上海科技大学
ShanghaiTech University

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

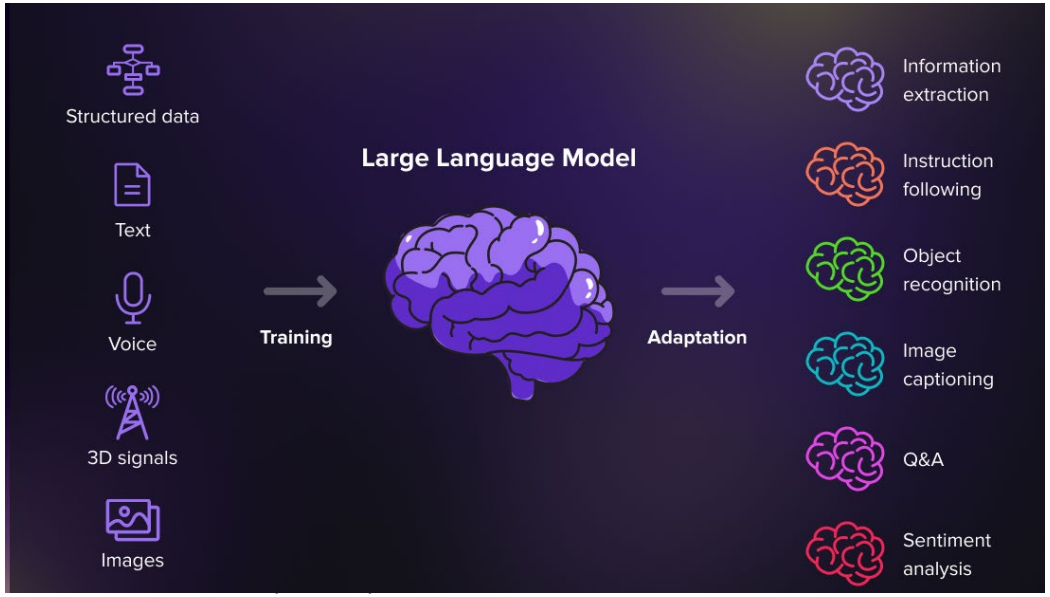
HISIR Lab 侯嘉玥

上海科技大学 ShanghaiTech University

Catalogs

Utility and Security in Large Language Model are Dialectical Issues

Utility



LLMs can be used in a wide range of applications

?

Security

LLMs are at risk of privacy breaches

There is a Delicate Balance Between the Two Characteristics

Research Objective

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).

Utility Assessment of the Model

This involves evaluating the model's **performance** in real-world medical applications

- ✓ **Performance metrics;**
- ✓ **Robustness and stability;**
- ✓ **Fairness and reasonableness;(options)**

Security Assessment of the Model

This involves evaluating the **risks** associated with how the AI system **handles data**

- ✓ **Privacy leakage risk=Attack;**
- ✓ **Data ownership and compliance;**
- ✓ **Data quality issues;(options)**

Q **What datasets are chosen?** → **How to design an experimental procedure?** → **What the final expected result looks like?**

DATASET

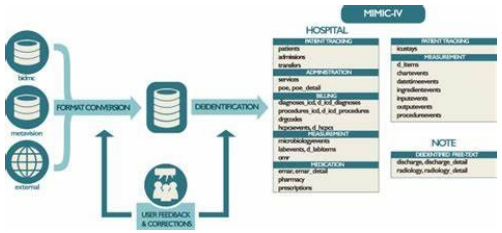
MIMIC-IV-Note

A collection of deidentified **free-text** clinical notes for patients included in the **MIMIC-IV clinical database**.

This structured table contains **discharge summaries** for hospitalizations.

Discharge summaries are long form narratives which describe the reason for

- **Patients' admission**
- **Their hospital course**
- **Relevant discharge instructions.**



MIMIC-IV-Note: Deidentified free-text clinical notes

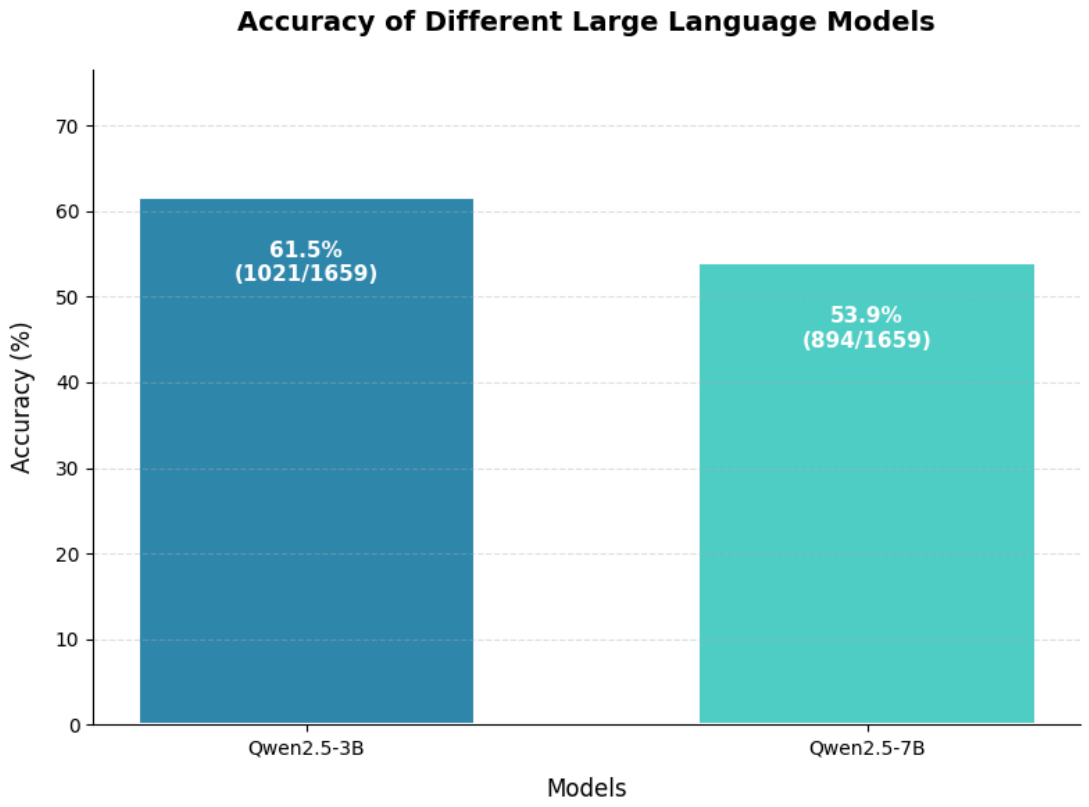
EHRNoteQA

EHRNoteQA dataset comprises **962** questions paired with the discharge summaries of 962 distinct patients.

Category	# of Discharge Summaries per patient	# of Questions	Total # of Discharge Summaries
Level1	1	264	264
	2	265	530
Level2	1	145	145
	2	144	288
	3	144	432
Total		962	1,659

EHRNoteQA: An LLM Benchmark for Real-World Clinical Practice Using Discharge Summaries.

Preliminary Results and Analysis



Analysis

Qwen2.5-3B: 1021 correct, 638 incorrect.

Qwen2.5-7B: 894 correct, 500 incorrect, 265 display errors.

Existing problems:

- Input tokens too long.
- The number of model parameters is small.
- Patient dataset and problem set are not reasonably combined.

➢ Final Results

Utility

- Adjusting data sets to improve accuracy

Robustness and stability	Fairness and reasonableness
• Assessing variations in model performance under different inputs, noise levels, and data subsets	• Identifying whether the model exhibits group bias or generates inappropriate or unjustifiable outputs

Security

- Checking that sensitive data is protected
- Testing the effect of incorrect labeling on accuracy
- Trying to design an MIA attack

Visualization

Results Showcase

- Evaluation Documents
- Core code
- Testing demo

上海科技大学
ShanghaiTech University

Thank you for listening

侯嘉玥

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