

Improving the Reliability of Retrieval-Augmented Generation (RAG) for Large Language Models

Abstract

Large Language Models (LLMs) have demonstrated strong capabilities in understanding and generating natural language, but they can still produce incorrect, unsupported, or hallucinated information. Retrieval-Augmented Generation (RAG) is an approach that improves LLM responses by retrieving relevant information from external knowledge sources before generating an answer. However, the reliability of RAG systems can still be affected by problems such as irrelevant or noisy retrieved information, poor retrieval quality, and unsupported generated answers.

This research will investigate approaches for improving the reliability of Retrieval-Augmented Generation for Large Language Models. The study will review and analyze existing research on RAG, with particular attention to retrieval quality, factuality, faithfulness, and the effect of retrieved information on generated answers. Existing RAG methods and evaluation approaches will be compared to identify their strengths, limitations, and research gaps. Based on the findings, the research will explore possible approaches for making RAG systems more reliable and reducing incorrect or unsupported responses. The presentation will discuss the background of RAG, existing approaches, research problems, research gaps, evaluation methods, and the proposed direction for improving RAG reliability.

Keywords

Retrieval-Augmented Generation (RAG), Large Language Models (LLMs), RAG Reliability, Retrieval Quality, Information Retrieval, Retrieval Noise, Factuality, Faithfulness, Hallucination, Answer Relevance, Context Relevance, RAG Evaluation, Robustness, Knowledge Grounding