SOVEREIGN: What is the impact of dynamic routing strategies on the energy efficiency of multimodal model inference on low

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

In the last few years, the deep learning (DL) computing paradigm has been deemed the Gold Standard in the machine learning (ML) community. Moreover, it has gradually become the most widely used computational approach in the field of ML, thus achieving outstanding results on several complex cognitive tasks, matching or even beating those provided by human performance. One of the benefits of DL is the ability to learn massive amounts of data. The DL field has grown fast in the last few years and it has been extensively used to successfully address a wide range of traditional applications. More i

1 Introduction

Analysis of: Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. Research goal: What is the impact of dynamic routing strategies on the energy efficiency of multimodal model inference on low-power GPU accelerators like NVIDIA T4 and L4?.

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

9 papers retrieved. 0 claims extracted, 0 verified. Tribunal: 2.7/10 $\rightarrow$ 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.

References

- https://doi.org/10.1146/annurev-fluid-010719-060214
- https://doi.org/10.1186/s40537-021-00444-8
- https://doi.org/10.1109/jproc.2019.2918951