SOVEREIGN: How does ExpertFlow's predictive expert caching mechanism affect inference latency and memory usage compared t

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

In this paper we report the set-up and results of the Multimodal Brain Tumor Image Segmentation Benchmark (BRATS) organized in conjunction with the MICCAI 2012 and 2013 conferences. Twenty state-of-the-art tumor segmentation algorithms were applied to a set of 65 multi-contrast MR scans of low- and high-grade glioma patients-manually annotated by up to four raters-and to 65 comparable scans generated using tumor image simulation software. Quantitative evaluations revealed considerable disagreement between the human raters in segmenting various tumor sub-regions (Dice scores in the range 74%-85

1 Introduction

Analysis of: The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). Research goal: How does ExpertFlow's predictive expert caching mechanism affect inference latency and memory usage compared to static expert allocation baselines on MMBench and SEED-Bench across different input distributions?.

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. 4 claims extracted, 4 verified. Tribunal: 9.0/10 → 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
Twenty state-of-the-art tumor segmentation algorithms were applied to a set of 65 multi-contrast MR scans of low- and hi	✓	0.36
Quantitative evaluations revealed considerable disagreement between the human raters in segmenting various tumor sub-reg	✓	0.36
Different algorithms worked best for different sub-regions	✓	0.24
No single algorithm ranked in the top for all sub-regions simultaneously	✓	0.23

References

- https://doi.org/10.1109/tmi.2014.2377694
- https://doi.org/10.7551/mitpress/9780262033589.001.0001
- https://doi.org/10.1007/s11704-026-60308-3