# SOVEREIGN: How does model accuracy and inference efficiency vary across different expert counts in sparse multimodal mode

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

**Abstract**

The deployment of large language models (LLMs) within the healthcare sector has sparked both enthusiasm and apprehension. These models exhibit the remarkable ability to provide proficient responses to free-text queries, demonstrating a nuanced understanding of professional medical knowledge. This comprehensive survey delves into the functionalities of existing LLMs designed for healthcare applications and elucidates the trajectory of their development, starting with traditional Pretrained Language Models (PLMs) and then moving to the present state of LLMs in the healthcare sector. First, we ex

## 1 Introduction

Analysis of: Large Language Models in Healthcare and Medical Domain: A Review. Research goal: How does model accuracy and inference efficiency vary across different expert counts in sparse multimodal models on VQAv2 and OK-VQA 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

10 papers retrieved. 8 claims extracted, 8 verified. Tribunal: 7.8/10 $\rightarrow$ APPROVE (revision_round=0). Policy: AUTO_APPROVE.

## 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 |
|---|---|---|
| Large language models (LLMs) can provide proficient responses to free-text queries and demonstrate a nuanced understandi | ✓ | 0.29 |
| The development of LLMs for healthcare applications started with traditional Pretrained Language Models (PLMs) and then | ✓ | 0.26 |
| LLMs can amplify the efficiency and effectiveness of diverse healthcare applications, particularly clinical language und | ✓ | 0.30 |
| Clinical language understanding tasks include named entity recognition, relation extraction, natural language inference, | ✓ | 0.33 |
| The paper conducts an extensive comparison of the most recent state-of-the-art LLMs in the healthcare domain. | ✓ | 0.23 |
| The paper assesses the utilization of various open-source LLMs and highlights their significance in healthcare applicati | ✓ | 0.19 |
| The paper presents essential performance metrics used to evaluate LLMs in the biomedical domain. | ✓ | 0.16 |
| The paper summarizes prominent challenges and constraints faced by large language models in healthcare. | ✓ | 0.22 |

## References


- https://doi.org/10.48550/arxiv.2402.05935
- https://doi.org/10.3390/informatics11030057
- https://doi.org/10.1109/access.2024.3365742