Responsible Conversational Artificial Intelligence for Sustainable Healthcare Futures

Abstract

The deployment of Large Language Models (LLMs) in healthcare introduces severe risks of miscalibration and spurious feature correlations, where non-expert medical advice precisely mimics clinical authority. To diagnose this representational collapse, we used model-agnostic explainers (LIME and ELI5) to audit baseline classification models. This audit revealed a critical performance asymmetry, achieving 83% precision for the authentic advice class but failing on the non-authentic class with only 24% precision. This proved that parametric models rely on lexical mimicry rather than clinical provenance. To structurally correct this reliance on spurious features, we developed a Human-in-the-Loop (HITL) MLOps pipeline called the Prompt Engineering and Auditing Platform (PEAP). The platform operationalises Reinforcement Learning from Human Feedback (RLHF) to generate the Clinically Audited Maternal Health Dialogue Corpus (CAMHeaD). Then we encoded this corpus with baseline Bidirectional Encoder Representations from Transformers (BERT). The encoding yielded a deceptively high 79.7% accuracy, yet exposed a high Expected Calibration Error (ECE) and poor out-of-distribution generalisation. To resolve this semantic rigidity, seven Retrieval-Knowledge Enhancement (RKE) architectures (including RAG, GRAFT, RAFT, GRAG, CAG, and KAG) were benchmarked against strict metrics, including calibration, bias, fairness, toxicity, accuracy, robustness, and inference time. On evaluation, Graph Retrieval-Augmented Fine-Tuning (GRAFT) was conclusively identified as the optimal architecture. This was because GRAFT achieved superior calibration and ethical scores by enforcing neuro-symbolic constraints over the Transformer's generative space to actively mitigate representational collapse. This work mandates PEAP as an architecture-agnostic auditing standard for deploying accountable, knowledge-grounded clinical AI, prioritising algorithmic alignment over uncalibrated statistical fluency.