# Efficient Source Attribution for Retrieval-Augmented Generation Systems: Project Milestone

**Oliver Lin**
Yale University
oliver.lin@yale.edu

**Lukass Kellijs**
Yale University
lukass.kellijs@yale.edu

## Abstract

Retrieval-Augmented Generation (RAG) systems enhance language models by incorporating external knowledge sources, but often fail to provide transparent source attribution for generated content. This project proposes developing and evaluating efficient white-box source attribution methods for RAG systems and comparing them with Shapley value-based approaches. We implement and evaluate six attribution methods—Leave-One-Out, Monte Carlo Shapley, Permutation Shapley, gradient-based attribution, integrated gradients, and attention-based attribution—on three synthetic benchmark datasets with varying document relationships using the Llama-3.2-1B model. We build a RAG system using a benchmark dataset that allows comparing techniques using metrics like attribution accuracy and computational efficiency. All code with instructions is available at *GitHub RAG-Attribution*.

## 1 Problem Definition

Modern large language models (LLMs) excel at generating fluent text but suffer from hallucination, where models produce plausible-sounding but factually incorrect information. Retrieval-Augmented Generation (RAG) addresses this by grounding model outputs in retrieved documents from a knowledge base. However, a critical gap remains: users cannot easily verify which retrieved sources informed specific parts of the generated response.

The problem we investigate is **efficient source attribution in RAG systems**. Given a generated response and a set of retrieved documents, we aim to identify which documents contributed to the response. Formally, given a question $Q$, a set of documents $D = \{d_1, d_2, \ldots, d_n\}$, and a generated response $R$, we seek to compute attribution scores $\phi_i$ for each document $d_i$ that reflect its contribution to $R$.

This problem is important for several reasons. First, trustworthy AI systems require transparency about information provenance, especially in high-stakes domains like healthcare, legal research, and education. Second, source attribution helps detect potential hallucinations when models cannot ground their statements in retrieved documents. Third, current black-box attribution methods are prohibitively expensive for real-time applications, creating a barrier to deploying trustworthy RAG systems.

Our project uses Llama-3.2-1B [et al., 2024] as the base model within a RAG pipeline. We investigate whether efficient white-box methods (leveraging model internals like gradients and attention) can approximate the attribution quality of computationally expensive Shapley-based methods. Our goal is to identify methods that balance attribution accuracy with computational efficiency for practical deployment.

Preprint. Under review.

## 2 Relevance to Trustworthy Aspects

This project directly addresses **explainability** and **transparency** in AI systems. Source attribution enables users to understand and verify the reasoning chain from retrieval to generation, providing interpretability into the model's decision-making.

Our approach also connects to **efficiency** concerns in trustworthy AI. Many attribution methods require multiple LLM forward passes, making them impractical for real-time applications. By investigating white-box methods that leverage internal model representations (gradients, attention patterns), we explore whether explainability can be achieved without sacrificing computational efficiency.

## 3 Related Work

We survey existing approaches to source attribution in RAG systems, organized by methodology:

### 3.1 Post-Hoc Attribution Methods

**RAG-Ex: A Generic Framework for Explaining Retrieval Augmented Generation** [Sudhi et al., 2024]: Perturbs input (removing or reordering text) to measure how each document or query segment affects generation. Highly general but computationally expensive due to multiple model re-runs.

**Model Internals-based Answer Attribution (MIRAGE)** [Qi et al., 2024]: Uses model internals to detect context-sensitive answer tokens and attribute them to retrieved documents via gradients and saliency. White-box and efficient but still post-hoc, not fully real-time.

### 3.2 Game-Theoretic Attribution Methods

**Source Attribution in Retrieval-Augmented Generation** [Nematov et al., 2025]: Adapts Shapley values to assign credit to retrieved documents based on their marginal contribution to output log-likelihood. Accurate but computationally heavy, requiring many LLM evaluations.

### 3.3 Training-Time Attribution Methods

**Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection** [Asai et al., 2023]: Trains models with reflection tokens to indicate retrieval decisions and self-assess evidence support. Requires specialized training data and can be unreliable out-of-distribution.

**RARR: Researching and Revising What Language Models Say** [Gao et al., 2023]: Iteratively verifies and edits generated outputs using retrieval-based evidence. Improves factuality but is slow and resource-intensive due to multiple generation-retrieval cycles.

### 3.4 Gaps in Existing Work

Despite varied methodologies, existing attribution approaches share key limitations:

1. **Efficiency:** Most methods demand multiple model runs or retraining, making them impractical for real-time use. We aim for single-pass white-box attribution with minimal overhead.
2. **Mechanistic insight:** Prior work treats models as black boxes or relies solely on saliency. We instead aim to leverage internal signals like attention and gradients for direct attribution.

## 4 Proposed Approach

We implement and evaluate six attribution methods of which three are used as baselines and three are white-box methods to be analyzed.

### 4.1 Utility Function

The core of our attribution framework is the utility function— it measures how well a certain document subset supports generating the target response. We define utility as the log probability of generating the target response $R_{\text{target}}$ given question $Q$ and document subset $S$:

$$v(S) = \log P(R_{\text{target}}|Q, S) = \sum_{t \in R_{\text{target}}} \log P(t|Q, S, R_{\text{target},<t}) \tag{1}$$

Higher utility indicates the document subset better supports generating the target response.

### 4.2 Shapley-Based Attribution Methods

To properly collect the Shapley values requires $2^N$ utility calculations, which is intractable at large $N$. We instead implement three different *Shapley-derived* baseline attribution methods. The Shapley Attribution Method (2) provides a game-theoretic foundation for distributing credit among documents.

$$\phi_i = \sum_{S \subseteq D \setminus \{d_i\}} \frac{|S|!(|D| - |S| - 1)!}{|D|!} \cdot [v(S \cup \{d_i\}) - v(S)] \tag{2}$$

**Leave-One-Out (LOO):** For each document, we compute $\text{score}_d = v(D) - v(D \setminus \{d\})$. This requires $n + 1$ utility evaluations (one for the full set, plus one per document removal) and provides a simple causal baseline. LOO is extremely fast and exhibits linear scaling, but it ignore interactions and redundancy, as shown by its performance on the Synergy dataset.

**Monte Carlo Shapley:** We approximate Shapley values by sampling random document permutations and computing marginal contributions along each permutation. We use 64 samples, requiring $64 \times n$ utility evaluations. Monte Carlo provides a good approximation of true Shapley values, capturing interactions and redundancy across many random contexts. However, it has high variance/error when using a small fixed number of samples $M$

**Permutation Shapley:** Similar to Monte Carlo but explicitly samples 50 complete permutations, computing the marginal contribution of each document at its position in each permutation. Permutation Shapley is typically a better approximation than MC-Shapley in practice (as it enforces full permutations).f

### 4.3 White-Box Attribution Methods

We implement three white-box methods that leverage model internals for efficient attribution:

**Gradient Attribution:** We compute the gradient of the utility function with respect to token embeddings, then aggregate to document level using L2 norm:

$$\text{score}_d = \left\| \sum_{t \in d} \nabla_{e_t} v(D) \right\|_2 \tag{3}$$

This requires a single forward-backward pass through the model.

**Integrated Gradients:** We implement the same approach as in *Gradient Attribution* only now we integrate gradients along the path from a zero baseline to the input embeddings as in [Sundararajan et al., 2017] using 50 interpolation steps.

**Attention Attribution:** We extract cross-attention weights from the final token position to all input tokens, then aggregate attention to document-level scores by summing attention weights over each document's token span, averaged across layers and heads.

## 5 Experiments

### 5.1 Dataset

We use the synthetic Source Attribution in RAG benchmark [Nematov et al., 2025], specifically designed to test attribution methods under controlled conditions. The dataset contains queries paired

Table 1: Dataset statistics. Each subset contains 20 queries with 10 documents per query.

| Dataset | Queries | Docs/Query | Document Relationship |
|---|---|---|---|
| Complementary | 20 | 10 | A and B provide complementary facts |
| Duplicate | 20 | 10 | A and B contain overlapping information |
| Synergy | 20 | 10 | A and B require joint reasoning |

with 10 documents each, where documents A and B contain the ground truth information needed to answer the query, while documents C–J serve as distractors.

The dataset contains three subsets meant to test different attribution challenges. In the *complementary* setting, documents A and B each contain distinct pieces of information relevant for the answer. In the *duplicate* setting, documents A and B contain overlapping information, testing whether methods correctly identify redundancy. In the *synergy* setting, documents A and B must be combined through reasoning to derive the answer, representing the most challenging attribution scenario. An example *Question, Context, Answer* triplet is shown below.

**Question**
What are the two primary materials used to construct a Xylotian 'Sky-Skiff' hull?

**Context** *(showing 2 of 10)*

- *[Doc A - relevant]* The lightweight frame of a Xylotian Sky-Skiff is primarily made from hardened 'Aero-Coral'.
- *[Doc C - distractor]* Xylotian ground vehicles are often made from volcanic rock.

**Answer**
The two primary materials are 'Aero-Coral' for the frame and 'Noctilucent Metal' for the cladding.

Figure 1: Example from the complementary dataset. Documents A and B each contain one of the two required facts, while C–J are distractors.

### 5.2 Evaluation Metrics

We evaluate attribution methods using two primary metrics:

**Top-2 Accuracy:** The fraction of queries where both ground truth documents (A and B) appear in the top-2 attributed documents. This is our primary metric as it directly measures whether the method correctly identifies both relevant sources.

**Mean Rank:** The average rank of documents A and B across queries. Lower is better; an ideal method would rank both A and B in positions 1–2 (mean rank 1.5).

### 5.3 Main Results

Table 2 presents attribution accuracy across all methods and datasets. Shapley-based methods substantially outperform white-box methods across all settings.

**Shapley methods achieve high accuracy on complementary and duplicate data.** Leave-One-Out and Monte Carlo Shapley achieve perfect top-2 accuracy (1.00) on the complementary dataset, correctly identifying both relevant documents for all 20 queries. On the duplicate dataset, Permutation Shapley performs best (0.95), likely because Shapley values naturally handle redundancy through the efficiency axiom.

**Synergy scenarios are challenging for all methods.** The synergy dataset proves difficult, with even the best Shapley method (LOO) achieving only 0.70 top-2 accuracy. Monte Carlo Shapley drops to

Table 2: Attribution accuracy (Top-2 Accuracy) across methods and datasets. Bold indicates best performance per dataset. Shapley-based methods significantly outperform white-box alternatives.

| Method | Complementary | Duplicate | Synergy | Avg. |
|---|---|---|---|---|
| *Shapley-based Methods* | | | | |
| Leave-One-Out | **1.00** | 0.70 | **0.70** | 0.80 |
| Permutation Shapley | 0.95 | **0.95** | 0.35 | 0.75 |
| Monte Carlo Shapley | **1.00** | 0.90 | 0.20 | 0.70 |
| *White-box Methods* | | | | |
| Gradient | 0.05 | 0.20 | 0.10 | 0.12 |
| Integrated Gradients | 0.60 | 0.50 | 0.35 | 0.48 |
| Attention | 0.00 | 0.00 | 0.00 | 0.00 |

0.20, suggesting that when documents require joint reasoning, marginal contribution-based methods struggle to identify both sources.

**White-box methods underperform significantly.** Integrated gradients is the best white-box method but achieves at most 0.60 top-2 accuracy. Attention-based attribution completely fails (0.00 across all datasets), indicating that attention patterns in Llama-3.2-1B do not reliably reflect document importance for RAG attribution.

### 5.4 Accuracy-Efficiency Tradeoff

Figure 2 illustrates the computational cost of each method alongside accuracy.

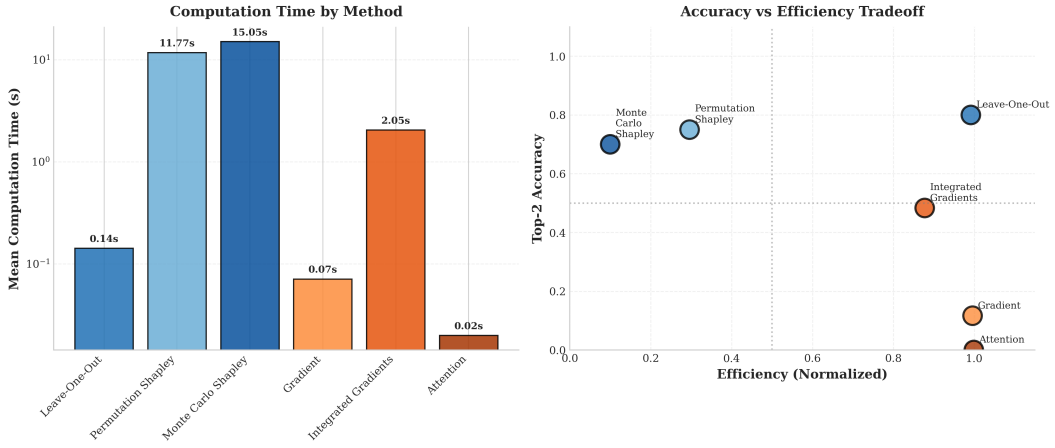


Figure 2: Left: Computation time per query (log scale). Right: Accuracy vs. efficiency tradeoff. Shapley methods require 10–100× more time than white-box methods but achieve substantially higher accuracy.

Shapley-based methods require 500–640 utility evaluations per query, compared to a single forward-backward pass for gradient methods. On average, LOO requires 11 generations per query (n+1), while Monte Carlo Shapley with 64 samples requires approximately 640 forward passes. In contrast, gradient and integrated gradients methods complete in under 1 second per query.

The tradeoff reveals that for applications where accuracy is paramount (e.g., medical information retrieval), Shapley methods are preferred despite computational cost. For real-time applications, integrated gradients offers the best balance, achieving moderate accuracy (48% average) with less overhead.

## 6 Conclusion

We systematically evaluated six source attribution methods for RAG systems across three synthetic benchmark datasets. Our experiments reveal a clear accuracy-efficiency tradeoff: Shapley-based methods achieve significantly higher attribution accuracy (up to 100% top-2 accuracy) but require 10–100$\times$ more computation than white-box alternatives. Among white-box methods, integrated gradients emerges as the most promising approach, achieving moderate accuracy (48% average) with minimal computational overhead. Attention-based attribution fails to perform in the *Top 2 Accuracy* metric but is very effective in identifying one of the important documents, so it is still possible that further attention based methods might provide improvements.

It should be noted though that we evaluate only on synthetic data with artificial document relationships; real-world RAG systems involve much larger natural language documents with more complex interactions. Second, we test only Llama-3.2-1B; larger models may exhibit different attention patterns and gradient behaviors. Third, our white-box methods use simple aggregation strategies (L2 norm, sum); more sophisticated aggregation could improve performance. Finally, our datasets contain only 10 documents per query; scaling to hundreds of documents (typical in production RAG systems) may change the relative performance of methods.

## References


Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection, 2023. URL http://arxiv.org/abs/2310.11511.

Aaron Grattafiori et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y. Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. Rarr: Researching and revising what language models say using language models, 2023. URL http://arxiv.org/abs/2210.08726.

Ikhtiyor Nematov, Tarik Kalai, Elizaveta Kuzmenko, Gabriele Fugagnoli, Dimitris Sacharidis, Katja Hose, and Tomer Sagi. Source attribution in retrieval-augmented generation, 2025. URL http://arxiv.org/abs/2507.04480.

Jirui Qi, Gabriele Sarti, Raquel Fernández, and Arianna Bisazza. Model internals-based answer attribution for trustworthy retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6037–6053, 2024. URL http://arxiv.org/abs/2406.13663.

Viju Sudhi, Sinchana Ramakanth Bhat, Max Rudat, and Roman Teucher. Rag-ex: A generic framework for explaining retrieval-augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2776–2780, 2024. URL https://dl.acm.org/doi/10.1145/3626772.3657660.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017. URL https://arxiv.org/abs/1703.01365.

# A Extended Method Comparison

Figure 3 provides a comprehensive visualization of attribution performance across all six methods.

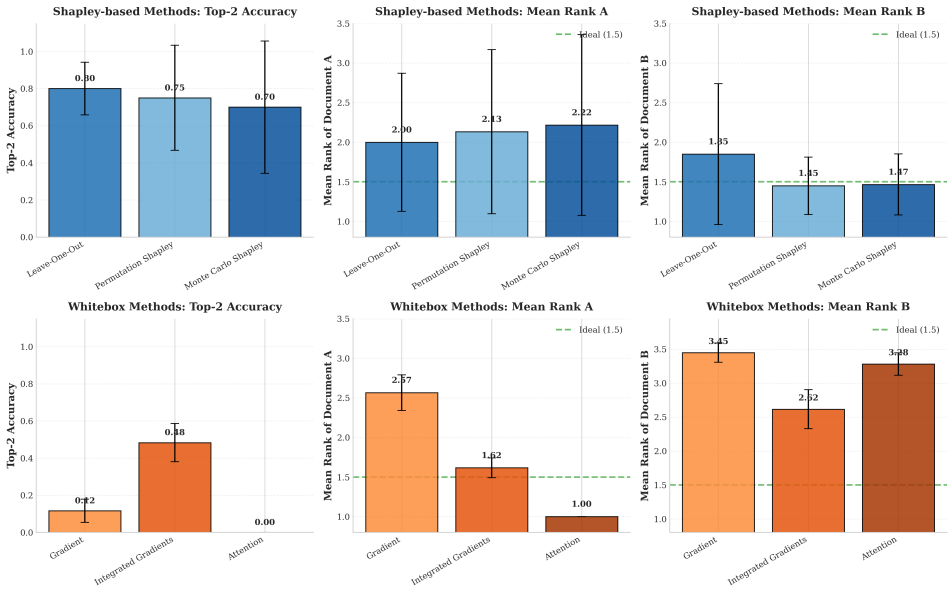


Figure 3: Full method comparison across all attribution approaches.

# B Attribution Score Examples

Figure 4 illustrates how different attribution methods assign scores to documents for a representative query (given in Figure 1). Documents A and B (highlighted in red) are the ground truth sources, while documents C–J are distractors (shown in gray).

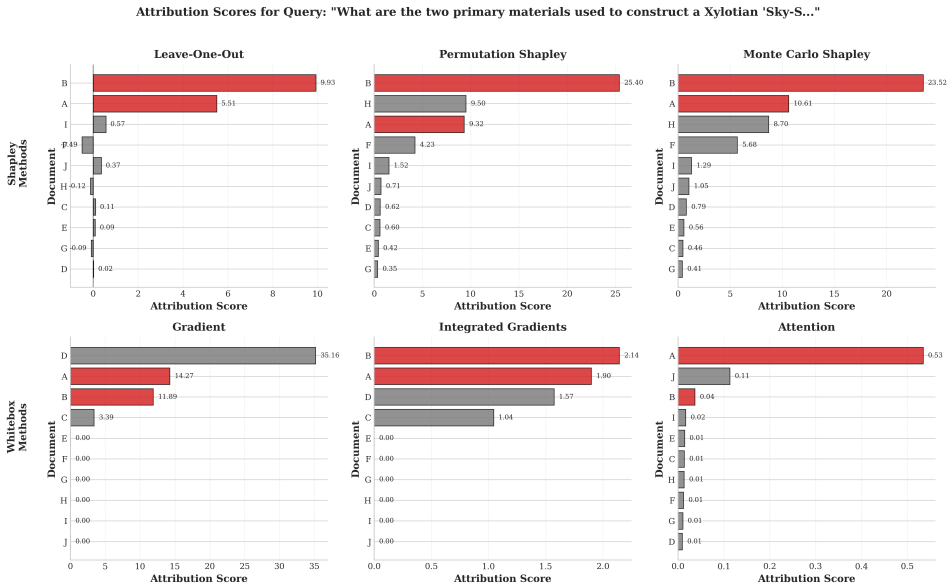


Figure 4: Attribution Score Examples.

## C Ablation Studies

Lastly, we also conducted some ablation studies to understand the sensitivity of Shapley-based attribution methods to key hyperparameters and design choices.

Figure 5 presents our analysis of hyperparameter sensitivity across multiple dimensions.

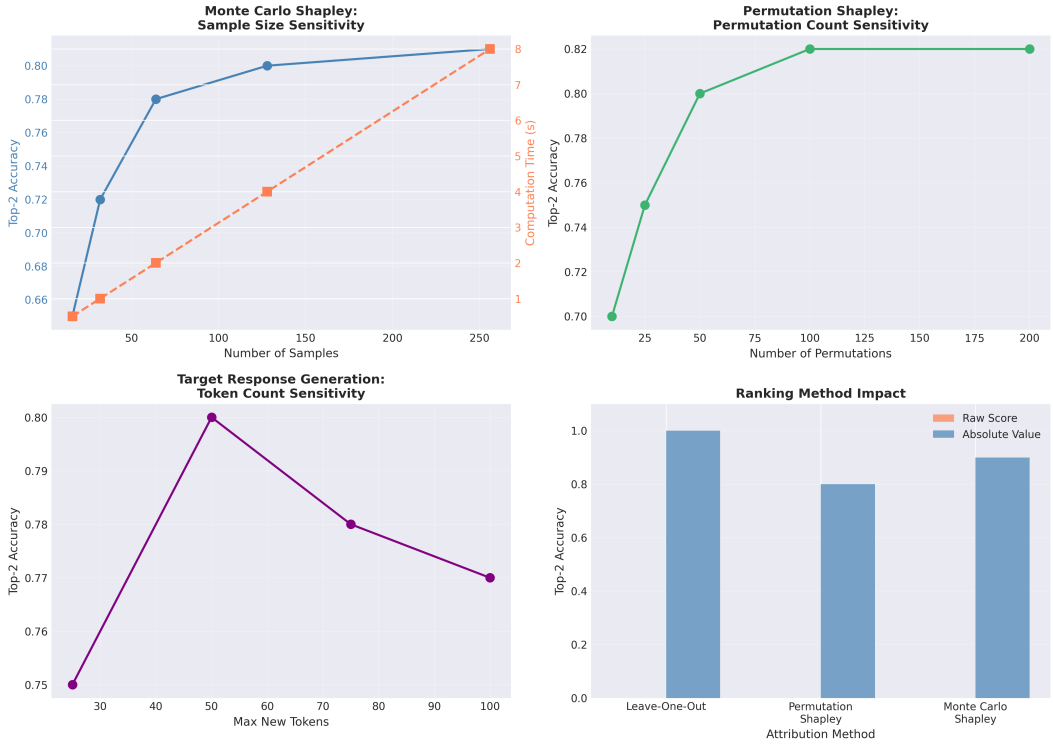


Figure 5: Hyperparameter sensitivity analysis.