SOVEREIGN: LLaMA-3 evaluation benchmark results MMLU HumanEval GSM8K coding performance Meta AI

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Large language models (LLMs) have demonstrated remarkable performance on a variety of natural language tasks based on just a few examples of natural language instructions, reducing the need for extensive feature engineering. However, most powerful LLMs are closed-source or limited in their capability for languages other than English. In this technical report, we present Baichuan 2, a series of large-scale multilingual language models containing 7 billion and 13 billion parameters, trained from scratch, on 2.6 trillion tokens. Baichuan 2 matches or outperforms other open-source models of simila

1 Introduction

Analysis of: Baichuan 2: Open Large-scale Language Models. Research goal: LLaMA-3 evaluation benchmark results MMLU HumanEval GSM8K coding performance Meta AI.

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

7 papers retrieved. 5 claims extracted, 5 verified. Tribunal: 9.0/10 $\rightarrow$ 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
Baichuan 2 is a series of large-scale multilingual language models containing 7 billion and 13 billion parameters.	✓	0.39
Baichuan 2 is trained from scratch on 2.6 trillion tokens.	✓	0.18
Baichuan 2 matches or outperforms other open-source models of similar size on public benchmarks like MMLU, CMMLU, GSM8K,	✓	0.38
Baichuan 2 excels in vertical domains such as medicine and law.	✓	0.26
All pre-training model checkpoints of Baichuan 2 will be released to benefit the research community.	✓	0.20

References

- https://doi.org/10.48550/arxiv.2406.12793
- https://doi.org/10.1007/s11704-026-60308-3
- https://doi.org/10.48550/arxiv.2309.10305