Explainability of LLMs & Their Ethical Implications

Final Presentation

1. Motivation

Large Language Models (LLMs) are increasingly deployed in healthcare environments where decisions can directly impact patient outcomes.

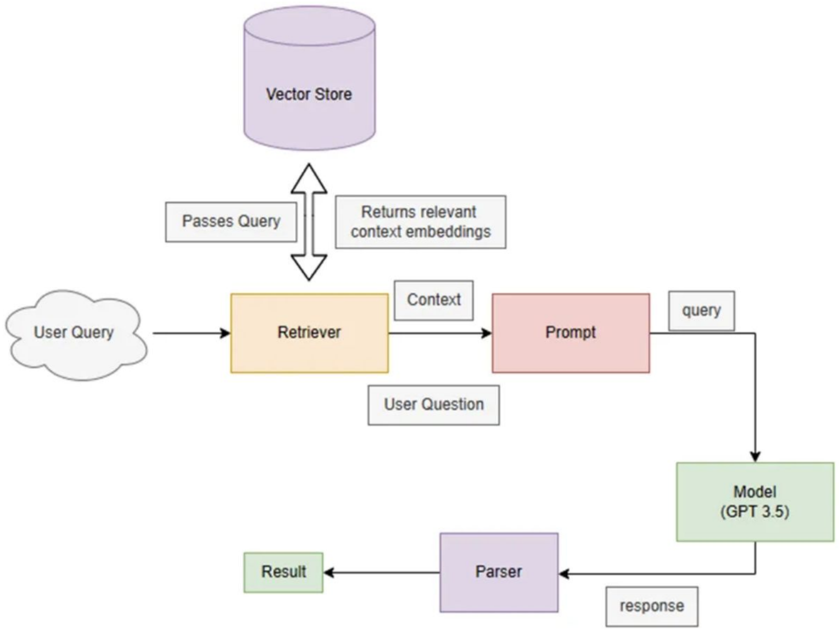
Why this matters:

- Traditional ML models have structured transparency tools, but LLMs are often seen as “black boxes.”
- In clinical decision-making, lack of explainability can lead to distrust, misdiagnosis, or unsafe prescriptions.
- High-stakes domains like healthcare require verifiable reasoning, especially for ethical and legal accountability.
- Explainability techniques like CoT and ReAct improve not just user trust but actual model performance.

I explore how prompting strategies can make LLMs safer and more interpretable for critical decisions.

2.Literature Review

2.1 SHAP LLM

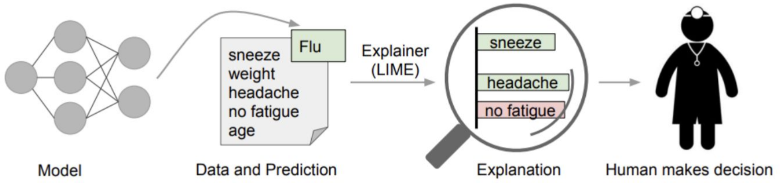


SHAP LLM workflow

2.2 SHAP LLM

Local Interpretable Model-agnostic Explanations (LIME)

LIME aims to identify an interpretable model over the interpretable representation that is locally faithful to the classifier.



3.Approach used

1. Zero-shot & Few-shot prompting

- In the zero-shot setting, the model receives no examples—just the question.
- Few-shot prompts include 1–3 manually crafted examples, which help orient the model.
- These are useful baselines, but often don’t provide any reasoning or explanation.

2. Chain of Thought (CoT) prompting

- This technique asks the model to "think step by step."
- For example, instead of just answering “yes” or “no” to a medical question, the model walks through the patient's condition, symptoms, and known drug interactions before concluding.
- I observed that this improves both performance and transparency in tasks like medication safety assessments.

3.Approach used

- 3. ReAct (Reason + Act)
- This approach combines reasoning with tool-use.

- The model reflects, then takes an action like querying a database or citing a document, then reflects again before answering.

- In my implementation, I simulated this behavior by integrating guideline-based logic through chained prompts and knowledge lookup via hardcoded responses.
-
- 4. RAG (Retrieval-Augmented Generation) (partially implemented)
- I experimented with integrating RAG-style templates where the model was supplied with retrieved context from a fake knowledge base.

- This showed how additional factual support can make the response more trustworthy, especially when citing guidelines or medical studies.
-

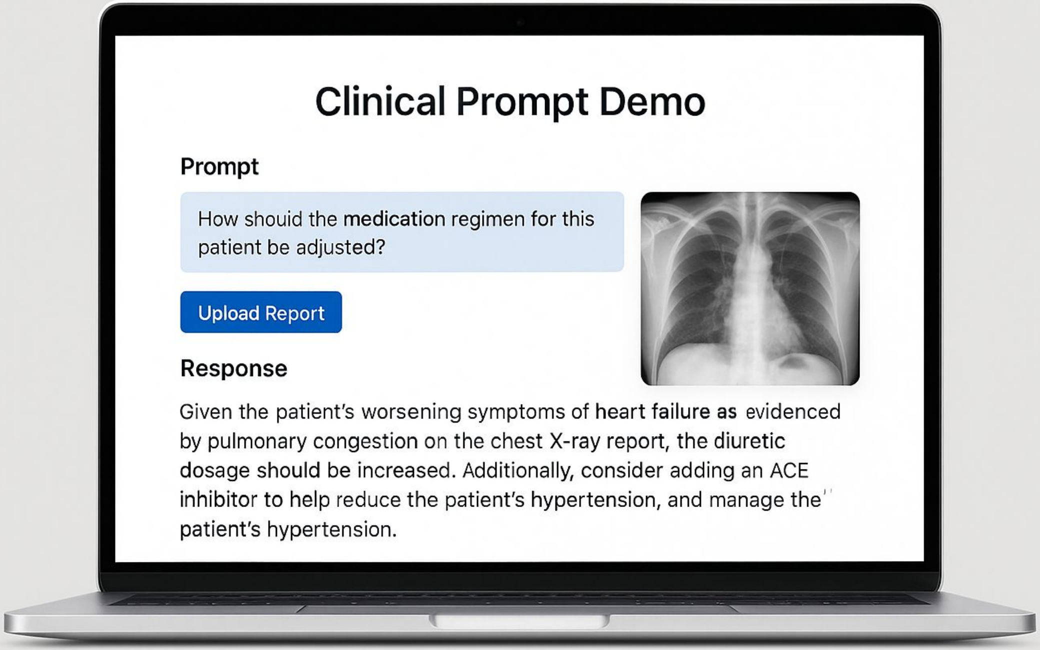
4.Demo

Example (Medical):

Q: Should this patient be prescribed Drug X?
Prompt: Let's think step by step.

Response:
Patient is 72, has liver issues.
Drug X stresses liver.
→ Do not prescribe.

Benefits:
- Transparent logic
- Ethical justifiability



References

1. Lundberg & Lee (2017). NeurIPS
2. Ribeiro et al. (2016). KDD
3. Wei et al. (2022). NeurIPS
4. Yao et al. (2022). arXiv
5. Lewis et al. (2020). arXiv
6. Zhang et al. (2023). arXiv
7. Wang et al. (2022). arXiv
8. https://eli5.readthedocs.io
9. https://www.analyticsvidhya.com/blog/2022/07/everything-you-need-to-know-about-lime/
10. https://medium.com/@davidacad10/using-llms-for-shap-explanation-f106da74dd75
11. https://arxiv.org/pdf/1602.04938.pdf