SOVEREIGN: Question Decomposition for Retrieval-Augmented Generation

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 2026

Abstract

Grounding large language models (LLMs) in verifiable external sources is a well-established strategy for generating reliable answers. Retrieval-augmented generation (RAG) is one such approach, particularly effective for tasks like question answering: it retrieves passages that are semantically related to the question and then conditions the model on this evidence. However, multi-hop questions, such as"Which company among NVIDIA, Apple, and Google made the biggest profit in 2023?,"challenge RAG because relevant facts are often distributed across multiple documents rather than co-occurring in on

1 Introduction

Analysis of: Question Decomposition for Retrieval-Augmented Generation. Research goal: What is the accuracy drop of LLM-based answer generation in multi-hop RAG systems when using adversarial query perturbations (e.g., synonym substitution, negation) compared to single-hop queries, evaluated on the BEIR benchmark with dense vs. sparse retrievers?.

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

14 papers retrieved. 6 claims extracted, 5 verified. Tribunal: 7.8/10 $\rightarrow$ APPROVE (revision_round=0). Policy: AUTO_APPROVE.

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
Retrieval-augmented generation (RAG) is a strategy for generating reliable answers by grounding large language models in	✓	0.26
Multi-hop questions, such as 'Which company among NVIDIA, Apple, and Google made the biggest profit in 2023?', challenge	✓	0.39
The proposed RAG pipeline incorporates question decomposition: an LLM decomposes the original query into sub-questions,	✓	0.41
Question decomposition effectively assembles complementary documents, while reranking reduces noise and promotes the mos	✓	0.37
Pairing an off-the-shelf cross-encoder reranker with LLM-driven question decomposition bridges the retrieval gap on mult	✓	0.44
The approach is evaluated on the MultiHop dataset.	×	0.08

References

- http://arxiv.org/abs/2502.11228v2
- https://www.semanticscholar.org/paper/a9bef6726f9e1be4e75376e0b4bb53945650971e
- https://www.semanticscholar.org/paper/dbfc272ba0ca458c85616ecc11e1faf5e9ce4c75