SOVEREIGN: Does content-adaptive tokenization enable more efficient inference latency per token processed compared to fix

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

The exceptionally rapid development of highly flexible, reusable artificial intelligence (AI) models is likely to usher in newfound capabilities in medicine. We propose a new paradigm for medical AI, which we refer to as generalist medical AI (GMAI). GMAI models will be capable of carrying out a diverse set of tasks using very little or no task-specific labelled data. Built through self-supervision on large, diverse datasets, GMAI will flexibly interpret different combinations of medical modalities, including data from imaging, electronic health records, laboratory results, genomics, graphs or

1 Introduction

Analysis of: Foundation models for generalist medical artificial intelligence. Research goal: Does content-adaptive tokenization enable more efficient inference latency per token processed compared to fixed-patch approaches on vision-language medical AI benchmarks?.

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

8 papers retrieved. 9 claims extracted, 1 verified. Tribunal: 1.0/10 $\rightarrow$ 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.

5 Extracted Claims

Claim	Verified	Confidence
Foundation models for generalist medical artificial intelligence can carry out diverse medical tasks using minimal task-	✓	0.30
	×	0.00
	×	0.00
	×	0.00
	×	0.00
	×	0.00
	×	0.00
	×	0.00
	×	0.00

References

- https://doi.org/10.1186/s40537-021-00444-8
- https://doi.org/10.1038/s41586-023-05881-4
- https://doi.org/10.1109/5.726791