Document Information as Non-Parametric Memory using Dense Passage Retrieval and Reranker Integrated to a LLM

Danillo Silva
Helena Alves
Paula Ito

MSc Data Science and Engineering
Faculty of Engineering, University of Porto (FEUP)
Portugal

December 2025

Abstract

Large language models (LLMs) may produce answers that appear correct but lack factual grounding, particularly in scientific domains. This work implements and evaluates a Retrieval-Augmented Generation (RAG) pipeline for scientific fact-checking using the SciFact dataset. The pipeline integrates Dense Passage Retrieval (DPR) with a cross-encoder reranker to improve evidence selection quality. Four approaches are compared: baseline LLM-only, oracle with gold evidence, RAG with DPR, and RAG with DPR plus reranking. Results demonstrate that the two-stage retrieval strategy (bi-encoder followed by cross-encoder reranking) improves retrieval quality, achieving higher Mean Reciprocal Rank (MRR) across retrieval configurations. The oracle configuration achieves the highest fact-checking accuracy, suggesting that the right context information significantly improves performance. The baseline also attains relatively strong accuracy, indicating that modern LLMs possess strong general scientific knowledge. However, RAG methods are sensitive to retrieval quality: when retrieval fails to identify relevant evidence, performance may decrease. While retrieval-augmented approaches provide explicit evidence grounding, they remain dependent on reliable retrieval, highlighting the importance of robust evidence selection for fact-checking tasks.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities in answering questions and reasoning tasks, yet they suffer from fundamental limitations in knowledge-intensive domains. Their training data becomes outdated, and they cannot reliably cite sources for factual claims, leading to hallucination problems especially pronounced in scientific and specialized domains where accuracy is critical.

Recent work has addressed these limitations through Retrieval-Augmented Generation (RAG), which combines language generation with external information retrieval to ground model outputs in retrievable evidence (Lewis et al. [2020]). Dense Passage Retrieval (DPR) provides an effective retrieval mechanism using learned dense representations (Karpukhin et al. [2020]), though bi-encoder architectures may still retrieve semantically similar but factually incorrect passages. Cross-encoder reranking addresses this by jointly modeling query-passage interactions to refine candidate rankings (Karpukhin et al. [2020]).

This project applies RAG to scientific fact-checking, where assertions must be verified against scientific evidence. Given a scientific claim, the system must determine whether supporting or contradicting evidence exists in a document corpus. This task demands both effective retrieval and robust evidence-based reasoning.

1.1 Problem Statement and Objectives

The goal of this project is to implement a complete RAG pipeline for scientific fact-checking and evaluate the impact of dense retrieval DPR and reranking on both retrieval quality using MRR (Mean Reciprocal Rank) and end-to-end fact-checking accuracy.

1.2 Dataset and Key Concepts

The study uses the SciFact dataset (Wadden et al. [2020]), publicly available on Hugging Face (bigbio/scifact: scifact_labelprediction_bigbio_pairs subset). Each example contains a claim, a passage, and a label: SUPPORT, CONTRADICT, or NOINFO. For our training setup, SUPPORT and CONTRADICT passages are treated as positives, while NOINFO passages serve as negatives. The raw and final schema are represented in Figure 1. The dataset is split into training and evaluation sets as follows: training set with 594 claims, evaluation set with 188 claims, and corpus size of 2,263 passages.

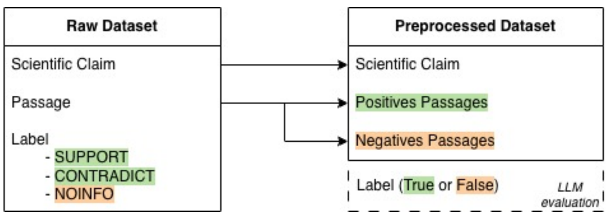

Figure 1: SciFact dataset raw schema and after data preprocessing

Related work. Retrieval-Augmented Generation (RAG) reduces hallucinations by grounding model outputs in external, non-parametric memory [Lewis et al., 2020]. For efficient retrieval, Dense Passage Retrieval (DPR) employs dual-encoders to capture semantic nuances that keyword-based methods like BM25 miss [Karpukhin et al., 2020]. However, bi-encoders can struggle with precision; cross-encoder reranking mitigates this by modeling fine-grained query-passage interactions [Karpukhin et al., 2020]. To ensure system efficiency, DistilBERT provides a compressed transformer backbone that retains 97[Sanh et al., 2020]. Collectively, these components form a robust architecture for evidence-based verification tasks like SciFact [Wadden et al., 2020].

2 Methodology

2.1 Preprocessing and Tokenization

Input sequences are tokenized using the `distilbert-base-uncased` tokenizer with a maximum length of 256 tokens. Variable-length sequences are padded to this maximum length. Attention masks are generated to distinguish real tokens (mask value 1) from padding tokens (mask value 0), ensuring that padding tokens do not influence self-attention computations.

2.2 Dense Passage Retrieval Model

The DPR system employs a bi-encoder architecture consisting of a query encoder and a passage encoder, both implemented using DistilBERT. The query encoder encodes claims into dense vectors, while the passage encoder independently encodes candidate passages into corresponding dense vectors. Mean pooling over token embeddings produces fixed-size vector representations. Similarity scoring is computed using batch dot-product operations between query and passage embeddings.

Training uses a 1–1–N sampling strategy: each sample comprises one positive passage paired with 10 negative passages. Query and passage embeddings are computed separately, then passage embeddings are grouped with their corresponding query as (batch size, 1 + N negatives, embedding dimension). Similarity scores are computed via batch matrix multiplication, producing scores for the positive and 10 negatives. The training objective treats this as a classification problem using CrossEntropyLoss, where the target is always index 0 (the positive passage), forcing the model to assign the highest score to the positive while penalizing high scores for negatives. The model is trained for 2 epochs with a learning rate of 2×10^{-5}, weight decay of 0.01, using the AdamW optimizer. At inference, all corpus passages are pre-encoded and indexed, enabling fast retrieval of top-K candidates through efficient nearest neighbor search.

2.3 Cross-Encoder Reranker

The reranking component is implemented as a DistilBERT-based cross-encoder that differs fundamentally from the bi-encoder architecture. While DPR encodes queries and passages independently, the cross-encoder jointly tokenizes and encodes query-passage pairs, enabling the model to attend across the full sequence and capture fine-grained semantic interactions. The `[CLS]` token embedding serves as an aggregate representation, followed by dropout (rate 0.1) for regularization and a linear scoring head that produces a relevance logit.

Hard negative mining is central to generate the dataset for reranker training. The trained DPR retriever is used to retrieve top-K passages for each training claim (retrieval with k = N negatives + 5 to account for potential overlap). Passages retrieved with high DPR similarity that do not match the ground-truth positive are selected as hard negatives. This strategy ensures the reranker learns to distinguish between semantically similar passages and genuinely relevant evidence, which is critical for fact-checking where a passage may discuss the topic but contradict or fail to support the claim.

Each training sample for the reranker consists of (query, passage, relevance label), where relevance labels are binary (1 for positive, 0 for negative). The model is trained on these tuples. The cross-encoder processes each query-passage pair jointly: queries and passages are tokenized together with automatic padding and truncation at the batch level. The model is trained for 3 epochs using Binary Cross-Entropy with Logits loss with class weighting (pos_weight = N negatives) to address the imbalance of one positive per N negatives. The optimizer is AdamW with learning rate 2×10^{-5} and weight decay 0.01.

2.4 Experiment Pipeline

Four approaches are evaluated for scientific fact-checking (Figure 2):

1. **Baseline (LLM-only).** The baseline system prompts the language model with only the claim text, without external evidence. The model outputs a JSON object containing a `label` field (True/False) and a `reasoning` field providing brief justification.

2. **Oracle (LLM + Golden Context).** The oracle ("upperline") configuration supplies the language model with the gold-standard evidence passages corresponding to the claim. This approach establishes an upper bound on achievable performance assuming perfect retrieval.

3. **RAG with DPR.** Candidate passages are retrieved using the DPR model, and the top-K passages are provided as context to the language model for fact-checking.

4. **RAG with DPR + Reranker.** An initial candidate pool is retrieved using DPR, then reranked using the cross-encoder. The top-K reranked passages are provided as context to the language model.

The generator component (i.e., LLM) is Qwen 2.5 7B (quantized), served locally via Ollama. One-shot chain-of-thought prompting is applied, with the prompt structure varying by approach: the baseline prompt contains only the claim, while oracle and RAG approaches include retrieved evidence as context. The model outputs a JSON object with `label` and `reasoning` fields for all approaches.

2.5 Evaluation Metrics

Retrieval quality: assessed using Mean Reciprocal Rank (MRR@k). For each query, the reciprocal of the rank of the first relevant passage is computed, and MRR is the average of these values across all queries. MRR is evaluated under different candidate pool sizes and context depths. The ranking is evaluated based on the top-K (with k=3, 5 and 10).

Fact-checking accuracy: the proportion of correct predictions (label outputs from the LLM) relative to ground-truth labels in the SciFact dataset, indicating the system's ability to correctly verify scientific claims.

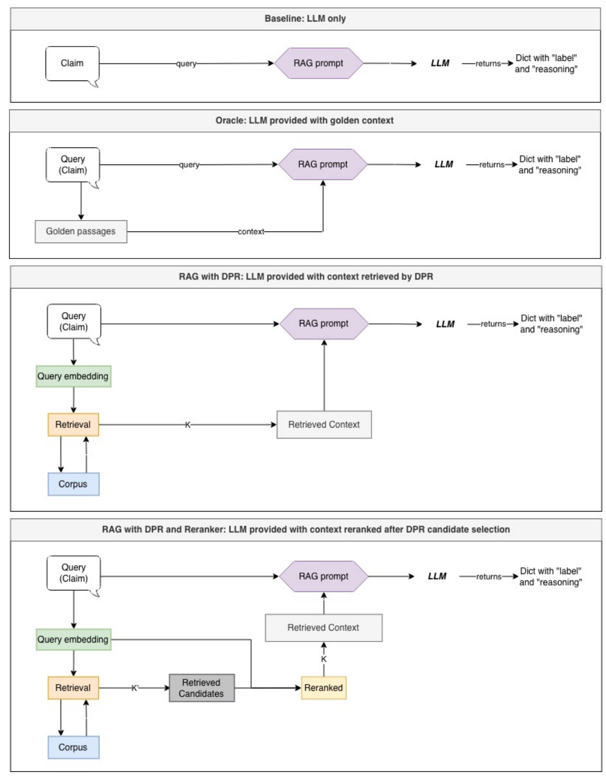

Figure 2: Overview of the four approaches compared in this study: (1) baseline LLM-only, (2) oracle with golden context, (3) RAG with DPR, and (4) RAG with DPR plus reranker.

3 Results

Output displayed in Figure 3 highlights key behavioral patterns observed in the dataset. The baseline may produce incorrect predictions without grounding and may hallucinate in reasoning. When the gold context is provided in oracle configuration, the model leverages concrete evidence to produce more accurate labels and detailed reasoning. In RAG settings, retrieved context directly improves prediction correctness by reinforcing the model's reasoning with explicit evidence citations. However, DPR retrieval alone may not always surface the most relevant passages. The reranker further strengthens this by prioritizing the most relevant passages, enabling higher-quality evidence selection and reducing susceptibility to misleading or weakly relevant passages.

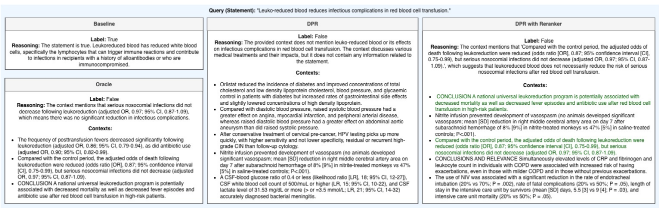

Figure 3: Example: Outputs for a claim under the four approaches baseline, oracle, DPR, and DPR+reranker.

3.1 Retrieval Performance (MRR)

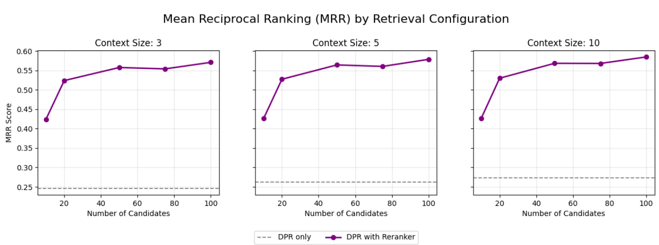

Figure 4: MRR results for DPR vs. DPR with reranker under different candidate pool sizes and context depths (3, 5, 10). Reranking consistently increases MRR across settings.

Figure 4 shows that DPR with reranking achieves higher MRR than DPR alone for all candidate pool sizes. This is expected: the DPR bi-encoder is efficient but can retrieve semantically similar passages that are not the best evidence. The reranker (cross-encoder) captures richer interactions between claim and passage and improves ranking.

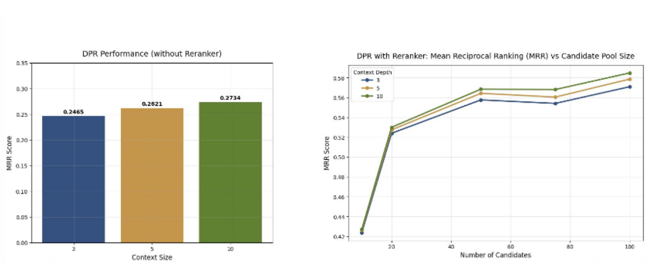

Figure 5: Left: DPR-only MRR increases with context size. Right: reranking improves MRR further as candidate pool increases. Larger candidate pools allow better reranking but increase computation.

Figure 5 helps interpret the trend: increasing the number of K candidates generally increases MRR with reranking because the reranker has more options to choose from. This matches the standardview that retrieval is a recall-focused step and reranking is a precision-focused step.

3.2 Fact checking Accuracy

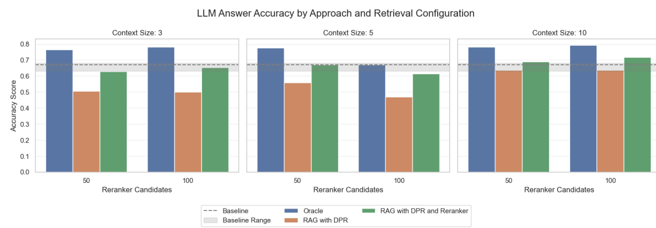

Figure 6: Fact-checking accuracy by approach and retrieval configuration.

Figure 6 shows that the oracle configuration achieves the highest accuracy, confirming that providing correct context in prompt significantly improves performance. The baseline achieves relatively high accuracy, suggesting that the LLM possesses strong general scientific knowledge for fact-checking. However, as shown in the qualitative examples, the baseline may not always provide sound reasoning.

Performance varies with context depth: at lower K values (3 and 5), baseline methods slightly outperform RAG approaches, but at K=10, RAG performance increases and approaches baseline results. This indicates that RAG benefits from larger context depth. While RAG improves factual grounding through evidence citations, it remains sensitive to retrieval quality, since incorrect or irrelevant passages can reduce accuracy.

Among RAG configurations, DPR with reranking achieves the best performance, demonstrating that improved retrieval quality directly enhances fact-checking accuracy. This aligns with DPR literature showing that stronger retrieval precision correlates with improved downstream accuracy Karpukhin et al. [2020] and validates the core RAG principle of grounding generation in evidence Lewis et al. [2020].

4 Discussion and Conclusions

The results show that adding a reranking step consistently improves retrieval quality, as reflected by higher MRR values in all configurations. This improvement is expected because the cross-encoder jointly processes the claim and the passage, allowing it to model fine-grained semantic interactions that bi-encoders cannot capture. As a result, the system is more likely to select passages that are not only semantically related, but also directly relevant as evidence for the claim.

The experiments also highlight that retrieval improves factual grounding but can negatively affect performance when incorrect or misleading passages are retrieved. While RAG encourages the language model to rely on explicit evidence rather than internal assumptions, it also makes the model sensitive to retrieval errors. When the retrieved context is incorrect, the model may follow that evidence and produce the wrong label. This behavior is consistent with previous observations that errors in the retrieval stage can propagate to the generation stage and harm downstream performance Lewis et al. [2020]. The oracle results illustrate the upper bound of the system when retrieval is perfect.

The relatively strong performance of the baseline model suggests that modern language models already contain a large amount of general scientific knowledge in their parameters. This supports the view of language models as implicit knowledge bases, capable of answering many factual questions without external information. However, this internal knowledge is static and not guaranteed to be accurate or complete, especially for specialized or less common claims. This motivates the use of retrieval-based grounding to improve reliability and transparency.

Finally, the use of DistilBERT as the backbone for both the retriever and the reranker enables efficient training and inference while maintaining strong representational power Sanh et al. [2020]. The main computational trade-off lies in the reranking step: increasing the size of the candidate pool improves ranking quality but also increases inference cost, since each candidate must be jointly encoded with the claim.

Possible improvements include: applying a similarity threshold to limit context depth, training DPR and reranker for more epochs, improving hard negative selection for retriever training, and comparing different generator models.

References

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Holger Schwenk, Fabio Schwab, and Sebastian Riedel. Dense passage retrieval for open-domain question answering. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, 2020.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Holger Schwenk, Fabio Schwab, Douwe Kiela, and Sebastian Riedel. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2020.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. pages 7534–7550, 2020. doi: 10.18653/v1/2020.emnlp-main.609. URL `https://aclanthology.org/2020.emnlp-main.609`.