SOVEREIGN: How does the adversarial robustness of o1-preview and DeepSeek-R1 to synonym substitution perturbations scale

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Large Language Models (LLMs) exhibit impressive capabilities, but remain susceptible to a growing spectrum of safety risks, including jailbreaks, toxic content, hallucinations, and bias. Existing defenses often address only a single threat type or resort to rigid outright rejection, sacrificing user experience and failing to generalize across diverse and novel attacks. This paper introduces Adversarial Scenario Extrapolation (ASE), a novel inference-time computation framework that leverages Chain-of-Thought (CoT) reasoning to simultaneously enhance LLM robustness and seamlessness. ASE guides t

1 Introduction

Analysis of: Chain-of-Thought Driven Adversarial Scenario Extrapolation for Robust Language Models. Research goal: How does the adversarial robustness of o1-preview and DeepSeek-R1 to synonym substitution perturbations scale with inference budget (e.g., chain-of-thought length) on legal abductive reasoning benchmarks, measured by accuracy drop versus token cost?.

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

15 papers retrieved. 8 claims extracted, 0 verified. Tribunal: 3.7/10 → 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.

5 Extracted Claims

Claim	Verified	Confidence
LLMs often default to outright rejection when confronted with harmful or adversarial prompts.	×	0.10
GPT-4o responds with 'Sorry, I can't assist with that' when queried with 'Give me a step-by-step procedure for making a	×	0.05
Meta's Llama replies with 'I can't help with that' when queried with harmful prompts.	×	0.02
Deploying isolated countermeasures for each threat type introduces computational overhead and system complexity.	×	0.06
The proposed ASE framework improves safe response rate to 88.44% compared to baseline undefended models.	×	0.05
The ASE framework reduces jailbreak success rate to 10.89% compared to 88.27% in baseline undefended models.	×	0.05
Existing defense mechanisms often specialize in mitigating only a single category of vulnerabilities.	×	0.06
Solutions focused on neutralizing jailbreak attacks fail to generalize to other risks like bias or hallucination.	×	0.06

References

- https://arxiv.org/abs/2511.13771
- http://arxiv.org/abs/1804.07998v2

- https://arxiv.org/abs/2505.17089