AgentRx: A Benchmark for Multimodal Clinical Forecasting with LLM Agents

Baraa Al Jorf [1] **Farah E. Shamout** [1]

Abstract

Effective clinical forecasting requires integrating heterogeneous multimodal data, including electronic health records, images, and clinical notes. While Large Language Model (LLM) agents present a promising solution to mitigate healthcare data fragmentation, their effectiveness in multimodal clinical risk forecasting remains largely unexplored. To address this, we introduce AgentRx, a systematic benchmark evaluating single and multi-agent LLM frameworks across unimodal and multimodal clinical prediction tasks using real-world data. Our findings highlight that single agent frameworks outperform naive multi-agent systems, are better at handling multimodal data, and are better calibrated.

1. Introduction

The integration of Artificial Intelligence (AI) into clinical decision support systems promises more optimized clinical workflows and better patient outcomes, particularly in resource-constrained settings such as the Intensive Care Unit (ICU). In such settings, forecasting a patient's trajectory relies on diverse data modalities routinely collected over time, such as Patient Summaries (PS), Electronic Health Records (EHR) data, Chest X-ray (CXR) images, Radiology Reports (RR), and Discharge Notes (DN). While State-Of-The-Art (SOTA) deep learning architectures successfully fuse these temporal streams to predict risk outcomes (Khader et al., 2023; Hayat et al., 2022; Al Jorf & Shamout, 2025), their real-world adoption is impeded by two key limitations. First, their black-box nature undermines clinical trust and interpretability (Catalina et al., 2023; von Eschenbach, 2021; Shuaib, 2024). Second, they typically have fixed data requirements, which restrict the transferability of these approaches across other modalities and clinical settings.

[1]New York University Abu Dhabi, UAE. Correspondence to: Baraa Al Jorf <baraa.al.jorf@nyu.edu>, Farah E. Shamout <fs999@nyu.edu>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

Large Language Models (LLMs) offer a compelling alternative for forward-looking clinical support (Tai-Seale et al., 2024). Unlike black-box models, LLMs can function as semantic translators, converting probabilistic risk trajectories into intuitive, natural language explanations (Lee et al., 2025). However, their predictive performance remains inconsistent. While some studies show promise on unimodal temporal tasks (Acharya et al., 2024; Jin et al., 2025), others report underperformance compared to supervised methods (Tan et al., 2024; Zhu et al., 2025). More importantly, all these studies have been unimodal with the potential of LLMs to handle multimodal data integration for robust clinical risk forecasting remaining largely unexplored.

To address this gap, we investigate the effectiveness of agentic systems at multimodal clinical prediction tasks. Our motivation stems from the fragmented nature of healthcare data, where clinical information is often dispersed across multiple isolated databases that ideally would be managed by modality-specific agents to avoid expensive data transfers. However, it remains unclear whether such decentralized, agentic approaches perform well across a variety of data availability settings. Hence, we introduce `AgentRx`, a comprehensive evaluation framework to analyze agent performance across three progressive settings. First, we establish performance baselines using only a single modality, specifically clinical notes encompassed within patient summaries. Second, we assess the capacity of a single agent to synthesize heterogeneous multimodal data within one context window, utilizing modality dropping ablations to measure robustness. Finally, we investigate whether multi-agent reasoning across specialized agents can improve performance compared to single agent approaches.

In summary, we make the following contributions:

- We provide a systematic benchmark (`AgentRx`) for evaluating LLM-based clinical prediction tasks using four modalities: EHR, RR, CXR, and PS.
- We analyze the impact of data heterogeneity on model discriminative ability and calibration by varying the modalities available to the agents.
- We publicly release our code and framework to support the advancement of agentic AI forecasting: `https://github.com/nyuad-cai/AgentRX`.

2. Related Work

LLMs and Health Agents. While generative LLMs and Vision-Language Models (VLMs) excel at static reasoning and medical QA (Singhal et al., 2023; Li et al., 2023; Afshar et al., 2025), their zero-shot ability to forecast temporal clinical outcomes often lags behind supervised baselines (Tan et al., 2024; Kalpelbe et al., 2025). To avoid expensive fine-tuning, inference-time reasoning frameworks like Medprompt (Nori et al., 2023) and AgentMD (Jin et al., 2025) combine strategies like Chain-of-Thought (Wei et al., 2022) and tool use to improve risk stratification. However, these methods primarily optimize isolated reasoning on cross-sectional data, rather than continuous forecasting over longitudinal timelines.

Multi-Agent Systems & Benchmarks. Multi-agent systems enhance reasoning by simulating collaboration to mitigate hallucinations (Hong et al., 2023; Liang et al., 2024). In healthcare, frameworks like MedAgents (Tang et al., 2024) achieve SOTA on static QA. Yet, recent benchmarks show agents frequently underperform specialized models on temporal forecasting tasks (Zhu et al., 2025). Crucially, current medical AI evaluation reveals a distinct divide: traditional deep learning focuses on forecasting endpoints (e.g., mortality, LOS) (Bae et al., 2023; Elsharief et al., 2025), while agentic benchmarks prioritize static retrieval (Kim et al., 2024). The few existing agentic prediction benchmarks are strictly unimodal (Tan et al., 2024; Zhu et al., 2025). This reveals a critical gap for evaluating collaborative agentic systems on multimodal, longitudinal forecasting, which forms the core motivation of `AgentRx`.

3. Methodology

3.1. Preliminaries & Datasets

To formalize the `AgentRx` benchmark, we assume for a given patient encounter p, the presence of multimodal historical data $\mathcal{X}_p = \{\mathbf{x}_{ps}, \mathbf{x}_{ehr}, \mathbf{x}_{cxr}, \mathbf{x}_{rr}\}$. Here, $\mathbf{x}_{ps}$ represents a textual patient summary (demographics, history), $\mathbf{x}_{ehr} \in \mathbb{R}^{d \times t}$ represents the multivariate time-series data from the EHR with d features over t time steps, $\mathbf{x}_{cxr} \in \mathbb{R}^{h \times w}$ represents a CXR image, and $\mathbf{x}_{rr}$ represents the textual radiology reports. The system's objective is to forecast a future clinical endpoint y.

We curated our dataset from MIMIC-IV (Johnson et al., 2023b), MIMIC-CXR (Johnson et al., 2019), and MIMIC-IV-Notes (Johnson et al., 2023a), aligning subject and admission identifiers. Patient summaries ($\mathbf{x}_{ps}$) are mandated for all samples as the base modality, representing the clinical context available upon admission. To prevent data leakage, we extracted only the information available prior to ICU admission. This includes the patient's medical, surgical, and family histories as well as demographic details such as age and sex, ensuring the agent operates strictly with the same context available to a clinician at the exact time of admission. Other modalities are paired if available within the initial observation window. Appendix A.1 details the exact modality splits across our dataset.

3.2. Forecasting Tasks

We evaluate the frameworks across two temporal forecasting tasks using the first 48 hours of ICU data, reporting AUROC, AUPRC, and Expected Calibration Error (ECE):

- **In-hospital Mortality:** A binary classification task forecasting the risk of mortality based on the initial 48-hour trajectory.
- **Length of Stay (LOS):** A binary classification task projecting whether a patient's stay will extend beyond 7 days.

All baselines were evaluated using a zero-shot paradigm on the test set which represented 20% of the full dataset. To accommodate LLM context limits, EHR data was serialized into a structured format (`[`T_0+ΔT`] Variable=Value`). High-frequency streams exceeding 500 time-steps were truncated to preserve the admission state and the most recent clinical trajectory prior to the forecasting horizon.

3.3. Agentic Frameworks & Baselines

We formalize an agent $\mathcal{A}(\cdot)$ as a functional unit processing clinical context to produce a forecast probability. We evaluate three primary settings (Appendix B.1):

1. **Single Agent Unimodal:** The agent relies solely on the Patient Summary ($\mathbf{x}_{ps}$).
2. **Single Agent Multimodal:** A generalist agent synthesizes all available modalities $\mathcal{X}_{available}$ simultaneously within a single context window.
3. **Multi-Agent Multimodal:** Specialized agents $\mathcal{A}_m$ independently process distinct data streams and generate intermediate probabilities, which are then aggregated.

We benchmark three VLM backbones (Qwen2.5-VL, HuatuoGPT-Vision, and LLaVA-Med) across various reasoning strategies (Zero-shot, Few-shot, CoT, CoT-SC, Self-Refine) and collaborative multi-agent architectures (Majority Vote, Debate, Meta-Prompting, Traj-CoA, MedAgents, and MDAgents). Appendix C defines all these baselines in more detail.

Crucially, we compare these agentic systems against two specialized deep learning architectures: BioBERT (Lee et al., 2020) for the unimodal setting, and MedPatch (Al Jorf

Table 1. **Single Agent Unimodal Results.** Comparison of agentic backbones against a supervised baseline (BioBERT) using only the PS. Metrics are reported with 95% Confidence Intervals. The best overall performance in each column is bolded.

Backbone	Method	In-Hospital Mortality			Length of Stay (>7 Days)		
		AUROC	AUPRC	ECE	AUROC	AUPRC	ECE
Supervised	BioBERT	0.680 (0.657 - 0.702)	0.228 (0.203 - 0.263)	**0.006**	**0.641 (0.621 - 0.661)**	**0.307 (0.282 - 0.337)**	**0.024**
Qwen	Zero-shot	0.667 (0.645 - 0.690)	0.234 (0.208 - 0.267)	0.025	0.603 (0.582 - 0.624)	0.261 (0.241 - 0.286)	0.105
	Few-shot	0.697 (0.678 - 0.718)	0.236 (0.213 - 0.270)	0.042	0.602 (0.580 - 0.621)	0.257 (0.237 - 0.280)	0.122
	CoT	0.650 (0.629 - 0.675)	0.214 (0.191 - 0.244)	0.060	0.592 (0.573 - 0.613)	0.260 (0.240 - 0.286)	0.480
	CoT-SC	0.679 (0.657 - 0.702)	0.236 (0.209 - 0.269)	0.029	0.593 (0.571 - 0.613)	0.259 (0.239 - 0.284)	0.454
	Self-Refine	0.666 (0.644 - 0.689)	0.212 (0.187 - 0.240)	0.080	0.580 (0.560 - 0.601)	0.236 (0.219 - 0.257)	0.198
HuaTuo	Zero-shot	0.692 (0.672 - 0.714)	**0.238 (0.213 - 0.268)**	0.166	0.599 (0.578 - 0.620)	0.256 (0.236 - 0.281)	0.203
	Few-shot	**0.700 (0.681 - 0.721)**	0.227 (0.205 - 0.258)	0.093	0.602 (0.580 - 0.621)	0.260 (0.240 - 0.284)	0.029
	CoT	0.670 (0.650 - 0.693)	0.219 (0.194 - 0.247)	0.211	0.595 (0.574 - 0.615)	0.258 (0.239 - 0.280)	0.599
	CoT-SC	0.670 (0.649 - 0.692)	0.218 (0.194 - 0.246)	0.211	0.594 (0.574 - 0.614)	0.258 (0.239 - 0.281)	0.599
	Self-Refine	0.671 (0.651 - 0.694)	0.213 (0.190 - 0.242)	0.106	0.581 (0.560 - 0.601)	0.249 (0.230 - 0.273)	0.385
Llava	Zero-shot	0.642 (0.621 - 0.666)	0.204 (0.182 - 0.235)	0.755	0.574 (0.552 - 0.594)	0.234 (0.216 - 0.255)	0.787
	Few-shot	0.684 (0.664 - 0.706)	0.233 (0.208 - 0.265)	0.776	0.605 (0.584 - 0.624)	0.253 (0.235 - 0.273)	0.780
	CoT	0.633 (0.610 - 0.655)	0.170 (0.154 - 0.186)	0.831	0.509 (0.499 - 0.518)	0.194 (0.183 - 0.205)	0.808
	CoT-SC	0.645 (0.622 - 0.666)	0.197 (0.174 - 0.221)	0.840	0.542 (0.523 - 0.560)	0.206 (0.194 - 0.221)	0.806
	Self-Refine	0.631 (0.608 - 0.655)	0.170 (0.154 - 0.187)	0.829	0.516 (0.504 - 0.526)	0.196 (0.185 - 0.208)	0.808

& Shamout, 2025), a SOTA multimodal fusion framework that utilizes a confidence-guided patching mechanism to effectively integrate heterogeneous data streams.

4. Results

4.1. Unimodal Forecasting Performance

In Table 1, agents are restricted to the PS modality to evaluate base clinical reasoning. For mortality forecasting, specialized LLMs exhibit the capacity to exceed supervised baselines. The medical-specific HuaTuo backbone surpasses BioBERT's AUROC (0.680) in the few-shot setting (0.700) and achieves the highest AUPRC in the zero-shot (0.238).

However, despite strong discriminative performance, generative models exhibit severe miscalibration. The supervised BioBERT baseline maintains a minimal ECE of 0.006, whereas all agentic setups yield considerably higher errors (e.g., HuaTuo Few-shot at 0.093, Llava exceeding 0.750). This indicates that while agents can effectively rank patient risk trajectories, their probabilistic confidence is substantially less reliable than supervised counterparts. For LOS, BioBERT maintains superiority across all metrics, suggesting that clinical LLM pre-training benefits diagnostic endpoints more than operational forecasting.

4.2. Multimodal Forecasting Performance

Table 2 details performance when integrating full heterogeneous data. A significant gap persists between agentic frameworks and the supervised MedPatch. MedPatch achieves a dominant AUROC of 0.877 for mortality and 0.844 for LOS, while the best agentic configuration trails at 0.763 (Qwen few-shot). This highlights a structural limitation: while LLMs possess strong semantic reasoning, they lack the dedicated fusion mechanisms optimized for temporal and visual inputs.

Despite this gap, multimodal integration yields considerable gains over unimodal agent baselines. Qwen's mortality AUROC improves from 0.667 (unimodal) to 0.756 (multimodal zero-shot). Interestingly, distributing this reasoning across multiple specialized agents often degrades outcomes. Qwen-based Debate (0.631) and Meta-Prompt (0.599) architectures severely underperform the single-agent zero-shot baseline. Traj-CoA remains the strongest multi-agent framework (0.762) specifically because it employs a final decision-maker that retains direct access to the aggregated multimodal context rather than relying purely on decentralized probabilities.

Furthermore, we conducted ablations assessing the impact of gradually adding modalities in the zero-shot single agent and the majority vote setups (in Appendix D.1). We note the the single-agent ECE remains consistent, whereas the multi-agent ECE increases as more modalities are added. We also observed a failure of consensus in collaborative settings (in Appendix E.1). In the Debate baseline, backbones like LLavaMed failed to reach consensus for over half of the patient cohort, defaulting to averaged probabilities. We noted a direct correlation between prolonged inter-agent debate and degraded AUROC performance, suggesting current LLM debate mechanics amplify noise.

Table 2. **Multimodal Results.** Comparison of Supervised, Single-Agent, and Multi-Agent architectures on the 4-modality dataset. Best overall performance in each column is bolded.

Backbone	Arch.	Method	In-Hospital Mortality			Length of Stay (>7 Days)		
			AUROC	AUPRC	ECE	AUROC	AUPRC	ECE
Supervised	-	MedPatch	**0.877 (0.864 - 0.888)**	**0.546 (0.504 - 0.585)**	**0.019**	**0.844 (0.830 - 0.857)**	**0.551 (0.517 - 0.587)**	**0.025**
Qwen	Single	Zero-shot	0.756 (0.737 - 0.776)	0.330 (0.297 - 0.368)	0.023	0.714 (0.694 - 0.731)	0.345 (0.320 - 0.373)	0.411
		Few-shot	0.763 (0.744 - 0.783)	0.325 (0.292 - 0.361)	0.032	0.682 (0.662 - 0.699)	0.318 (0.294 - 0.346)	0.479
		CoT	0.733 (0.713 - 0.752)	0.274 (0.244 - 0.308)	0.049	0.683 (0.663 - 0.702)	0.311 (0.288 - 0.336)	0.676
		CoT-SC	0.762 (0.742 - 0.782)	0.337 (0.300 - 0.379)	0.039	0.698 (0.678 - 0.715)	0.340 (0.313 - 0.370)	0.661
	Multi	Majority Vote	0.748 (0.727 - 0.768)	0.315 (0.282 - 0.355)	0.111	0.710 (0.690 - 0.728)	0.352 (0.324 - 0.383)	0.046
		Debate	0.631 (0.607 - 0.656)	0.210 (0.185 - 0.243)	0.091	0.644 (0.623 - 0.664)	0.281 (0.259 - 0.307)	0.121
		Meta-Prompt	0.599 (0.573 - 0.625)	0.179 (0.160 - 0.204)	0.051	0.537 (0.514 - 0.558)	0.226 (0.208 - 0.249)	0.086
		Traj-CoA	0.762 (0.743 - 0.780)	0.318 (0.283 - 0.355)	0.039	0.708 (0.688 - 0.726)	0.336 (0.310 - 0.365)	0.190
		MDAgents	0.624 (0.601 - 0.647)	0.192 (0.169 - 0.221)	0.110	0.584 (0.563 - 0.603)	0.240 (0.222 - 0.262)	0.138
		MedAgents	0.662 (0.641 - 0.686)	0.206 (0.185 - 0.235)	0.034	0.634 (0.613 - 0.654)	0.285 (0.261 - 0.310)	0.019
HuaTuo	Single	Zero-shot	0.762 (0.743 - 0.782)	0.325 (0.296 - 0.362)	0.049	0.704 (0.686 - 0.721)	0.331 (0.307 - 0.358)	0.418
		Few-shot	0.763 (0.744 - 0.782)	0.324 (0.291 - 0.361)	0.167	0.703 (0.684 - 0.720)	0.332 (0.308 - 0.358)	0.514
		CoT	0.697 (0.678 - 0.720)	0.233 (0.208 - 0.266)	0.306	0.639 (0.617 - 0.658)	0.282 (0.260 - 0.306)	0.671
		CoT-SC	0.696 (0.678 - 0.719)	0.233 (0.208 - 0.265)	0.305	0.639 (0.617 - 0.658)	0.282 (0.261 - 0.306)	0.669
	Multi	Majority Vote	0.711 (0.690 - 0.730)	0.245 (0.217 - 0.279)	0.050	0.691 (0.671 - 0.709)	0.322 (0.296 - 0.352)	0.087
		Debate	0.628 (0.604 - 0.652)	0.188 (0.165 - 0.216)	0.109	0.621 (0.599 - 0.642)	0.273 (0.250 - 0.300)	0.061
		Meta-Prompt	0.636 (0.614 - 0.657)	0.179 (0.160 - 0.202)	0.428	0.558 (0.534 - 0.578)	0.228 (0.210 - 0.250)	0.080
		Traj-CoA	0.744 (0.725 - 0.766)	0.295 (0.266 - 0.333)	0.065	0.701 (0.680 - 0.718)	0.326 (0.301 - 0.353)	0.323
		MDAgents	0.490 (0.467 - 0.514)	0.125 (0.112 - 0.142)	0.309	0.571 (0.551 - 0.590))	0.237 (0.218 - 0.260)	0.059
		MedAgents	0.606 (0.584 - 0.629)	0.165 (0.148 - 0.187)	0.335	0.584 (0.563 - 0.606)	0.230 (0.213 - 0.252)	0.486
Llava	Single	Zero-shot	0.741 (0.721 - 0.761)	0.268 (0.242 - 0.303)	0.835	0.613 (0.596 - 0.630)	0.242 (0.226 - 0.259)	0.803
		Few-shot	0.684 (0.662 - 0.706)	0.225 (0.199 - 0.255)	0.704	0.581 (0.561 - 0.602)	0.233 (0.217 - 0.252)	0.786
		CoT	0.676 (0.658 - 0.693)	0.184 (0.167 - 0.202)	0.819	0.553 (0.537 - 0.567)	0.210 (0.197 - 0.223)	0.807
		CoT-SC	0.691 (0.670 - 0.714)	0.215 (0.192 - 0.240)	0.806	0.583 (0.563 - 0.602)	0.226 (0.211 - 0.242)	0.806
	Multi	Majority Vote	0.710 (0.686 - 0.732)	0.280 (0.248 - 0.316)	0.812	0.674 (0.654 - 0.693)	0.291 (0.268 - 0.316)	0.765
		Debate	0.495 (0.470 - 0.519)	0.121 (0.109 - 0.137)	0.293	0.506 (0.485 - 0.525)	0.192 (0.178 - 0.210)	0.437
		Meta-Prompt	0.659 (0.638 - 0.683)	0.211 (0.189 - 0.237)	0.810	0.576 (0.555 - 0.594)	0.231 (0.214 - 0.250)	0.780
		Traj-CoA	0.694 (0.673 - 0.718)	0.249 (0.221 - 0.283)	0.758	0.608 (0.589 - 0.626)	0.239 (0.223 - 0.257)	0.795
		MDAgents	0.565 (0.544 - 0.588)	0.143 (0.130 - 0.160)	0.771	0.492 (0.473 - 0.512)	0.194 (0.181 - 0.210)	0.759
		MedAgents	0.609 (0.589 - 0.633)	0.164 (0.147 - 0.184)	0.683	0.545 (0.525 - 0.563)	0.207 (0.193 - 0.223)	0.782

5. Discussion

In this work, we introduced `AgentRx`, a systematic benchmark for evaluating LLM-based agentic systems on high-stakes clinical forecasting endpoints. By evaluating across progressive data settings, we yield a critical insight: multi-agent multimodal systems consistently underperform single-agent multimodal systems. This degradation is driven by a fundamental divergence in calibration; decentralized voting and debate mechanisms fail to capture the uncertainty reduction inherent in fusing distinct temporal modalities, leading to systemic overconfidence.

Our findings align with recent benchmarks like `MedAgentBoard` (Zhu et al., 2025), confirming that agentic systems trail state-of-the-art supervised fusion networks. However, we identify a modality-dependent nuance. While LLMs struggle to encode high-dimensional, structured EHR time-series, our unimodal results demonstrate that specialized agents can actually outperform supervised baselines on free-text (PS). This suggests the performance gap is not intrinsic to the task of temporal forecasting, but rather a consequence of the models' inability to natively fuse non-textual data streams without robust architectural support.

Despite these contributions, this study has limitations. Our reliance on the single-center MIMIC database poses generalizability constraints. Furthermore, serializing high-frequency EHR data into text-based context windows remains a bottleneck, limiting the agents' ability to capture subtle physiological trajectories. Future work will expand `AgentRx` to diverse clinical endpoints and investigate hybrid architectures that merge the interpretable semantic reasoning of LLMs with the predictive precision of frozen, multimodal supervised encoders.

Acknowledgements

This work was supported by ASPIRE, the technology program management pillar of Abu Dhabi's Advanced Technology Research Council (ATRC), via the ASPIRE Precision Medicine Research Institute Abu Dhabi (ASPIREPMRIAD) award grant number VRI-20-10, the NYUAD Center for Artificial Intelligence and Robotics, funded by Tamkeen under the NYUAD Research Institute Award CG010, the Meem Foundation, and the ADIA Lab Health Sciences Grant. The research was carried out on the High Performance Computing resources at New York University Abu Dhabi.

References

Acharya, A., Shrestha, S., Chen, A., Conte, J., Avramovic, S., Sikdar, S., Anastasopoulos, A., and Das, S. Clinical risk prediction using language models: benefits and considerations. *Journal of the American Medical Informatics Association: JAMIA*, 31(9):1856–1864, September 2024. ISSN 1527-974X. doi: 10.1093/jamia/ocae030.

Afshar, M., Ryan Baumann, M., Resnik, F., Hintzke, J., Gravel Sullivan, A., Wills, G., Lemmon, K., Dambach, J., Mrotek, L. A., Quinn, M., Abramson, K., Kleinschmidt, P., Brazelton, T. B., Leaf, M. A., Twedt, H., Kunstman, D., Patterson, B., Liao, F., Rasmussen, S., Burnside, E. S., Goswami, C., and Gordon, J. A Pragmatic Randomized Controlled Trial of Ambient Artificial Intelligence to Improve Health Practitioner Well-Being. *NEJM AI*, 2(12):AIoa2500945, November 2025. doi: 10.1056/AIoa2500945. URL https://ai.nejm.org/doi/10.1056/AIoa2500945. Publisher: Massachusetts Medical Society.

Al Jorf, B. and Shamout, F. E. MedPatch: Confidence-Guided Multi-Stage Fusion for Multimodal Clinical Data. 2025. URL https://www.mlforhc.org/s/QBWQPiaXvm_camera_ready-Baraa-Al-Jorf.pdf.

Bae, S., Kyung, D., Ryu, J., Cho, E., Lee, G., Kweon, S., Oh, J., Ji, L., Chang, E., Kim, T., and Choi, E. EHRXQA: A Multi-Modal Question Answering Dataset for Electronic Health Records with Chest X-ray Images. *Advances in Neural Information Processing Systems*, 36:3867–3880, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/0c007ebef1d11fd48da6ce4f54687db6-Abstract-Datasets_and_Benchmarks.html.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language Models are Few-Shot Learners, May 2020. URL https://arxiv.org/abs/2005.14165v4.

Catalina, Q. M., Fuster-Casanovas, A., Vidal-Alaball, J., Escalé-Besa, A., Marin-Gomez, F. X., Femenia, J., and Solé-Casals, J. Knowledge and perception of primary care healthcare professionals on the use of artificial intelligence as a healthcare tool. *DIGITAL HEALTH*, 9: 20552076231180511, January 2023. ISSN 2055-2076. doi: 10.1177/20552076231180511. URL https://doi.org/10.1177/20552076231180511. Publisher: SAGE Publications Ltd.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pp. 11733–11763, Vienna, Austria, 2024. JMLR.org.

Elsharief, S., Shurrab, S., Jorf, B. A., Lopez, L. J. L., Geras, K. J., and Shamout, F. E. MedMod: Multimodal Benchmark for Medical Prediction Tasks with Electronic Health Records and Chest X-Ray Scans. In *Proceedings of the sixth Conference on Health, Inference, and Learning*, pp. 781–803. PMLR, July 2025. URL https://proceedings.mlr.press/v287/elsharief25a.html.

Hayat, N., Geras, K. J., and Shamout, F. E. MedFuse: Multimodal fusion with clinical time-series data and chest X-ray images. In *Proceedings of the 7th Machine Learning for Healthcare Conference*, pp. 479–503. PMLR, December 2022. URL https://proceedings.mlr.press/v182/hayat22a.html. ISSN: 2640-3498.

Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Zhang, C., Wang, J., Wang, Z., Yau, S. K. S., Lin, Z., Zhou, L., Ran, C., Xiao, L., Wu, C., and Schmidhuber, J. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework, August 2023. URL https://arxiv.org/abs/2308.00352v7.

Hou, Y., Dong, H., Wang, X., Li, B., and Che, W. MetaPrompting: Learning to Learn Better Prompts. In Calzolari, N., Huang, C.-R., Kim, H., Pustejovsky, J., Wanner, L., Choi, K.-S., Ryu, P.-M., Chen, H.-H., Donatelli, L., Ji, H., Kurohashi, S., Paggio, P., Xue, N., Kim, S., Hahm, Y., He, Z., Lee, T. K., Santus, E., Bond, F., and Na, S.-H. (eds.), *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 3251–3262, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.287/.

Jin, Q., Wang, Z., Yang, Y., Zhu, Q., Wright, D., Huang, T., Khandekar, N., Wan, N., Ai, X., Wilbur, W. J., He, Z., Taylor, R. A., Chen, Q., and Lu, Z. AgentMD: Empowering language agents for risk prediction with large-scale clinical tool learning. *Nature Communications*, 16(1):9377, October 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-64430-x. URL https://www.nature.com/articles/s41467-025-64430-x.

Johnson, A., Pollard, T., Horng, S., Celi, L. A., and Mark, R. MIMIC-IV-Note: Deidentified free-text clinical notes, 2023a. URL https://physionet.org/content/mimic-iv-note/2.2/.

Johnson, A. E. W., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Mark, R. G., and Horng, S. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, December 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0322-0. URL https://www.nature.com/articles/s41597-019-0322-0. Publisher: Nature Publishing Group.

Johnson, A. E. W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., Lehman, L.-w. H., Celi, L. A., and Mark, R. G. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, January 2023b. ISSN 2052-4463. doi: 10.1038/s41597-022-01899-x. URL https://www.nature.com/articles/s41597-022-01899-x.

Kaesberg, L. B., Becker, J., Wahle, J. P., Ruas, T., and Gipp, B. Voting or Consensus? Decision-Making in Multi-Agent Debate. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11640–11671, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.606. URL https://aclanthology.org/2025.findings-acl.606/.

Kalpelbe, B. C., Adaambiik, A. G., and Peng, W. Vision Language Models in Medicine, February 2025. URL http://arxiv.org/abs/2503.01863. arXiv:2503.01863 [cs].

Khader, F., Kather, J. N., Müller-Franzes, G., Wang, T., Han, T., Tayebi Arasteh, S., Hamesch, K., Bressem, K., Haarburger, C., Stegmaier, J., Kuhl, C., Nebelung, S., and Truhn, D. Medical transformer for multimodal survival prediction in intensive care: integration of imaging and non-imaging data. *Scientific Reports*, 13(1):10666, July 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-37835-1. URL https://www.nature.com/articles/s41598-023-37835-1.

Kim, Y., Park, C., Jeong, H., Chan, Y. S., Xu, X., McDuff, D., Lee, H., Ghassemi, M., Breazeal, C., and Park, H. W. MDAgents: an adaptive collaboration of LLMs for medical decision-making. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, volume 37 of *NIPS '24*, pp. 79410–79452, Red Hook, NY, USA, December 2024. Curran Associates Inc. ISBN 979-8-3313-1438-5.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, February 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz682. URL https://doi.org/10.1093/bioinformatics/btz682.

Lee, S., Cho, W. I., Lee, Y., Kim, D. J., Nam, K. H., Lee, S., Suh, J., and Ko, T. A prompt framework for enhancing LLM-based explainability of medical machine learning models: an intensive care unit application. *BMC Medical Informatics and Decision Making*, 25(1):430, November 2025. ISSN 1472-6947. doi: 10.1186/s12911-025-03239-6. URL https://doi.org/10.1186/s12911-025-03239-6.

Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., and Gao, J. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day, June 2023. URL http://arxiv.org/abs/2306.00890. arXiv:2306.00890 [cs].

Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Shi, S., and Tu, Z. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17889–17904, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.992. URL https://aclanthology.org/2024.emnlp-main.992/.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Gupta, S., Majumder, B. P., Hermann, K., Welleck, S., Yazdanbakhsh, A., and Clark, P. Self-Refine: Iterative Refinement with Self-Feedback, May 2023. URL http://arxiv.org/abs/2303.17651. arXiv:2303.17651 [cs].

Nori, H., Lee, Y. T., Zhang, S., Carignan, D., Edgar, R., Fusi, N., King, N., Larson, J., Li, Y., Liu, W., Luo, R., McKinney, S. M., Ness, R. O., Poon, H., Qin, T., Usuyama, N., White, C., and Horvitz, E. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine, November 2023. URL http://arxiv.org/abs/2311.16452. arXiv:2311.16452 [cs].

Shuaib, A. Transforming Healthcare with AI: Promises, Pitfalls, and Pathways Forward. *International Journal of General Medicine*, 17:1765–1771, May 2024. doi: 10.2147/IJGM.S449598. URL https://www.dovepress.com/transforming-healthcare-with-ai-promises-pitfalls-and-pathways-forward-peer-reviewed-fulltext-article-IJGM. Publisher: Dove Press.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., Agüera y Arcas, B., Webster, D., Corrado, G. S., Matias, Y., Chou, K., Gottweis, J., Tomasev, N., Liu, Y., Rajkomar, A., Barral, J., Semturs, C., Karthikesalingam, A., and Natarajan, V. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, August 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06291-2. URL https://www.nature.com/articles/s41586-023-06291-2.

Tai-Seale, M., Baxter, S. L., Vaida, F., Walker, A., Sitapati, A. M., Osborne, C., Diaz, J., Desai, N., Webb, S., Polston, G., Helsten, T., Gross, E., Thackaberry, J., Mandvi, A., Lillie, D., Li, S., Gin, G., Achar, S., Hofflich, H., Sharp, C., Millen, M., and Longhurst, C. A. AI-Generated Draft Replies Integrated Into Health Records and Physicians' Electronic Communication. *JAMA Network Open*, 7(4): e246565, April 2024. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2024.6565. URL https://doi.org/10.1001/jamanetworkopen.2024.6565.

Tan, M., Merrill, M. A., Gupta, V., Althoff, T., and Hartvigsen, T. Are language models actually useful for time series forecasting? In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, volume 37 of *NIPS '24*, pp. 60162–60191, Red Hook, NY, USA, December 2024. Curran Associates Inc. ISBN 979-8-3313-1438-5.

Tang, X., Zou, A., Zhang, Z., Li, Z., Zhao, Y., Zhang, X., Cohan, A., and Gerstein, M. MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 599–621, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.33. URL https://aclanthology.org/2024.findings-acl.33/.

von Eschenbach, W. J. Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philosophy & Technology*, 34(4):1607–1622, December 2021. ISSN 2210-5441. doi: 10.1007/s13347-021-00477-0. URL https://doi.org/10.1007/s13347-021-00477-0.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-Consistency Improves Chain of Thought Reasoning in Language Models, March 2022. URL https://arxiv.org/abs/2203.11171v4.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pp. 24824–24837, Red Hook, NY, USA, November 2022. Curran Associates Inc. ISBN 978-1-7138-7108-8.

Zeng, S., Fu, Y., Zhou, S., Yu, Z., Liu, L. J., Wen, J., Thompson, M., Etzioni, R., and Yetisgen, M. Traj-CoA: Patient Trajectory Modeling via Chain-of-Agents for Lung Cancer Risk Prediction.

Zhu, Y., He, Z., Hu, H., Zheng, X., Zhang, X., Wang, Z., Gao, J., Ma, L., and Yu, L. MedAgentBoard: Benchmarking Multi-Agent Collaboration with Conventional Methods for Diverse Medical Tasks. October 2025. URL https://openreview.net/forum?id=BPpG4qQaNj.

A. Dataset

We extracted the EHR data from MIMIC-IV (Johnson et al., 2023b), the CXR images from MIMIC-CXR (Johnson et al., 2019), and the RR and PS from MIMIC-IV-Notes (Johnson et al., 2023a). The MIMIC-IV dataset includes de-identified data from over $315,460$ patients with ICU stays at the Beth Israel Deaconess Medical Center between 2008 and 2019. The MIMIC-CXR dataset consists of over $377,000$ chest radiographs. MIMIC-IV-Note supplements these two datasets with unstructured textual data, including both 331,794 DNs and 2,321,355 RRs. We constructed the multimodal dataset by aligning the subject, stay, and admission identifiers across all three datasets. We mandate the existence of PS for all samples, while other modalities are paired if available.

After multimodal pairing, the dataset was split into training (70%), validation (10%), and testing (20%) sets. The exact dataset distribution per task is reported in Table A.1.

A.1. Multimodal Data Processing and Pairing

Since patient summaries are considered to be the primary base modality, we ensure that every patient in our cohort has one. We generated patient summaries by processing the raw DNs from MIMIC-IV-Note. To prevent data leakage, we extracted only the information available prior to ICU admission. This includes the patient's medical, surgical, and family histories as well as the demographic details such as age and sex, ensuring the agent operates with the same context available to a clinician at the time of admission.

For EHR, we used a set of 17 clinical variables consistent with previous work (Hayat et al., 2022; Al Jorf & Shamout, 2025). These include 5 categorical variables (capillary refill rate, Glasgow coma scale eye opening, motor response, verbal response, and total) and 12 continuous variables (diastolic blood pressure, fraction of inspired oxygen, glucose, heart rate, height, mean blood pressure, oxygen saturation, respiratory rate, systolic blood pressure, temperature, weight, and pH). To format the EHR data for the LLM agents, we serialized it into a structured text format `[`T_0+ΔT`] Variable=Value` where T_0 is the time of admission, recording only observed measurements at each timestamp to minimize token usage and handle irregular sampling rates efficiently. The observation window was strictly limited to the first 48 hours of ICU admission. To strictly adhere to context window limits, high-frequency streams exceeding 500 time-steps were truncated to preserve the initial admission state (first 100 steps) and the most recent clinical trajectory (last 400 steps).

For the same cohort, we included CXR images that were collected within the first 48 hours of ICU admission. We restricted the selection to Anterior-Posterior (AP) views, as these are standard for portable ICU bedside imaging. For patients with multiple scans within the window, we selected the latest valid scan to capture the most recent clinical state prior to the prediction horizon.

We extracted reports corresponding to patients available in the dataset. MIMIC-IV-Note includes reports for various imaging modalities (CT, MRI, Ultrasound). We concatenated all relevant reports for each patient that were collected within 48 hours from admission into a single report aggregate that is passed as the full RR modality.

Table A.1. **Dataset statistics per modality and task.** We describe the sample counts for the training and test splits across the mortality and length of stay prediction tasks. The counts are stratified by modality.

	Mortality		Length of Stay	
Modality	**Training**	**Test**	**Training**	**Test**
PS	17,773	4,925	17,476	4,845
EHR	17,773	4,925	17,476	4,845
RR	16,128	4,454	15,865	4,380
CXR	4,259	1,174	4,171	1,153

B. Agentic Frameworks

Figure B.1 shows the overall frameworks followed by the different agentic setups evaluated.

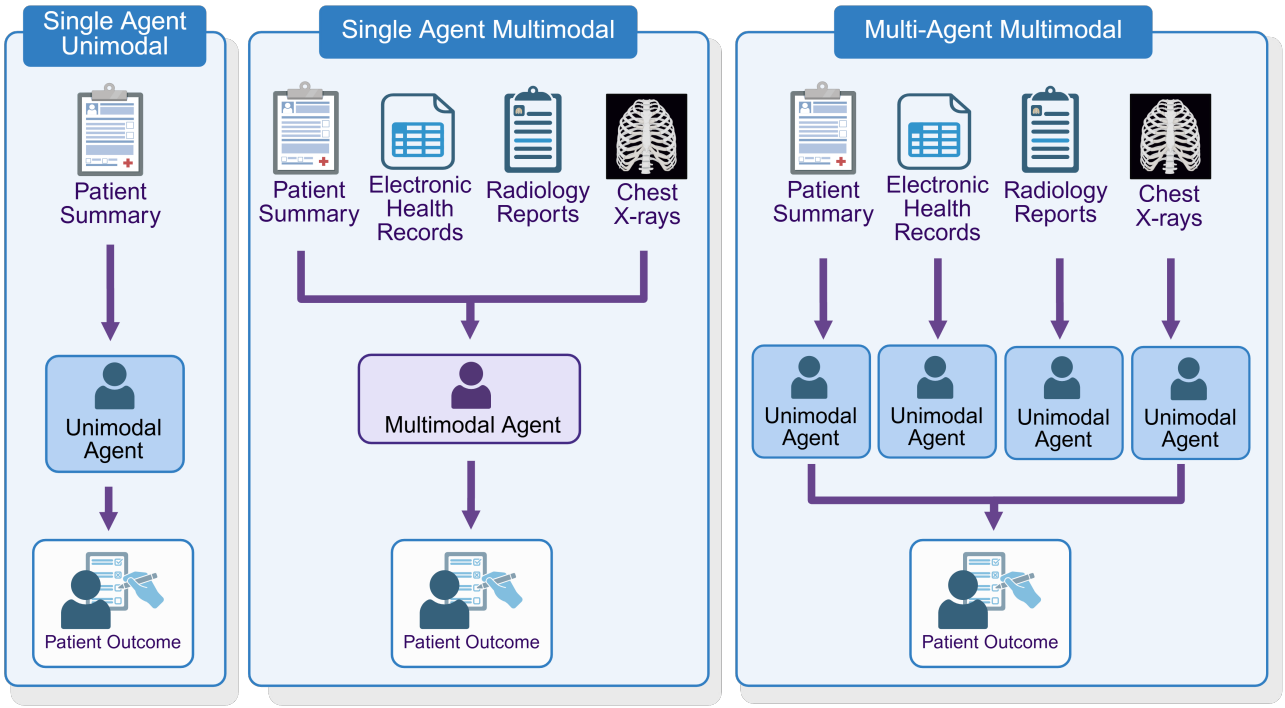

Figure B.1. Overview of the agentic evaluation frameworks considered within the `AgentRx` benchmark, spanning Single-Agent (Unimodal/Multimodal) and Multi-Agent settings. The figure was created in BioRender (Al Jorf, B. (2025) https://BioRender.com/jl2gr7s).

C. Baseline Definitions

To assess the efficacy of different agentic architectures, we compare performance across the following single-agent and multi-agent baselines.

C.1. Supervised Baselines

We compare our agentic frameworks against two specialized deep learning architectures:

1. BioBERT (Unimodal): For the text-only setting, we utilize BioBERT (Lee et al., 2020), a language model pre-trained on biomedical corpora (PubMed). We freeze the backbone and fine-tune a linear classification head on the token embeddings of the PS to establish a strong supervised baseline.
2. MedPatch (Multimodal): For the multimodal setting, we employ MedPatch (Al Jorf & Shamout, 2025), a SOTA fusion architecture that utilizes a confidence-guided patching mechanism to effectively integrate heterogeneous modalities.

C.2. Single-Agent Baselines - Unimodal and Multimodal

1. Zero-shot: Vanilla baseline where we feed the model all the setting's available modalities and ask it for a prediction.
2. Few-shot: A baseline where we feed the model one positive example and one negative example (data + labels from the training set) and then ask it to predict the outcome for a test sample (Brown et al., 2020).
3. Chain-of-Thought (CoT): A baseline where we allow the model to first generate a reasoning step, then use that reasoning to produce a prediction (Wei et al., 2022).
4. Self-Consistency + Chain-of-Thought (CoT-SC): A baseline where the model runs 3 parallel reasoning pathways and then makes a decision using all three paths via voting (probability averaging) (Wang et al., 2022).

5. Self-Refinement: A baseline where the model generates a prediction with reasoning, then self-evaluates its reasoning before making a final prediction based on the evaluation feedback (Madaan et al., 2023).

C.3. Multi-Agent Baselines

1. Majority Vote: A baseline where unimodal agents independently analyze their specific data and vote on the outcome (Kaesberg et al., 2025).
2. Debate: A baseline where unimodal agents debate with each other until they reach a consensus (Du et al., 2024).
3. Meta-Prompting: A baseline where a meta-agent evaluates the data and either makes a prediction or instantiates other expert agents to help refine its task (Hou et al., 2022).
4. Traj-CoA + Multimodal Judge: A baseline that uses multiple worker agents to construct an EHR memory. Each worker receives a chunk of EHR data. The EHR memory and the other modalities are then passed to a multimodal judge agent that makes a final prediction. This baseline combines a unimodal and a multimodal agent setup (Zeng et al.).
5. MDAgents: A baseline where a diverse ensemble of agents is intialized depending on patient case severity (Kim et al., 2024).
6. MedAgents: A baseline where a agents follow predefined roles and collaborate to create a patient state summary to enable outcome prediction (Tang et al., 2024).

D. Ablations

Table D.1 reveals that single-agent architectures improve both predictive performance and calibration (showing higher AUROC and lower ECE) with increased modality integration by leveraging cross-modal dependencies. Conversely, majority voting mechanisms suffer from degrading calibration; aggregating hard labels without shared context leads to system overconfidence and fails to capture the uncertainty reduction benefits of multimodal data.

Table D.1. **Modality Ablations.** Comparing zero-shot vs. majority vote settings for in-hospital mortality. Note that while AUROC/AUPRC scale with modality density, Single Agent calibration (ECE) improves with more data while majority vote calibration deteriorates. Best overall performance in each column is bolded.

Backbone	Arch.	Modalities	In-Hospital Mortality		
			AUROC	AUPRC	ECE
Qwen	Single (ZS)	PS	0.667 (0.645 - 0.690)	0.234 (0.208 - 0.267)	0.025
		PS + CXR	0.671 (0.649 - 0.695)	0.236 (0.209 - 0.267)	0.027
		PS + CXR + RR	0.732 (0.712 - 0.753)	0.273 (0.244 - 0.305)	0.027
		All (PS+EHR+CXR+RR)	0.756 (0.737 - 0.776)	**0.330 (0.297 - 0.368)**	0.023
	Multi (MV)	PS + CXR	0.666 (0.643 - 0.689)	0.221 (0.197 - 0.252)	0.068
		PS + CXR + RR	0.725 (0.704 - 0.745)	0.266 (0.237 - 0.301)	0.089
		All (PS+EHR+CXR+RR)	0.748 (0.727 - 0.768)	0.315 (0.282 - 0.355)	0.111
HuaTuo	Single (ZS)	PS	0.692 (0.672 - 0.714)	0.238 (0.213 - 0.268)	0.166
		PS + CXR	0.695 (0.675 - 0.716)	0.236 (0.212 - 0.266)	0.205
		PS + CXR + RR	0.742 (0.723 - 0.763)	0.277 (0.248 - 0.310)	0.269
		All (PS+EHR+CXR+RR)	**0.762 (0.743 - 0.782)**	0.325 (0.296 - 0.362)	0.049
	Multi (MV)	PS + CXR	0.649 (0.629 - 0.670)	0.201 (0.177 - 0.229)	0.068
		PS + CXR + RR	0.689 (0.667 - 0.709)	0.220 (0.195 - 0.252)	**0.014**
		All (PS+EHR+CXR+RR)	0.711 (0.690 - 0.730)	0.245 (0.217 - 0.279)	0.050
Llava	Single (ZS)	PS	0.642 (0.621 - 0.666)	0.204 (0.182 - 0.235)	0.755
		PS + CXR	0.658 (0.636 - 0.681)	0.213 (0.188 - 0.243)	0.766
		PS + CXR + RR	0.734 (0.712 - 0.756)	0.270 (0.241 - 0.304)	0.779
		All (PS+EHR+CXR+RR)	0.741 (0.721 - 0.761)	0.268 (0.242 - 0.303)	0.835
	Multi (MV)	PS + CXR	0.621 (0.597 - 0.647)	0.184 (0.164 - 0.211)	0.804
		PS + CXR + RR	0.694 (0.672 - 0.717)	0.256 (0.227 - 0.291)	0.800
		All (PS+EHR+CXR+RR)	0.711 (0.690 - 0.730)	0.245 (0.217 - 0.279)	0.812

E. Debate Consensus Analysis

To further understand the failure modes of multi-agent systems, we analyze the generated traces and conduct more ablations. Table E.1 shows the percentage of samples in the debate baseline traces that reached consensus at each round, or never reached consensus and had their probabilities averaged (MAX rounds). The results show a big variety depending on the backbone used. For instance, the Qwen backbone reached consensus for all the samples from the first round (indicating no inter-agent debate), while LLavaMed never reached a consensus for half of the patients. Notably, the models show a trend where having more samples with more rounds of debate correlates with degraded performance in AUROC. This suggests that current debate capabilities are lacking and do not consistently improve agentic reasoning.

Table E.1. Consensus analysis showing the percentage of samples reaching consensus at each round and AUROC.

Backbone	Round 1	Round 2	Round 3	MAX	AUROC
LlaVaMed	47.1%	0.2%	0.2%	52.6%	0.495
Huatuo	96.8%	2.1%	0.3%	0.8%	0.628
Qwen	100%	0%	0%	0%	0.631