SOVEREIGN: How does the accuracy of DeepSeek-R1 and o1-preview scale with chain-of-thought length (number of reasoning to

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Despite increasing discussions on open-source Artificial Intelligence (AI), existing research lacks a discussion on the transparency and accessibility of state-of-the-art (SoTA) Large Language Models (LLMs). The Open Source Initiative (OSI) has recently released its first formal definition of open-source software. This definition, when combined with standard dictionary definitions and the sparse published literature, provide an initial framework to support broader accessibility to AI models such as LLMs, but more work is essential to capture the unique dynamics of openness in AI. In addition,

1 Introduction

Analysis of: Comprehensive Analysis of Transparency and Accessibility of ChatGPT, DeepSeek, And other SoTA Large Language Models. Research goal: How does the accuracy of DeepSeek-R1 and o1-preview scale with chain-of-thought length (number of reasoning tokens) across Chinese vs. English legal benchmarks, and what is the token-efficiency trade-off (accuracy per inference cost)?.

2 Methodology

Multi-query arXiv search (1 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

3 papers retrieved. 4 claims extracted, 3 verified. Tribunal: 7.5/10 → 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
Open Source Initiative (OSI) has recently released its first formal definition of open-source software	✓	0.29
Some models labeled as open-source do not report model training data, code, and key metrics such as weight accessibility	✓	0.32
This study examines SoTA LLMs from the last five years including ChatGPT, DeepSeek, LLaMA, and Grok	✓	0.22
This is the first study that sys	×	0.10

References

- https://doi.org/10.20944/preprints202502.1608.v1
- https://doi.org/10.36227/techrxiv.176834456.68708065/v1
- https://doi.org/10.3390/electronics14183583