Trustworthy LLM in Clinical Notes: A Dual Evaluation of Utility and Security

Background

With the widespread adoption of artificial intelligence in the healthcare domain—particularly in the analysis of clinic notes—large language models (LLMs) are increasingly becoming vital tools for clinical decision support. However, the trustworthiness of these systems still faces multiple risks stemming from both the models themselves and the underlying data. Current research primarily focuses on improving model performance, while systematic evaluations of whether LLMs can truly be trusted remain relatively limited. Therefore, the goal of my course project is to conduct a comprehensive study and practical evaluation of trustworthy LLMs. By assessing their utility and security across multiple dimensions in clinical text processing scenarios, I aspire to promote the safe and reliable integration of AI technology into real-world medical practice.

Main Tasks

This course project is centered around clinic notes as the primary data type and aims to build a comprehensive evaluation framework for trustworthy large language models (LLMs). The evaluation will focus on the following two core dimensions:

1. Utility Assessment of the Model

This involves evaluating the model's performance and risks in real-world applications, including but not limited to:

Performance metrics: such as accuracy, recall, F1 score, and other standard indicators;

Robustness and stability: assessing variations in model performance under different inputs, noise levels, and data subsets;

Fairness and reasonableness: identifying whether the model exhibits group bias or generates inappropriate or unjustifiable outputs.

2. Security Assessment of Data Usage

This involves evaluating the risks associated with how the AI system handles data, including:

Data ownership and compliance: determining whether protected or sensitive data is used in accordance with relevant regulations;

Data quality issues: testing how incorrect labels or incomplete records impact the model's behavior;

Privacy leakage risk: assessing the potential for recovering training data from the model through techniques such as Membership Inference Attack (MIA).

Timeline

Phase	Date Range	Tasks
Preparation	May 10 – May 11	- Define project objectives and deliverables -Draft the initial trustworthy AI evaluation framework-Organize and preprocess the Clinic Notes dataset
Framework Finalization & Experimental Design	May 12 – May 15	-Define specific evaluation metrics and methods for each dimension -Design and begin implementing evaluation experiments
Experiments & Analysis	May 16 – May 25	-Run model evaluation experiments -Collect and analyze results from multiple perspectives -Use indicators to assess privacy risks via Membership Inference Attacks -Conduct interpretability or fairness analysis
Demo and Materials Preparation	May 26 – May 28	-Develop and record a complete demo video
Report Writing & Revision	May 28 – May 30	-Write a comprehensive report covering the framework and findings
Presentation Preparation	June 3 – June 6	Organize and prepare the final presentation