Title: Seven Failure Points When Engineering a Retrieval Augmented Generation System

Authors: Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, Mohamed Abdelrazek

Summary:
Retrieval-Augmented Generation (RAG) systems combine information retrieval with large language models (LLMs) to enhance responses by providing contextually relevant external knowledge. This paper presents an experience report identifying seven key failure points encountered in RAG systems across three domains: research, education, and biomedical applications. The study highlights the practical challenges engineers face when designing and implementing RAG systems, along with lessons learned from real-world deployments.

The identified failure points include issues such as missing content, improper ranking of relevant documents, retrieval failures due to context limitations, incorrect information extraction, formatting errors, inadequate specificity, and incomplete responses. These challenges arise due to the complexity of processing unstructured knowledge, the limitations of LLMs in handling extracted content, and the difficulties in optimizing retrieval and query processes.

The study also explores strategies to mitigate these issues, emphasizing the importance of chunking and embeddings, the trade-offs between RAG and fine-tuning, and the necessity of continuous testing and monitoring. The authors argue that RAG system robustness evolves over time rather than being fully designed from the start. They conclude by proposing research directions to improve RAG implementation, including optimizing chunking strategies, enhancing retrieval accuracy, and developing better evaluation methods for RAG system performance.

Reference:
Barnett, S., Kurniawan, S., Thudumu, S., Brannelly, Z., & Abdelrazek, M. (2024). *Seven Failure Points When Engineering a Retrieval Augmented Generation System*. Proceedings of the 3rd International Conference on AI Engineering — Software Engineering for AI (CAIN 2024). ACM. https://doi.org/10.1145/nnnnnnn.nnnnnnn