QA-Prompting Outperforms State-of-the-Art Strategies for Graduate-Level Science Questions

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: What prompting strategies maximize language model accuracy on graduate-level science questions. 16 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: QA-prompting: Improving Summarization with Large Language Models using Question-Answering. Research question: What prompting strategies maximize language model accuracy on graduate-level science questions.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

15 papers retrieved. 16 claims extracted; 1 independently verified. Quality review score: 4.5/10.

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
QA-prompting outperforms baseline and other state-of-the-art methods, achieving up to 29% improvement in ROUGE scores.	✓	0.30
The evaluation metrics used are ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore F1 (with Roberta large).	×	0.03
BERTScore analysis is important as LM outputs sometimes don't match at the n-gram level but are still semantically corre	×	0.02
ROUGE has known limitations, and some works have emphasized using LLM as a judge to evaluate.	×	0.02
The results were analyzed on values of k from 0 to 5.	×	0.02
QA-prompting consistently outperforms vanilla prompting and in-context learning (ICL).	×	0.13
Mistral-7B and Gemma-3-12B perform unacceptably bad using vanilla prompting but improve with in-context examples.	×	0.05
For small models ($\leq$1B) like Llama-3.2-1B and Qwen-2.5-0.5B, the ROUGE-L gain from ICL to QA-prompting is only 13.02% and	×	0.04
The increase between ICL to QA-prompting remains high, with 18.29% gain on Gemma-3-12B and 29.75% gain on the Llama-3.1-	×	0.05
Mistral-7B and Qwen-2.5-7B witness an increase of 17.10% and 22.95% respectively with QA-prompting.	×	0.03
The optimal k roughly increases as model complexity increases.	×	0.02
Increasing model size increases the extent of improvement with QA-prompting.	×	0.07
The first step in QA-prompting is to sample candidate questions that will aid the generation of effective summaries.	×	0.08
A set of 10 manually crafted questions are used as a starting point for sampling candidate questions.	×	0.01
The overlap precision Pi(r, ai) is used to rank the candidate questions.	×	0.02
The overlap precision Pi(r, ai) is defined as the ratio of the number of intersecting words in the generated answer ai a	×	0.02

References

- http://arxiv.org/abs/2505.14347v2
- http://arxiv.org/abs/2605.07422v1
- http://arxiv.org/abs/2510.18892v1