SOVEREIGN: Does content-adaptive tokenization improve robustness and accuracy on the MMBench and MME benchmarks under var

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Large Vision-Language Models (LVLMs) are capable of handling diverse data types such as imaging, text, and physiological signals, and can be applied in various fields. In the medical field, LVLMs have a high potential to offer substantial assistance for diagnosis and treatment. Before that, it is crucial to develop benchmarks to evaluate LVLMs' effectiveness in various medical applications. Current benchmarks are often built upon specific academic literature, mainly focusing on a single domain, and lacking varying perceptual granularities. Thus, they face specific challenges, including limited

1 Introduction

Analysis of: GMAI-MMBench: A Comprehensive Multimodal Evaluation Benchmark Towards General Medical AI. Research goal: Does content-adaptive tokenization improve robustness and accuracy on the MMBench and MME benchmarks under varying image resolutions compared to fixed-patch baselines when scaling the vision encoder from ViT-L to ViT-g?.

2 Methodology

Multi-query arXiv search (1 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

3 papers retrieved. 10 claims extracted, 0 verified. Tribunal: 1.0/10 → 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
GMAI-MMBench consists of 284 diverse clinical-related datasets from worldwide sources, covering 38 modalities.	×	0.09
GMAI-MMBench includes 18 clinical VQA tasks and 18 clinical departments.	×	0.13
The benchmark provides varying degrees of perceptual details from image to region level.	×	0.05
GMAI-MMBench spans multiple levels of perception including image, region, and pixel levels.	×	0.03
GMAI-MMBench contains 26K data instances.	×	0.04
GMA1-MMBench evaluates 38 data modalities.	×	0.04
The benchmark supports multi-perceptual granularity evaluation from image to region level.	×	0.11
GMAI-MMBench is organized into a lexical tree structure with 18 clinical departments.	×	0.15
GMAI-MMBench contains data from 284 different datasets.	×	0.06
The benchmark includes data from both public and hospital sources.	×	0.03

References

- http://arxiv.org/abs/2010.01177v4
- http://arxiv.org/abs/2408.03361v7
- http://arxiv.org/abs/1710.05833v2