# Evaluating Reasoning Capabilities of Large Language Models for Medical Coding and Hospital Readmission Risk Stratification with Zero Shot Prompting

Parvati Naliyatthaliyazchayil, Raajitha Muthyala, Judy Wawira Gichoya, Saptarshi Purkayastha

Submitted to: Journal of Medical Internet Research
on: March 21, 2025

# *Table of Contents*

# Evaluating Reasoning Capabilities of Large Language Models for Medical Coding and Hospital Readmission Risk Stratification with Zero Shot Prompting

Parvati Naliyatthaliyazchayil[1] BDS, MS; Raajitha Muthyala[1] BPharm; Judy Wawira Gichoya[2] MD, MS; Saptarshi Purkayastha[1] PhD

[1]Department of Biomedical Engineering and Informatics Luddy School of Informatics, Computing and Engineering Indiana University Indianapolis Indianapolis US
[2]Emory University School of Medicine Department of Radiology and Imaging Sciences Emory University Atlanta US

**Corresponding Author:**
Saptarshi Purkayastha PhD
Department of Biomedical Engineering and Informatics
Luddy School of Informatics, Computing and Engineering
Indiana University Indianapolis
535 W Michigan St.
Indianapolis
US



## *Abstract*

**Background:** The proliferation of large language models (LLMs) through accessible chatbot interfaces has created unprecedented opportunities in healthcare, with state-of-the-art models such as ChatGPT-4, LLaMA-3·1, Gemini-1·5, DeepSeek-R1and OpenAI-O3, offering artificial intelligence-driven clinical support. Some studies showcase the potential of LLMs in managing complex healthcare tasks, while others emphasize concerns regarding their accuracy, reliability, and compliance with the rigorous standards of clinical settings. This study was conducted to better understand their true potential and identify areas where they can be most effective in healthcare.

**Objective:** This study presents a comprehensive comparative analysis of leading reasoning and non-reasoning LLMs - ChatGPT-4, LLaMA-3·1, Gemini-1·5, DeepSeek-R1and OpenAI-O3 - evaluated across three critical healthcare tasks using the Medical Information Mart for Intensive Care IV (MIMIC-IV) dataset.

**Methods:** We assessed the model capabilities in: (1) generating primary diagnoses, (2) mapping diagnoses to ICD-9 codes, and (3) predicting hospital readmission risk stratification through zero-shot prompting protocols. The study utilized a cohort of 300 randomly selected subjects from MIMIC-IV, with standardized prompts systematically generated from discharge summary sections. Each prompt was engineered to incorporate both patient clinical information and specific task requirements in a unified input format. To enhance result interpretability, we implemented explicit rationale elicitation within the prompting structure, requiring models to articulate their reasoning process for diagnostic and prognostic predictions. Since this is a zero-shot prompt approach, the prompt is not tested, repeating the same multiple times.

**Results:** In our comparative analysis among non-reasoning models, LLaMA-3·1 demonstrated superior aggregate performance across all evaluation metrics, with 85% correctness in Primary Diagnosis prediction, 42·6% in ICD-9 code prediction, and 41·3% in hospital readmission risk prediction. Reasoning models DeepSeek-R1 and OpenAI-O3 showed similar performance, with O3 achieving slightly higher accuracy in primary diagnosis (90%) and ICD-9 prediction (45·3%), while R1 performed slightly better in readmission risk prediction (72·66%).

**Conclusions:** Our findings show that none of the evaluated models met clinical standards across all tasks, with medical coding showing the weakest performance. This aligns with few of the literature findings indicating that pretrained LLMs struggle with medical coding. This underscores the need for further refinement of these models to enhance their clinical applicability.

(JMIR Preprints 21/03/2025:74142)
DOI: https://doi.org/10.2196/preprints.74142

## Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone.

# Original Manuscript

# Evaluating Reasoning Capabilities of Large Language Models for Medical Coding and Hospital Readmission Risk Stratification with Zero Shot Prompting

Naliyatthaliyazchayil P [1], Muthyala R [1], Gichoya J [2], Purkayastha S [1]

[1] Department of Biomedical Engineering and Informatics, Indiana University Indianapolis

[2] Department of Radiology and Imaging, School of Medicine, Emory University, Atlanta

**ABSTRACT**

The proliferation of large language models (LLMs) through accessible chatbot interfaces has created unprecedented opportunities in healthcare, with state-of-the-art models such as ChatGPT-4, LLaMA-3·1, Gemini-1·5, DeepSeek-R1and OpenAI-O3, offering artificial intelligence-driven clinical support. Some studies showcase the potential of LLMs in managing complex healthcare tasks, while others emphasize concerns regarding their accuracy, reliability, and compliance with the rigorous standards of clinical settings. This study was conducted to better understand their true potential and identify areas where they can be most effective in healthcare.

This study presents a comprehensive comparative analysis of leading reasoning and non-reasoning LLMs - ChatGPT-4, LLaMA-3·1, Gemini-1·5, DeepSeek-R1and OpenAI-O3 - evaluated across three critical healthcare tasks using the Medical Information Mart for Intensive Care IV (MIMIC-IV) dataset. We assessed the models' capabilities in: (1) generating primary diagnoses, (2) mapping diagnoses to ICD-9 codes, and (3) predicting hospital readmission risk stratification through zero-shot prompting protocols. The study utilized a cohort of 300 randomly selected subjects from MIMIC-IV, with standardized prompts systematically generated from discharge summary sections. Each prompt was engineered to incorporate both patient clinical information and specific task requirements in a unified input format. To enhance result interpretability, we implemented explicit rationale elicitation within the prompting structure, requiring models to articulate their reasoning process for diagnostic and prognostic predictions. Since this is a zero-shot prompt approach, the prompt is not tested repeating the same multiple times.

In our comparative analysis among non-reasoning models, LLaMA-3·1 demonstrated superior aggregate performance across all evaluation metrics, with 85% correctness in Primary Diagnosis prediction, 42·6% in ICD-9 code prediction, and 41·3% in hospital readmission risk prediction. Reasoning models DeepSeek-R1 and OpenAI-O3 showed similar performance, with O3 achieving slightly higher accuracy in primary diagnosis (90%) and ICD-9 prediction (45·3%), while R1 performed slightly better in readmission risk prediction (72·66%). Our findings show that none of the evaluated models met clinical standards across all tasks, with medical coding showing the weakest performance. This aligns with few of the literature findings indicating that pretrained LLMs struggle with medical coding. This underscores the need for further refinement of these models to enhance their clinical applicability.

No Funding was used for this research.

**BACKGROUND**

The rapid evolution of reasoning and non-reasoning Large Language Models (LLMs) has sparked widespread interest in their potential applications across various domains[1], particularly healthcare[2]. Alongside established non-reasoning models like ChatGPT-4, LLaMA-3.1, and Gemini-1.5, new reasoning models, such as DeepSeek-R1 and OpenAI-O3, have also emerged, with reasoning capabilities embedded in their design, enabling more logical, step-by-step decision-making. However, the existing body of research presents a nuanced perspective on these pre-trained LLMs. While some studies highlight the promising ability of these LLMs to handle complex healthcare tasks[3], others raise critical concerns about their accuracy, reliability, and adherence to the high standards required in clinical settings[4]. This duality highlights the need for careful evaluation of their utility and reliability in real-world clinical environments. This leads us to a key question in this rapidly advancing field: which of these pre-configured LLMs is most suitable for addressing the unique challenges of healthcare tasks, and do newer reasoning models outperform non-reasoning models in this context? Their capacity to respond to open-ended questions without task-specific training through chatbot interface has generated both excitement and skepticism[5].

To address this question, our study systematically compares the performance of five models, prominent non-reasoning LLMs ChatGPT-4, LLaMA-3.1, and Gemini-1.5, as well as reasoning models DeepSeek-R1 and OpenAI-O3 across key healthcare tasks. In the rest of this study, we will refer to them as ChatGPT, Llama, Gemini, R1 and O3. Specifically, we evaluated their aggregated ability to generate primary diagnoses, code it to the International Classification of Diseases, Ninth Revision (ICD-9) codes, and predict risk stratification for hospital readmission using zero-shot prompting. To enhance transparency, we generated reasoning to explain the selection of primary

diagnoses and readmission risk predictions, specifically for non-reasoning models, as they lack embedded reasoning capabilities. In our study context, we would also like to define our above stated key concepts as follows: Primary diagnosis refers to the main condition that is chiefly responsible for a patient's current hospitalization. To ensure consistency across healthcare systems, diagnosis is coded to ICD-9 or 10 as a standard practice. ICD-9 and 10 are standardized coding systems used globally for categorizing diseases, conditions, and medical procedures[6,7]. Each diagnosis is assigned a unique numeric or alphanumeric code that is crucial in coding diagnoses for medical records. Further, we define hospital readmission as the likelihood of a patient being readmitted to hospital after discharge within the full timeframe covered by the dataset.

Controlled access to the Medical Information Mart for Intensive Care (MIMIC-IV) dataset was used in this study[8]. Given that these LLMs are predominantly trained on open-source internet data such as publicly available medical texts, research articles, health system websites, and accessible health information podcasts and videos[9] - this study seeks to evaluate their performance against limited access, real-world clinical data. MIMIC-IV data from patient details, admission details, diagnoses, and discharge are used. Discharge summaries, also referred to as clinical notes, were used to extract sections such as chief complaints, past medical history, surgical history, labs, imaging, and primary diagnosis. All these sections, except primary diagnosis, were used to create prompts for LLM evaluation. These prompts were then employed in zero-shot testing, which has garnered increasing interest in the literature to assess LLMs' generalization capabilities without task-specific fine-tuning[10]. The extracted primary diagnoses served as the ground truth for semantically evaluating the diagnosis predictions made by the LLMs. ICD-9 Codes from diagnosis were used as ground truth to assess the accuracy of ICD-9 code predictions, while counts of patient readmission using hadm_id were used as the ground truth for evaluating the models' ability to predict hospital readmission risk.

This study contributes to existing literature through the following: first, by using the controlled access MIMIC-IV dataset, we provide a clinically relevant evaluation based on authentic, complex, real-world healthcare data. Second, our focus on key medical tasks such as diagnosis generation, ICD-9 coding, and hospital readmission risk prediction offers a granular assessment of LLMs in high-value areas. Third, we emphasize transparency by generating models' reasoning, a crucial requirement for explainable AI in healthcare. Fourth, our use of zero-shot prompting reflects the practical constraints of real-world deployment, where task-specific fine-tuning is often expensive and time-consuming. Lastly, the comparative analysis of state-of-the-art models, including the non-reasoning ChatGPT, LLaMA, and Gemini, alongside the reasoning models R1 and O3, offers actionable insights into which LLM may be the best fit for "one-stop-shop" healthcare applications.

These models are not designed to replace human expertise but to assist healthcare professionals by reducing time spent on labor-intensive tasks, minimizing errors, and uncovering subtle diagnostic insights that might otherwise be overlooked. LLMs can potentially improve patient outcomes by augmenting diagnostic processes and enhancing decision-making, provided their safety and reliability are rigorously validated[11].

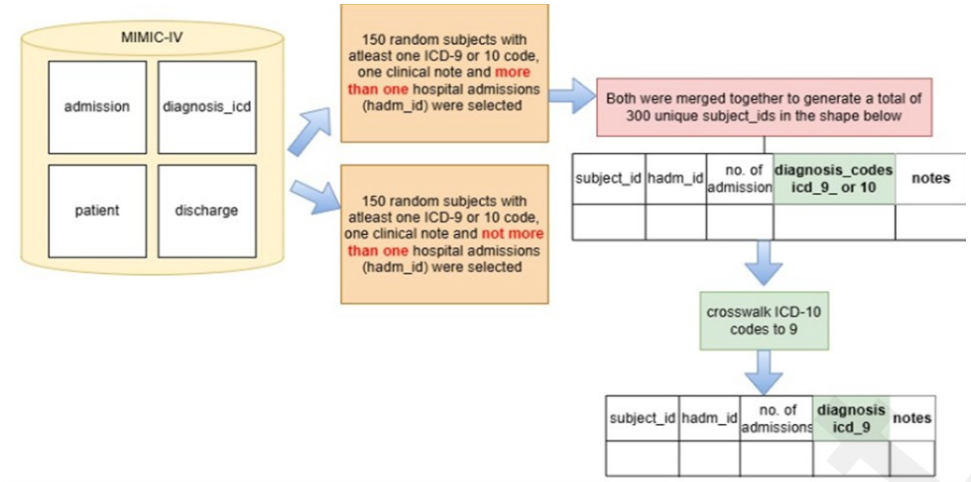
**METHODS:**

This study employed a multi-step approach to evaluate the performance of LLMs in addressing key healthcare tasks, including the ability to predict primary diagnoses, assign ICD-9 codes, and stratify hospital readmission risks, along with explanations for diagnosis and risk classification. The methodology is organized into four key phases, summarized as follows:

1. **Sample collection**

   Clinical data was obtained from the controlled access MIMIC-IV dataset. It is a de-identified dataset containing detailed health information from patients admitted to the emergency department or intensive care units at Beth Israel Deaconess Medical Center in Boston, Massachusetts. It includes structured clinical data such as demographics, diagnoses, procedures, laboratory test results, etc., as well as unstructured data like labeled clinical notes[12]. Key files related to admissions, patient, diagnoses_icd, and discharge were utilized in conjunction to create our sample. A sample of 300 unique Patient IDs was selected, ensuring each patient had valid diagnosis codes and at least one available discharge summary. Each patient's ICD-9 and 10 Codes were extracted as a comma-separated list (CSV), along with the first discharge note for each patient. Since there were more ICD-9 Codes in our sample than ICD-10, we cross walked all the ICD-10 to ICD-9 to keep any kind of data loss during this conversion to minimal. To evaluate hospital readmission risk, each patient's total number of admissions is calculated using *hadm_id*, and admission dates. In the 300 patient sample, 150 patients had more than one readmission, while the remaining 150 did not experience any readmission.

   **Figure 1: Sample collection of 300 unique subject_ids**

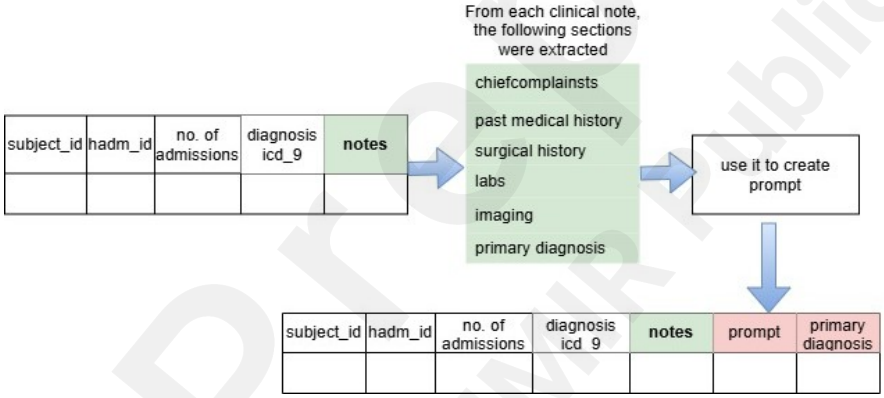



*This figure shows the sample of 300 subject_ids was created from the MIMIC-IV dataset and then any ICD-10 codes in the sample were crosswalked to respective ICD-9 using UMLS crosswalk. The tables show the structure of the output for ease of understanding.*

## 2. Prompt template and creation

Prominent sections from labeled discharge summaries in the MIMIC-IV Notes database were utilized to draft prompts. MIMIC-IV Notes provides preprocessed and labeled clinical notes, offering a rich source of information about a patient's clinical journey[13]. For each patient, the following sections were extracted: chief complaints, past medical history, surgical history, labs, and imaging. Since these sections were labeled, we used regex to extract these sections as shown in Figure 2.

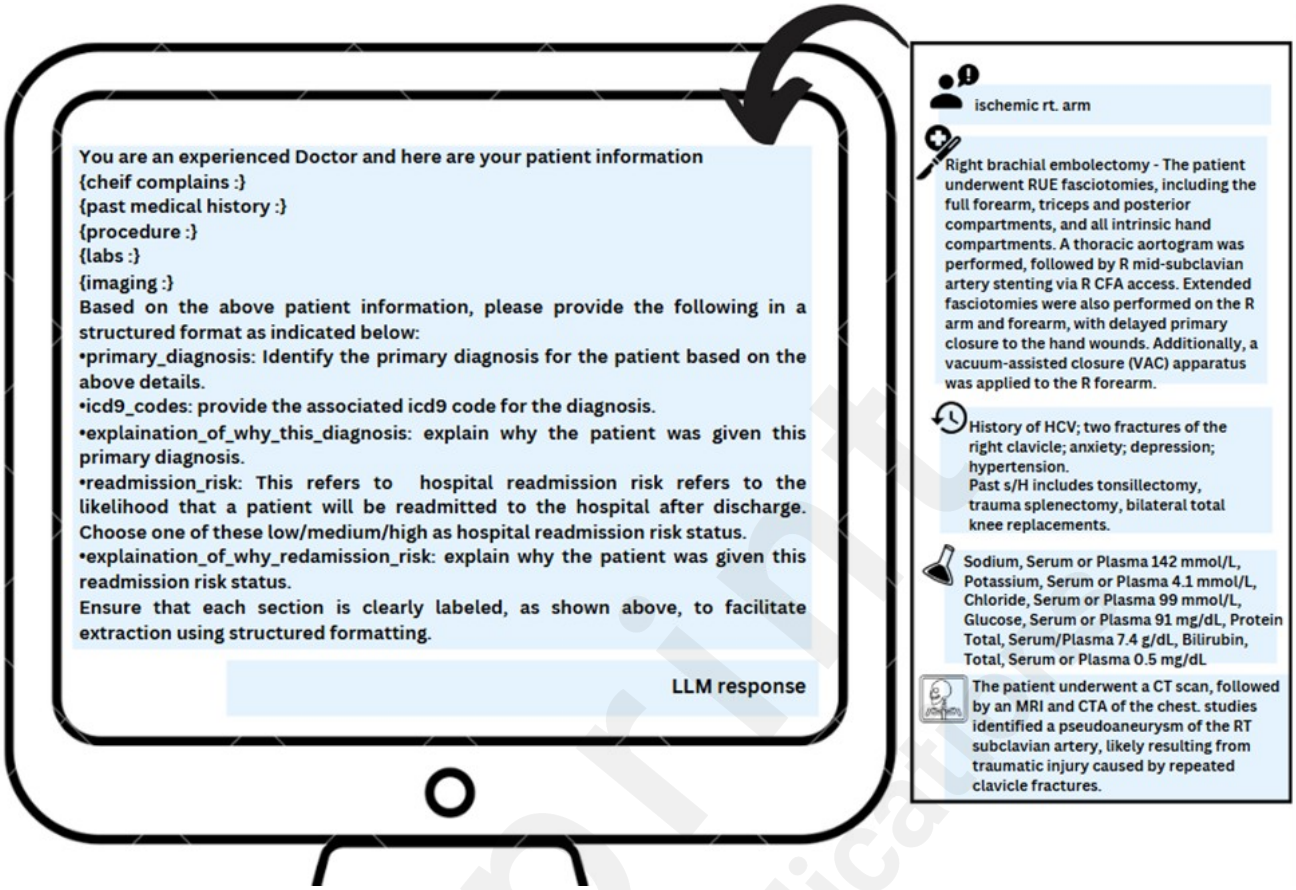
**Figure 2: Creation of prompt using sections from discharge summaries/clinical notes**




*This figure shows how the output from Figure 1 is further utilized. The key sections from MIMIC-IV clinical notes were utilized in prompt creation and extracting primary diagnosis of sample.*

These extracted sections, except for primary diagnosis, were then programmatically integrated into a structured "prompt template" for large language model (LLM) evaluation, as shown in Figure 3. Though the primary diagnosis is extracted, this was not included in the prompt design as mentioned; instead, it served as ground truth data for later evaluating the accuracy of the LLM-generated responses.

**Figure 3: Prompt Template**




*This figure shows the prompt template that is systematically populated for each subject_id from their notes that were extracted. On the right you see an example representation of MIMIC-IV note, which is then populated into its respective sections within the prompt. The note here is an example and not an actual record from MIMIC-IV.*

**3. Collecting and processing the response**

All prompts were systematically generated and input into each AI chatbot through their respective interfaces. This study focuses on testing the zero-shot capabilities of the freely available out-of-the-box chatbot interface rather than the Application Programming Interface(API), as API implementation may incur additional costs based on token usage and require specific setup skills, emphasizing low setup needs while maximizing output. Each prompt was given to each chatbot only once, without repetition, and the memory of the chatbox was turned off to prevent them from learning from each prompt. Once the prompts were given, the responses generated for each subject were collected in a CSV file, where previously recorded subject metadata was stored.

Since the chatbot responses followed a structured format, the relevant information was extracted into individual columns for each subject. This process yielded the following responses from the LLMs:

- The primary diagnosis generated by the LLM,
- A list of ICD-9 Codes associated with the primary diagnosis,
- The predicted readmission risk status,
- Explanations for the selected primary diagnosis, and
- Justifications for the predicted readmission risk status.

These responses were consolidated into the CSV file, creating a dataset ready for detailed evaluation against the ground truth data.

## RESULTS

This study provides a comparative evaluation of leading LLMs, ChatGPT, Llama, Gemini, R1 and O3 in terms of their ability to perform healthcare-specific tasks. The prompt was created from the MIMIC-IV clinical notes' key sections. The responses produced by LLM were extracted into their individual structured columns for analysis and compared against the ground truth from MIMIC-IV data. The results highlight notable and interesting variations in performance across tasks.

1. **Comparing the prediction of Primary diagnosis**

The Primary diagnosis from each LLM's response was compared against the Primary diagnosis we extracted from MIMIC-IV Clinical Notes. We utilized SciBERT, a pre-trained model specifically designed for scientific and medical contexts[14]. This makes it particularly adept at processing and understanding domain-specific language, which is essential

for comparing medical terminologies.

The allenai/scibert_scivocab_uncased variant of SciBERT[15], implemented through the SentenceTransformer framework, was employed to generate embeddings for both the ground truth primary diagnosis (from MIMIC-IV clinical notes) and the LLM-predicted diagnosis. The process involved:

1. Embedding Generation: Both the reference diagnosis and the LLM-generated text were converted into high-dimensional embeddings using SciBERT.
2. Cosine Similarity Computation: Cosine similarity was calculated between the two embeddings to quantify their semantic similarity. A threshold of 0·7 was established to classify predictions:
    - Scores ≥0·7 were considered semantically aligned with the ground truth.
    - Scores <0·7 were categorized as incorrect or divergent predictions.

Among non-reasoning models, Llama and ChatGPT exhibited comparable performance, with semantic match rates of 85% (255 of 300) and 84·9% (254 of 300), respectively. This marginal difference suggests that both models are similarly capable of aligning with the ground truth diagnoses, outperforming Gemini, which achieved a match rate of 79% (237 of 300). Between the reasoning models, O3 exhibited higher performance with 90% (270 of 300) match rate whereas R1 showed 85%(255 of 300) match rate. Reasoning models performed better than the non-reasoning models.

### 2. Comparing the Prediction of ICD-9 Code

To evaluate the accuracy of ICD-9 Code predictions by the LLMs, we performed a systematic comparison against the ground truth codes from the MIMIC-IV dataset, which includes both ICD-9 Codes and ICD-10 Codes. We crosswalked the ICD-10 Codes to ICD-9 using the UMLS ICD-9 to ICD-10 crosswalk[7,8]. The decision to crosswalk was driven by the relatively small number of ICD-10 codes present in our sample, ensuring that the majority of original diagnostic codes could be consistently represented for comparison.

Both the ground truth ICD-9 Codes and LLM-generated codes were converted into comma-separated lists to ensure uniformity. We then conducted a row-wise comparison to identify matches between the predicted and ground truth ICD-9 Codes.

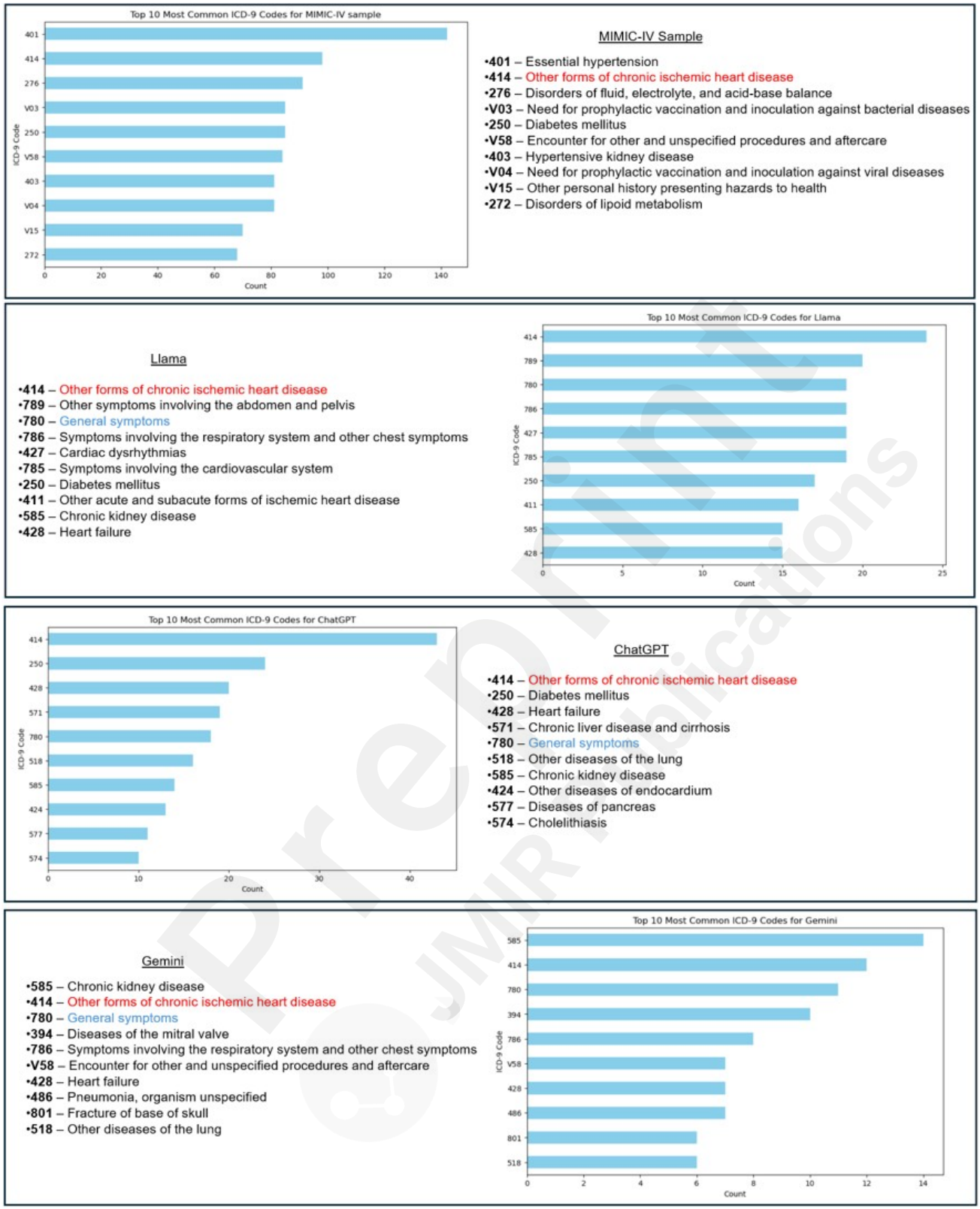
In evaluating the ability of the non-reasoning LLMs to predict ICD-9 Codes for primary diagnoses, Llama correctly predicted ICD-9 Codes for 128 out of 300 subjects. ChatGPT followed, correctly predicting ICD-9 Codes for 122 out of 300 subjects. Gemini lagged behind, predicting ICD-9 Codes for 44 out of 300 subjects. These results indicate that LLaMA and ChatGPT are comparably effective,but their performance still falls short of the accuracy required for reliable medical coding applications and this finding aligns with few studies in literature[4]. Between the reasoning models, O3 correctly predicted ICD-9 Codes for 136 out of 300 subjects whereas R1 correctly predicted ICD-9 Codes for 121 out of 300 subjects. The medical coding skills for the reasoning models also lagged far behind the standards expected for clinical practice. Further refinement and training may be needed to enhance the models' effectiveness in this domain.

### 3. Top ten ICD-9 Codes in MIMIC-IV Sample and three non-reasoning LLMs

We also found the top ten ICD-9 Codes from the MIMIC-IV sample and the three non-reasoning LLM-generated ICD-9 Codes, as shown in Figure 4. Each subject_id can have more than one ICD-9 Code. For this analysis, we implemented an ICD-9 hierarchical rollup by aggregating detailed diagnosis codes to their respective three-digit parent categories. For example, specific codes like 414.0 (Coronary atherosclerosis) and 414.00 (Coronary atherosclerosis of unspecified type of vessel) were rolled up to their broader parent category, 414 (Other forms of chronic ischemic heart disease). Top ten ICD-9 codes were calculated after this rollup.

We found that ICD-9 codes associated with the parent category 414 (Other forms of chronic ischemic heart disease) were present across all three LLMs and the MIMIC-IV sample as one of the top two. In contrast, another parent category, 780 (General Symptoms), appeared in all three LLMs but was absent in the MIMIC-V sample. This suggests that the LLMs were coding many symptoms differently from clinical practice, highlighting an area for potential improvement. Additionally, the parent category for Diabetes Mellitus was observed in the MIMIC-IV sample, Llama, and ChatGPT, but not in Gemini, which aligns with our findings of ICD-9 code predictions, where Gemini underperformed.

**Figure 4: Non-reasoning LLMs and MIMIC-IV sample Top 10 ICD-9 Codes**




*This figure shows the top 10 ICD-9 codes from MIMIC-IV sample and the three non-reasoning LLMs. Such graphs can help us show patterns. Here we see a pattern of Ischemic Heart diseases showing in sample and LLM whereas the category of General symptoms were only seen in all three LLMs and not MIMIC-IV, showing that it might be an area for scope of improvement. Diabetes Mellitus is seen in MIMIC-IV sample, Llama and ChatGPT and not in Gemini, which aligns with our findings of ICD-9 code predictions where Gemini underperformed*

### 4. Comparing the Prediction of Hospital Readmission Risk Status

To evaluate the predictive performance of the LLMs for hospital readmission risk status, the numeric ground truth values were derived as the total number of readmissions for each subject. Since the LLM-generated responses classified risk as Low, Medium, or High, the numeric ground truth was converted into comparable categorical labels using quantile-based

thresholds. This approach ensured consistency between the ground truth and the model responses. Numeric readmission counts were categorized as follows:

- Low Risk: Values in the bottom 25% (≤ first quantile).
- Medium Risk: Values falling between the first and third quantiles (middle 50%).
- High Risk: Values in the top 25% (> third quantile).

After aligning the LLM responses with the ground truth categories and evaluation, among non-reasoning models, Llama had 41·3% (124 out of 300) correct predictions, followed by Gemini with 40·7% (122 out of 300) and ChatGPT with 33% (99 out of 300). While Llama and Gemini demonstrated moderate alignment with the ground truth categories, the overall results suggest significant room for improvement. Between the reasoning models, R1 performed slightly better with 72·66% correct risk predictions (218 out of 300) than O3 with 70·66% correct risk predictions (212 out of 300). This shows reasoning models perform better than non-reasoning models for readmission risk prediction.

### 5. F1 Score for ICD-9 Code Prediction and Readmission Risk Status

We calculated the Multiclass Multilabel F1 score for ICD-9 Code prediction and the Macro averaged F1 score for readmission risk stratification for all three LLMs. F1 score for ICD-9 Code prediction helps to evaluate how well the model is identifying correct codes while avoiding incorrect ones, and for readmission risk prediction will show how effectively the LLM balances identifying patients at risk (e.g., "high risk") versus avoiding unnecessary false alarms. Though the F1 scores were low for both reasoning and non-reasoning models as seen in Table 1, due to the increased number of false negatives, out of the three non-reasoning LLMs, Llama had a higher F1 score for ICD-9 Code prediction and readmission risk stratification. Between the reasoning models, O3 has a collective higher F1 score for ICD-9 Code prediction and readmission risk stratification combined. It is an interesting finding that shows both reasoning and non-reasoning models did not exhibit exceptional F1 scores showing both generated false negatives and false positives.

**Table 1: F1 scores for Llama, ChatGPT, Gemini**

| Chatbot | F1-Score ICD-9 Code prediction | F1-Score readmission prediction |
|---|---|---|
| Llama | 0·083 | 0·412 |
| ChatGPT | 0·081 | 0·322 |
| Gemini | 0·024 | 0·408 |
| DeepSeek-R1 | 0·091 | 0·422 |
| OpenAI-O3 | 0·122 | 0·414 |

*This table shows the Multiclass multilabel F1 score for ICD-9 prediction and F1 score for hospital readmission risk prediction. F1 scores take into consideration true positives, true negatives, false positives and false negatives. The F1 scores for ICD-9 code prediction is low for all LLMs due to the increased false negatives than the true positives.*

## DISCUSSION

The existing literature presents mixed results on the ability of LLMs to perform healthcare tasks, primarily medical coding and predicting diagnosis. A significant study by Soroush A et al. highlights the poor performance of LLMs in medical coding[4], while another research by Kwan K et al. suggests that LLMs can perform well with some additional augmentation[3]. Petro J et al. demonstrates that while LLMs may make mistakes, they are also capable of identifying errors, emphasizing that the technology is still a work in progress[9]. Another interesting study by Zhu Y et al. shows a framework where the prompts incorporate longitudinal health records and improve predictive accuracy[10]. Furthermore, a systematic review by Zhou S et al. indicates that prompt engineering and fine-tuning with high-quality data can lead to the development of robust diagnostic systems using LLMs[21]. With these diverse findings in mind, we sought to evaluate the performance of three prominent, out-of-the-box LLMs for aggregated high-value healthcare tasks such as predicting primary diagnoses, generating ICD-9 codes, and stratifying hospital readmission risk using a dataset that these LLMs are not already trained on.

In this study, we evaluated a sample of 300 subjects from the MIMIC-IV dataset[17], leveraging sections of clinical notes to generate prompts systematically. LLM API was not used in this study; instead, we focused on evaluating the zero-shot capabilities of the pre-configured chatbot interface for diagnostic prompting. Each prompt was presented only once, without repetition to the LLM, and chat memory was disabled to prevent the chatbot from learning from previous interactions. A sample size of 300 subjects is larger than some of the existing studies[18] and provides sufficient dataset to evaluate the AI chatbot version across all five LLMs effectively. While we want to assess the aggregate capability for high-value tasks, evaluating the capability of these five LLMs (three non-reasoning and two reasoning) out-of-the-box via zero-shot prompting specifically for hospital readmission risk prediction, using all relevant information from the patient discharge summary, is a unique approach that we believe is a valuable discussion to literature. Employing zero-shot prompts and prompt-based methods presents distinct advantages of being user-friendly, saving time and

computational resources, and less setup[19-21]. This approach evaluates the baseline performance of these models, highlighting strengths, limitations, and gaps for targeted improvements through fine-tuning or prompt engineering.

In evaluating the performance of non-reasoning LLMs for predicting primary diagnoses, Llama demonstrated improved accuracy, achieving 85% correctness in a zero-shot prompting scenario. While not outstanding, this level of performance demonstrates the model's capability to support clinical decision-making without task-specific fine-tuning. Between the reasoning models, O3 demonstrated higher performance with 90% correctness. Our approach aims to enhance efficiency and decision-making through AI-human collaboration. Additionally, we generated explanations for each prediction in both reasoning and non-reasoning models to ensure transparency in the model's reasoning. In future work, we plan to conduct further analysis of the reasoning provided by these LLMs. By leveraging o3's diagnostic predictive capabilities alongside the expertise of healthcare practitioners, clinical workflows can be streamlined.

The performance of reasoning and non-reasoning LLMs in predicting ICD-9 Codes was suboptimal in this study, with Llama (42·6%) and ChatGPT (40·6%) performing relatively better than Gemini(14·6%). O3(45·3%) and R1(40·3%) also lagged. The F1 scores of these LLMs were also relatively low (0·083, 0·081, 0·024, 0·122, 0·091 for Llama, ChaGPT, Gemini, O3 and R1 respectively), highlighting areas for improvement, particularly in reducing false positives. One of the crucial concerns of such models is the potential risk of these models having hallucinations. One of the types of hallucinations is the "faithfulness problem," where LLM generates non-factual or unfaithful information [22,23]. Addressing this is essential for improving the reliability and trustworthiness of AI-driven systems in clinical decision-making. Our finding for predicting ICD-9 codes aligns with literature[4] showing these LLMs fall short, and that models specifically trained for medical coding tasks, such as those fine-tuned on clinical text datasets[24,25] can assist coders by suggesting the most relevant ICD codes based on the content of clinical notes, reduce human errors, improve patient-care follow-ups and ensure compliance with regulatory requirements which aligns with our findings.

In evaluating readmission risk prediction, Llama (41·3%) outperformed both Gemini (40·7%) and ChatGPT (33%) among non-reasoning models. Reasoning models performed better than non-reasoning with R1 achieving 72·6% and O3 with 70·6% correct risk predictions. F1 score of 0·412 for Llama demonstrates there is a need for the LLM to be trained further. F1 score for O3 was also low of 0·122 but better than non-reasoning LLMs. None of the models performed at an ideal level as one might expect. A limitation of this analysis can be the potential variability in the readmission information within the dataset. While the models' performance was suboptimal, we believe including generated explanations for the risk predictions is a valuable addition. These explanations can enhance transparency, offering healthcare providers insights into how the models arrived at their conclusions. This will be one of the key areas of our future work where we can tune a model to better predict the hospital readmission risk and conduct further analysis on the LLM's reasoning capabilities.

Another observation was that reasoning models produced more verbose "explanations" for Primary diagnosis and readmission risk than non-reasoning models. Non-reasoning models generated an average of 70 words for Primary diagnosis explanations and 54 words for readmission risk explanations. In contrast, reasoning models like R1 averaged 418 words for Primary diagnosis explanations and 612 words for readmission risk explanations. O3 generated an average of 713 words for Primary diagnosis explanations and 1,112 words for readmission risk explanations.

Gemini made an interesting observation during the response collection phase. When prompted with scenarios involving psychiatric information, the Gemini model consistently replied with a safety-focused message instead of performing the specific task in the prompt: "Call or text 988 for support." This behavior highlights the model's prioritization of user safety in sensitive contexts, potentially reflecting ethical programming to handle mental health-related prompts. This was not noted in the other two LLMs.

## CONCLUSION

In conclusion, the LLMs in this study demonstrated varied performance across healthcare tasks using zero-shot diagnostic prompting. No model achieved satisfactory results out of the box, aligning with the literature suggesting the need for more specific semantic training with medical datasets. Despite suboptimal performance, this study highlights the potential of ready-to-use reasoning and non-reasoning LLMs to assist healthcare providers with minimal setup and no prior expertise in fine-tuning. LLaMA outperformed the other non-reasoning models in all predictions, while O3 outperformed both reasoning models and LLaMA, proving itself as a powerful out-of-the-box tool for healthcare professionals. These findings suggest that future chatbot LLMs in healthcare need further training to improve interaction, reduce hallucinations, and enhance robustness for real-world deployment. Our future work will focus on refining models for hospital readmission risk prediction and analyzing their reasoning capabilities to better support clinical workflows.

**References:**

1. Minaee S, Mikolov T, Nikzad N, et al. Large language models: a survey. arXiv [Preprint]. 2024;2402.06196. doi:10.48550/arXiv.2402.06196.
2. Yang R, Tan TF, Lu W, Thirunavukarasu AJ, Ting DSW, Liu N. Large language models in health care: development, applications, and challenges. Health Care Sci. 2023 Jul 24;2(4):255-63. doi:10.1002/hcs2.61. PMID:38939520; PMCID:PMC11080827.
3. Kwan K. Large language models are good medical coders, if provided with tools. arXiv [Preprint]. 2024;2407.12849. doi:10.48550/arXiv.2407.12849.

4. Soroush A, Glicksberg BS, Zimlichman E, Barash Y, et al. Large language models are poor medical coders-benchmarking of medical code querying. NEJM AI. 2024;1(5):AIdbp2300040. doi:10.1056/AIdbp2300040.
5. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. Nat Med. 2023;29(12):1930-40. doi:10.1038/s41591-023-02448-8.
6. Cooke DT, Broghammer JA. From residency to retirement: building a successful career in thoracic surgery. Thorac Surg Clin. 2011;21(3):413-9.
7. Government of British Columbia. Diagnostic code descriptions: ICD-9. Available from: https://www2.gov.bc.ca/gov/content/health/practitioner-professional-resources/msp/physicians/diagnostic-code-descriptions-icd-9.
8. Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R. MIMIC-IV (version 2.0). PhysioNet. 2022. doi:10.13026/7vcr-e114.
9. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med. 2023;388(13):1233-9. doi:10.1056/NEJMsr2214184.
10. Zhu Y, Wang Z, Gao J, et al. Prompting large language models for zero-shot clinical prediction with structured longitudinal electronic health record data. arXiv [Preprint]. 2024;2402.01713. doi:10.48550/arXiv.2402.01713.
11. Xie W, Xiao Q, Zheng Y, et al. LLMs for doctors: leveraging medical LLMs to assist doctors, not replace them. arXiv [Preprint]. 2024;2406.18034. doi:10.48550/arXiv.2406.18034.
12. Johnson A, Bulgarelli L, Pollard T, et al. MIMIC-IV (version 3.1). PhysioNet. 2024. doi:10.13026/kpb9-mt58.
13. Aali A, Van Veen D, et al. MIMIC-IV-Ext-BHC: labeled clinical notes dataset for hospital course summarization (version 1.1.0). PhysioNet. 2024. doi:10.13026/41et-8342.
14. Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. In: Inui K, Jiang J, Ng V, Wan X, editors. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); 2019 Nov; Hong Kong, China. Stroudsburg (PA): Association for Computational Linguistics; 2019. p. 3615-20. doi:10.18653/v1/D19-1371.
15. Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2019; Hong Kong, China. Stroudsburg (PA): Association for Computational Linguistics; 2019.
16. National Library of Medicine. Unified Medical Language System (UMLS). Bethesda (MD): U.S. National Library of Medicine; [2024].
17. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. 2023;10:1. doi:10.1038/s41597-022-01899-x.
18. Chiu W, Ko W, Cho W, Hui S, Chan W, Kuo M. Evaluating the diagnostic performance of large language models on complex multimodal medical cases. J Med Internet Res. 2024;26:e53724. Available from: https://www.jmir.org/2024/1/e53724. DOI: 10.2196/53724
19. DataCamp. Zero-shot prompting. Available from: https://www.datacamp.com/tutorial/zero-shot-prompting.
20. Zhou S, Xu Z, Zhang M, et al. Large Language Models for Disease Diagnosis: A Scoping Review. arXiv. 2024 Sep 19. Available from: https://arxiv.org/html/2409.00097v2
21. Symbio6. Zero-shot prompting benchmarking. Available from: https://symbio6.nl/en/blog/zero-shot-prompting-benchmarking.
22. Xie Q, Schenck EJ, Yang HS, Chen Y, Peng Y, Wang F. Faithful AI in Medicine: A systematic review with large language models and beyond. medRxiv [Preprint]. 2023 Jul 1;2023.04.18.23288752. doi: 10.1101/2023.04.18.23288752. PMID: 37398329; PMCID: PMC10312867.
23. Li W, Wu W, Chen M, Liu J, Xiao X, Wu H. Faithfulness in natural language generation: A systematic survey of analysis, evaluation, and optimization methods. arXiv [Preprint]. 2022 Mar 10. Available from: https://doi.org/10.48550/arXiv.2203.05227.
24. Carberry J, Xu H. A hierarchical fine-grained deep learning model for automated medical coding. In: Proceedings of the 2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI); 2024; Mt Pleasant, MI, USA. IEEE; 2024. p. 1-6. doi:10.1109/ICMI60790.2024.10585710.
25. Caralt MH, Ng CBL, Rei M. Continuous predictive modeling of clinical notes and ICD codes in patient health records. In: Demner-Fushman D, Ananiadou S, Miwa M, Roberts K, Tsujii J, editors. Proceedings of the 23rd Workshop on Biomedical Natural Language Processing; 2024 Aug; Bangkok, Thailand. Association for Computational Linguistics; 2024. p. 243-55. doi:10.18653/v1/2024.bionlp-1.19.


**Contributors**

Conceptualization: Parvati Naliyatthaliyazchayil

Formal analysis and visualization: Parvati Naliyatthaliyazchayil, Raajitha Mutyala, Judy Gichoya and Saptarshi Purkayastha

Funding acquisition: N/A

Project administration: Parvati Naliyatthaliyazchayil, Saptarshi Purkayastha

Supervision: Saptarshi Purkayastha, Judy Gichoya
All authors had access to the data, have read and approved the final manuscript, and accepted responsibility for the decision to submit it for publication. All authors have verified the data.

**Data Sharing**

MIMIC-IV is a controlled access ICU data provided to researchers only after completion of certain CITI training, hence they have not been shared directly in a repository. This data can be made available to the authors upon request. All the subject_ids used in this study are provided in supplemental material for reproducibility. All the scripts and queries used for data extraction, analysis, and visualization are shared via GitHub repo- https://github.com/pnaliyatthaliyazchayil/evaluate_chatbot_llms_for_healthcare

**Declaration of interests**

We declare no competing interests.

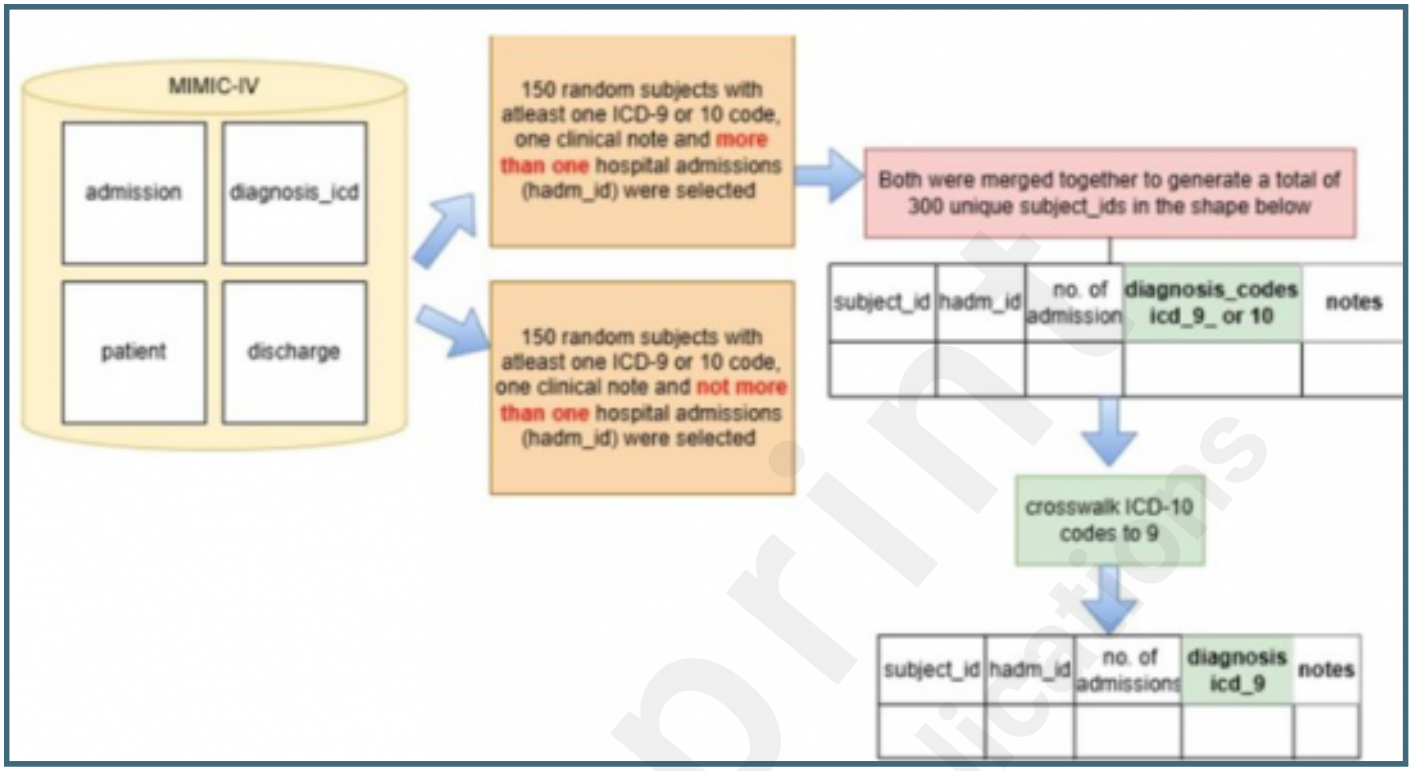
**Disclosure**

Naliyatthaliyazchayil P hereby discloses that she is currently employed by ConcertAI and has volunteered at Indiana University for this research. The research presented in this paper is independent of her work at ConcertAI. This disclosure applies solely to Naliyatthaliyazchayil P and does not extend to any of the other authors of this paper.

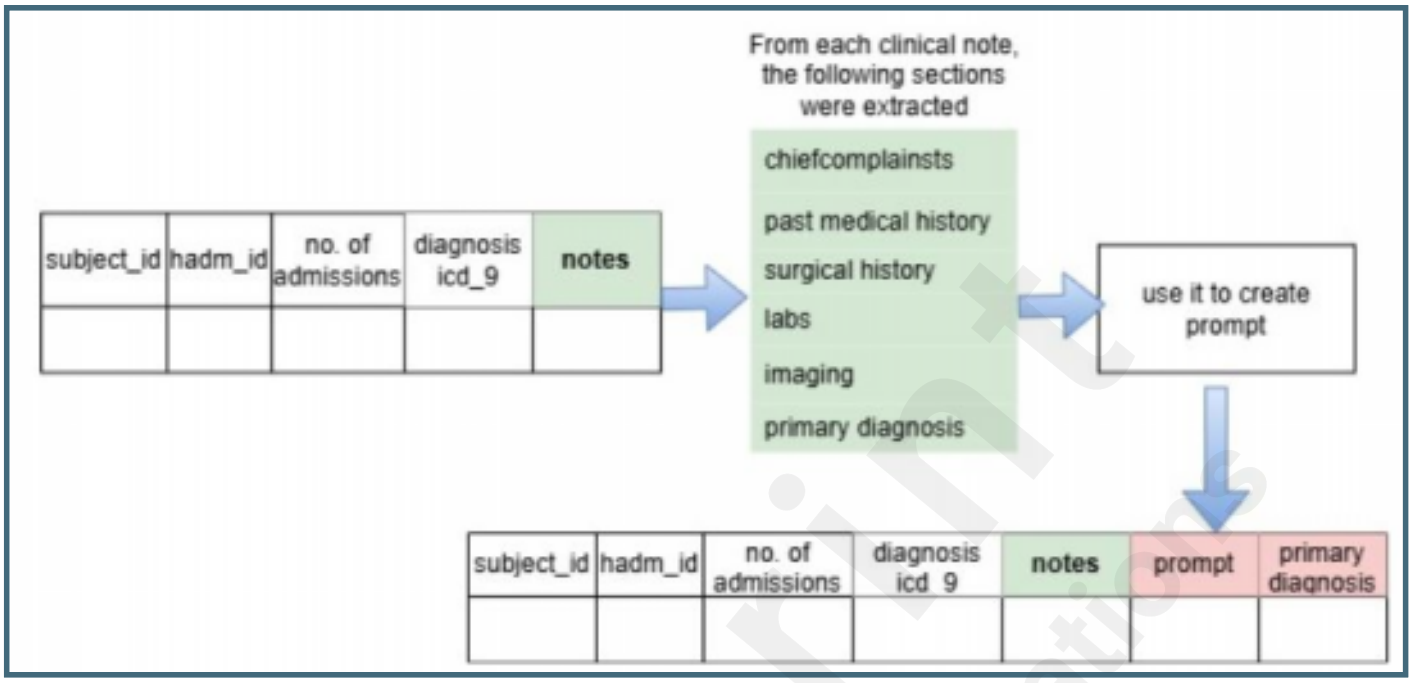
# Supplementary Files

# Figures

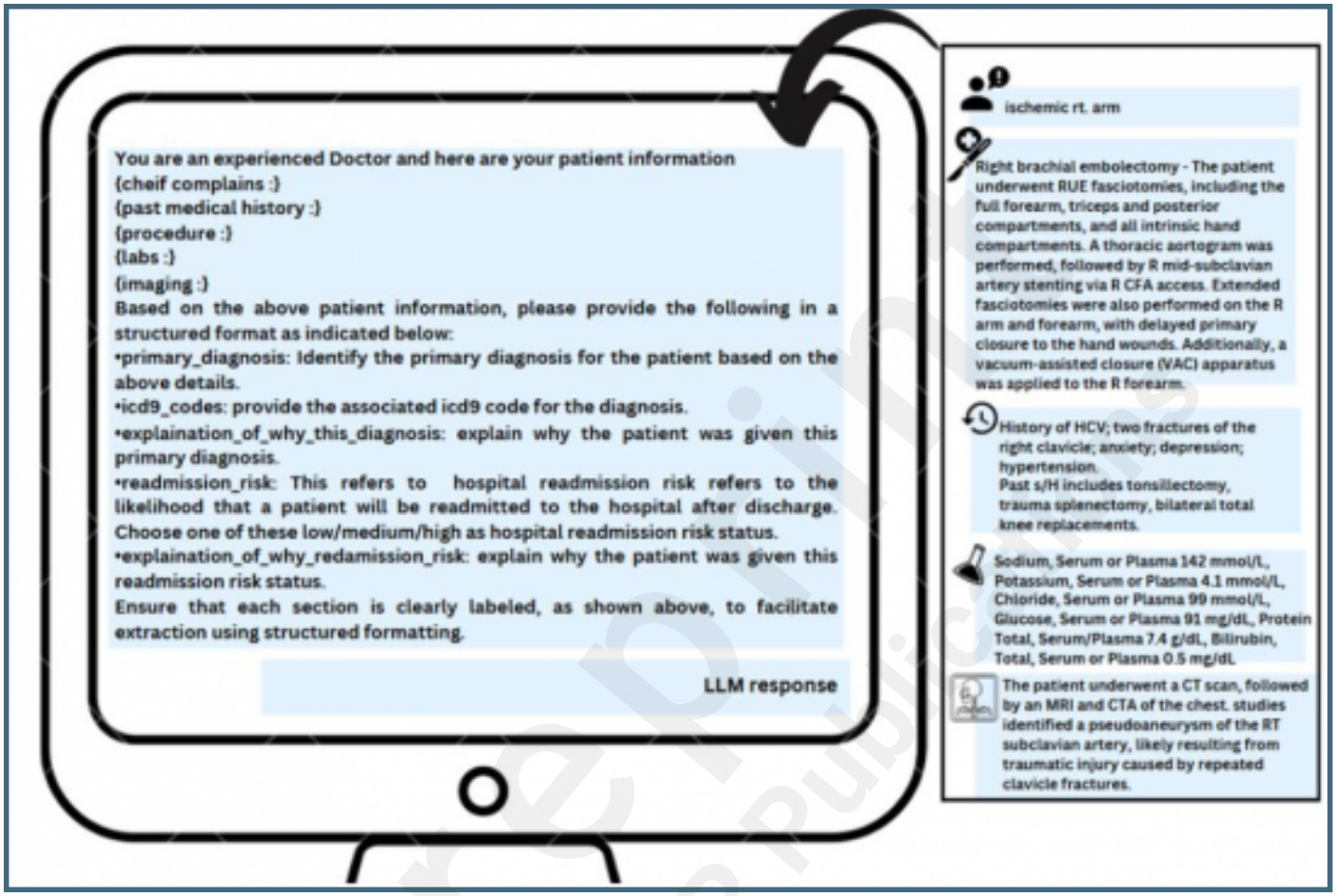
A sample of 300 subject_ids was created from the MIMIC-IV dataset, and then any ICD-10 codes in the sample were crosswalked to respective ICD-9 using UMLS crosswalk. The tables show the structure of the output for ease of understanding.



Shows how the output from Figure 1 is further utilized. The key sections from MIMIC-IV clinical notes were utilized in prompt creation and extracting primary diagnosis of the sample.



The prompt template is systematically populated for each subject_id from their notes that were extracted. On the right, you see an example representation of the MIMIC-IV note, which is then populated into its respective sections within the prompt. (The note here is an example and not an actual record from MIMIC-IV).



Shows the top 10 ICD-9 codes from the MIMIC-IV sample and the three non-reasoning LLMs. Such graphs can help us show patterns. Here, we see a pattern of Ischemic Heart diseases showing in the sample and LLM, whereas the category of General symptoms was only seen in all three LLMs and not MIMIC-IV, showing that it might be an area for a scope of improvement. Diabetes Mellitus is seen in the MIMIC-IV sample, Llama, and ChatGPT and not in Gemini, which aligns with our findings of ICD-9 code predictions where Gemini underperformed.