Iterative Consensus Ensemble Accuracy Gains on GPQA Diamond Across Frontier LLMs

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the Iterative Consensus Ensemble (ICE) framework impact accuracy on the GPQA Diamond benchmark compared to standard chain-of-thought prompting across diverse frontier LLMs. 12 claims were extracted from source literature; 9 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 7.4/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Large language models encode clinical knowledge. Research question: How does the Iterative Consensus Ensemble (ICE) framework impact accuracy on the GPQA Diamond benchmark compared to standard chain-of-thought prompting across diverse frontier LLMs?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.4/10.

3 Results

14 papers retrieved. 12 claims extracted; 9 independently verified. Quality review score: 7.4/10.

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
MultiMedQA is a benchmark combining six existing medical question answering datasets and a new dataset called HealthSear	✓	0.22
HealthSearchQA is a new dataset of medical questions searched online.	✓	0.21
The proposed human evaluation framework assesses model answers along axes including factuality, comprehension, reasoning	✓	0.23
PaLM is a 540-billion parameter large language model.	✓	0.20
Flan-PaLM is the instruction-tuned variant of PaLM.	✓	0.16
Flan-PaLM achieves state-of-the-art accuracy on the MedQA, MedMCQA, PubMedQA, and MMLU clinical topics datasets within M	✓	0.29
Flan-PaLM achieved 67.6% accuracy on the MedQA dataset.	×	0.14
Flan-PaLM's accuracy on MedQA surpasses the prior state of the art by more than 17%.	✓	0.17
Human evaluation of Flan-PaLM revealed key gaps in performance despite high multiple-choice accuracy.	✓	0.15
Instruction prompt tuning is a parameter-efficient approach for aligning LLMs to new domains using a few exemplars.	✓	0.28
Med-PaLM is the model resulting from applying instruction prompt tuning.	×	0.13
Med-PaLM performs inferior to clinicians according to the study's evaluation.	×	0.12

References

- https://doi.org/10.48550/arxiv.2204.14198
- https://doi.org/10.1038/s41586-023-06291-2
- https://doi.org/10.1101/2024.12.25.24319629