SOVEREIGN: What is the robustness of test-time scaling gains for o1-preview and DeepSeek-R1 under adversarial legal input

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

In an era dominated by Large Language Models (LLMs), understanding their capabilities and limitations, especially in high-stakes fields like law, is crucial. While LLMs such as Meta's LLaMA, OpenAI's ChatGPT, Google's Gemini, DeepSeek, and other emerging models are increasingly integrated into legal workflows, their performance in multilingual, jurisdictionally diverse, and adversarial contexts remains insufficiently explored. This work evaluates LLaMA and Gemini on multilingual legal and non-legal benchmarks, and assesses their adversarial robustness in legal tasks through character and word-

1 Introduction

Analysis of: Evaluating the Limits of Large Language Models in Multilingual Legal Reasoning. Research goal: What is the robustness of test-time scaling gains for o1-preview and DeepSeek-R1 under adversarial legal input perturbations (e.g., ambiguous statutes, contradictory clauses) on the 17-task benchmark compared to standard prompting?.

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

8 papers retrieved. 4 claims extracted, 0 verified. Tribunal: 4.3/10 → REVISE (revision_round=1). 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.

5 Extracted Claims

Claim	Verified	Confidence
The paper evaluates the performance of large language models using classification metrics including Accuracy, Precision,	×	0.07
The evaluation uses text generation metrics such as ROUGE, BLEU, and METEOR to assess the quality of summaries and free-	×	0.03
The study employs robustness and reliability metrics like variance measures, consistency, entropy, Gini Index, and confi	×	0.03
The study uses LLM-as-judge scores to evaluate quality judgments beyond surface similarity.	×	0.06

References

- http://arxiv.org/abs/2503.16040v2
- http://arxiv.org/abs/2509.22472v1
- http://arxiv.org/abs/2310.05276v1