Comparing LLMs in RAG: A Multi-Metric Evaluation

TOOPLOOX

Karol Szymański
karol.szymanski@tooploox.com

Szymon Płaneta
szymon.planeta@tooploox.com

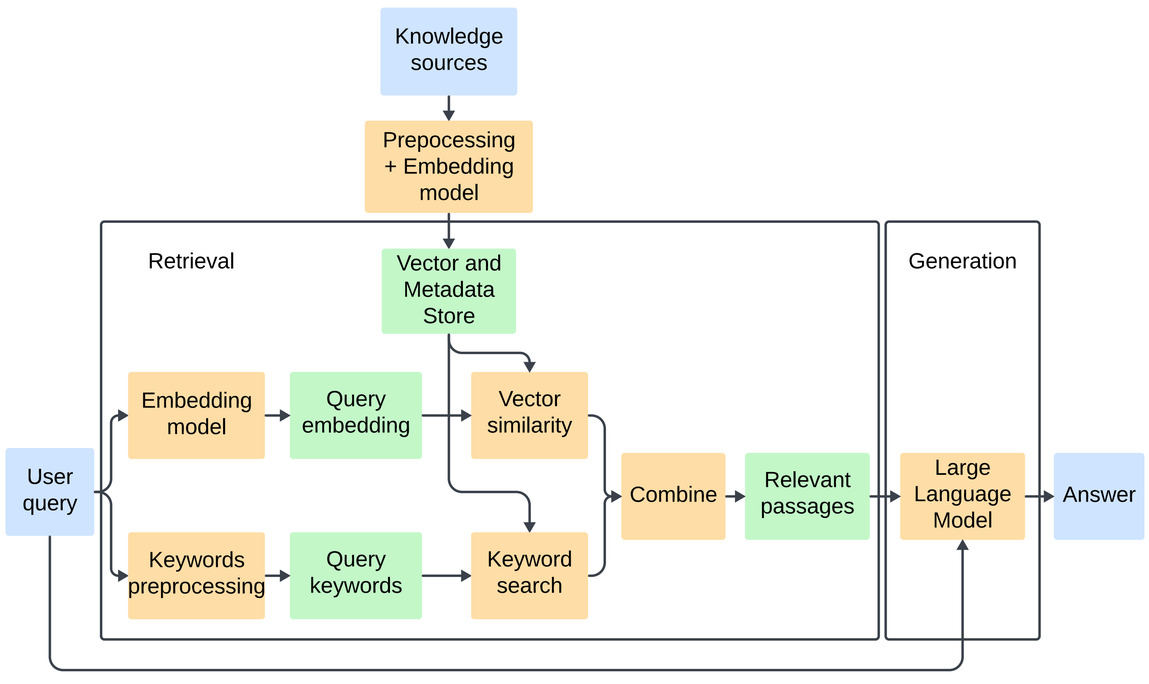
ML in PL CONFERENCE 2024

Introduction

This study presents a comparative analysis of several large language models (LLMs) within a Retrieval-Augmented Generation (RAG) framework. OpenAI's GPT-4, GPT-4 Turbo, and GPT-3.5 models [1], as well as popular open-source models LLaMA2 13B [2] and Mistral 7B [3], were assessed for their performance in generating contextually accurate answers. A synthetic test dataset was generated by LLMs using a selection of internal company documents. Performance was measured across multiple quality metrics, with metric values also determined by LLMs to enable an automated evaluation process.

RAG

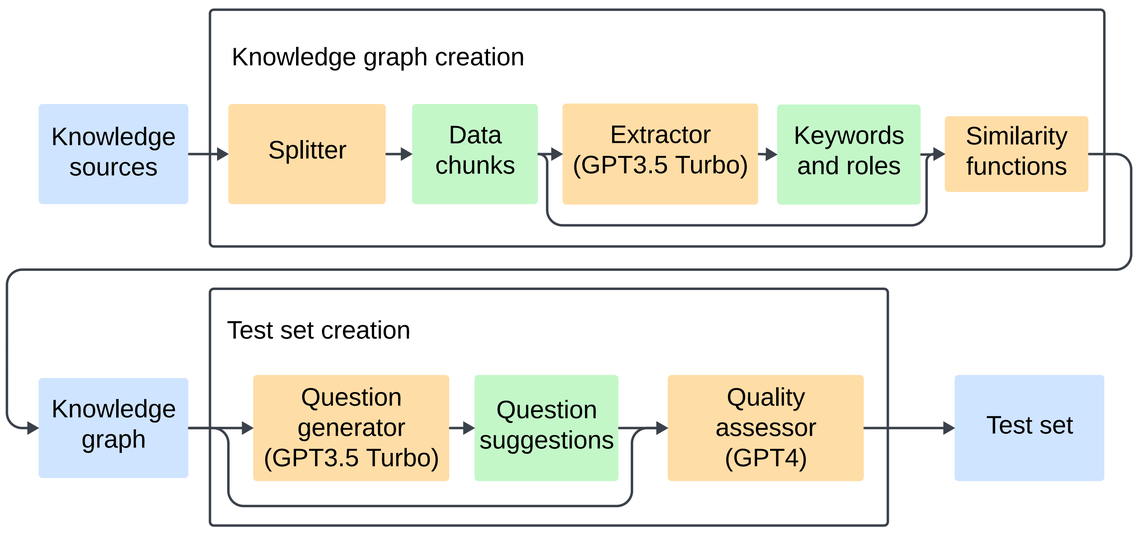
In this experiment, the Retrieval-Augmented Generation (RAG) system was implemented with the Danswer framework [4], using the Vespa search engine [5] as the vector database. Text embeddings were generated with the intfloat/multilingual-e5-large model [6] to support multilingual capabilities.



RAG architecture

Dataset creation

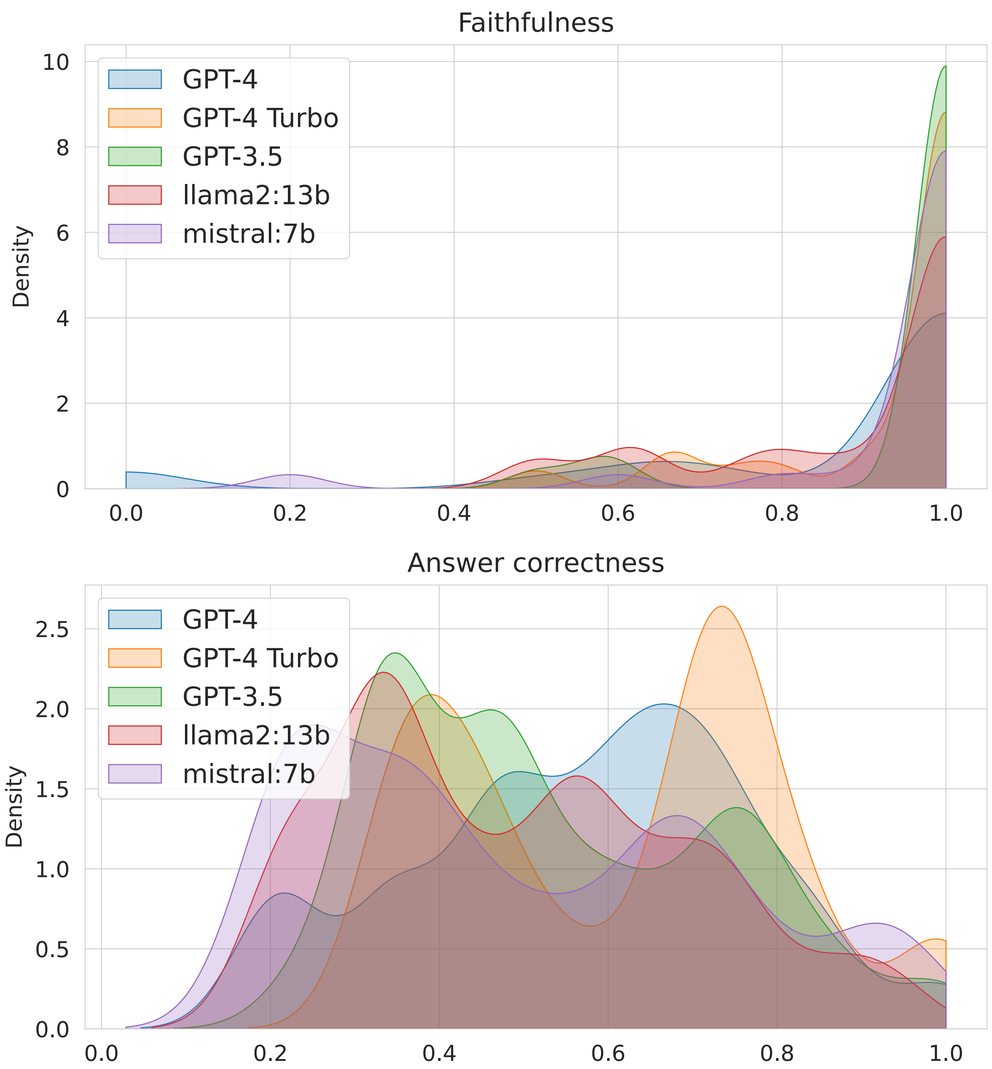
A synthetic test set was generated following Ragas guidelines [7] using GPT models. A small set of internal company documents, written partially in Polish and partially in English, was used as the basis for this dataset. The final test set included 27 questions based on 14 documents to assess model performance.

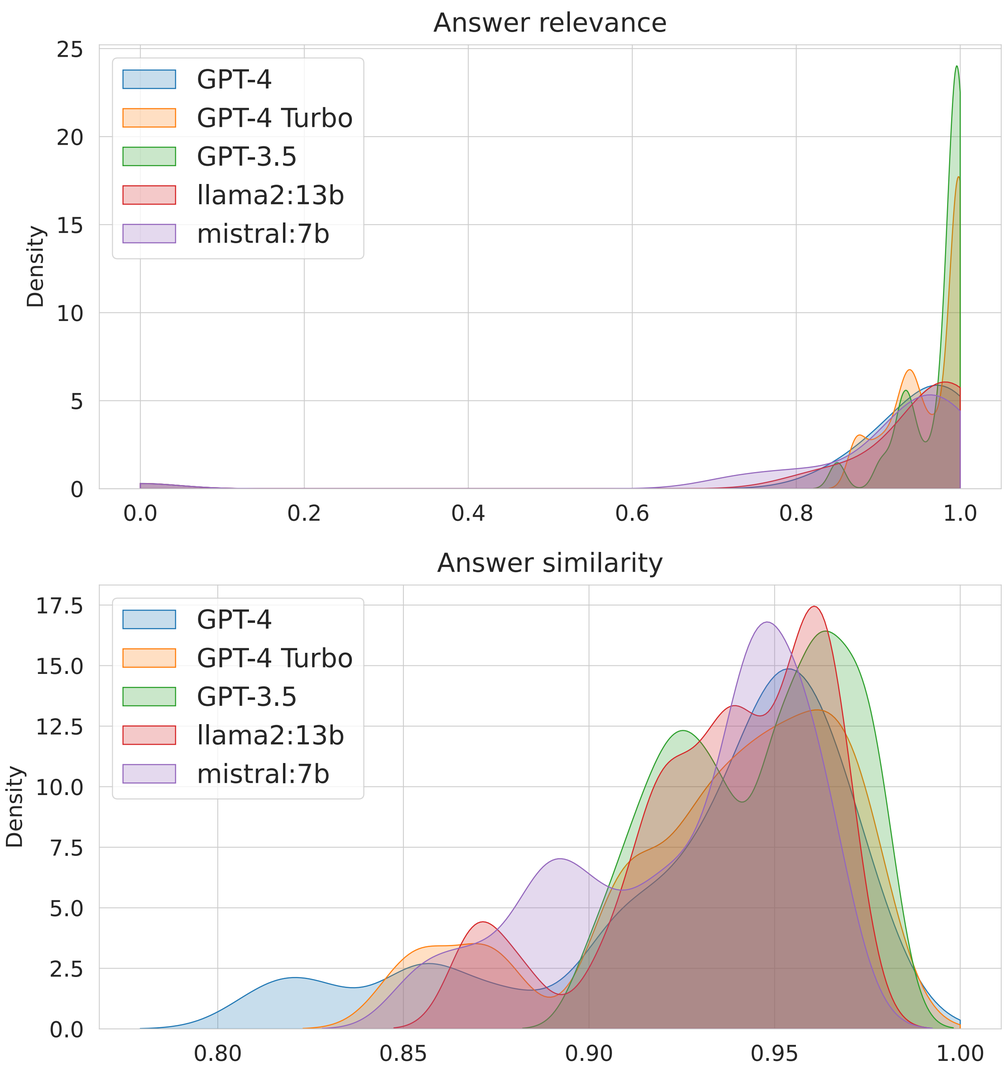


Test set generation from knowledge base

Metrics

- Faithfulness score: Proportion of claims in the generated answer that can be inferred from the provided context, relative to the total number of claims in the answer.
- Answer relevancy: Average cosine similarity between the embeddings of generated questions based on generated answers and the embedding of the original questions.
- Answer correctness: Count of true positive (TP), false positive (FP), and false negative (FN) claims relative to the reference answer.
- Answer similarity: Cosine similarity between the embedding of the generated answer and the embedding of the reference answer.





Conclusions

- The experiment used a small knowledge base, as RAG on a larger dataset frequently failed to provide the LLM with sufficient context to accurately answer the questions.
- GPT-4 Turbo achieved the highest average score in Answer Correctness, making it the top performer in this primary metric.
- On average, open-source models produced less correct answers compared to commercial models, indicating a quality gap in accuracy.
- Despite being a smaller model, GPT-3.5 achieved the highest faithfulness and relevance, suggesting that less powerful models may outperform larger ones in tasks requiring strict adherence to source content

References

[1] OpenAI. GPT Models. URL: https://platform.openai.com/docs/models
[2] Meta AI. LLaMA2. URL: https://ai.facebook.com/llama
[3] Mistral. Mistral 7B. URL: https://mistral.ai
[4] Danswer. Open Source RAG Platform.
URL: https://github.com/danswer-ai/danswer
[5] Vespa. Big Data Serving Engine. URL: https://vespa.ai
[6] intfloat. multilingual-e5-large Embeddings.
URL: https://huggingface.co/intfloat/multilingual-e5-large
[7] Exploding Gradients. Ragas.
URL: https://github.com/explodinggradients/ragas