SOVEREIGN: Does the adversarial robustness gap between DeepSeek-R1 and o1-preview on legal reasoning tasks generalize to

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

In this report, we introduce the Gemini 2.X model family: Gemini 2.5 Pro and Gemini 2.5 Flash, as well as our earlier Gemini 2.0 Flash and Flash-Lite models. Gemini 2.5 Pro is our most capable model yet, achieving SoTA performance on frontier coding and reasoning benchmarks. In addition to its incredible coding and reasoning skills, Gemini 2.5 Pro is a thinking model that excels at multimodal understanding and it is now able to process up to 3 hours of video content. Its unique combination of long context, multimodal and reasoning capabilities can be combined to unlock new agentic workflows. G

1 Introduction

Analysis of: Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. Research goal: Does the adversarial robustness gap between DeepSeek-R1 and o1-preview on legal reasoning tasks generalize to multimodal code generation benchmarks like APPS when using chain-of-thought prompting and the S* selection mechanism?.

2 Methodology

Multi-query arXiv search (1 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

3 papers retrieved. 0 claims extracted, 0 verified. Tribunal: 1.7/10 → 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/2507.06261v6
- http://arxiv.org/abs/1411.4413v2
- http://arxiv.org/abs/2603.25938v1