ABSTRACT

REDUCING HALLUCINATIONS IN DOMAIN-SPECIFIC QA THROUGH A HYBRID ARCHITECTURE OF FINE-TUNED ENCODER-ONLY AND DECODER-ONLY LARGE LANGUAGE MODELS

by

Izzeldin Rayyan Bastian
21/477936/PA/20718

Retrieval-Augmented Generation (RAG) is commonly used in LLM-based question-answering systems to include external domain knowledge without additional fine-tuning. However, implementations that only rely on the generative model to interpret retrieved information remain prone to factual contradictions, fabrication, and context inconsistency due to the probabilistic nature of deep learning models. This research examines whether separating reasoning from generation can reduce these issues by proposing a RAG-based architecture that uses a dedicated extractive question-answering module.

The proposed system follows a modular RAG architecture, where answer selection is handled by an encoder-only model performing extractive QA over the retrieved context. A decoder-only model then rewrites the extracted span into a natural response. Retrieval used standard dense vector search, *roberta-large-squad2* for extraction, and *TinyLlama-1.1B-Chat-v1.0* for generation. Both models were fine-tuned at their respective tasks, with full-model fine-tuning for the encoder and LoRA applied to the decoder. The selected domain for fine-tuning and evaluation was the Computer Science major at Universitas Gadjah Mada.

Evaluation on a 207-sample test set showed statistically significant improvements over a baseline Naive RAG, suggesting improved factuality and faithfulness. LERC, a metric that approximates human judgment of response correctness, improved from 2.63 to 2.99, which is an 8.79% increase on the 1-5 scale. BERTScore F1 improved from 0.89 to 0.93, a 4.08% increase on the 0-1 scale. NLI probabilities also shifted from 40.95% to 62.01% for entailment, 51.57% to 32.26% for neutral, and 7.47% to 5.63% for contradiction.

Keywords: Large Language Models, Retrieval-Augmented Generation, Closed-Domain QA, Extractive QA, Hallucination Reduction, Domain Adaptation