

ClinNoteAgents: An LLM Multi-Agent System for Predicting and Interpreting Heart Failure 30-Day Readmission from Clinical Notes

Rongjia Zhou¹, Chengzhuo Li¹, Carl Yang, PhD^{1*}, Jiaying Lu, PhD^{1*}
¹Emory University, Atlanta, GA, USA

ABSTRACT

Heart failure (HF) is one of the leading causes of rehospitalization among older adults in the United States. Although clinical notes contain rich, detailed patient information and make up a large portion of electronic health records (EHRs), they remain underutilized for HF readmission risk analysis. Traditional computational models for HF readmission often rely on expert-crafted rules, medical thesauri, and ontologies to interpret clinical notes, which are typically written under time pressure and may contain misspellings, abbreviations, and domain-specific jargon. We present ClinNoteAgents, an LLM-based multi-agent framework that transforms free-text clinical notes into (1) structured representations of clinical and social risk factors for association analysis and (2) clinician-style abstractions for HF 30-day readmission prediction. We evaluate ClinNoteAgents on 3,544 notes from 2,065 patients (readmission rate=35.16%), demonstrating high extraction fidelity for clinical variables (conditional accuracy $\geq 90\%$ for multiple vitals), key risk factor identification, and preservation of predictive signal despite 60–90% text reduction. By reducing reliance on structured fields and minimizing manual annotation and model training, ClinNoteAgents provides a scalable and interpretable approach to note-based HF readmission risk modeling in data-limited healthcare systems.

INTRODUCTION

Heart failure (HF) remains a major global health challenge, affecting more than 55 million individuals worldwide, with nearly 80% of cardiovascular deaths occurring in low- and middle-income nations.¹ Approximately 25% of HF patients are readmitted within 30 days,² imposing substantial clinical and financial burdens on health systems. Multiple studies have identified diverse contributors to readmission of HF, including HF exacerbation,³ comorbidities such as *chronic obstructive pulmonary disease*, *chronic kidney disease*,⁴ and socioeconomic factors.⁵ Given the complex, multi-factorial nature of HF, computational risk modeling for HF readmission often requires comprehensive longitudinal patient data.⁶ Electronic health record (EHR) data, which capture demographics, diagnoses, laboratory results, and medications, have therefore become a central data source for studying HF outcomes, including 30-day readmission.^{7,8} While containing valuable multimodal health information for HF patients, EHRs are often incomplete or unavailable in developing countries⁹ due to financial, technological, and organizational barriers.¹⁰ This challenge is particularly evident in developing countries in Asia and Africa. Many hospitals in Bangladesh and Indonesia continue to rely on handwritten or locally stored digital notes due to financial constraints, limited IT capacity, and poor interoperability infrastructure.^{11,12} Similarly, healthcare facilities in Kenya, Uganda, and Ghana frequently experience unstable internet connectivity and shortages of trained health informatics personnel.^{13,14} In these settings, unstructured clinical notes often serve as the primary source of documented patient information. Despite widespread adoption of EHR in the U.S., around 80% of clinical information remains embedded in free-text notes.¹⁵ Therefore, clinical notes offer a pragmatic strategy for constructing predictive models when access to structured EHR data is limited.

Early HF readmission models relied primarily on structured EHR data such as demographics, socioeconomic status, medical history, and laboratory measurements.^{16,17} More recent work has incorporated unstructured clinical notes to capture richer contextual and temporal information,¹⁸ with hybrid models combining structured variables and note-derived embeddings yielding further gains.¹⁹ This shift underscores the potential of text-based modeling for early detection of readmission risk. Growing evidence also highlights the importance of social determinants of health (SDOHs), defined by the World Health Organization as the conditions in which individuals live and work.²⁰ SDOHs have been repeatedly linked to HF outcomes, with employment status, housing stability, and social support identified as major contributors to readmission risk.²¹ However, SDOHs are rarely structured in EHRs, limiting their use in predictive systems. Recent advances in natural language processing (NLP) and large language models (LLMs) have

*Corresponding Authors: Carl Yang (j.carlyang@emory.edu), Jiaying Lu (jiaying.lu@emory.edu).

enabled the extraction of both clinical and social risk factors from unstructured notes, with machine learning and transformer models achieving state-of-the-art performance in HF readmission prediction using discharge notes.^{22,23} Summarization-based preprocessing is shown to further enhance predictive signal quality,²⁴ and recent studies show that LLMs can extract SDOH with near-clinician accuracy.^{25,26} Despite these advances, most existing approaches rely on predefined SDOH taxonomies or focus solely on social factors, often overlooking clinical predictors for readmission risk.²⁷ These gaps highlight the need for a unified framework that jointly extracts and harmonizes social and clinical determinants from discharge notes to enable scalable readmission modeling.

To address these limitations, we present ClinNoteAgents, a LLM-based modular framework that transforms unstructured discharge notes into structured, interpretable representations of clinical and social risk factors for HF 30-day readmission. Our framework integrates (1) a structural extractor of clinical and social risk factors for statistical analysis of their relationship to readmission outcomes, and (2) a summarizer that produces qualitative or mixed-evidence summaries for predictive modeling. By reducing reliance on structured EHR and minimizing manual annotation, ClinNoteAgents enables scalable HF risk analysis from discharge documentation in data-limited healthcare settings.

METHOD

Study Design

In this study, we leverage discharge notes, a rich clinical text source containing patients’ demographics, chief complaints, comorbidities, clinical measurements, social determinants of health and other critical risk factors, as the primary data source for building computational risk analysis models of HF 30-day readmission. This design highlights the importance of clinical notes as both a practical alternative to structured EHR data in resource-limited settings and an under-exploited information source in data-rich health systems. To achieve HF 30-day readmission risk analysis, it involves (1) predicting the readmission risk; and (2) identifying the driving factors of that risk. We therefore formalize the scientific problem as two interdependent subtasks.

Heart failure readmission risk prediction. Formally, let $\mathbf{X}_i = \{x_1, x_2, \dots, x_l\}$ denote the discharge note of length l textual sequence for the i -th hospital admission, and let $y_i \in \{0, 1\}$ be the binary indicator of whether the patient was readmitted within 30 days ($y_i = 1$) or not ($y_i = 0$). The goal is to learn a predictive model $f : \mathbf{X}_i \rightarrow y_i$, that predicts whether a patient will be readmitted within 30 days based on the input discharge note.

Heart failure readmission risk factors mining. To systematically quantify the associations between risk factors \mathbf{R}_i and 30-day HF readmission outcome y_i , the first step is to identify a structured set of risk factors from each discharge note and then conduct statistical analyses to evaluate their relationships with readmission risk. Specifically, the discharge note-based risk factors extraction process can be defined as $\mathbf{R}_i = h(\mathbf{X}_i)$, where $h(\cdot)$ denotes the risk factor extractor, and $\mathbf{R}_i = \{(r_j, v_j)\}_{j=1}^k$ denotes a set of risk factors with each factor type r_j associated with a risk factor value v_j . Examples of extracted tuples (r_j, v_j) include (body temperature, 97), (condition, renal cell cancer), (employment, retired). After obtaining structured risk factor representations \mathbf{R}_i for each discharge note, the second step is to conduct statistical analysis to quantify their associations with 30-day readmission. Depending on the data type and distribution of each risk factor, appropriate statistical tests will be applied (e.g., chi-square test for categorical variables and logistic regression for continuous variables) to obtain both p-values and effect sizes.

Data Source

This study used the publicly available Medical Information Mart for Intensive Care III (MIMIC-III) database.²⁸ Following established conventions in prior studies,²⁷ heart failure (HF) patients were identified using ICD-9 diagnosis codes (398.91, 402.01, 402.11, 402.91, 404.01, 404.03, 404.11, 404.13, 404.91, 404.93, and all codes beginning with 428). For each patient with multiple hospitalizations, we constructed readmission pairs by linking each index admission with its subsequent readmission. The discharge summary from the earlier admission served as the input note \mathbf{X}_i for modeling, while the outcome label y_i was determined by the time interval between discharge and the next admission ($y_i = 1$ if the interval ≤ 30 days; otherwise $y_i = 0$). Table 1 summarizes the cohort used in this study.

Table 1: Summary statistics of the heart failure cohort.

| # Patients | ReAdm (%) | # Notes | # Discharge Notes | % Female | Median Age (Q1–Q3) | Median LoS (Q1–Q3) |
|------------|-----------|---------|-------------------|----------|--------------------|--------------------|
| 2065 | 35.16% | 3604 | 3544 | 46.39% | 72.9 (62.6–81.3) | 7.0 (4.0–13.0) |

Proposed LLM-based Multi-agent System

We propose ClinNoteAgents, an LLM-based multi-agent system for comprehensive clinical notes analytics. Our system comprises three agents: the Risk Factor Extractor, the Risk Factor Normalizer, and the Note Summarizer. The overall model framework is outlined in Figure 1. All LLM agents were implemented using Qwen3-14B²⁹ with thinking mode enabled. The extractor extracts clinical and SDOH variables; the normalizer and labeler standardize heterogeneous SDOH expressions into LLM-derived categories; and the summarizer produces clinician-style abstracts for downstream modeling.

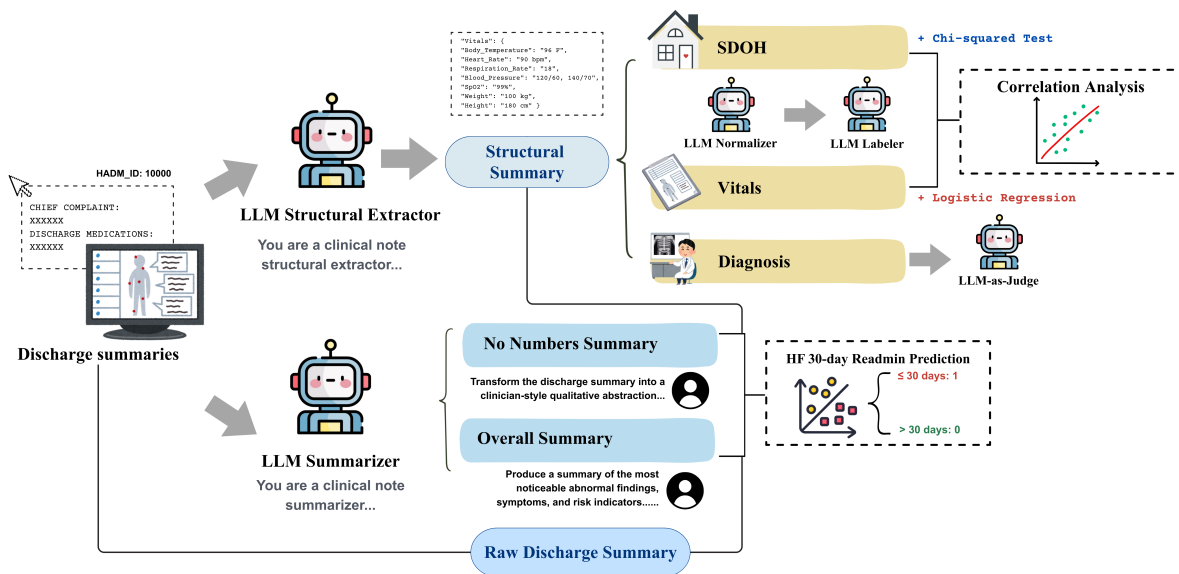


Figure 1: Overview of the ClinNoteAgents system for comprehensive clinical note analytics.

Risk factor extractor: We designed domain-specific LLM prompts to extract structured information from discharge summaries across three categories: (1) charted SDOH—gender, age, primary language, and marital status; (2) uncharted SDOH—alcohol, tobacco, and drug use, transportation, housing, employment, parental status, and social support; and (3) clinical measurements—vital signs (temperature, heart rate, respiratory rate, oxygen saturation, height, weight, blood pressure), chief complaint, and diagnoses.

Prompt example for LLM structural extractor

You are a clinical expert in extracting structured information from discharge summaries. Extract only the specific clinical factors that are important for predicting 30-day readmission in heart failure (HF) patients.
Output Schema Example:

```
{
  "Charted_SDOHs": {"Gender": "M", "Age": "50", "Language": "null", "Marital_Status": "Married"},
  "NonCharted_SDOHs": {"Alcohol_use": "No", "Tobacco_use": "1 ppd x 35y, quit 3 months ago", "Drug_use":
    ↪ "No", "Transportation": "null", "Housing": "null", "Parental": "null", "Employment_Status": "Retired",
    ↪ "Social_Support": "null"},
  "Clinical_Info": {
```

```
"Vitals": {"Body_Temperature": "96 F", "Heart_Rate": "90 bpm", "Respiration_Rate": "18", "Blood_Pressure":  
→ "120/60", "SpO2": "99%", "Weight": "100 kg", "Height": "180 cm"}},  
"Chief_Complaint": {"Symptoms": "...", "Description": "..."},  
"Diagnoses": [{"Condition": "Renal cell cancer", "Details": "..."}, ...]  
}
```

Risk factor normalizer. Both charted and uncharted SDOH were often documented in heterogeneous free-text, limiting their use in quantitative analysis. To address this, we developed an LLM-based normalization module that standardized these entries into categorical values through a two-stage process: the first LLM (the normalizer) generated a concise set of normalized categories for each variable, and the second LLM (the labeler) assigned each extracted value to one of these categories. For charted SDOHs, only *language* and *marital status* required normalization, as *gender* was already standardized (male and female). All uncharted SDOHs underwent the same normalization pipeline. Results were compared to existing reference taxonomies (MIMIC-SBDH³⁰ and LLM-SDOH²¹) using predefined categories.

Note summarizer. To assess whether LLM-based summarization improves HF readmission prediction, we implemented and compared two summarization methods: (1) an overall summary, which condenses the discharge note into a narrative emphasizing the most relevant clinical findings and risk indicators. Note that an overall summary can include numerical values, which contrasts to (2) a no-number summary, where all numerical values are replaced with qualitative descriptors to mitigate LLM instability in interpreting raw numeric data.

Prompt example (overall summary)

You are a clinical expert in abstracting information from EHR discharge summaries. Given the discharge summary, produce a summary of the most noticeable abnormal findings, symptoms, and risk indicators that could be related to 30-day hospital readmission for heart failure (HF) patients.

Prompt example (no-number summary)

[Same as overall summary above]. Transform the discharge summary into a clinician-style qualitative abstraction, preserving section headers, while removing numbers and converting them to qualitative descriptors, focused on the potential 30-day readmission risk for HF patients.

RESULTS

Evaluation on HF Readmission Factors Mining

We evaluated the risk-factor mining pipeline across extraction, normalization, and correlation analysis. Risk-factor mining transforms each discharge note into structured clinical and social variables and assesses their correlations with HF readmission. We first assessed extraction fidelity for clinical variables and charted SDOH using structured EHR as a surrogate ground truth, and applied an LLM-as-a-judge framework³¹ for diagnoses. Next, we evaluated the LLM-based normalization module, which standardizes heterogeneous SDOH expressions into analyzable categories. Finally, we conducted statistical association analyses to identify risk factors significantly correlated with HF readmission.

Structured EHR as surrogate ground truth. Clinical measurements (vitals) and charted SDOH were evaluated against corresponding EHR tables to assess extraction coverage and conditional accuracy. Coverage, defined as the proportion of patients with non-null LLM-extracted values (“% Extracted” in Table 2), ranged from 4.03% for *Height* to 89.25% for *Heart Rate*. Among SDOHs, *Gender* achieved the highest coverage (99.18%), whereas *Language* and *Marital Status* were less frequently detected (6.07% and 25.73%, respectively). For each non-null extraction, values were compared with ground truth within variable-specific tolerance ranges (“Tolerance Range” in Table 2) to compute conditional accuracy (“Cond Acc” in Table 2). Tolerance ranges were designed to account for variability in clinical documentation, including rounding differences, unit conversions, or discrepancies between EHR and discharge notes. Mixed-unit variables, namely *Temperature*, *Height*, and *Weight*, were compared in their native units when aligned. When discrepancies occurred, both LLM-extracted and structured EHR values were converted to canonical units (°C,

cm, and kg) before comparison. Other vitals were consistently recorded and did not require unit harmonization. For each vital within an admission, extracted and ground-truth values were converted to canonical units, aggregated, and summarized using medians. Mean absolute error and mean absolute percentage error (“MAE” and “MAPE” in Table 2) were computed without tolerance adjustments to reflect raw deviation magnitudes.

Table 2: Agreement with structured EHR ground truth for extracted clinical variables and charted SDOH.

| Variable | % Extracted | Tolerance Range | Cond Acc | MAE | MAPE |
|---------------------|-------------|--|----------|-------|--------|
| <i>Vitals</i> | | | | | |
| Temperature | 77.14% | $\pm 0.5^\circ\text{F}, \pm 0.3^\circ\text{C}$ | 84.24% | 1.40 | 1.42% |
| HR | 89.25% | ± 5 bpm | 88.59% | 11.42 | 13.74% |
| RR | 81.04% | ± 1 breath/min | 93.77% | 3.94 | 19.14% |
| SPO2 | 85.92% | $\pm 1\%$ | 94.16% | 2.77 | 2.93% |
| Height | 4.03% | ± 2 cm, ± 1 inch | 68.50% | 2.91 | 1.73% |
| Weight | 15.74% | ± 2 kg, ± 5 lbs | 57.90% | 5.48 | 7.03% |
| BP_SYS | 88.83% | ± 5 mmHg | 90.85% | 15.94 | 13.24% |
| BP_DIA | 88.83% | ± 5 mmHg | 91.24% | 11.55 | 19.99% |
| <i>Charted SDOH</i> | | | | | |
| Gender | 99.18% | – | 99.94% | – | – |
| Age | 89.31% | – | 93.38% | – | – |
| Language | 6.07% | – | 88.89% | – | – |
| Marital_Status | 25.73% | – | 77.89% | – | – |

Evaluation via LLM-as-a-judge. Diagnosis extraction was evaluated using an LLM-as-a-judge framework,³¹ comparing LLM-extracted diagnoses with associated ICD codes. For each patient, the judge assigned a score from 0 (lowest) to 5 (highest) based on the semantic similarity between the extracted diagnoses and the ICD-9 codes. We computed two metrics: conditional accuracy, the proportion of correctly identified diagnoses among those extracted by the LLM, and absolute accuracy, the proportion of all ICD-9 diagnoses correctly recovered. As summarized in Table 3, the LLM extracted fewer diagnoses per patient than the structured ICD-9 list. The average similarity score was 3.04, and the conditional accuracy was 62.27%.

Table 3: LLM-as-a-judge evaluation of diagnosis extraction.

| Avg. # ICD-9 | Avg. # LLM-extracted | Mean Score | Median Score | Cond Acc | Abs Acc |
|--------------|----------------------|------------|--------------|----------|---------|
| 15.14 | 5.91 | 3.04 | 3.00 | 62.67% | 25.25% |

Normalization results of clinical variables. LLM-based normalization agent was applied to reduce the high variability in free-text entries from SDOH extractions, thereby enabling direct correlation analyses. We first used k-medoids clustering to group semantically similar entries, selecting $k = 200$ to balance coverage and granularity ($k = 300$ produced overly fragmented clusters). Cluster medoids were then provided to an LLM to generate standardized category labels and descriptions, which a second LLM used to label each free-text entry as one of the normalized categories. The LLM-normalized categories of selected clinical variables are presented in Table 4.

Correlation analysis. We evaluated associations across variable types, using logistic regression for numerical variables (vitals and age) and chi-square tests for categorical variables (charted and uncharted SDOH). Results for both analyses are reported in Table 5 and Table 6. Logistic regression identified three statistically significant variables: age, weight, and blood pressure. Age and BP are positively associated with HF readmission risk, whereas weight showed a negative association. The chi-square test showed housing as the only statistically significant SDOH variable.

Evaluation of LLM-based Note Summarization

We assessed whether transforming discharge notes into structured, clinician-style abstracts improves downstream readmission prediction. The summaries were generated using Qwen3-14B, and the prediction performance was evaluated using three classifiers: TF-LDF with logistic regression (LR), ClinicalBERT, and a LoRA-finetuned Qwen3-8B. The

Table 4: LLM-normalized charted and uncharted SDOH categories.

| Variable | Categories |
|-------------------|---|
| Marital Status | Married; Widowed; Divorced/Separated; Single/Never Married; Unknown/Other |
| Alcohol Use | Abstinent/No Use; Current Heavy Use; Current Moderate/Social Use; Former Heavy Use; Former Moderate Use; Occasional/Rare Use; Unknown/Other; Past Use, Not Current |
| Tobacco Use | Never Smoker; Current Smoker; Former Smoker; Remote Tobacco Use; Occasional/Intermittent Use; Past Tobacco Use; High Pack-Year History; Unknown/Other |
| Transportation | Self-Driven; Non-Driver; Primary Transportation Method; Multiple Transportation Aids; Arranged Transportation; Assisted by Companion; Transportation Limitations; Unknown/Other |
| Housing | Living Alone; Living with Family Members; Institutional/Long-Term Care; Homelessness/Sheltered Living; Senior Housing/Retirement Communities; Residential Housing Type; Living with Non-Family Members; Home with 24/7 Care Services; Housing Instability/Unsafe Environment; Unknown/Other |
| Employment Status | Retired; Employed (Full-Time); Employed (Part-Time); Unemployed; On Disability; Self-Employed/Own Business; Student/Other Education; Unknown/Other |
| Social Support | Family Caregivers; Professional Caregivers; Social/Emotional Support; Living Arrangements; Lack of Social Support; Mixed Support Systems; Community/Non-Family Resources; Unknown/Other |

Table 5: Logistic regression correlation analysis for LLM-extracted clinical risk factors.

| Variable | Coef | p-value | OR | OR 95% CI |
|------------------|--------|------------------|-------|----------------|
| Temperature | -0.019 | 0.650 | 0.981 | (0.904, 1.065) |
| HR | 0.035 | 0.380 | 1.035 | (0.958, 1.118) |
| RR | 0.047 | 0.239 | 1.048 | (0.969, 1.134) |
| SpO ₂ | 0.006 | 0.871 | 1.006 | (0.933, 1.085) |
| Height | 0.244 | 0.198 | 1.276 | (0.881, 1.848) |
| Weight | -0.251 | 0.010 | 0.778 | (0.643, 0.942) |
| BP_SYS | -0.204 | <0.001 | 0.816 | (0.738, 0.902) |
| BP_DIA | -0.085 | 0.037 | 0.918 | (0.847, 0.995) |
| Age | 0.001 | 0.008 | 1.008 | (1.002, 1.015) |

Table 6: Chi-square test results for LLM-extracted charted and uncharted SDOH.

| Variable | N | Unique Values | Normalized Levels | Chi ² | p-value |
|-----------------------|------|---------------|-------------------|------------------|--------------|
| <i>Charted SDOH</i> | | | | | |
| Gender | 3515 | 2 | 2 | 2.62 | 0.106 |
| Language | 214 | 41 | 5 | 3.50 | 0.478 |
| Marital Status | 912 | 44 | 5 | 2.88 | 0.577 |
| <i>Uncharted SDOH</i> | | | | | |
| Alcohol use | 1930 | 851 | 8 | 6.94 | 0.435 |
| Tobacco use | 2487 | 1517 | 8 | 10.28 | 0.173 |
| Drug use | 864 | 315 | 5 | 3.56 | 0.468 |
| Transportation | 37 | 36 | 8 | 9.08 | 0.247 |
| Housing | 1031 | 554 | 10 | 21.13 | 0.012 |
| Parental | 278 | 240 | 6 | 5.44 | 0.365 |
| Employment | 1162 | 610 | 8 | 8.32 | 0.305 |
| Social support | 954 | 769 | 8 | 6.90 | 0.439 |

classification results are summarized in Figure 2. Across all models, raw discharge notes achieved the highest AUROC (LR: 0.6535; ClinicalBERT: 0.6095; LoRA: 0.6064). Performance declines after summarization were moderate despite severe text reduction. Among the summarization methods, no-number summary (61.36% word reduction) performed the best, with AUROCs remaining close to the raw baseline (LR: 0.6434; ClinicalBERT: 0.6046; LoRA: 0.5634). The overall summary (83.49% word reduction) showed a larger decrease (LR: 0.5866; ClinicalBERT: 0.5986; LoRA: 0.5588) but remained competitive with ClinicalBERT. The structural extraction summary (91.44% word reduction) produced the largest performance drop (LR: 0.5735; ClinicalBERT: 0.5708; LoRA: 0.5595).

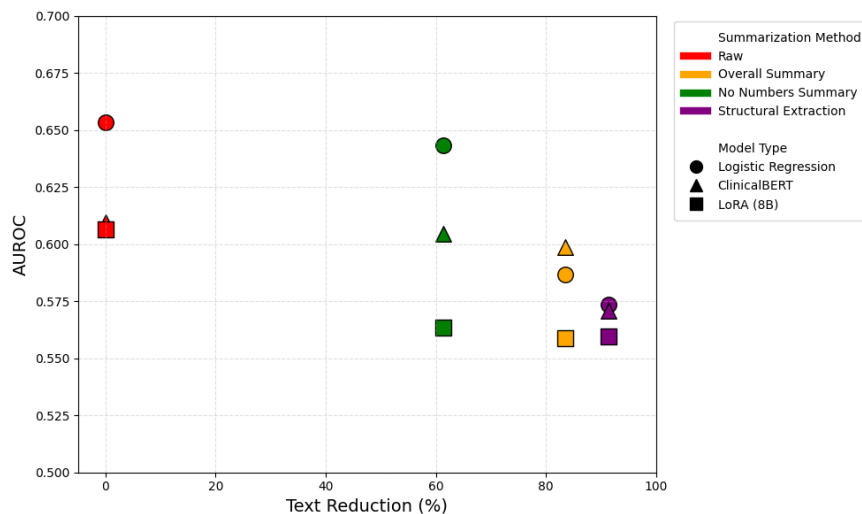


Figure 2: HF 30-day readmission classification performance across summarization methods and model types.

DISCUSSION

LLM Risk Factor Extractor. Accuracy within tolerance was high across most vitals. Conditional accuracy was highest for cardiorespiratory measures—*SpO₂* (94.16%), *Respiratory Rate* (93.77%), and *Blood Pressure* ($\geq 90\%$)—whereas anthropometric variables such as *Height* (68.50%) and *Weight* (57.90%) performed noticeably worse. These reductions were likely caused by heterogeneous reporting formats in discharge notes. Height may be recorded in centimeters or feet/inches, and weight in kilograms or pounds. Such variations may increase unit-conversion errors. Because LLMs exhibit unstable quantitative reasoning, and units in free-text form cannot be reliably validated, LLM-driven unit conversion was minimized during extraction. Nevertheless, values in uncommon units may still be misinterpreted. These patterns are consistent with evidence that LLMs hallucinate unsupported or incorrect clinical details during information extraction from discharge notes.³² MAPE remained below 15% for most vitals, indicating close agreement with the EHR surrogate ground truth. Small deviations occurred for *Temperature* (1.40%) and *SpO₂* (2.77%), while more variable measures such as *Blood Pressure* (13.24%-19.99%) and *Respiration Rate* (19.14%) showed larger discrepancies, consistent with their expected within-admission fluctuation. Several vitals and SDOHs showed low extraction coverage, largely due to under-documentation in discharge notes rather than extraction failure. Association analyses were therefore conducted only on patients with non-null values for each variable. While this approach may introduce selection effects, the findings should be interpreted as exploratory signals rather than population-level estimates. The LLM-as-a-judge evaluation showed that the model extracted fewer diagnoses per patient than the ICD-9 documentation. This is consistent with prior works that showed discrepancies between ICD coding and discharge notes.³³ Despite the difference, the LLM achieved a moderately strong similarity score (mean 3.04) and a conditional accuracy of 62.67%. This suggests that the extracted diagnoses were often semantically aligned with the ICD codes. These findings indicate LLM-based diagnosis extraction is feasible for scaling chart reviews, though discrepancies between clinical notes and coding systems warrant caution when using extracted diagnoses for downstream modeling.

LLM Normalization and Correlation Analysis. The LLM-based normalization agent enabled correlation analysis that was otherwise infeasible on raw extractions. Prior works typically map SDOHs onto predefined taxonomies such as

MIMIC-SBDH³⁰ and LLM-SDoH.²¹ However, these schemes are coarse and often collapse clinically distinct expressions into a small set of categories (e.g., binary community factors and limited alcohol/tobacco categories). This limits cohort-specific nuance and may attenuate statistical associations. In contrast, the LLM-generated categories provided concise and cohort-specific representations while preserving relevant contextual detail. Logistic regression identified *age*, *weight*, and *blood pressure* (BP) as significantly associated with HF readmission. Age showed a positive association, consistent with established epidemiology in which older patients are at higher risk of readmission.²⁷ Systolic and diastolic BP exhibited negative associations, aligning with reports that low BP is associated with acute heart failure.³⁴ Lower discharge weight may similarly reflect HF risk. The observed correlations underscore the potential of LLM-based extraction to support early identification and monitoring of high-risk patients. Other vital signs were not significant, suggesting that their discharge-time values provide limited discriminative value for readmission. Chi-square analysis found housing to be the only significant SDOH. Other social factors commonly reported as influential, including smoking, social support, and marital status,³⁵ were not significant in our cohort. These discrepancies reflect the under-documentation of SDOH in discharge notes or limitations in LLM extraction, highlighting the challenge of using discharge notes alone to capture SDOH.

30-Day Readmission Predication using LLM Summaries. LLM-generated summaries preserved most predictive signal needed for HF readmission modeling. Although summarization did not improve classification performance as initially expected, performance declines were modest across models, even with 60–90% text reduction. This suggests the LLM preserved the majority of risk-relevant information, such as comorbidities, discharge stability, and key physiological measurements, while removing redundant content and noise. Substantial compression reduces token requirements for transformer models and produces more concise inputs for downstream review. Our results demonstrate that LLM summarization offers a balance between information preservation and computational efficiency, enabling scalable representation of clinical text while maintaining clinically meaningful signals for risk prediction.

Model Choice in ClinNoteAgents. We selected LLM agents over pre-trained language models (PLMs) such as ClinicalBERT because our note-to-structure tasks (extraction, normalization, and summarization) require generative reasoning and minimal supervision. LLMs are known to extract clinical entities and numeric data from unstructured clinical notes with high fidelity in zero- and few-shot settings,^{36,37} enabling direct extraction of heterogeneous risk factors from discharge notes. LLMs also support SDOH extraction²¹ and, unlike PLMs, can perform normalization without requiring task-specific pretraining.²⁶ Moreover, LLMs provide substantially longer context windows than PLMs and is able to achieve summarization quality approaching expert benchmarks.³⁸ We employed Qwen3 across agents to balance computing efficiency and performance, given its strong results in clinical information extraction benchmarks.³²

Limitations. One limitation of our study is the indirect evaluation of LLM-extracted risk factors. For SDOH variables and most vital signs, we relied on structured EHR fields as surrogate ground truth and assessed extraction quality using tolerance-based accuracy and an LLM-as-judge approach. Another limitation is that we did not observe substantial performance gains from summarization alone. Finally, we did not include clinician-led evaluations of LLM extraction or summarization results, which limits formal assessment of clinical fidelity and real-world applicability.

Ethical Considerations. This study involves secondary analysis of electronic health records and discharge summaries, all of which were fully de-identified in accordance with HIPAA standards. Access to the data required completion of the mandated training and certification, and all analyses were conducted under approved data-use conditions. Because de-identified clinical text may still contain sensitive contextual information, all large language models used in this work were deployed locally within secure computing environments. No patient data were sent to external, public, or third-party services. Although LLM demonstrated reliable extraction performance, it remains prone to omissions and hallucinations. Inaccurate outputs could disrupt diagnostic reasoning, risk stratification, or care planning. Thus, LLM-based clinical systems should be used as decision-support tools, not standalone sources of truth. Responsible deployment requires rigorous validation and clear guardrails to ensure safe and appropriate use in clinical workflows.

CONCLUSION

The study introduced ClinNoteAgents, an LLM-based multi-agent framework that operationalizes two core tasks: (1) HF readmission risk-factor mining and (2) HF readmission risk prediction. By transforming unstructured discharge notes into structured representations of clinical measurements, social determinants of health, and diagnoses,

the system provides an end-to-end pipeline that directly supports predictive modeling and risk-factor analysis. For the readmission-factor mining task, the coordinated extraction and normalization agents recovered clinically relevant information with high fidelity, standardized heterogeneous medical term expressions, and enabled downstream statistical association analysis that would otherwise have been infeasible on raw free-text. These results demonstrate the feasibility of using LLMs to generate analyzable structural representations in limited structured EHR environments. For the readmission prediction task, abstractions produced by the LLM summarizer preserved most predictive signals from raw discharge notes despite substantial text compression. This indicates that the summarizer can distill long and noisy clinical narratives into concise representations that reduce computational cost while retaining clinical utility. Taken together, ClinNoteAgents provides a unified framework that enables scalable extraction, normalization, and summarization of free-text documentation to support HF readmission risk modeling. The system produces structured, interpretable abstractions compatible with rule-based pipelines, temporal reasoning modules, and clinical decision-support tools. In a clinical decision-support setting, ClinNoteAgents could be integrated into the EHR to generate readmission risk predictions alongside extracted risk factors and structured note summaries. Outputs would remain solely for advisory purposes and require clinician verification to mitigate hallucination and omission risks.

ACKNOWLEDGMENT

This research was partially supported by the internal funds and GPU servers provided by School of Nursing's Center for Data Science of Emory University, Computer Science Department of Emory University, the US National Science Foundation under Award Numbers 2442172, 2312502, 2319449, and the US National Institute of Health under Award Numbers K25DK135913, RF1NS139325, R01DK143456 and U18DP006922.

References

1. Yan S. Global, Regional, and National Burden of Heart Failure and Its Risk Factors between 1990 and 2021 and Projections to 2050: An Analysis of the Global Burden of Disease Study; 2021. Preprint.
2. Khan MS, Sreenivasan J, Lateef N, et al. Trends in 30- and 90-Day Readmission Rates for Heart Failure. *Circulation: Heart Failure*. 2021;14(4):e008335.
3. Eltelbany M, Chan S, Gottlieb S. Specific Causes of 30-Day and 1-Year Readmissions in Heart Failure Patients. *Journal of Cardiac Failure*. 2019;25(8 Suppl):S131.
4. Jain A, Arora S, Patel V, Raval M, Modi K, Arora N, et al. Etiologies and Predictors of 30-Day Readmission in Heart Failure: An Updated Analysis. *International Journal of Heart Failure*. 2023;5(3):159-68.
5. Reddy YNV, Borlaug BA. Readmissions in Heart Failure: It's More Than Just the Medicine. *Mayo Clinic Proceedings*. 2019;94(10):1919-21.
6. Umehara T, Katayama N, Tsunematsu M, Kakehashi M. Factors affecting hospital readmission heart failure patients in Japan: a multicenter retrospective cohort study. *Heart and Vessels*. 2020;35(3):367-75.
7. Mahajan SM, Burman P, Newton A, Heidenreich PA. A validated risk model for 30-day readmission for heart failure. In: *MEDINFO 2017: Precision Healthcare through Informatics*. IOS Press; 2017. p. 506-10.
8. Yu MY, Son YJ. Machine learning-based 30-day readmission prediction models for patients with heart failure: a systematic review. *European Journal of Cardiovascular Nursing*. 2024;23(7):711-9.
9. Murray CJL. The Global Burden of Disease Study at 30 Years. *Nature Medicine*. 2022;28(10):2019-26.
10. Bostan S, Johnson OA, Jaspersen LJ, Randell R. Contextual Barriers to Implementing Open-Source Electronic Health Record Systems for Low- and Lower-Middle-Income Countries: Scoping Review. *Journal of Medical Internet Research*. 2024;26:e45242.
11. Hossain MK, Sutanto J, Handayani PW, et al. An exploratory study of electronic medical record implementation and recordkeeping culture: the case of hospitals in Indonesia. *BMC Health Services Research*. 2025;25:249.
12. Taher A, Shimul MMH, Khan S, et al. Adoption challenges of digital transformation of human resource management in Bangladesh's healthcare system: a cross-sectional mixed-methods evaluation. *BMC Health Services Research*. 2025;25:1383.
13. Alzghaibi H, Hutchings HA. Barriers to the implementation of large-scale electronic health record systems in primary healthcare centers: a mixed-methods study in Saudi Arabia. *Frontiers in Medicine*. 2025;12:1516714.
14. Mensah NK, Adzakpah G, Kissi J, et al. Health professionals' perceptions of electronic health records system: a

- mixed method study in Ghana. *BMC Medical Informatics and Decision Making*. 2024;24:254.
15. Kong HJ. Managing unstructured big data in healthcare system. *Healthcare Informatics Research*. 2019;25:1-2.
 16. Frizzell JD, Liang L, Schulte PJ, Yancy CW, Heidenreich PA, Hernandez AF, et al. Prediction of 30-Day All-Cause Readmissions in Patients Hospitalized for Heart Failure: Comparison of Machine Learning and Other Statistical Approaches. *JAMA Cardiology*. 2017;2(2):204-9.
 17. Pishgar M, Theis J, Del Rios M, Ardatti A, Anahideh H, Darabi H. Prediction of unplanned 30-day readmission for ICU patients with heart failure. *BMC Medical Informatics and Decision Making*. 2022;22(1):117.
 18. Liu X, Chen Y, Bae J, Li H, Johnston J, Sanger T. Predicting Heart Failure Readmission from Clinical Notes Using Deep Learning. *arXiv*. 2019.
 19. Golas SB, Shibahara T, Agboola S, Otaki H, Sato J, Nakae T, et al. A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data. *BMC Medical Informatics and Decision Making*. 2018;18(1):44.
 20. World Health Organization. Social Determinants of Health; 2025. Accessed: 2025-10-12. https://www.who.int/health-topics/social-determinants-of-health#tab=tab_1.
 21. Guevara M, Chen S, Thomas S, Chaunzwa TL, Franco I, Kann BH, et al. Large language models to identify social determinants of health in electronic health records. *NPJ digital medicine*. 2024;7(1):6.
 22. Alnomasy N, Pangket P, Mostoles R, et al. Predictive Performance of Machine Learning Models for Heart Failure Readmission: A Systematic Review. *Biomedicines*. 2025;13(9):2111.
 23. Huang K, Altosaar J, Ranganath R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission; 2020. *arXiv preprint arXiv:1904.05342*.
 24. Boll HO, Boll AO, Boll LP, Abu Hanna A, Calixto I. DistillNote: LLM-Based Clinical Note Summaries Improve Heart Failure Diagnosis; 2025. *arXiv preprint arXiv:2506.16777*.
 25. Consoli B, Wang H, Wu X, et al. SDoH-GPT: Using Large Language Models to Extract Social Determinants of Health. *Journal of the American Medical Informatics Association*. 2025.
 26. Gu Z, He L, Naeem A, et al. SBDH-Reader: An LLM-Powered Method for Extracting Social and Behavioral Determinants of Health from Clinical Notes; 2025. Preprint.
 27. Shao M, Kang Y, Hu X, Kwak HG, Yang C, Lu J. Mining Social Determinants of Health for Heart Failure Patient 30-Day Readmission via Large Language Model. In: *Studies in Health Technology and Informatics*; 2025. .
 28. Johnson AE, Pollard TJ, Shen L, Lehman LwH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. *Scientific data*. 2016;3(1):1-9.
 29. Team Q. Qwen3 technical report. *arXiv preprint arXiv:250509388*. 2025.
 30. Ahsan H, Ohnuki E, Mitra A, You H. MIMIC-SBDH: a dataset for social and behavioral determinants of health. In: *Machine Learning for Healthcare Conference*. PMLR; 2021. p. 391-413.
 31. Face H. LLM Judge: Building Automatic Evaluators with Large Language Models; 2024. https://huggingface.co/learn/cookbook/en/llm_judge.
 32. Das AB, Ahmed S, Sakib SK. Hallucinations and Key Information Extraction in Medical Texts: A Comprehensive Assessment of Open-Source Large Language Models. In: *IEEE EMBS BHI 2025*; 2025. .
 33. Kim HN, Lee M, Lee JH, Lee S, Lee J. Diagnostic Accuracy of ICD Codes Versus Discharge Summary Text. *Healthcare*. 2021;9(12):1714.
 34. Kim HJ, Jo SH. Effect of low blood pressure on prognosis of acute heart failure. *Scientific Reports*. 2024;14:15605.
 35. Calvillo-King L, Arnold D, Eubank KJ, Lo M, Yunyongying P, Stieglitz H, et al. Impact of Social Factors on Hospital Readmissions: A Systematic Review. *Annals of Internal Medicine*. 2013;159(5):327-38.
 36. Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D. Large Language Models Are Few-Shot Clinical Information Extractors. In: Goldberg Y, Kozareva Z, Zhang Y, editors. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics; 2022. p. 1998-2022.
 37. Adam H, Lin J, Ghassemi M. Clinical Information Extraction with Large Language Models: A Case Study on Organ Procurement. *Journal of the American Medical Informatics Association*. 2024.
 38. Van Veen D, Van Uden C, Blankemeier L, Delbrouck JB, Aali A, Bluethgen C, et al. Adapted Large Language Models Can Outperform Medical Experts in Clinical Text Summarization. *Nature Medicine*. 2024;30(4):1134-42.