SOVEREIGN: NOVA: A Benchmark for Anomaly Localization and Clinical Reasoning in Brain MRI

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 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 expert bridging maintain or improve performance on out-of-distribution samples from Conceptual Captions and LAION datasets compared to full fine-tuning baselines?.

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, 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
NOVA is an evaluation-only benchmark of approximately 900 brain MRI scans.	✓	0.19
The benchmark spans 281 rare pathologies.	×	0.12
Each case includes clinical narratives and double-blinded expert bounding-box annotations.	✓	0.23
NOVA enables joint assessment of anomaly localisation, visual captioning, and diagnostic reasoning.	✓	0.22
NOVA is never used for training.	×	0.14
Baseline results are reported for GPT-4o, Gemini 2.0 Flash, and Qwen2.5-VL-72B.	✓	0.19
Foundation and vision-language models are pretrained on large and diverse datasets with the expectation of broad genera	✓	0.28
Benchmarking models on test sets with only a few common outlier types silently collapses the evaluation back to a closed	✓	0.38

References

- https://www.semanticscholar.org/paper/52fade7eee36054b1919b54d20996dfda07510fc
- https://www.semanticscholar.org/paper/91ebed0cc7b4af570f46f3cdcb11e023fdd77b8c
- http://arxiv.org/abs/2512.13855v2