Language Models in Multi-Hop Scientific Reasoning: A Systematic Synthesis

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How do language models handle multi-hop reasoning chains in scientific question answering v13. 14 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Automatic Inter-document Multi-hop Scientific QA Generation. Research question: How do language models handle multi-hop reasoning chains in scientific question answering v13.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.0/10.

3 Results

13 papers retrieved. 14 claims extracted; 0 independently verified. Quality review score: 3.0/10.

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
Each query in the IM-SciQA dataset is associated with a predefined paper cluster containing exactly 30 papers.	×	0.09
Each paper cluster in IM-SciQA consists of one gold target paper and 29 distractor papers.	×	0.07
The full corpus used for the 'w/o Cluster' retrieval setting contains 8,211 papers.	×	0.04
The retrieval evaluation uses only the Retrieval Question as input, excluding the Combined QA.	×	0.05
All retrieval experiments were conducted on an NVIDIA A100 (80GB) GPU.	×	0.04
Retrieval performance under the paper cluster setting (N=30) is measured using Hit@1, Hit@3, and MRR@5.	×	0.03
Retrieval performance under the full paper setting (N=8211) is measured using Hit@1, Hit@50, and MRR@50.	×	0.03
In the Realistic Setting for IM-QA evaluation, the top-1 retrieved document is provided as the Target Paper.	×	0.08
In the Oracle Setting for IM-QA evaluation, the gold Target Paper is supplied regardless of retrieval outcomes.	×	0.05
Human Experts achieved an Accuracy of 0.975 in the Oracle Setting.	×	0.04
The model gpt-5 achieved an Accuracy of 0.835 in the Oracle Setting.	×	0.03
The model DeepSeek-R1 achieved an Accuracy of 0.820 but an F1 score of 0.150 in the Oracle Setting.	×	0.04
The criterion 'Cross-reference Necessity' evaluates whether both papers are essential to answer the Complete QA.	×	0.05
The criterion 'Relational Appropriateness' evaluates whether two papers are conceptually suitable to be combined into a	×	0.04

References

- http://arxiv.org/abs/2603.14257v1
- http://arxiv.org/abs/2510.25621v1
- http://arxiv.org/abs/2404.14464v1