Head-Tail-Aware KL Divergence Scaling in LLM Distillation and Alignment Metrics

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How does head-tail-aware KL divergence scaling affect alignment metrics in large language models compared to standard KL divergence during distillation. Standard Knowledge Distillation (KD) compresses Large Language Models (LLMs) by optimizing final outputs, yet it typically treats the teacher's intermediate layer's thought process as a black box. While feature-based distillation attempts to bridge this gap, existing methods. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: DistillLens: Symmetric Knowledge Distillation Through Logit Lens. Research question: How does head-tail-aware KL divergence scaling affect alignment metrics in large language models compared to standard KL divergence during distillation?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.0/10.

3 Results

15 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 3.0/10.

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.

References

- http://arxiv.org/abs/2602.13567v1
- http://arxiv.org/abs/2406.19227v1
- http://arxiv.org/abs/2504.20445v2