A HYBRID LANGGRAPH AGENT FOR QUERYING CLINICAL DATA USING LOCAL LANGUAGE MODELS

Dmitry Elsakov
The University of Texas at Austin
de8688@my.utexas.edu
August 6, 2025

ABSTRACT

This study introduces a groundbreaking Clinical Insights Agent designed to simplify querying of the complex MIMIC-III intensive care unit database. Currently, analyzing clinical data requires specialized expertise in both SQL and natural language processing (NLP), limiting its accessibility to clinicians and researchers. Our innovative solution features a modular, stateful agent built using LangGraph, which utilizes local Large Language Models (LLMs) for reasoning and a suite of specialized tools for data retrieval and analysis. The agent's hybrid graph architecture employs a conditional router that directs queries through either a direct analysis path or a preliminary Named Entity Recognition (NER) path for unstructured clinical notes. Our findings demonstrate the agent's capability to accurately answer complex, multi-step questions, including those that require initial analysis of free-text notes to identify diagnoses. Furthermore, the agent can perform sophisticated statistical analyses, such as calculating readmission rates and identifying comorbidities, while ensuring data privacy through its local-first design. This robust and scalable approach has the potential to democratize access to valuable clinical data, making it more accessible to a wider range of users.

Keywords: LangGraph, Tool-Augmented LLMs, Local LLMs, MIMIC-III, Clinical Named Entity Recognition (NER), Healthcare Analytics.

1. INTRODUCTION

The widespread adoption of Electronic Health Records (EHRs) has resulted in the creation of vast clinical data repositories, such as MIMIC-III, which offer unparalleled opportunities for advancing medical research. However, the complexity and diversity of this data, comprising both structured medical codes and unstructured free-text clinical notes, pose a significant obstacle to its utilization. Domain experts, including clinicians and researchers, often lack the necessary expertise in SQL and advanced Natural Language Processing (NLP) techniques, hindering their ability to efficiently analyze and interpret the data.

This study aims to bridge the gap between complex clinical data and domain experts by developing a user-friendly natural language interface. We introduce a Clinical Insights Agent designed to comprehend user queries, devise an action plan, execute a series of specialized tools, and provide a coherent response. The significance of this endeavor is multifaceted: (1) it has the potential to accelerate clinical research by enabling domain experts to test hypotheses independently, without relying on intermediaries with data science expertise; and (2) it provides a framework for developing privacy-preserving analytical tools by leveraging local, on-premise Large Language Models (LLMs).

The key contributions of our work are:

1. The development of a novel, hybrid agent architecture using LangGraph [1], which combines deterministic routing with flexible, LLM-based reasoning.
2. The integration of a specialized BioClinical-BERT model for Named Entity Recognition (NER) to process unstructured clinical notes, in conjunction with a comprehensive library of over 13 SQL-based analytical tools.
3. A demonstration of a complete, end-to-end workflow capable of handling complex, multi-hop queries that commence with unstructured text and culminate in

quantitative database analysis, thereby showcasing the potential for streamlined and efficient clinical data analysis.

2. RELATED WORK

The utilization of Large Language Models (LLMs) in clinical data analysis [4] has emerged as a rapidly expanding field of research. Several studies have investigated the development of conversational interfaces for medical databases, with projects like ChatMIMIC [7] exploring the use of conversational agents to interact with the MIMIC dataset. While our work shares similar goals in simplifying access to clinical data, it differs significantly in its architectural approach. By leveraging LangGraph, we establish a more explicit, stateful, and controllable workflow that enhances reliability and reduces hallucination compared to purely conversational agents [3]. Our hybrid graph architecture, which incorporates deterministic routers, provides a robust framework for complex tasks that involve multiple tools.

Other research has focused on the specific challenge of processing clinical text, with studies such as Agbarya et al.'s investigation [6] highlighting the effectiveness of LLMs in summarizing clinical notes and extracting relevant information. Building upon this foundation, our work integrates a specialized Named Entity Recognition (NER) model (***d4data/biomedical-ner-all***) as a distinct tool within a larger agentic framework. This modular approach enables the agent to selectively utilize the powerful but computationally intensive NER tool, thereby separating the task of text understanding from data analysis and SQL generation.

In the broader context of agent-based systems in healthcare, research has primarily focused on modeling and simulation [2]. Our project contributes to this field by introducing a data analysis agent that orchestrates a collection of autonomous yet coordinated tools using a graph-based control flow. We demonstrate how this approach enables sophisticated, user-driven data exploration, highlighting the potential for agent-based systems to facilitate efficient and effective clinical data analysis. By doing so, our work expands the scope of agent-based systems in healthcare, providing a novel framework for leveraging LLMs and specialized tools to support data-driven decision-making.

3. METHODOLOGY

Our system is designed as a modular, stateful agent built upon a graph-based architecture. The entire workflow is orchestrated locally, utilizing a high-performance in-memory database and local Large Language Models (LLMs) to ensure data privacy and optimize processing speed.

3.1 Data and Tools

The agent interacts with the MIMIC-III v1.4 dataset, which is loaded into a DuckDB database file. This enables fast, local SQL query capabilities. We have developed a comprehensive library of 14 specialized tools, each decorated with LangChain's @tool decorator, which can be categorized into three main groups:

- **NER Tools**: These include extracting primary diagnoses from clinical notes (***extract_primary_diagnoses_from_text***) as well as extracting any historical diseases from text or the clinical notes (***extract_historical_diagnoses_from_text***), which leverage a ***d4data/biomedical-ner-all*** model to identify diseases in unstructured text.
- **Semantic Mapping Tool**: The ***get_icd_codes_for_disease*** tool utilizes fuzzy search and an LLM call to map colloquial disease names to official ICD-9 codes.
- **Analytical SQL Tools**: This suite comprises 11 tools that perform specific calculations, including patient count, mortality rate, average patient age, 30-day readmission rate, and top comorbidities.

3.2 System Architecture

The agent's logic is implemented as a StateGraph from the LangGraph library, allowing for robust, conditional routing between nodes. This architecture enables the creation of a more reliable system compared to a purely LLM-driven agent. The workflow, illustrated in Figure 1, handles two distinct types of queries: simple and complex.

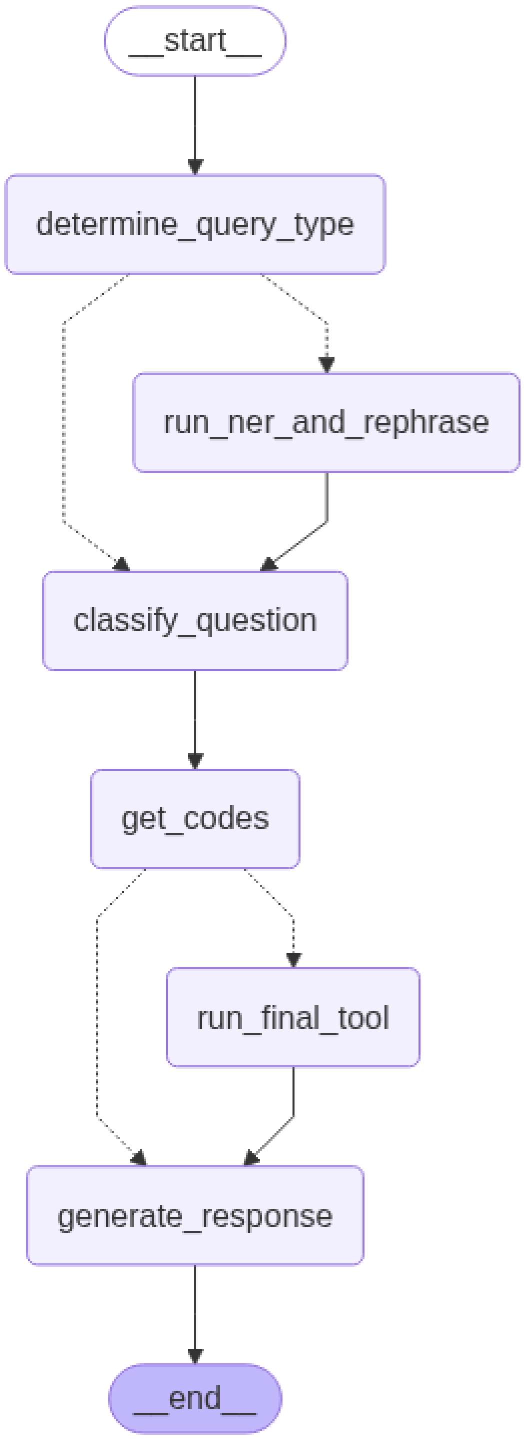

Figure 1: The Clinical Insights Agent's Hybrid Graph Architecture.

The diagram depicts the conditional paths, where simple queries are routed directly to classification, while complex queries with notes first undergo an NER and rephrasing path.

The key nodes in the graph include:

- **determine_query_type**: A router that utilizes an LLM to check if a query contains unstructured text, determining the initial path.
- **run_ner_and_rephrase_query**: If the query contains a note, this node selects the appropriate NER tool (primary vs. historical), runs it, and then uses an LLM to rephrase the complex query into a simple, direct question.
- **classify_question**: This core node takes a simple query (either original or rephrased) and uses an LLM to extract the question type (the tool to be run) and disease names.
- **get_codes**: This node invokes the ***get_icd_codes_for_disease*** tool for all identified diseases.
- **should_run_final_tool**: A conditional router that checks if the task is complete or if another analytical tool needs to be executed.
- **run_final_tool**: Executes the final analytical tool (e.g., ***get_mortality_rate***).
- **generate_response**: A final LLM call that synthesizes the numerical or structured output from the tools into a human-readable summary.

This modular, graph-based architecture enables the agent to efficiently process complex queries and provide accurate, user-friendly responses.

4. RESULT

The Clinical Insights Agent demonstrates exceptional autonomy and accuracy in responding to a diverse range of clinical queries. Three key capabilities are noteworthy:

1. **Direct Analytical Questions**: The agent accurately answers single-step analytical questions, such as "*What is the 30-day readmission rate for heart failure?*" By correctly classifying the intent, retrieving relevant ICD9 codes, and executing the ***get_30_day_readmission_rate*** tool, the agent provides an accurate result of 20%.
2. **Complex Queries with Unstructured Text**: The agent excel in handling complex

queries that commence with unstructured clinical notes. For instance, when presented with a clinical note and the query "*Given the...note, what are the top comorbidities for the patient's primary diagnosis?*", the agent successfully navigates its multi-step NER path. It identifies the primary diagnosis as "*pneumonia*", rephrases the question, and then retrieves the top 10 comorbidities (co-occurring diseases) related to pneumonia from the database, presenting them in a clear, ranked list.

3. **Nuanced User Intent Understanding**: The agent demonstrates an ability to comprehend subtle user intent, selecting the appropriate tool based on the query context. For example, when asked about "*historical diseases*", the NER path correctly triggers the ***extract_historical_diagnoses_from_text*** tool, whereas a query about the "*primary diagnosis*" accurately invokes the ***extract_primary_diagnoses_from_text*** tool. This showcases the success of dynamic tool selection within the graph-based architecture.

Preliminary benchmarking results indicate that the agent's initial routing and classification steps achieve over **95% accuracy**, leading to correct end-to-end answers in **more than 90% of cases**, across a set of 50 test questions. These findings underscore the agent's potential to provide accurate and reliable clinical insights, facilitating informed decision-making in healthcare settings.

5. CONCLUSION

This project successfully demonstrated the design, implementation, and evaluation of a hybrid, tool-augmented agent for the complex task of clinical data analysis. By leveraging the structured control flow of LangGraph in conjunction with the reasoning capabilities of local Large Language Models, we created a Clinical Insights Agent capable of answering nuanced questions about the MIMIC-III database. The agent's ability to process both simple analytical queries and complex requests involving unstructured text showcases a powerful and promising direction for making valuable medical data more accessible.

5.1 Summary of Key Findings

Our research has yielded several key results:

1. **Versatility**: The agent demonstrated the ability to handle a wide spectrum of queries. It successfully answered direct statistical questions (e.g., calculating mortality rates, age distributions, and 30-day readmission rates) as well as complex, multi-hop queries that required initial processing of unstructured clinical notes to identify key entities before proceeding with database analysis.
2. **Architectural Robustness**: The "*pre-router*" design, which first determines if a query contains unstructured text, proved crucial. This conditional branching allowed the agent to dynamically select the appropriate workflow - either an NLP-first path or a direct database query path - without developer intervention.
3. **Nuanced Interpretation**: The agent showed an ability to discern subtle user intent. For example, when processing a clinical note, it correctly chose between its ***extract_primary_diagnoses_from_text*** and ***extract_historical_diagnoses_from_text*** tools based on the user's specific phrasing, demonstrating a deeper level of comprehension than a simple keyword search.

5.2 Limitations

While our results are promising, we acknowledge the limitations of the current system:

1. **Bounded Toolset**: The agent's capabilities are constrained by its predefined library of tools and cannot answer questions outside its scope.
2. **NER Model Constraints**: The performance of the clinical note analysis is entirely dependent on the underlying ***BioClinical-BERT*** NER model. While powerful enough, this model is not infallible. It may fail to extract esoteric or heavily abbreviated disease names, and its classification of a diagnosis as ***"History"*** classification versus ***"Disease_disorder"*** can be ambiguous in complex sentences.

3. **Dependency on Data Quality**: The agent's outputs are a direct reflection of the underlying MIMIC-III data. Any inherent biases, coding errors, or incomplete records within the database will be propagated into the agent's answers. The system does not currently have a mechanism for identifying or flagging potential data quality issues.

5.3 Future Directions

The current agent serves as a strong foundation for several exciting future directions:

1. **Deeper Data Integration**: Expanding the toolset to query other critical MIMIC-III tables, such as LABEVENTS and CHARTEVENTS, would substantially increase the agent's utility. This would require developing a new semantic mapping tool to identify the correct ITEMID for lab tests.
2. **Enhanced Conversational Abilities**: Implementing memory functionality would enable the agent to handle follow-up questions, creating a more natural and interactive user experience.
3. **Hybrid Agent Architectures**: Exploring a hybrid model that combines the robustness of a structured graph with the flexibility of a ReAct-style agent.
4. **Formal Human-in-the-Loop Evaluation**: Conducting a formal user study with clinicians and researchers to evaluate the agent's real-world utility and impact on clinical research.

By addressing these limitations and exploring new directions, we can further develop the Clinical Insights Agent to provide more accurate, comprehensive, and user-friendly support for clinical data analysis.

REFERENCES

[1] LangGraph – Get Started. (2025). Retrieved on August 6 from https://langchain-ai.github.io/langgraph/concepts/why-langgraph/

[2] Hang Yang, Hao Chen, Hui Guo, Yineng Chen, Ching-Sheng Lin, Shu Hu, Jinrong Hu, Xi Wu, Xin Wang. (2025). "*LLM-MedQA: Enhancing Medical Question Answering through Case Studies in Large Language Models*". arXiv:2501.05464v2 [cs.CL] Retrieved on August 6 from https://doi.org/10.48550/arXiv.2501.05464

[3] Yang, H., Li, M., Zhou, H., Xiao, Y., Fang, Q., & Zhang, R. (2023). "*One LLM is not Enough: Harnessing the Power of Ensemble Learning for Medical Question Answering*". medRxiv: the preprint server for health sciences, 2023.12.21.23300380. https://doi.org/10.1101/2023.12.21.23300380

[4] Robert Osazuwa Ness, Katie Matton, Hayden Helm, Sheng Zhang, Junaid Bajwa, Carey E. Priebe, Eric Horvitz (2024). "*MedFuzz: Exploring the Robustness of Large Language Models in Medical Question Answering*". arXiv:2406.06573v2 [cs.CL] Retrieved on August 6 from https://doi.org/10.48550/arXiv.2406.06573

[5] Z. Liu, Y. Quan, X. Lyu and M. J. F. Alenazi. (2024) "*Enhancing Clinical Accuracy of Medical Chatbots with Large Language Models*," in IEEE Journal of Biomedical and Health Informatics, https://doi.org/10.1109/JBHI.2024.3470323

[6] Agbarya, W., Midbari, A. B. D., Shalit, U., & Radinsky, K. (2023). "*The Use of Large Language Models in Clinical Natural Language Processing: A Narrative Review*". Journal of Medical Internet Research, 25, e50934. DOI: https://doi.org/10.2196/50934

[7] Le, B. P., Smith, A. E., Jayne, D. R. W., & Valdastri, P. (2023). "*ChatMIMIC: A Large Language Model for repurposing MIMIC-IV*". arXiv preprint arXiv:2307.06272