SOVEREIGN: What is the impact of evidence gap identification mechanisms in FAIR-RAG on downstream task performance measur

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

The advent of Large Language Models (LLMs) has revolutionized Natural Language Processing, yet their application in high-stakes, specialized domains like religious question answering is hindered by challenges like hallucination and unfaithfulness to authoritative sources. This issue is particularly critical for the Persian-speaking Muslim community, where accuracy and trustworthiness are paramount. Existing Retrieval-Augmented Generation (RAG) systems, relying on simplistic single-pass pipelines, fall short on complex, multi-hop queries requiring multi-step reasoning and evidence aggregation.

1 Introduction

Analysis of: FARSIQA: Faithful and Advanced RAG System for Islamic Question Answering. Research goal: What is the impact of evidence gap identification mechanisms in FAIR-RAG on downstream task performance measured through F1 scores and exact match accuracy on complex multi-hop QA benchmarks like HotpotQA and 2WikiMultihopQA?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

9 papers retrieved. 10 claims extracted, 4 verified. Tribunal: 6.5/10 → APPROVE (revision_round=1). Policy: ESCALATE_TO_OWNER.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
FARSIQA achieves an Answer Correctness score of 74.3% as evaluated via LLM-as-Judge on multi-hop queries.	×	0.14
FARSIQA substantially outperforms standard baselines across metrics of relevance, correctness, and robustness.	×	0.04
FARSIQA is built upon the FAIR-RAG architecture: a Faithful, Adaptive, Iterative Refinement framework for RAG.	✓	0.26
FAIR-RAG employs a dynamic, self-correcting process that adaptively decomposes complex queries, critically assesses the	✓	0.28
FARSIQA operates on a curated knowledge base of over one million documents from authoritative Islamic sources.	✓	0.16
FARSIQA achieves a 97.0% in Negative Rejection, a 40-point improvement over standard baselines.	✓	0.18
The knowledge base includes approximately 431,000 unique documents from eleven reputable online Persian Islamic encyclop	×	0.07
The knowledge base includes approximately 304,000 question-answer pairs from religious Q&A platforms IslamQuest.net and	×	0.05
All sources were crawled ethically, respecting copyrights and ensuring diversity across Islamic perspectives to mitigate	×	0.02
A recursive chunking strategy was applied to encyclopedic articles, first splitting by paragraphs to maintain semantic c	×	0.02

References

- http://arxiv.org/abs/2404.14464v1
- http://arxiv.org/abs/2510.22344v1
- http://arxiv.org/abs/2510.25621v1