SOVEREIGN: How does the inference throughput-accuracy trade-off differ between o1-preview and DeepSeek-R1 under constrain

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 does the inference throughput-accuracy trade-off differ between o1-preview and DeepSeek-R1 under constrained token budgets when performing abductive reasoning on legal case summaries, as measured by F1 score and latency per query?.

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

9 papers retrieved. 0 claims extracted, 0 verified. Tribunal: 2.2/10 → REJECT (revision_round=0). Policy: ESCALATE_TO_OWNER.

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.

References

- http://arxiv.org/abs/2503.16040v2
- http://arxiv.org/abs/2304.06912v2
- http://arxiv.org/abs/2310.05276v1