What is the impact of varying token misalignment thresholds in TAE on downstream task performance (e.g., MMLU,

Assignee Research

May 29, 2026

Abstract

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. Nevertheless, akin to a double-edged sword, LLMs also present potential risks. They could suffer from private data leaks or yield inappropriate, harmful, or misleading content. Additionally, the rapid progress of LLMs raises concerns about the potential emergence of superintelligent systems without adequate safeguards. To effectively capitalize on LLM capacities as well as ensure their safe and

1 Introduction

This paper examines: Evaluating Large Language Models: A Comprehensive Survey. Research question: What is the impact of varying token misalignment thresholds in TAE on downstream task performance (e.g., MMLU, HellaSwag) when applied to both Baichuan 2 and Vicuna-13B models?.

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 8.7/10.

3 Results

14 papers retrieved. 11 claims extracted; 11 independently verified. Quality review score: 8.7/10.

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 remarkable capabilities across a broad spectrum of tasks.	✓	0.26
LLMs have attracted significant attention and been deployed in numerous downstream applications.	✓	0.22
LLMs could suffer from private data leaks or yield inappropriate, harmful, or misleading content.	✓	0.26
The rapid progress of LLMs raises concerns about the potential emergence of superintelligent systems without adequate sa	✓	0.29
To effectively capitalize on LLM capacities as well as ensure their safe and beneficial development, it is critical to c	✓	0.37
This survey endeavors to offer a panoramic perspective on the evaluation of LLMs.	✓	0.26
We categorize the evaluation of LLMs into three major groups: knowledge and capability evaluation, alignment evaluation	✓	0.32
We collate a compendium of evaluations pertaining to LLMs' performance in specialized domains.	✓	0.25
We discuss the construction of comprehensive evaluation platforms that cover LLM evaluations on capabilities, alignment,	✓	0.35
This comprehensive overview will stimulate further research interests in the evaluation of LLMs.	✓	0.27
The ultimate goal is to make evaluation serve as a cornerstone in guiding the responsible development of LLMs.	✓	0.22

References

- https://doi.org/10.48550/arxiv.2307.04657
- https://doi.org/10.48550/arxiv.2310.19736
- https://doi.org/10.48550/arxiv.2402.06196