Large Language Models vs. Human Experts on Professional Knowledge Benchmarks Microbiome-Specialized LLMs and Human Expertise in Metabolic Science Benchmarking Language Models Against Human Experts in Scientific Knowledge Domains METAB

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 7 peer-reviewed papers addressing the following research question: How do language models compare to human experts on professional knowledge and science benchmarks v20. 10 claims were extracted from source literature; 9 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 7.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Development of Large Language Model Specialized into Microbiome Datasets: an Application of Self-Evaluation and Scoring Comparison with Conventional Natural Language Processing Markers. Research question: How do language models compare to human experts on professional knowledge and science benchmarks v20.

2 Methodology

Systematic literature search across multiple databases yielded 7 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.2/10.

3 Results

7 papers retrieved. 10 claims extracted; 9 independently verified. Quality review score: 7.2/10.

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
The gut microbiome plays a fundamental role in host metabolism, immune regulation, and disease development.	✓	0.27
METABOLISM is a microbiome-specialized Large Language Model (LLM).	✓	0.20
METABOLISM was fine-tuned on 160,000 scientific abstracts.	✓	0.21
METABOLISM was optimized using LoRA-based parameter-efficient training.	✓	0.19
Model performance was evaluated using automated Phi-4 scoring for relevance, informativeness, and fluency.	✓	0.20
Model performance was evaluated using structured human expert rubric assessments involving 20 domain specialists.	✓	0.24
METABOLISM achieved mean relevance and clarity scores greater than 7.5 with a standard deviation of 0.06.	×	0.11
METABOLISM achieved superior relevance and clarity scores compared to Gemma-3-12B-IT and ChatGPT-4o.	✓	0.22
Correlation analysis revealed a relationship of R = –0.65 ($p < 0.0001$) between traditional NLP metrics (BLEU, ROUGE) and	✓	0.26
Traditional NLP metrics (BLEU, ROUGE) show a weak to moderate negative relationship with human expert rubric scores.	✓	0.21

References

- https://www.semanticscholar.org/paper/f287736c43e2de064ee68e9529993a77ad2f4555
- https://arxiv.org/abs/2410.18344
- http://arxiv.org/abs/2603.12895v1