Scaling Effects of RLHF-Aligned Models on LawBench Legal Knowledge Performance

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: What is the impact of model size scaling (e.g., 7B vs. 13B vs. 30B) on the LawBench benchmark performance of RLHF-aligned models, particularly in the Legal knowledge level, and does the performance. 15 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: LawBench: Benchmarking Legal Knowledge of Large Language Models. Research question: What is the impact of model size scaling (e.g., 7B vs. 13B vs. 30B) on the LawBench benchmark performance of RLHF-aligned models, particularly in the Legal knowledge level, and does the performance gap narrow with increased model capacity?.

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 4.5/10.

3 Results

12 papers retrieved. 15 claims extracted; 1 independently verified. Quality review score: 4.5/10.

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 evaluation process for LawBench involves two steps: answer extraction and metric computation.	×	0.06
Answer extraction is necessary because many LLMs do not generate output directly comparable with gold labels.	×	0.04
For Article Number Extraction, the delimiter " is used to separate the prediction text into chunks, and the cn2an libr	×	0.03
For Prison Term Extraction, Chinese numerals are converted to Arabic numerals, and digits followed by time intervals in	×	0.01
For Criminal Damages Extraction, all numbers appearing in the prediction text are extracted using regular expressions.	×	0.02
For Named-Entity Recognition, entity types are found in the model prediction, and a regular expression is applied to ext	×	0.02
For Trigger Word Extraction, the model prediction is split by the delimiter " and the split array is treated as a list	×	0.03
For Option Extraction, all possible options are checked against the prediction text, and the set of options that occur i	×	0.03
For tasks 1-1, 2-1, 2-5, 2-7, 3-2, and 3-8, the model prediction is taken as the answer without performing any extractio	×	0.04
The benchmark table includes models such as MPT, LLaMA, Alpaca-v1.0, Vicuna-v1.3, WizardLM, StableBeluga2, ChatGPT, GPT-	×	0.03
GPT-4 and ChatGPT are accessed via API, while other models are accessed via weights.	×	0.02
The performance of multilingual LLMs, Chinese-oriented LLMs, and legal-specific LLMs is compared in the benchmark tables	✓	0.22
GPT-4 and ChatGPT show higher performance compared to other models in the benchmark tables.	×	0.04
The performance of models on tasks 1-1, 1-2, 2-1 to 2-10, and 3-2 to 3-8 is detailed in the benchmark tables.	×	0.05
The performance of models in zero-shot and one-shot settings is compared in the benchmark tables.	×	0.08

References

- http://arxiv.org/abs/2508.15478v2
- http://arxiv.org/abs/2308.04332v1
- http://arxiv.org/abs/2309.16289v1