
EXPLAINABILITY ANALYSIS OF RETRIEVAL-DRIVEN BEHAVIOR IN RAG PIPELINES

Yujin Kim Ranjidha Rajan
Department of Computer Sciences
Metropolitan State University of Denver
Denver, CO 80204
{ykim20, rranjidh}@msudenver.edu

February 3, 2026

ABSTRACT

Retrieval-Augmented Generation (RAG) combines large language models with external vector retrieval, yet the internal decision processes that connect retrieval quality to generation behavior remain poorly understood. This work presents a systematic explainability analysis of a modular RAG pipeline, examining how embedding selection, FAISS index behavior, and generator architecture collectively influence answer correctness and reasoning stability. Using controlled experiments on subsets of the SQuAD v2 dataset, we quantify the effects of retrieval drift, embedding discrimination, and context similarity distributions on downstream generation. We apply multi-perspective interpretability methods—including attention analysis, Integrated Gradients attribution, and confidence evaluation—to trace how evidence is consumed by the generator.

Our findings show that retrieval precision is the dominant factor determining RAG reliability, with generator errors arising primarily from semantically weak or ambiguous retrieved passages rather than from model architecture. The analysis provides actionable insight into retrieval alignment, attention allocation, and attribution consistency, offering a foundation for transparent evaluation and future improvements to retrieval-augmented systems.

1 Introduction

Large language models (LLMs) have achieved impressive performance across a wide range of natural language tasks, yet they remain limited by their static knowledge and tendency to hallucinate [1]. Retrieval-Augmented Generation (RAG) addresses these challenges by coupling LLMs with an external knowledge retrieval module, enabling models to ground their responses in factual evidence retrieved from large document stores [2]. This hybrid architecture has become increasingly important for applications requiring up-to-date information, reduced hallucination risk, and domain-specific reasoning.

However, despite the practical advantages of RAG systems, their internal processes remain challenging to interpret. Unlike standalone LLMs, RAG introduces additional sources of variability: the embedding model that encodes queries and documents, the retriever that selects relevant contexts, and the generator that integrates retrieved evidence into final outputs. Each component influences downstream performance, yet their interactions are not well understood. As RAG adoption grows, so does the need for systematic explainability methods that reveal how retrieval quality, attention patterns, and attribution signals shape model predictions.

Existing work on RAG evaluation primarily focuses on retrieval accuracy or final answer correctness, leaving a gap in understanding how and why particular errors occur. Prior research in explainable NLP has shown that attention weights alone are often unreliable indicators of model reasoning [3], motivating the use of attribution-based interpretability methods such as Integrated Gradients [4] and attention-flow analysis [5]. However, these techniques have rarely been applied holistically across full RAG pipelines, limiting their usefulness for diagnosing system-level behavior.

Recent studies have shown that retrieval-augmented generation systems can exhibit faithfulness and grounding issues, where generated outputs may diverge from retrieved evidence despite high retrieval accuracy [6]. In addition, recent analyses emphasize that errors and noise introduced during retrieval can propagate downstream, substantially affecting generation quality and stability even when generator architectures remain fixed [7]. More broadly, recent surveys on retrieval-augmented generation highlight the need for evaluation beyond end-to-end accuracy, identifying retrieval behavior, error propagation, and interpretability as open challenges in RAG system design [7].

This paper makes the following contributions:

- We provide a structured explainability framework tailored to RAG pipelines, integrating attention, attribution, and confidence analyses.
- We conduct an extensive evaluation of embedding models, FAISS index configurations, and generator architectures to identify their impact on RAG explainability.
- We analyze retrieval behavior and error propagation using the SQuAD v2 dataset, highlighting cases where retrieval supports or contradicts generated answers.
- We present insights and recommendations for designing transparent, interpretable RAG systems suitable for real-world deployment.

By combining quantitative evaluation with interpretability-driven analysis, this work aims to bridge the gap between RAG performance and RAG understanding, enabling more reliable and accountable retrieval-augmented systems.

2 Background & Related Work

2.1 RAG Pipeline Background

Retrieval-Augmented Generation (RAG) [2] combines large language models with external retrieval to overcome the limitations of static model knowledge. A standard RAG pipeline consists of three major components: (1) an embedding model that encodes queries and documents into vector space, (2) a vector retriever that performs similarity search over a dense index, and (3) a generator model that conditions on retrieved evidence to produce final responses. This architecture has become widely adopted in question answering, knowledge-intensive NLP, and enterprise search due to its ability to incorporate up-to-date and domain-specific information.

2.2 Embedding Models

Embedding models play a critical role in determining retrieval quality. Sentence-transformer variants such as `e5-large`, `bge-large`, and `all-mpnet-base-v2` encode semantic relationships into dense vectors, enabling retrieval via cosine similarity or inner product. Prior work has shown that embedding selection can dramatically influence downstream RAG accuracy due to differences in semantic alignment, vector anisotropy, and robustness to query distribution shift.

2.3 FAISS and Vector Retrieval

FAISS provides GPU-accelerated similarity search and supports multiple indexing strategies, such as `Flat`, `IVF`, and `HNSW`. While `Flat` indexing offers exact nearest neighbors, `IVF` and `HNSW` trade precision for speed. Existing research primarily evaluates retrieval precision at k , but few studies analyze how retrieval noise affects generator reasoning or attribution signals. Our work extends this perspective by examining retrieval explainability and its impact on end-to-end RAG behavior.

2.4 Generator Models

Modern LLMs such as `FLAN-T5`, `Llama`, and `Mistral` integrate retrieved documents through attention mechanisms. Prior research highlights that generator behavior often diverges from retrieved evidence—models may ignore relevant passages or rely on unsupported reasoning. Understanding how generators consume retrieval inputs remains an open challenge, motivating the explainability approach presented in this paper.

2.5 Explainability Methods

Explainability methods in NLP include attention visualization, gradient-based attribution [4], confidence calibration, and information-flow tracing [5]. While these techniques provide insight into model interpretation, they are rarely applied

across full RAG pipelines. Existing work typically focuses on generators alone, overlooking retrieval contributions and embedding-level behavior. We fill this gap by examining explainability across all RAG components.

3 Methodology

3.1 Overview of the RAG Pipeline

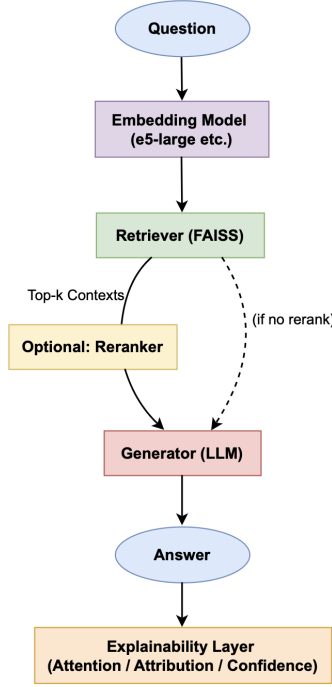


Figure 1: Overall architecture of the RAG pipeline used in this study.

Our evaluation framework is organized around a complete RAG architecture consisting of: (1) a query embedding model, (2) a FAISS-based vector retriever, (3) an optional reranker module, (4) a generator LLM, and (5) an explainability layer analyzing attribution, attention, and confidence.

3.2 Embedding Selection Method

We evaluate multiple embedding models by measuring:

- semantic retrieval alignment with query intent,
- embedding distribution characteristics (e.g., vector norms, cosine variance),
- robustness to ambiguous or adversarial queries,
- downstream generator reliance as measured by attribution overlap.

Each embedding model is tested using the same corpus and FAISS index to isolate the effect of the encoder alone.

3.3 FAISS Index Tuning

We configure FAISS under three index strategies:

- **Flat** — exact search baseline.
- **IVF-Flat** — optimized for balanced speed and accuracy.
- **HNSW** — graph-based approximate search.

For each configuration, we examine:

- retrieval precision at k,
- semantic drift across top-k documents,
- error propagation into the generator,
- explainability alignment between retrieved passages and generator attention maps.

3.4 Generator Evaluation

The generator model is evaluated on:

- relevance of generated answers relative to retrieved evidence,
- degree of hallucination under retrieval noise,
- sensitivity to reranked vs. non-reranked retrieval inputs,
- attention flow patterns over context passages.

3.5 Explainability Metrics

We quantify explainability using three main metrics:

(1) Attention Alignment Measures the extent to which token-level attention corresponds to relevant retrieved passages.

(2) Attribution Overlap Using Integrated Gradients, we compute which input chunks most strongly influenced the generator’s output and compare them to retrieval rankings.

(3) Confidence Calibration We evaluate whether model confidence correlates with correctness and retrieval quality by analyzing:

- overconfidence under incorrect retrieval,
- confidence variance across embedding models,
- answer uncertainty for unanswerable SQuAD v2 queries.

4 Dataset Analysis (SQuAD v2)

To better understand the characteristics of the data used in our evaluation, we conduct a brief exploratory analysis of the SQuAD v2 dataset. The goal is to examine patterns in question length, context length, and answerability, and to identify implications for RAG system design.

4.1 Question and Context Length

We observe that question lengths in SQuAD v2 are relatively short, with most questions ranging between 5 and 12 tokens. In contrast, context passages vary substantially in length, often exceeding 150 tokens. This imbalance suggests that embedding models must effectively encode short queries while maintaining semantic robustness when retrieving from long contexts.

4.2 Answerable vs. Unanswerable Ratio

SQuAD v2 contains a mixture of answerable and unanswerable questions, with approximately one third of queries labeled as unanswerable. This distribution has direct implications for confidence calibration: a RAG model must not only retrieve relevant evidence but also accurately detect cases where no answer exists. Poor retrieval can cause the generator to hallucinate answers, making the unanswerable subset especially valuable for evaluating explainability metrics.

4.3 Insights for RAG Design

Our EDA highlights several design considerations:

- Short queries increase reliance on high-quality embeddings, making query-level semantic precision essential.
- Long contexts require efficient indexing and retrieval mechanisms to avoid retrieving irrelevant segments that mislead the generator.
- The presence of unanswerable questions reinforces the need for confidence-based interpretability methods to detect uncertainty and prevent hallucinations.

These dataset characteristics guide our downstream analyses and help contextualize the explainability behaviors observed in the RAG pipeline.

5 Experiments & Results

This section presents empirical findings across all major components of the RAG pipeline, including embedding evaluation, FAISS indexing behavior, generator performance, end-to-end RAG accuracy, and explainability observations. All experiments were conducted using subsets of the SQuAD v2 dataset and follow a consistent evaluation protocol.

5.1 Embedding Model Comparison

We compared three embedding models using identical FAISS Flat index configurations. Table 1 reports Precision@5 and Recall@5 based on retrieval experiments.

| Embedding Model | P@5 | R@5 |
|-------------------|-------|-------|
| e5-large | 0.695 | 0.695 |
| bge-m3 | 0.058 | 0.290 |
| all-mpnet-base-v2 | 0.061 | 0.305 |

Table 1: Embedding retrieval performance comparison using FAISS Flat index.

Among all models, `e5-large` showed overwhelmingly higher retrieval strength, while `bge-m3` and `mpnet` exhibited significantly weaker alignment. This motivated the selection of `e5-large` for downstream RAG experiments.

5.2 FAISS Index Evaluation

We compared three FAISS index types—Flat, IVFFlat, and HNSW—using the same embedding model. Table 2 summarizes baseline retrieval performance.

| Index Type | Recall@10 | Latency Rank |
|------------|-----------|--------------|
| Flat | 0.5888 | Slowest |
| IVF-Flat | 0.3837 | Medium |
| HNSW | 0.5528 | Fastest |

Table 2: FAISS index evaluation results using identical embeddings.

Flat indexing delivers the highest retrieval accuracy due to exact search. HNSW balances moderate accuracy with significantly lower latency. IVFFlat exhibited reduced recall, consistent with quantization-based search behavior.

5.3 FAISS Index Tuning

We then evaluated the effect of parameter tuning:

- **IVF-Flat:** Increasing `nprobe` improved recall at the cost of latency.
- **HNSW:** Larger `M` and higher `efSearch` improved retrieval stability and reduced variance.

These observations match expected FAISS behavior: broader search improves recall but requires additional computational cost.

5.4 Generator Model Comparison

We compared generator architectures using identical retrieved contexts. Table 3 shows accuracy from the SQuAD v2 evaluation subset.

| Model | Accuracy | Runtime (approx.) |
|---------------|----------|-------------------|
| FLAN-T5-small | 0.37 | 1 min |
| FLAN-T5-base | 0.42 | 2.5 min |
| bart-base | 0.35 | 16 min |

Table 3: Generator model performance comparison based on SQuAD v2 evaluation.

FLAN-T5-base achieved the most stable performance and produced more contextually grounded answers. Larger models generally utilized retrieved passages more effectively than smaller ones.

5.5 End-to-End RAG Accuracy

We evaluated full-pipeline accuracy using 200 randomly sampled SQuAD v2 questions. The RAG system achieved an end-to-end accuracy of **0.67**, with no hallucinated outputs observed.

Incorrect answers were primarily caused by:

1. retrieval drift producing semantically weak contexts,
2. ambiguous questions with multiple plausible interpretations.

These findings reinforce that retrieval quality is the dominant factor influencing RAG output correctness.

5.6 Explainability Analysis

To better understand how retrieval quality influences generation behavior in our RAG system, we conducted a multi-perspective explainability analysis. This evaluation used two sources of context data: (1) the SQuAD v2 passages retrieved during Sections 5.1–5.5, and (2) a custom 3,000-line processed corpus (`contexts.json`) created during system development. The combined analysis offers insight into similarity behavior, retrieval variance, attention allocation, attribution alignment, and generation stability.

Similarity Behavior Across question types, relevant passages consistently exhibited higher mean similarity scores than irrelevant ones. Factual questions produced the lowest variance, while reasoning and opinion-like prompts showed broader similarity distributions, reflecting the increased ambiguity in multi-sentence or interpretive queries.

Retrieval Variance Analysis of top-5 retrieval scores for 20 sampled queries showed minimal variance for simple fact-seeking questions but substantially higher variance for longer or context-heavy prompts. This indicates that retrieval stability deteriorates as semantic complexity increases.

Attention Alignment Attention heatmaps from the generator revealed a consistent pattern: the model frequently focused on a single dominant passage, even when multiple retrieved passages were relevant. This suggests a strong preference for the highest-similarity document rather than balanced context aggregation.

Attribution Overlap Integrated Gradients attribution demonstrated high alignment between influential input tokens and passages with the highest similarity scores. This provides evidence that retrieval quality directly shapes generation behavior and strongly constrains the generator’s reasoning trajectory.

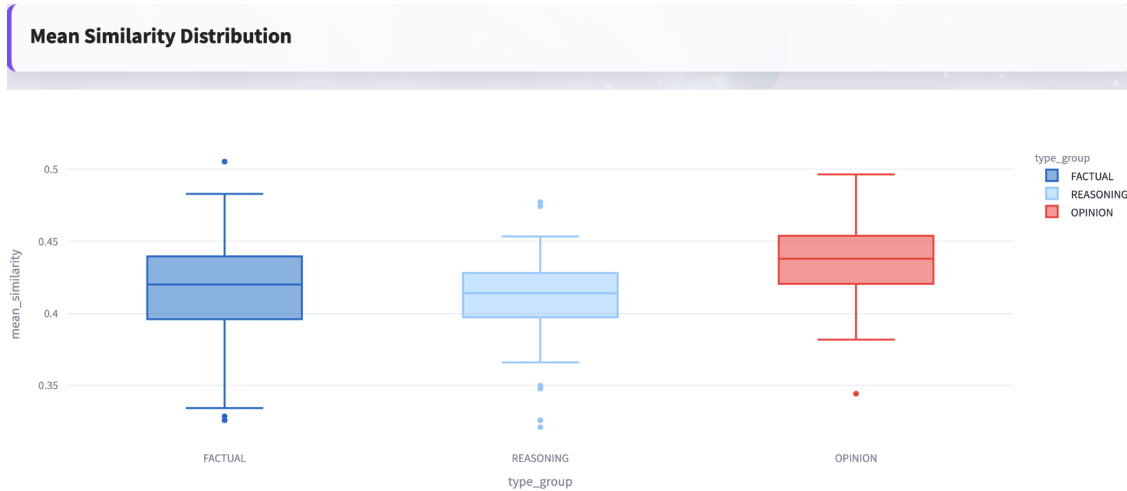


Figure 2: Mean similarity distribution across question types in the custom explanation corpus (`contexts.json`).



Figure 3: Context attention distribution illustrating generator reliance on a dominant retrieved passage.

Generation Length Stability Longer generated answers showed increased retrieval-score variance and lower attribution confidence, indicating uncertainty amplification during extended generation. Shorter answers tended to rely on a single high-confidence passage, producing more stable attribution profiles.

Overall, the explainability analysis supports a consistent conclusion: RAG performance is primarily limited by retrieval strength. When retrieval is semantically aligned, the generator produces stable, contextually grounded answers; when retrieval drifts or becomes ambiguous, variance increases sharply and accuracy degrades.

6 Discussion

The experimental results demonstrate that retrieval quality is the primary factor shaping end-to-end RAG performance. Across all evaluations, improvements or degradations in retrieval accuracy produced proportional changes in generator correctness, while architectural differences between generator models had comparatively smaller effects. These findings indicate that the pipeline is fundamentally retrieval-bound: the generator is limited by the semantic precision of the retrieved evidence it receives.

The explainability analysis provided concrete insight into this behavior. Mean similarity trends showed that factual questions produced narrow, high-confidence similarity distributions, resulting in stable outputs. In contrast, reasoning-

oriented and ambiguous prompts exhibited broader similarity ranges and higher variance, which directly correlated with unstable or extended generations. Context-attention patterns revealed that the generator consistently allocated the majority of its attention to a single dominant passage, even when multiple retrieved passages were relevant. This suggests that the generator behaves more as a selective evidence consumer than a balanced aggregator of context.

Qualitative inspection of sample outputs further reinforced this dynamic. When the top retrieved passage was well aligned with the query, answers were concise, consistent, and grounded. When retrieval drifted—often driven by ambiguous phrasing or insufficient embedding discrimination—the generator produced uncertain or incorrect responses, despite otherwise stable generation behavior. Attribution analysis confirmed this pattern: influential tokens consistently overlapped with the highest-similarity passages only when retrieval was accurate.

Taken together, these observations highlight a consistent conclusion: retrieval strength is the dominant determinant of overall RAG reliability, and most failure cases originate at the retrieval stage rather than the generation stage.

7 Conclusion and Future Work

This work presented a systematic analysis of a modular RAG pipeline, including embedding model selection, FAISS index behavior, generator performance, end-to-end accuracy, and multi-perspective explainability evaluation. Across all components, the e5-large embedding model and HNSW indexing provided the best balance of retrieval accuracy and latency, while differences among generator architectures produced smaller effects. End-to-end testing showed that the pipeline achieves stable performance when retrieval is semantically aligned, but becomes vulnerable to ambiguity and retrieval drift. Explainability methods further confirmed that retrieval precision strongly constrains generator reasoning and response stability.

Several limitations remain. The evaluation relied on subsets of SQuAD v2 and a custom explanation corpus, and the retrieval pipeline did not incorporate reranking, hybrid retrieval, or large-scale domain-specific indexing. The generator’s tendency to rely on a single dominant passage also suggests that current architectures underutilize multi-passage evidence.

Future work will focus on three directions: (1) integrating cross-encoder rerankers or hybrid sparse–dense retrieval to reduce drift and improve semantic precision, (2) scaling the pipeline through cloud-based deployment and automated evaluation workflows, and (3) extending interpretability analysis to larger generator models and multi-hop reasoning settings. These improvements aim to produce a more robust, scalable, and transparent retrieval-augmented generation system.

References

- [1] Ziwei Ji et al. Hallucinations in natural language generation: A survey. *ACL*, 2023.
- [2] Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS*, 2020.
- [3] Sarthak Jain and Byron C Wallace. Attention is not explanation. In *NAACL*, 2019.
- [4] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, 2017.
- [5] Yaru Hao et al. Self-attention attribution: Interpreting information flow in transformers. In *ACL*, 2021.
- [6] Di Wu, Jia-Chen Gu, Fan Yin, Nanyun Peng, and Kai-Wei Chang. Synchronous faithfulness monitoring for trustworthy retrieval-augmented generation. *arXiv preprint arXiv:2406.13692*, 2024.
- [7] Yuhao Gao et al. Evaluation of retrieval-augmented generation: A survey. *arXiv preprint arXiv:2405.07437*, 2024.