SOVEREIGN: Evaluating Test-Time Scaling LLMs for Legal Reasoning: OpenAI o1, DeepSeek-R1, a

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 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: What is the scaling behavior of test-time compute (chain-of-thought length) versus accuracy gains for DeepSeek-R1 and o1-preview across multilingual legal reasoning tasks (e.g., Chinese vs. English) on the 17-task 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

7 papers retrieved. 8 claims extracted, 8 verified. Tribunal: 8.2/10 → 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 and OpenAI's o1 extend the chain of thought during inference to improve general reasoning performance.	✓	0.29
The paper presents the first systematic evaluation of 12 LLMs across 17 Chinese and English legal tasks.	✓	0.21
The authors curated a bilingual chain-of-thought dataset for legal reasoning through distillation from DeepSeek-R1.	✓	0.26
Legal-R1 is an open-source model specialized for the legal domain.	✓	0.25
Legal-R1 delivers competitive performance across diverse tasks.	✓	0.24
DeepSeek-R1 exhibits clear advantages in Chinese legal reasoning.	✓	0.29
OpenAI's o1 achieves comparable results on English tasks.	✓	0.25
Recurring issues include outdated legal knowledge, limited capacity for legal interpretation, and susceptibility to fact	✓	0.26

References

- http://arxiv.org/abs/2503.16040v2
- http://arxiv.org/abs/2505.18673v1
- http://arxiv.org/abs/2509.22472v1