SOVEREIGN: What is the impact of domain-specific fine-tuning on BEIR-NL datasets on downstream task performance measured

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Since the inception of the Transformer architecture in 2017, Large Language Models (LLMs) such as GPT and BERT have evolved significantly, impacting various industries with their advanced capabilities in language understanding and generation. These models have shown potential to transform the medical field, highlighting the necessity for specialized evaluation frameworks to ensure their effective and ethical deployment. This comprehensive survey delineates the extensive application and requisite evaluation of LLMs within healthcare, emphasizing the critical need for empirical validation to ful

1 Introduction

Analysis of: A Comprehensive Survey on Evaluating Large Language Model Applications in the Medical Industry. Research goal: What is the impact of domain-specific fine-tuning on BEIR-NL datasets on downstream task performance measured by MRR and R@100 improvements?.

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. 0 claims extracted, 0 verified. Tribunal: 2.0/10 $\rightarrow$ 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.

References

- https://doi.org/10.18653/v1/2023.eacl-main.203
- https://doi.org/10.17863/cam.30462
- https://doi.org/10.48550/arxiv.2404.15777