SOVEREIGN: How does the Baichuan 2 model's performance in low-resource inference settings compare to Meta AI's LLaMA-3 on

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Intervening the internal activations of large language models (LLMs) provides an effective inference-time alignment approach to mitigate undesirable behaviors, such as generating erroneous or harmful content, thereby ensuring safe and reliable applications of LLMs.However, previous methods neglect the misalignment discrepancy among varied tokens, resulting in deviant alignment direction and inflexible editing strength.To address these issues, we propose a token-aware editing (TAE) approach to fully utilize token-level alignment information in the activation space, therefore realizing superior

1 Introduction

Analysis of: Token-Aware Editing of Internal Activations for Large Language Model Alignment. Research goal: How does the Baichuan 2 model's performance in low-resource inference settings compare to Meta AI's LLaMA-3 on the TrustLLM benchmark in terms of accuracy and latency trade-offs?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

2 papers retrieved. 11 claims extracted, 8 verified. Tribunal: 7.3/10 $\rightarrow$ REVISE (revision_round=1). Policy: SOFT_APPROVE.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
Intervening the internal activations of large language models (LLMs) provides an effective inference-time alignment appr	✓	0.37
Previous methods neglect the misalignment discrepancy among varied tokens.	✓	0.22
Neglecting token-level misalignment discrepancy results in deviant alignment direction and inflexible editing strength.	✓	0.25
The proposed Token-Aware Editing (TAE) approach utilizes token-level alignment information in the activation space.	✓	0.29
The Mutual Information-guided Graph Aggregation (MIG) module develops an MI-guided graph to exploit tokens' informative	✓	0.35
The MIG module improves alignment probing and facilitates intervention.	×	0.11
The Misalignment-aware Adaptive Intervention (MAI) module perceives token-level misalignment degree from token represent	✓	0.31
The MAI module guides the adaptive adjustment of editing strength.	✓	0.16
Experiments were conducted on three alignment capabilities.	×	0.09
TAE surpasses the baseline by 25.8% on the primary metric of truthfulness.	✓	0.15
TAE achieves performance improvements with minimal cost.	×	0.09

References

- https://doi.org/10.18653/v1/2025.emnlp-main.480
- https://doi.org/10.36227/techrxiv.176620610.03288677/v1