

SOVEREIGN: What is the impact of privacy-preserving representations on inference latency when scaling model sizes from 7B

SOVEREIGN Research Kernel

Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Abstract The rapid evolution of large language models (LLMs) has driven a transformative shift in artificial intelligence (AI), reshaping both research paradigms and practical applications. Distinguished from their predecessors by unprecedented scale and advanced capabilities, LLMs necessitate new frameworks for understanding their development, behavior, and societal impact. This survey systematically reviews recent advancements in LLM techniques across four key dimensions: (1) pre-training methodologies, which establish core model capabilities through large-scale self-supervised training, arc

1 Introduction

Analysis of: A Survey of Large Language Models. Research goal: What is the impact of privacy-preserving representations on inference latency when scaling model sizes from 7B to 70B parameter models for legal domain applications?.

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. 0 claims extracted, 0 verified. Tribunal: 6.5/10 \$\rightarrow\$ APPROVE (revision_round=1). 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/10.1109/jproc.2021.3060483>
- <https://doi.org/10.1007/s11704-026-60308-3>
- <https://doi.org/10.1561/22000000083>