SOVEREIGN: How do the test-time compute scaling curves (accuracy vs. inference FLOPs) for DeepSeek-R1 and o1-preview diff

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Recent advances in test-time scaling of large language models (LLMs), exemplified by DeepSeek-R1 and OpenAI's o1, show that extending the chain of thought during inference can significantly improve general reasoning performance. However, the impact of this paradigm on legal reasoning remains insufficiently explored. To address this gap, we present the first systematic evaluation of 12 LLMs, including both reasoning-focused and general-purpose models, across 17 Chinese and English legal tasks spanning statutory and case-law traditions. In addition, we curate a bilingual chain-of-thought dataset

1 Introduction

Analysis of: Evaluating Test-Time Scaling LLMs for Legal Reasoning: OpenAI o1, DeepSeek-R1, and Beyond. Research goal: How do the test-time compute scaling curves (accuracy vs. inference FLOPs) for DeepSeek-R1 and o1-preview differ across legal sub-domains (e.g., contract interpretation vs. criminal law) on the multilingual benchmark, and do cross-lingual transfer effects emerge?.

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

10 papers retrieved. 4 claims extracted, 3 verified. Tribunal: 7.6/10 → 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
DeepSeek-R1 demonstrates superior performance in Chinese legal reasoning tasks.	✓	0.19
OpenAI's o1 achieves comparable results on English legal reasoning tasks.	✓	0.28
Legal-R1 outperforms baseline models on the majority of Chinese and English legal tasks.	×	0.15
Legal-R1 addresses the main obstacles in legal-domain LLMs including outdated legal knowledge and factual hallucinations	✓	0.23

References

- http://arxiv.org/abs/2503.16040v2
- http://arxiv.org/abs/2509.22472v1
- http://arxiv.org/abs/2406.14887v1