Language Models vs. Human Experts on Professional Knowledge Benchmarks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How do language models compare to human experts on professional knowledge and science benchmarks v14. 11 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Knowledge-Driven Agentic Scientific Corpus Distillation Framework for Biomedical Large Language Models Training. Research question: How do language models compare to human experts on professional knowledge and science benchmarks v14.

2 Methodology

Systematic literature search across multiple databases yielded 12 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

12 papers retrieved. 11 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
The framework trains a large language model as an automatic evaluation agent using knowledge-guided cold-start prefere	×	0.08
The evaluation agent LLM is optimized by minimizing the negative log-likelihood of the preferred candidate pair.	×	0.04
The training process for the Question Evaluation Agent does not require human-annotated data.	×	0.06
The Answer Generation Agent uses GPT-4o to generate answers for identified optimal question-context pairs.	×	0.04
The m-KAILIN framework optimizes the question generator through Direct Preference Optimization (DPO).	×	0.02
The framework initializes two distinct Question Generation Agents: one based on BioMistral (domain-specific) and one bas	×	0.06
The Question Generation Agent is fine-tuned on the BioASQ QA dataset.	×	0.05
The objective function for fine-tuning the question generator is cross-entropy loss defined as the negative log-likeliho	×	0.03
During inference, the trained question generator samples biomedical documents from PubMed to generate candidate question	×	0.05
The benchmark results show a score of 72.8 for lpaca(7B).	×	0.03
The benchmark results show a score of 74.6 for MEDBITRON (Llama 7B).	×	0.02

References

- http://arxiv.org/abs/2404.17000v1
- http://arxiv.org/abs/2504.19565v3
- http://arxiv.org/abs/2509.12382v1