Asking the Right Questions: Adapting LLMs to Analyze Clinical Notes from Multiple Care-Domains

Anonymous ACL submission

Abstract

Clinical notes contain fine-grained and domain-specific information that can significantly improve patient risk estimation. However, these notes differ widely across care-domains such as nursing, physician, and radiology as each subscribe to a particular viewpoint when documenting patient's status. Although Large Language Models (LLMs) can reason over long clinical narratives, their performance depends heavily on prompting, and fixed or manually crafted prompts often fail to reflect the linguistic and semantic variations across note types. Our empirical analysis shows that notes from distinct care-domains exhibit large differences in topic distributions, underscoring the need for viewpoint-aware modeling. To tackle this problem, we propose an end-to-end framework that learns optimal guiding questions for each viewpoint, enabling LLMs to extract clinically meaningful and interpretable risk factors tailored to each. The guiding questions are optimized using only supervision from downstream prediction tasks, without any instruction tuning of the base LLM. Across two real-world EHR datasets and three prediction tasks, our framework outperforms domain-agnostic prompting and demonstrates that viewpoint-specific guiding questions are crucial for accurate and explainable patient risk estimation.

1 Introduction

Clinical notes in Electronic Health Records (EHRs) contain rich, unstructured information that is crucial for patient risk prediction and clinical decision-making. Unlike structured data such as laboratory values or medical codes, these narratives capture nuanced patient trajectories, provider observations, clinical reasoning, and interventions (Tayefi et al., 2021; Kwon et al., 2024). They also reflect the temporal evolution of patient states across care-domains (e.g., physicians, nurses, pharmacists), providing context often missing from coded data. Prior work (Huang et al., 2021; Song et al., 2022; Wang et al., 2024; Choudhuri et al., 2023; Wang et al., 2023) demonstrates that clinical notes alone or combined with structured EHR data substantially improves clinical outcome prediction.

Note Type	Note Type	JS Distance
Nursing	Physician	0.989859
Nursing	Radiology	0.967678
Physician	Radiology	1.0

Nursing Notes Topics
- action, response, plan, tracheal, hr
- evaluation, stent, found, plan, transferred
- placed, episode, plan, hospital, vein

Physician Notes Topics
- icu, present, pulse,assessed,meq,cmh2o
- fentanyl, heparin, hour, continue

Radiology Notes Topics
- chest, clip, portable, atelectasis, interval
- contrast, ct, lower lobe, evidence, right
- reason, pfi report, vein, dvt, peg

Figure 1: **Analysis of Notes from SUBJECT ID 55672 in MIMIC-III Dataset**: Jensen–Shannon (JS) distances between topic distributions of note care-domains (top), where higher values indicate greater dissimilarity. The most salient topics for each viewpoints are shown at the bottom.

However, clinical notes originating from different care-domains vary substantially in content, semantics, and focus. Physician notes typically document disease progression and treatment rationale, nursing notes emphasize continuous monitoring, care planning, and patient responses, while radiology notes focus on imaging findings and anatomical evidence. Although leveraging such rich textual information has improved downstream predictive modeling, heterogeneous writing styles, variable content quality, and *viewpoint* specific emphasis pose significant challenges for automated analysis, motivating robust methods for extracting clinically relevant knowledge.

To quantify these differences, we analyzed all clinical notes from five patients in MIMIC-III (see Appendix A.1 for details). Figure 1 presents results for SUBJECT ID 55672, the patient with the largest number of notes. The high JS distances indicate that note types are highly dissimilar. Nursing notes focus on care planning and management (*action, plan, assessment, evaluation*) and patient state transitions, physician notes emphasize ICU-level clinical management (*icu, pulse, fentanyl, cmh2o*), and radiology notes center on imaging findings (*ct, contrast, atelectasis*), thoracic anatomy (*chest, lower lobe*), and thrombotic conditions (*vein, dvt*). These patterns are consistent across the remaining patients (Appendix A.1). Overall, these findings underscore the need to model clinical notes in a viewpoint aware manner. Ignoring viewpoint specific differences can conflate heterogeneous signals, weakening representations and degrading downstream patient risk prediction.

Recent work has leveraged Large Language Models (LLMs) to extract information from clinical notes, driven by their ability to process long contexts more effectively than prior models such as Clinical-Longformer (Li et al., 2022; Beltagy et al., 2020). Studies show that LLMs can summarize, reason over, and infer patient risk factors from unstructured clinical narratives (Choudhuri et al., 2024; Hager et al., 2024). However, LLMs are highly sensitive to prompts (Zhao et al., 2021; Webson and Pavlick, 2022), and their clinical performance depends on how well they are guided toward task-relevant medical concepts. Fixed or manually designed prompts often fail to capture viewpoint specific nuances across physician, nursing, and radiology notes (Mashima et al., 2024; Liu et al., 2024b). This motivates methods that automatically learn optimal, viewpoint specific guiding questions, enabling transparent and clinically meaningful risk factor extraction.

To this end, we focus on explainable patient risk prediction from clinical notes by learning optimal guiding questions tailored to different healthcare care-domains. Our primary contributions are:

- We model care-domain specific nuances by treating guiding questions as learnable natural language parameters, explicitly leveraging multi-view progress notes rather than undifferentiated note pools or discharge summaries.
- We learn separate guiding questions for each viewpoint (originating from individual care domains) using an LLM-based optimizer, guided solely by downstream risk prediction supervision and without instruction-tuning the base LLM.
- We evaluate our approach on two real-world EHR datasets across three prediction tasks, demonstrating that viewpoint specific guiding questions significantly improve downstream patient risk prediction.

2 Related Works

Clinical notes for risk prediction, heterogeneity, and long-document handling: Clinical notes enhance risk prediction beyond structured EHR data by encoding rich clinical signals (Liu et al., 2018; Feller et al., 2018; Choudhuri et al., 2023; Brown et al., 2022; Garriga et al., 2023). However, existing techniques treat all notes from a visit as a single block of information. This narrows insights about how different provider types contribute to the model's predictions (Si and Roberts, 2021) and do not fully exploit the rich diversity inherent in clinical notes in style, vocabulary, and clinical focus (Wen et al., 2021). Notes also contain repetitive boilerplate text that can distort meaningful signals and make isolating clinically relevant information difficult (Liu et al., 2022).

Large language models for clinical reasoning and risk estimation: LLMs are increasingly used for clinical text understanding tasks such as summarization, question answering, and clinical decision support (Singhal et al., 2023a,b; Yang et al., 2022; Guevara et al., 2024). For risk prediction, common approaches either prompt an LLM to directly predict an outcome (Gu et al., 2024), or use prompts that elicit an explicit outcome probability, often alongside free-text clinical reasoning (Zhu et al., 2024; Small et al., 2025). However, fine-tuning may be limited by the data that they were trained on (Pingua et al., 2025), which leads to the recent use of Retrieval-Augmented Generation (RAG) methods (Sohn et al., 2025; Zakka et al., 2024; Jeong et al., 2024).

Prompt sensitivity and automatic prompt learning: Prompt methods are sensitive to how instructions are phrased, leading to research on optimizing prompts through discrete search, soft tuning, and iterative refinement (Ramnath et al., 2025; Deng et al., 2022; Lu et al., 2024; Chen et al., 2023). However, existing prompt tuning approaches often learn a single prompt for uniform inputs, which can lead to reduced performance in heterogeneous, multi-domain scenarios (Dun et al., 2025). Recent research has also focused on learning positive and negative prompts (Ahmed et al., 2024), meta-learning (Das et al., 2012) and reinforcement learning (Chen et al., 2025) in healthcare.

3 Method

Our proposed framework focuses specifically on learning-guiding questions for each viewpoint to query clinical notes. The architecture is described in Figure 2. We will first describe the predictive task, followed by the details of our overall framework in the subsequent subsections.

3.1 Problem Setup

We denote the EHR data to consist of a set of patients $\mathcal{P}$. Each patient $p \in \mathcal{P}$ has a corresponding set of hospital or critical care unit visits $\mathcal{V}_p = \{v_1, v_2, \ldots, v_{|\mathcal{V}_p|}\}$. Each visit $v_j \in \mathcal{V}_p$ contains a timestamped collection of multi-view clinical progress notes $N_j = \{n_1, n_2, \ldots, n_{|N_j|}\} \subset \mathcal{N}$, where $\mathcal{N}$ denotes the total set of clinical notes for patient p. Note that, contrary to most prior works, our work excludes discharge summaries and only deals with progress notes. To explicitly represent the viewpoint structure, we divide the notes in each visit into viewpoint specific subsets $N_j = N_j^{(s_1)} \cup N_j^{(s_2)} \cup \cdots N_j^{(s_{|D|})}$, where $N_j^{(s_i)}$ denotes the set of notes written by a member of viewpoint s_i. To reduce notational redundancies in the following sections, we only describe the approach using a single visit by a patient.

Given the set of multi-view clinical notes N_j for visit v_j for patient p, we aim to predict the patient's clinical outcome y_j. This is modeled as a binary/multi-class classification in our work. In our experiments y_j represents length of stay categories, ICU readmission and transfer. Akin to other supervised classification tasks, the overall data represented by note-label tuple $\{\{N_j, y_j\}_{\mathcal{V}_p}\}_{p \in \mathcal{P}}$ is randomly split into train, validation, and test sets.

3.2 Task-Specific Clinical Summaries Extraction

Clinical notes typically suffer from redundant and boilerplate information across daily progress entries, which can dilute the clinically salient signal for predictive modeling (Liu et al., 2022). With the limitation of context length in transformer-based models (Menasalvas and Gonzalo-Martin, 2016; Si and Roberts, 2021), we first extract concise, viewpoint specific representations from the unstructured text in $N_j^{(s_i)}$ for each viewpoint $s_i \in D$. All clinical notes have variable contributions towards different downstream tasks (Liu et al., 2022; Gao et al., 2023). To extract a subset of the clinical notes that help in predicting the particular downstream task label, we design a natural language task-specific prompt for task t $\text{prompt}_t = \beta(t)$. Here $\beta(\cdot)$ used in our experiments was 'GPT-4o'. This prompt, along with all the multi-view clinical note summaries $S_j^{(s_i)}$ for each viewpoint are converted to embeddings as given by $h_p = f(\text{prompt}_t)$ and $h_{S_j^{(s_i)}} = f(S_j^{(s_i)})$.

For each summary, we then construct a similarity score as $\alpha_j^{(s_i)} = \text{sim}(h_{S_j^{(s_i)}}, h_p)$, where $\text{sim}(\cdot)$ denotes a similarity function such as cosine similarity between the task prompt embedding h_p and the embedding of the summary $h_{S_j^{(s_i)}}$. $\alpha_j^{(d)}$ quantifies the alignment of each viewpoint specific summary with the task objective. We then select the k-most important notes for every viewpoint for each visit by choosing the notes with the top-k scores as $\tilde{N}_j^{(s_i)} = \text{Top-}k\big(N_j^{(s_i)}, \alpha_j^{(s_i)}\big)$, where $\tilde{N}_j^{(s_i)}$ denotes the subset of viewpoint s_i notes selected from $N_j^{(s_i)}$ that have the highest scores.

3.3 Learning Optimal Guiding Questions

Clinical notes authored by healthcare entities with varying viewpoints differ significantly semantically (Liu et al., 2024b; Mashima et al., 2024), which are not always interchangeable (Weng et al., 2017), leading to the need for viewpoint aware guidance. To integrate this idea in a computationally inexpensive manner, we try to enable LLMs to extract patient risk factors without any instruction tuning. We enable a medical LLM to learn explicit guiding questions to extract risk factors from clinical notes from each viewpoint inspired by a prior work (Ye et al., 2025) on video anomaly detection. For each viewpoint, we initialize a set of guiding natural language questions $Q_{s_i} = \{q_{s_i}^1, q_{s_i}^2, \cdots, q_{s_i}^m\}$, where $q_{s_i}^j$ is the j-th question $\{1 \le i \le m\}$ and m is the total number of questions. The overall framework considers the guiding questions Q_{s_i} as learnable parameters while the LLM itself is frozen. The training data for learning Q_{s_i} is $\tilde{N}_j^{(s_i)}$ with the corresponding true label being y_j.

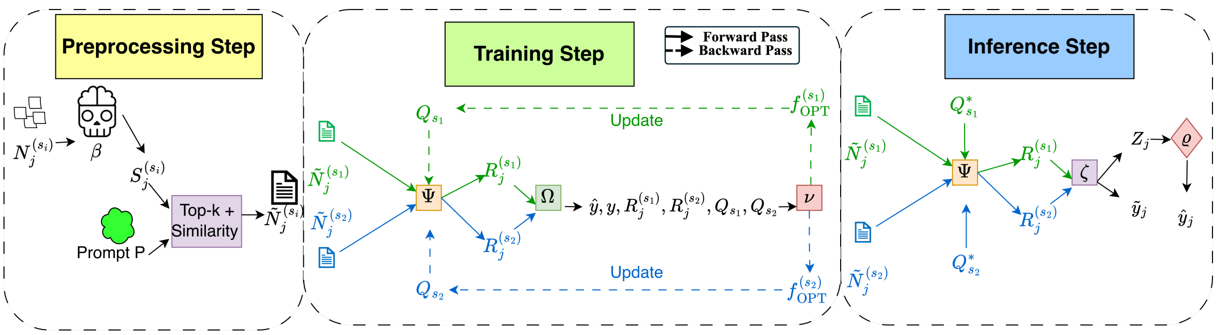

Figure 2: **Our Framework**: Our overall framework uses clinical notes (*Preprocessing Step*) by generating summaries from clinical notes and choosing top-k notes from each viewpoint. These notes are then used to learn guiding questions (*Training Step*) to query each important note summary to extract the risk factors from notes using the viewpoint-specific Learner $\boldsymbol{\Psi}$ and Aggregate Learner $\boldsymbol{\Omega}$, and optimized by the Optimizer ν. Finally, for test data, the viewpoint-specific Learner $\boldsymbol{\Psi}$ generates risk factors, which are then fed to the Inference Predictor ζ, which produces soft predictions and reasons behind them. The reasons are finally fed to ϱ to get predictions. Note that the squares in the figure represent LLMs and the diamond is a predictive ML model.

Henceforth, we will use $f(x, \upsilon)$ to denote an LLM parametrized by the input data x and the instructions/prompts υ. The goal in our framework is to learn the best possible viewpoint specific υ for each x. Since υ is a free-text and x has a specific viewpoint, we define a **viewpoint-specific learner agent** $\Psi(\cdot, \cdot)$ for viewpoint-aware risk-factor extraction. To combine all viewpoints, we also define **aggregate learner agent** $\Omega(\cdot)$) and an **optimizer agent** (denoted by function $\nu(\cdot)$) analogous to numerical optimizers (Bilmes et al., 1997; J and Ba, 2014). Our overall training process is illustrated in Figure 2 can be broken down into three parts, as defined below.

Viewpoint-Specific Learner Agent: To effectively extract important patient risk factors from the clinical note summaries using viewpoint specific prompts, we feed the frozen LLM $\Psi(\cdot, \cdot)$ with the viewpoint specific clinical note summary $\tilde{N}_j^{(s_i)}$ and guiding questions Q_{s_i} to generate patient risk factors. The overall prompt template is shown in Figure 8 in the Appendix. As shown in the figure, we set the overall context of the model under 'Model Description'. The learned guiding questions for each viewpoint are input into the placeholder along with the corresponding top-k clinical note summaries for every patient visit. This generates the risk factors for the visit based on the notes for that particular viewpoint, given as $R_j^{(s_i)} = \Psi(\tilde{N}_j^{(s_i)}, Q_{s_i})$.

Aggregate Learner Agent: After extracting the viewpoint-specific risk factors from the clinical notes, the risk factors are concatenated as $R_j = (R_j^{(s_1)} \parallel R_j^{(s_2)} \parallel \ldots \parallel R_j^{(s_{|D|})})$ where $|D|$ represents cardinality of the unique viewpoints set. The aggregated risk factors are then sent to the Aggregate Learner to produce a classified task-specific label $\hat{y}_j$ as $\hat{y}_j = \Omega(R_j, P_\Omega)$. P_Ω denotes the prompt given to $\Omega(\cdot, \cdot)$ (example shown in Figure 9 in Appendix) to generate the predictive class label. The aggregate learner operates as a frozen LLM that performs reasoning over the concatenated textual risk factors from all viewpoints. In this way, $\Omega(\cdot, \cdot)$ acts as a meta-reasoner by bridging the semantic gap between viewpoint-specific observations and clinical decision outcomes by serving as the final decision stage. The learner executes the 'forward pass', where it predicts a label for the given input data.

Optimizer Agent: The optimizer agent, unlike the learner, executes a so-called 'backward pass', to update the viewpoint-specific questions Q_{s_i} via a minibatch. Given the viewpoint-specific risk factors R_j^{batch}, predicted outputs $\hat{y}_j^{\text{batch}}$, viewpoint-specific guiding questions Q_{s_i} generated for the batch along with corresponding task labels y_j^{batch}, the optimizer generates the set of new viewpoint-specific guiding questions $Q'_{s_i} = \nu(R_j^{\text{batch}}, \hat{y}_j^{\text{batch}}, y_i^{\text{batch}}, \{Q_{s_i}\}_{s_i \in D})$. The prompt template for the optimizer is shown in Figure 8 in the Appendix.

The updated guiding questions explicitly emphasize clinically salient concepts, suppress irrelevant information, or re-frame queries to elicit more discriminative evidence from the same notes. In this way, learning emerges as a feedback-driven process over free-text prompts, where the questions progressively specialize to the clinical priorities of each viewpoint. This enables the overall framework to learn what to ask from each viewpoint rather than relying on manually crafted heuristics or static instruction templates. Unlike instruction tuning, which requires GPU-intensive finetuning, the Optimizer Agent optimizes via prompt engineering in a zero-shot or few-shot manner. Compared to gradient-based methods, it handles non-differentiable text directly. This avoids data privacy issues associated with full model training, as only questions are updated while LLMs remain frozen.

3.4 Inference

We denote the viewpoint specific guiding questions that gave the highest accuracy for the validation set as $\{Q^*_{s_i}\}_{s_i \in D}$ or Q^* without loss of generality. During inference, given Q^*, our framework makes prediction $\hat{y}_{\text{test}}$ for the clinical notes N_{test} for patient visit v_{test} through the following steps:

Step 1: Summarize and Select Notes: Similar to the training step, we first summarize all the notes as shown in Section 3.2 to get $S^{(s_i)}_{\text{test}}$ along with the training and validation sets $S^{(s_i)}_{\text{train}}$ and $S^{(s_i)}_{\text{val}}$. Then select the top-k note summaries for each viewpoint and obtain $\{\tilde{N}^{(s_i)}_{\text{test}}\}$, $\{\tilde{N}^{(s_i)}_{\text{val}}\}$, and $\{\tilde{N}^{(s_i)}_{\text{train}}\}$.

Step 2: Generate Reasons from Notes: Given the note summaries $\{\tilde{N}^{(s_i)}_{\text{test}}\}$, $\{\tilde{N}^{(s_i)}_{\text{val}}\}$, and $\{\tilde{N}^{(s_i)}_{\text{train}}\}$ and optimal guiding questions Q^*, generate risk factors for each viewpoint, as shown in Section 3.3 by $\{R^{(s_i)}_{\text{test}}\} = \{\Psi(\tilde{N}^{(s_i)}_{\text{test}}, Q^*_{s_i})\}$, $\{R^{(s_i)}_{\text{val}}\} = \{\Psi(\tilde{N}^{(s_i)}_{\text{val}}, Q^*_{s_i})\}$ and $\{R^{(s_i)}_{\text{train}}\} = \{\Psi(\tilde{N}^{(s_i)}_{\text{train}}, Q^*_{s_i})\}$. Using these risk factors, we make LLM $\zeta(\cdot)$ make a soft label prediction along with the reasoning behind the prediction represented as $Z_{\text{test}}, \tilde{y}_{\text{test}} = \zeta(\{R^{(s_i)}_{\text{test}}\}, P_\zeta)$. Similarly, for training and validation datasets, repeat the process to get $Z_{\text{train}}, \tilde{y}_{\text{train}}$ and $Z_{\text{val}}, \tilde{y}_{\text{val}}$ Here Z_{data} is the reason and $\tilde{y}_{\text{data}}$ is the soft prediction while P_ζ is the prompt shown in Figure 9. As prior works have shown that LLMs struggle to make accurate point predictions (Hager et al., 2024; Choudhuri et al., 2024), we ignore the soft predictions and focus on the reasoning Z_{test} for the next step.

Step 3: Make Predictions: Given the reasons Z_{test}, Z_{val}, and Z_{train} obtained from the previous step, we use an encoder $\varrho(\cdot)$ to obtain embeddings as $(h_{\text{train}}, h_{\text{val}}, h_{\text{test}})$.

4 Experiments

In this section, we will discuss the results of our experimental evaluation on two patient risk estimation tasks on two real-world EHR datasets. Our experiments aim to answer the following research questions (**RQ**s), which are: **RQ1:** Does our overall framework improve the performance in downstream tasks? **RQ2:** Which viewpoint has a dominating effect over others in downstream tasks? **RQ3:** Does learning guiding questions for each viewpoint improve performance in downstream patient risk prediction tasks? **RQ4:** Do the learned questions improve the explainability of the overall risk prediction task?

To answer the above questions, we will first briefly describe the overall setting, followed by the baseline methods used in the experiments, and the results and discussions. The codes and supplementary materials are available for peer-review[1].

4.1 Datasets and Tasks

4.1.1 Datasets

In our experiments, we used two real-world EHR datasets. Further details about both the datasets are provided in A.4. The descriptions of both datasets are as follows:

- **MIMIC-III** (Johnson et al., 2016): We used the popularly used open-source MIMIC-III (Johnson et al., 2016) EHR dataset for our study. We only used information about the patients who were admitted to the Intensive Care Units (ICU) and stayed there for more than 2 days for each admission to the ICU.
- **PHOP**: PHOP is obtained from a large tertiary care teaching hospital in the U.S. midwestern region. Analysis of the dataset is permitted under institutional IRB.

4.1.2 Tasks

We evaluate our models on three predictive tasks:

- **Length of Stay (LOS) Category Prediction**: We formulate ICU LOS prediction as a multi-class classification task (Xian et al., 2023; Choudhuri et al., 2024), with LOS categorized as *low* (2–4 days), *medium* (4–7 days), and *high* ($\geq$ 7 days). Experiments are conducted on MIMIC-III and PHOP; LOS is predicted per ICU visit in MIMIC-III and per hospital visit in PHOP.
- **ICU Transfer Prediction**: Using clinical notes from the first 48 hours of hospital admission, we predict whether a patient will be admitted to the

[1] https://anonymous.4open.science/r/health-agents-F07A

ICU during the same visit. This binary task is evaluated on PHOP only, and visits with ICU transfers within the first 48 hours are excluded to avoid information leakage.

- **ICU Readmission Prediction**: Using clinical notes from the first 48 hours of the prior ICU stay, we predict ICU readmission during the same hospital visit (Papadakis et al., 2025). This binary task is evaluated on MIMIC-III only, and only readmission occurring at least 48 hours after the previous ICU discharge are included to prevent information leakage.

Elaborate details about the tasks are provided in the Appendix A.6. Note that for answering **RQs 2-4**, we considered LOS Category Prediction as the primary prediction task. We primarily share the results on MIMIC-III in the main text and relegate the analysis on PHOP to the appendix to aid readability and reproducibility.

Evaluation Metrics: Both the ICU Transfer and ICU Readmission Prediction tasks exhibit highly imbalanced labels, as detailed in Tables 4 and 5 in the Appendix. Following prior work on binary classification with imbalanced datasets (Choudhuri et al., 2023, 2025; Jang et al., 2022; Choudhuri et al., 2024), we report the area under the ROC curve (ROC-AUC) for both binary tasks. Additionally, consistent with (Jang et al., 2022), we include the macro F1 score for these tasks. For the multi-class classification problem of length-of-stay (LOS) category prediction, we use overall accuracy and macro F1 score, in line with previous studies (Choudhuri et al., 2024).

4.2 Baseline Methods

We compare our approach against several baselines.

1. **Traditional Models:** These use classic NLP techniques that capture lexical or statistical features of the text. Term Frequency-Inverse Document Frequency (TF-IDF) (Salton and Buckley, 1988) encodes term relevance based on frequency statistics, whereas Word2Vec (Mikolov et al., 2013) and Doc2Vec (Le and Mikolov, 2014) learn distributed embeddings capturing semantic relationships at the word and document levels, respectively.
2. **Deep Learning Models:** These methods utilize contextualized embeddings and transformer-based architectures for richer text understanding. BERT (Devlin et al., 2019) leverages bidirectional context via masked language modeling, and ClinicalBERT (Alsentzer et al., 2019) adapts BERT to the clinical domain by pretraining on large-scale medical corpora such as MIMIC-III.
3. **Large Language Models (LLMs):** These are large-scale transformer models pretrained on broad general or medical corpora. LLaMA-3.1 (Dubey et al., 2024) is a general-purpose LLM optimized for instruction following, while HuatuoGPT-o1 (Chen et al., 2024) is a medical-domain LLM fine-tuned from LLaMA-3.1 and designed to enhance clinical reasoning.

4.3 RQ1: Downstream Patient Risk Prediction Performance

LOS Category Prediction: The results of the experiments are presented in Table 1. We observe that for MIMIC-III, traditional models like TF-IDF and Word2Vec outperform the Deep Learning Models like BERT, LLAMA, and HuatuoGPT across both metrics. However, we notice that our framework, which uses HuatuoGPT as a backbone, obtains a performance improvement of more than 17% on average over the LLM alone. Additionally, our proposed framework outperforms the nearest baseline by about 11% in terms of accuracy and about 2% in terms of macro F-1 score on average on the MIMIC-III dataset. On the other hand, in the PHOP dataset, where the clinical notes are not deidentified like MIMIC-III, we notice that LLM-based methods dominate the performance. This hints that LLMs can leverage additional information when the entire data is provided, whereas count-based methods may do better in deidentified data. In PHOP, our method outperforms the closest performing baseline HuatuoGPT by about 15% in terms of mean accuracy and about 8% in terms of macro F-1 score.

ICU Transfer and Readmission Prediction: The results of the experiments are presented in Table 2. We observe that for the ICU Readmission Task (MIMIC-III), transformer-based deep learning models like BERT outperform the other baselines in terms of AUC-ROC score. However, traditional models like Word2Vec and TF-IDF dominate performance akin to the trend in Table 1 for MIMIC-III. On the other hand, we notice that for PHOP, deep learning and LLM-based methods significantly outperform traditional models, particularly in F-1 scores, where ClinicalBERT achieves the highest baseline performance, and HuatuoGPT-

Table 1: LOS category prediction performance on MIMIC-III and PHOP.

Model	MIMIC-III		PHOP	
	Accuracy	F-1	Accuracy	F-1
TF-IDF	52.26±0.07	47.42±0.06	38.94±0.17	18.31±1.09
Word2Vec	52.03±0.45	48.77±0.46	38.57±0.87	21.04±0.03
Doc2Vec	47.14±1.07	44.59±1.08	33.16±0.26	19.53±0.11
BERT	51.38±1.13	45.74±0.58	36.83±2.09	20.04±0.27
ClinicalBERT	47.31±0.86	43.31±0.81	39.61±0.60	22.18±0.91
LLaMA-3.1 8B	51.35±0.65	41.65±2.69	41.47±1.03	23.05±0.85
HuatuoGPT-o1 8B	50.00±0.82	45.67±0.47	43.16±0.59	23.99±0.74
Ours	**57.66±1.07**	**49.43±0.73**	**49.50±0.83**	**25.97±0.92**

Table 2: Binary classification performance on MIMIC-III and PHOP.

Model	MIMIC-III		PHOP	
	AUC-ROC	F-1	AUC-ROC	F-1
TF-IDF	61.41±1.09	51.90±1.78	61.44±0.59	26.41±0.20
Word2Vec	60.96±0.27	52.16±0.06	58.20±1.08	31.15±0.52
Doc2Vec	53.40±1.82	49.95±0.68	59.18±0.29	30.02±0.41
BERT	64.41±0.28	48.00±2.78	58.77±0.57	33.95±0.62
ClinicalBERT	56.40±4.48	48.96±1.25	60.97±0.26	45.83±0.19
LLaMA-3.1 8B	50.48±3.15	48.71±0.02	62.50±0.58	43.77±0.64
HuatuoGPT-o1 8B	49.77±1.71	49.61±1.24	65.28±1.21	44.19±0.93
Ours	**69.01±0.57**	**53.19±0.39**	**67.37±0.78**	**50.63±1.07**

o1 leads in AUC-ROC. This pattern mirrors the observations from the LOS category prediction on PHOP, suggesting that more advanced models can better leverage the richer, non-deidentified clinical notes in this dataset to capture nuanced patient trajectories. In contrast, traditional methods like TF-IDF and Word2Vec show lower performance, likely due to their reliance on simpler lexical features that struggle with the complexity and variability of full hospital-wide data.

Our proposed framework demonstrates superior results across both datasets for binary classification tasks. On MIMIC-III (ICU Readmission Prediction), it achieves a 7% improvement over the best baseline for AUC-ROC, and an improvement of 2% in terms of F-1 score. On PHOP (ICU Transfer Prediction), it reaches has a 3% improvement in terms of AUC-ROC and an F-1 improvement of 10%. These gains highlight the robustness of our approach in handling imbalanced binary outcomes by integrating domain-specific reasoning with efficient feature extraction.

4.4 RQ2: Ablation Study

Figure 3 reports the impact of removing viewpoint-specific note subsets on LOS prediction performance for MIMIC-III, while Figure 7 in the Appendix reports the same for PHOP. For both datasets, removing any viewpoint reduces performance relative to using all notes, confirming that each viewpoint provides complementary clinical cues. In MIMIC-III, removing **Radiology** notes yields the largest drop, suggesting that imaging-related observations carry unique signals for deterioration risk in the ICU. Removing **Doctor** or **Nurse** notes also degrades performance, but less severely, indicating partial redundancy between these narrative styles. Analysis of the results for PHOP is provided in A.2. Overall, these results demonstrate that viewpoint-specific notes encode non-overlapping information that is essential for robust LOS prediction, and removing any viewpoint weakens model performance.

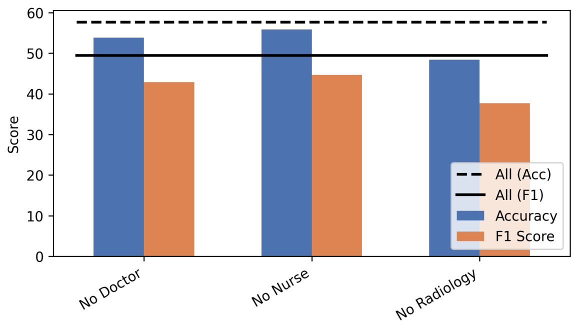

Figure 3: **Ablation Study:** Performance on MIMIC-III Dataset on removing each viewpoint.

4.5 RQ3: Feasibility Study

To empirically assess the benefit of viewpoint-specific guiding questions, we conducted an additional experiment in which a single, shared set of guiding questions was applied to clinical notes from all viewpoints. To ensure a fair comparison and avoid exceeding the context length of the back-

Table 3: Viewpoint-specific versus shared guiding questions for MIMIC-III and PHOP

Setting	Accuracy	F-1
MIMIC-III		
Shared GQs	51.19 ± 0.16	48.51 ± 0.21
Separate GQs	57.66 ± 1.07	49.43 ± 0.73
PHOP		
Shared GQs	46.94 ± 0.17	24.08 ± 0.73
Separate GQs	49.50 ± 0.83	25.97 ± 0.92

bone LLM, we fixed the total number of guiding questions to 8 (in contrast, our proposed framework uses 5 guiding questions per viewpoint). The prompts are slightly modified to account for this change of data form. We use the same input data for both models. The resulting test performance on the MIMIC-III and PHOP datasets is in Table 3.

We notice that for PHOP, using separate Guiding Questions leads to superior performance in the LOS category prediction task, both in terms of accuracy and F-1 score. The same trend also holds for MIMIC-III, where separate guiding questions lead to higher gain in accuracy than the F-1 score. This empirically validates the broader motivation of our work. The guiding questions found for both MIMIC-III and PHOP for both methods are provided in A.7.

4.6 RQ 4: Importance of Learned Questions for Long-Term Procedure Prediction

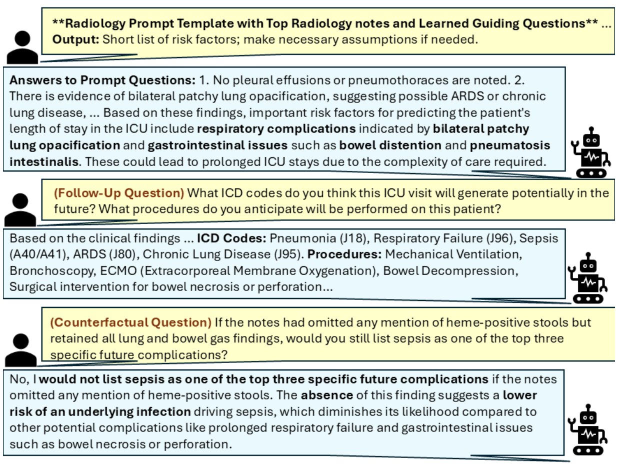

Figure 4: **Case Study:** Ability of our framework to answer open-ended questions.

During inference, we have different risk factors for clinical notes coming over time, which are combined to make soft predictions along with reasoning. As our framework relies on LLMs to generate these risk factors, it subsequently allows humans to further interact with the LLMs. As our framework does not rely on finetuning on the backbone LLM weights, it is imbibed with the flexibility to handle general questions compared to its fine-tuned counterparts, which suffer from catastrophic forgetting and related problems (Liu et al., 2024a; Zhai et al., 2024). To demonstrate the general question-answering ability of our framework, we looked at the data for the patient with ICUSTAY ID 208712 (patient in the test set with the largest ICU Length of Stay) in MIMIC-III. This patient was a newborn who was admitted to the NICU and only had radiology notes during the first 48 hours of stay, with the overall ICU length of stay being 123.28 days. The summaries of the top 10 important notes are shown in Figure 10 in the Appendix. The results of the follow-up conversation are shown in Figure 4. We can notice that the task-specific radiology guiding questions Q^*_{radio} can steer reasoning in the frozen LLM, while allowing it to answer follow-up and counterfactual questions.

Additionally, on looking at the clinical notes created during this ICU stay (as procedures were absent in the data for this patient), we found strong evidence of mechanical ventilation and bowel decompression, and moderate evidence of surgery for bowel necrosis/perforation (we provide the clinical notes in the supplementary material). This demonstrates that while the LLM over-predicts high-acuity rescue procedures (ECMO, bronchoscopy) based on severe diagnoses (ARDS, sepsis), it accurately predicts most other procedures. This leads to future opportunities for exploring calibration versus plausibility in procedure prediction.

5 Conclusion

In this work, we showed that effectively leveraging clinical notes for patient risk prediction requires explicitly accounting for the heterogeneity across note viewpoints. We introduced an end-to-end framework that treats viewpoint-specific guiding questions as learnable natural language parameters. Through experiments on two EHR datasets and multiple prediction tasks, we demonstrated consistent improvements over several baselines. In conclusion, this work highlights the need to ask the right view-aware questions when adapting LLMs to heterogeneous clinical narratives and opens avenues for future research on calibration, longitudinal reasoning, and procedure-level inference.

6 Limitations

This study shares limitations common to AI-based clinical modeling. Our evaluation is retrospective and relies on observational EHR data, which may encode documentation biases, missing information, and institution-specific practices that limit causal interpretation and real-world generalization. As our framework builds on large language models, it inherits known issues such as prompt sensitivity, hallucinated or over-confident reasoning, and imperfect calibration, particularly when predicting high-acuity outcomes. Finally, our experiments are limited to a small number of tasks, datasets, and care-domains, and future work is needed to assess generalization, calibration, and prospective clinical utility.

References

Awais Ahmed, Mengshu Hou, Rui Xi, Xiaoyang Zeng, and Syed Attique Shah. 2024. Prompt-eng: Healthcare prompt engineering: Revolutionizing healthcare applications with precision prompts. In *Companion Proceedings of the ACM Web Conference 2024*, pages 1329–1337.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Jeff Bilmes, Krste Asanovic, Chee-Whye Chin, and Jim Demmel. 1997. Using phipac to speed error backpropagation learning. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 4153–4156. IEEE.

Jeremiah R Brown, Iben M Ricket, Ruth M Reeves, Rashmee U Shah, Christine A Goodrich, Glen Gobbel, Meagan E Stabler, Amy M Perkins, Freneka Minter, Kevin C Cox, and 1 others. 2022. Information extraction from electronic health records to predict readmission following acute myocardial infarction: does natural language processing using clinical notes improve prediction of readmission? *Journal of the American Heart Association*, 11(7):e024198.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024. Huatuogpt-o1, towards medical complex reasoning with llms. *Preprint*, arXiv:2412.18925.

Yanda Chen, Chen Zhao, Zhou Yu, Kathleen McKeown, and He He. 2023. On the relation between sensitivity and accuracy in in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 155–167.

Yinda Chen, Yangfan He, Jing Yang, Dapeng Zhang, Zhenlong Yuan, Muhammad Attique Khan, Jamel Baili, and Lip Yee. 2025. Empower: Evolutionary medical prompt optimization with reinforcement learning. *IEEE Journal of Biomedical and Health Informatics*.

Akash Choudhuri, Hankyu Jang, Alberto M Segre, Philip M Polgreen, Kishlay Jha, and Bijaya Adhikari. 2023. Continually-adaptive representation learning framework for time-sensitive healthcare applications. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 4538–4544.

Akash Choudhuri, Philip Polgreen, Alberto Segre, and Bijaya Adhikari. 2024. Summarizing clinical notes using llms for icu bounceback and length-of-stay prediction. In *2024 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 859–866. IEEE.

Akash Choudhuri, Hieu Vu, Kishlay Jha, and Bijaya Adhikari. 2025. Domain knowledge augmented contrastive learning on dynamic hypergraphs for improved health risk prediction. In *Proceedings of the 2025 SIAM International Conference on Data Mining (SDM)*, pages 476–486. SIAM.

JJ Cyr and GA Haley. 1983. Use of demographic and clinical characteristics in predicting length of psychiatric hospital stay: a final evaluation. *Journal of Consulting and Clinical Psychology*, 51(4):637.

Barnan Das, Diane J Cook, Maureen Schmitter-Edgecombe, and Adriana M Seelye. 2012. Puck: an automated prompting system for smart environments: toward achieving automated prompting—challenges involved. *Personal and ubiquitous computing*, 16(7):859–873.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric Xing, and Zhiting Hu. 2022. Rlprompt: Optimizing discrete text prompts with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3369–3391.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Chen Dun, Mirian Del Carmen Hipolito Garcia, Guoqing Zheng, Ahmed Hassan Awadallah, Robert Sim, and Anastasios Kyrillidis. 2025. Sweeping heterogeneity with smart mops: Mixture of prompts for llm task adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 16426–16434.

Daniel J Feller, Jason Zucker, Michael T Yin, Peter Gordon, and Noémie Elhadad. 2018. Using clinical notes and natural language processing for automated hiv risk assessment. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 77(2):160–166.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, Matthew M Churpek, Ozlem Uzuner, and Majid Afshar. 2023. Progress note understanding—assessment and plan reasoning: overview of the 2022 n2c2 track 3 shared task. *Journal of biomedical informatics*, 142:104346.

Roger Garriga, Teodora Sandra Buda, João Guerreiro, Jesús Omaña Iglesias, Iñaki Estella Aguerri, and Aleksandar Matić. 2023. Combining clinical notes with structured electronic health records enhances the prediction of mental health crises. *Cell Reports Medicine*, 4(11).

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Bowen Gu, Rishi J Desai, Kueiyu Joshua Lin, and Jie Yang. 2024. Probabilistic medical predictions of large language models. *npj Digital Medicine*, 7(1):367.

Marco Guevara, Shan Chen, Spencer Thomas, Tafadzwa L Chaunzwa, Idalid Franco, Benjamin H Kann, Shalini Moningi, Jack M Qian, Madeleine Goldstein, Susan Harper, and 1 others. 2024. Large language models to identify social determinants of health in electronic health records. *NPJ digital medicine*, 7(1):6.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, and 1 others. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622.

Kexin Huang, Tamryn F Gray, Santiago Romero-Brufau, James A Tulsky, and Charlotta Lindvall. 2021. Using nursing notes to improve clinical outcome prediction in intensive care patients: a retrospective cohort study. *Journal of the American Medical Informatics Association*, 28(8):1660–1666.

Kingma DP Ba J and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 1412(6).

Hankyu Jang, Sulyun Lee, DM Hasibul Hasan, Philip M Polgreen, Sriram V Pemmaraju, and Bijaya Adhikari. 2022. Dynamic healthcare embeddings for improving patient care. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 52–59. IEEE.

Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. 2024. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics*, 40(Supplement_1):i119–i129.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Yeonsu Kwon, Jiho Kim, Gyubok Lee, Seongsu Bae, Daeun Kyung, Wonchul Cha, Tom Pollard, Alistair Johnson, and Edward Choi. 2024. Ehrcon: Dataset for checking consistency between unstructured notes and structured tables in electronic health records. *Advances in Neural Information Processing Systems*, 37:89334–89345.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. 2022. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *arXiv preprint arXiv:2201.11838*.

Chengyuan Liu, Yangyang Kang, Shihang Wang, Lizhi Qing, Fubang Zhao, Chao Wu, Changlong Sun, Kun Kuang, and Fei Wu. 2024a. More than catastrophic forgetting: Integrating general capabilities for domain-specific llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7531–7548.

Jinghui Liu, Daniel Capurro, Anthony Nguyen, and Karin Verspoor. 2022. Note bloat impacts deep learning-based nlp models for clinical prediction tasks. *Journal of biomedical informatics*, 133:104149.

Jinghui Liu, Daniel Capurro, Anthony Nguyen, and Karin Verspoor. 2024b. Uncovering variations in clinical notes for nlp modeling. In *MEDINFO 2023—The Future Is Accessible*, pages 1460–1461. IOS Press.

Jingshu Liu, Zachariah Zhang, and Narges Razavian. 2018. Deep ehr: Chronic disease prediction using medical notes. In *Machine Learning for Healthcare Conference*, pages 440–464. PMLR.

Sheng Lu, Hendrik Schuff, and Iryna Gurevych. 2024. How are prompts different in terms of sensitivity? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5833–5856.

Yukinori Mashima, Masatoshi Tanigawa, and Hideto Yokoi. 2024. Information heterogeneity between progress notes by physicians and nurses for inpatients with digestive system diseases. *Scientific Reports*, 14(1):7656.

Ernestina Menasalvas and Consuelo Gonzalo-Martin. 2016. Challenges of medical text and image processing: Machine learning approaches. In *Machine Learning for Health Informatics: State-of-the-Art and Future Challenges*, pages 221–242. Springer.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Eleftherios Papadakis, Athanasia Proklou, Sofia Kokkini, Ioanna Papakitsou, Ioannis Konstantinou, Aggeliki Konstantinidi, Georgios Prinianakis, Stergios Intzes, Marianthi Symeonidou, and Eumorfia Kondili. 2025. Icu readmission and in-hospital mortality rates for patients discharged from the icu—risk factors and validation of a new predictive model: The worse outcome score (woscore). *Journal of Personalized Medicine*, 15(10):479.

Bhagyajit Pingua, Adyakanta Sahoo, Meenakshi Kandpal, Deepak Murmu, Jyotirmayee Rautaray, Rabindra Kumar Barik, and Manob Jyoti Saikia. 2025. Medical llms: Fine-tuning vs. retrieval-augmented generation. *Bioengineering*, 12(7):687.

Kiran Ramnath, Kang Zhou, Sheng Guan, Soumya Smruti Mishra, Xuan Qi, Zhengyuan Shen, Shuai Wang, Sangmin Woo, Sullam Jeoung, Yawei Wang, and 1 others. 2025. A systematic survey of automatic prompt optimization techniques. *arXiv preprint arXiv:2502.16923*.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Yuqi Si and Kirk Roberts. 2021. Three-level hierarchical transformer networks for long-sequence and multiple clinical documents classification. *arXiv preprint arXiv:2104.08444*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023a. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, and 1 others. 2023b. Towards expert-level medical question answering with large language models. 2023. *arXiv preprint arXiv:2305.09617*, 10.

William R Small, Jonathan Austrian, Luke O'Donnell, Jesse Burk-Rafel, Katherine A Hochman, Adam Goodman, Jonah Zaretsky, Jacob Martin, Stephen Johnson, Vincent J Major, and 1 others. 2025. Evaluating hospital course summarization by an electronic health record–based large language model. *JAMA Network Open*, 8(8):e2526339–e2526339.

Jiwoong Sohn, Yein Park, Chanwoong Yoon, Sihyeon Park, Hyeon Hwang, Mujeen Sung, Hyunjae Kim, and Jaewoo Kang. 2025. Rationale-guided retrieval augmented generation for medical question answering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12739–12753.

Jiyoun Song, Mollie Hobensack, Kathryn H Bowles, Margaret V McDonald, Kenrick Cato, Sarah Collins Rossetti, Sena Chae, Erin Kennedy, Yolanda Barrón, Sridevi Sridharan, and 1 others. 2022. Clinical notes: an untapped opportunity for improving risk prediction for hospitalization and emergency department visit during home health care. *Journal of biomedical informatics*, 128:104039.

Maryam Tayefi, Phuong Ngo, Taridzo Chomutare, Hercules Dalianis, Elisa Salvi, Andrius Budrionis, and Fred Godtliebsen. 2021. Challenges and opportunities beyond structured data in analysis of electronic health records. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(6):e1549.

Xiaochen Wang, Junyu Luo, Jiaqi Wang, Ziyi Yin, Suhan Cui, Yuan Zhong, Yaqing Wang, and Fenglong Ma. 2023. Hierarchical pretraining on multimodal electronic health records. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2023, page 2839.

Yuanlong Wang, Changchang Yin, and Ping Zhang. 2024. Multimodal risk prediction with physiological signals, medical images and clinical notes. *Heliyon*, 10(5).

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 2300–2344.

Zhi Wen, Pratheeksha Nair, Chih-Ying Deng, Xing Han Lu, Edward Moseley, Naomi George, Charlotta Lindvall, and Yue Li. 2021. Mining heterogeneous clinical notes by multi-modal latent topic model. *PloS one*, 16(4):e0249622.

Wei-Hung Weng, Kavishwar B Wagholikar, Alexa T McCray, Peter Szolovits, and Henry C Chueh. 2017. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. *BMC medical informatics and decision making*, 17(1):155.

Chengqian Xian, Camila PE de Souza, and Felipe F Rodrigues. 2023. Health outcome predictive modelling in intensive care units. *Operations Research for Health Care*, 39:100409.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, and 1 others. 2022. A large language model for electronic health records. *NPJ digital medicine*, 5(1):194.

Muchao Ye, Weiyang Liu, and Pan He. 2025. Vera: Explainable video anomaly detection via verbalized learning of vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8679–8688.

Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R Dalal, Jennifer L Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan Ashley, and 1 others. 2024. Almanac—retrieval-augmented language models for clinical medicine. *Nejm ai*, 1(2):AIoa2300068.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. 2024. Investigating the catastrophic forgetting in multimodal large language model fine-tuning. In *Conference on Parsimony and Learning*, pages 202–227. PMLR.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR.

Yinghao Zhu, Junyi Gao, Zixiang Wang, Weibin Liao, Xiaochen Zheng, Lifang Liang, Miguel O Bernabeu, Yasha Wang, Lequan Yu, Chengwei Pan, and 1 others. 2024. Clinicrealm: Re-evaluating large language models with conventional machine learning for non-generative clinical prediction tasks. *arXiv preprint arXiv:2407.18525*.

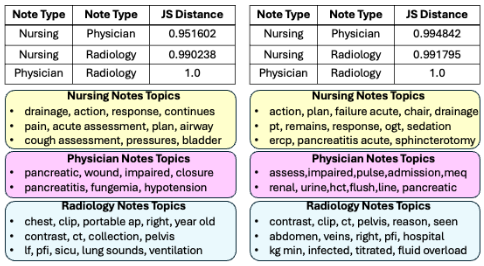

Figure 5: **Analysis of Notes from SUBJECT ID 78076 (left) and SUBJECT ID 77614 (right)**: The Jensen-Shannon distance between the topic distributions for each note category (top). Note that lower values indicate that the topics of the notes are similar, while higher values indicate high differences. The top topics for each note category (bottom).

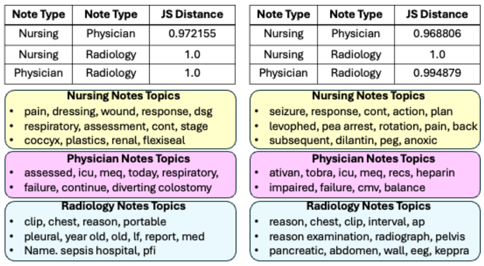

Figure 6: **Analysis of Notes from SUBJECT ID 27427 (left) and SUBJECT ID 99383 (right)**: The Jensen-Shannon distance between the topic distributions for each note category (top). Note that lower values indicate that the topics of the notes are similar, while higher values indicate high differences. The top topics for each note category (bottom).

A Appendix

A.1 Analysis of Clinical Note Categories of Patients

We excluded the discharge summaries and took all other types of notes. To quantify the degree of difference in the categories of clinical notes, we first performed standard preprocessing on the note texts, initialized a vectorizer with unigrams and bigrams, and fit BERTTopic (Grootendorst, 2022) to the notes to get topic clusters. Finally, we computed the Jensen-Shannon distance between topic distributions for each category of note specialty.

In addition to analyzing the clinical notes for SUBJECT ID 55672, which had 1,415 clinical notes, we also analyzed 4 other SUBJECT IDs, namely 78076 (1,409 notes), 77614 (1,349 notes), 27427 (1,289 notes), and 99383 (1,273 notes). The results of our overall analysis are presented in Figures 5 and 6. While we can clearly see that the clinical notes written by healthcare entities of different specialities are very different from each other (similar to the result in Figure 1) for the other four SUBJECT IDs, we can notice that the topics are slightly different.

Nursing notes predominantly revolve around **direct patient care and bedside management**, with recurring emphasis on **acute clinical events and hands-on interventions**. These include care actions and patient responses (e.g., *action, plan, response, drainage*), pain and wound management, airway and respiratory support (*cough, tracheal,*

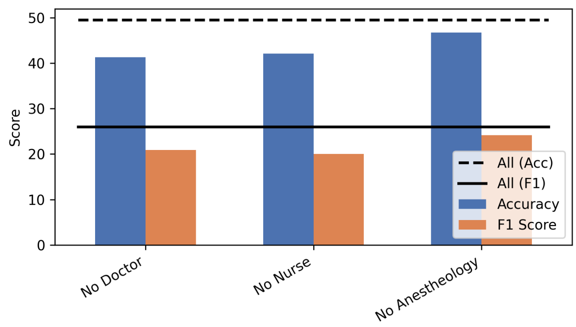

Figure 7: **Ablation Study**: Performance on PHOP Dataset on removing each specialty.

sedation), neurological monitoring (*seizure, response*), and procedural or tube-related care (*PEG, Flexi-Seal, ERCP, sphincterotomy*).

In contrast, **physician notes focus on high-level ICU management and clinical oversight**, emphasizing patient assessment, physiologic monitoring, and ongoing critical care therapies. The extracted topics highlight routine evaluation of vital and respiratory parameters (*ICU, pulse, assessed, cmH$_2$O, mEq*), along with continuous medication management, particularly analgesia, sedation, and anticoagulation (*fentanyl, heparin, hour, continue*). This reflects physician decision-making centered on treatment continuation, stability, and overall clinical trajectory rather than bedside task execution.

Radiology notes primarily center on **diagnostic imaging interpretation**, with a focus on chest radiographs and CT imaging (*chest, portable, atelectasis, CT, contrast, lower lobe*). These notes document imaging indications and longitudinal comparisons (*reason, interval*), as well as vascular and device-related findings (*vein, DVT, PEG*), supporting diagnostic confirmation and disease progression assessment.

A.2 Ablation Study for PHOP

On the other hand, for PHOP (shown in Figure 7), where the data come from overall hospital operations rather than ICU operations, the same trend holds. Retaining all notes yields the best performance. The largest degradation arises when removing **Doctor** notes for Accuracy and when removing **Nurse** notes for F-1 score, while discarding **Anesthesiology** notes has a moderate but noticeable effect. Compared to MIMIC-III, the relative drops are more uniform across specialties, suggesting that PHOP notes exhibit a more distributed signal contribution across provider types.

A.3 Prompts for Agents

The prompts for the domain-specific learner agent and the optimizer agent are presented in Figure 8.

On the other hand, the prompts for the inference predictor and the domain aggregator agent prompt templates are present in Figure 9.

A.4 Details about the Datasets

As mentioned in the main text, we used two datasets in our experiments. Their details are as follows:

- **MIMIC-III:** This is de-identified healthcare operations data from patients who were admitted to the critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. The dataset contains data from heterogeneous sources, including demographic information, International Classification of Diseases codes (ICD-9), hourly vital signs, laboratory tests, microbiological culture results, medication administrations, and survival statistics.
- **PHOP:** In this data, each patient visit includes a list of diagnoses, a timestamped record of room transfers, physician-performed procedures, and prescription medications. We extracted patient visit data along with corresponding clinical notes between December 1, 2007, and August 31, 2008.

A.5 Details of the Experiments

We provide more details of the experimental procedure that will aid reproducibility. For our model, the logistic regression function was used as $\varrho(\cdot)$. Similarly, this was the model used on top of the embeddings for TF-IDF, Word2Vec, and Doc2Vec.

Table 4: Label distribution for length-of-stay (LOS) prediction tasks across datasets.

Label	**MIMIC-III**	**PHOP**
Low (2–4 days)	3682	463
Medium (4–7 days)	1982	595
High (7+ days)	2664	845
Total	**8328**	**1903**

A.6 Further Details about the Predictive Tasks

In our experiments, we evaluated model predictions on two primary tasks:

Viewpoint-Specific Learner Prompt Template ψ

Model Description: You are a clinical expert designed to identify important risk factors from clinical notes.
The input consists of summaries of the most important clinical notes of domain d for the first 48 hours following a patient's admission into the ICU, with stopwords and punctuation removed. Your task is to identify patient risk factors in these notes. If a normal human medical professional has level 10 knowledge, you will have level 3000 knowledge in this role. As it's crucial for the patient's well-being, please ensure accurate assessments because if you don't, the patient's condition could worsen. Take pride in your work and give it your best. Do not tell your limitations. Do not echo my prompt. Please do not remind me what I asked you for. Do not apologize. Do not self-reference. Get to the point precisely and accurately. First, answer the prompt questions below.
Prompt Questions: Answer the following questions based on what you observe from the clinical notes and provide an explanation in one sentence.

[Guiding Questions Q_t]

Based on the analysis above, please conclude your answer to "What are the important findings from the physician notes that may help to predict the patient's length of stay in the ICU?"

Input: **[Clinical Notes of domain d]**

Please give your output strictly in the following format:

- **Answers to Prompt Questions:** Provide your analysis by answering the questions listed in Prompt Questions.
- **Output:** Short list of risk factors; make necessary assumptions if needed

Please ONLY reply according to this format, don't give me any other words or repeat my words.

Optimizer Prompt Template ν

You are the optimizer for a model. Your goal is to learn the best prompt questions to properly classify the length of stay of the ICU visit for the model. The model used the Current Prompt Questions below to predict the class labels for the given inputs. You are given the target labels. Please optimize the Current Prompt Questions for better prediction.

Inputs: **[A batch of risk factors extracted from important note summaries for each Domain]**

Model Descriptions: You are designed to do multi-class classification. The input is important summaries of nurse, radiology and physician clinical notes (stopwords removed) for the first 48 hours following a patient's admission into the ICU for identifying what the patient's length of stay will be for the current ICU visit; you need to output the class label, i.e., an integer in the set {0, 1, 2}. 0 represents short length of stay (between 2-4 days), and 1 represents medium length of stay (between 4-7 days), and 2 represents long length of stay (7+ days). Please answer the prompt questions.
Current Prompt Questions: Answer the following questions based on what you observe from the clinical notes and provide an explanation in one sentence.

[Guiding Questions for all Domains]

Based on the analysis above, please conclude your answer to 'What is the length of stay category for this patient visit?' in 'It is label 0 (2-4 days)' or 'It is label 1 (4-7 days)' or 'It is label 2 (7+ days)'.

The model predictions: **[A batch of learner model predictions]**

The targets: **[A batch of true labels]**

If the model is doing well, you can keep using the current prompt questions. However, if the model is not performing well, please update the model by improving upon the 'Current Prompt Questions', which should result in lower classification error both on the current and the next batch of i.i.d. data. Limit your 'New Prompt Questions' to no more than five questions! Please think step by step and give your outputs strictly in the following format:

- **Reasoning:** [Be explicit and verbose, improve the Current Prompt Questions by yourself; please show your work and use the features in the videos; note that you don't have access to computers]
- **New Nurse Prompt Questions:** [Put your new nurse prompt questions here, the questions MUST be based on the features in the input summaries. please limit prompt questions to be at most five!]
- **New Doctor Prompt Questions:** [Put your new doctor prompt questions here, the questions MUST be based on the features in the input summaries. please limit prompt questions to be at most five!]
- **New Radiology Prompt Questions:** [Put your new radiology prompt questions here, the questions MUST be based on the features in the input summaries. please limit prompt questions to be at most five!]

Please ONLY reply according to this format. Don't give me any other words.

Figure 8: The Viewpoint-Specific Learner Prompt Template (left) and the Optimizer Prompt Template (right).

Inference Predictor Prompt Template ζ

You are a classification model. Given the input, output ONLY the integer class label: 0, 1, or 2.
Do NOT include any explanation, words, or punctuation.
Your output must be a single digit: 0, 1, or 2.
If you understand, reply with only the number.
Model Description: You are designed to do multi-class classification. The input is important summaries of nurse, physician and radiology clinical notes for the first 48 hours following a patient's admission into the ICU for identifying what the patient's length of stay will be for the current ICU visit; you need to output the class label, i.e., an integer in the set {0, 1, 2}. 0 represents short length of stay (between 2-4 days), and 1 represents medium length of stay (between 4-7 days), and 2 represents long length of stay (7+ days).
Based on the information from the notes, what is the length of stay category for this patient visit? Feel free to make necessary assumptions, but PREDICT A NUMBER BETWEEN 0,1 AND 2.
Input:

[$Data_n]
[$Data_r]
[$Data_d]

Please give your output strictly in the following format:

Output:
[Provide the overall risk factors with a prediction between 0, 1, or 2, indicating the ICU length of stay of the patient. Always provide a prediction.]
Reason:
[Provide a brief reasoning for your prediction, citing specific details from the notes that influenced your decision. Only include relevant information that supports your prediction.]

Please ONLY reply according to this format, DON'T INCLUDE ANY OTHER WORDS.

Aggregate Learner Prompt Template Ω

You are a classification model. Given the input, output ONLY the integer class label: 0, 1, or 2.
Do NOT include any explanation, words, or punctuation.
Your output must be a single digit: 0, 1, or 2.
If you understand, reply with only the number.
Model Description: You are designed to do multi-class classification. The input is important summaries of nurse, physician and radiology clinical notes for the first 48 hours following a patient's admission into the ICU for identifying what the patient's length of stay will be for the current ICU visit; you need to output the class label, i.e., an integer in the set {0, 1, 2}. 0 represents short length of stay (between 2-4 days), and 1 represents medium length of stay (between 4-7 days), and 2 represents long length of stay (7+ days).
Based on the information from the notes, what is the length of stay category for this patient visit? Feel free to make necessary assumptions but PREDICT A NUMBER BETWEEN 0,1 AND 2.
Input:

[$Data_n]
[$Data_r]
[$Data_d]

Please give your output strictly in the following format:
Output:
[ONLY the integer class label : 0, 1, or 2. NO WORDS, NO EXPLANATION, NO PUNCTUATION.]

Please ONLY reply according to this format, DON'T INCLUDE ANY OTHER WORDS.

Figure 9: **All Learner Prompts**: Inference Predictor (left) and Aggregate Learner (right Prompt Templates.

- **Length of Stay Category Prediction**: The first predictive task we present is the prediction of the total length of stay (LOS) for each patient visit in the ICU. Although this problem can be posed as a regression problem (Cyr and Haley, 1983), our study presents it as a multi-class classification problem similar to (Xian et al., 2023; Choudhuri et al., 2024), with different classes representing different ICU stay categories. LOS between 2-4 days was categorized as 'low', between 4-7 days was classified as 'medium', and 7 days and above was categorized as 'high'. We perform this task on both MIMIC-III and PHOP datasets. However, for MIMIC-III, the task is to predict the LOS category for each ICU visit, while for PHOP, the task is to predict the LOS category for

Note 1: key clinical findings note include infant prematurity elective reintubation rul collapse previous film patient evaluated lung fields underwent chest exam showed improved aeration lungs residual hazy right upper lobe opacity chronic underlying patchy predominantly perihilar reticular opacities also noted treatment recommendations provided note
Note 2: key clinical finding endotracheal tube placed verify ett placement patient infant prematurity elective reintubation distal tip endotracheal tube terminates distal thoracic inlet diffuse granular opacities persist throughout lungs focal air space opacities visualized upper abdominal bowel gas pattern unremarkable treatment recommendations provided note
Note 3: key clinical findings note include 24 week ga newborn reintubated apnea cardiac murmur reason examination check ett position evaluate heart lungs findings showed et tube withdrawn several millimeters likely terminates approximately 3 4 mm carina bilateral granular predominantly perihilar opacities persisting likely representing diffuse microatelectasis treatment recommendation provided given information
Note 4: key clinical findings note patient 24 week infant mild ventriculomegaly underwent follow study rule pvl scans showed normal cerebral cerebellar architecture evidence hydrocephalus hemorrhage edema mass small cyst left choroid plexus posterior foramen last name un 565 measuring 4 x 5 x 4 mm size treatment recommendations provided note
Note 5: key clinical findings patient chronic lung disease prematurity bilateral patchy lung opacification clip projecting expected position ductus arteriosis bowel gas pattern quite normal paucity gas right abdomen evidence bowel distention free intraperitoneal air treatment recommendations cannot provided based information given note
Note 6: key clinical findings infant chronic lung disease concerning bowel gas pattern hemetest positive stools patient underwent single frontal portal view abdomen showed several uniformly distended gas filled loops bowel without focal obstruction definite radiographic evidence pneumatosis portal venous gas nasogastric tube also noted treatment recommendations provided given information
Note 7: key clinical findings note include newborn reintubated apnea cardiac murmur increase atelectatic opacities seen right upper lobe right lower lobe left upper lobe treatment recommendation clinically correlate findings however note provide specific treatment recommendations\n
Note 8: patient examined cause respiratory acidosis found et tube end carina nasogastric tube end stomach heart normal size coarse opacities lungs suggestive chronic lung disease also mild distention bowel loops abdomen significant change seen lungs lines tubes present described mild distention bowel loops nonspecific finding treatment recommendations given provided clinical note
Note 9: key clinical findings note include 20 day old former 24 weeker history pda ligation presented abdominal distention ap view abdomen showed distribution gas abdomen suggest bowel obstruction evidence pneumatosis intestinalis however previous babygram showed concern possible bowel obstruction therefore recommended closely monitor patient condition consider evaluation necessary
Note 10: based discharge summary key clinical finding infant chronic lung disease hemetest positive stools reason examination r pneumotosis x ray showed distended gas filled loops bowel right mid lower abdomen definite radiographic evidence pneumatosis portal venous gas clinical suspicion nec persists close follow films performed treatment recommendation provided given discharge summary

Figure 10: Pre-processed Summaries of top 10 Radiology Notes for ICUSTAY ID 208712.

Table 5: Label distribution for binary label prediction tasks across datasets.

Label	**MIMIC-III**	**PHOP**
False	7901	3339
True	427	143
Total	**8328**	**3482**

the overall hospital visit.

- **ICU Transfer Prediction**: The second predictive task we present is the prediction of whether a patient will be admitted to the ICU in the current hospital visit based on the first 48 hours of clinical notes from the patient's admittance to the hospital. We perform this experiment only on PHOP and treat it as a binary classification problem with labels indicating 'admittance' and 'not admittance'. We preprocess the visits to include those where ICU transfers took place after 48 hours following patient admittance to prevent information leakage.
- **ICU Readmission Prediction**: The third predictive task we present is the prediction of whether a patient will be readmitted to the ICU in the current hospital visit based on the clinical notes from the first 48 hours of the previous ICU visit for the current hospital visit. We perform this experiment only on MIMIC-III and treat it as a binary classification problem with labels indicating 'readmission' and 'not readmission'. We preprocess the ICU visits to only include those readmissions that took place 48 hours following the previous ICU admission to prevent information leakage.

A.7 Guiding Questions

The optimal guiding questions learned by our proposed model in the experiments are as follows:

- **MIMIC-III**
  - **LOS Category Prediction**
    * Physician:
      1. Has the patient undergone any recent surgeries, such as CABG? Recent surgeries complicate recovery and may increase the length of stay.

2. Are there any medication or treatment plans in place to address respiratory issues? This informs us about the complexity of care required.
3. How is the patient's overall neurological status? Changes in mental status can significantly impact the length of stay.
4. Are there any cardiovascular issues, such as arrhythmias or hypotension? These can complicate the patient's condition and influence the length of stay.
5. Is there any mention of renal failure or need for dialysis? Renal impairment can prolong the ICU stay.

* Nurse:
1. Is the patient experiencing respiratory failure? This is crucial because it determines the need for advanced respiratory support.
2. What type of respiratory support is the patient receiving? This provides insight into the severity of their condition and the level of care required.
3. Are there any signs of pneumonia or atelectasis? These symptoms indicate the extent of respiratory compromise and impact on the patient's recovery.
4. Is the patient on mechanical ventilation? This is a strong indicator of critical illness and potential for extended ICU stay.
5. Are there any mentions of sepsis or systemic inflammatory responses? These can complicate the patient's condition and prolong their stay.

* Radiology:
1. Are there any pleural effusions or pneumothoraces visible on imaging? These findings can indicate complications affecting the patient's respiratory status.
2. Is there evidence of ARDS or other pulmonary infiltrates? These suggest severe respiratory involvement and potential for prolonged ICU stay.
3. Are there any gastrointestinal abnormalities, such as bowel obstructions or perforations? These can complicate the patient's condition and impact recovery.
4. Is there any mention of cardiac issues, such as cardiomegaly or pulmonary edema? These can contribute to the complexity of the patient's condition.
5. Are there any signs of septic shock or systemic inflammation? These can lead to prolonged ICU stays due to the complexity of care required.

– **ICU Readmission Prediction**

* Physician:
1. Are there any critical lab values or vital signs indicating instability?
2. Is there a diagnosis of a life-threatening condition or organ dysfunction?
3. Are there any plans for invasive procedures or interventions?
4. Is there any mention of potential complications or contraindications?
5. Are there any concerns regarding the patient's mental status or neurological function?

* Nurse:
1. Does the patient have any severe illnesses like DKA or sepsis?
2. Are there any recent surgical interventions or complications mentioned?
3. Is there any indication of cardiovascular issues or vascular problems?
4. Are there any medications listed that could affect the patient's condition?
5. Is there any mention of respiratory distress or failure?

* Radiology:
1. Are there any structural abnormalities or pleural effusions visible?
2. Is there any evidence of vascular compromise or DVT?

3. Are there any signs of respiratory distress or pulmonary edema?
4. Is there any mention of gastrointestinal issues or obstructions?
5. Are there any other notable findings that could impact the patient's condition?

- **PHOP**
  - **LOS Category Prediction**
    * Physician
      1. How do the patient's cognitive impairments affect their ability to participate in their own care, and what strategies can be employed to enhance their engagement?
      2. What are the implications of the patient's medication regimen on their overall condition, and how can potential interactions be mitigated?
      3. What immediate medical concerns, such as infections or complications, require urgent attention, and how can they be addressed effectively?
      4. How does the patient's cardiovascular history, including their history of strokes and hypertension, influence their recovery and length of stay?
      5. What approaches can be taken to effectively manage the patient's pain and discomfort during their stay?
    * Nurse:
      1. What are the specific surgical risks associated with the patient's condition, and how can they be minimized through preoperative and intraoperative management?
      2. How does the patient's medical history, particularly their history of strokes and hypertension, impact their suitability for surgery and potential recovery outcomes?
      3. What are the potential complications that could arise from the procedure, and how can they be prevented or managed?
      4. How does the patient's current cognitive state affect their ability to participate in postoperative care and recovery?
      5. What is the expected length of stay after the procedure, considering the patient's complex medical history?
    * Anesthesiologist:
      1. What are the implications of the patient's complex medical history for their anesthesia management during the procedure, and how can these be addressed?
      2. How does the patient's cardiovascular history, including their history of strokes and hypertension, affect their risk for perioperative complications?
      3. What are the potential effects of the patient's diabetes on their anesthesia and recovery, and how can these be managed?
  - **ICU Transfer Prediction**
    * Physician:
      1. What are the patient's vital signs indicating, and are there any concerning trends?
      2. Which medications is the patient currently taking, and how might they impact their condition?
      3. What procedures is the patient undergoing, and what are the associated risks?
      4. Are there any signs of infection or bleeding related to the port placement?
      5. How is the patient's pain being managed, and is it effective?
    * Nurse:
      1. How often does the patient experience difficulty breathing, and what are the associated symptoms?
      2. What is the nature of the patient's pain, and how is it being managed?
      3. Are there any changes in the patient's mental state, such as confusion or disorientation?
      4. What is the patient's oxygen saturation, and are there any concerns

about oxygenation?

5. Are there any signs of complications from the recent port placement procedure?

* Anesthesiologist:
  1. What is the patient's overall condition, and are there any anesthesia-related risks?
  2. How does the patient respond to anesthesia, and are there any complications to consider?
  3. Is the patient stable enough to avoid ICU care post-surgery?
  4. Are there any concerns about the patient's airway or breathing during recovery?
  5. How is the patient's pain being managed post-operatively, and is it sufficient?