Language Models vs. Human Experts on Professional Knowledge and Science Benchmarks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How do language models compare to human experts on professional knowledge and science benchmarks v9. 15 claims were extracted from source literature; 7 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.6/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Can GPT-4o Evaluate Usability Like Human Experts? A Comparative Study on Issue Identification in Heuristic Evaluation. Research question: How do language models compare to human experts on professional knowledge and science benchmarks v9.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.6/10.

3 Results

13 papers retrieved. 15 claims extracted; 7 independently verified. Quality review score: 6.6/10.

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
GPT-4o identified 21.2% of the issues identified by human experts in a heuristic evaluation.	✓	0.26
GPT-4o found 27 new issues not identified by human experts.	✓	0.22
GPT-4o performed better for heuristics related to aesthetic and minimalist design and match between system and real worl	✓	0.31
GPT-4o had difficulty identifying issues in heuristics related to flexibility, control, and user efficiency.	✓	0.29
GPT-4o generated several false positives due to hallucinations and attempts to predict issues.	✓	0.24
GPT-4o identified a total of 111 issues (mean=5.55, SD=1.14) using 20 screenshots.	×	0.05
43 (38.7%) of the issues identified by GPT-4o were considered duplicates.	×	0.08
27 (24.3%) of the issues identified by GPT-4o were considered false positives.	×	0.12
The sample of consolidated issues identified by GPT-4o was 41 (36.9%).	×	0.08
The study compared the outcomes of a heuristic evaluation performed by human experts with those produced by GPT-4o on th	✓	0.20
The study aimed to address three research questions: agreement between experts and GPT-4o, characteristics of issues fou	×	0.11
The study used guidelines provided by Lazar et al. [20] to define the steps of the methodology.	×	0.02
The study used a mixed analysis method to verify agreement in identifying issues and the characteristics of the issues a	×	0.12
The study compared results from a previously published heuristic evaluation performed by three human experts with result	✓	0.19
The three human experts included two with PhDs in the field of HCI and one master's student who was also a specialist in	×	0.08

References

- http://arxiv.org/abs/2506.16345v1
- http://arxiv.org/abs/2402.14800v2
- http://arxiv.org/abs/2603.12895v1