Document Information as Non-Parametric Memory using Dense Passage Retrieval and
Custom 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) can answer many questions from internal knowledge, but their knowledge is static and they may hallucinate, especially in scientific domains. Retrieval-Augmented Generation (RAG) addresses this by combining a parametric generator with a non-parametric memory (a document corpus) that can be retrieved at inference time. In this project, we implement and evaluate a RAG pipeline for **scientific fact-checking** using the SciFact dataset. We compare four approaches: (1) baseline LLM-only, (2) oracle LLM with golden evidence, (3) RAG with Dense Passage Retrieval (DPR), and (4) RAG with DPR plus a cross-encoder reranker. We evaluate retrieval quality with Mean Reciprocal Rank (MRR) and evaluate end-to-end performance with fact-checking accuracy (True/False label prediction). Results show that reranking improves retrieval quality and that retrieval-augmented approaches improve factual grounding, although incorrect retrieval can still harm performance compared to the oracle.

1 Introduction

Open-domain question answering and related knowledge-intensive tasks require systems to access relevant information from large corpora. Traditional methods often relied on sparse lexical matching (e.g., BM25), while more recent approaches use large language models (LLMs) that generate outputs directly from their parameters. However, purely generative models may produce answers that sound correct but are not supported by evidence, especially in specialized domains.

Recently, the Retrieval-Augmented Generation (RAG) paradigm emerged as a strong approach to combine retrieval and generation Lewis et al. [2020]. RAG systems retrieve top-K documents from a corpus and provide them as context to the generator. This reduces hallucination and makes outputs easier to inspect, since the model can cite evidence. In the original RAG work, retrieval is performed with Dense Passage Retrieval (DPR) and the generator is a sequence-to-sequence model that conditions on retrieved passages Lewis et al. [2020].

Dense Passage Retrieval Karpukhin et al. [2020] is a learned retrieval method that maps queries and passages into dense vectors and retrieves by similarity in embedding space. Compared to sparse retrieval, DPR can capture semantic similarity and match paraphrases. DPR uses a dual-encoder (bi-encoder) architecture, which is efficient because passage embeddings can be precomputed.

Despite strong retrieval performance, bi-encoders may still retrieve semantically similar but incorrect passages. A common improvement is to add a reranking step using a cross-encoder, which jointly encodes query and passage to score relevance more precisely (at higher computational cost) Karpukhin et al. [2020].

Although much RAG literature focuses on open-domain QA, this project applied RAG to scientific fact-checking. Given a scientific claim, the system must decide if it is supported or contradicted by scientific evidence. This requires retrieving relevant passages and reasoning over them. Our system outputs a label (True/False) and a short reasoning text.

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 and reranking. Specifically, we aim to:

1. Implement a DPR-based retrieval component using DistilBERT encoders
2. Implement a cross-encoder reranker and use hard negatives from retrieval to train it
3. Compare baseline, oracle, DPR-based RAG, and DPR combined with reranker RAG approaches
4. Evaluate retrieval quality with Mean Reciprocal Rank (MRR) and end-to-end accuracy

1.2 Dataset and Key Concepts

We use the SciFact dataset (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 are used as negatives. The dataset split used in this project is:

- Train: 594 claims
- Evaluation: 188 claims
- Corpus size: 2,263 passages

2 Related Work

Retrieval-Augmented Generation (RAG). Lewis et al. Lewis et al. [2020] propose combining parametric generation with non-parametric memory accessed through retrieval. Their work highlights that retrieval can reduce hallucinations, help with factuality, and allow knowledge updates by changing the index. They show strong results in open-domain QA and also discuss fact verification tasks.

Dense Passage Retrieval (DPR). Karpukhin et al. Karpukhin et al. [2020] introduce a dual-encoder framework trained with contrastive learning. DPR outperforms BM25 in open-domain QA retrieval settings because it matches semantically similar text rather than only exact keywords. DPR uses dot-product similarity between query and passage embeddings and can be scaled with approximate nearest neighbor search.

Efficient Transformers (DistilBERT). DistilBERT **?** is a compressed version of BERT that keeps most of the performance while being smaller and faster. This makes it appropriate for retrieval systems and student projects with limited resources. In our pipeline, DistilBERT is used as the backbone encoder for both DPR and the reranker.

3 Methodology

3.1 Data Preparation

The SciFact dataset contains claim–passage pairs. We preprocess it to build training tuples for retrieval:

- **Query:** the claim
- **Positive passages:** passages labeled SUPPORT or CONTRADICT
- **Negative passages:** randomly selected NOINFO passages

We also build a retrieval corpus of 2,263 unique passages (after removing duplicates). This corpus is used to evaluate retrieval and to provide context in the RAG pipeline.

3.2 Tokenization and Attention Masks

Inputs are tokenized using `distilbert-base-uncased` with maximum length 256. Variable-length sequences are padded to this maximum length. Attention masks are generated so that real tokens have mask value 1 and padding tokens have value 0. This ensures padding tokens do not influence the self-attention computation.

3.3 Dense Passage Retrieval (DPR) Model

3.3.1 Architecture

The DPR retriever is a bi-encoder:

- **Query encoder:** DistilBERT encodes the claim into a dense vector.
- **Passage encoder:** DistilBERT encodes passages into dense vectors.
- **Pooling:** mean pooling over token embeddings produces a fixed-size vector.

3.3.2 Scoring and Retrieval

We use dot-product similarity between claim and passage embeddings. At inference time, we encode all corpus passages and retrieve top candidates by similarity.

3.3.3 Training Objective and Configuration

Each training sample follows a 1–1–N structure: one query, one positive passage, and 10 negatives. We train using a cross-entropy objective over the positive vs. negatives. Training configuration (as in the slides):

- epochs = 2
- learning rate = 2×10^{-5}
- weight decay = 0.01
- optimizer = AdamW
- number of negatives per query = 10

3.4 Cross-Encoder Reranker

3.4.1 Architecture

The reranker is a DistilBERT-based cross-encoder:

- Claim and candidate passage are concatenated and jointly encoded.
- We use the `[CLS]` embedding as an aggregate representation.
- A dropout layer (0.1) is applied to reduce overfitting.
- A linear scoring head outputs a relevance logit.

3.4.2 Hard Negative Mining and Training

Hard negatives are candidates retrieved by DPR with high similarity but that are factually incorrect (wrong label/evidence). For each training claim, we take the top 10 DPR candidates and select incorrect ones as hard negatives. Training configuration:

- epochs = 3
- learning rate = 2×10^{-5}
- weight decay = 0.01
- optimizer = AdamW
- loss = Binary Cross-Entropy with Logits

3.5 Fact-Checking Pipeline

We compare four approaches (as in the slides):

3.5.1 Approach 1: Baseline (LLM-only)

The baseline prompts the LLM using only the claim. It outputs a JSON object with:

- `label` (True/False)
- `reasoning` (short justification)

3.5.2 Approach 2: Oracle (LLM + Golden Context)

The oracle provides the gold evidence passages to the LLM. This represents an upper bound, assuming perfect retrieval.

3.5.3 Approach 3: RAG with DPR

We retrieve candidates using DPR and provide the top-K passages as context to the LLM.

3.5.4 Approach 4: RAG with DPR + Reranker

We retrieve an initial candidate pool with DPR, rerank them with the cross-encoder, and provide the top-K reranked passages to the LLM.

3.5.5 Generator Configuration and Prompting

We use Qwen 2.5 7B quantized and served locally via Ollama. We apply one-shot chain-of-thought prompting:

- prompt without context (baseline)
- prompt with context (oracle and RAG approaches)

The model output is a JSON object with `label` and `reasoning`.

3.6 Evaluation

3.6.1 Retrieval Metric: Mean Reciprocal Rank (MRR)

We evaluate retrieval using Mean Reciprocal Rank (MRR). For each query, we compute the rank of the first relevant passage and average reciprocal ranks across queries. We evaluate MRR under different candidate pool sizes and context depths (as in the results slides).

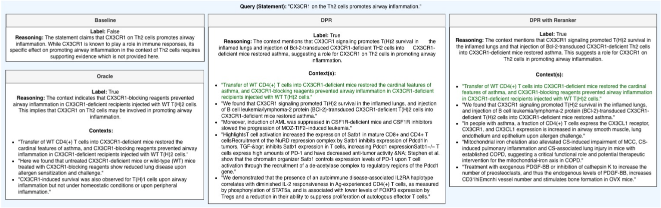
3.6.2 End-to-End Metric: Fact-Checking Accuracy

We evaluate accuracy by comparing the predicted label (True/False) with the ground-truth label in SciFact.

4 Results

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

4.1 Qualitative Examples



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."

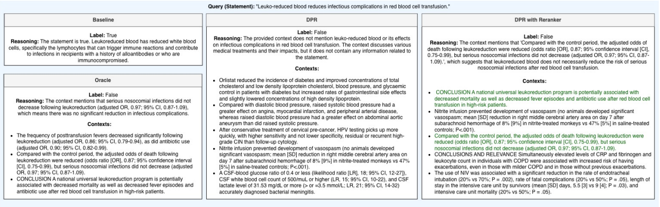
Figure 1: Outputs for a claim under baseline, oracle, DPR, and DPR+reranker. The baseline relies only on parametric knowledge and may hallucinate. Oracle uses gold evidence. DPR retrieval improves grounding, and reranking provides more relevant evidence, improving reasoning quality.

Outputs displayed in Figure 1 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 2 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].

4.2 Retrieval Performance (MRR)

Figure 3 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.



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 2: 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.

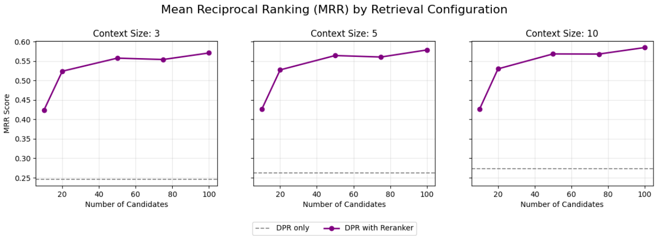


Figure 3: MRR results for DPR vs. DPR with reranker under different candidate pool sizes and context depths (3, 5, 10). Reranking consistently increases MRR across settings.

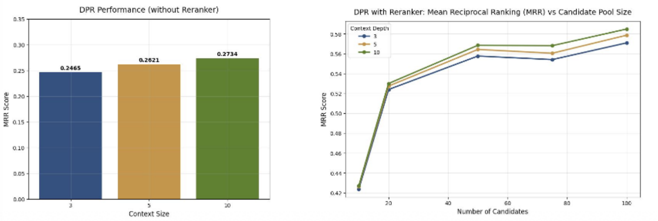


Figure 4: Left: DPR-only MRR increases with context size. Right: reranking improves MRR further as candidate pool increases. Larger candidate pools allow better reranking but increase computation.

Figure 4 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

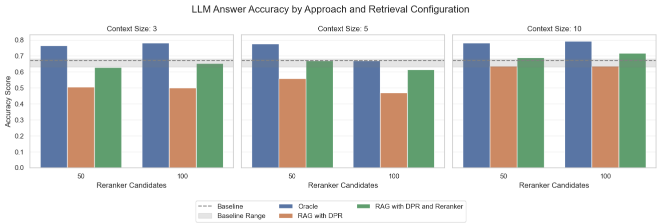


Figure 5: Fact-checking accuracy by approach and retrieval configuration. Oracle is the upper bound. RAG improves over baseline when retrieved evidence is correct; reranking helps by selecting better contexts.

Figure 5 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 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 **?**. 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

Wayne Chang, Jie Yu, Yuqing Zhang, Zheng Zhou, Joshua B Tenenbaum, and Regina Barzilay. Pre-trained models for natural language processing: A survey. *arXiv preprint arXiv:2003.08271*, 2020.

Vladimir Karpukhin, Barlas O0̆308uz, 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.

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.