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

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: How do language models handle multi-hop reasoning chains in scientific question answering v10. 11 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: CaresAI at BioCreative IX Track 1 – LLM for Biomedical QA. Research question: How do language models handle multi-hop reasoning chains in scientific question answering v10.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.8/10.

3 Results

11 papers retrieved. 11 claims extracted; 0 independently verified. Quality review score: 3.8/10.

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
Exact Match (EM) is a metric where a prediction is considered correct if it exactly matches the gold standard answer aft	×	0.04
Concept-level evaluation assesses whether the submitted answer is semantically equivalent to the gold answer, regardless	×	0.05
On the validation set, all three approaches achieved approximately 0.5 in EM score and around 0.8 in concept-level accur	×	0.05
Zero-shot inference using general-purpose models such as LLaMA 8B and Qwen Instruct 7B resulted in near-zero EM scores.	×	0.06
LLaMA 8B often failed to follow the prompt to generate a short 1–2 phrase response, producing verbose or off-topic compl	×	0.04
Qwen 7B Instruct demonstrated better adherence to the prompt format and instruction following.	×	0.01
Both LLaMA 8B and Qwen 7B occasionally managed to answer surface-level factoid questions but frequently hallucinated ans	×	0.06
Questions asking about the appropriate medical field or specialty were answered with approximately 80% accuracy by LLaMA	×	0.02
In the testing phase, the first approach yielded an EM score of 0.2, while the second and third approaches achieved an E	×	0.03
The two-stage inference pipeline involves generating an initial response to the question and then extracting the exact a	×	0.09
If the model fails to produce a valid short answer across three sampled attempts, the system defaults to using the longe	×	0.05

References

- http://arxiv.org/abs/2504.19565v3
- http://arxiv.org/abs/2509.00806v1
- http://arxiv.org/abs/2507.16746v2