Technical Test Submission : Evaluation Methodology

Iman Bensalami

May 2025

To evaluate the quality of responses produced by a Retrieval-Augmented Generation (RAG) system, I propose a practical and structured methodology combining automated tools, LLM-based reasoning, and a transparent scoring rubric. It is inspired by recent work like the RAGAS framework [1] and the vRAG-Eval rubric [2] , ensuring both explainability and scalability.My evaluation focuses on three key dimensions:

- Faithfulness : Are the answer's claims supported by the retrieved context?
- Answer Relevance : Does the answer actually respond to the user’s question?
- Context Relevance : Is the retrieved information focused and useful?

To evaluate faithfulness, I break the answer as a sequence of factual statements using a language model, and test whether each factual content is supported by the retrieved context. The faithfulness score corresponds to the fraction of justified claims. This approach is consistent with the FEVER-style fact-checking task and is also implemented within RAGAS. For answer retrieval or relevance, I generate paraphrased versions of the answer's implied question and compare these to the original in terms of semantic similarity (text-embedding-ada-002) to to ensure thematic alignment. For context relevance, I ask an LLM to extract only the necessary parts of the context used to answer the question, and measure the precision, a crucial factor for long-document RAG.

To complement these metrics, I apply a simple 5-point scoring rubric:

- 1 = Hallucinated/off-topic
- 2 = Honest admission of missing info
- 3 = Relevant but inaccurate
- 4 = Correct but incomplete
- 5 = Accurate and comprehensive

This rubric is easy to apply manually or automatically, and has shown high consistency with expert judgments (82% agreement with GPT-4 scoring).

Finally, I use RAGAS to automatically run the entire evaluation pipeline without needing model answers. It is easy to integrate with tools such like LangChain ans LlamaIndex.

Bibliography

[1] J. J. L. E.-A. S. S. Shahul Es, «RAGAS: Automated Evaluation of Retrieval Augmented Generation,» 2023.

[2] A. G. H. R. K. N. K. Yang Wang, «Evaluating Quality of Answers for Retrieval-Augmented Generation: A Strong LLM Is All You Need,» 2024.