Language Models vs. Human Experts on Professional Knowledge Benchmarks

Assignee Research

June 7, 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 v17. 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: Human-Centered Evaluation of an LLM-Based Process Modeling Copilot: A Mixed-Methods Study with Domain Experts. Research question: How do language models compare to human experts on professional knowledge and science benchmarks v17.

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
LLMs can support the generation and refinement of process models from textual descriptions.	×	0.05
Commercial tools such as the Camunda BPMN Copilot [6] and BPMN Assistant [20] reflect growing interest in LLM-assisted p	×	0.14
A comprehensive benchmark consisting of 20 diverse business processes was developed to evaluate 16 state-of-the-art LLMs	×	0.05
Process model quality is inherently multidimensional, including pragmatic, human-centered aspects.	×	0.11
Trust in AI systems is a critical factor in effective human-AI collaboration.	×	0.10
Participants expressed moderate confidence that the system could perform the task better than an inexperienced human use	×	0.02
Reliability received the lowest rating (Q3: M=1.8, SD=0.45), indicating that the participants could not consistently rel	×	0.01
Predictability was also rated low (Q2: M=2.4, SD=0.55), as was the sense of security when relying on the system (Q4: M=2	×	0.01
The system's suitability for decision-making received moderate ratings (Q8: M=2.6, SD=0.89).	×	0.01
Participants showed low suspicion of the system (Q6: M=3.8, SD=1.10) and moderate efficiency ratings (Q5: M=3.2, SD=0.84	×	0.02
The mean score for the tool-specific quality assessment was 54.4% (SD=17.8%), indicating moderate performance.	×	0.06

References

- http://arxiv.org/abs/2504.19565v3
- http://arxiv.org/abs/2603.12895v1
- http://arxiv.org/abs/2408.06717v3