Automated evaluation and parameter optimisation for retrieval augmented generation systems

Master's Thesis Computational Science and Engineering

Carried out at the Faculty of Informatics TU Vienna
Supervised by Prof. Dr. A. Rauber

Author

Github

Lichess

Simon König BSc

- **Bachelors:** Physics
- **Key Area Informatics**
  - Machine Learning (A. Rauber)
  - Heuristic Optimization Techniques (G. Raidl)
  - Adv. Regression and Classification (P. Filzmoser)
- **Key Area Building Science**
  - Adv. Numerical Methods in Building Science (T. Bednar)
  - Digital Twins for Buildings and Cities (T. Bednar)
- **Current Occupation:** Research Engineer at Austrian Institute of Technology
- **Favorite Lectures:** NumPDE, Heuristic Optimisation Techniques, Machine Learning

Introduction

This Thesis introduces a lightweight Retrieval-Augmented Generation (RAG) evaluation framework that uses two complementary metrics—an efficient ROUGE-based score and a more detailed LLM-based judge—and proposes the RAG Triad approach for managing unlabeled data. By offering a structured methodology for both evaluation and system configuration, this framework aims to advance a more scientific, scalable design of RAG systems.

Implementation and Evaluation

RAG systems gained popularity shortly after the rise of LLMs to general awareness [1]. The basic process is structured into four parts. (i) First domain data is processed and stored in a vector index for later retrieval. (ii) Second, the system receives a query/question and retrieves information from a vector database [2]. (iii) Third, the query is fused with the retrieved information via an instructive prompt. (iv) Fourth, based on query, information and prompt, a LLM produces the final RAG output [3].

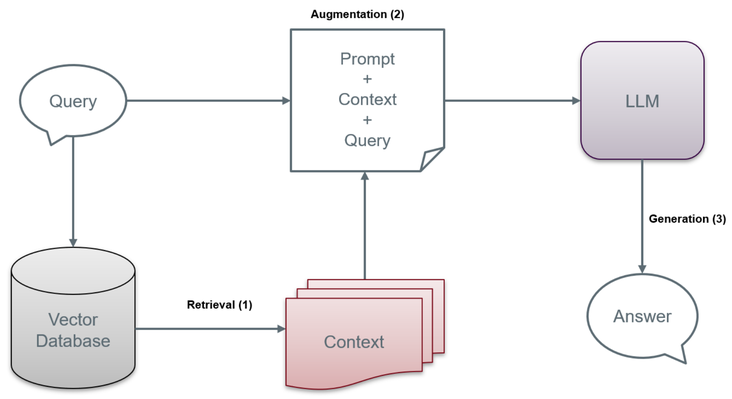


Figure: Naive RAG

The graphic overview below illustrates the evaluation workflow: starting with a given dataset, progressing through the RAG pipeline, entering the evaluation framework and yielding results.

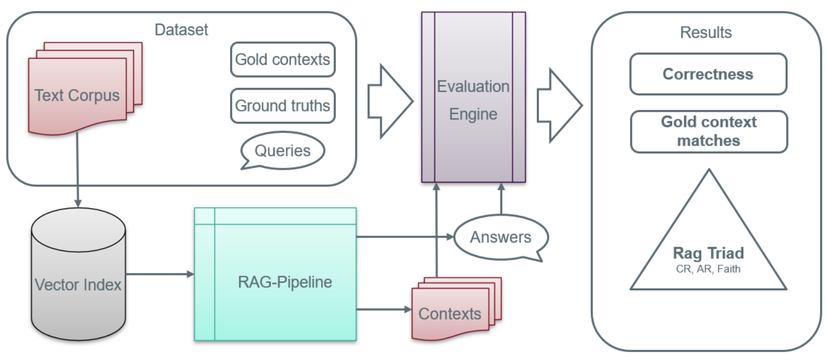


Figure: Evaluation Framework

RAG Triad

The RAG Triad is a innovative performance indicator addressing the relationships between RAG sub-components and intermediate results without relying on ground truth answers or gold standard resources. The hypothesis of the RAG Triad states that if the intermediate results among the sub-components of a RAG system — query, resources, and answer — are sufficiently correct, then the RAG system yields sufficiently correct answers.

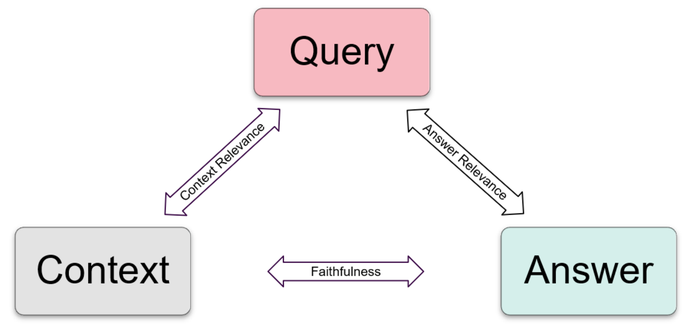


Figure: RAG Triad

$$\text{RAG Triad} = \frac{1}{|Q|} \sum_{q_i \in Q} \left(\max_{c_i} \text{CR}(q_i, c_i) + \text{AR}(q_i, a_i) + \max_{c_i} \text{F}(a_i, c_i) \right)$$

Advanced RAG

The advanced RAG pipeline updates the naive RAG pipeline, enriching the RAG ecosystem with additional options illustrated in the figure below.

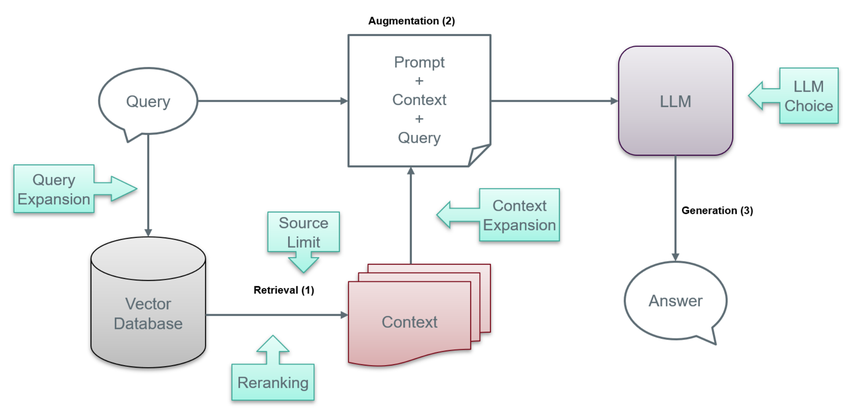
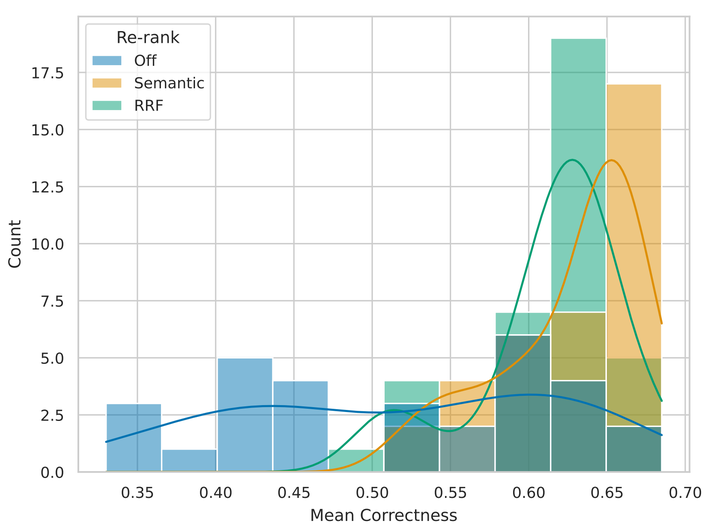
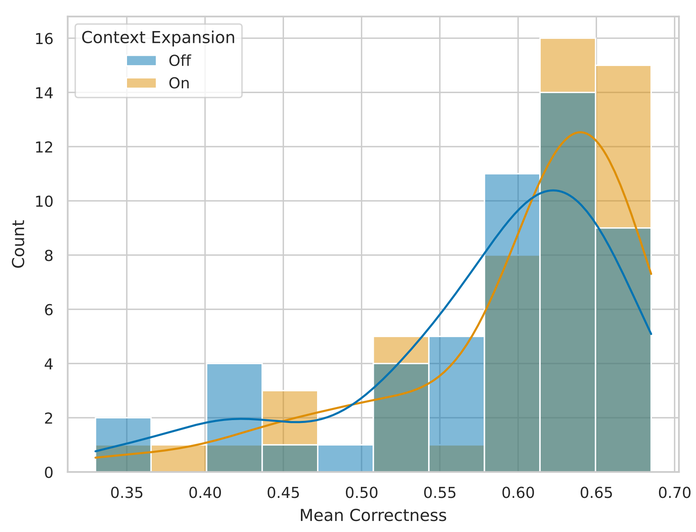
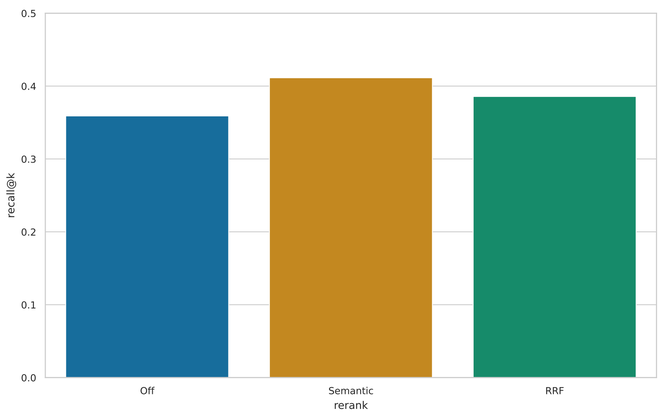


Figure: Advanced RAG

Results Overview







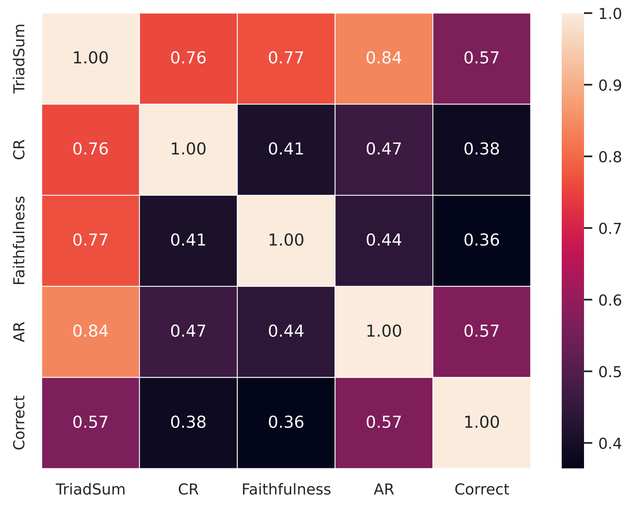


Figure: Evaluation Results

Conclusion

We evaluated various optimizations in a RAG pipeline, including re-ranking, query expansion, context expansion, source selection, and LLM choice. Re-ranking (semantic or reciprocal rank fusion) improved answer correctness at different computational costs, while query expansion introduced noise and failed to enhance retrieval. Context expansion added valuable information at minimal cost. The RAG Triad emerged as a useful metric in the absence of ground truths, enabling performance assessment of individual pipeline components. Overall, the proposed evaluation framework can handle labeled and unlabeled data, guiding a more efficient and accurate RAG pipeline design.

References

[1] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. *Retrieval-Augmented Generation for Large Language Models: A Survey*. arXiv:2312.10997 [cs]. Mar. 2024.

[2] Toni Taipalus. "Vector database management systems: Fundamental concepts, use-cases, and current challenges". In: *Cognitive Systems Research* 85 (June 2024). arXiv:2309.11322 [cs], p. 101216. ISSN: 13890417.

[3] Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. *Retrieval-Augmented Generation for AI-Generated Content: A Survey*. arXiv:2402.19473 [cs]. Apr. 2024.