FlowKV Selective Eviction Fine-Tuning for Domain-Specific Robustness in LongBench

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: Can FlowKV's selective eviction be fine-tuned for domain-specific robustness in LongBench, and if so, how does it affect the accuracy decay of Llama-3-70b on domain-shifted tasks (e.g., from legal to. 16 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.6/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: A Llama walks into the 'Bar': Efficient Supervised Fine-Tuning for Legal Reasoning in the Multi-state Bar Exam. Research question: Can FlowKV's selective eviction be fine-tuned for domain-specific robustness in LongBench, and if so, how does it affect the accuracy decay of Llama-3-70b on domain-shifted tasks (e.g., from legal to medical) compared to a domain-agnostic approach?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.6/10.

3 Results

11 papers retrieved. 16 claims extracted; 1 independently verified. Quality review score: 4.6/10.

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
Llama 3 demonstrates an initial accuracy of 0.48 before fine-tuning on the Multi-state Bar Exam dataset.	✓	0.17
The learning curve for Llama 3 performance as a function of training samples has an R-squared value of 0.626.	×	0.02
Llama 2 starts near random chance accuracy but exhibits steady gains with an R-squared value of 0.991.	×	0.02
The human baseline performance on the evaluated task is 0.675.	×	0.03
Llama 3 achieves a maximum accuracy of 0.54 after training on only 20 samples.	×	0.02
Llama 2 achieves a peak accuracy of 0.37 at 225 training samples.	×	0.02
Both Llama 2 and Llama 3 models failed to reach the passing threshold for the bar exam in this study.	×	0.10
Llama 3 performance plateaus beyond 20 samples per domain in the training set.	×	0.04
Katz et al. (2024) did not employ standardized grading rubrics typically used in actual bar examinations for their essay	×	0.02
Katz et al. (2024) based essay comparisons on 'representative' good answers published by the state of Maryland.	×	0.01
Legal experts evaluating responses in Katz et al. (2024) were not specifically experienced in UBE evaluation.	×	0.06
Katz et al. (2024) presented only the best performances of GPT-4 without reporting performance variability across differ	×	0.03
The training dataset consists of previous bar exam questions gathered from online study materials.	×	0.09
The study verified no overlap between the curated SFT dataset and the test set licensed from JD Advising.	×	0.09
Every question in the training set contains the question body, four possible options, and an explanation justifying the	×	0.04
Explanations in the original training dataset are predominantly written in an informal and unstructured tone.	×	0.02

References

- http://arxiv.org/abs/2409.08687v4
- http://arxiv.org/abs/2111.12525v5
- http://arxiv.org/abs/2504.04945v1