SOVEREIGN: How does the cross-domain generalization performance of DeepSeek-R1 and o1-preview models vary when evaluated

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 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: How does the cross-domain generalization performance of DeepSeek-R1 and o1-preview models vary when evaluated on multilingual legal benchmarks with different time-compute budgets.

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. 0 claims extracted, 0 verified. Tribunal: 3.5/10 $\rightarrow$ 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/2509.22472v1
- http://arxiv.org/abs/2505.03970v1
- http://arxiv.org/abs/2503.16040v2