Baichuan-2 Fine-Tuning on Legal and Biomedical Data: TruthfulQA and HellaSwag Performance

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the fine-tuning of Baichuan-2 on in-domain legal datasets compare to biomedical datasets in terms of TruthfulQA alignment scores and reasoning accuracy on the HellaSwag benchmark. Large language models (LLMs) have demonstrated remarkable capabilities across a broad spectrum of tasks. They have attracted significant attention and been deployed in numerous downstream applications. 7 claims were extracted from source literature; 7 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Evaluating Large Language Models: A Comprehensive Survey. Research question: How does the fine-tuning of Baichuan-2 on in-domain legal datasets compare to biomedical datasets in terms of TruthfulQA alignment scores and reasoning accuracy on the HellaSwag benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.0/10.

3 Results

14 papers retrieved. 7 claims extracted; 7 independently verified. Quality review score: 8.0/10.

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
Large language models (LLMs) have demonstrated capabilities across a broad spectrum of tasks.	✓	0.20
LLMs have been deployed in numerous downstream applications.	✓	0.16
LLMs present potential risks including private data leaks.	✓	0.18
LLMs can yield inappropriate, harmful, or misleading content.	✓	0.18
The survey categorizes LLM evaluation into three major groups: knowledge and capability evaluation, alignment evaluation	✓	0.24
The survey collates a compendium of evaluations pertaining to LLMs' performance in specialized domains.	✓	0.20
The survey discusses the construction of comprehensive evaluation platforms covering capabilities, alignment, safety, an	✓	0.23

References

- https://doi.org/10.48550/arxiv.2402.06196
- https://doi.org/10.1093/jamia/ocad259
- https://doi.org/10.48550/arxiv.2310.19736