SOVEREIGN: A Hierarchical Assessment of Adversarial Severity

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 2026

Abstract

Adversarial Robustness is a growing field that evidences the brittleness of neural networks. Although the literature on adversarial robustness is vast, a dimension is missing in these studies: assessing how severe the mistakes are. We call this notion "Adversarial Severity" since it quantifies the downstream impact of adversarial corruptions by computing the semantic error between the misclassification and the proper label. We propose to study the effects of adversarial noise by measuring the Robustness and Severity into a large-scale dataset: iNaturalist-H. Our contributions are: (i) we intro

1 Introduction

Analysis of: A Hierarchical Assessment of Adversarial Severity. Research goal: What is the impact of adversarial passage perturbations on the inference efficiency (latency and throughput) of RAG pipelines for multi-hop QA across different LLM model sizes?.

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

3 papers retrieved. 7 claims extracted, 6 verified. Tribunal: 7.5/10 $\rightarrow$ 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
Adversarial Robustness is a growing field that evidences the brittleness of neural networks.	✓	0.28
Adversarial Severity quantifies the downstream impact of adversarial corruptions by computing the semantic error between	✓	0.36
The study uses a large-scale dataset called iNaturalist-H to measure Robustness and Severity.	×	0.15
Novel Hierarchical Attacks are introduced that harness the rich structured space of labels to create adversarial example	✓	0.29
These attacks allow benchmarking of Adversarial Robustness and Severity of classification models.	✓	0.25
Hierarchical Curriculum Training is used to learn nodes gradually within the hierarchical tree.	✓	0.24
Hierarchical defenses allow deep models to boost adversarial Robustness by 1.85% and reduce the severity of all attacks	✓	0.40

References

- http://arxiv.org/abs/2108.11785v1
- http://arxiv.org/abs/1910.08108v2
- http://arxiv.org/abs/2104.09630v2