Retrieval-Augmented Language Models in Knowledge-Intensive Task Performance

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: How does retrieval augmentation improve language model performance on knowledge-intensive tasks v18. 8 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: Knowledge-Augmented Reasoning Distillation for Small Language Models in Knowledge-Intensive Tasks. Research question: How does retrieval augmentation improve language model performance on knowledge-intensive tasks v18.

2 Methodology

Systematic literature search across multiple databases yielded 16 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

16 papers retrieved. 8 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
The MedQA-USMLE dataset contains 12,723 4-option multiple-choice question answering problems from US medical licensing e	×	0.10
98% of the questions in the MedQA-USMLE dataset simulate realistic clinical settings by presenting patient cases that re	×	0.06
The StrategyQA dataset involves 2,780 yes/no questions that demand sophisticated multi-step reasoning skills and the abi	×	0.03
The OpenbookQA dataset consists of 5,957 elementary-level science questions with 4 multiple-choice options.	×	0.03
KARD consistently outperforms all the baselines on the MedQA-USMLE dataset on both encoder-decoder (Flan-T5) and decoder	×	0.08
KARD exhibits a substantial positive effect on smaller models, as evident from the significant performance gain of the F	×	0.10
Gluco corti coids are the first -line trea tment for Graves' dis ease.	×	0.00
The best sin gle treat ment op tion for Graves' dis ease is me thi maz ole.	×	0.01

References

- http://arxiv.org/abs/2502.20988v2
- http://arxiv.org/abs/2503.16581v1
- http://arxiv.org/abs/2305.18395v2