Language Models and Human Experts on Professional Knowledge Benchmarks

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: How do language models compare to human experts on professional knowledge and science benchmarks v18. 8 claims were extracted from source literature; 1 was 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: Is Self-knowledge and Action Consistent or Not: Investigating Large Language Model's Personality. Research question: How do language models compare to human experts on professional knowledge and science benchmarks v18.

2 Methodology

Systematic literature search across multiple databases yielded 4 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

4 papers retrieved. 8 claims extracted; 1 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
Human respondents show a basic consistency in self-knowledge and an ability to align self-knowledge with action in real-	✓	0.20
The average Cosine Similarity and Spearman's Rank Correlation Coefficient for LLMs are substantially below those of huma	×	0.03
The Value Mean Difference for LLMs averages around 1.52, indicating a substantial divergence in self-knowledge between t	×	0.09
For most LLMs, the proportion of consistent pairs falls below 55%.	×	0.04
16, 10, and 35 native Chinese speakers were recruited as reviewers and respondents based on their IELTS, CET 6 exam resu	×	0.01
16 individuals with distinct MBTI types were recruited for the study.	×	0.03
Reviewers were asked to rate whether practical scenario cases were consistent with their corresponding personality knowl	×	0.06
The study used the 16 Personalities Test, MBTI-M Test, and TDA-100 Test for personality assessment.	×	0.04

References

- http://arxiv.org/abs/2405.11357v3
- http://arxiv.org/abs/2403.09676v1
- http://arxiv.org/abs/2402.14679v2