Does the optimal token misalignment threshold for maximizing alignment safety differ between Baichuan 2 and Vi

Assignee Research

May 29, 2026

Abstract

Large Language Models (LLMs) have drawn a lot of attention due to their strong performance on a wide range of natural language tasks, since the release of ChatGPT in November 2022. LLMs' ability of general-purpose language understanding and generation is acquired by training billions of model's parameters on massive amounts of text data, as predicted by scaling laws \cite{kaplan2020scaling,hoffmann2022training}. The research area of LLMs, while very recent, is evolving rapidly in many different ways. In this paper, we review some of the most prominent LLMs, including three popular LLM families

1 Introduction

This paper examines: Large Language Models: A Survey. Research question: Does the optimal token misalignment threshold for maximizing alignment safety differ between Baichuan 2 and Vicuna-13B when evaluated on harmful content generation datasets?.

2 Methodology

Systematic literature search across multiple databases yielded 9 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 9.0/10.

3 Results

9 papers retrieved. 10 claims extracted; 10 independently verified. Quality review score: 9.0/10.

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.

5 Extracted Claims

Claim	Verified	Confidence
Large Language Models (LLMs) have drawn a lot of attention due to their strong performance on a wide range of natural la	✓	0.40
LLMs' ability of general-purpose language understanding and generation is acquired by training billions of model's param	✓	0.39
The research area of LLMs, while very recent, is evolving rapidly in many different ways.	✓	0.28
This paper reviews some of the most prominent LLMs, including three popular LLM families (GPT, LLaMA, PaLM).	✓	0.28
The paper discusses the characteristics, contributions, and limitations of popular LLMs.	✓	0.16
The paper gives an overview of techniques developed to build and augment LLMs.	✓	0.20
The paper surveys popular datasets prepared for LLM training, fine-tuning, and evaluation.	✓	0.27
The paper reviews widely used LLM evaluation metrics.	✓	0.19
The paper compares the performance of several popular LLMs on a set of representative benchmarks.	✓	0.21
The paper concludes by discussing open challenges and future research directions.	✓	0.18

References

- https://doi.org/10.48550/arxiv.2401.05561
- https://doi.org/10.48550/arxiv.2401.17256
- https://doi.org/10.48550/arxiv.2402.06196