Emergency Room Cost Prediction Chatbot: Mistral vs Llama Model Comparison

Camila Grubb, Abhyuday Singh, Ben Ge, Edward Smith
Faculty Advisor: Gondy Leroy
The University of Arizona Eller College of Management – Tucson

Introduction

Many patients who require medical attention delay or avoid seeking care due to affordability concerns. This issue stems from the uncertainty surrounding medical costs, especially in systems where pricing transparency is limited or deemed too complex for vulnerable patients.

A potential solution involves leveraging technology to improve cost transparency and empower patients to make informed decisions about their healthcare. We aim to develop a RAG-LLM Emergency Room cost prediction chatbot that patients can use to help mitigate financial confusion. An individual can query the chatbot about their situation, describe any external cost factors such as location, insurance, pain severity, etc, and receive a cost estimation and/or range for an ER visit that is more personalized to their circumstances.

By training models on synthetic ER cost scenarios and evaluating responses to queries from users describing their specific situation, we can evaluate initial performance accuracy as well as strengths and weaknesses between models for greater perspective on limits beyond training data.

Methodology

We compared two models: Mistral-based model and a Llama-based model using the steps as follows.

- Connect Mistral and Llama APIs to a Python script for real time querying.
- Developed a RAG architecture as shown in Figure 1 where Python code references a precompiled file of synthetic ER procedures, symptoms, demographics, and costs (500 data points).
- The queries are parsed, extracting the diagnoses and contextualizing information.
- The RAG finds ER cases similar to the prompts and retrieves the cost estimates based on matching similarity of circumstances.
- LLM returns ER cost estimates to the user.

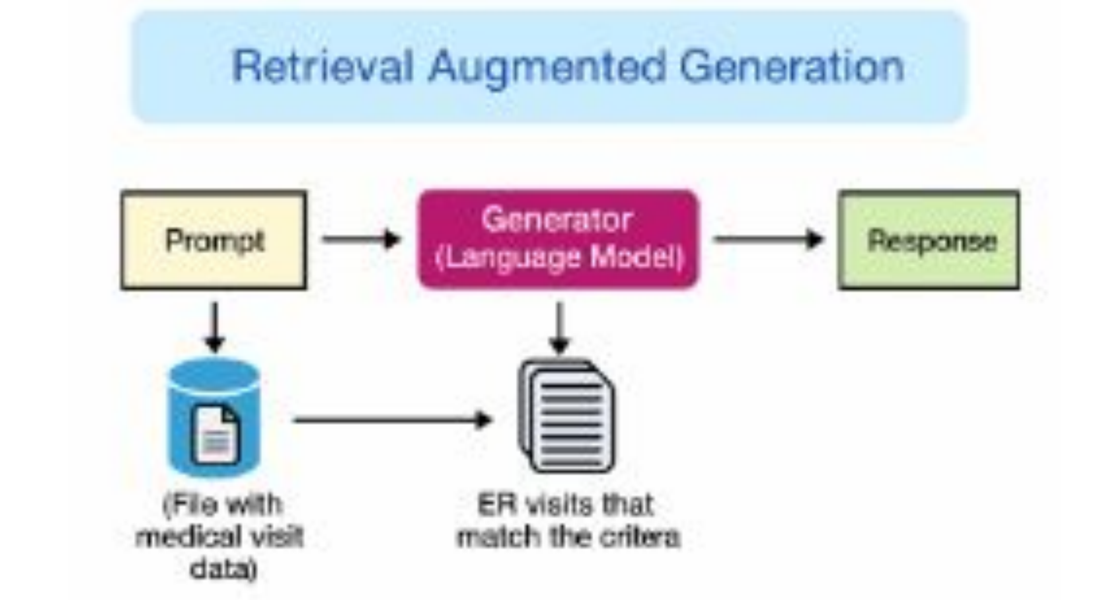

Figure 1. RAG-LLM Architecture

We compare two RAG LLM setups using Mistral and Llama on diagnostic and ER cost prediction tasks using 20 prompts with 2 types of prompts (10 vague, 10 specific). This study uses a small, 500 synthetic data point database.

- **'Vague' queries**:
  - -E.g. "I have a broken finger." excluding any detailed elaboration.
- **'Specific' queries**:
  - -E.g. "I contracted COVID. I have a high fever and I'm sweating a lot. I'd rate my pain a 1/10. I currently live in Pasadena, California, 91051. I currently have medicare."

'Vague' queries exclude any detailed elaboration. In this way, we attempted to test how the models respond to lack of context and how they determine a cost from limited information. Specific' queries contain a diagnosis, details on the user, and cost influences. Specific queries were developed to glean how much the models incorporate context into responses. Given the external factors regarding the severity of the injury and the location causing ER cost fluctuations, results of specific queries determine if the cost estimates are adjusted for the mentioned indicators.

Results

Prompt Type	Metric	Mistral	Llama
Vague	Time (t)	0.726s	2.477s
Vague	Char. Len (ct)	130 chars	90 chars
Vague	Accuracy (%)	80%	40%
Vague	Deviation (%)	0%	0%
Specific	Time (t)	0.77s	2.366s
Specific	Char. Len (ct)	121 chars	77 chars
Specific	Accuracy (%)	60%	0%
Specific	Deviation (%)	2.05%	23.58%

Table 1. Mistral/Llama Results (Vague and Specific Prompts)

- In vague scenarios, Mistral significantly outperformed Llama, retrieving data for 8/10 scenarios versus Llama's 4/10.
- Both models provided 100% accurate cost averages when data was successfully retrieved for vague queries.
- For specific prompts, Mistral retrieved data for 8/10 queries but accurately answered only 6 of them, with failures occurring when 'private' or 'uninsured' insurance types were present.
- Llama performed poorly with specific prompts, retrieving data for only 2/10 queries with a 23.58% average deviation from expected estimates.

Figure 2. Mock Representation of Chatbot Interface

Conclusion

Key Observations: ER Visits are a complicated space to research, with very limited publicly available data. Our study focused on understanding what key factors affect ER visit pricing and how to best use factors such as location, symptoms, and insurance options
Limitations: Even with our RAG architecture, we find it important to note that both the LLMs we used – Mistral and Llama – were trained on very different datasets. We would also like to highlight that our RAG setup still has some scope of improvement especially on retrieving best matched visit cases from the database as well as greater range of data points to train models.
Future Work: Evaluate model performance using the full NEDS HCUP dataset by running both Mistral and LLama models on real emergency department records to assess diagnostic accuracy and cost prediction reliability at scale. Integrate the best modell (Mistral or Llama) with a chatbot interface to be used by future users.

Acknowledgements

We would like to thank the Eller Partnerships Office and the Center for Management Innovations in Healthcare for making this project possible. We would also like to thank Dr Gondy Leroy for her support this semester.