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

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: How do language models compare to human experts on professional knowledge and science benchmarks. 13 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Signal or Noise? Evaluating Large Language Models in Resume Screening Across Contextual Variations and Human Expert Benchmarks. Research question: How do language models compare to human experts on professional knowledge and science benchmarks.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.0/10.

3 Results

16 papers retrieved. 13 claims extracted; 0 independently verified. Quality review score: 4.0/10.

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
Robust statistical methodologies are fundamental for evaluating artificial intelligence (AI) systems in recruitment, esp	×	0.10
Analysis of variance (ANOVA) is commonly used to test for statistically significant differences across multiple AI model	×	0.09
Studies comparing machine learning models for recruitment—such as Random Forest, Neural Networks, and Gradient Boosting—	×	0.05
Effect size measures are crucial for understanding the magnitude of observed differences in recruitment AI evaluation.	×	0.04
Paired statistical tests, such as paired t-tests, are valuable for evaluating the impact of specific interventions—like	×	0.06
Techniques such as the false discovery rate (FDR) correction help maintain statistical power while reducing the likeliho	×	0.04
Context sensitivity is crucial for language models in recruitment tasks, affecting their ability to match candidates to	×	0.03
Models struggle with information position in long inputs, performing best when key details are at the beginning or end,	×	0.02
Irrelevant information degrades performance, highlighting the need for careful input curation.	×	0.01
Three large language models were evaluated in this study: Claude (developed by Anthropic), GPT (developed by OpenAI), an	×	0.08
Three experienced human recruitment professionals participated as expert evaluators.	×	0.05
Expert 1 possessed over eight years of experience in recruitment within multinational corporations, with particular expe	×	0.02
Expert 2 specialized in startup recruitment with six years of experience, bringing knowledge of agile hiring practices a	×	0.03

References

- http://arxiv.org/abs/2507.08019v1
- http://arxiv.org/abs/2307.13692v2
- http://arxiv.org/abs/2503.20786v1