SOVEREIGN: How does Baichuan 2 perform in low-resource inference settings compared to Meta AI's LLaMA-3 in terms of laten

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Large language models (LLMs), exemplified by ChatGPT, have gained considerable attention for their excellent natural language processing capabilities. Nonetheless, these LLMs present many challenges, particularly in the realm of trustworthiness. Therefore, ensuring the trustworthiness of LLMs emerges as an important topic. This paper introduces TrustLLM, a comprehensive study of trustworthiness in LLMs, including principles for different dimensions of trustworthiness, established benchmark, evaluation, and analysis of trustworthiness for mainstream LLMs, and discussion of open challenges and f

1 Introduction

Analysis of: TrustLLM: Trustworthiness in Large Language Models. Research goal: How does Baichuan 2 perform in low-resource inference settings compared to Meta AI's LLaMA-3 in terms of latency and throughput on a standardized benchmark?.

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

13 papers retrieved. 10 claims extracted, 10 verified. Tribunal: 8.7/10 $\rightarrow$ APPROVE (revision_round=0). Policy: AUTO_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
Large language models (LLMs) have gained considerable attention for their excellent natural language processing capabili	✓	0.24
LLMs present many challenges, particularly in the realm of trustworthiness.	✓	0.22
TrustLLM is a comprehensive study of trustworthiness in LLMs, including principles for different dimensions of trustwort	✓	0.45
TrustLLM proposes a set of principles for trustworthy LLMs that span eight different dimensions.	✓	0.25
TrustLLM establishes a benchmark across six dimensions including truthfulness, safety, fairness, robustness, privacy, an	✓	0.27
TrustLLM presents a study evaluating 16 mainstream LLMs in TrustLLM, consisting of over 30 datasets.	✓	0.26
Trustworthiness and utility (i.e., functional effectiveness) are positively related.	✓	0.22
Proprietary LLMs generally outperform most open-source counterparts in terms of trustworthiness.	✓	0.30
A few open-source LLMs come very close to proprietary ones in terms of trustworthiness.	✓	0.29
Some LLMs may be overly calibrated towards exhibiting trustworthiness.	✓	0.18

References

- https://doi.org/10.48550/arxiv.2404.14294
- https://doi.org/10.48550/arxiv.2401.05561
- https://doi.org/10.48550/arxiv.2401.14656