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, upperline LLM-only with golden context, 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 precision and retrieval quality, with reranking achieving higher Mean Reciprocal Rank (MRR) across configurations. While retrieval-augmented approaches improve factual grounding compared to the baseline, performance remains sensitive to retrieval errors, highlighting the importance of robust evidence selection for fact-checking tasks.

1 Introduction

Large language models 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 Their training data becomes outdated, and they cannot reliably cite sources for factual claims, leading to hallucination problems especially pronounced in scientific and specialized domains where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 neural 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]).

Although most RAG work focuses on open-domain question answering, 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 was to implement a complete RAG pipeline for scientific fact-checking and evaluate the impact of dense retrieval DPR and reranking on both retrieval quality (Mean Reciprocal Rank) and end-to-end fact-checking accuracy. Specifically, the study compare four approaches:

1. Baseline: using LLM-only.
2. Upperline (Oracle): LLM provided with golden context
3. RAG with DPR: LLM provided with context retrieved by DPR
4. RAG with DPR and Reranker: LLM provided with context reranked after DPR candidate selection

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 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.

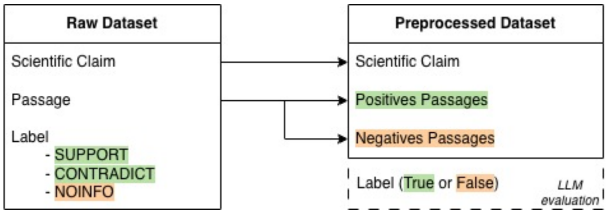


Figure 1: SciFact dataset raw schema and after data preprocessing

2 Related Work

Retrieval-Augmented Generation (RAG). Lewis et al. Lewis et al. [2020] combine parametric generation with non-parametric memory through retrieval, reducing hallucinations and improving factuality for knowledge-intensive tasks. RAG enables knowledge updates by modifying the retrieval index without retraining and has demonstrated effectiveness in open-domain QA and fact verification.

Dense Passage Retrieval (DPR). Karpukhin et al. Karpukhin et al. [2020] introduce DPR, a dual-encoder framework trained with contrastive learning that captures semantic similarity between queries and passages. DPR outperforms sparse methods like BM25 by leveraging dense representations and enables efficient retrieval through precomputed passage embeddings and approximate nearest neighbor search.

Cross-Encoder Reranking. Cross-encoders jointly encode query-passage pairs to compute precise relevance scores, improving upon bi-encoders by modeling fine-grained query-passage interactions Karpukhin et al. [2020]. In fact-checking, reranking is particularly valuable for distinguishing semantically relevant but factually incorrect passages from genuine supporting evidence. The two-stage architecture (recall-focused bi-encoder followed by precision-focused cross-encoder) improves retrieval quality without significant computational overhead.

Efficient Transformers (DistilBERT). DistilBERT Sanh et al. [2020] is a distilled version of BERT Devlin et al. [2018] that retains 97% of performance while being 40% smaller and 60% faster. Its efficiency makes it suitable for resource-constrained retrieval systems. In this work, DistilBERT serves as the backbone encoder for both DPR and the reranker.

3 Methodology

This section describes the technical approach to implementing the RAG pipeline for scientific fact-checking. The methodology covers: (1) input preprocessing and tokenization using DistilBERT, (2) training and deployment of the DPR retriever with contrastive learning, (3) development of a cross-encoder reranker using hard negative mining, (4) the four computational approaches evaluated for fact-checking (baseline, oracle, RAG with DPR, and RAG with DPR + reranker), and (5) evaluation metrics for both retrieval quality and end-to-end accuracy.

3.1 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.

3.2 Dense Passage Retrieval (DPR) 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.

3.3 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 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.

Training data consists of (query, positive, negative 1, . . . , negative N) 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.

3.4 Experiment Pipeline

Four computational approaches are evaluated for scientific fact-checking:

Approach 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.

Approach 2: Oracle (LLM + Golden Context). The oracle 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.

Approach 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.

Approach 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 language model component utilizes 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.

3.5 Evaluation Metrics

Retrieval quality: assessed using Mean Reciprocal Rank (MRR). 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.

End-to-end fact-checking performance: measured using accuracy, computed as the proportion of predictions matching the ground-truth labels from the SciFact dataset.

4 Results

This section presents both qualitative and quantitative results. We first analyze example outputs, then report retrieval metrics and end-to-end accuracy.

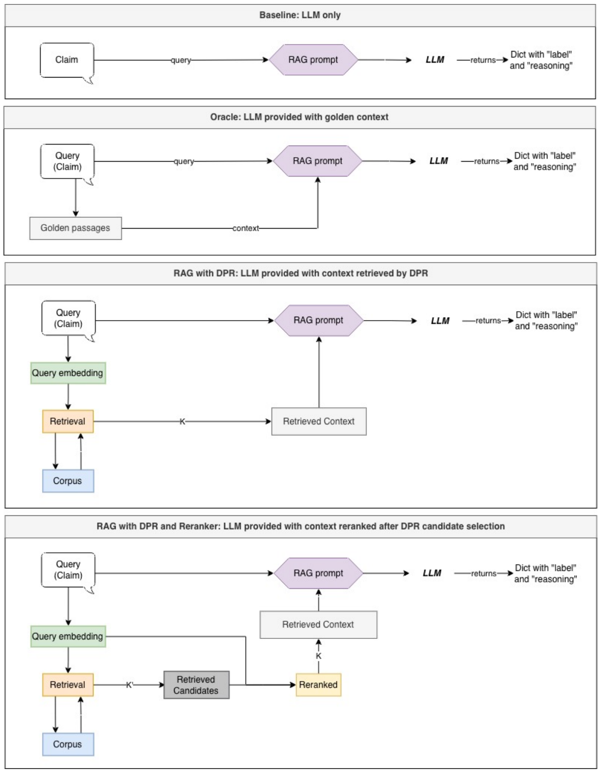
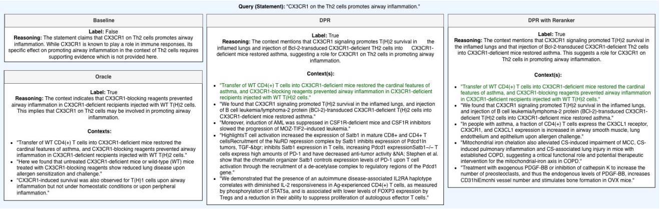

Baseline: LLM only
Claim
query
RAG prompt
LLM
returns
Dict with "label" and "reasoning"
Oracle: LLM provided with golden context
Query (Claim)
query
RAG prompt
LLM
returns
Dict with "label" and "reasoning"
Golden passages
context
RAG with DPR: LLM provided with context retrieved by DPR
Query (Claim)
RAG prompt
LLM
returns
Dict with "label" and "reasoning"
Query embedding
Retrieval
K
Retrieved Context
Corpus
RAG with DPR and Reranker: LLM provided with context reranked after DPR candidate selection
Query (Claim)
RAG prompt
LLM
returns
Dict with "label" and "reasoning"
Retrieved Context
Query embedding
K
Retrieval
K
Retrieved Candidates
Reranked
Corpus

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.

4.1 Qualitative Examples

Outputs displayed in Figure 3 highlights the main behavior observed across the dataset. The baseline sometimes produces confident answers with weak evidence. When correct passages are provided (oracle), the reasoning becomes more detailed and aligned with evidence. In the RAG settings, the quality of the retrieved context directly affects correctness: when retrieval includes relevant evidence, the model's reasoning becomes more concrete; when retrieval includes partially relevant or misleading passages, the model can be pushed toward the wrong label.

In Figure 4 outputs, it shows that even when the baseline predicts correctly, it often produces generic reasoning. With RAG, the reasoning is more grounded, because the model can point to retrieved sentences. This supports the core motivation of RAG: reducing hallucination and improving transparency Lewis et al. [2020].



Query (Statement): "CX3CR1 on the Th2 cells promotes airway inflammation."

Baseline

Label: False
Reasoning: The statement claims that CX3CR1 on Th2 cells promotes airway inflammation. While CX3CR1 is known to play a role in immune responses, its specific effect on promoting airway inflammation in the context of Th2 cells requires supporting evidence which is not provided here.

Oracle

Label: True
Reasoning: The context indicates that CX3CR1-blocking reagents prevented airway inflammation in CX3CR1-deficient recipients injected with WT T(H)2 cells. This implies that CX3CR1 on Th2 cells may be involved in promoting airway inflammation.

Contexts:

- "Transfer of WT CD4(+) T cells into CX3CR1-deficient mice restored the cardinal features of asthma, and CX3CR1-blocking reagents prevented airway inflammation in CX3CR1-deficient recipients injected with WT T(H)2 cells."
- "Here we found that untreated CX3CR1-deficient mice or wild-type (WT) mice treated with CX3CR1-blocking reagents show reduced lung disease upon allergen sensitization and challenge."
- "CX3CR1-induced survival was also observed for T(H)1 cells upon airway inflammation but not under homeostatic conditions or upon peripheral inflammation."

DPR

Label: True
Reasoning: The context mentions that CX3CR1 signaling promotes T(H)2 survival in the inflamed lungs and injection of Bcl-2-transduced CX3CR1-deficient TH2 cells into CX3CR1-deficient mice restored asthma, suggesting a role for CX3CR1 on Th2 cells in promoting airway inflammation.

Context(s):

- "Transfer of WT CD4(+) T cells into CX3CR1-deficient mice restored the cardinal features of asthma, and CX3CR1-blocking reagents prevented airway inflammation in CX3CR1-deficient recipients injected with WT T(H)2 cells."
- "We found that CX3CR1 signaling promoted T(H)2 survival in the inflamed lungs, and injection of B cell leukemia/lymphoma-2 protein (BCl-2)-transduced CX3CR1-deficient T(H)2 cells into CX3CR1-deficient mice restored asthma."
- "Moreover, induction of AML was suppressed in CSF1R-deficient mice and CSF1R inhibitors slowed the progression of MOZ-TIF2–induced leukemia."
- "HighlightsT cell activation increased the expression of Satb1 in mature CD8+ and CD4+ T cellsRecruitment of the NuRD repression complex by Satb1 inhibits expression of Pdcd1In tumors, TGF-&bgr; inhibits Satb1 expression in T cells, increasing Pdcd1 expressionSatb1−/− T cells express high amounts of PD-1 and have decreased anti-tumor activity &NA; Stephen et al. show that the chromatin organizer Satb1 controls expression levels of PD-1 upon T cell activation through the recruitment of a de-acetylase complex to regulatory regions of the Pdcd1 gene."
- "We demonstrated that the presence of an autoimmune disease-associated IL2RA haplotype correlates with diminished IL-2 responsiveness in Ag-experienced CD4(+) T cells, as measured by phosphorylation of STAT5a, and is associated with lower levels of FOXP3 expression by Tregs and a reduction in their ability to suppress proliferation of autologous effector T cells."

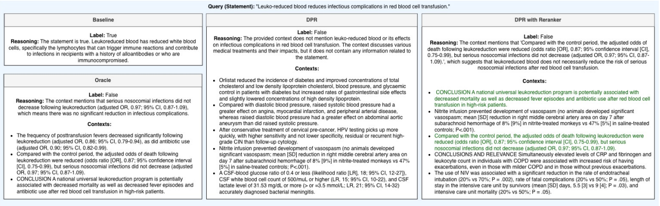
DPR with Reranker

Label: True
Reasoning: The context mentions that CX3CR1 signaling promoted T(H)2 survival in the inflamed lungs and that injection of Bcl-2-transduced CX3CR1-deficient Th2 cells into CX3CR1-deficient mice restored asthma. This suggests a role for CX3CR1 on Th2 cells in promoting airway inflammation.

Context(s):

- "Transfer of WT CD4(+) T cells into CX3CR1-deficient mice restored the cardinal features of asthma, and CX3CR1-blocking reagents prevented airway inflammation in CX3CR1-deficient recipients injected with WT T(H)2 cells."
- "We found that CX3CR1 signaling promoted T(H)2 survival in the inflamed lungs, and injection of B cell leukemia/lymphoma-2 protein (BCl-2)-transduced CX3CR1-deficient T(H)2 cells into CX3CR1-deficient mice restored asthma."
- "In people with asthma, a fraction of CD4(+) T cells express the CX3CL1 receptor, CX3CR1, and CX3CL1 expression is increased in airway smooth muscle, lung endothelium and epithelium upon allergen challenge."
- "Mitochondrial iron chelation also alleviated CS-induced impairment of MCC, CS-induced pulmonary inflammation and CS-associated lung injury in mice with established COPD, suggesting a critical functional role and potential therapeutic intervention for the mitochondrial-iron axis in COPD."
- "Treatment with exogenous PDGF-BB or inhibition of cathepsin K to increase the number of preosteoclasts, and thus the endogenous levels of PDGF-BB, increases CD31hiEmcnhi vessel number and stimulates bone formation in OVX mice."

Figure 3: Outputs for a claim under baseline, oracle, DPR, and DPR+reranker. The baseline relies only on parametric knowledge and may hallucinate. Oracle uses gold evidence. DPR retrieval improves grounding, and reranking provides more relevant evidence, improving reasoning quality.



Query (Statement): "Leuko-reduced blood reduces infectious complications in red blood cell transfusion."

Baseline

Label: True
Reasoning: The statement is true. Leukoreduced blood has reduced white blood cells, specifically the lymphocytes that can trigger immune reactions and contribute to infections in recipients with a history of alloantibodies or who are immunocompromised.

Oracle

Label: False
Reasoning: The context mentions that serious nosocomial infections did not decrease following leukoreduction (adjusted OR, 0.97; 95% CI, 0.87-1.09), which means there was no significant reduction in infectious complications.

Contexts:

- The frequency of posttransfusion fevers decreased significantly following leukoreduction (adjusted OR, 0.86; 95% CI, 0.79-0.94), as did antibiotic use (adjusted OR, 0.90; 95% CI, 0.82-0.99).
- Compared with the control period, the adjusted odds of death following leukoreduction were reduced (odds ratio [OR], 0.87; 95% confidence interval [CI], 0.75-0.99), but serious nosocomial infections did not decrease (adjusted OR, 0.97; 95% CI, 0.87-1.09).
- CONCLUSION A national universal leukoreduction program is potentially associated with decreased mortality as well as decreased fever episodes and antibiotic use after red blood cell transfusion in high-risk patients.

DPR

Label: False
Reasoning: The provided context does not mention leuko-reduced blood or its effects on infectious complications in red blood cell transfusion. The context discusses various medical treatments and their impacts, but it does not contain any information related to the statement.

Contexts:

- Orlistat reduced the incidence of diabetes and improved concentrations of total cholesterol and low density lipoprotein cholesterol, blood pressure, and glycaemic control in patients with diabetes but increased rates of gastrointestinal side effects and slightly lowered concentrations of high density lipoprotein.
- Compared with diastolic blood pressure, raised systolic blood pressure had a greater effect on angina, myocardial infarction, and peripheral arterial disease, whereas raised diastolic blood pressure had a greater effect on abdominal aortic aneurysm than did raised systolic pressure.
- After conservative treatment of cervical pre-cancer, HPV testing picks up more quickly, with higher sensitivity and not lower specificity, residual or recurrent high-grade CIN than follow-up cytology.
- Nitrite infusion prevented development of vasospasm (no animals developed significant vasospasm; mean [SD] reduction in right middle cerebral artery area on day 7 after subarachnoid hemorrhage of 8% [9%] in nitrite-treated monkeys vs 47% [5%] in saline-treated controls; P<.001).
- A CSF-blood glucose ratio of 0.4 or less (likelihood ratio [LR], 18; 95% CI, 12-27]), CSF white blood cell count of 500/muL or higher (LR, 15; 95% CI, 10-22), and CSF lactate level of 31.53 mg/dL or more (> or =3.5 mmol/L; LR, 21; 95% CI, 14-32) accurately diagnosed bacterial meningitis.

DPR with Reranker

Label: False
Reasoning: The context mentions that 'Compared with the control period, the adjusted odds of death following leukoreduction were reduced (odds ratio [OR], 0.87; 95% confidence interval [CI], 0.75-0.99), but serious nosocomial infections did not decrease (adjusted OR, 0.97; 95% CI, 0.87-1.09).', which suggests that leukoreduced blood does not necessarily reduce the risk of serious nosocomial infections after red blood cell transfusion.

Contexts:

- CONCLUSION A national universal leukoreduction program is potentially associated with decreased mortality as well as decreased fever episodes and antibiotic use after red blood cell transfusion in high-risk patients.
- Nitrite infusion prevented development of vasospasm (no animals developed significant vasospasm; mean [SD] reduction in right middle cerebral artery area on day 7 after subarachnoid hemorrhage of 8% [9%] in nitrite-treated monkeys vs 47% [5%] in saline-treated controls; P<.001).
- Compared with the control period, the adjusted odds of death following leukoreduction were reduced (odds ratio [OR], 0.87; 95% confidence interval [CI], 0.75-0.99), but serious nosocomial infections did not decrease (adjusted OR, 0.97; 95% CI, 0.87-1.09).
- CONCLUSIONS AND RELEVANCE Simultaneously elevated levels of CRP and fibrinogen and leukocyte count in individuals with COPD were associated with increased risk of having exacerbations, even in those with milder COPD and in those without previous exacerbations.
- The use of NIV was associated with a significant reduction in the rate of endotracheal intubation (20% vs 70%; P = .002), rate of fatal complications (20% vs 50%; P = .05), length of stay in the intensive care unit by survivors (mean [SD] days, 5.5 [3] vs 9 [4]; P = .03), and intensive care unit mortality (20% vs 50%; P = .05).

Figure 4: Outputs for a claim about CO_2 reduction. The baseline may guess correctly but with generic reasoning. RAG improves evidence-based explanation, especially when reranking selects passages closer to the claim semantics.

4.2 Retrieval Performance (MRR)

Figure 5 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 6 helps interpret the trend: increasing the number of candidates generally increases MRR with reranking because the reranker has more options to choose from. This matches the standard view that retrieval is a recall-focused step and reranking is a precision-focused step.

4.3 End-to-End Accuracy

Figure 7 shows that the oracle configuration achieves the highest accuracy, confirming that correct evidence leads to strong downstream performance. The baseline performance is relatively high, suggesting that the LLM has strong general scientific knowledge, but it is not always reliable. The RAG approaches improve grounding, but

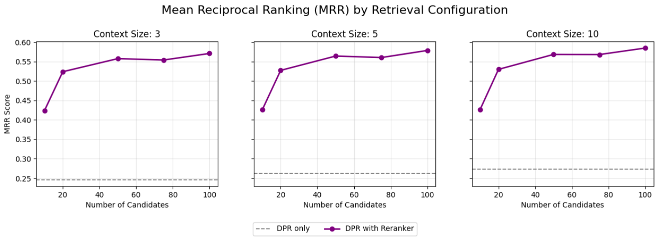


Figure 5: MRR results for DPR vs. DPR with reranker under different candidate pool sizes and context depths (3, 5, 10). Reranking consistently increases MRR across settings.

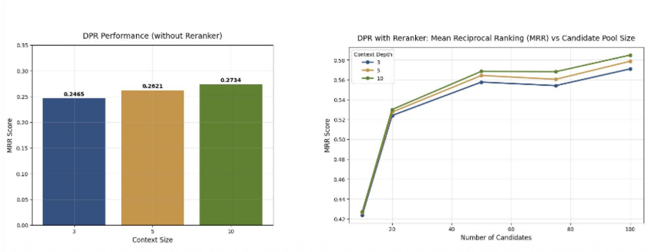


Figure 6: Left: DPR-only MRR increases with context size. Right: reranking improves MRR further as candidate pool increases. Larger candidate pools allow better reranking but increase computation.

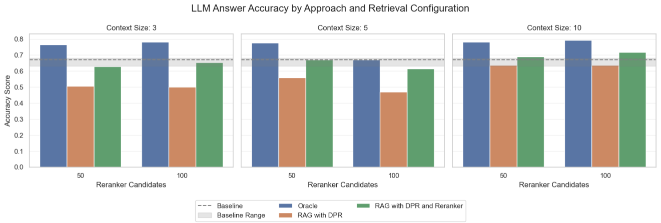


Figure 7: Fact-checking accuracy by approach and retrieval configuration. Oracle is the upper bound. RAG improves over baseline when retrieved evidence is correct; reranking helps by selecting better contexts.

they can sometimes reduce accuracy when wrong contexts are retrieved and passed to the generator. This is also mentioned in the slides as a limitation: “RAG can compromise results when wrong contexts are passed”.

Overall, the best configuration combines DPR and reranking, showing that improving retrieval quality improves end-to-end correctness. This is consistent with DPR literature, where stronger retrieval precision is associated with improved downstream QA accuracy Karpukhin et al. [2020], and with the RAG motivation of reducing hallucination by providing evidence Lewis et al. [2020].

5 Discussion

The results show that adding a reranking step consistently improves retrieval quality, as reflected by higher MRR values across all tested 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.

6 Conclusion

We implemented and evaluated a RAG pipeline for scientific fact-checking on SciFact. We compared baseline LLM-only, oracle with golden evidence, RAG with DPR, and RAG with DPR plus a reranker. The results show that reranking improves retrieval quality (higher MRR) and generally improves end-to-end accuracy by providing better evidence to the generator. However, incorrect retrieval can still harm results, highlighting the importance of reliable evidence selection.

Possible improvements (as also suggested in the slides) 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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

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.

A Implementation Details

A.1 Hardware and Software

The experiments were conducted using Python with PyTorch. Main libraries include Hugging Face Datasets, Sentence-Transformers, and Ollama for serving the LLM locally.

A.2 Reproducibility

The training and evaluation code is implemented in Jupyter notebooks with fixed random seeds and explicit hyperparameter settings for retriever and reranker training.