Language Models and Multi-Hop Reasoning in Scientific Question Answering with MuISQA

Assignee Research

June 5, 2026

Abstract

This report synthesises findings from 2 peer-reviewed papers addressing the following research question: How do language models handle multi-hop reasoning chains in scientific question answering v8. 19 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: MuISQA: Multi-Intent Retrieval-Augmented Generation for Scientific Question Answering. Research question: How do language models handle multi-hop reasoning chains in scientific question answering v8.

2 Methodology

Systematic literature search across multiple databases yielded 2 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.3/10.

3 Results

2 papers retrieved. 19 claims extracted; 3 independently verified. Quality review score: 5.3/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce

errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The MuISQA method consistently outperforms conventional approaches on the MuISQA benchmark and other general RAG dataset	✓	0.27
The MuISQA method demonstrates superior performance particularly in retrieval accuracy and evidence coverage compared to	✓	0.20
TriviaQA (Joshi et al., 2017) and Natural Questions (Kwiatkowski et al., 2019) annotate a single gold span per query.	×	0.02
Metrics like nDCG and Recall@K are designed for datasets that annotate a single gold span per query.	×	0.03
Existing RAG systems tend to focus on one dominant answer, repeatedly retrieving redundant evidence while overlooking co	×	0.08
The MuISQA benchmark covers five scientific domains: biology, chemistry, geography, medicine, and physics.	×	0.06
In the MuISQA benchmark, each question is annotated for diverse sub-intents and their corresponding answers.	×	0.13
MuISQA introduces evaluation metrics across three dimensions: Query formulation, Passage retrieval, and Answer generatio	×	0.04
The MuISQA Query formulation metric measures the ability to capture distinct intents.	×	0.03
The MuISQA Passage retrieval metric assesses coverage over different subtopics.	×	0.06
The MuISQA Answer generation metric evaluates both accuracy and completeness of final responses.	×	0.04
The proposed intent-aware retrieval framework uses LLMs to hypothesize potential answers and decompose them into intent-	✓	0.27
Traditional query-rewriting methods generate semantically similar variants, whereas the proposed approach injects distin	×	0.00
HyDE (Gao et al., 2023a) relies on a single synthetic passage to guide retrieval.	×	0.02
The proposed method promotes retrieval diversification by decomposing multiple hypotheses into independent queries.	×	0.04
Retrieved chunks in the proposed method are aggregated and re-ranked using Reciprocal Rank Fusion (RRF).	×	0.15
The paper introduces a metric called vector entropy to quantify the informational complexity of query representations.	×	0.00
Vector entropy is computed by normalizing embedding vectors into a probability distribution	×	0.00

References

- http://arxiv.org/abs/2511.16283v1
- http://arxiv.org/abs/2601.11327v2