Reverse-KL Regularization Effects on LLM Reasoning in Low-Resource MMLU Settings

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What is the impact of the KL-divergence constraint in the reverse-KL regularized contextual bandit formulation on the reasoning performance of aligned LLMs, as measured by the MMLU benchmark in. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Sample-Efficient Alignment for LLMs. Research question: What is the impact of the KL-divergence constraint in the reverse-KL regularized contextual bandit formulation on the reasoning performance of aligned LLMs, as measured by the MMLU benchmark in low-resource settings?.

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 3.5/10.

3 Results

14 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 3.5/10.

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.

References

- http://arxiv.org/abs/2411.01493v2
- http://arxiv.org/abs/2509.16679v1
- http://arxiv.org/abs/2505.17508v4