CARE-AGENT: MULTI-AGENT COLLABORATION WITH CONFLICT-AWARE ROUTING MECHANISM FOR DIAGNOSIS PREDICTION

Pengxiang Zhan[1] *Binteng Cai*[1] *Jinxu Zhang*[1] *Hang Lv*[1] *Yanchao Tan*[1*]
Zhigang Lin[2] Xiping Chen[3] Carl Yang[4]

[1]Fuzhou University, [2]The First Affiliated Hospital of Fujian Medical University,
[3]Hangzhou Bywin Technology Co., Ltd., [4]Emory University

ABSTRACT

Diagnosis prediction from electronic health records (EHRs) requires reasoning over clinical sequences and semantics to ensure reliable outcomes. Deep temporal models excel at capturing sequential patterns but offer limited interpretability, whereas large language models (LLMs) provide contextual clinical reasoning and explanation yet struggle with structured EHR inputs. To bridge these complementary strengths, we present CARE-Agent, a multi-Agent collaboration with a Conflict-Aware Routing mEchanism for accurate and reliable diagnosis prediction, which coordinates various deep predictors and LLMs. First, deep models act as sequential agents to generate candidate diagnoses, and a router identifies inter-agent conflicts. Then, LLMs serve as clinical agents with EHR-based, role-specialized prompting to synthesize patient context and deliver the final decision for ambiguous cases. Extensive experiments on two real-world EHR datasets demonstrate that CARE-Agent consistently outperforms state-of-the-art methods, achieving superior accuracy, robustness, and reliability. Our code is released at https://github.com/YYYTDMS/CARE-Agent.

***Index Terms*—** Electronic health record, Large language models, Multi-agent collaboration

1. INTRODUCTION

Accurate diagnosis prediction from electronic health records (EHRs) requires modeling clinical sequences while capturing semantic context to ensure reliable and interpretable outcomes. Deep temporal predictors efficiently capture sequence regularities but operate as black boxes with limited interpretability [1, 2, 3], whereas large language models (LLMs) provide contextual clinical reasoning and explanation but struggle with structured EHR inputs [4, 5]. Fig. 1 highlights the potential of a hybrid multi-agent paradigm that exploits the complementary strengths of deep predictors and LLMs to achieve accurate and reliable diagnosis prediction.

Recent years have seen rapid progress in multi-agent healthcare frameworks [6, 7, 8]. MDAgents [9] simulates multidisciplinary consultations to improve diagnostic reliability, MedAgents [10] leverages role specialization and debate to facilitate consensus, and ReConcile [11] introduces structured deliberation among LLM agents to enhance reasoning consistency. Yet these approaches are still dominated by a single backbone, either deep models or LLMs, leaving their complementary potential underexplored.

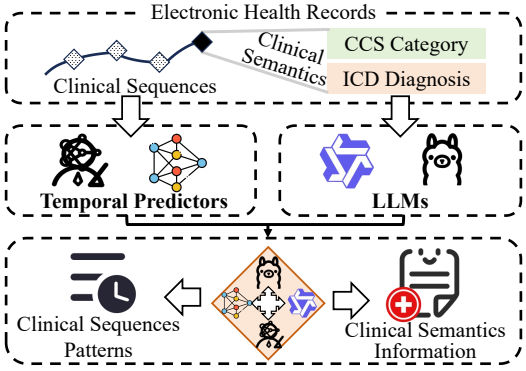


Fig. 1: A toy example illustrating a multi-agent framework that integrates deep temporal predictors and LLMs to capture clinical sequence patterns and clinical semantic information.

In this paper, we present **CARE-Agent**, a multi-agent collaboration framework with a Conflict-Aware Routing mEchanism for diagnosis prediction. CARE-Agent operates in two stages. First, **Conflict-Aware Candidate Generation**: multiple deep models act as sequential agents to produce candidate diagnoses, while a detection router monitors conflicts between agents to identify cases that merit collaboration. Then, **Clinical Decision Synthesis with LLMs**: LLMs are repurposed as clinical agents via EHR-based and role-specialized prompting to synthesize patient context and deliver the final decision for ambiguous cases.

Our contributions can be summarized as follows: (i) We introduce a novel framework that models prediction conflicts as actionable signals, enabling targeted collaboration between deep predictors and LLMs for diagnosis prediction; (ii) We design a conflict-aware routing mechanism and EHR-based prompting strategies that allow LLMs to function as role-specific clinical agents, improving contextual reasoning on

*Corresponding author

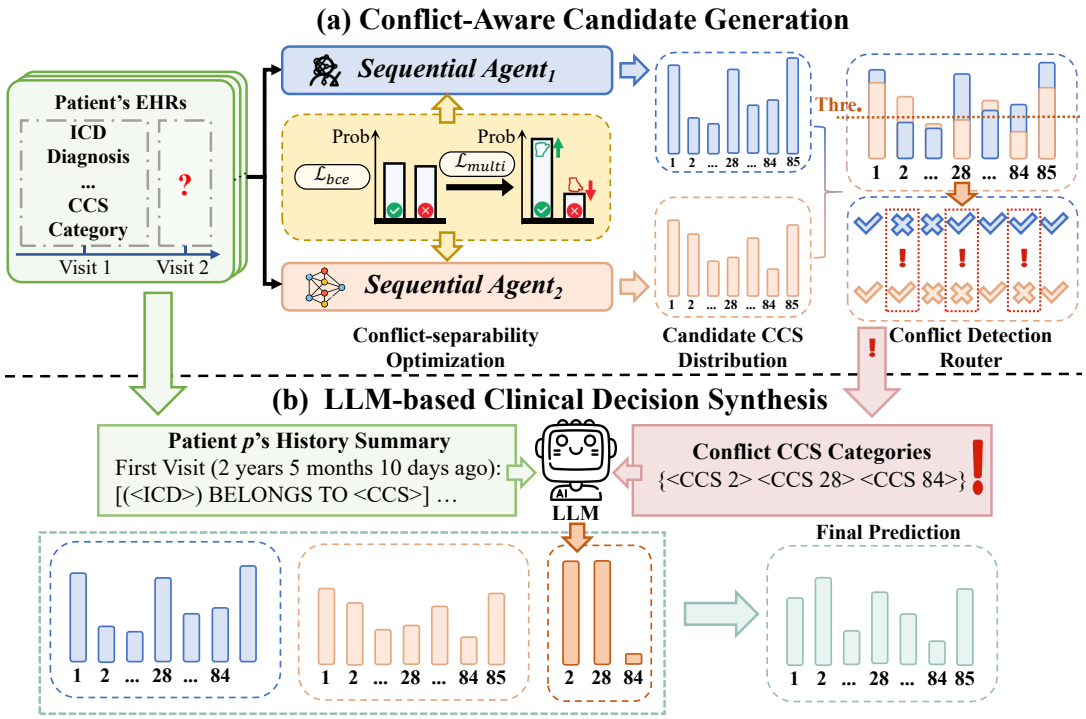


Fig. 2: The overall framework of CARE-Agent. (a) Deep models act as sequential agents to generate candidate diagnoses, while a router detects conflicts between agents. (b) LLMs act as clinical agents to synthesize patient context and resolve conflicts.

ambiguous cases; and (iii) Experiments on two real-world EHR datasets show that CARE-Agent achieves superior predictive accuracy and robustness.

2. METHODOLOGY

2.1. Problem Formulation and CARE-Agent Overview

We formalize the EHR of patient p as a sequence of visits $\{v_1, \ldots, v_t\}$, where each visit v_t contains ICD diagnosis codes $d_t \in \mathcal{D}$ and their corresponding CCS clinical categories $c_t \in \mathcal{C}$. The goal is to predict the CCS codes of the next visit v_{t+1} based on the historical EHR data.

We summarize the two main components of the CARE-Agent framework in Fig. 2. First, *Conflict-Aware Candidate Generation*, employs multiple deep temporal models as sequential agents to generate candidate diagnoses, while a detection router monitors conflicts between agents to identify cases that require further collaboration. Second, *LLM-based Clinical Decision Synthesis*, repurposes LLMs as role-specific clinical agents through EHR-based prompting, enabling them to synthesize patient context and resolve conflicting predictions to produce the final diagnosis.

2.2. Conflict-Aware Candidate Generation

Deep temporal models capture sequential patterns from longitudinal EHRs and enable diagnosis prediction. To leverage their complementary strengths, we employ multiple deep temporal models as sequential agents to generate candidate diagnoses. However, architectural differences often lead to divergent outputs on the same case.

To address this issue, we propose *conflict detection* on candidate CCS categories generated by multiple sequential agents. A key challenge is that predictive models trained with standard binary cross-entropy loss produce probability outputs with limited separability, making conflicts difficult to identify across sequential agents. Inspired by advances in medication recommendation [12, 13], we adopt a *multi-label margin loss* to improve the discriminability of predictions. The objective function is defined:

$$\mathcal{L}_{\text{BCE}} = -\sum_{i=1}^{|\mathcal{C}|} \left(y_i^p \log \hat{y}_i^p + (1 - y_i^p) \log(1 - \hat{y}_i^p)\right), \quad (1)$$

$$\mathcal{L}_{\text{Multi}} = \sum_{\{i|y_i^p=1\}} \sum_{\{j|y_j^p=0\}} \frac{\max\left(1 - (\hat{y}_i^p - \hat{y}_j^p), 0\right)}{|\mathcal{C}|}, \quad (2)$$

$$\mathcal{L}_{\text{Pred}} = \mathcal{L}_{\text{BCE}} + \lambda \mathcal{L}_{\text{Multi}}, \quad (3)$$

where $\hat{y}_i^p$ denotes the predicted probability of CCS category i for patient p, $y_i^p \in \{0, 1\}$ is the ground-truth label, and λ is a hyperparameter controlling discriminative separability. By enlarging margins between positive and negative categories, probability outputs become more separable.

After obtaining sufficiently discriminative CCS probability distributions, we apply a *conflict-aware routing mechanism* to direct predictions into either *conflict* or *consensus* categories. Formally, for patient p and CCS category i, let $z_{i,A_S}^p = \mathbf{1}\left[\hat{y}_{i,A_S}^p \geq \tau\right]$ denote the binary decision of a se-

Table 1: Statistics of the datasets used in our experiments.

Dataset	MIMIC-III	MIMIC-IV
# of patients	5,449	79,393
# of visits	14,141	329,605
Avg. # CCS per visit	12.08	11.30
Avg. # visits per patient	2.60	4.15
Max. # visits per patient	29	169
# of unique diagnoses	3,874	37,917
# of CCS codes	285	842

quential agent A_S. Given two agents A_S^1 and A_S^2, the conflict and consensus sets are defined as:

$$\begin{aligned} \mathcal{C}^p_{\text{conf}} &= \{\, i \mid z^p_{i,A_S^1} \neq z^p_{i,A_S^2} \,\}, \\ \mathcal{C}^p_{\text{cons}} &= \{\, i \mid z^p_{i,A_S^1} = z^p_{i,A_S^2} \,\}, \end{aligned} \tag{4}$$

where τ denotes the threshold and $\mathbf{1}[\cdot]$ the indicator function.

2.3. LLM-based Clinical Decision Synthesis

Conflicting CCS categories with opposite model predictions are difficult to resolve reliably. Recent studies highlight the potential of LLMs in diagnosis prediction [4, 5], owing to their strong contextual reasoning and semantic alignment.

To leverage this capability, we repurpose LLMs as *role-specific clinical agents* through EHR-based prompting, enabling them to synthesize patient context and resolve conflicting predictions. Specifically, we construct a *patient history summary* from longitudinal EHR data, where each visit is represented by its timestamp, ICD diagnosis names, and mapped CCS categories through CCS–ICD ontology alignment. We then extract the conflicting CCS set $\mathcal{C}^p_{\text{conf}}$ and integrate it with the patient history summary to form a structured prompt. This EHR-based prompt guides the LLM to synthesize clinical context and revise conflicting predictions, producing corrected binary outputs. Formally, for a patient p, the corrected decision for category $i \in \mathcal{C}^p_{\text{conf}}$ is given by

$$\hat{y}^p_{i,\,\text{LLM}} = f_{\text{LLM}}(\text{History}(p),\, \mathcal{C}^p_{\text{conf}})\,, \tag{5}$$

where $\text{History}(p)$ is the patient history summary and f_{LLM} outputs binary decisions for all conflicting categories.

Finally, we integrate the outputs of sequential agents and clinical agents to produce the final diagnosis decision:

$$\hat{y}^p_{i,\text{final}} = \frac{\hat{y}^p_{i,A_S^1} + \hat{y}^p_{i,A_S^2} + \alpha_i\, \hat{y}^p_{i,\text{LLM}}}{3}, \alpha_i \in \{0,1\}, \tag{6}$$

where α_i indicates whether the LLM contributes ($\alpha_i = 1$ for conflicts, $\alpha_i = 0$ otherwise).

3. EXPERIMENT

3.1. Datasets and Evaluation Protocols

We evaluate CARE-Agent on two real-world EHR datasets: MIMIC-III [14] and MIMIC-IV [15]. We select patients with at least two visits and predict the CCS codes for the next visit of patients. The statistics are summarized in Table 1. For evaluation, we adopt visit-level Precision@k (P@k) and code-level Accuracy@k (Acc@k), following [2]. P@k measures effectiveness as the proportion of correct codes in the top-k predictions. Acc@k measures reliability as the fraction of correctly predicted diagnoses at the code level.

3.2. Baselines and Implementation Details

We compare CARE-Agent against three categories of baselines: (1) **Deep Predictive Models**: RETAIN (R) [1], TRANS (T) [2], StageNet (S) [3], CGL [16], and SHy [17]; (2) **LLM-based Methods**: LLaMA3.1-8B [18] and Qwen3-8B [19]; (3) **Multi-agent Methods**: MedAgents [10] and ensembles of multiple predictive agents with unweighted averaging (Avg.).

Both datasets are split into train/validation/test sets (7:1:2) with patient-level partitioning, following [2]. All models are trained with Adam and hyperparameters as in the original papers. For LLMs, we adopt LLaMA-Factory [20] for deployment and evaluation, and for MedAgents, we employ Qwen3-8B to ensure fairness. The embedding dimension is fixed at 16, and λ and τ are set to 0.03 and 0.5, respectively. We adopt 5-fold cross-validation for robust evaluation. All experiments are conducted on two NVIDIA RTX 3090 Ti GPUs.

3.3. Overall Performance

Table 2 demonstrates CARE-Agent's consistent superiority over all baselines across both MIMIC-III and MIMIC-IV.

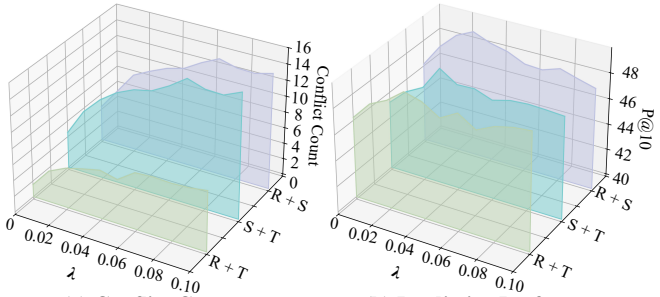
Compared with individual deep models, CARE-Agent leverages conflict-aware routing to detect inter-model disagreements and invokes LLM clinical agents for refinement, achieving significant performance gains and enhanced stability. On MIMIC-III, CARE-Agent (R + T) achieves 49.29 P@10, outperforming RETAIN and TRANS by an average of 5.67%, while CARE-Agent (R + S) achieves 49.72 P@10, exceeding StageNet and TRANS by 9.92% on average.

Compared with LLM baselines, CARE-Agent integrates multiple deep models to provide reliable probability estimates and uses structured prompts to guide LLM reasoning only on conflicting categories, reducing hallucinations and improving efficiency. On MIMIC-IV, CARE-Agent achieves an improvement of 9.58% in P@10 over Qwen3-8B.

Compared with multi-agent methods, CARE-Agent demonstrates stronger reliability and efficiency. Unlike MedAgents, which relies on full deliberation among LLM agents and incurs high cost, CARE-Agent selectively invokes LLM agents only on conflicts, enabling targeted refinement and higher accuracy, improving P@10 by 4.02% on MIMIC-III. Against simple averaging ensembles, it further avoids noise from poorly calibrated agents. Even under the weaker TRANS backbone in MIMIC-IV, CARE-Agent (R + T) still surpasses Avg. (R + T) by 6.24% in P@10.

Table 2: Experimental results for diagnosis prediction (%) on the MIMIC-III and MIMIC-IV datasets. The best results are highlighted in **bold** while the second best are underlined.

Dataset	MIMIC-III				MIMIC-IV			
Metric	Visit-Level		Code-Level		Visit-Level		Code-Level	
	P@10	P@20	Acc@10	Acc@20	P@10	P@20	Acc@10	Acc@20
RETAIN (R)	$47.82_{\pm 0.17}$	$54.40_{\pm 0.17}$	$34.05_{\pm 0.16}$	$52.13_{\pm 0.18}$	$48.45_{\pm 0.09}$	$54.67_{\pm 0.11}$	$35.23_{\pm 0.09}$	$50.68_{\pm 0.10}$
TRANS (T)	$45.53_{\pm 0.20}$	$52.05_{\pm 0.21}$	$32.43_{\pm 0.17}$	$49.84_{\pm 0.20}$	$39.43_{\pm 0.10}$	$46.19_{\pm 0.08}$	$29.09_{\pm 0.11}$	$43.56_{\pm 0.09}$
StageNet (S)	$44.94_{\pm 0.20}$	$50.82_{\pm 0.23}$	$31.74_{\pm 0.17}$	$48.18_{\pm 0.14}$	$51.15_{\pm 0.09}$	$57.42_{\pm 0.10}$	$36.67_{\pm 0.08}$	$52.88_{\pm 0.11}$
CGL	$47.84_{\pm 0.21}$	$54.78_{\pm 0.25}$	$34.02_{\pm 0.20}$	$52.18_{\pm 0.21}$	$49.07_{\pm 0.11}$	$55.06_{\pm 0.13}$	$34.99_{\pm 0.13}$	$51.20_{\pm 0.09}$
SHy	$45.37_{\pm 0.19}$	$51.62_{\pm 0.21}$	$32.25_{\pm 0.20}$	$49.12_{\pm 0.20}$	$48.56_{\pm 0.12}$	$54.20_{\pm 0.11}$	$35.08_{\pm 0.09}$	$50.13_{\pm 0.10}$
LLaMA3.1-8B	$41.19_{\pm 2.11}$	$42.06_{\pm 1.03}$	$30.46_{\pm 0.84}$	$39.88_{\pm 0.71}$	$38.32_{\pm 1.02}$	$43.62_{\pm 1.36}$	$31.22_{\pm 0.63}$	$33.82_{\pm 0.72}$
Qwen3-8B	$47.03_{\pm 1.05}$	$48.61_{\pm 1.12}$	$32.73_{\pm 0.66}$	$46.10_{\pm 1.02}$	$48.04_{\pm 0.75}$	$50.99_{\pm 1.20}$	$34.12_{\pm 0.69}$	$46.06_{\pm 0.78}$
MedAgents	$47.41_{\pm 0.72}$	$48.98_{\pm 1.80}$	$33.01_{\pm 0.82}$	$46.65_{\pm 1.01}$	$48.39_{\pm 0.61}$	$51.17_{\pm 1.58}$	$34.46_{\pm 0.61}$	$47.26_{\pm 0.91}$
Avg. (R + S)	$47.85_{\pm 0.12}$	$53.88_{\pm 0.14}$	$33.80_{\pm 0.13}$	$51.36_{\pm 0.12}$	$51.36_{\pm 0.14}$	$57.81_{\pm 0.15}$	$36.85_{\pm 0.12}$	$52.96_{\pm 0.14}$
Avg. (S + T)	$47.30_{\pm 0.15}$	$52.95_{\pm 0.15}$	$33.38_{\pm 0.13}$	$50.43_{\pm 0.13}$	$51.26_{\pm 0.16}$	$56.66_{\pm 0.15}$	$36.19_{\pm 0.13}$	$52.10_{\pm 0.14}$
Avg. (R + T)	$47.34_{\pm 0.14}$	$53.98_{\pm 0.15}$	$33.72_{\pm 0.12}$	$52.76_{\pm 0.14}$	$46.93_{\pm 0.15}$	$53.00_{\pm 0.14}$	$34.23_{\pm 0.11}$	$49.08_{\pm 0.13}$
CARE-Agent (R + S)	$\mathbf{49.72}_{\pm 0.11}$	$\mathbf{56.37}_{\pm 0.11}$	$35.04_{\pm 0.12}$	$\mathbf{53.82}_{\pm 0.11}$	$\mathbf{52.64}_{\pm 0.12}$	$\mathbf{59.58}_{\pm 0.11}$	$\mathbf{37.78}_{\pm 0.10}$	$\mathbf{54.72}_{\pm 0.12}$
CARE-Agent (S + T)	$48.94_{\pm 0.14}$	$55.40_{\pm 0.13}$	$34.54_{\pm 0.10}$	$52.94_{\pm 0.11}$	$52.52_{\pm 0.14}$	$58.96_{\pm 0.13}$	$37.59_{\pm 0.12}$	$54.05_{\pm 0.11}$
CARE-Agent (R + T)	$49.29_{\pm 0.10}$	$55.60_{\pm 0.12}$	$\mathbf{35.05}_{\pm 0.10}$	$53.41_{\pm 0.12}$	$49.86_{\pm 0.13}$	$56.14_{\pm 0.11}$	$35.74_{\pm 0.10}$	$51.93_{\pm 0.11}$



(a) Conflict Count **(b) Predictive Performance**

Fig. 3: Impact of λ on conflict detection: (a) conflict count and (b) predictive performance across different model pairs on the MIMIC-III dataset.

3.4. Varying the Conflict-Aware Router

Fig. 3 illustrates the impact of varying hyperparameter λ on conflict counts (a) and predictive performance (b). Increasing λ enlarges probability margins to flag more conflicts , but excessive values degrade performance by introducing noise to LLM revisions via over-sensitive routing. A moderate setting ($\lambda = 0.03$) achieves the best balance, capturing meaningful conflicts without diluting the reliability of sequential agents, thereby maintaining both sensitivity and predictive stability.

3.5. Varying Agents for Decision Synthesis

Table 3 validates CARE-Agent's generalization, demonstrating consistent superiority over standalone baselines across both LLaMA3.1-8B and Qwen3-8B. LLaMA3.1-8B shows higher relative gains (+14.73%) than Qwen3-8B (+5.71%), indicating the conflict-aware design especially benefits weaker LLMs. Selective LLM invocation for conflicting categories reduces inference time by 90% over LLM-only baselines, balancing accuracy and efficiency.

Table 3: Generalization ability of CARE-Agent across different LLMs on the MIMIC-III dataset.

LLM	Deep	P@10	Acc@10	Avg. Time (s)
LLaMA3.1	None	41.19	30.46	10.52
	R + S	48.55	34.04	2.13
	S + T	47.87	33.76	1.97
	R + T	48.93	34.38	1.75
Qwen3	None	47.03	32.73	10.21
	R + S	49.72	35.04	1.35
	S + T	48.94	34.54	1.27
	R + T	49.29	35.05	0.98

4. CONCLUSION

In this paper, we presented CARE-Agent, a multi-agent collaboration framework for diagnosis prediction that first employs sequential agents with a conflict-aware router to flag inter-agent disagreements and then repurposes LLMs as clinical agents through EHR-grounded prompting to resolve them. Experiments on two real-world EHR datasets demonstrate that CARE-Agent consistently outperforms individual deep models, LLM-only baselines, and multi-agent methods, achieving more accurate and robust diagnosis prediction.

5. ACKNOWLEDGEMENTS

This work was supported by the Fujian Provincial Artificial Intelligence Industry Development Technology Project under Grant (2025H0042), Fujian Provincial Natural Science Foundation of China under Grants (2025J01540), National Natural Science Foundation of China under Grants (62302098). Carl Yang was not supported by any fund from China.

6. REFERENCES

[1] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart, "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism," *NeurIPS*, vol. 29, 2016.

[2] Jiayuan Chen, Changchang Yin, Yuanlong Wang, and Ping Zhang, "Predictive modeling with temporal graphical representation on electronic health records," in *IJCAI*, 2024, pp. 5763–5771.

[3] Junyi Gao, Cao Xiao, Yasha Wang, Wen Tang, Lucas M Glass, and Jimeng Sun, "Stagenet: Stage-aware neural networks for health risk prediction," in *WWW*, 2020, pp. 530–540.

[4] Xuanzhong Chen, Xiaohao Mao, Qihan Guo, Lun Wang, Shuyang Zhang, and Ting Chen, "Rarebench: can llms serve as rare diseases specialists?," in *SIGKDD*, 2024, pp. 4850–4861.

[5] Taeyoon Kwon, Kai Tzu-iunn Ong, Dongjin Kang, Seungjun Moon, Jeong Ryong Lee, Dosik Hwang, Beomseok Sohn, Yongsik Sim, Dongha Lee, and Jinyoung Yeo, "Large language models are clinical reasoners: Reasoning-aware diagnosis framework with prompt-generated rationales," in *AAAI*, 2024, vol. 38, pp. 18417–18425.

[6] Weijieying Ren, Jingxi Zhu, Zehao Liu, Tianxiang Zhao, and Vasant Honavar, "A comprehensive survey of electronic health record modeling: From deep learning approaches to large language models," *arXiv preprint arXiv:2507.12774*, 2025.

[7] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang, "Large language model based multi-agents: A survey of progress and challenges," in *IJCAI*, 2024.

[8] Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Jiaming Ji, Wenting Chen, Xiang Li, and Yixuan Yuan, "A survey of llm-based agents in medicine: How far are we from baymax?," in *ACL*, 2025.

[9] Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park, "Mdagents: An adaptive collaboration of llms for medical decision-making," *NeurIPS*, vol. 37, pp. 79410–79452, 2024.

[10] Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein, "Medagents: Large language models as collaborators for zero-shot medical reasoning," in *ICLR 2024 Workshop*.

[11] Justin Chen, Swarnadeep Saha, and Mohit Bansal, "Reconcile: Round-table conference improves reasoning via consensus among diverse llms," in *ACL*, 2024, pp. 7066–7085.

[12] Chuang Zhao, Hongke Zhao, Xiaofang Zhou, and Xiaomeng Li, "Enhancing precision drug recommendations via in-depth exploration of motif relationships," *TKDE*, 2024.

[13] Nianzu Yang, Kaipeng Zeng, Qitian Wu, and Junchi Yan, "Molerec: Combinatorial drug recommendation with substructure-aware molecular representation learning," in *WWW*, 2023, pp. 4075–4085.

[14] 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, "Mimic-iii, a freely accessible critical care database," *Sci. Data*, vol. 3, no. 1, pp. 1–9, 2016.

[15] Alistair EW Johnson, David J Stone, Leo A Celi, and Tom J Pollard, "The mimic code repository: enabling reproducibility in critical care research," *J. Am. Med. Inform. Assoc.*, vol. 25, no. 1, pp. 32–39, 2018.

[16] Chang Lu, Chandan K Reddy, Prithwish Chakraborty, Samantha Kleinberg, and Yue Ning, "Collaborative graph learning with auxiliary text for temporal event prediction in healthcare," in *IJCAI*, 2021.

[17] Leisheng Yu, Yanxiao Cai, Minxing Zhang, and Xia Hu, "Self-explaining hypergraph neural networks for diagnosis prediction," *arXiv preprint arXiv:2502.10689*, 2025.

[18] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al., "The llama 3 herd of models," *arXiv e-prints*, pp. arXiv–2407, 2024.

[19] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al., "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[20] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo, "Llamafactory: Unified efficient fine-tuning of 100+ language models," in *ACL*, 2024, pp. 400–410.