

Fachbereich II Computerlinguistik und Digital Humanities

The Impact of Retrieval Strategies on Hallucination and Accuracy in RAG-Based Question Answering

Term Paper

In Module MA2NLP2002 Natural Language Understanding - P1240M0010e

Lecturer: Dr. phil. Simon Werner

vorgelegt von

Ahmad Abdullah| 1737077

S2adabdu@uni-trier.des

Summer Semester 2025 Submission on 29/30/2025

Universität Trier

Declaration of Originality

Student Name:	Ahmad Abdullah	Matriculation Number:	1737077
Date of Birth:	14.01.2002		
Subject of Study: Course Title:	Trends in Natural language Processing		
Topic of Work:	The Impact of Retrieval Strategies on Hallucination and Accuracy in RA Based Question Answering		

I declare herewith, that this above-mentioned work (essay, project, thesis etc.) is my own original work.

Furthermore, I confirm that:

- this work has been composed by me without assistance;
- I have clearly referenced in accordance with departmental requirements, in both the text and the bibliography or references, all sources (either from a printed source, internet or any other source) used in the work;
- all data and findings in the work have not been falsified or embellished;
- this work has not been previously, or concurrently, used either for other courses or within other exam processes as an exam work;
- this work has not been published.

I appreciate that any false claim in respect of this work will result in disciplinary action in accordance with university or departmental regulations.

I confirm that I understand that my work may be electronically checked for plagiarism by the use of plagiarism detection software and stored on a third party's server for eventual future comparison.

Signature Ahmad Abdullah …….. **Date** 30.09.2025 …….

This Declaration must be filled out, signed and attached to your student work.

Contents

Abstract

This study examines the impact of different retrieval strategies on the accuracy and reliability of Retrieval-Augmented Generation (RAG) systems. Specifically, we compare the performance of sparse retrieval using BM25, dense retrieval using DPR, and a hybrid retriever that combines both methods. The evaluation was conducted in two stages: first, by measuring retrieval accuracy in isolation, and second, by assessing hallucination and correctness when retrieval was combined with a large language model (LLM). Our findings show that hybrid retrieval consistently achieves the best balance between accuracy and factual grounding, while fine-tuning significantly improves the effectiveness of dense retrievers. Moreover, the amount of available training data plays a critical role in determining which retrieval strategy performs best, with sparse methods excelling in smaller datasets and dense or hybrid approaches gaining advantages in larger corpora. These results highlight the importance of retrieval design choices for minimizing hallucination and ensuring reliable outputs in RAG-based question answering.

1 Introduction

Large Language Models (LLMs) such as GPT-4 and LLaMA have transformed natural language processing by enabling coherent and contextually appropriate text generation. These models have advanced applications including dialogue systems, summarization, and open-domain question answering. Despite their linguistic fluency, LLMs often produce responses that are factually inaccurate or entirely fabricated—a phenomenon known as hallucination. Hallucinations undermine user trust, particularly in question answering tasks where precision and factual grounding are critical.

To address this, RAG models have been proposed as a more efficient alternative, integrating external knowledge sources during inference to provide up-to-date and accurate information (Lewis et al., 2020; Borgeaud et al., 2022; Lee et al., 2024). Rather than relying solely on the model's parametric memory, RAG retrieves relevant documents from a corpus and conditions the LLM on this evidence when generating answers. By incorporating supporting context, RAG reduces hallucination and improves factual accuracy.

However, the effectiveness of RAG depends heavily on the retrieval step. If retrieved documents are irrelevant, incomplete, or noisy, the generator may produce inaccurate or misleading responses. Retrieval strategies—sparse, dense, and hybrid—introduce distinct trade-offs in recall, precision, and robustness. Sparse methods rely on exact keyword matching, dense methods leverage semantic embeddings, and hybrid methods combine both approaches. Because the quality of retrieval directly shapes the model's output, evaluating the impact of these strategies on accuracy and hallucination is essential for developing reliable RAG systems.

2 Background

2.1 Retrieval-Augmented Generation (RAG) Architecture

The RAG framework augments a generative language model with a retrieval module that provides external evidence during inference. Formally, given a user query (q), the system proceeds through the following pipeline:

1. **Query Formulation:** A natural-language query q is submitted by the user. This query is encoded either as a sparse vector (e.g., TF-IDF, BM25), a dense embedding (e.g., DPR, Contriever), or both in the case of hybrid retrieval.

2. **Retrieval Step:** The retriever searches the indexed document corpus D and returns the top-k candidate passages $(d_1, d_2, \ldots, d_k)$ ranked by relevance score. Sparse retrievers rely on term overlap, dense retrievers compute similarity in a continuous embedding space, and hybrid retrievers combine both signals.

3. **Chunking and Preprocessing:** : According to Oche et al. (2025), large documents are often divided into smaller, self-contained passages, such as paragraphs, to enable more efficient indexing and retrieval in RAG systems. Chunking increases the probability that critical information is included in the retrieval set.

4. **Context Injection:** Retrieved chunks are integrated into the input sequence for the language model. Common strategies include:

    - **Concatenation:** Directly appending retrieved passages to the query.
    - **Attention-based fusion:** Encoding retrieved passages separately and allowing the model to attend to them dynamically during generation.

5. **Answer Generation:** The LLM conditions its output on both the query and the injected retrieval context. Ideally, the generation process leverages external evidence to produce factually accurate responses, while still relying on the model's reasoning abilities for synthesis.

2.2 Accuracy and Hallucination in RAG

Accuracy in RAG-based systems refers to the degree to which generated responses match the ground truth. It is typically measured using both automatic metrics—such as Exact Match, F1 Score, BLEU, ROUGE, or semantic similarity—and human evaluations for nuanced judgments of correctness.

Hallucination, by contrast, describes the generation of fluent but unsupported or factually incorrect content. Two primary types are commonly distinguished

- **Intrinsic hallucination:** Output contradicts retrieved passages.
- **Extrinsic hallucination:** The output introduces information that is not present in or supported by the retrieved context

Evaluating hallucination remains a significant challenge. Common approaches include fact-checking model outputs against retrieved evidence, attribution scoring, truthfulness benchmarks, automatic metrics such as QAGS, and human annotation.

3 Retrieval Strategies in RAG

Sparse retrieval methods, such as BM25 and TF-IDF, rely on keyword overlap to identify relevant passages. They are efficient, scalable, and easy to interpret, which makes them effective for domain-specific or keyword-heavy queries. However, their inability to capture semantic similarity means that paraphrased queries often fail, increasing the risk of hallucination when critical evidence is missed.

Dense retrieval instead maps queries and documents into a shared embedding space, enabling semantic similarity matching. Oche et al. (2025) describe DPR as a bi-encoder framework in which a question encoder maps queries into vector space and a passage encoder projects documents into the same space, with relevance determined by a similarity function such as the dot product. Approaches like DPR and Contriever excel at capturing paraphrases and ambiguous queries, which improves recall and reduces reliance on parametric memory. Still, they are computationally expensive, harder to interpret, and sometimes retrieve semantically similar but irrelevant passages, which can introduce noise. As Karpukhin et al. (2020) explain, synonyms or paraphrases with different tokens can still be represented by vectors that are close in the embedding space. Hybrid retrieval combines the two approaches, either by rescoring sparse candidates with dense embeddings or merging results from both systems. This strategy typically balances precision and recall, leading to stronger factual grounding and fewer hallucinations across diverse queries. The trade-off, however, is increased computational cost and more complex tuning. Despite these challenges, hybrid retrievers often provide the most robust performance for knowledge-intensive tasks.

4 Experiments

To evaluate the impact of different retrieval strategies on both retrieval performance and the overall performance of a Retrieval-Augmented Generation (RAG) system, we conducted two separate experiments. The first experiment focuses solely on the retriever to isolate its performance, while the second experiment integrates the retriever with a generative language model to evaluate end-to-end RAG performance, including both accuracy and hallucination.

4.1 Experiment 1: Retriever-Only Evaluation

The first experiment isolates the retrieval component to measure recall, defined as the proportion of relevant documents successfully retrieved from the corpus. This approach allows us to evaluate the effectiveness of different retrieval strategies independently of the generative model. We compare three retrieval strategies: BM25 (sparse), DPR (dense), and a hybrid retriever that combines BM25 and DPR signals.

Two datasets were used for this evaluation: SQuAD v1 and SQuAD v2. For SQuAD v1, we indexed approximately 20,000 examples, while SQuAD v2 used around 80,000 examples, reflecting a larger and more diverse dataset. We set the retrieval parameter (K) to 5 for SQuAD v1 and 9 for SQuAD v2, meaning that the top 5 or 9 passages were retrieved for each query, respectively.

Recall was chosen as the primary metric because it quantifies the ability of the retriever to capture relevant passages, independent of the generative model. According to Oche et al. (2025), modern RAG systems often integrate dense retrieval with lightweight filtering or hybrid approaches (e.g., BM25 combined with embeddings) to improve recall on challenging queries. Exact match was avoided since answers can vary in wording yet still contain the necessary information. Instead, semantic similarity was computed between retrieved passages and the gold standard answers using the FIASS metric to evaluate relevance. The results of this experiment are provided in Table 1.

Dataset	K	Retriever	Recall
SQuAD v1	5	BM25	0.69
SQuAD v1	5	DPR	0.54
SQuAD v1	5	Hybrid	0.78
SQuAD v2	9	BM25	0.66
SQuAD v2	9	DPR (Fine-tuned)	0.76
SQuAD v2	9	Hybrid	0.83

Table 1: Retriever-only comparison on SQuAD v1 and SQuAD v2.

Observations from these results as shown in Table 1, indicate that the smaller SQuAD v1 dataset favored BM25, likely due to its reliance on exact keyword matches and limited training data for DPR. For SQuAD v2, fine-tuning the dense DPR retriever, increasing (K), and indexing a larger dataset improved recall for both DPR and the hybrid approach. These results confirm that hybrid retrieval benefits from both semantic understanding and keyword matching, achieving the highest recall across datasets.

4.2 Experiment 2: Retriever + LLM Evaluation

The second experiment evaluates the end-to-end RAG system, combining the retrieval strategies with a generative language model to measure both accuracy and hallucination. We used the gpt-neo-125M model as our LLM. The same three retrievers—BM25, fine-tuned DPR, and hybrid—were compared to assess how retrieval quality influences the generated answers' correctness and factuality.

Accuracy measures the proportion of correct answers relative to the total number of questions, while hallucination quantifies the fraction of outputs that are fluent but unsupported or factually incorrect. These metrics together provide insight into how well the RAG system produces reliable, grounded responses.

Retriever + LLM	Accuracy	Hallucination
BM25 + LLM	0.60	0.25
DPR (Fine-tuned) + LLM	0.65	0.30
Hybrid + LLM	0.78	0.18

Table 2: End-to-end evaluation on SQuAD v1 benchmark.

In Table 2, it is evident that the hybrid retriever provides the best performance when combined with the LLM, achieving the highest accuracy and lowest hallucination. Fine-tuning DPR improves accuracy, particularly on larger datasets, though BM25 remains effective for keyword-heavy queries but lacks semantic coverage.

Together, the two experiments illustrate the importance of retriever selection and dataset size. Evaluating retrievers independently using recall allows for optimization before integrating the generative model, while the end-to-end evaluation demonstrates how retrieval quality directly affects answer correctness and hallucination in the RAG system

5 Discussion

To evaluate retrieval strategies in RAG systems, we first examined the retriever alone to understand its standalone effectiveness. The retriever is responsible for ranking and providing candidate passages, and if its accuracy is poor, even a powerful LLM cannot compensate, causing the RAG system to fail. In the first experiment, we observed that BM25 performs well on small, coherent datasets, while on larger, more scattered datasets, sparse retrieval underperforms, and dense or hybrid retrievers provide better coverage. Fine-

tuning the dense retriever further improved its performance, highlighting the importance of adapting retrievers to dataset characteristics. In the second experiment, we evaluated the full RAG system with the LLM, measuring both accuracy and hallucination. Hybrid retrieval consistently achieved the best balance, yielding the highest accuracy and lowest hallucination, while BM25 remained limited to more homogeneous datasets. This showed that hybrid retrieval, even though computationally expensive, can work very well even when the dataset is mixed and not specific.

Beyond these observations, several practical considerations emerge. One is the choice of evaluation criteria for retrieved answers. While exact matching provides a strict measure of correctness, it often underestimates the utility of passages in open-ended tasks. Evaluating semantic similarity against the gold answer captures contextually relevant but non-identical responses, significantly boosting measured accuracy and recall. This approach is well-suited for open-domain QA or chat-based applications, though in scenarios demanding highly precise answers—such as medical or financial datasets— where the precise answer is a need, exact matching remains necessary. Another critical factor is the retrieval parameter K, which determines the number of top passages retrieved. For generic datasets with ambiguous terms, sparse retrievers may require a higher K to ensure relevant content appears in the retrieved set, allowing the LLM to identify the most appropriate answer even from lower-ranked passages. Conversely, dense or hybrid retrievers that leverage semantic relationships can often operate effectively with a lower K, as the top passages are already highly relevant. These considerations illustrate the interplay between retriever selection, dataset characteristics, evaluation criteria, and retrieval parameters in optimizing both factual accuracy and robustness in RAG-based question answering. Another important observation is that the choice of retriever should be guided by the characteristics of the dataset, including its specificity to one or multiple topics, as well as the intended use case. Additionally, effective preprocessing of the dataset significantly enhances retriever performance; when the data is clean and well-structured, all retrieval methods tend to operate more accurately and efficiently.

6 Conclusion

The choice of retrieval strategy significantly affects RAG system performance. Sparse retrievers work well for small, coherent datasets, while dense and hybrid retrievers excel on larger, diverse, or semantically complex data. Fine-tuning dense retrievers improves recall and accuracy. Evaluating both retriever-only and end-to-end performance is essential, and parameters like top-k retrieval, dataset preprocessing, and evaluation methods im-

pact results. Overall, hybrid retrieval with careful tuning offers the best balance between accuracy and reduced hallucination, supporting reliable, knowledge-grounded question answering.

7 Future Work

Future research could explore several avenues to further improve RAG-based question answering systems. First, investigating the effect of chunk size on retrieval and generation performance could provide insights into optimizing the trade-off between context granularity and computational efficiency, particularly when balancing small versus large chunks with limited processing power. Second, alternative hit detection methodologies beyond semantic similarity or exact match could enhance the relevance of retrieved passages, especially for open-ended or complex queries. Third, introducing learnable parameters within retrievers may allow the system to dynamically adapt to the dataset and query distribution, improving overall retrieval quality. Fourth, further exploration of ranking methodologies could help prioritize the most contextually relevant passages for the LLM, reducing hallucination and improving accuracy. Additional directions could include multi-hop retrieval to capture information spanning multiple documents, and evaluation of cross-lingual retrieval strategies to extend RAG capabilities to multilingual datasets. Finally, integrating feedback loops from downstream LLM output to retriever tuning could enable iterative improvement of retrieval quality over time.

Appendix A: Retrieval Examples.

Example 1:

```
Query: 'Which private companies are leading reusable rocket development?'


==================== BM25 RESULTS ====================

1. Score: 4.6657
   Text: In the 21st century , private companies have transformed the landscape of space

==================== DPR RESULTS ====================

1. Score: 0.5622
   Text: In the 21st century , private companies have transformed the landscape of space

==================== FUSION RESULTS ====================

1. Score: 1.0000
   Text: In the 21st century , private companies have transformed the landscape of space
   BM25: 4.6657, DPR: 0.5622
```

Displaing the highest scored hit.

Example 2 :

Displaying how the k=3 looks like for a query.

```
Query: 'Which presidents were involved in the Civil War?'


==================== BM25 RESULTS ====================

1. Score: 4.2215
   Text: 1973 ) Allende regime in Chile-C.I.A . ( 1973 ) Agnew resigns

2. Score: 2.6269
   Text: of Tonkin Resolution ( 1964 ) , Tet Offensive ( 1968 ) Mr. Jar

3. Score: 2.3283
   Text: Dallas , TX ( Nov. 1963 ) by Lee Harvey Oswald 36 . Lyndon B.

==================== DPR RESULTS ====================

1. Score: 0.5311
   Text: U.S. threatens to take Cuba by force if deal is not made 15 .

2. Score: 0.4608
   Text: Proclamation ( 1863 ) Ten Percent Plan Lincoln ’ s assassinati
```

```
2. Score: 0.4608
   Text: Proclamation ( 1863 ) Ten Percent Plan Lincoln ’ s assassination-April 14 , 1865 b

3. Score: 0.4034
   Text: Clayton-Bulwer Treaty ( 1850 ) -Britain and the U.S. agree not to expand in Centra

=================== FUSION RESULTS ===================

1. Score: 0.7246
   Text: 1973 ) Allende regime in Chile-C.I.A . ( 1973 ) Agnew resigns ( 1973 ) Watergate S
   BM25: 4.2215, DPR: 0.3370

2. Score: 0.6891
   Text: U.S. threatens to take Cuba by force if deal is not made 15 . James Buchanan , 185
   BM25: 1.5969, DPR: 0.5311

3. Score: 0.4945
   Text: .P.-Chester A. Arthur Secretary of State-James A. Blaine Major Items : Garfield ’
   BM25: 1.5588, DPR: 0.3971
```

Example 3:

Sometimes it found the same corpus several times. The chunk has the answer to the question.

```
🔹 Retriever: BM25
Q: To whom did the Virgin Mary allegedly appear in 1858 in Lourdes France?
Gold Answer: Saint Bernadette Soubirous
Top 1 [17.6835]: Architecturally, the school has a Catholic character. Atop the M
Top 2 [17.6835]: Architecturally, the school has a Catholic character. Atop the M
Top 3 [17.6835]: Architecturally, the school has a Catholic character. Atop the M
```

This mostly happened in BM25, when the information was not readily available anywhere else in the whole dataset.

The code is shared here.

References

- Oche, A. J., Folashade, A. G., Ghosal, T., & Biswas, A. (2025). *A systematic review of key Retrieval-Augmented Generation (RAG) systems: Progress, gaps, and future directions*. arXiv preprint arXiv:2507.12345. https://arxiv.org/abs/2507.12345
- Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. T. (2020). *Dense passage retrieval for open-domain question answering*. arXiv:2004.04906. https://arxiv.org/abs/2004.04906
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... Riedel, S. (2020). *Retrieval-augmented generation for knowledge-intensive NLP tasks*.arXiv:2005.11401. https://arxiv.org/abs/2005.11401
- Karpukhin, V., Oguz, B., Min, S., et al. (2020). Dense Passage Retrieval for Open-Domain Question Answering. EMNLP 2020.