Quantify, Collect, and Correct: Evidence-Driven Multi-Agent Framework for Critical Evidence Refinement in Emergency Rooms

Haoyu Cao[1*], Yueze Fu[1*], Wenbo Gao[1,3], Yuxin Lin[1], Xiangyu Li[2], and Wei Wang[1(✉)]

[1] College of Informatics, Harbin Institute of Technology, Shenzhen, China
wangwei2019@hit.edu.cn
[2] Faculty of Computing, Harbin Institute of Technology, Harbin, China
[3] The Hong Kong Polytechnic University, Hong Kong SAR, China

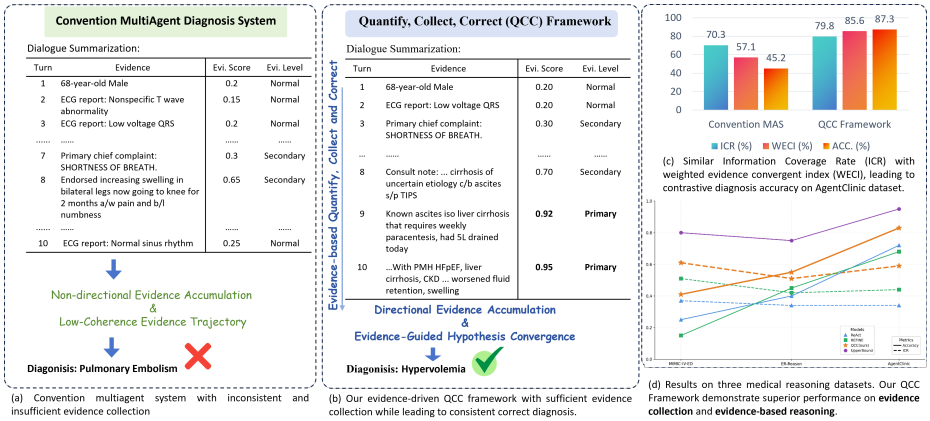
Abstract. Emergency-room (ER) decision-making is characterized by high clinical risk, strict time constraints, and rapidly evolving yet incomplete evidence. Recently, multi-agent systems (MAS) have shown promise for interactive clinical reasoning. However, we identify a key limitation in ER-style workflows: multi-turn dialogues often terminate with insufficient evidence coverage, and similar coverage ratios can still lead to markedly different outcomes when a small set of decision-critical evidence is missing. To make evidence a first-class optimization target, we propose a closed-loop Quantify, Collect, and Correct (QCC) multi-agent framework for emergency reasoning. In the Quantify stage, we introduce the Weighted Evidence Convergent Index (WECI), which importance-weights collected evidence to measure how well MAS capture decision-critical signals during interaction. In the Collect stage, an Evidence-Driven Planning Module leverages current evidence and retrieval-augmented medical knowledge to infer plausible diagnostic directions, plan a small set of prioritized missing evidence targets for subsequent turns, while the Doctor autonomously determines when the accumulated evidence is sufficient to conclude the consultation. In the Correct stage, we perform evidence-driven inference and apply evidence–diagnosis cross-checking to detect inconsistencies and revise conclusions when necessary. Experiments on two ER datasets and one general diagnosis dataset demonstrate that QCC improves emergency decision quality and robustness under noisy and insufficient evidence, consistently outperforming strong MAS baselines across multiple backbone LLMs and evaluation settings. Codes are available at `https://github.com/HaoyuCao/QCC_MAS`.

Keywords: Multi-Agent · Evidence-Driven Diagnosis · Emergency Room

1 Introduction

Multi-agent systems (MAS) have recently emerged as a promising paradigm for clinical decision support [20, 18], where specialized agents collaborate through

[*] Haoyu Cao and Yueze Fu contributed equally.



(a) Convention multiagent system with inconsistent and insufficient evidence collection

(b) Our evidence-driven QCC framework with sufficient evidence collection while leading to consistent correct diagnosis.

(c) Similar Information Coverage Rate (ICR) with weighted evidence convergent index (WECI), leading to contrastive diagnosis accuracy on AgentClinic dataset.

(d) Results on three medical reasoning datasets. Our QCC Framework demonstrate superior performance on **evidence collection** and **evidence-based reasoning**.

Fig. 1. Existing multi-agent diagnostic systems collect evidence in a non-directional and weakly coherent manner, undermining progressive hypothesis refinement. Our QCC framework performs direction-guided selection of diagnostically salient evidence, producing accurate predictions with coherent evidence chains.

multi-turn interactions to perform history taking, differential diagnosis, information retrieval, and verification [21, 19, 4]. Although MAS have shown strong potential in surgical planning [29], dental assessment [2], and medical image analysis [11], emergency-room (ER) decision-making remains underexplored from an evidence-centric perspective. In ER settings, clinicians must act under strict time constraints with incomplete and rapidly evolving evidence [1, 5], making evidence efficiency a primary concern for interactive MAS.

Existing MAS frameworks such as ReAct-style [27] and REFINE-style [10] pipelines treat evidence as unstructured context to be progressively expanded. We evaluate these frameworks with an evidence coverage ratio and identify two key issues (Fig. 1(c)): (1) multi-turn dialogues frequently terminate while substantial critical evidence remains missing, and (2) even similar coverage ratios can yield divergent outcomes when a small set of high-significance evidence is omitted. These findings underscore the need to explicitly model and weight evidence utility [17], making evidence an optimization target rather than incidental context.

To address these challenges, we propose a closed-loop [12, 25] Quantify–Collect–Correct (QCC) multi-agent framework that makes evidence a first-class optimization target for emergency reasoning. Rather than merely adding another multi-agent workflow optimized toward final diagnostic accuracy, QCC shifts the focus from answer-only optimization to evidence-oriented optimization, aiming to acquire decision-critical and clinically relevant evidence during interactive dialogue. In the Quantify stage, we introduce the Weighted Evidence Convergent Index (WECI), an importance-weighted metric that measures how well a MAS captures decision-critical evidence during interaction. In the Collect stage, we present an Evidence-Driven Planning Module (EDPM) that periodically lever-

ages the currently collected evidence to infer plausible diagnostic directions at the body-system level and generate prioritized Prospective Primary Evidence targets for subsequent turns, while preserving the Doctor's clinical autonomy to explore freely and to determine when sufficient evidence has been gathered. In the Correct stage, an Evidence-Based Verifier Module [6, 15] cross-checks evidence–diagnosis consistency and planning direction integrity, issuing targeted feedback to both the Doctor and the Planner when inconsistencies or reasoning biases are detected, improving robustness and interpretability. Our QCC framework efficiently collects logically coherent, decision-critical evidence and supports reliable conclusions, as shown in Fig. 1(a)–(b).

Experiments on one general diagnosis benchmark and two ER datasets show that QCC consistently outperforms strong MAS baselines across multiple backbone LLMs, especially under noisy and insufficient evidence conditions (Fig. 1(d)).

2 Methodology

As motivated by the evidence bottlenecks identified in Sec. 1, we propose a closed-loop Quantify–Collect–Correct (QCC) framework that makes evidence a first-class optimization target throughout interaction. QCC couples multi-turn dialogue with evidence-level feedback: (i) **Quantify** introduces the Weighted Evidence Convergent Index (WECI) to importance-weight collected evidence and measure the capture of decision-critical signals; (ii) **Collect** uses an Evidence-Driven Planning Module that periodically infers plausible diagnostic directions from current evidence and retrieval-augmented medical knowledge, and outputs a small set of Prospective Primary Evidence targets for subsequent turns; (iii) **Correct** performs evidence-driven inference after termination and applies an Evidence-Based Verifier to cross-check evidence–diagnosis consistency, triggering revision when conflicts or missing key evidence are detected. The three components are detailed in Secs. 2.1–2.3.

2.1 Weighted Evidence Convergent Index

To instantiate the Quantify stage in QCC, we propose the Weighted Evidence Convergent Index (WECI) to measure how well an interactive multi-agent system captures *decision-critical* evidence during multi-turn diagnosis. Unlike coverage-only metrics that treat all evidence equally, WECI accounts for heterogeneous clinical importance: missing a small number of high-significance cues can be more detrimental than omitting many low-impact details. Formally, we define

$$\mathrm{WECI} = \frac{\sum_{e \in E \cap E^*} w_e}{\sum_{e \in E^*} w_e}, \tag{1}$$

where E is the collected evidence set, E^* is the ground-truth evidence set curated by specialist clinicians, and $w_e \in [0, 1]$ denotes the importance of evidence item e, estimated by an evidence evaluation agent $\phi_{EA}(\cdot)$ [30] conditioned on retrieval-augmented specialist knowledge.

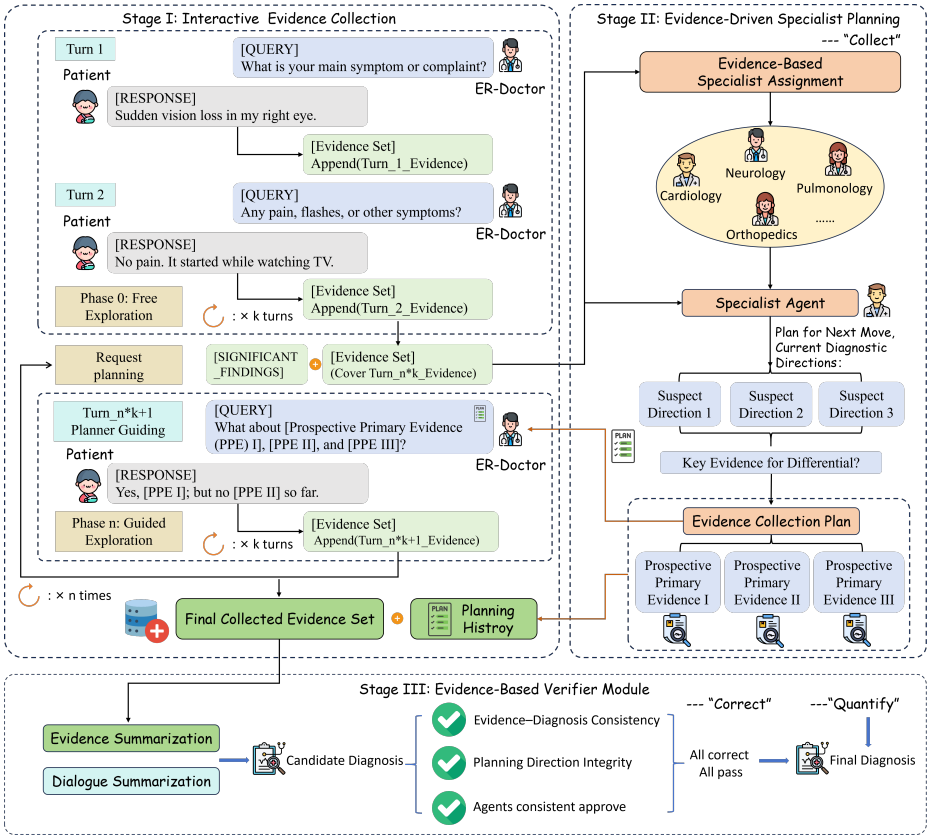


Fig. 2. Overview of our QCC framework.

WECI can be interpreted as an *importance-weighted recall* over decision-critical evidence: it quantifies the fraction of clinically salient evidence mass in E^* that is successfully acquired through interaction. A higher WECI indicates that the dialogue captures a larger share of decisive cues required for correct diagnosis and management, while a lower WECI reveals that the interaction fails to retrieve key evidence even if the raw coverage appears comparable. Therefore, WECI complements PC/EC/ICR by directly evaluating whether the collected evidence aligns with clinical decision needs, and serves as a principled target for evidence-oriented planning and verification in QCC.

2.2 Evidence-Driven Planning Module

In emergency departments, physicians must cover a broad spectrum of clinical conditions. While such generalist scope ensures wide applicability, the depth of specialty-specific reasoning is often limited, leading to two common issues in multi-agent diagnostic systems: (1) insufficient evidence collection (Fig. 1(c))

and (2) non-directional evidence accumulation (Fig. 1(a)). To address these, we introduce an Evidence-Driven Planning Module (EDPM, Fig. 2 Stage II), which restructures evidence collection through periodic specialist-guided planning.

Initial Free Exploration. The ER-Doctor Agent first conducts K turns of unconstrained information gathering, yielding an initial observed evidence set $\mathcal{S}_{obs} = \{s_1, s_2, \ldots, s_n\}$. This open-ended exploration allows the Doctor to establish a broad clinical picture before EDPM intervenes.

Specialist-Guided Planning. EDPM maintains a candidate pool of specialist agents:

$$\mathcal{A} = \{A_{card}, A_{neuro}, A_{pulm}, A_{ortho}, \ldots\}. \tag{2}$$

Given $\mathcal{S}_{obs}$, the system computes a relevance score for each specialist: $\text{Score}(A_i \mid \mathcal{S}_{obs}) = \psi(\mathcal{S}_{obs}, \mathcal{K}_{A_i})$, where $\mathcal{K}_{A_i}$ is the relevant specialist knowledge database. The specialist $\arg\max_{A_i \in \mathcal{A}} \text{Score}(A_i \mid \mathcal{S}_{obs})$ is assigned as the **Planner** to guide subsequent reasoning, transforming the process from a broad generalist perspective into a focused specialty-driven hypothesis space.

To preserve the Doctor's autonomous exploratory capability and avoid premature diagnostic anchoring [22], the Planner A^* provides only m system-level diagnostic directions (default $m = 3$): $\mathcal{D}_{dir} = \{d_1, d_2, \ldots, d_m\}$, representing the most plausible clinical systems to investigate without specifying concrete candidate diseases. For each direction d_j, EDPM identifies a yet-uncollected, diagnostically critical evidence target termed Prospective Primary Evidence (PPE):

$$\mathcal{P}lan = \{\text{PPE}_1, \text{PPE}_2, \ldots, \text{PPE}_m\}, \tag{3}$$

where each $\text{PPE}_j \notin \mathcal{S}_{obs}$ and possesses high discriminative power for confirming or differentiating direction d_j.

Periodic Guided Exploration. The PPE plan guides the ER-Doctor Agent in subsequent turns. After every K turns, EDPM re-intervenes to update diagnostic directions based on newly acquired evidence and retrieval-augmented medical knowledge [9] and to issue a revised PPE plan. This cycle repeats until the Doctor determines that sufficient evidence has been gathered. The Doctor may also flag an unexpected significant finding at any turn to trigger an immediate EDPM re-assessment, enabling the Planner to adjust its diagnostic directions in response to pivotal clinical clues [26]. The complete interaction produces a Final Collected Evidence Set and a Planning History for downstream verification.

Through this periodic planning mechanism, EDPM shifts the system from passive questioning toward hypothesis-driven evidence acquisition [26], promoting directional evidence convergence and improved diagnostic consistency.

2.3 Evidence-Based Verifier Module

Upon the Doctor signaling consultation completion, the EBVM activates to validate the candidate diagnosis. The Summarizer generates a structured evidence summary from the dialogue history, and the Diagnostician analyzes this summary alongside the Planner's diagnostic directions $\mathcal{D}_{dir}$ to produce a candidate diagnosis with a confidence level.

Table 1. Performance on AgentClinic. Interactive baselines are evaluated by PC/EC/ICR, WECI, and Acc. CoT is the upper bound that observes all evidence.

Method	Model	AgentClinic				
		PC (%)	EC (%)	ICR (%)	WECI (%)	Acc. (%)
COT	Qwen3-Max	-	-	-	-	86.0
	GPT-5-mini	-	-	-	-	95.0
RolePlay	Qwen3-Max	6.3	24.5	16.4	15.4	42.5
	GPT-5-mini	44.2	22.7	32.3	50.1	67.5
SC	Qwen3-Max	4.2	31.8	19.3	30.2	67.5
	GPT-5-mini	38.8	33.0	35.4	54.5	69.5
REFINE	Qwen3-Max	5.4	35.5	21.9	20.6	69.5
	GPT-5-mini	41.4	47.8	44.4	64.1	67.5
REACT	Qwen3-Max	4.3	21.3	13.5	17.9	46.5
	GPT-5-mini	37.8	30.6	33.5	52.5	71.5
QCC (Ours)	Qwen3-Max	**10.8**	**46.9**	**30.3**	**42.1**	**76.5**
	GPT-5-mini	**49.6**	**72.2**	**61.5**	**68.9**	**82.5**

The Verifier then evaluates the candidate against three criteria: (1) **Evidence–Diagnosis Consistency**, which checks whether the collected evidence sufficiently supports the diagnosis without internal contradictions; (2) **Planning Direction Integrity**, which examines whether the Planner's reasoning remained balanced throughout the diagnostic process, rather than narrowing prematurely onto a single hypothesis while overlooking contradictory evidence;(3) **Agents Consistent Approve** [23, 3], which requires agreement between the Diagnostician's confidence and the Verifier's independent assessment.

3 Experiment

3.1 Datasets

We evaluate QCC on AgentClinic [16], a controlled multi-agent diagnosis benchmark with relatively complete case descriptions, and two real-world ED sources, MIMIC-IV-ED [7] and ER-REASON [14], which better reflect noisy and fragmented emergency documentation. For evidence-level evaluation, we use an external LLM (GPT-5.2) to convert raw medical records into structured, evidence-based cases by separating patient background from examination/test information and decomposing each part into non-overlapping atomic evidence units[13]. These patient and exam evidence sets are progressively revealed to the patient/reporter agents during interaction, enabling dynamic evidence re-weighting and correction in multi-turn reasoning. Following common benchmark construction and evaluation practice, we uniformly sample 200 cases from each data source to balance coverage and evaluation cost.

Table 2. Performance comparison on MIMIC-IV-ED and ER-REASON. CoT performs as the performance upper bound for all interactive multiagent systems.

Method	Model	MIMIC-IV-ED					ER-REASON				
		PC (%)	EC (%)	ICR (%)	WECI (%)	Acc. (%)	PC (%)	EC (%)	ICR (%)	WECI (%)	Acc. (%)
CoT	Qwen3-Max	-	-	-	-	60.0	-	-	-	-	70.0
	GPT-5-mini	-	-	-	-	80.0	-	-	-	-	80.0
RolePlay	Qwen3-Max	10.0	0.0	3.8	9.1	0.0	15.0	17.5	16.2	30.2	10.0
	GPT-5-mini	37.5	33.5	35.0	47.9	10.0	36.4	29.9	33.9	59.7	30.0
SC	Qwen3-Max	13.3	7.5	9.7	14.5	15.0	23.6	26.5	24.5	46.7	30.0
	GPT-5-mini	35.8	64.0	53.4	64.2	5.0	35.7	35.4	35.2	68.7	20.0
REFINE	Qwen3-Max	15.8	12.0	13.4	17.5	10.0	24.3	26.2	25.0	43.5	30.0
	GPT-5-mini	34.2	62.0	51.6	60.6	15.0	34.3	56.5	43.0	78.7	45.0
ReACT	Qwen3-Max	10.0	16.5	14.1	13.4	0.0	21.4	24.2	22.4	39.3	20.0
	GPT-5-mini	35.8	56.0	48.4	58.2	25.0	32.1	34.2	32.3	71.5	40.0
QCC (Ours)	Qwen3-Max	**23.8**	**47.0**	**34.5**	**29.3**	**45.0**	**28.6**	**31.5**	**35.8**	**51.1**	**50.0**
	GPT-5-mini	**43.8**	**68.5**	**59.2**	**68.7**	**52.5**	**42.1**	**66.5**	**44.7**	**82.1**	**72.5**

Table 3. Ablation study of evidence planning and reasoning modules.

Model	Variant	Components		MIMIC-IV-ED					ER-REASON				
		Plan	Reason	PC (%)	EC (%)	ICR (%)	WECI (%)	Acc. (%)	PC (%)	EC (%)	ICR (%)	WECI (%)	Acc. (%)
Qwen3-Max	Baseline	-	-	15.8	12.0	13.4	17.5	10.0	24.3	26.2	25.0	43.5	30.0
	+Evi. Plan	✓	-	21.6	47.0	33.8	27.6	32.5	24.8	29.6	32.4	49.7	37.5
	+Evi. Plan+Verifier	✓	✓	**23.8**	**47.0**	**34.5**	**29.3**	**45.0**	**28.6**	**31.5**	**35.8**	**51.1**	**50.0**
GPT-5-mini	Baseline	-	-	34.2	62.0	51.6	60.6	15.0	34.3	56.5	43.0	78.7	45.0
	+Evi. Plan	✓	-	42.1	62.3	56.6	58.2	37.5	39.5	60.2	42.8	68.2	62.5
	+Evi. Plan+Verifier	✓	✓	**43.8**	**68.5**	**59.2**	**68.7**	**52.5**	**42.1**	**66.5**	**44.7**	**82.1**	**72.5**

3.2 Baselines and Evaluation Protocol

We focus on interactive multi-agent diagnosis settings, where the system iteratively queries and aggregates clinical evidence through multi-turn dialogue. We compare our QCC framework against representative interactive frameworks, including ReAct [28], Summarization Context [8], RolePlay, and REFINE [10]. In addition, we include a non-interactive upper-bound baseline, CoT [24], which performs one-shot reasoning by providing the model with all available evidence at once, thereby removing the evidence-seeking constraint.

To evaluate both evidence acquisition and final decision quality, we adopt coverage-based metrics to quantify evidence collection performance during interaction. Specifically, Patient Facts Coverage (PC) measures the proportion of collected patient-background facts relative to the labeled patient facts, while Exam Facts Coverage (EC) measures the proportion of collected examination/test facts relative to the labeled exam facts. We further report Information Convergent Rate (ICR) as the overall fraction of collected evidence that matches the labeled evidence, reflecting the framework's holistic evidence-gathering ability.

Beyond coverage, we assess whether the collected evidence is directionally useful for clinical decision-making. We use an external LLM, conditioned on the ground-truth diagnosis, to assign importance weights to each evidence item, and compute the WECI metrics based on the weighted evidence. WECI captures the significance and diagnostic alignment of the acquired evidence, complement-

ing coverage-oriented metrics. Finally, we report diagnostic Accuracy (Acc.) to evaluate the correctness of the final diagnosis produced by each method.

3.3 Results on General Diagnostic Dataset

As shown in Table 1, QCC achieves the strongest interactive performance on AgentClinic, demonstrating that our evidence-oriented interaction generalizes beyond ED-specific benchmarks. With GPT-5-mini, QCC markedly improves evidence acquisition and convergence over the best interactive baselines, increasing PC from 44.2 to 49.6, EC from 47.8 to 72.2, and ICR from 44.4 to 61.5. Importantly, QCC also attains a higher WECI (68.9 vs. 64.1), indicating superior capture of decision-critical evidence rather than merely collecting more facts. This evidence advantage translates into substantially better final decisions, with QCC reaching 82.5% accuracy and outperforming ReAct (71.5%) by a wide margin. Similar gains are observed with Qwen3-Max, where QCC delivers the best interactive accuracy (76.5%) alongside consistently stronger evidence coverage.

3.4 Results on Real-World ER Reasoning

Table 2 reports results on two real-world ED benchmarks, where noisy and incomplete documentation makes evidence acquisition a first-order bottleneck. While CoT serves as an upper bound by observing all evidence, QCC consistently dominates interactive baselines across evidence coverage, convergence, evidence quality, and final accuracy. With GPT-5-mini, QCC achieves 52.5% accuracy on MIMIC-IV-ED and 72.5% on ER-REASON, simultaneously attaining the highest PC/EC/ICR and the highest WECI among interactive methods (68.7 and 82.1), confirming that QCC not only collects more evidence but also captures more decision-critical signals. QCC remains robust with Qwen3-Max, improving accuracy to 45.0% and 50.0% on MIMIC-IV-ED and ER-REASON, respectively, while maintaining the strongest evidence convergence among interactive competitors.

3.5 Ablation Study

Table 3 establishes a clear causal chain from evidence planning to diagnostic correctness. Enabling the Plan module substantially strengthens evidence acquisition and convergence (higher PC/EC and ICR) on both MIMIC-IV-ED and ER-REASON, which directly translates into large accuracy gains (Qwen3-Max: 10.0→32.5 and 30.0→37.5; GPT-5-mini: 15.0→37.5 and 45.0→62.5). Building on this enriched evidence set, adding the Verifier further increases decision-critical evidence capture (WECI) and delivers another consistent boost in accuracy (e.g., GPT-5-mini: 37.5→52.5 and 62.5→72.5), demonstrating that verification is most effective when guided by sufficient, salient evidence.

4 Conclusion

We presented QCC, a closed-loop multi-agent diagnostic framework that treats evidence as a first-class optimization target in interactive clinical reasoning. By jointly quantifying evidence importance, guiding direction-aware evidence collection, and verifying evidence–diagnosis consistency, QCC enables coherent evidence accumulation and more reliable decision-making under incomplete and noisy emergency conditions. Extensive experiments across general diagnostic and real-world ED benchmarks demonstrate that evidence-oriented interaction substantially improves both evidence quality and diagnostic accuracy, highlighting the importance of explicitly modeling evidence dynamics in medical multi-agent systems.

Acknowledgment. This work was supported by Shenzhen Medical Research Fund C2501016, Science and Technology Innovation Committee of Shenzhen Municipality under grant No. JCYJ20250604145426036, the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University under grant No. VRLAB2026A01, and Sanming Project of Medicine in Shenzhen under grant No. SZSM202411032.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Croskerry, P.: The importance of cognitive errors in diagnosis and strategies to minimize them. Academic Medicine **78**(8), 775–780 (2003). https://doi.org/10.1097/00001888-200308000-00003
2. Deng, X., Miletic, V., Trinh, E., Gao, J., Xu, C., Liu, D.: Denteval: Fine-tuning-free expert-aligned assessment in dental education via llm agents. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 144–153. Springer (2025)
3. Du, Y., Li, S., Torralba, A., Tenenbaum, J.B., Mordatch, I.: Improving factuality and reasoning in language models through multiagent debate. arXiv preprint arXiv:2305.14325 (2023)
4. Fan, Z., Wei, L., Tang, J., Chen, W., Siyuan, W., Wei, Z., Huang, F.: Ai hospital: Benchmarking large language models in a multi-agent medical interaction simulator. In: Proceedings of the 31st International Conference on Computational Linguistics. pp. 10183–10213 (2025)
5. Graber, M.L., Franklin, N., Gordon, R.: Diagnostic error in internal medicine. Archives of Internal Medicine **165**(13), 1493–1499 (2005). https://doi.org/10.1001/archinte.165.13.1493
6. Guyatt, G.H., Sackett, D.L., Sinclair, J.C., Hayward, R., Cook, D.J., Cook, R.J.: Evidence-based medicine: A new approach to teaching the practice of medicine. JAMA **268**(17), 2420–2425 (1992)

7. Johnson, A., Bulgarelli, L., Pollard, T., Celi, L.A., Mark, R., Horng, S.: MIMIC-IV-ED. PhysioNet (Jan 2023). https://doi.org/10.13026/5ntk-km72, https://doi.org/10.13026/5ntk-km72, version 2.2
8. Johri, S., Jeong, J., Tran, B.A., Schlessinger, D.I., Wongvibulsin, S., Cai, Z.R., Daneshjou, R., Rajpurkar, P.: Craft-md: A conversational evaluation framework for comprehensive assessment of clinical llms. In: AAAI 2024 Spring Symposium on Clinical Foundation Models (2024)
9. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advances in Neural Information Processing Systems. vol. 33 (2020)
10. Long, Z., Bao, Z., Wei, Z.: Strong reasoning isn't enough: Evaluating evidence elicitation in interactive diagnosis. arXiv preprint arXiv:2601.19773 (2026)
11. Lyu, X., Liang, Y., Chen, W., Ding, M., Yang, J., Huang, G., Zhang, D., He, X., Shen, L.: Wsi-agents: A collaborative multi-agent system for multi-modal whole slide image analysis. In: 28th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2025). pp. 680–690. Springer (2026)
12. Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., et al.: Self-refine: Iterative refinement with self-feedback. In: Advances in Neural Information Processing Systems. vol. 36 (2023)
13. Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.t., Koh, P.W., Iyyer, M., Zettlemoyer, L., Hajishirzi, H.: FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (2023)
14. Molina, M., Mehandru, N., Golchini, N., Alaa, A.: ER-REASON: A Benchmark Dataset for LLM-Based Clinical Reasoning in the Emergency Room. PhysioNet (Oct 2025). https://doi.org/10.13026/55s7-3c27, https://doi.org/10.13026/55s7-3c27, version 1.0.0
15. Sackett, D.L., Rosenberg, W.M., Gray, J.A.M., Haynes, R.B., Richardson, W.S.: Evidence based medicine: What it is and what it isn't. BMJ **312**(7023), 71–72 (1996). https://doi.org/10.1136/bmj.312.7023.71
16. Schmidgall, S., Ziaei, R., Harris, C., Reis, E., Jopling, J., Moor, M.: Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. arXiv preprint arXiv:2405.07960 (2024)
17. Shinn, N., Cassano, F., Berman, E., Gopinath, A., Narasimhan, K., Yao, S.: Reflexion: Language agents with verbal reinforcement learning. In: Advances in Neural Information Processing Systems. vol. 36 (2023)
18. Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., et al.: Towards expert-level medical question answering with large language models. arXiv preprint arXiv:2305.09617 (2023)
19. Tang, X., Zou, A., Zhang, Z., Zhao, Y., Zhang, X., Cohan, A., Gerstein, M.: MedAgents: Large language models as collaborators for zero-shot medical reasoning. arXiv preprint arXiv:2311.10537 (2023)
20. Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., Ting, D.S.W.: Large language models in medicine. Nature Medicine **29**(8), 1930–1940 (2023)
21. Tu, T., Palepu, A., Schaekermann, M., Saab, K., Freyberg, J., Tanno, R., Wang, A., Li, B., Amin, M., et al.: Towards conversational diagnostic AI. arXiv preprint arXiv:2401.05654 (2024)

22. Tversky, A., Kahneman, D.: Judgment under uncertainty: Heuristics and biases. Science **185**(4157), 1124–1131 (1974). https://doi.org/10.1126/science.185.4157.1124
23. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. In: International Conference on Learning Representations (2023)
24. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
25. Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., et al.: AutoGen: Enabling next-gen LLM applications via multi-agent conversation. arXiv preprint arXiv:2308.08155 (2023)
26. Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T.L., Cao, Y., Narasimhan, K.: Tree of thoughts: Deliberate problem solving with large language models. In: Advances in Neural Information Processing Systems (2023)
27. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., Cao, Y.: ReAct: Synergizing reasoning and acting in language models. In: International Conference on Learning Representations (ICLR) (2023)
28. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K.R., Cao, Y.: React: Synergizing reasoning and acting in language models. In: The eleventh international conference on learning representations (2022)
29. Zhang, J., Xu, M., Wang, Y., Dou, Q.: Csap-assist: Instrument-agent dialogue empowered vision-language models for collaborative surgical action planning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 139–148. Springer (2025)
30. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., et al.: Judging LLM-as-a-judge with MT-bench and chatbot arena. In: Advances in Neural Information Processing Systems. vol. 36 (2023)