Question Answering System for Quran, Tafseer and Ahadith

Abstract

Grounded Language Generation (GLG) is a cornerstone of trustworthy artificial intelligence, ensuring that model-generated responses are not only coherent but also verifiable against authoritative sources. The importance of GLG becomes paramount in high-stakes domains such as religion, law, and medicine, where interpretive drift or unsupported claims can cause serious consequences. Among these, *Islamic scholarship* poses unique challenges: its textual corpus is vast, multi-sourced, and semantically interdependent, encompassing the Qur'an, Tafseer, Hadith, and jurisprudential works. Producing long-form, contextually precise, and faithfully grounded answers in this domain requires deep semantic understanding, multi-source integration, and transparent evidence linkage.

This dissertation investigates how Long-Form Question Answering (LFQA) systems can generate faithful, traceable, and semantically grounded answers in such complex and sensitive domains. It progresses through three major contributions, each representing a chapter-level milestone.

The first contribution introduces **LCQA-Islamic**, a large-scale benchmark of more than 73,000 question–answer pairs, each linked with authoritative context passages from multiple Islamic sources. This chapter details the data acquisition, context engineering, and validation pipeline that converts standard QA pairs into question – context – answer triplets, enabling both supervised fine-tuning and retrieval-augmented generation. The benchmark establishes the first standardized evaluation protocol for long-form QA in the religious domain, incorporating automatic and human assessment of semantic alignment, faithfulness, and interpretability.

The second contribution benchmarks domain-adapted transformer models such as T5, BART, and Falcon on LCQA-Islamic. Experimental analyses reveal that although these models produce fluent responses, they often diverge from retrieved evidence and hallucinate unverifiable content. The findings underscore a critical limitation of monolithic architectures when addressing semantically dense and multi-source contexts.

Building on these insights, the third and central contribution presents **CARGO++ (Citation-Aware Routing for Grounded Output)**—a modular retrieval-augmented generation (RAG) framework that enforces both context precision and citation fidelity. CARGO++ partitions the corpus into semantically coherent retrievers (Qur'an,

Tafseer, Hadith, and jurisprudence) and employs Dynamic Multi-Select Routing (DMSR) to direct each query toward the most relevant retrievers. This routing mechanism mitigates the effect of verbosity or noise in queries by prioritizing the retrieval of semantically aligned evidence. The Reference-Preserving Chunking (RPC) mechanism further ensures that retrieved passages retain their citation structure, enabling explicit traceability of each claim in the generated answer.

Comprehensive evaluation demonstrates consistent and significant improvements across automatic and human metrics. CARGO++ raises BERTScore from 0.73 to 0.81 and rAGAs grounding score from 0.61 to 0.74, while human evaluation indicates a 23% reduction in hallucination and a rise in citation fidelity from 0.62 to 0.88. Annotators also rate it higher in coherence (0.92 vs. 0.38), correctness (0.98 vs. 0.71), and helpfulness (0.98 vs. 0.70) compared to baseline systems.

To assess its generalizability, the framework is further applied to BibleQA and Legal-UQA datasets, representing distinct textual and reasoning paradigms. In both domains, CARGO++ delivers remarkable performance, maintaining high grounding and low hallucination rates without domain-specific retraining. These results validate its adaptability to varied high-stakes contexts where verifiable evidence and interpretability are essential.

Collectively, the findings demonstrate that structured retrieval routing and evidence-linked generation can substantially enhance the faithfulness, transparency, and usability of long-form QA systems. Beyond the immediate application in Islamic scholarship, this research establishes CARGO++ as a reusable framework for citation-aware generation in any domain where the credibility of answers depends on traceable evidence. It thus lays the methodological and empirical foundation for future advancements in responsible and grounded language generation, bridging the gap between fluency, faithfulness, and factual trust.

Keywords: Retrieveal Augmented Generation (RAG), Long Form Question Answering (LFQA), Grounded Language Generation (GLG), Faithfulness, Large Language Models (LLMs), Hallucination, Citation Fidelity