An Ablation Study of Retrieval Strategies for Stock-News RAG Systems

Om Chimurkar

2026

Abstract

Retrieval-Augmented Generation (RAG) systems rely heavily on retrieval quality to ground large language model outputs. In this work, we conduct a controlled ablation study comparing semantic retrieval, hybrid retrieval, and hybrid retrieval with Maximum Marginal Relevance (MMR) re-ranking on a stock-news question answering task. Using a fixed dataset snapshot and Gemini for both generation and automatic evaluation, we analyze retrieval relevance, grounding, and coverage. Our results show that hybrid retrieval improves coverage compared to semantic retrieval, while MMR further increases contextual diversity without significantly improving grounding. These findings highlight that retrieval diversity and grounding are distinct concerns in RAG system design.

1 Introduction

Large Language Models (LLMs) often produce confident but incorrect answers due to hallucinations and lack of access to up-to-date information. Retrieval-Augmented Generation (RAG) mitigates this limitation by conditioning generation on external documents [2]. However, retrieval quality remains a dominant failure mode in RAG systems, with poor recall, redundancy, and limited coverage leading to incomplete or misleading answers.

Hybrid retrieval approaches that combine semantic similarity with keyword-based retrieval improve recall but frequently retrieve redundant documents describing the same events. This motivates the use of diversity-aware re-ranking techniques such as Maximum Marginal Relevance (MMR) [1]. In this work, we investigate whether improving retrieval diversity through MMR leads to better RAG answer quality, particularly in terms of coverage, without directly improving grounding.

Contributions

The main contributions of this work are:

- A controlled ablation study comparing semantic, hybrid, and diversity-aware (MMR) retrieval strategies for stock-news RAG systems.
- An empirical analysis showing that retrieval diversity improves coverage but does not directly address grounding failures.
- A reproducible evaluation framework using a fixed dataset snapshot and LLM-based automatic scoring.

2 Methodology

2.1 Data Collection

Stock-related news articles were collected using a financial news API and frozen into a dataset snapshot prior to experimentation. This snapshot-based approach ensures reproducibility and prevents evaluation variance caused by changing external data sources.

2.2 Retrieval Strategies

We compare three retrieval strategies:

- **Semantic Retrieval**: Document retrieval using sentence embeddings and cosine similarity [3].
- **Hybrid Retrieval**: A combination of semantic similarity and BM25 keyword-based scoring [4].
- **Hybrid + MMR**: Hybrid retrieval followed by Maximum Marginal Relevance re-ranking to encourage diversity among retrieved documents [1].

MMR is applied strictly as a re-ranking step and does not modify the underlying retrieval scores.

2.3 Generation

Retrieved documents are concatenated into a context window and provided to a Gemini-based language model. The model is instructed to answer strictly using the provided context and to abstain when insufficient information is available.

2.4 Evaluation

Evaluation is performed using a Gemini-based LLM-as-a-judge framework. Generated answers are scored on three criteria:

- **Retrieval Relevance (0–2)**
- **Grounding (0–2)**
- **Coverage (0–2)**

3 Results

Hybrid retrieval consistently achieved higher coverage scores compared to semantic retrieval, indicating improved recall of relevant information. However, hybrid retrieval often produced redundant contexts containing multiple documents describing the same event.

Applying MMR re-ranking reduced redundancy and further improved coverage scores by promoting diversity among retrieved documents. Grounding scores remained largely unchanged across retrieval methods, suggesting that retrieval diversity alone does not substantially reduce hallucinations.

Table 1: Average Evaluation Scores by Retrieval Strategy

Retrieval Method	Relevance	Grounding	Coverage	Total (/6)
Semantic Retrieval	1.4	1.3	1.0	3.7
Hybrid Retrieval	1.6	1.3	1.4	4.3
Hybrid + MMR	1.6	1.3	**1.7**	**4.6**

3.1 Quantitative Results

4 Discussion

The results demonstrate that retrieval improvements primarily affect coverage rather than grounding. While hybrid retrieval increases recall, redundancy limits its effectiveness for multi-factor reasoning. MMR alleviates this issue by encouraging diversity, leading to more balanced contextual inputs.

However, grounding failures persist even with improved retrieval diversity. This indicates that hallucination appears to be more strongly influenced by generation behavior and prompt constraints than by retrieval strategy alone. Retrieval optimization and grounding enforcement should therefore be treated as complementary but distinct challenges in RAG system design.

5 Limitations

This study is limited by the size of the evaluation set and the use of free-tier API quotas, which constrained the number of experimental runs. Automatic evaluation using an LLM-as-a-judge introduces potential bias and variability in scoring. Additionally, this work focuses on retrieval strategies and does not explore advanced grounding mechanisms such as citation enforcement or constrained decoding.

6 Future Work

Future work includes expanding the evaluation dataset, incorporating human-in-the-loop evaluation, and comparing MMR against alternative re-ranking strategies. Integrating explicit grounding mechanisms and exploring retrieval-generation co-optimization may further reduce hallucinations. Applying this evaluation framework to other domains such as legal or medical document retrieval is another promising direction.

7 Conclusion

This work shows that hybrid and diversity-aware retrieval strategies improve contextual coverage in RAG systems but do not directly address grounding failures. Effective RAG system design therefore requires treating retrieval diversity and grounding as separate but complementary concerns.

Code Availability

The full implementation and experimental setup are available at: `https://github.com/Omc12/RAG-Evaluation---Ablation-Study`

Reproducibility

All experiments were conducted on a fixed dataset snapshot to ensure reproducibility. The retrieval, generation, and evaluation pipelines are deterministic given the same inputs, subject to external API constraints. Randomness was minimized by using fixed prompts and temperature settings during generation and evaluation.

References

[1] Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. *Proceedings of SIGIR*, 1998.

[2] Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 2020.

[3] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *Proceedings of EMNLP*, 2019.

[4] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 2009.