SOVEREIGN: How does content-adaptive tokenization affect the accuracy-efficiency trade-off on the MMMU and MathVista data

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Recent language models have shown impressive multilingual performance, even when not explicitly trained for it. Despite this, there are concerns about the quality of their outputs across different languages. In this paper, we show how disparity in the treatment of different languages arises at the tokenization stage, well before a model is even invoked. The same text translated into different languages can have drastically different tokenization lengths, with differences up to 15 times in some cases. These disparities persist even for tokenizers that are intentionally trained for multilingual

1 Introduction

Analysis of: Language Model Tokenizers Introduce Unfairness Between Languages. Research goal: How does content-adaptive tokenization affect the accuracy-efficiency trade-off on the MMMU and MathVista datasets for multimodal models with 7B and 13B parameter language backbones, measured by FLOPs per sample and exact match scores?.

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

12 papers retrieved. 2 claims extracted, 0 verified. Tribunal: 2.7/10 → 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
A 10-fold reduction in the vocabulary would result in only 30% longer sequences for English	×	0.02
With one-third of the vocabulary, English sequences will become just 10% longer for ChatGPT/GPT-4	×	0.03

References

- http://arxiv.org/abs/2605.17152v1
- http://arxiv.org/abs/2510.13759v3
- http://arxiv.org/abs/2305.15425v2