Evaluation of Large Language Models as Emergency Department Revisit Predictors

Emma Chen[1,2†], Luyang Luo[1†], Fatma Gunturkun[3], Sraavya Sambara[1], Rushil Arora[3], Boyang Tom Jin[3], Pranav Rajpurkar[1,*], and David A Kim[3,*]

[1]*Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA*
[2]*Harvard John A. Paulson School Of Engineering And Applied Sciences, Cambridge, MA, USA*
[3]*Stanford University, Stanford, CA, USA*
[†]*These authors contributed equally to this work*
[*]*Corresponding authors*
E-mail: yingchen@g.harvard.edu, pranav_rajpurkar@hms.harvard.edu, davidak@stanford.edu

Large Language Models (LLMs) have shown promise in clinical reasoning and question answering, yet their effectiveness for real-world clinical prediction remains an open question. We present the first large-scale study evaluating LLMs for predicting 30-day emergency department (ED) revisits using 138,010 visits from the Adult Emergency Department at Stanford. We assessed two modeling paradigms: (1) direct prediction, where the LLM generates revisit risk assessments in natural language, and (2) embedding-based approaches that leverage LLM-derived vector representations (LLM2Vec) of patient data for downstream modeling. Retrieval augmentation improved direct prediction performance (e.g., Claude 3.7 F1 from 0.3755 (95% CI [0.3647, 0.3864]) to 0.4160 (95% CI [0.4024, 0.4294])), and embedding-based methods consistently outperformed direct approaches, with LLM2Vec achieving F1=0.4505 (95% CI [0.4345, 0.4666]). Despite having access to comprehensive structured and unstructured clinical data, all LLM approaches (F1=0.3022-0.4505) failed to exceed a traditional LightGBM model using only structured data (F1=0.4614 (95% CI [0.4496, 0.4789])). Through systematic analysis of the reasoning chains in 17,488 predictions, we suggest potential failure patterns: reasoning may systematically degrade performance through overweighting medical histories and similar visits, neglecting protective factors, and risk aversion. Our work establishes essential baseline performance while revealing fundamental limitations in current-generation LLMs for clinical prediction tasks.

Keywords: Large Language Models, Emergency Medicine, Clinical Prediction

1. Introduction

Large Language Models (LLMs) have demonstrated impressive medical knowledge, with models like Med-PaLM 2 surpassing physician performance on medical licensing examinations.[1,2] These advances have generated considerable interest in applying LLMs to clinical prediction tasks, where their ability to process unstructured clinical text and leverage pretrained medical knowledge could address longstanding challenges in healthcare.

The Promise and Challenge of LLMs for Clinical Prediction. Emergency department (ED) revisit prediction represents a compelling application for LLMs. Twenty percent

of discharged ED patients return within 30 days, often due to incomplete evaluation, premature discharge, or inadequate follow-up.[3] Predicting these revisits is challenging because the underlying drivers are multifactorial—spanning clinical, behavioral, and social factors often embedded in unstructured clinical notes and distributed across temporally separated encounters. LLMs are uniquely positioned to address this complexity through their ability to: (1) leverage rich medical knowledge from pretraining, (2) process unstructured inputs like clinical notes and radiology reports, and (3) model longitudinal patient histories across multiple visits.[4–6]

Research Questions and Approach. To evaluate LLM potential for clinical prediction, this study addresses three fundamental questions: (1) What performance can current frontier LLMs achieve in 30-day ED revisit, and how does that compare to a traditional machine learning approach? (2) What LLM modeling paradigms are suitable for predicting 30-day ED revisit? (3) What clinical knowledge do LLMs use when making predictions, and how does reasoning affect this process? Our systematic evaluation across multiple state-of-the-art models and integration strategies provides crucial insights into both the capabilities and limitations of current LLM approaches for complex clinical prediction tasks.

Key Contributions. Our work makes four contributions to understanding LLMs in clinical prediction:

- **Comprehensive evaluation framework:** Using 138,010 ED visits, we establish the first large-scale benchmark comparing direct prediction and embedding-based approaches across multiple state-of-the-art models.
- **Open challenge identification:** We demonstrate that current-generation LLMs cannot exceed traditional machine learning approaches in 30-day ED revisit prediction, even when given access to both structured and unstructured data while the traditional approach uses only structured features.
- **Modeling strategy insights:** We highlight the importance of retrieval augmentation for improving LLM performance. We also show that embedding-based approaches consistently outperform direct prediction, suggesting that LLMs' primary value may lie in representation learning rather than end-to-end reasoning for complex prediction tasks.
- **Systematic analysis of reasoning failures:** Through detailed examination of 17,488 predictions, we identify patterns suggesting how reasoning may degrade LLM performance through systematic biases, including risk aversion and medical history overweighting.

2. Related Work

2.1. *Structured Electronic Health Record (EHR) Foundation Models*

Transformer-based foundation models trained on structured EHR data have emerged as a promising approach to clinical prediction. These models treat EHR sequences as analogous to natural language, encoding events as tokens and leveraging pretraining objectives such as masked event prediction or time-to-event modeling. For example, CLMBR (Clinical Language Model-Based Representations) demonstrated that pretraining on longitudinal structured data improves robustness and sample efficiency in downstream clinical tasks.[7] Similarly, MOTOR

(Many Outcome Time Oriented Representations) introduces a self-supervised time-to-event modeling framework and achieves state-of-the-art performance across 19 clinical tasks using over 9 billion clinical events from 55 million patients.[8] These methods offer robustness to temporal distribution shifts and perform well under limited supervision. However, they require large-scale, site-specific pretraining, and typically exclude unstructured data such as clinical notes or radiology reports.

2.2. *LLMs for Clinical Reasoning and Outcome Prediction*

In parallel, LLMs pretrained on general-domain text have demonstrated impressive capabilities in clinical reasoning and medical question answering.[9] Med-PaLM and Med-PaLM 2, for example, surpass the passing threshold on the US Medical Licensing Examination (USMLE) and outperform physicians on key clinical axes such as reasoning, factual accuracy, and safety.[1,2] Similarly, NYUTron shows that LLMs trained on unstructured clinical notes can serve as general-purpose predictive engines, achieving AUCs between 78.7-94.9% across five clinical and operational tasks within a real-world health system.[10] These results underscore the potential of LLMs to extract high-value representations from unstructured data and generalize across a wide range of clinical contexts with minimal task-specific tuning.[5,11]

The potential of LLMs in clinical prediction is enhanced by recent developments that convert LLMs into effective embedding generators. These techniques, such as LLM2Vec, transform decoder-only LLMs into bidirectional text encoders capable of generating dense vector representations from multimodal EHR inputs for downstream ML tasks.[12] In parallel, specialized text embedding models like Voyage-3-Large (Voyage AI Innovations, Inc.) have emerged as strong baselines for general-purpose text representation learning, offering state-of-the-art performance across diverse text understanding tasks.

2.3. *LLMs for Clinical Outcome Prediction in Emergency Settings*

Several recent studies have explored the use of LLMs for clinical outcome prediction in emergency settings. Gebrael et al. evaluated ChatGPT's ability to triage patients with metastatic prostate cancer in the ED and found high sensitivity in identifying cases requiring admission, along with improved comprehensiveness and accuracy of diagnostic suggestions compared to physicians.[13] Glicksberg et al. investigated GPT-4 for predicting hospital admissions from ED encounters, comparing its performance to traditional ML models with and without prompting and numerical probabilities.[14] Both studies highlight LLMs' potential to support clinical decision-making in emergency care, especially when structured and unstructured data are combined. However, these works rely on naïve prompting strategies and do not explore the use of LLM-derived embeddings or fine-grained modeling of revisit risk. Moreover, they focus on admission prediction rather than revisit forecasting, which is a simpler task whose outcome is known within hours rather than days or weeks.

3. Methods

To evaluate the readiness of LLMs for clinical prediction tasks, we evaluated five approaches for predicting 30-day ED revisit risk. Our methods fall into two main categories: (1) direct

prediction, where the LLM generates revisit risk assessments through its generated text outputs, and (2) embedding-based approaches that use LLM-generated vector representations of patient data for modeling. (Figure 1)

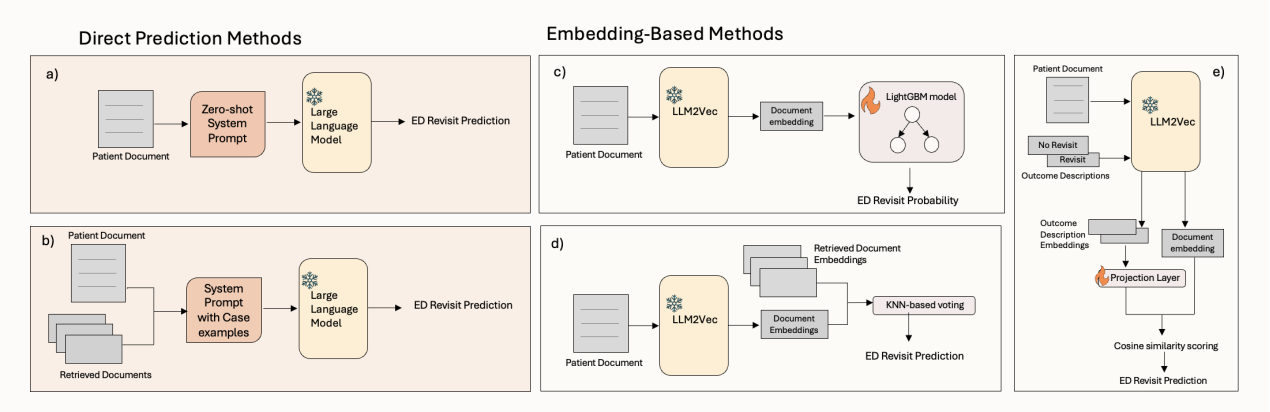

Fig. 1. Overview of Large Language Model (LLM) utilization for 30-day Emergency Department (ED) revisit prediction. Our approach explores two primary paradigms: (a, b) Direct Prediction Methods, where LLMs generate revisit risk assessments directly via text outputs, leveraging (a) zero-shot prompting or (b) retrieval-augmented prompting with case examples. (c, d, e) Embedding-Based Methods generate LLM-derived vector representations (e.g., using LLM2Vec) of patient visit data for downstream modeling, including (c) LightGBM classification, (d) k-Nearest Neighbors (KNN) based lookup, and (e) cosine similarity scoring with retrieved outcome descriptions. Each method processes a "Patient Document" representing the patient's visit data.

3.1. *Data*

We studied 138,010 discharged ED visits from 87,112 adult ED patients, from the Adult Emergency Department at Stanford , between February 1st, 2021, and February 28th, 2024. The dataset is split chronologically into 100,987, 19,535, and 17,488 visits for training, validation, and test sets, with no patient overlap between splits. 30-day ED revisit rates in the train, validation, and test splits are 17.45%, 15.77%, and 16.45%, respectively.

The dataset contains both structured and unstructured elements from the EHR. Visit-level data includes patient demographics (age, gender, race, ethnicity), socioeconomic indicators (payor information), and visit outcomes (length of stay, disposition). Vital signs (heart rate, respiratory rate, oxygen saturation, temperature, systolic blood pressure, diastolic blood pressure) are summarized by minimum, maximum, triage, and final values throughout the visit. Orders data captures procedure and medication orders placed during the visit. Laboratory results contain both categorical interpretations (normal/abnormal) and numerical values with reference ranges. Home medication data includes generic and brand names with medication classifications. Radiology data consists of study types and de-identified free-text narrative impressions. Past medical history combines ICD-10 diagnosis codes and descriptions, Clinical Classification Software (CCS) categories, and diagnosis dates. Past ED visit history includes temporal patterns of ED utilization, with chief complaints, diagnoses, and dispositions. Hospital admission data includes admit service, level of care, and hospital length of stay. While we only predict revisits for ED patients discharged home, our predictions draw on informa-

tion from all their previous ED visits, regardless of whether those visits ended in discharge or admission. (Supplement: Table 1)

3.2. *Prediction Using Structured EHR Data*

To provide a traditional machine learning baseline, we trained a LightGBM[15] model using a binary classification objective on the structured EHR only, totaling 2,221 features. Specifically, we included five demographic variables, admission types, diagnosis codes mapped to 278 Clinical Classifications Software categories, 46 distinct chief complaints, 1,675 clinical orders (e.g., medications, procedures), seven triage measurements, 181 laboratory test results (each encoded as abnormal = 1 or normal = 0), and 36 summary features derived from vital signs (e.g., minimum or maximum heart rate). Additionally, two aggregated temporal features were derived to represent the number of ED visits and inpatient admissions within the six months preceding each index visit. The model's outputs are thresholded at the value that maximizes F1 on the validation set.

3.3. *Visit Documents*

We developed a rule-based approach to convert the raw EHR data, including both structured and unstructured elements, into a comprehensive visit document for each ED encounter. Each document begins with patient demographics and medical history, followed by current medications. Previous ED visits from the past six months are summarized with their chief complaints, diagnoses, and dispositions, including hospital admissions when applicable. The current visit section includes arrival information, chief complaint, vital signs (with triage, final, minimum, and maximum values), visit timeline (arrival, rooming, disposition decision, and departure times), procedural and medication orders, laboratory and imaging results, diagnosis, and disposition details. For visits resulting in admission, additional details about admit time, service, and level of care are included. Laboratory results are categorized by abnormality status (critical, abnormal, normal) to highlight significant findings. Timing information is in the format of time, day of the week, and month. (Supplement: Table 1, Figure 1)

3.4. *Direct LLM Prediction Methods*

We implemented two approaches where the LLM directly predicts revisit risk through text generation. We evaluated several state-of-the-art LLMs with different characteristics. Reasoning models included Claude 3.7 (claude-3-7-sonnet-20250219-v1), a multimodal model with an optional reasoning mode designed to enhance step-by-step analytical thinking, which we evaluated in both standard and reasoning mode; OpenAI o3-mini (2024-12-01-preview), a smaller, more efficient variant from OpenAI's o3 series; and Deepseek R1 (v1), a specialized model from Deepseek AI that employs structured reasoning techniques. Models without explicit reasoning modes included GPT-4o (2024-10-21), OpenAI's multimodal model optimized for instruction-following and responsiveness, but without a dedicated reasoning mode that emphasizes step-by-step analytical processing.

Zero-Shot Direct Prediction In our baseline direct prediction approach, we prompted the LLM to assess ED revisit risk without providing examples. Additionally, we calculated the prevalence of 30-day revisits in the training data and included that in the prompt:

```
You are an expert in emergency department patient risk assessment. Given a
    patient's current ED visit data, evaluate their risk of returning to the ED
    within 30 days. Assess the patient's 30-day ED revisit risk as HIGH or LOW.
    If the revisit risk is HIGH, provide the most likely chief complaint for a
    potential return visit. Consider that the 30-day ED revisit prevalence in
    this hospital is 17.45%. Pay special attention to patterns in previous visits
    , particularly repeated visits for similar complaints. Return your assessment
     in JSON format: {"risk": "HIGH" or "LOW", "risk_factors": {"list key factors
     driving the risk assessment"}}.
```

Retrieval-Augmented Generation We enhanced the direct prediction approach by implementing Retrieval-Augmented Generation (RAG), where each prediction was informed by similar cases. We used a pretrained embedding model (LLM2Vec-Meta-Llama-3-8B-Instruct-mntp-unsup-simcse) to generate embeddings for visit documents and retrieved the 10 most relevant cases from the training set for each test patient based on cosine similarity. These retrieved cases were then provided to the LLM as few-shot examples to augment the same prompt used by zero-shot direct prediction:

```
Following are 10 examples:
###
{patient's EHR data using the same template}
30-day revisit: {'Yes' or 'No'}
####
...
Now, assess the following case:
```

3.5. *Text Embedding-Based Prediction Methods*

We implemented three distinct methods using LLM-derived embeddings to predict ED revisit disposition. For our primary experiments, we used LLM2Vec-Meta-Llama-3-8B-Instruct-mntp-unsup-simcse, which converts a Llama 3 8B instruction-tuned model into a text embedding model. To establish a performance baseline, we repeated all experiments using Voyage-3-Large, a state-of-the-art general-purpose text embedding model.

Embedding Similarity Classification Our first approach uses semantic similarity to classify outcomes by comparing visit document embeddings with outcome description embeddings. We created natural language descriptions for each possible outcome and measured how closely each visit's embedding matched these outcome descriptions. The outcome descriptions were:

- No revisit with 30 days: "No return visits to the Emergency Department occur within 30 days following the current Emergency Department visit."
- Revisit within 30 days: "A return visit to the Emergency Department occurs within 30 days following the current Emergency Department visit for any reason."

Predictions were made by computing cosine similarities between the patient visit embedding and each outcome description embedding, with the highest similarity score determining the predicted outcome. Formally, for patient visit embedding $\mathbf{p}$ and outcome description em-

bedding $\mathbf{o}$, the cosine similarity is calculated as:

$$\cos(\mathbf{p}, \mathbf{o}) = \frac{\mathbf{p} \cdot \mathbf{o}}{\|\mathbf{p}\|\|\mathbf{o}\|} = \frac{\sum_{i=1}^{n} p_i o_i}{\sqrt{\sum_{i=1}^{n} p_i^2} \sqrt{\sum_{i=1}^{n} o_i^2}} \quad (1)$$

where $\mathbf{p} \cdot \mathbf{o}$ represents the dot product of the two vectors, and $||\mathbf{p}||$ and $||\mathbf{o}||$ represent their Euclidean norms (magnitudes). To enhance the discriminative power of our model, we train a single linear projection layer for the outcome embeddings. This projection is implemented as $\mathbf{o}_{\text{proj}} = W_{\text{projection}} \cdot \mathbf{o}$, where $W_{\text{projection}}$ is a learnable weight matrix. This projection helps to align the outcome embeddings with the relevant semantic features in the patient data space.

LLM Embeddings with LightGBM For our second approach, we trained a LightGBM model on the patient document embeddings to predict 30-day ED revisit. We selected LightGBM[15] due to its widespread use and demonstrated strong performance in predicting hospital admissions and readmissions across various clinical settings.[16–18]

K-Nearest Neighbor Prediction Our third embedding-based approach used k nearest-neighbor (KNN) retrieval to identify similar cases for prediction. We created patient document embeddings for visits in the training set. These embeddings were stored in a FAISS (Facebook AI Similarity Search) vector database for efficient similarity search. During prediction, we retrieved the k most similar cases based on cosine similarity between the document embedding of the patient in query and those in the vector store. Note that because there is no patient overlap between splits, the retrieved cases will only belong to patients different from the one we are making predictions for. To predict the probability of 30-day ED revisits, we calculated the proportion of similar cases in the training set that resulted in revisits, with Laplace smoothing applied to avoid zero probabilities:

$$P(\text{revisit}) = \frac{\text{count of revisit neighbors} + 1}{\text{total neighbors} + 2}$$

We experimented with multiple k values (10, 20, 30, 50, 100, 500, 1000, 2000, 5000, 10000) to find the optimal neighborhood size using the validation set.

4. Results

4.1. *Overall Performance Comparison*

All performance metrics are reported with 95% confidence intervals computed using bootstrap resampling with 1,000 iterations. Among LLM-based approaches, embedding methods consistently outperformed direct prediction. LLM-derived embeddings (LLM2Vec) achieved competitive performance at F1=0.4505 (95% CI [0.4345, 0.4666]) compared to Voyage-3-Large (F1=0.4608 (95% CI [0.4375, 0.4667])). Notably, RAG provided consistent improvements across all direct prediction models. However, we also found that a structured EHR baseline achieved performance (F1=0.4614 (95% CI [0.4496, 0.4789])) that exceeded LLM-based methods (F1=0.3022-0.4505). (Figure 2, Supplement: Table 2-7)

4.2. *Direct Prediction Performance*

Figure 3 demonstrate the performance of various LLMs in directly predicting ED revisits without additional context. F1 scores ranged from 0.3022 to 0.3889, with o3-mini achieving

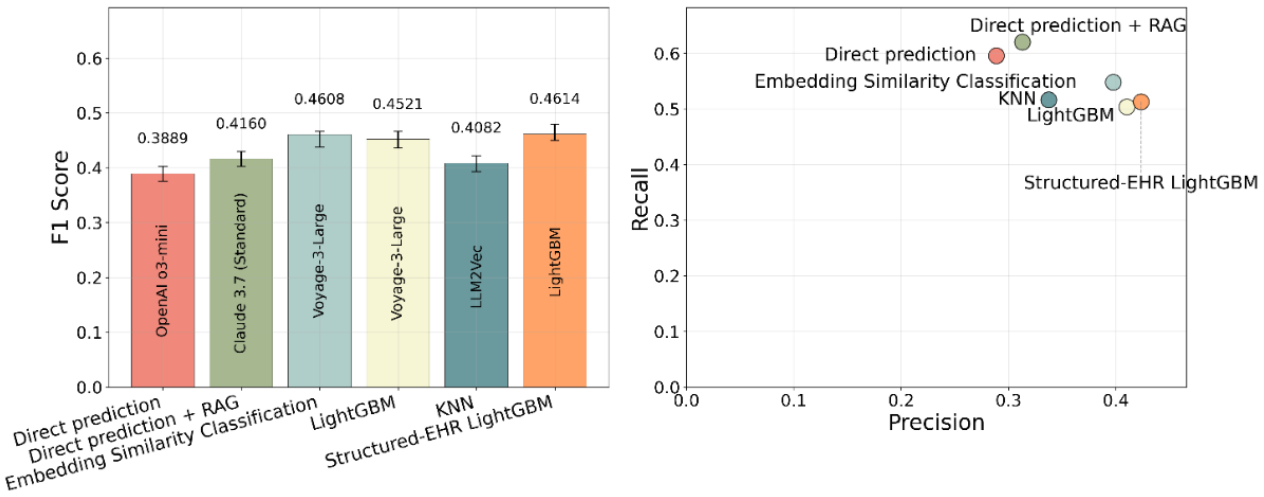

Fig. 2. Comparison of the best model across modeling paradigms.

the highest performance among zero-shot approaches. The precision-recall breakdown reveals that models exhibited different operational characteristics, with Deepseek R1 and Claude 3.7 (reasoning) favoring high recall (>85%) at the expense of precision (<20%), while GPT-4o and Claude 3.7 (standard) achieved a more balanced precision-recall profile.

Incorporating retrieval augmentation with the top 10 similar cases substantially improved performance across all models. The most significant gains were observed in GPT-4o (from F1=0.3403 to 0.3867) and Claude 3.7 (standard), with the latter achieving the highest overall F1 score of 0.4160 (95% CI [0.4024, 0.4294]) among direct prediction methods. This improvement suggests that providing LLMs with similar cases enables them to identify relevant patterns that enhance prediction accuracy, potentially mirroring how clinicians draw upon their experience with similar patients. (Figure 3; Supplement: Table 2,3)

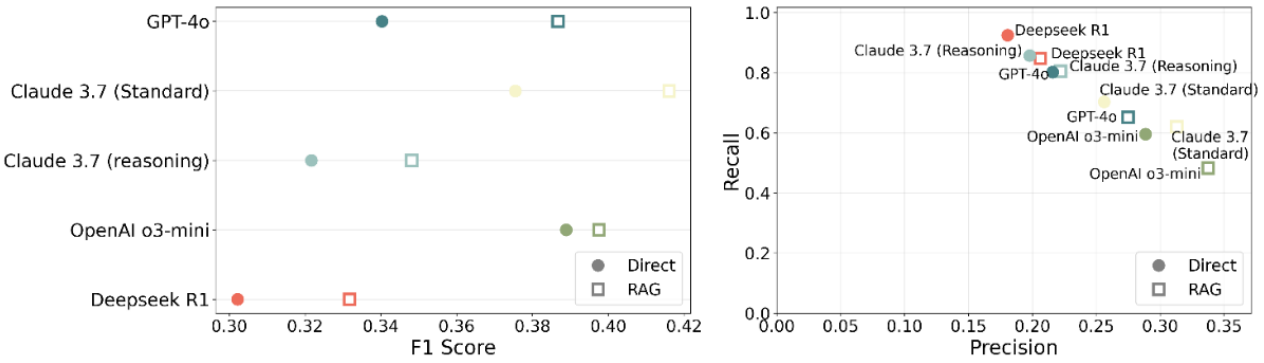

Fig. 3. Performance of direct prediction with and without retrieval augmentation (RAG) across LLMs.

4.3. *Embedding-Based Method Performance*

Figure 4 present the performance of different embedding approaches across three prediction methods: embedding similarity classification, LightGBM, and KNN-based prediction. Embedding similarity classification achieved the highest performance across all metrics (F1,

precision@recall=0.95, precision@recall=0.85) for both embedding models. Voyage-3-Large achieved the highest performance in both similarity classification (F1=0.4608 (95% CI [0.4375, 0.4667])) and LightGBM (F1=0.4521 (95% CI [0.4358, 0.4667])) approaches. However, LLM2Vec demonstrated competitive performance with F1=0.4505 (95% CI [0.4345, 0.4666]) in similarity classification and F1=0.4386 (95% CI [0.4231, 0.4531]) in LightGBM, showing that the performance gap between LLM-derived and specialized embedding models is relatively modest. (Supplement: Table 4-6)

For our KNN approach, we observed performance variations across different neighborhood sizes. With LLM2Vec embeddings, performance improved as k increased from 10 (F1=0.3688 (95% CI [0.3536, 0.3838])) to 500 (F1=0.4082 (95% CI [0.3927, 0.4218])), before declining with larger k values. Similarly, with Voyage-3-Large embeddings, performance peaked at k=100 (F1=0.4002 (95% CI [0.3849, 0.4137])). The inverted-U performance curve across k values suggests a trade-off between capturing relevant cases and introducing noise, indicating that while a moderate number of similar cases provides useful signal for revisit prediction, extremely large neighborhood sizes dilute this signal by incorporating less relevant cases. The competitive performance of simple KNN approaches (F1 >0.4) validates that both LLM2Vec and Voyage-3-Large capture clinically relevant patient similarity—when patients are similar in embedding space, their revisit outcomes are also similar, making the prevalence of revisits among retrieved neighbors a reliable predictor. (Supplement: Table 6)

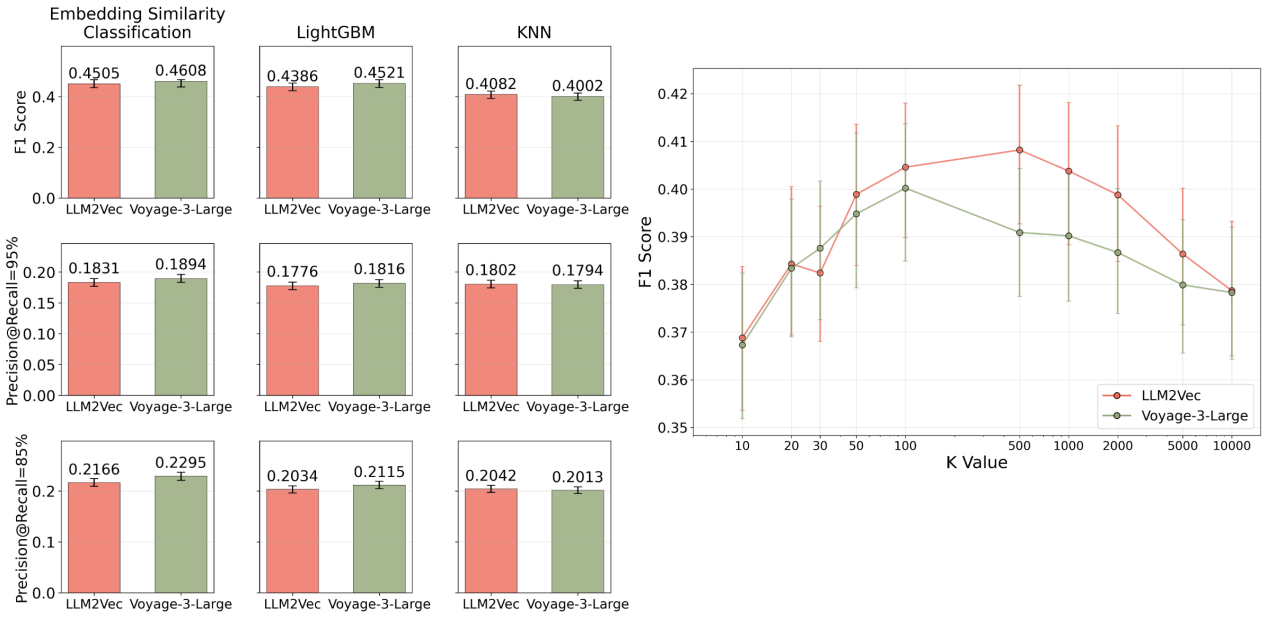

Fig. 4. Performance of embedding-based methods (similarity classification, LightGBM, KNN) using LLM2Vec and Voyage-3-Large (left) and the impact of k on performance for KNN (right).

4.4. *The Challenge of Reasoning: When Sophisticated Models Perform Worse*

Contrary to expectations, explicit reasoning modes didn't consistently improve prediction performance. This finding is particularly striking given that reasoning capability is often con-

sidered a hallmark of advanced LLMs. We conducted a quantitative and qualitative analysis on the responses of Claude 3.7 with RAG in the reasoning mode because it allows a controlled comparison against the standard mode to evaluate the impact of reasoning. (Supplement C,D)

4.4.1. *Reasoning Mode Performance Patterns*

Claude 3.7's standard mode (F1=0.3755 (95% CI [0.3647, 0.3864])) outperformed its reasoning counterpart (F1=0.3217 (95% CI [0.3125, 0.3311])). To understand this phenomenon, we analyzed prediction changes when reasoning was enabled, focusing on the 4,891 cases where Claude 3.7's predictions differed between standard and reasoning modes.

The analysis revealed a systematic bias: the reasoning mode predominantly changed predictions toward higher risk. Of cases where predictions were flipped after enabling reasoning:

- Correct improvements: 542 cases changed from Low to High risk correctly, 69 cases changed from High to Low risk correctly
- Incorrect changes: 4,271 cases changed from Low to High risk incorrectly, 9 cases changed from High to Low risk incorrectly

4.4.2. *Systematic Analysis of Reasoning Failures*

We conducted a comprehensive qualitative analysis using GPT-4o to examine Claude 3.7's reasoning chains on cases where reasoning changed the prediction. Summarizing the qualitative analysis with Claude Sonnet 4 suggest several factors that may cause reasoning-mode failures:

When Reasoning Hurts Performance (4271 cases): The most common failure occurred when correct low-risk predictions became incorrect high-risk predictions. Key patterns included:

- *Medical history overweighting*: Excessive emphasis on chronic conditions and complex medical histories while undervaluing protective factors, including no prior ED visits, normal test results, stable vital signs, and etc. (87.59% cases)
- *Risk aversion*: The step-by-step reasoning made the model more conservative, favoring high-risk predictions to avoid missing potential problems (76.96% cases)
- *Statistical misinterpretation*: Focus on small percentages of similar cases that experienced revisits while ignoring that the majority did not return (41.65% cases)

When Reasoning Helps Performance (542 cases): Reasoning successfully identified missed high-risk cases by:

- *Capturing subtle risk factors*: Noticing demographic and social predictors, such as insurance status, medication adherence issues, that affect follow-up (90.04%)
- *Identifying untreated conditions*: Finding abnormal laboratory results with no or inadequate corresponding interventions (63.28% cases)
- *Leveraging temporal patterns*: Finding escalating patterns by connecting current visit to recent ED visits for related complaints (26.38% cases)
- *Catching concerning vital signs*: Noticing worsening vitals at discharge, such as low SpO2, persistent tachycardia, that signal instability (25.46% cases)
- *Recognizing diagnostic mismatches*: Identifying cases where discharge diagnoses didn't match presenting symptoms (10.52% cases)

4.4.3. *Factors Underlying LLM Predictions*

To characterize what factors drive LLM predictions, we used OpenAI's text-embedding-3-large to embed the "*risk_factors*" in Claude 3.7's responses and applied k-means clustering, identifying 56 factor groupings. The full list with random samples from each grouping can be found in Supplement Table 8. The five most frequently considered factors align with established revisit prediction literature: (1) Population Revisit Patterns - comparisons to similar cases and hospital baseline rate (n=3530), (2) Very Recent Prior Visits - ED visits within days/weeks of current presentation (n=2755), (3) Frequent ED Utilization - patterns of multiple recent visits (n=1857), (4) ED Length of Stay - duration of encounters (n=1821), and (5) Ongoing Treatment Protocols - continued therapies requiring follow-up (n=1757).[19–21]

To understand whether the LLM utilized the provided similar cases for prediction, we prompted GPT-4o to classify any references to the retrieved cases in the reasoning chain, showing that 99.15% of the cases have references. Besides, the k-mean clustering above shows that comparisons to similar cases and hospital baseline revisit rates is the most used risk factor, confirming again that the LLM really reasons over similar cases.

5. Discussion

Current LLMs Cannot Exceed Traditional Approaches Despite Capability to Process Richer Data Our study reveals a striking finding that despite providing LLMs with comprehensive structured and unstructured clinical data, all LLM-based approaches (F1=0.3022-0.4505) failed to exceed a simple LightGBM model trained only on structured EHR features (F1=0.4614 (95% CI [0.4496, 0.4789])). While LLMs showed varying effectiveness depending on integration strategies, the fundamental limitation persists across all approaches, challenging assumptions about LLMs' readiness for complex clinical prediction tasks.

Zero-Shot Prediction Demonstrates High Recall but Limited Precision Direct zero-shot prediction using LLMs yielded modest performance (F1 scores 0.3022-0.3889), with these models demonstrating a tendency toward high recall at the expense of precision. This overestimate of revisit risk is potentially caused by the characteristics of ED patients who typically have multiple comorbidities and acute presentations. The clinical consequence of such behavior would be an excessive number of false positives, limiting practical utility in resource-constrained healthcare environments where targeted interventions must be allocated efficiently.

Retrieval-Based Methods Offer Interpretability and Clinically Meaningful Context Notably, our simple KNN approaches using LLM embeddings achieved competitive performance (F1=0.4082 (95% CI [0.3927, 0.4218]) and 0.4002 (95% CI [0.3849, 0.4137]) for LLM2Vec and Voyage-3-Large respectively), suggesting the similar cases are clinically informative. This approach combines reasonable performance with high interpretability, as similar cases can be presented alongside predictions to provide clinicians with concrete examples informing the risk assessment. Such transparency may enhance clinical trust and adoption while enabling physicians to incorporate their judgment when similar cases reveal important nuances not captured in the prediction itself.

Retrieval augmentation with similar cases significantly improved LLM performance across

all models evaluated. This enhancement was particularly pronounced for Claude 3.7 (standard mode with reasoning disabled) and GPT-4o. The consistent benefit of retrieval augmentation suggests that LLMs can effectively leverage pattern recognition from relevant cases—a process that mirrors clinical reasoning, where physicians draw upon their experience with similar patients to inform prognostication. This finding aligns with recent work showing that case-based reasoning can enhance clinical decision support systems.[14]

Embedding-Based Models Yield Strong Performance with Tradeoffs in Interpretability Embedding-based approaches demonstrated superior performance compared to direct prediction methods, with LLM2Vec embeddings achieving F1 scores exceeding 0.45 when combined with embedding similarity classification. However, this performance gain comes with a tradeoff in interpretability—while direct LLM predictions can include explanatory text and reference similar cases, methods built on embeddings function as relative black boxes from a clinical perspective.

Our analysis of LLM reasoning suggests a potential paradox The very capability that makes LLMs appear more "intelligent"—their ability to engage in step-by-step reasoning—may systematically degrade prediction performance in clinical contexts. Through automated analysis of reasoning chains, we observed that explicit reasoning appears more risk averse and may overly focus on small percentages of similar cases that experienced revisits while potentially ignoring that the majority did not return. It also seems to overweight complex medical histories and seek confirmatory evidence for identified risk factors, while potentially discounting contradictory evidence and protective factors. These potential confirmation biases in LLM reasoning may mirror well-documented challenges in human clinical decision-making,[22] though further validation with human expert review is needed to confirm these patterns.

Limitations and Future Directions Our dataset includes EHR components unavailable in MIMIC-IV-ED, the only publicly available ED dataset, precluding external validation. We focused on binary revisit prediction rather than more clinically actionable outcomes such as preventable returns or severity-stratified revisits. Our outcome data captures only Stanford ED revisits, potentially missing visits to other emergency departments; however, this limitation affects all methods equally and should not bias comparative performance. We did not prospectively evaluate the impact of these predictions on clinical decision-making or patient outcomes. Our models use complete ED visit data available at discharge, limiting immediate applicability for pre-emptive interventions during the visit, though many interventions can be delivered after discharge (e.g., expedited clinic or telemedicine follow-up), and LLMs' flexible input format enables adaptation to earlier time points during visits at potential cost to accuracy. While our findings reveal fundamental limitations of current LLMs for this task, targeted post-training approaches warrant investigation. Supervised fine-tuning or reinforcement learning from human feedback (RLHF/DPO[23,24]) incorporating physician evaluations could potentially improve both prediction accuracy and clinical alignment.

Supplement `https://github.com/dkimlab/PSB_2026_EDLLM_Supplement`

Acknowledgement This research was supported in part by the National Institutes of Health, grant number 1R01HL172794.

References

1. K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal, M. Schaekermann, A. Wang, M. Amin, S. Lachgar, P. Mansfield, S. Prakash, B. Green, E. Dominowska, B. Aguera y Arcas, N. Tomasev, Y. Liu, R. Wong, C. Semturs, S. S. Mahdavi, J. Barral, D. Webster, G. S. Corrado, Y. Matias, S. Azizi, A. Karthikesalingam and V. Natarajan, Towards Expert-Level Medical Question Answering with Large Language Models (2023), arXiv preprint arXiv:2305.09617. https://arxiv.org/abs/2305.09617.
2. K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, A. Babiker, N. Schärli, A. Chowdhery, P. Mansfield, D. Demner-Fushman, B. Agüera y Arcas, D. Webster, G. S. Corrado, Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. Barral, C. Semturs, A. Karthikesalingam and V. Natarajan, Large language models encode clinical knowledge, *Nature* **620**, 172 (2023), https://www.nature.com/articles/s41586-023-06291-2.
3. J. S. Dhaliwal and A. K. Dang, Reducing Hospital Readmissions, in *StatPearls [Internet]*, (StatPearls Publishing, Treasure Island (FL), 2024) Last updated June 7, 2024. https://www.ncbi.nlm.nih.gov/books/NBK606114/.
4. H. Ahsan, D. J. McInerney, J. Kim, C. Potter, G. Young, S. Amir and B. C. Wallace, Retrieving Evidence from EHRs with LLMs: Possibilities and Challenges, *Proceedings of Machine Learning Research* **248**, 489 (2024), PMID: 39224857, PMCID: PMC11368037. https://pubmed.ncbi.nlm.nih.gov/39224857/.
5. G. Wang, G. Yang, Z. Du, L. Fan and X. Li, CLINICALGPT: Large Language Models Finetuned with Diverse Medical Data and Comprehensive Evaluation (2023), arXiv:2306.09968 [cs.CL]. https://arxiv.org/abs/2306.09968.
6. Q. Shen, X. Zhang, H. Ren, Q. Guo and Z. Yi, Knowledge-embedded large language models for emergency triage, *Knowledge-Based Systems* **283**, p. 110574 (2025), https://www.sciencedirect.com/science/article/abs/pii/S0950705125004782.
7. E. Steinberg, K. Jung, J. A. Fries, C. K. Corbin, S. R. Pfohl and N. H. Shah, Language Models Are An Effective Patient Representation Learning Technique For Electronic Health Record Data (2020), arXiv:2001.05295 [cs.CL]. https://arxiv.org/abs/2001.05295.
8. E. Steinberg, J. Fries, Y. Xu and N. Shah, MOTOR: A Time-To-Event Foundation Model For Structured Medical Records (2023), arXiv:2301.03150 [cs.LG]. https://arxiv.org/abs/2301.03150.
9. C. Y. K. Williams, B. Y. Miao, A. E. Kornblith and A. J. Butte, Evaluating the use of large language models to provide clinical recommendations in the Emergency Department, *Nature Communications* **15**, 1 (2024), https://www.nature.com/articles/s41467-024-52415-1.
10. L. Y. Jiang, X. C. Liu, N. Pour Nejatian, M. Nasir-Moin, D. Wang, A. Abidin, K. Eaton, H. A. Riina, I. Laufer, P. Punjabi *et al.*, Health system-scale language models are all-purpose prediction engines, *Nature* **619**, 357 (2023), https://www.nature.com/articles/s41586-023-06160-y.
11. H. Wang, C. Liu, N. Xi, Z. Qiang, S. Zhao, B. Qin and T. Liu, HuaTuo: Tuning LLaMA Model with Chinese Medical Knowledge (2023), arXiv:2304.06975 [cs.CL]. https://arxiv.org/abs/2304.06975.
12. P. BehnamGhader, V. Adlakha, M. Mosbach, D. Bahdanau, N. Chapados and S. Reddy, LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders, in *Proceedings of the Conference on Language Modeling (COLM)*, 2024. arXiv preprint arXiv:2404.05961. https://arxiv.org/abs/2404.05961.
13. G. Gebrael, K. K. Sahu, B. Chigarira, N. Tripathi, V. M. Thomas, N. Sayegh, B. L. Maughan, N. Agarwal, U. Swami and H. Li, Enhancing Triage Efficiency and Accuracy in Emergency Rooms for Patients with Metastatic Prostate Cancer: A Retrospective Analysis of Artificial Intelligence-

Assisted Triage Using ChatGPT 4.0, *Cancers* **15**, p. 3717 (2023), https://pubmed.ncbi.nlm.nih.gov/37509379/.
14. B. S. Glicksberg, P. Timsina, D. Patel, A. Sawant, A. Vaid, G. Raut, A. W. Charney, D. Apakama, B. G. Carr, R. Freeman *et al.*, Evaluating the accuracy of a state-of-the-art large language model for prediction of admissions from the emergency room, *Journal of the American Medical Informatics Association* **31**, 1921 (2024).
15. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, in *Advances in Neural Information Processing Systems 30*, 2017.
16. Y. Hu, F. Ma, M. Hu, B. Shi, D. Pan and J. Ren, Development and validation of a machine learning model to predict the risk of readmission within one year in HFpEF patients: Short title: Prediction of HFpEF readmission, *International Journal of Medical Informatics* **194**, p. 105703 (2025).
17. J. Miao, C. Zuo, H. Cao, Z. Gu, Y. Huang, Y. Song and F. Wang, Predicting ICU readmission risks in intracerebral hemorrhage patients: Insights from machine learning models using MIMIC databases, *Journal of the Neurological Sciences* **456**, p. 122849 (2024).
18. C. Brossard, C. Goetz, P. Catoire, L. Cipolat, C. Guyeux, C. Gil Jardine, M. Akplogan and L. Abensur Vuillaume, Predicting emergency department admissions using a machine-learning algorithm: a proof of concept with retrospective study, *BMC Emergency Medicine* **25**, p. 3 (2025).
19. S. Janda and J. Sansuk, Factors Associated with Re-attendance at Emergency Departments Among Older Adults: A Cross-Sectional Analytical Study, *Inquiry* **62**, p. 469580251349652 (2025), https://doi.org/10.1177/00469580251349652.
20. I. Dufour, M. C. Chouinard, N. Dubuc, J. Beaudin, S. Lafontaine and C. Hudon, Factors associated with frequent use of emergency-department services in a geriatric population: a systematic review, *BMC Geriatrics* **19**, p. 185 (2019), https://doi.org/10.1186/s12877-019-1197-9.
21. P. Tangkulpanich, C. Yuksen, W. Kongchok and C. Jenpanitpong, Clinical predictors of emergency department revisits within 48 hours of discharge; a case control study, *Archives of academic emergency medicine* **9**, p. e1 (2020).
22. B. Sætrevik, V. T. Seeligmann, T. F. Frotvedt and Ø. Keilegavlen Bondevik, Anchoring, confirmation and confidence bias among medical decision-makers, *Collabra: Psychology* **10**, p. 126223 (2024).
23. P. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg and D. Amodei, Deep reinforcement learning from human preferences (2023).
24. R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning and C. Finn, Direct preference optimization: Your language model is secretly a reward model (2024).