SOVEREIGN: How does the cross-domain robustness of LLM-based retriever evaluation strategies (e.g., using an LLM as a jud

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0%, respectively, which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully connected layers with a final 1000-way softmax. To make training faster, we used non-saturat

1 Introduction

Analysis of: ImageNet classification with deep convolutional neural networks. Research goal: How does the cross-domain robustness of LLM-based retriever evaluation strategies (e.g., using an LLM as a judge) compare to traditional retrieval metrics (e.g., recall@k) when scaling the number of hops in multi-hop RAG on HotPotQA, measured by correlation with downstream QA accuracy?.

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. 7 claims extracted, 7 verified. Tribunal: 7.0/10 → 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
The neural network has 60 million parameters and 650,000 neurons	✓	0.29
The neural network consists of five convolutional layers, some of which are followed by max-pooling layers, and three fu	✓	0.43
On the ImageNet LSVRC-2010 test data, the model achieved top-1 error rate of 37.5%	✓	0.21
On the ImageNet LSVRC-2010 test data, the model achieved top-5 error rate of 17.0%	✓	0.21
The model achieved a winning top-5 test error rate of 15.3% in the ILSVRC-2012 competition	✓	0.24
The second-best entry in the ILSVRC-2012 competition achieved a top-5 test error rate of 26.2%	✓	0.24
The model was trained to classify 1.2 million high-resolution images into 1000 different classes	✓	0.23

References

- https://doi.org/10.1145/3065386
- https://doi.org/10.3389/neuro.06.004.2008
- https://doi.org/10.3390/app11146421