ColaCare: Enhancing Electronic Health Record Modeling through Large Language Model-Driven Multi-Agent Collaboration

Anonymous Author(s)

ABSTRACT

We introduce ColaCare, a framework that enhances Electronic Health Record (EHR) modeling through multi-agent collaboration driven by Large Language Models (LLMs). Our approach seamlessly integrates domain-specific expert models with LLMs to bridge the gap between structured EHR data and text-based reasoning. Inspired by the Multidisciplinary Team (MDT) approach used in clinical settings, ColaCare employs two types of agents: DoctorAgents and a MetaAgent, which collaboratively analyze patient data. Expert models process and generate predictions from numerical EHR data, while LLM agents produce reasoning references and decision-making reports within the MDT-driven collaborative consultation framework. The MetaAgent orchestrates the discussion, facilitating consultations and evidence-based debates among DoctorAgents, simulating diverse expertise in clinical decision-making. We additionally incorporate the Merck Manual of Diagnosis and Therapy (MSD) medical guideline within a retrieval-augmented generation (RAG) module for medical evidence support, addressing the challenge of knowledge currency. Extensive experiments conducted on three EHR datasets demonstrate ColaCare's superior performance in clinical mortality outcome and readmission prediction tasks, underscoring its potential to revolutionize clinical decision support systems and advance personalized precision medicine. The code, complete prompt templates, case studies are publicly available at the anonymous link: https://colacare.netlify.app.

1 INTRODUCTION

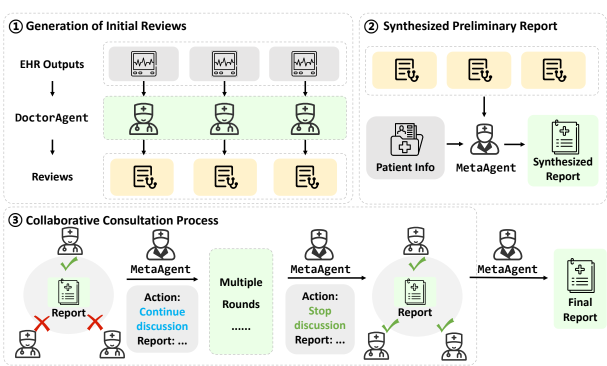

Figure 1: ColaCare*'s multi-agent collaborative consultation.*

The Web has become an indispensable platform for facilitating the integration and analysis of Electronic Health Records (EHR) data, playing a pivotal role in prognosis prediction and clinical decision-making. By leveraging web-based technologies, EHR modeling not only enhances data accessibility but also fosters the development of data-driven healthcare advancements [12]. In recent years, deep learning has achieved remarkable success in EHR modeling, particularly through structured data repositories and interoperable platforms that are part of the broader web ecosystem [29, 31, 47]. However, these efforts have primarily relied on purely data-driven, end-to-end methods that operate independently of external, web-based semantic knowledge. Consequently, these models often fail to fully capture the clinical significance of the recorded features, treating them as mere variables without a comprehensive semantic context [49]. This limitation hinders the real-world applicability of such "black box" methods, as they lack interpretability and do not leverage the rich, interlinked knowledge available on the Web, making it difficult for human physicians to trust them in real-world clinical practice. *(See Appendix A for more WWW relevance details.)*

Most interpretability models primarily rely on traditional feature importance analysis techniques, such as attention mechanisms [30], SHAP (SHapley Additive exPlanations) [38], and activation-level visualizations [43]. While these techniques provide a basic level of interpretability, they still fall short of aligning models with real-world knowledge. Several existing methods attempt to incorporate external knowledge to improve interpretability. For instance, some approaches embed knowledge from external knowledge graphs [4, 28], or construct knowledge graphs from patient sequential visit data [20]. However, deploying these methods in practical settings remains challenging due to their dependence on manually crafted knowledge representations and the slow pace of knowledge updates, which often fail to keep pace with the latest medical research or updated guidelines. Moreover, physicians increasingly require deeper insights into the reasoning behind deep learning models for interpretability purposes.

Given the impressive capabilities of Large Language Models (LLMs) in handling general tasks [1], including medical applications [37] such as medical question-answering (Q&A) tasks that reason over unstructured clinical notes [2, 17, 40], we are motivated to explore their potential for enhancing structured EHR modeling, which remains less explored. Specifically, we aim to meet the requirements of knowledge fusion and interpretable reasoning that address the limitations of conventional methods. We have further identified and summarized the key limitations of existing LLM-driven works in structured EHR modeling as follows:

(1) *Structured EHR comprehension*: Although some works have demonstrated LLMs' reasoning ability in structured EHR analysis under few-shot settings [15], there remains a notable performance gap compared to conventional methods [48]. While LLMs excel at mining natural language contexts, their ability to analyze or make predictions based on structured EHR data is still limited [48].
(2) *Lack of medical reference*: While recent research has expanded beyond single-model approaches by introducing multiple LLMs to address complex problems [2, 17, 40], most previous works lack trustworthy medical guidelines for clinical decision-making.

These continuously updated guidelines are crucial when the inherent knowledge within LLMs is insufficient for diagnosis.

To address these challenges, we propose ColaCare, a Retrieval-Augmented Generation (RAG)-enhanced framework inspired by the Multidisciplinary Team (MDT) approach employed by physicians in clinical settings [35]. ColaCare integrates domain-specific EHR modeling methods (expert models) into LLM-driven agents that fulfill two roles: doctor agents (referred to as DoctorAgent) and a meta doctor agent (referred to as MetaAgent). The MetaAgent orchestrates MDT discussions, facilitating consultations and evidence-based debates among DoctorAgents. Overall, ColaCare simulates the collaborative decision-making process among physicians with diverse expertise by leveraging LLMs' reasoning and role-playing capabilities, alongside the strengths of expert models in processing and predicting from structured EHR data.

ColaCare has the potential to revolutionize clinical decision-making and advance the field of personalized precision medicine. Our primary contributions are summarized as follows:

- **Insightfully**, we introduce the MDT approach to EHR modeling by incorporating external knowledge via RAG, enabling a prediction process that is both driven by EHR data and enriched with up-to-date external knowledge, complemented by self-examination.
- **Technically**, we develop ColaCare, which outputs clinical decision evidence from multiple DoctorAgents and the MetaAgent. These agents may hold differing opinions, thereby enhancing transparency and providing human-understandable evidence to support physicians in their diagnostic reasoning.
- **Experimentally**, we conduct extensive experiments on three EHR datasets, demonstrating the superior performance of ColaCare in predicting clinical outcomes such as mortality and readmission, with relative improvements on AUPRC of 0.86%, 2.50%, 2.00% and 4.49% on four tasks. Case studies highlight the reasonableness and interpretability of the reports generated by ColaCare, offering healthcare professionals detailed and understandable insights into each prediction.

2 RELATED WORK

2.1 LLMs in Medical Tasks

Large Language Models (LLMs) have demonstrated significant success in the medical domain, particularly in medical question-answering (Q&A) tasks [3, 10, 44]. They have also excelled in medical evidence summarization [39, 41]. Notably, the advanced LLM GPT-4 [1] has outperformed medical students on standard medical board exams [23]. While these achievements primarily involve textual clinical notes, recent research has begun exploring LLMs' capabilities in handling structured Electronic Health Record (EHR) data. Approaches include prompting LLMs directly [15, 48] or ensembling machine learning models' outputs [13, 19]. These studies reveal that GPT-4 shows potential for zero-shot prediction on structured EHR data, although a significant performance gap remains compared to conventional deep learning methods trained on full datasets [48]. Additionally, LLM's direct outputs face challenges with hallucination, where generated content may not strictly adhere to instructions, resulting in unexpected outputs beyond numerical values [45].

2.2 LLM-Driven Multi-Agent Collaboration in Medical Field

The development of LLM-driven agent systems, where multiple agents with distinct roles collaborate and utilize external tools [11], has garnered increasing attention in medical domains. Examples include AI Hospital [9] and Agent Hospital [25], which simulate real hospital environments through collaboration among several agents. Recent studies have also explored adversarial collaboration, incorporating debates and negotiations among multiple agents. MedAgents [40] proposes a medical collaboration framework where doctor agents vote on diagnoses, while ArgMed-Agents [17] constructs conflict relationship graphs and employs formal deduction to generate coherent Q&A conclusions. Further advancements include ReConcile [2], which employs a confidence-weighted voting mechanism for better consensus.

In summary, most prior work focuses on medical Q&A tasks, with LLMs generating text-based diagnoses. However, real-world scenarios often require specific numerical outputs, like disease mortality risk predictions. While approaches like Multi-Agent Debate (MAD) [7, 14, 26] exist, current implementations often use multiple instances of the same LLM, leading to homogeneous reasoning. This falls short of clinical requirements for diverse diagnostic reasoning. Therefore, there is a need for frameworks that better integrate structured EHR and address complex, quantitative medical tasks.

3 PROBLEM DEFINITION

3.1 EHR Datasets Formulation

The EHR datasets are structured as multivariate time-series data with multiple features, denoted as $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T]^\top \in \mathbb{R}^{T \times F}$, encapsulating information across T visits and F features, which includes static features (e.g., sex and age) and dynamic features (e.g., lab tests and vital signs).

3.2 Predictive Objective Formulation

The prediction task is defined as a binary classification problem aimed at predicting patients' mortality outcomes or 30-day readmission. Our goal is to extract knowledge from EHR data, supplemented by auxiliary external medical knowledge (e.g., medical guidelines), to enhance predictive modeling of EHRs. Thus, the predictive objective is formulated as:

$$\hat{y} = \texttt{Framework}(\boldsymbol{x}_{EHR}, \mathit{MedicalKnowledge}) \tag{1}$$

where $\hat{y}$ represents the predicted outcome.

For the mortality prediction task, the outcome $\hat{y}$ is a binary variable where 0 indicates the patient is alive, and 1 indicates the patient is deceased. In the case of the readmission prediction task, the model predicts whether the patient will be readmitted within 30 days of discharge, with 0 representing no readmission and 1 indicating readmission within the specified time frame.

4 METHODOLOGY

Fig. 1 and Fig. 2 demonstrates our proposed LLM-based multi-agent collaboration ColaCare framework, where the key module: Multi-Agent Collaborative Consultation Module is detailedly illustrated in Fig. 1. It gathers a group of doctor agents and a meta doctor agent

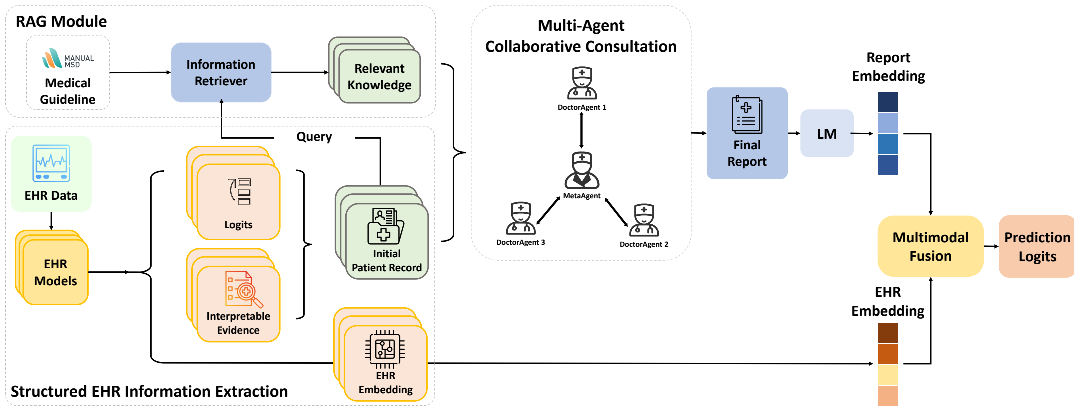

Figure 2: *Overall architecture of our proposed* `ColaCare` *framework.*

for medical discussion about a certain patient's condition. Doctor agents reach an agreement after multiple rounds of debate and the meta doctor agent proposes a report finally.

4.1 Structured EHR Information Extraction

Given the EHR data of a patient, $\boldsymbol{x}_{EHR}$, we utilize the EHR model, denoted as Model, to encode this temporally linked information:

$$\boldsymbol{h}_{EHR} = \mathrm{Model}(\boldsymbol{x}_{EHR}) \tag{2}$$

Then we use a MLP layer to predict the output logit z and the SHAP strategy to obtain feature importance weights $\boldsymbol{\alpha}$:

$$\begin{aligned} z &= \mathrm{MLP}(\boldsymbol{h}_{EHR}) \\ \boldsymbol{\alpha} &= \mathrm{SHAP}(\mathrm{Model}, \boldsymbol{x}_{EHR}) \end{aligned} \tag{3}$$

The prediction logit z and weights $\boldsymbol{\alpha}$ by deep learning models are used to support the LLM-based collaborative consultation.

4.2 Multi-Agent Collaborative Consultation

As shown in Fig. 1, in the multi-agent collaborative consultation, we define two distinct roles: the doctor agent (`DoctorAgent`) and the meta doctor agent (`MetaAgent`). Each `DoctorAgent` is linked to a domain-specific expert model.

The collaboration begins with each `DoctorAgent` providing an initial review of a patient's condition. Subsequently, the `MetaAgent` synthesizes these reviews to generate a comprehensive report and orchestrates the collaborative consultation process. During this iterative process, all `DoctorAgents` express their opinions on the current report. The `MetaAgent` then considers the feedback, accordingly revises the report, and determines whether more rounds of consultation are necessary. Fig. 1 demonstrates the overview pipeline of the collaborative consultation.

4.2.1 Generation of Initial Reviews. We first utilize output logits z and feature importance weights $\boldsymbol{\alpha}$ of EHR models, as well as basic information of a certain patient (e.g. Sex and Age), to build an initial patient record $\boldsymbol{x}_{record}$:

$$\boldsymbol{x}_{record} = Prompt(z, \boldsymbol{\alpha}, Info_{\mathrm{patient}}) \tag{4}$$

Next, the retriever measures the cosine similarity between embeddings of documents in the corpus and the patient record, and select the top-K relevant documents:

$$\bigcup Docs = Retriever(\boldsymbol{x}_{record}) \tag{5}$$

Lastly, we instruct the `DoctorAgent` to generate a review:

$$\boldsymbol{r}_{doctor} = \mathtt{DoctorAgent}(Prompt(\boldsymbol{x}_{record}, \bigcup Docs)) \tag{6}$$

> ***Prompt Template of*** `DoctorAgent`***'s Initial Review:***
> 1. **Here is the relevant evidence** ($\bigcup$Docs):
> Document [0] (Title: Diabetic Nephropathy - Genitourinary Disorders ...
> Document [1] ......
> ...
> 2. **Here is the patient record, including the patient's basic information and analysis results of AI models** ($\boldsymbol{x}_{record}$):
> **(**$Info_{patient}$**)** This male patient, aged 65.73, is an End-Stage Renal Disease (ESRD) patient with original disease Diabetic Nephropathy, and basic disease Diabetes ...
> **(Logit** z**)** The mortality prediction risk for the patient from AdaCare model is **0.01 out of 1.0**.
> **(Feature** $\boldsymbol{\alpha}$**)** We pay great attention to these features: Carbon dioxide binding power: with shap value of 0.165. The feature value ...
> 3. **You need to analyze the patient's condition based on the above information and generate an analytical review** ($\boldsymbol{r}_{meta}$).

4.2.2 Synthesized Preliminary Report. The `MetaAgent` is instructed to leverage the patient's basic information and initial reviews from all `DoctorAgents` to generate a synthesized report:

$$\boldsymbol{r}_{meta} = \mathtt{MetaAgent}(Prompt(\sum \boldsymbol{r}_{doctor}, Info_{patient})) \tag{7}$$

> ***Prompt Template of*** `MetaAgent`***'s Synthesized Report:***
> 1. **First, please read the patient's basic information carefully** ($Info_{patient}$):
> This male patient, aged 65.73, is an End-Stage Renal Disease (ESRD) patient with original disease Diabetic Nephropathy, and basic disease Diabetes.
> 2. **All doctors made a diagnosis on the patient's condition**

and gave their reasons as follows ($\sum \boldsymbol{r}_{doctor}$):
Doctor 0: The mortality risk of the patient is: 0.01. The patient's basic condition is: a 65-year-old male with End-Stage Renal Disease due to ... The retrieved evidence: ...
Doctor 1: ...
3. **You need to consider all doctors' opinions carefully and write a synthesized report ($\boldsymbol{r}_{meta}$).**

For the report, the `MetaAgent` is instructed to first assess and articulate the patient's mortality risk, categorizing it as either high or low and then select and incorporate pertinent comments and supporting evidence from the `DoctorAgents`' reviews.

4.2.3 Collaborative Consultation Process. The `DoctorAgents` are instructed to evaluate their initial assessments and state their agreement or disagreement with the current report. In cases of disagreement, they are required to provide detailed rationales and support their arguments with relevant documents retrieved from the corpus:

$$\boldsymbol{r}^{j}_{doctor} = \texttt{DoctorAgent}(Prompt(\boldsymbol{r}^{j-1}_{doctor}, \boldsymbol{r}^{j-1}_{meta}, \bigcup Docs)) \quad (8)$$

where j stands for the j-th round in the process.

Prompt Template of `DoctorAgents`' Collaboration:
1. **Here is your initial review, you may adjust it based on the meta doctor's report ($\boldsymbol{r}^{j-1}_{doctor}$) :**
The patient's basic condition is: a 65-year-old male with End-Stage Renal Disease due to ...
2. **Here is the synthesized report generated by the meta doctor ($\boldsymbol{r}^{j-1}_{meta}$):**
(1) In my opinion, the patient has a high risk of mortality.
(2) The patient exhibits significant clinical markers of severe renal dysfunction and metabolic imbalances......
3. **Here is the relevant document newly retrieved by the retriever module ($\bigcup Docs$):**
Document [0] (Title: Hemodialysis ...
Document [1] ...
4. **You need to consider the synthesized report carefully and provide your own opinions ($\boldsymbol{r}^{j}_{doctor}$).**

Subsequently, the `MetaAgent` aggregates and analyzes the feedback from all `DoctorAgents` to determine whether further discussion is necessary:

$$Action^{j} = \texttt{MetaAgent}(Prompt(\sum \boldsymbol{r}^{j-1}_{doctor})) \quad (9)$$

The `MetaAgent` evaluates the statements of the `DoctorAgents`, focusing on their agreement or disagreement with the current report. If unanimous agreement is reached, the `MetaAgent` concludes that further discussion is unnecessary. However, in cases of disagreement, the `MetaAgent` conducts a more detailed analysis by carefully examining the evidence presented by the dissenting `DoctorAgents` and assessing the validity and relevance of this evidence within the context of the current report. If any opposing statements and their supporting evidence are considered meritorious, the `MetaAgent` continues the discussion.

Prompt Template of `MetaAgent`'s Next Action:
1. **Several doctors put forward their own opinions and reasons ($\sum \boldsymbol{r}^{j-1}_{doctor}$):**
Doctor 0's statement is: I **agree with** the meta doctor. The reason is: the patient's condition, as described in the report, aligns with the high risk...
Doctor 1's statement is: I **disagree with** the meta doctor. The reason is ... The evidence is ...
2. **Next, you need to judge whether the next round of discussion is needed based on each doctor's statement** ($Action^{j}$).

If the action infers a new round, the `MetaAgent` considers the opinions and relevant documents by all `DoctorAgents` and refine the report:.

$$\boldsymbol{r}^{j}_{meta} = \texttt{MetaAgent}(Prompt(\boldsymbol{r}^{j-1}_{meta}, \sum \boldsymbol{r}^{j-1}_{doctor})) \quad (10)$$

Once again, `DoctorAgents` express their statements towards the new report and try to convince the `MetaAgent`. The `MetaAgent` considers their statements and further revises the report, until all `DoctorAgents` reach an agreement or the `MetaAgent` calls it a day.

Prompt Template of `MetaAgent`'s Revised Report:
1. **In the previous discussion, you gave the report ($\boldsymbol{r}^{j-1}_{meta}$):**
The patient has a high risk of mortality. Key concerns include significantly low systolic and diastolic blood pressures...
2. **All the doctors offered new perspectives ($\sum \boldsymbol{r}^{j-1}_{doctor}$):**
Doctor 0's statement is: the mortality risk of the patient is 0.15. The patient shows a low mortality risk of ... The supporting evidence is: Document [3] (Title: Anemia of Renal Disease ... Document ...
Doctor 1's statement is: ......
3. **You need to consider all the doctors' new ideas and modify your original report ($\boldsymbol{r}^{j}_{meta}$).**

4.3 Multimodal Fusion Network

We begin by leveraging hidden representations of EHR data from previous N EHR-specific expert models, denoted as $\boldsymbol{h}^{i}_{EHR}$. Subsequently, we utilize a medical-domain pretrained language model, LM, to encode the final consensus report embedding:

$$\boldsymbol{h}_{Report} = \text{LM}(\boldsymbol{x}_{Report}) \quad (11)$$

We then concatenate the representations from both modalities and apply an MLP layer to obtain the final prediction, $\hat{\boldsymbol{y}}$:

$$\hat{\boldsymbol{y}} = \text{MLP}\left(\text{Concat}\left[\boldsymbol{h}^{1}_{EHR}, \boldsymbol{h}^{2}_{EHR}, \ldots, \boldsymbol{h}^{N}_{EHR}, \boldsymbol{h}_{Report}\right]\right) \quad (12)$$

The loss function employed is the Binary Cross-Entropy (BCE) Loss for binary classification:

$$\mathcal{L}(\hat{\boldsymbol{y}}, \boldsymbol{y}) = -\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{y}_i \log(\hat{\boldsymbol{y}}_i) + (1 - \boldsymbol{y}_i)\log(1 - \hat{\boldsymbol{y}}_i)) \quad (13)$$

where N represents the number of patients in a batch, $\hat{\boldsymbol{y}} \in [0, 1]$ is the predicted probability, and $\boldsymbol{y}$ is the ground truth.

5 EXPERIMENTAL SETUPS

5.1 Experimented Datasets and Utilized Medical Guideline

We adopt three real-world datasets: MIMIC-IV, CDSL, and ESRD datasets and additionally introduce the MSD (Merck Manual of Diagnosis and Therapy) medical guideline to enhance our experimental framework.

- **MIMIC-IV** [22]: MIMIC-IV is part of the Medical Information Mart for Intensive Care project. It contains comprehensive EHR data for intensive care unit patients, including demographic information, vital signs, laboratory results, procedures, medications, clinical notes, and mortality statistics. For our study, we focus on the clinical notes, demographic data, and laboratory test features.
- **CDSL** [18]: This dataset is derived from the HM Hospitales EHR system and consists of anonymized records of 4,479 patients admitted with a confirmed or suspected diagnosis of COVID-19. CDSL offers a rich variety of medical features, including comprehensive details on diagnoses, treatments, admissions, ICU stays, diagnostic imaging tests, laboratory results, and patient discharge or death status.
- **ESRD** [30]: The end-stage renal disease (ESRD) dataset comprises data from 656 patients, including 13,091 visit records collected over a 12-year period, from January 1, 2006, to January 1, 2018. This longitudinal dataset features patients' baseline information, visit records, and clinical outcomes.

We incorporate the Merck Manual of Diagnosis and Therapy (MSD) medical guideline [36][1] into our framework `ColaCare`. The MSD guideline is a comprehensive medical reference that provides detailed information on various diseases, corresponding diagnoses, and treatment protocols.

We adhere to the established EHR benchmark pipeline [12, 16, 30] for preprocessing time-series data, ensuring consistency and comparability in our data preparation across all datasets. The test set size is limited to approximately 1,000 samples due to the time cost associated with each sample. We suggest a sample of 1,000 patients provides sufficient representation for benchmarking and evaluation, offering diverse, information-rich records that effectively test model generalization in realistic healthcare scenarios [48, 50]. The statistics of dataset splits and label distributions for the three datasets are presented in Tab. 1.

Table 1: *Statistics of the experimented datasets after preprocessing.* The # Samples column shows the number of samples and their percentage of the entire dataset, indicating data splits (train, val, test). The # Label$_{Out.}$ = 1 and # Label$_{Re.}$ = 1 columns provides the count and percentage of patients with adverse outcomes within each data split. "Out." denotes "mortality outcome", "Re." denotes "Readmission".

Dataset	Split	# Samples	# Label$_{Out.}$ = 1	# Label$_{Re.}$ = 1
MIMIC-IV	Train	17,397 (90.00%)	2,067 (11.88%)	2,685 (15.43%)
	Val	966 (5.00%)	118 (12.22%)	157 (16.25%)
	Test	968 (5.01%)	115 (11.88%)	153 (15.81%)
CDSL	Train	2,127 (49.98%)	270 (12.69%)	-
	Val	1,064 (25.01%)	135 (12.69%)	-
	Test	1,064 (25.01%)	135 (12.69%)	-
ESRD	Train	379 (57.77%)	157 (41.42%)	-
	Val	131 (19.97%)	47 (35.88%)	-
	Test	146 (22.26%)	57 (39.04%)	-

5.2 Evaluation Metrics

We employ three widely-used evaluation metrics for binary classification tasks, all of which are interpreted as "higher is better":

- **AUROC, AUPRC**: These complementary metrics assess model performance across various classification thresholds. AUROC is valued in clinical settings [33], while AUPRC is particularly useful for imbalanced datasets [24]. Together, they provide a comprehensive view of the model's discriminative ability.
- **min(+P, Se)**: This metric takes the minimum value between precision (+P) and sensitivity (Se), offering a balanced assessment of model performance by correctly identifying positive cases and minimizing false positives [32].

5.3 Baseline Models

5.3.1 EHR-specific Baselines. We include the following established EHR-specific deep learning-based models as baselines: AdaCare [29], ConCare [31], and RETAIN [5]. These models employ various architecture, such as attention mechanisms, feature extraction, and recalibration, to address different aspects of EHR data analysis and patient health representation.

Additionally, we implement two ensemble-based approaches: a simple mean, which averages the predictions from AdaCare, ConCare, and RETAIN, and a weighted mean, which assigns weights to each model's prediction based on AUPRC score.

5.3.2 LLM-based Baselines. We incorporate baselines utilizing single or multiple LLMs. For those using a single LLM, we employ three different prompt strategies: zero-shot, few-shot, and self-consistency [42]. Additionally, we include three multi-agent collaboration approaches: multi-agent debate [26], MedAgents [40], and ReConcile [2]. All of these LLM-based baselines directly process EHR data as input, relying on large language models to interpret the data and generate final prediction results.

5.4 Implementation Details

5.4.1 Hardware and Software Configuration. All experiments are conducted on a single Nvidia RTX 3090 GPU with CUDA 12.5. The server's RAM size is 128GB. We implement the model in Python 3.9.19, PyTorch 2.3.1 [34], PyTorch Lightning 2.3.3 [8].

5.4.2 Model Training and Hyperparameters. AdamW [27] optimizer is employed with a batch size of 128 patients. All models are trained for 50 epochs with an early stopping strategy based on AUPRC after 10 epochs without improvement.

For EHR-specific baseline models, the learning rate $\{0.01, 0.001, 0.0001\}$ and hidden dimensions $\{64, 128\}$ are tuned using a grid search strategy on the validation set. For LLM-based baselines, the few-shot approach simulates two sample examples, one positive and one negative, following the approach in previous work [48]. The self-consistency method selects three possible reasoning paths, followed by a consistency evaluation.

For `ColaCare`, the maximum number of rounds for the collaborative consultation process is set to 3. The fusion network's hidden dimension is 128 with a learning rate of 0.001. The K for the retrieval process is set to 3. Performance is reported as mean±std, calculated by bootstrapping all test set samples 100 times for all three datasets. Experiments are conducted from June 30th, 2024, to October 15th, 2024.

5.4.3 Utilized (Large) Language Models. `ColaCare` incorporates both Language Models (LMs) and Large Language Models (LLMs) within its framework. For language models, we employ MedCPT [21]

[1]Access link: https://www.msdmanuals.com/professional

Table 2: *Overall performance of in-hospital mortality and 30-day readmission prediction results on MIMIC-IV, CDSL, and ESRD datasets.* Bold indicates the best performance. Performance is reported in the form of mean±std. All metric scores are multiplied by 100 for readability purposes.

Methods	MIMIC-IV Outcome			MIMIC-IV Readmission			CDSL Outcome			ESRD Outcome		
	AUPRC (↑)	AUROC (↑)	min(+P, Se) (↑)	AUPRC (↑)	AUROC (↑)	min(+P, Se) (↑)	AUPRC (↑)	AUROC (↑)	min(+P, Se) (↑)	AUPRC (↑)	AUROC (↑)	min(+P, Se) (↑)
AdaCare	52.67±4.50	87.56±1.37	53.62±3.91	50.51±4.24	78.82±1.94	50.01±3.37	82.76±3.74	95.54±1.02	77.20±3.23	60.44±6.94	69.11±5.53	59.23±6.46
ConCare	49.71±4.83	87.21±1.41	52.96±3.74	46.66±4.05	79.06±1.82	47.00±3.20	80.77±3.59	94.47±1.24	74.17±2.80	56.65±6.91	68.84±4.85	58.47±5.78
RETAIN	51.89±4.22	87.87±1.27	49.71±3.83	50.89±3.90	80.73±1.82	50.42±3.33	74.46±4.13	93.32±1.30	70.09±3.30	60.02±6.32	70.82±4.32	59.26±5.41
$\text{Ensemble}_{\text{Mean}}$	53.59±4.52	88.60±1.26	54.08±3.66	50.48±4.68	86.97±1.45	50.36±3.58	83.54±3.36	95.63±1.04	76.24±3.48	60.90±7.08	73.41±4.80	59.98±6.02
$\text{Ensemble}_{\text{WeightedMean}}$	53.44±4.51	88.58±1.26	53.98±3.72	50.64±4.69	87.05±1.45	**50.68±3.63**	84.08±3.34	95.88±0.98	76.48±3.42	61.07±7.11	**73.49±4.81**	60.09±5.99
$\text{LLM}_{\text{ZeroShot}}$	31.01±3.68	75.20±2.67	34.39±3.71	21.63±1.89	63.98±2.08	26.47±2.71	29.71±2.82	78.80±2.12	33.39±2.63	6.71±2.88	49.93±12.30	7.72±3.28
$\text{LLM}_{\text{FewShot}}$	26.28±2.91	70.94±2.30	40.39±3.74	29.09±3.25	65.72±1.83	34.43±3.62	41.58±4.13	78.61±1.83	42.88±4.53	2.78±2.54	54.54±38.33	2.78±2.54
$\text{LLM}_{\text{SelfConsistency}}$	27.39±3.38	72.36±2.56	30.77±3.55	24.73±2.33	62.84±2.46	27.95±2.79	36.24±4.09	75.37±2.07	32.44±3.96	9.09±4.44	58.57±10.70	11.60±6.06
MAD	30.95±5.59	78.43±3.09	29.62±4.48	24.33±2.66	65.09±2.47	24.14±2.38	42.53±5.38	78.55±2.94	40.02±5.05	5.14±2.63	52.47±18.77	6.59±3.47
MedAgents	27.73±3.38	74.38±2.19	27.80±3.31	20.15±2.46	58.25±2.66	19.39±2.29	33.02±3.34	81.21±1.75	33.48±3.12	4.47±2.21	53.64±18.31	5.26±2.80
ReConcile	27.91±3.50	74.51±2.43	29.91±3.99	24.00±2.86	65.39±2.35	25.45±3.11	35.04±3.17	80.39±1.87	34.31±2.86	17.37±13.54	62.20±20.94	24.62±16.54
`ColaCare`	**54.05±4.88**	**88.80±1.31**	**54.44±3.80**	**51.91±4.99**	**87.72±1.37**	50.16±4.17	**85.76±2.99**	**96.23±0.91**	**78.93±3.21**	**63.81±6.43**	71.61±4.77	**63.33±5.54**

in the RAG system's biomedical information retrieval process. MedCPT is contrastively pre-trained on 255 million PubMed literature entries. Additionally, we utilize the GatorTron [46] model, which is pretrained on 82 billion de-identified clinical texts for final clinical note embedding computation. For large language models, we employ the DeepSeek-V2.5 [6] to act as the reasoning engine.

6 EXPERIMENTAL RESULTS AND ANALYSIS

This section evaluates the `ColaCare` framework by addressing the following research questions (**RQs**):

(1) **RQ1: Overall Performance** How does `ColaCare` compare to other EHR-specific deep learning models and LLM-based frameworks in clinical downstream tasks?
(2) **RQ2: Ablation Study** What is the contribution of each proposed module to the performance?
(3) **RQ3: Sensitivity to Number of Agents** How does the number of agents affect `ColaCare`'s performance?
(4) **RQ4: Sensitivity to Different LLMs** How does `ColaCare`'s performance vary with different LLMs?
(5) **RQ5: Case Study** Does `ColaCare` generate reliable clinical reports and summaries for interpretability?
(6) **RQ6: Cost Analysis** What are the implementation costs of `ColaCare` in clinical practice, considering token usage, etc., across datasets and tasks?

6.1 RQ1: Overall Performance

To address RQ1, we conduct mortality and readmission prediction tasks on the MIMIC-IV dataset and mortality prediction tasks on the CDSL and ESRD datasets. The overall performance of `ColaCare` is shown in Tab. 2. The results demonstrate that `ColaCare` consistently outperforms all baseline models in most cases, with notable improvements in the AUPRC metric. Specifically, `ColaCare` surpasses each expert model, the ensemble results of expert models (as `ColaCare` can be viewed as an LLM-based ensemble method), and LLM-driven approaches that are instructed to directly output prediction results. This superior performance highlights `ColaCare`'s potential for practical application in clinical decision-making.

6.2 RQ2: Ablation Study

For RQ2, we assess each module's impact on performance (see Tab. 3) by testing three variants: (1) without the RAG module, where relevant documents are not provided to agents; (2) without the fusion network, which relies on the LLM by prompting it to produce prediction results based on the final report; and (3) without integrating expert models, where information from EHR-specific domain models is omitted and the LLM is prompted to generate prediction results based on raw EHR data, replicating the MedAgents approach. The results indicate that excluding expert models significantly reduces performance. Additionally, the direct output approach of LLMs proves ineffective, as evidenced by the performance drop when the fusion network is removed and in comparison to other LLM-based baseline methods. Furthermore, incorporating the RAG system enhances overall performance.

Table 3: *Ablation study results for each module.* Bold indicates the best performance. Performance is reported in the form of mean±std. All metric scores are multiplied by 100 for readability purposes.

Methods	MIMIC-IV Outcome			MIMIC-IV Readmission		
	AUPRC (↑)	AUROC (↑)	min(+P, Se) (↑)	AUPRC (↑)	AUROC (↑)	min(+P, Se) (↑)
w/o RAG	51.57±4.76	88.09±1.30	49.25±3.91	48.00±4.58	86.74±1.63	**51.43±3.85**
w/o Fusion Network	42.61±4.02	87.53±1.27	49.72±3.69	49.03±4.38	79.85±1.90	51.26±3.30
w/o Expert Models	27.73±3.38	74.38±2.19	27.80±3.31	20.15±2.46	58.25±2.66	19.39±2.29
`ColaCare`	**54.05±4.88**	**88.80±1.31**	**54.44±3.80**	**51.91±4.99**	**87.72±1.37**	50.16±4.17

6.3 RQ3: Sensitivity to Number of Agents

For RQ3, we explore the impact of the number of agents participating in the collaborative consultation, as shown in Tab. 4. Our findings indicate that having one or two `DoctorAgents` has little to no impact or slightly decreases performance. This is because a small number of `DoctorAgents` can easily reach a consensus, even if their final opinion lacks robustness. As the number of `DoctorAgents` increases, performance improves due to the incorporation of a wider range of perspectives and medical evidence, resulting in more comprehensive and reliable reports.

6.4 RQ4: Sensitivity to Different LLMs

To address RQ4, we evaluate the performance of `ColaCare` when utilizing different LLMs as the reasoning engine. Specifically, we compare `ColaCare` instantiated with DeepSeek-V2.5, GPT-4o-Mini, GPT-4o, Qwen-Turbo, and Doubao-Pro. The performance presented in Tab. 5 shows that all these LLMs can reason with EHR data, with DeepSeek-V2.5 and GPT-4o-Mini performing slightly better.

Table 4: ***Performance of different numbers of agents in the in-hospital mortality prediction task on MIMIC-IV datasets.*** **Bold indicates the best performance. Performance is reported in the form of mean±std. All metric scores are multiplied by 100 for readability purposes.**

#Agents	Models			Metric		
	AdaCare	ConCare	RETAIN	AUPRC (↑)	AUROC (↑)	min(+P, Se)
0	✓	-	-	52.67±4.50	87.56±1.37	53.62±3.91
	-	✓	-	49.71±4.83	87.21±1.41	52.96±3.74
	-	-	✓	51.89±4.22	87.87±1.27	49.71±3.83
1	✓	-	-	52.96±4.54	87.59±1.34	53.79±3.86
	-	✓	-	50.55±4.67	87.24±1.42	52.78±3.56
	-	-	✓	50.44±4.61	87.31±1.34	49.39±4.02
2	✓	✓	-	52.61±4.85	87.96±1.33	53.96±3.8
	✓	-	✓	51.24±4.60	86.58±1.53	50.99±3.50
	-	✓	✓	52.14±4.74	87.91±1.34	51.98±3.94
3	✓	✓	✓	**54.05±4.88**	**88.80±1.31**	**54.44±3.80**

Table 5: ***Performance of different LLMs of in-hospital mortality and 30-day readmission prediction results on MIMIC-IV datasets.*** **Bold indicates the best performance. Performance is reported in the form of mean±std. All metric scores are multiplied by 100 for readability purposes.**

Methods	MIMIC-IV Outcome			MIMIC-IV Readmission		
	AUPRC (↑)	AUROC (↑)	min(+P, Se) (↑)	AUPRC (↑)	AUROC (↑)	min(+P, Se) (↑)
GPT-4o-Mini	**55.27±4.92**	88.66±1.38	**55.74±3.75**	51.01±5.05	85.00±1.88	**51.09±4.09**
GPT-4o	54.13±4.84	88.59±1.36	54.38±3.96	50.93±4.81	86.89±1.48	49.44±3.63
Qwen-Turbo	52.88±4.76	88.57±1.28	54.68±3.70	49.95±4.13	80.10±1.81	50.38±3.29
Doubao-Pro	52.14±4.65	88.04±1.34	53.37±3.69	49.70±4.04	79.92±1.80	50.00±3.33
ColaCare (DeepSeek-V2.5)	54.05±4.88	**88.80±1.31**	54.44±3.80	**51.91±4.99**	**87.72±1.37**	50.16±4.17

6.5 RQ5: Case Study

To evaluate whether ColaCare provides reliable clinical reports and summaries for interpretable analysis, we present a case study based on a patient from the ESRD dataset's mortality prediction task. Fig. 3 illustrates each step in the ColaCare pipeline during the collaborative consultation process.

In Step 1, each DoctorAgent is provided with the patient's records, including basic information and multivariate time-series EHR data, results from expert models such as mortality risk prediction logits, feature importance, and population-level statistics, as well as relevant retrieved documents. The DoctorAgents are then prompted to generate an initial review for the patient. DoctorAgent 1 assesses a moderate risk by focusing on the patient's carbon dioxide binding power and albumin levels. DoctorAgent 2 determines a low risk by concentrating on diastolic blood pressure and blood chlorine levels, both within normal ranges for patients with end-stage renal disease (ESRD). In contrast, DoctorAgent 3 identifies significantly low blood potassium levels, indicating a precarious condition. All DoctorAgents cite evidence from authoritative documents. In Step 2, the MetaAgent synthesizes a report based on the reviews from the three DoctorAgents, primarily integrating the analyses of DoctorAgent 1 and DoctorAgent 3 to conclude a high mortality risk and provide a comprehensive analysis. The MetaAgent highlights key factors such as carbon dioxide binding power, blood potassium, and albumin levels, all of which are abnormal and pose serious risks. Subsequently, in Step 3, each DoctorAgent reviews, votes on, and comments on the MetaAgent's synthesized report. DoctorAgent 1 and DoctorAgent 2 concur with the MetaAgent's perspective, revising their initial assessments and addressing previously overlooked critical factors. Upon reaching consensus among all DoctorAgents, the MetaAgent performs a final summary and delivers the ultimate report.

6.6 RQ6: Cost Analysis

To evaluate the practical viability of the ColaCare framework in real-world clinical settings, we conduct an analysis of its computational costs, as illustrated in Tab. 6. The cost analysis demonstrates that ColaCare is effective, particularly when leveraging domain-specific LLMs like DeepSeek-V2.5, making ColaCare a practical tool for integration into clinical workflows.

Table 6: ***Average time and token cost of* ColaCare *in in-hospital mortality and 30-day readmission prediction tasks on MIMIC-IV, CDSL, and ESRD datasets.***

Cost	MIMIC-IV Outcome	MIMIC-IV Readmission	CDSL Outcome	ESRD Outcome
Prompt Tokens	59,113.06	62,759.91	124,074.52	72,151.11
Output Tokens	4,435.91	4,570.47	5,456.79	6,156.30
# API Requests	8.56	8.37	10.39	11.86

7 LIMITATIONS AND FUTURE WORK

ColaCare shows promise in EHR modeling but has several limitations and areas for future exploration:

- **Generalizability:** We focus on mortality and readmission predictions using three datasets and models like DeepSeek-V2.5 and GPT-4o. Evaluating ColaCare on diverse clinical tasks and EHR datasets, and incorporating additional LLMs such as Claude, Gemini, LLaMA, and Qwen, will validate its generalizability.
- **Human Evaluation:** Although ColaCare generates interpretable reports reviewed by expert doctors, broader evaluation with a larger cohort of clinicians is necessary.
- **Continuous Learning:** Current EHR model parameters remain fixed after training. Developing mechanisms to continuously learn from LLM feedback and real-world clinical data can enhance its adaptability in dynamic healthcare environments.

8 CONCLUSION

This paper presents ColaCare, a framework that enhances EHR modeling through LLM-driven multi-agent collaboration. ColaCare delivers human-interpretable, personalized predictions with patient-specific evidence and can identify and rectify potential errors in predictions and evidence gathering. It integrates a knowledge-infused Multidisciplinary Team (MDT) ensemble within an LLM-driven multi-agent system, mirroring the collaborative and iterative nature of real-world clinical decision-making processes. Experiments on three real-world EHR datasets demonstrate ColaCare's superior performance in mortality and readmission predictions. ColaCare advances personalized precision medicine and has the potential to transform clinical decision support systems and healthcare delivery.

ETHICAL USE OF DATA

All EHR data used in this study are de-identified EHR datasets. We adhere to the data usage policy. The MIMIC-IV dataset is processed using secure Azure OpenAI API and human review of the data has been waived.

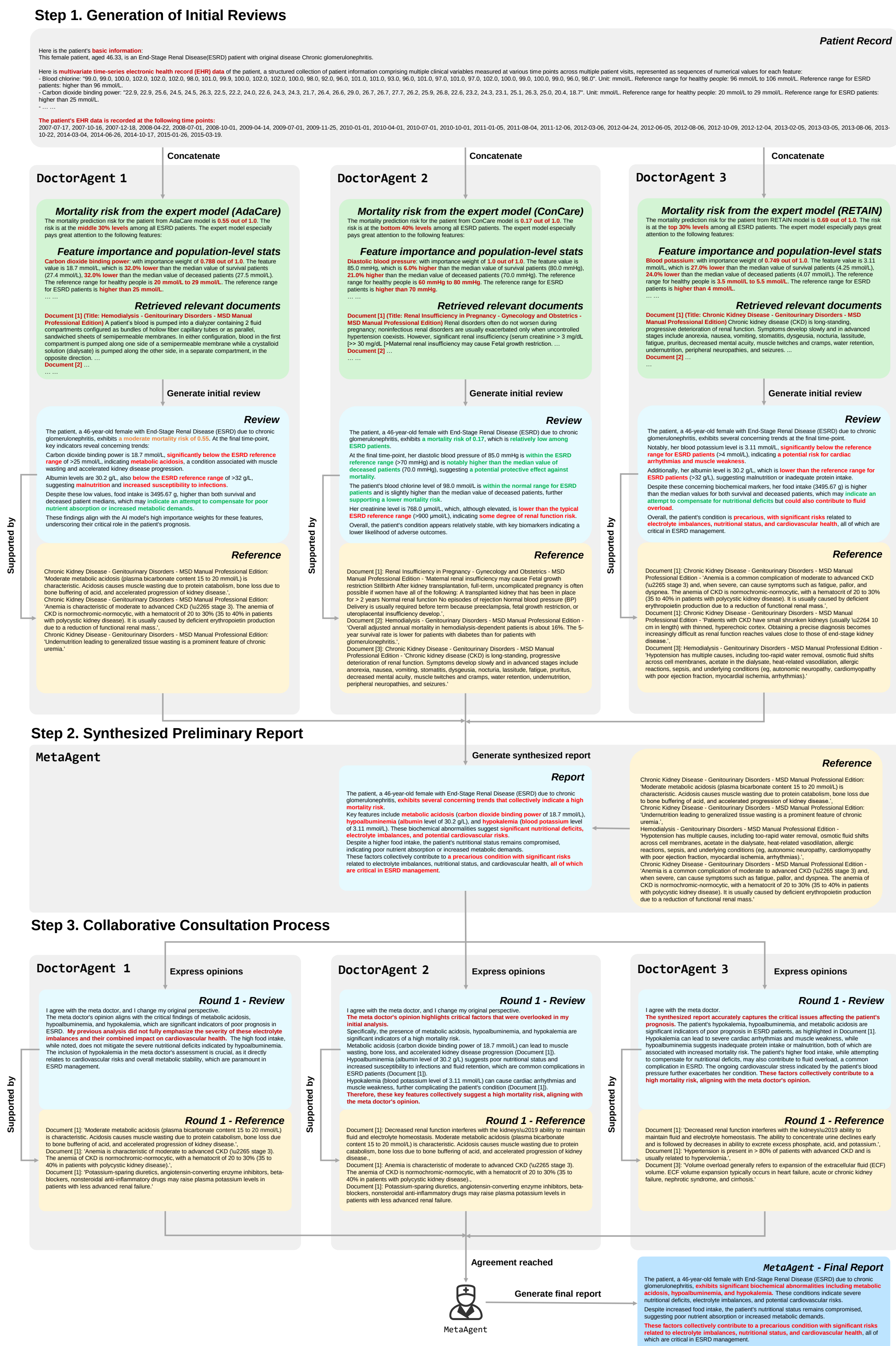

Figure 3: *The case of a patient from ESRD dataset, who deceased within one year after the last follow-up visit.* Important indicators are shown in dark red color. Diseases are shown in red color. Healthy indicators are shown in green color.

REFERENCES

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Justin Chen, Swarnadeep Saha, and Mohit Bansal. 2024. ReConcile: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 7066–7085. https://doi.org/10.18653/v1/2024.acl-long.381

[3] Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079* (2023).

[4] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. 2017. GRAM: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining.* 787–795.

[5] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems* 29 (2016).

[6] DeepSeek-AI. 2024. DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model. arXiv:2405.04434 [cs.CL]

[7] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325* (2023).

[8] William A Falcon. 2019. Pytorch lightning. *GitHub* 3 (2019).

[9] Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou. 2024. AI Hospital: Benchmarking Large Language Models in a Multi-agent Medical Interaction Simulator. arXiv:2402.09742 [cs.CL] https://arxiv.org/abs/2402.09742

[10] Hui Feng, Francesco Ronzano, Jude LaFleur, Matthew Garber, Rodrigo de Oliveira, Kathryn Rough, Katharine Roth, Jay Nanavati, Khaldoun Zine El Abidine, and Christina Mack. 2024. Evaluation of large language model performance on the Biomedical Language Understanding and Reasoning Benchmark. *medRxiv* (2024), 2024–05.

[11] Mehmet Fırat and Saniye Kuleli. 2023. What if GPT4 became autonomous: The Auto-GPT project and use cases. *Journal of Emerging Computer Technologies* 3, 1 (2023), 1–6.

[12] Junyi Gao, Yinghao Zhu, Wenqing Wang, Guiying Dong, Wen Tang, Hao Wang, Yasha Wang, Ewen M Harrison, and Liantao Ma. 2024. A Comprehensive Benchmark for COVID-19 Predictive Modeling Using Electronic Health Records in Intensive Care. *Patterns* (2024). https://doi.org/10.1016/j.patter.2024.100951

[13] Benjamin S Glicksberg, Prem Timsina, Dhaval Patel, Ashwin Sawant, Akhil Vaid, Ganesh Raut, Alexander W Charney, Donald Apakama, Brendan G Carr, Robert Freeman, Girish N Nadkarni, and Eyal Klang. 2024. Evaluating the accuracy of a state-of-the-art large language model for prediction of admissions from the emergency room. *Journal of the American Medical Informatics Association* (05 2024), ocae103. https://doi.org/10.1093/jamia/ocae103 arXiv:https://academic.oup.com/jamia/advance-article-pdf/doi/10.1093/jamia/ocae103/57728551/ocae103.pdf

[14] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680* (2024).

[15] Changho Han, Dong Won Kim, Songsoo Kim, Seng Chan You, Jin Young Park, SungA Bae, and Dukyong Yoon. 2024. Evaluation of GPT-4 for 10-year cardiovascular risk prediction: Insights from the UK Biobank and KoGES data. *Iscience* 27, 2 (2024).

[16] Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. 2019. Multitask learning and benchmarking with clinical time series data. *Scientific data* 6, 1 (2019), 96.

[17] Shengxin Hong, Liang Xiao, Xin Zhang, and Jianxia Chen. 2024. ArgMed-Agents: Explainable Clinical Decision Reasoning with Large Language Models via Argumentation Schemes. *arXiv preprint arXiv:2403.06294* (2024).

[18] HM Hospitales. 2020. Covid Data Save Lives. https://www.hmhospitales.com/prensa/notas-de-prensa/comunicado-covid-data-save-lives. Accessed: 2024-06-05.

[19] Danqing Hu, Bing Liu, Xiaofeng Zhu, and Nan Wu. 2024. The Power of Combining Data and Knowledge: GPT-4o is an Effective Interpreter of Machine Learning Models in Predicting Lymph Node Metastasis of Lung Cancer. *arXiv preprint arXiv:2407.17900* (2024).

[20] Pengcheng Jiang, Cao Xiao, Adam Richard Cross, and Jimeng Sun. 2024. GraphCare: Enhancing Healthcare Predictions with Personalized Knowledge Graphs. In *The Twelfth International Conference on Learning Representations.*

[21] Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. MedCPT: Contrastive Pre-trained Transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. *Bioinformatics* 39, 11 (2023), btad651.

[22] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data* 10, 1 (2023), 1.

[23] Uriel Katz, Eran Cohen, Eliya Shachar, Jonathan Somer, Adam Fink, Eli Morse, Beki Shreiber, and Ido Wolf. 2024. GPT versus Resident Physicians—A Benchmark Based on Official Board Scores. *NEJM AI* 1, 5 (2024), AIdbp2300192.

[24] Misuk Kim and Kyu-Baek Hwang. 2022. An empirical evaluation of sampling methods for the classification of imbalanced data. *PLoS One* 17, 7 (2022), e0271260.

[25] Junkai Li, Siyu Wang, Meng Zhang, Weitao Li, Yunghwei Lai, Xinhui Kang, Weizhi Ma, and Yang Liu. 2024. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957* (2024).

[26] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118* (2023).

[27] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[28] Fenglong Ma, Quanzeng You, Houping Xiao, Radha Chitta, Jing Zhou, and Jing Gao. 2018. Kame: Knowledge-based attention model for diagnosis prediction in healthcare. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management.* 743–752.

[29] Liantao Ma, Junyi Gao, Yasha Wang, Chaohe Zhang, Jiangtao Wang, Wenjie Ruan, Wen Tang, Xin Gao, and Xinyu Ma. 2020. AdaCare: Explainable Clinical Health Status Representation Learning via Scale-Adaptive Feature Extraction and Recalibration. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 01 (Apr. 2020), 825–832. https://doi.org/10.1609/aaai.v34i01.5427

[30] Liantao Ma, Chaohe Zhang, Junyi Gao, Xianfeng Jiao, Zhihao Yu, Yinghao Zhu, Tianlong Wang, Xinyu Ma, Yasha Wang, Wen Tang, Xinju Zhao, Wenjie Ruan, and Tao Wang. 2023. Mortality prediction with adaptive feature importance recalibration for peritoneal dialysis patients. *Patterns* 4, 12 (2023). https://doi.org/10.1016/j.patter.2023.100892

[31] Liantao Ma, Chaohe Zhang, Yasha Wang, Wenjie Ruan, Jiangtao Wang, Wen Tang, Xinyu Ma, Xin Gao, and Junyi Gao. 2020. Concare: Personalized clinical feature embedding via capturing the healthcare context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 833–840.

[32] Xinyu Ma, Yasha Wang, Xu Chu, Liantao Ma, Wen Tang, Junfeng Zhao, Ye Yuan, and Guoren Wang. 2022. Patient Health Representation Learning via Correlational Sparse Prior of Medical Features. *IEEE Transactions on Knowledge and Data Engineering* (2022).

[33] Matthew B. A. McDermott, Lasse Hyldig Hansen, Haoran Zhang, Giovanni Angelotti, and Jack Gallifant. 2024. A Closer Look at AUROC and AUPRC under Class Imbalance. *arXiv preprint arXiv:2401.06091* (2024).

[34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

[35] Brindha Pillay, Addie C Wootten, Helen Crowe, Niall Corcoran, Ben Tran, Patrick Bowden, Jane Crowe, and Anthony J Costello. 2016. The impact of multidisciplinary team meetings on patient assessment, management and outcomes in oncology settings: a systematic review of the literature. *Cancer treatment reviews* 42 (2016), 56–72.

[36] Robert S Porter and Justin L Kaplan. 2011. *The Merck manual of diagnosis and therapy.* Merck Sharp & Dohme Corp.

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

[38] Mukund Sundararajan and Amir Najmi. 2020. The many Shapley values for model explanation. In *International conference on machine learning.* PMLR, 9269–9278.

[39] Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F Rousseau, et al. 2023. Evaluating large language models on medical evidence summarization. *NPJ digital medicine* 6, 1 (2023), 158.

[40] Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning. In *Findings of the Association for Computational Linguistics ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand and virtual meeting, 599–621. https://doi.org/10.18653/v1/2024.findings-acl.33

[41] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine* 30, 4 (2024),

1134–1142.

[42] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171* (2022).

[43] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*. 3–19.

[44] Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, et al. 2024. Me-llama: Foundation large language models for medical applications. *arXiv preprint arXiv:2402.12749* (2024).

[45] Hao Xue and Flora D Salim. 2023. Promptcast: A new prompt-based learning paradigm for time series forecasting. *IEEE Transactions on Knowledge and Data Engineering* (2023).

[46] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, et al. 2022. GatorTron: a large clinical language model to unlock patient information from unstructured electronic health records. *arXiv preprint arXiv:2203.03540* (2022).

[47] Chaohe Zhang, Xin Gao, Liantao Ma, Yasha Wang, Jiangtao Wang, and Wen Tang. 2021. GRASP: Generic Framework for Health Status Representation Learning Based on Incorporating Knowledge from Similar Patients. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 1 (May 2021), 715–723. https://doi.org/10.1609/aaai.v35i1.16152

[48] Yinghao Zhu, Junyi Gao, Zixiang Wang, Weibin Liao, Xiaochen Zheng, Lifang Liang, Yasha Wang, Chengwei Pan, Ewen M Harrison, and Liantao Ma. 2024. Is larger always better? Evaluating and prompting large language models for non-generative medical tasks. *arXiv preprint arXiv:2407.18525* (2024).

[49] Yinghao Zhu, Changyu Ren, Zixiang Wang, Xiaochen Zheng, Shiyun Xie, Junlan Feng, Xi Zhu, Zhoujun Li, Liantao Ma, and Chengwei Pan. 2024. EMERGE: Integrating RAG for Improved Multimodal EHR Predictive Modeling. *arXiv preprint arXiv:2406.00036* (2024).

[50] Yinghao Zhu, Zixiang Wang, Junyi Gao, Yuning Tong, Jingkun An, Weibin Liao, Ewen M Harrison, Liantao Ma, and Chengwei Pan. 2024. Prompting large language models for zero-shot clinical prediction with structured longitudinal electronic health record data. *arXiv preprint arXiv:2402.01713* (2024).

A DETAILED DISCUSSION ON THE WORK'S RELEVANCE TO THE WEB AND THE TRACK

The Web Conference (WWW) is recognized for its pioneering role in exploring the Web's critical enablement of new research and applications, particularly within the realm of machine learning. As a distinct scholarly discipline, it boasts its own research methodologies, tools, and challenges, with a specific focus on enhancing the Web's technical infrastructure, including core platforms, and democratizing access to Web content and technologies to ensure they are more accessible, equitable, and inclusive.

Our submission aligns with the WWW's research scope in several fundamental respects:

A.1 Improving Accessibility of Web-Based Healthcare Services

For the majority of patients lacking expertise in the medical field, accessing web-based healthcare services to support the assessment and analysis of their health conditions presents significant accessibility challenges. Patients often rely on broad disease keywords to search for general medical knowledge on the web, which, while available, lacks the specificity and depth needed for accurate health assessments:

(1) **Vague Keyword Usage:** The keywords used by patients are often vague, leading to the retrieval of general medical information that is too broad and not tailored to their specific health conditions. This makes it difficult for patients to understand and apply the information to their situation.
(2) **Limitations of Traditional Search Engines:** Patients' electronic medical records contain more targeted information in the form of structured test and examination values. Identifying which values are indicative of their health prognosis is challenging and beyond the capabilities of traditional search engines or general large language models (LLMs), which lack data-driven analysis and modeling capabilities.
(3) **Generic Advice from LLMs:** Even when using general LLMs to analyze health conditions, the advice given is often too broad and generic, lacking specific direction. Moreover, the black-box nature of these models makes the results unreliable and inexplicable.

Our proposed framework addresses these issues by modeling electronic medical records data in a medical domain-specific, data-driven manner. It utilizes multi-agent, multi-view selection to identify the most critical features indicative of a patient's health condition, along with their abnormal descriptions. By leveraging the capabilities of large language models to retrieve highly targeted information on the web, our framework provides analysis results supported by high-credibility evidence. This enhances the accessibility of web-based healthcare services, serving a broader range of patients and empowering them to conduct health assessments based on their specific test data, thereby significantly improving the accessibility of medically relevant knowledge.

A.2 Facilitating Collaborative Prognosis for Physicians

Our method facilitates collaborative prognosis tasks for physicians worldwide who are web users. Designed as a web-based health monitoring framework and trained on web-accessed electronic medical records and medical corpora, our method advances personalized medicine, especially in underdeveloped countries. We have developed an online AI-Doctor interactive system that is fully internationalized and accessible, further indicating that our work falls within the scope of WWW. Related research has been previously published in WWW, as evidenced by the following citations:

(1) (WWW'21) Distilling knowledge from publicly available online EMR data to emerging epidemics for prognosis, https://dl.acm.org/doi/abs/10.1145/3442381.3449855
(2) (WWW'20) StageNet: Stage-Aware Neural Networks for Health Risk Prediction, https://dl.acm.org/doi/10.1145/3366423.3380136
(3) (WWW'19) LAVA: Longitudinal Adversarial Attack on Electronic Health Records Data, https://dl.acm.org/doi/abs/10.1145/3308558.3313528
(4) (WWW'19) Improving Medical Code Prediction from Clinical Text via Incorporating Online Knowledge Sources, https://dl.acm.org/doi/10.1145/3308558.3313485
(5) (WWW'17) Blood Pressure Prediction via Recurrent Models with Contextual Layer, https://dl.acm.org/doi/10.1145/3038912.3052604

A.3 Integrating Structured Data Analysis with Large Language Models

Large language models have become a fundamental engine for web-based services, revolutionizing the processing, analysis, and retrieval of purely natural language-based information. However, there exists a gap in the industry for supporting users in analyzing and utilizing structured data on the web. Our work bridges this gap by combining state-of-the-art domain-specific deep learning models for data-driven analysis and modeling of structured data with large language models through multi-agent collaboration. This offers a new perspective for the integration of structured data analysis into web-based services. By publicly releasing the code, our framework `ModelName` contributes to the conference's mission of promoting transparency in web applications for healthcare. This openness supports educational initiatives and the development of trustworthy web tools.

In summary, this work not only advances the field of interpretable clinical decision-making but also fits seamlessly within the research scope of the WWW conference. It embodies the conference's vision of creating AI tools that are both data-centric and web-centric, offering a significant contribution towards building trustworthy and transparent web applications in healthcare.