SOVEREIGN: To what extent does NOVA's anomaly localization accuracy degrade when tested on out-of-distribution brain MRI

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

In many real-world applications, deployed models encounter inputs that differ from the data seen during training. Out-of-distribution detection identifies whether an input stems from an unseen distribution, while open-world recognition flags such inputs to ensure the system remains robust as ever-emerging, previously $unknown$ categories appear and must be addressed without retraining. Foundation and vision-language models are pre-trained on large and diverse datasets with the expectation of broad generalization across domains, including medical imaging. However, benchmarking these models on t

1 Introduction

Analysis of: NOVA: A Benchmark for Anomaly Localization and Clinical Reasoning in Brain MRI. Research goal: To what extent does NOVA's anomaly localization accuracy degrade when tested on out-of-distribution brain MRI data from different scanner manufacturers and imaging protocols compared to in-distribution performance?.

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

10 papers retrieved. 5 claims extracted, 5 verified. Tribunal: 8.0/10 → 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
NOVA is a benchmark of approximately 900 brain MRI scans that span 281 rare pathologies and heterogeneous acquisition pr	✓	0.26
Each case in NOVA includes rich clinical narratives and double-blinded expert bounding-box annotations.	✓	0.24
NOVA enables joint assessment of anomaly localisation, visual captioning, and diagnostic reasoning.	✓	0.20
NOVA is never used for training, serving as an extreme stress-test of out-of-distribution generalisation.	✓	0.20
Baseline results with leading vision-language models (GPT-4o, Gemini 2.0 Flash, and Qwen2.5-VL-72B) reveal substantial p	✓	0.26

References

- http://arxiv.org/abs/2510.02155v1
- http://arxiv.org/abs/2203.06060v1
- http://arxiv.org/abs/2505.14064v1