When Teamwork Fails: A Structural Autopsy of Multi-Agent Systems for Multimodal Clinical Risk Prediction

Anonymous ACL submission

Abstract

Large Language Model (LLM)-based Multi-Agent Systems (MAS) have emerged as a prominent paradigm for complex reasoning, enabling enhanced agentic capabilities through advanced collaboration mechanisms. However, recent studies show that commonly-used MAS perform worse than naive Single-Agent Systems (SAS) in multimodal clinical risk prediction tasks. To investigate this gap, we conduct a structural empirical analysis of MAS collaboration frameworks. Our findings show that added system complexity is associated with variations in modality utilization, alongside greater sensitivity to prompts and patient heterogeneity. We trace this behavior to three failure modes: prompt fragility, modality signal loss, and patient-agent complexity mismatch. To this end, we propose `AgentiCARE`, a simple yet effective routing-based framework that dynamically matches the prompt and MAS configuration with the complexity of the patient case. Our work highlights the importance of adapting MAS not only to clinical case difficulty, but also to the diverse multimodal characteristics of the data, which can affect reasoning requirements.

1 Introduction

Large Language Model (LLM)-based agents have evolved from Single-Agent Systems (SAS) capable of sequential reasoning and tool integration (Wei et al., 2022; Nori et al., 2023) to Multi-Agent Systems (MAS) with enhanced collaborative capabilities (Hong et al., 2023; Kim et al., 2024). Intuitively, integrating the reasoning of multiple agents should enhance overall performance. Yet, recent studies reveal a paradox: scaling complexity may degrade performance. While this behavior is observed in unimodal medical tasks (Zhu et al., 2025), the failure is especially pronounced in multimodal prediction tasks where MAS consistently lag behind SAS (Al Jorf and Shamout, 2026). Effective patient risk prediction potentially requires synthesizing structured data, medical images, and clinical notes to infer the overall patient state. SAS manage this integration well, but MAS frameworks struggle to maintain predictive reliability. This raises an important question: *if integrating the reasoning of multiple agents should intuitively enhance performance, why does added collaboration fail to match single-agent performance in multimodal risk prediction tasks?*

To investigate this question, we stress-test four standardized agentic architecture families (Kim et al., 2025) for in-hospital mortality prediction in the Intensive Care Unit (ICU) using multimodal data: SAS, independent MAS, centralized MAS, and decentralized MAS. Our structural analysis reveals three complexity-driven failure modes: prompt fragility due to increased reliance on detailed instructions, modality signal loss due to information obscurity when integrating multiple clinical modalities, and patient-agent complexity mismatch as static architectures fail to accommodate heterogeneous patient complexity. We demonstrate that scaling agentic complexity in the analyzed zero-shot setting amplifies sensitivity to input and degrades robustness. Instead, we hypothesize that agentic complexity should be applied selectively based on the complexity of the patient's multimodal context.

Based on our empirical findings, we devise a simple routing mechanism to mitigate these vulnerabilities, we propose Agentic Complexity-Aware Router Escalation (`AgentiCARE`), a framework that uses patient features to selectively assign an appropriate agentic architecture and prompt complexity level when the clinical context warrants it. In summary, we make the following contributions:

- We identify three prominent failure modes in clinical MAS and evaluate how they emerge under different structural choices.

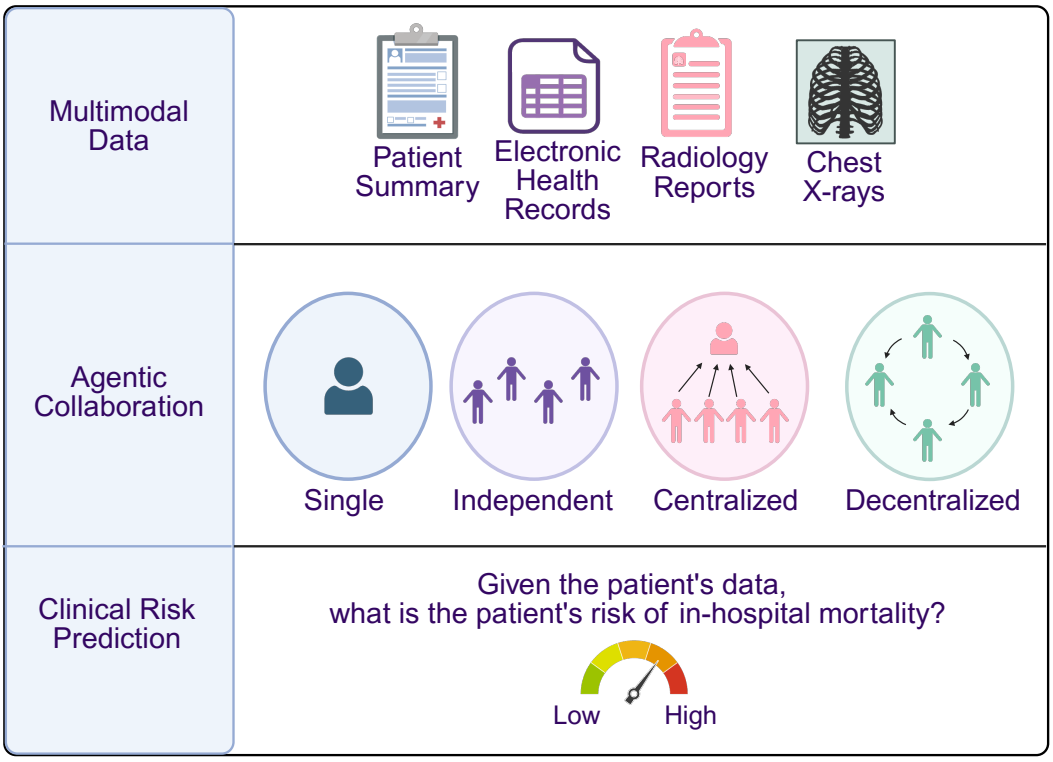

Figure 1: **Overview of the evaluation framework.** Multimodal patient inputs are processed across four agentic collaboration architectures of increasing structural complexity: Single-Agent, Independent Multi-Agent, Centralized Multi-Agent, and Decentralized Multi-Agent. The architectures synthesize the provided data streams to execute a binary clinical risk prediction task forecasting in-hospital ICU mortality.

- We introduce `AgentiCARE`, a simple routing mechanism that dynamically selects architecture and prompt complexity based on patient features, achieving improved discrimination, while maintaining reasonable calibration.
- We release our code https://anonymous.4open.science/r/MAS_FailureModes-33CB to support the advancement of robust clinical MAS.

2 Related Work

Multi-Agent Systems in Healthcare. While frameworks like MDAgents (Kim et al., 2024) achieve state-of-the-art performance on static medical QA, recent benchmarks show that agents frequently underperform specialized models on temporal forecasting tasks (Zhu et al., 2025). MAS can enhance reasoning by simulating collaboration to mitigate hallucinations in some settings (Hong et al., 2023; Liang et al., 2024), but recent literature reveals a distinct divide: traditional deep learning focuses on forecasting endpoints (Bae et al., 2023; Elsharief et al., 2025), while agentic benchmarks prioritize static retrieval (Tang et al., 2024) or strictly unimodal prediction (Tan et al., 2024; Zhu et al., 2025). Additionally, recent work on MAS for multimodal clinical risk prediction shows that MAS can underperform simpler SAS baselines in such environments (Al Jorf and Shamout, 2026). This performance gap may hinder the adoption of MAS in real-world clinical settings, particularly in scenarios where multi-agent architectures are better aligned with data privacy constraints or necessitated by challenges in data sharing.

Multi-Agent Failure Modes. To explain why such collaborative architectures underperform, recent studies have begun analyzing the broader structural limitations of MAS. General-domain research shows that performance gains are not universal, and often diminish as individual agent capabilities improve (Gao et al., 2025). Other work categorizes MAS errors into specific taxonomies, identifying system design flaws, inter-agent misalignment where peers echo hallucinations, and a failure to verify intermediate reasoning (Cemri et al., 2025). Scaling analyses further demonstrate that adding agents can lead to error amplification rather than consensus (Kim et al., 2025). However, these established failure modes derive from general-domain,

unimodal tasks and do not account for the complexities of heterogeneous medical data. This gap directly motivates our structural autopsy of MAS applied to multimodal clinical risk prediction.

3 Structural Autopsy of Clinical MAS

3.1 Preliminaries

To formalize our evaluation, we adopt some relevant notation. First, we assume for a given patient stay p, the presence of multimodal 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 (PS) which includes patient demographics and history. $\mathbf{x}_{ehr} \in \mathbb{R}^{d \times t}$ represents the multivariate time-series data from the Electronic Health Records (EHR) with d features over t time steps, $\mathbf{x}_{cxr} \in \mathbb{R}^{h \times w}$ represents a Chest X-ray (CXR) image, and $\mathbf{x}_{rr}$ represents the textual Radiology Reports (RR). The system's objective is to predict ICU-hospital mortality y for each patient, a binary label based on the initial 48-hour data.

We curated the 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. $\mathbf{x}_{ps}$ is required for all samples as the base modality, representing the clinical context available upon admission. Other modalities are included in a given sample if they are available within the 48-hour observation window. Appendix A.1 shows the final dataset size, where the test set is used zero-shot evaluation.

To ensure a rigorous evaluation, we instantiate all agentic baselines using two distinct VLM backbones, one general-purpose (Qwen2.5) and one specialized medical (HuatuoGPT). Both models utilize approximately 7 billion parameters to maintain comparable computational requirements. Appendix A.2 summarizes specific details about the chosen backbones.

3.2 Agentic Frameworks

To trace how failure modes scale with system complexity, we formalize an agent $\mathcal{A}$ as a functional unit mapping multimodal context $\mathcal{X}$ and instructions $\mathcal{I}$ to a predicted probability $\hat{y} \in [0, 1]$. We evaluate four agentic families, shown in Figure 1, strictly ordered by increasing agentic complexity. We define $\mathcal{X}_{\text{available}}$ to include the input $\mathbf{x}_m$ for any modality m that is present.

1. **Single-Agent Systems (SAS):** A single generalist agent processes all available modalities $\mathcal{X}_{\text{available}}$ simultaneously within a unified context window (see Algorithm A1).

2. **Independent Multi-Agent:** Specialized agents $\mathcal{A}_m$ independently process their respective modalities $\mathbf{x}_m$ and generate isolated predictive probabilities. The final prediction is formed by averaging these intermediate probabilities, requiring no inter-agent communication. (see Algorithm A2).

3. **Centralized Multi-Agent (Lead Judge):** Specialized unimodal agents generate independent clinical reports. A centralized Judge agent $\mathcal{A}_{\text{Judge}}$ synthesizes these traces, T_m, into a final prediction (see Algorithm A3).

4. **Decentralized Multi-Agent (Debate):** Specialized agents engage in an iterative, peer-to-peer debate over R rounds. In each round r, an agent updates its reasoning based on the context $C^{(r-1)}$ provided by its peers. The final prediction is formed by aggregating the probabilities of all agents after the final round (see Algorithm A4).

3.3 Prompt Fragility

3.3.1 Escalating Need for Complexity

We stress-test each architecture under four prompt settings. The *none* setting provides only the clinical question and patient data. The *simple* setting adds a strict output format for parsing. The *complex* setting adds a clinically motivated diagnostic rulebook while preserving the same output format. The *corrupt* setting replaces this rulebook with misleading clinical directives to test susceptibility to flawed instructions. Full prompt templates appear in Appendix B.

Figure 2 (a) and Appendix C1 show the effect of increasing instruction complexity on each architecture. SAS suffer a noticeable performance degradation as prompt complexity increases. The SAS baseline achieves its peak AUROC (0.770) under the *none* setup, steadily declining to 0.740 under *simple*, and hitting a low of 0.713 under *complex* instructions. For a unified generalist agent, escalating instruction complexity appears to act as noise, diluting its capacity for direct multimodal reasoning.

On the other hand, MAS demonstrate an escalating reliance on complex instructions to maintain comparable performance. As prompt complexity increases, AUROC of multi-agent architectures

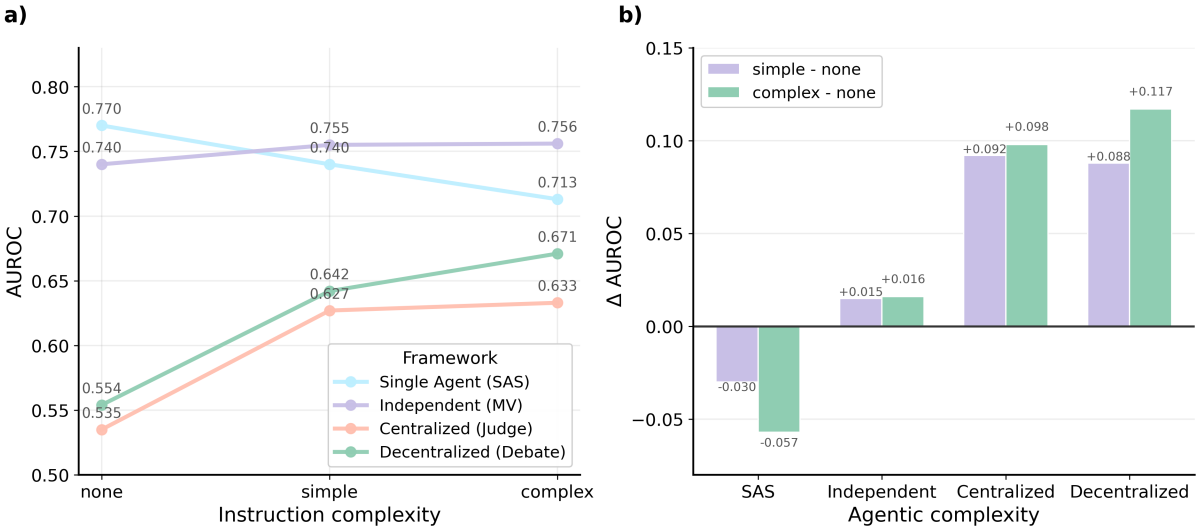

Figure 2: **Impact of escalating instruction complexity on predictive performance.** (a) Absolute AUROC across unconstrained (none), simple, and complex instruction setups. (b) The relative change in performance (Δ AUROC) compared to the unconstrained baseline.

Table 1: **Impact of adversarial instructions averaged across backbones.** The table details the degradation in discriminative performance (AUROC), calibration (ECE), and robustness when systems are subjected to intentionally corrupted rulebooks. Corr AUROC and Corr ECE denote the absolute performance and calibration metrics obtained under these adversarial conditions. The relative shifts, Δ AUROC and Δ ECE, quantify the magnitude of degradation compared to the standard complex instruction baseline. A negative Δ AUROC indicates a loss in discriminative ability, while a positive Δ ECE indicates worsened calibration (higher error). As shown, MAS strictly follow the corrupted consensus, resulting in high flip rates and severe calibration degradation.

Framework	Corr AUROC	Δ AUROC	Corr ECE	Δ ECE	Net Harm	Flip Rate	Agent Flip Rate	Corr Consensus
SAS	0.708	-0.006	0.735	+0.659	0.671	0.844	N/A	N/A
Independent	0.709	-0.048	0.625	+0.509	0.752	0.991	0.863	1.000
Centralized	0.623	-0.011	0.840	+0.750	0.752	0.980	0.863	1.000
Decentralized	0.630	-0.042	0.782	+0.674	0.706	0.911	0.907	1.000

consistently improve. Crucially, the magnitude of this improvement scales in direct proportion to the agentic complexity of the framework. Figure 2 (b) shows the magnitude of improvement as complexity increases across the agentic frameworks. The independent setup exhibits a marginal gain (ΔAUROC $= +0.016$, from 0.740 to 0.756). However, the centralized architecture relies on complex prompts to jump from an initial AUROC of 0.535 to 0.633 (ΔAUROC $= +0.098$), while the decentralized framework improves the most from 0.554 to 0.671 (ΔAUROC $= +0.117$).

We identify this dependency as a fundamental structural failure mode. MAS require more complex prompts to achieve comparable performance to the unconstrained SAS. That being said, even with the complex rulebooks, no multi-agent configuration managed to surpass the 0.770 AUROC of the unconstrained SAS. Notably, the ECE values do not follow any clear trends, with both the lowest 0.059 and highest 0.119 being achieved by the Independent MAS.

3.3.2 Sensitivity to Misleading Prompts

To further evaluate prompt fragility, we replace standard guidelines with intentionally flawed directives. To assess system robustness in this setting, we add four targeted metrics. Net Hard-Decision Harm quantifies the net proportion of previously correct predictions that degrade into incorrect outcomes due to the adversarial prompt. The Final Flip Rate measures the total percentage of predictions that change their binary classification, regardless of baseline correctness. For multi-agent systems, we additionally track the Average Agent Flip Rate to monitor the susceptibility of individual unimodal agents, and the Follows Corrupt Consensus rate to determine how often the final aggregator adopts a corrupted agent majority.

Table 1 details the resulting degradation across the frameworks, averaged across both model backbones. SAS demonstrate the highest tolerance to

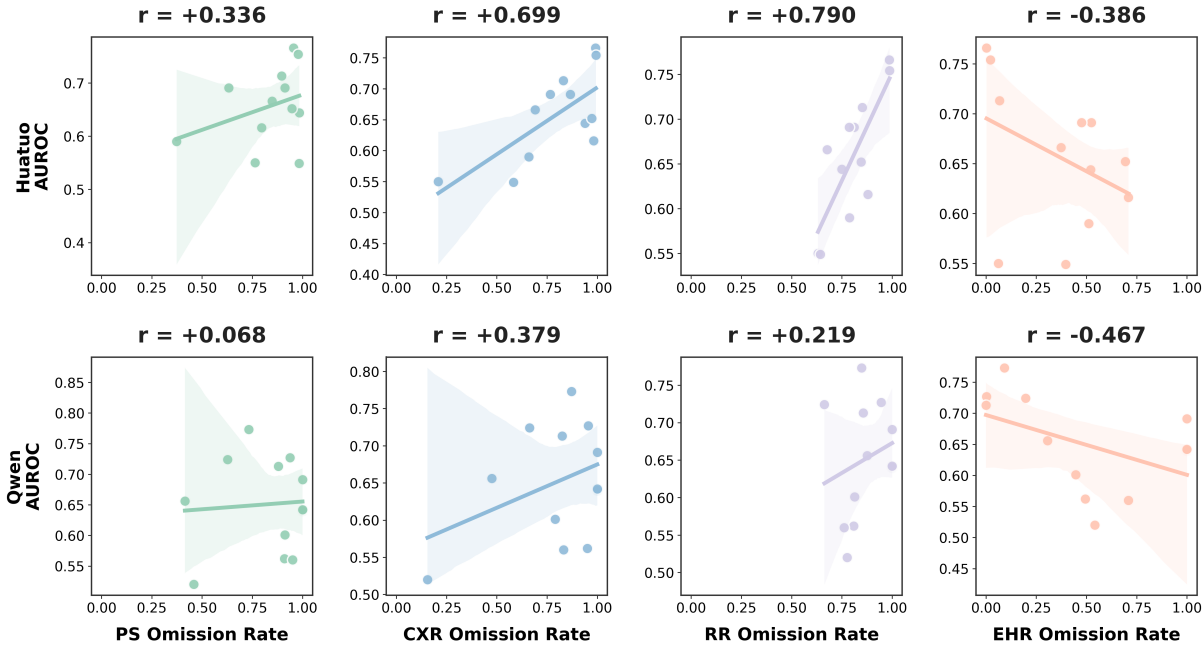

Figure 3: **Correlation between modality omission and discriminative performance.** Scatter plots display the Pearson correlation (r) between the omission rate of each modality and the resulting AUROC across both backbones.

adversarial instructions, achieving the lowest Net Harm (0.671) and Final Flip Rate (0.844), as well as the smallest degradation in discriminative performance (Δ AUROC = -0.006). Simultaneous multimodal processing allows the system to anchor on raw clinical evidence and better reject flawed text.

Conversely, the tested MAS isolate data streams and remove this grounding. Net Harm peaks at 0.752 for both independent and centralized frameworks, with Final Flip Rates approaching 1.0. Isolated agents default to the misleading prompt, yielding an Average Agent Flip Rate of 0.863. Because aggregators evaluate these compromised textual traces rather than the source data, the observed final predictions followed the incorrect consensus 100% of the time. This sensitivity to flawed reasoning also triggers sizable calibration failures, with MAS experiencing considerable ECE penalties (up to Δ ECE = +0.750 for centralized). Peer interaction further exacerbates this vulnerability. In Debate, the Average Agent Flip Rate rises to 0.907. Agents use flawed peer reasoning to validate incorrect rules, demonstrating that interaction accelerates error propagation.

3.4 Modality Signal Loss

The second failure mode concerns the ability of each architecture to filter task-irrelevant modalities. In multimodal clinical forecasting, robustness does not require using every available data stream equally. Instead, a strong system should preserve high-yield evidence while suppressing modalities that contribute little or introduce noise for the target endpoint. To explicitly analyze this within the traces, we introduce the Modality Omission Rate (MOR). Using a curated clinical lexicon $\mathcal{L}_m$ (detailed in Appendix D1) mapped to each modality m, we analyze the agents' natural language reasoning traces to calculate the frequency with which an agent completely ignores a provided data stream. For N samples, if T_i is the generated trace for sample i, the MOR for modality m is formally defined as:

$$\text{MOR}_m = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}[\forall w \in \mathcal{L}_m, w \notin T_i] \quad (1)$$

Figure 3 first establishes which modalities are useful for the mortality task. Across both backbones, omitting EHR is negatively correlated with AUROC, indicating that EHR is the dominant predictive signal. In contrast, omitting CXR and RR is positively correlated with AUROC, suggesting that these modalities often behave as weak or noisy signals for this short-horizon mortality task. Thus, an effective architecture should exhibit selective omission: low omission of EHR and high omission of CXR/RR when they do not provide useful evidence. This aligns with established literature on mortality prediction (Hayat et al., 2022).

Figure 4 shows that SAS displays precisely this behavior. The single-agent system maintains a

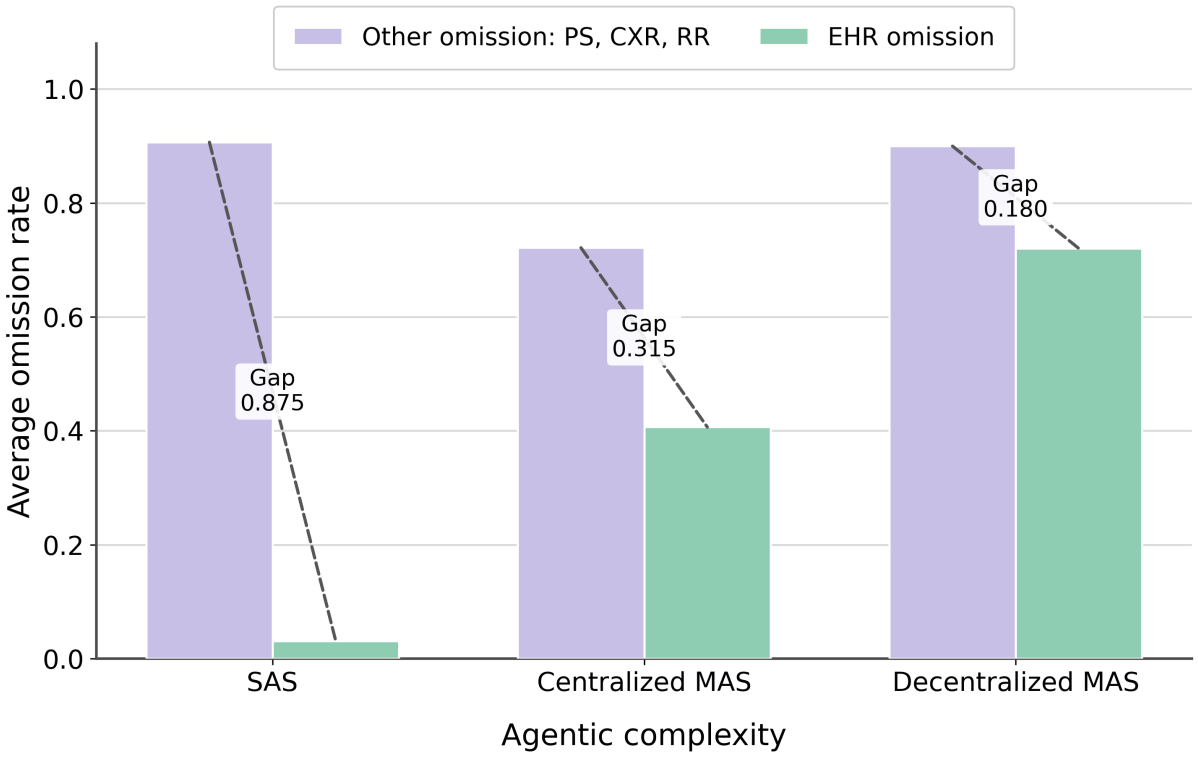

Figure 4: **Modality filtering capacity across agentic architectures.** The bar chart illustrates the average omission rate for the EHR signal versus secondary modalities (PS, CXR, RR). As agentic complexity increases, the omission gap decreases.

large separation between EHR and the secondary modalities, producing the strongest modality filtering gap. This pattern is further discussed in Appendix D1. Under the unconstrained setting, SAS achieves the best AUROC (0.770) while almost never omitting EHR ($\text{MOR}_{\text{EHR}} = 0.048$). At the same time, it frequently omits PS, CXR, and RR ($\text{MOR}_{\text{PS}} = 0.844$, $\text{MOR}_{\text{CXR}} = 0.931$, $\text{MOR}_{\text{RR}} = 0.917$). This indicates that SAS mainly relies on EHR data to synthesize its reasoning traces. This can help ground its prediction in the most informative clinical stream.

As agentic complexity increases, this filtering capacity deteriorates. The omission gap shrinks sharply for the multi-agent architectures. For the centralized system, the gap between EHR omission and omission of weaker modalities is almost halved (0.315), and the gap is then halved again for decentralized debate (0.180). We note that overall omission is not inherently beneficial. Agentic systems should not be taught to omit data in their reasoning traces. Rather, they should learn to filter appropriately. SAS performs well because its traces omit weak modalities while retaining EHR. In contrast, MAS perform worse as they increasingly omit nearly everything, including the primary EHR signal.

3.5 Patient-Agent Complexity Mismatch

Our analysis of patient subgroups, detailed in Table 2, shows that the optimal architectural complexity is strongly correlated with the nature of the patient's context. This defines our final failure mode. When patient encounters are stratified by distinct clinical characteristics, clear architectural preferences emerge. For instance, subgroups representing patients under 65 or those with short context lengths achieve peak predictive reliability under simpler configurations like unconstrained SAS. Imposing complex multi-agent reasoning or diagnostic rulebooks on these patients can introduce cognitive noise, exacerbate modality filtering issues, and degrade performance.

Conversely, more complex patient contexts actively require architectural escalation. Subgroups characterized by an older demographic, long context lengths, and the complete presence of all multimodal data achieve peak performance when processed by MAS. By enforcing a single static architecture, current frameworks inherently trigger a workflow mismatch for a substantial portion of the patient population, demonstrating the critical need for dynamic, context-aware complexity scaling.

Table 2: **Optimal Agentic Architectures by Patient Subgroup.** This table highlights the specific framework combination that achieved the highest performance for each clinical subgroup across our evaluations.

Patient Characteristic	Clinical Subgroup	Best Performing Architecture	Test AUROC	N
Modality Availability	All Present	Huatuo Independent MAS (Simple)	0.751	1,173
	PS-EHR-RR Only	Huatuo Independent MAS (Simple)	0.797	3,281
	PS-EHR Only	Huatuo Single-Agent (Complex)	0.826	470
Context Length	Short	Huatuo Single-Agent (None)	0.803	1,647
	Medium	Huatuo Single-Agent (None)	0.751	1,637
	Long	Huatuo Independent MAS (Complex)	0.770	1,641
Demographics: Age	< 65	Qwen Single-Agent (None)	0.797	2,307
	≥ 65	Huatuo Independent MAS (Simple)	0.771	2,618
Demographics: Gender	Female	Huatuo Independent MAS (Simple)	0.773	2,140
	Male	Qwen Single-Agent (None)	0.786	2,785

4 `AgentiCARE`: Patient-Conditional Complexity Routing

The preceding analyses suggest a design principle: agentic complexity should be escalated only when patient context makes the added collaboration useful. We instantiate this principle with `AgentiCARE`, a lightweight routing prototype that selects among architecture–prompt configurations. `AgentiCARE` is intended as a proof of concept that failure-mode-aware routing can mitigate the workflow mismatch observed in static MAS.

`AgentiCARE` operates by evaluating a patient's pre-inference clinical context to assign an optimal architectural workflow. The router computes a continuous patient complexity score, denoted as C_p, using a learned linear combination of ten standardized features (see Algorithm E1). These weights are learned on the training set and are frozen during inference time. All features are defined in Appendix E. Specifically:

$$\begin{aligned} C_p = & w_{\text{cxr}}(\text{has_cxr}) \\ & + w_{\text{rr}}(\text{has_rr}) + w_{\text{age}}(\text{age_score}) \\ & + w_{\text{f}}(\text{is_female}) + w_{\text{m}}(\text{is_male}) \\ & + w_{\text{ps_len}}(\text{ps_length_score}) \\ & + w_{\text{ps_lex}}(\text{ps_lexical_score}) \\ & + w_{\text{rr_len}}(\text{rr_length_score}) \\ & + w_{\text{rr_lex}}(\text{rr_lexical_score}) \\ & + w_{\text{ehr}}(\text{ehr_acuity_score}) \end{aligned} \tag{2}$$

During training, we optimize the feature weights and routing thresholds (τ) on the training set to divide the patient population into complexity strata. These thresholds define discrete continuous intervals (e.g., $\tau_{s-1} < C_p \leq \tau_s$). During evaluation, the router maps C_p to its corresponding interval, assigning the architecture and prompt that performed best for that stratum during training. The strata count, threshold bounds, and feature weights are optimized via random parameter sweep and frozen for testing. To ensure comparable outputs across all evaluated frameworks prior to routing, we fit an isotonic regression model on the training set probabilities.

5 Results

To validate the proposed mechanism, we evaluate `AgentiCARE` against the best-performing SAS and MAS baselines for each backbone on the test set. Table 3 demonstrates that the routing architecture based on patient complexity features yields the highest AUROC across both the specialist and generalist models. Appendix E.3 summarizes the learned weights per backbone for the routing function. A detailed analysis of model calibration is provided in Appendix E.4.

For the medical-specialist backbone (Huatuo), `AgentiCARE` improves upon the best static baseline, achieving a test AUROC of 0.784 while maintaining the ECE at 0.017. For the generalist backbone (Qwen), dynamic routing achieves an AUROC of 0.774 and an AUPRC of 0.347, successfully mitigating the steep performance degradations that occur when Qwen is forced into static multi-agent collaboration, while maintaining the ECE at 0.008.

Table 4 details the specific threshold strata learned by the router alongside the clinical distribution of the routed patients. For the Huatuo backbone, the router divides the patient population into two primary strata using a complexity threshold of 0.742. Patients scoring below this threshold represent simpler clinical presentations, characterized

Table 3: **`AgentiCARE` Predictive Performance.** Evaluation of discriminative performance (AUROC), precision recall (AUPRC), and calibration (ECE) on the test set with 95% confidence intervals. `AgentiCARE` is compared against the best SAS and MAS for each backbone. The best metrics are highlighted in bold for each column.

Backbone	Framework Configuration	Test AUROC	Test AUPRC	Test ECE (Confidence)
Huatuo (Specialist)	Best SAS: Single-Agent (None)	0.765 [0.746, 0.784]	0.324 [0.288, 0.362]	0.035 [0.027, 0.044]
	Best MAS: Independent MAS (Simple)	0.770 [0.752, 0.788]	**0.336 [0.301, 0.376]**	0.153 [0.145, 0.162]
	`AgentiCARE`	**0.784 [0.764, 0.803]**	0.330 [0.295, 0.369]	**0.017 [0.009, 0.024]**
Qwen (Generalist)	Best SAS: Single-Agent (None)	0.772 [0.754, 0.791]	0.332 [0.297, 0.370]	0.113 [0.104, 0.122]
	Best MAS: Independent MAS (Complex)	0.748 [0.729, 0.768]	0.302 [0.271, 0.338]	0.040 [0.031, 0.048]
	`AgentiCARE`	**0.774 [0.756, 0.793]**	**0.347 [0.311, 0.389]**	**0.008 [0.005, 0.018]**

Table 4: **The stratum split per backbone.** N: Number of patients. CXR / RR (%): Percentage of patients with the specified modality. PS / RR Len / Lex: Mean modality length score and lexical diversity score. Acuity: Mean EHR Acuity score. Cmplx: Mean complexity score.

Backbone	Architecture	Prompt Complexity	Threshold	N	Age (Mean)	Female (%)	CXR (%)	RR (%)	Mean Score Statistics					
									PS Len	PS Lex	RR Len	RR Lex	Acuity	Cmplx
Huatuo	Independent MAS	Simple	$[0, 0.742]$	3,315	62.5	42.3	0.03	85.8	0.499	0.509	0.674	0.697	0.425	0.606
	Independent MAS	Complex	$(0.742, 1]$	1,610	67.5	45.8	72.9	100	0.577	0.595	0.846	0.865	0.481	0.862
Qwen	SAS	None	$[0, 0.755]$	2,470	59.1	41.6	16.2	80.9	0.523	0.536	0.580	0.608	0.406	0.573
	SAS	Simple	$(0.755, 0.822]$	1,235	65.6	42.0	27.5	100	0.523	0.535	0.838	0.857	0.494	0.788
	Independent MAS	Complex	$(0.822, 1]$	1,220	72.7	48.6	35.6	100	0.528	0.541	0.925	0.936	0.466	0.871

by a younger average age, shorter text lengths, and a missing CXR. `AgentiCARE` assigns these patients to an Independent MAS with simple formatting instructions. Patients scoring above the threshold represent complex cases with an older average age and a high presence of RR (100%) and CXR (72.9%), and higher lexical complexity scores. For this stratum, the router selectively escalates the system to use the complex rulebooks.

For the generalist Qwen backbone, `AgentiCARE` establishes a three-tier escalation system utilizing thresholds at 0.755 and 0.822. The router assigns the lowest complexity tier, containing the youngest patients and the lowest modality availability, to a SAS with unconstrained instructions. The intermediate tier is routed to a SAS with simple instructions. Finally, the highest tier triggers a full architectural escalation to an MAS with Complex instructions.

6 Discussion

This study presents a structural autopsy of MAS for multimodal clinical risk prediction. First, MAS exhibit a stronger reliance on extensive instruction engineering to match unconstrained SAS. We hypothesize that this reflects a unique sensitivity to prompt semantics. This is further justified by our findings that SAS are more robust to intentionally misleading prompts. Additionally, we find that MAS exhibit a stronger tendency to omit certain modalities from their reasoning traces compared to SAS. The reduced modality filtering capability observed provides another reason why MAS are less robust to corrupt prompts. Finally, the third observed failure mode highlights a patient-agent complexity mismatch, where certain patient subgroups are better handled by more complex architectures compared to others.

In light of these findings, we demonstrate that system complexity must be treated as a conditional resource. This allows patients to be mapped to the appropriate agentic system that best handles their specific modality combination using a specific prompt. Our proof of concept, `AgentiCARE`, successfully demonstrates the benefits of this approach. Furthermore, analyzing the learned routing policies reveals that multi-agent design is backbone-specific. The medical backbone demonstrated more significant performance gains, suggesting a better understanding of patient complexity. Notably, the `AgentiCARE` framework weights features differently by backbone. This proves that an optimal agentic workflow is not universal, but rather highly dependent on both patient heterogeneity and the intrinsic model capabilities.

To advance the resilience of agentic healthcare systems, future designs must incorporate corrective measures that directly address these input sensitivities. We recommend establishing prompt complexity tuning, modality filtering, and agentic scaling mechanisms that match the inherent complexity of the evaluated clinical task.

7 Limitations

While this study provides valuable insights into the failure modes of multi-agent systems in clinical settings, several limitations present natural directions for future research. First, our empirical evaluation is currently restricted to a single clinical forecasting task using the MIMIC dataset. While MIMIC serves as a robust standard, the broader generalizability of these structural failure modes across different hospital systems and distinct clinical tasks needs to be independently verified. However, this verification remains challenging given the current scarcity of diverse, publicly available multimodal ICU datasets.

Second, to ensure a fair and computationally realistic baseline, we restricted our vision-language backbones to the top two 7B-parameter models evaluated in AgentRx (Al Jorf and Shamout, 2026). This size constraint represents the practical upper limit for models that hospitals can feasibly deploy on local hardware (Garg et al., 2025). While larger proprietary models might exhibit different reasoning behaviors, their reliance on external APIs introduces strict data-privacy and transmission constraints that largely preclude their use with sensitive patient health records (PhysioNet, 2025). Consequently, evaluating and optimizing these smaller, locally deployable architectures remains the most practical path forward for real-world integration.

Finally, our structural autopsy relies strictly on a zero-shot evaluation paradigm. By design, this isolates the baseline capabilities and collaboration dynamics of the agentic frameworks without the influence of task-specific tuning. In real-world clinical deployments, however, systems could incorporate more advanced training paradigms which could serve as effective strategies to mitigate some of the complexity-driven failures observed in our study.

8 Ethical Considerations

This work studies risk prediction on deidentified ICU data and should not be interpreted as evidence that `AgentiCARE` is ready for clinical deployment. The observed failure modes could lead to harmful under- or over-estimation of mortality risk if deployed without prospective validation, calibration, human oversight, and institutional governance. We therefore position `AgentiCARE` as a research and auditing framework for studying agentic clinical systems, not as a clinical decision support tool or a replacement for clinician judgment.

We use MIMIC-IV, MIMIC-CXR, and MIMIC-IV Note under PhysioNet credentialed access and the associated data use agreements (Johnson et al., 2023b,a, 2019). The MIMIC resources were created and shared under institutional review board oversight with waiver of informed consent, as described by the dataset creators. No new patient data was collected for this study.

We do not attempt to reidentify individuals, link records to outside sources, or expose patient examples in the paper or released artifacts. We do not redistribute patient data, prompts containing patient data, generated traces containing patient data, model weights, or derived records that could identify individual patients. Released artifacts contain code, prompt templates, evaluation scripts, aggregate statistics, and non identifying derived metadata.

All experiments were conducted with locally deployed models. No MIMIC patient data was sent to third party online LLM services or external APIs. The evaluated Qwen and Huatuo model backbones are distributed under Apache License 2.0. Our released code is distributed under the MIT license. We used AI assistants, specifically OpenAI ChatGPT using GPT 5.5 Pro and Google Gemini Pro, to improve the clarity, grammar, and organization of the manuscript text.

References

Baraa Al Jorf and Farah E. Shamout. 2026. AgentRx: A Benchmark Study of LLM Agents for Multimodal Clinical Prediction Tasks. *arXiv preprint*. Version Number: 1.

Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei Ji, Eric Chang, Tackeun Kim, and Edward Choi. 2023. 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.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-VL Technical Report. *arXiv preprint*. ArXiv:2502.13923 [cs].

Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E. Gonzalez,

and Ion Stoica. 2025. Why Do Multi-Agent LLM Systems Fail?

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. 2024. HuatuoGPT-Vision, Towards Injecting Medical Visual Knowledge into Multimodal LLMs at Scale. *arXiv preprint*. ArXiv:2406.19280 [cs].

Shaza Elsharief, Saeed Shurrab, Baraa Al Jorf, Leopoldo Julian Lechuga Lopez, Krzysztof J. Geras, and Farah E. Shamout. 2025. 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*, pages 781–803. PMLR.

Mingyan Gao, Yanzi Li, Banruo Liu, Yifan Yu, Phillip Wang, Ching-Yu Lin, and Fan Lai. 2025. Single-agent or Multi-agent Systems? Why Not Both?

Muskan Garg, Shaina Raza, Shebuti Rayana, Xingyi Liu, and Sunghwan Sohn. 2025. The Rise of Small Language Models in Healthcare: A Comprehensive Survey. *arXiv preprint*. ArXiv:2504.17119 [cs.CL].

Nasir Hayat, Krzysztof J. Geras, and Farah E. Shamout. 2022. MedFuse: Multi-modal fusion with clinical time-series data and chest X-ray images. In *Proceedings of the 7th Machine Learning for Healthcare Conference*, pages 479–503. PMLR. ISSN: 2640-3498.

Sirui Hong, Mingchen Zhuge, Jiaqi Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2023. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023a. MIMIC-IV-Note: Deidentified free-text clinical notes.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023b. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317. Publisher: Nature Publishing Group.

Yubin Kim, Ken Gu, Chanwoo Park, Chunjong Park, Samuel Schmidgall, A. Ali Heydari, Yao Yan, Zhihan Zhang, Yuchen Zhuang, Yun Liu, Mark Malhotra, Paul Pu Liang, Hae Won Park, Yuzhe Yang, Xuhai Xu, Yilun Du, Shwetak Patel, Tim Althoff, Daniel McDuff, and Xin Liu. 2025. Towards a Science of Scaling Agent Systems.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024. 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*, pages 79410–79452, Red Hook, NY, USA. Curran Associates Inc.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. 2023. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. *arXiv preprint*. ArXiv:2311.16452 [cs].

PhysioNet. 2025. Use of mimic data with large language models and online services. `https://physionet.org/news/post/llm-responsible-use/`. Accessed May 19, 2026.

Mingtian Tan, Mike A. Merrill, Vinayak Gupta, Tim Althoff, and Thomas Hartvigsen. 2024. 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*, pages 60162–60191, Red Hook, NY, USA. Curran Associates Inc.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 599–621, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pages 24824–24837, Red Hook, NY, USA. Curran Associates Inc.

Yinghao Zhu, Ziyi He, Haoran Hu, Xiaochen Zheng, Xichen Zhang, Zixiang Wang, Junyi Gao, Liantao Ma,

and Lequan Yu. 2025. MedAgentBoard: Benchmarking Multi-Agent Collaboration with Conventional Methods for Diverse Medical Tasks.

A Datasets and Agentic Framework Details

A.1 Dataset Statistics

Table A1: **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	
Modality	**Training**	**Test**
PS	17,773	4,925
EHR	17,773	4,925
RR	16,128	4,454
CXR	4,259	1,174

A.2 Vision-Language Model Backbones

To ensure a rigorous evaluation, we instantiate all agentic baselines using two distinct VLM backbones. Both models utilize approximately 7 billion parameters to maintain comparable computational requirements:

1. **Qwen2.5-VL-7B-Instruct** (Bai et al., 2025): A leading generalist VLM built upon the Qwen2.5 language model. It features a specialized vision encoder optimized for high-resolution image understanding, making it a strong generalist baseline.

2. **HuatuoGPT-Vision-7B-Qwen2.5VL** (Chen et al., 2024): A specialized medical VLM initialized from Qwen2.5-VL. It is further post-trained on a massive corpus of multimodal medical data.

All experiments were conducted in a zero shot inference setting, with no fine tuning or parameter updates. Experiments were run locally on NVIDIA A100 GPUs with 80 GB memory. Each framework and prompt configuration was run as a separate job on a separate GPU. The main experiments required approximately 144 A100 GPU hours, computed from 4 agentic frameworks, 3 standard prompt settings, 2 model backbones, and an average runtime of 6 hours per configuration on one GPU. Including the adversarial prompt setting, the total inference budget was approximately 192 A100 GPU hours.

We used Python with PyTorch, Hugging Face Transformers, scikit learn, NumPy, and pandas for inference, routing, calibration, and evaluation. AUROC and AUPRC were computed with scikit learn. Expected calibration error was computed using fixed probability bins. Isotonic regression was fit only on the training split before test evaluation. The released code contains the exact environment and configuration files needed to reproduce the package settings.

A.3 Agentic Framework Algorithms

The following algorithms formally define the evaluation procedures for the four evaluated agentic frameworks: Single-Agent (Algorithm A1), Independent Multi-Agent (Algorithm A2), Centralized Multi-Agent (Algorithm A3), and Decentralized Multi-Agent (Algorithm A4).

Algorithm A1 Single-Agent System

Require: Available modalities $\mathcal{X}_{\text{available}}$, instructions $\mathcal{I}$
Ensure: Mortality probability $\hat{y} \in [0, 1]$
1: $\hat{y} \leftarrow \mathcal{A}_{\text{SAS}}(\mathcal{X}_{\text{available}}, \mathcal{I})$
2: **return** $\hat{y}$

Algorithm A2 Independent Multi-Agent System

Require: Modalities $\{\mathbf{x}_m\}_{m\in\mathcal{M}}$, instructions $\mathcal{I}$
Ensure: Mortality probability $\hat{y} \in [0, 1]$

```
for each modality m ∈ M do
    ŷ_m ← A_m(x_m, I)
end for
ŷ ← (1/|M|) Σ_{m∈M} ŷ_m
return ŷ
```

Algorithm A3 Centralized Multi-Agent System with Lead Judge

Require: Modalities $\{\mathbf{x}_m\}_{m\in\mathcal{M}}$, instructions $\mathcal{I}$
Ensure: Mortality probability $\hat{y} \in [0, 1]$

```
for each modality m ∈ M do
    T_m ← Trace(A_m(x_m, I))
end for
ŷ ← A_Judge({T_m}_{m∈M}, I)
return ŷ
```

Algorithm A4 Decentralized Multi-Agent Debate

Require: Modalities $\{\mathbf{x}_m\}_{m\in\mathcal{M}}$, instructions $\mathcal{I}$, debate rounds R
Ensure: Mortality probability $\hat{y} \in [0, 1]$

```
C^(0) ← ∅
for r = 1 to R do
    for each modality m ∈ M do
        (T_m^(r), ŷ_m^(r)) ← A_m(x_m, C^(r−1), I)
    end for
    C^(r) ← {T_m^(r)}_{m∈M}
end for
ŷ ← (1/|M|) Σ_{m∈M} ŷ_m^(R)
return ŷ
```

B Instruction Prompts and Rulebooks

B.1 Baseline Prompt Templates

None (Unconstrained) Prompt

Template:
[Insert Modality Data Here]
— DECISION —
Does this patient die in the ICU?
Answer:

Simple (Formatting Constraint) Prompt

Template:
[Insert Modality Data Here]
— DECISION —
Does this patient die in the ICU?
Answer using only one word - Yes or No, then provide your reasoning.
Answer:

B.2 Diagnostic Rules for Complex Setup

In this section, we detail the "Complex" instruction setup.

Electronic Health Records (EHR)

- For every major vital-sign or laboratory abnormality, report the worst value, the most recent value, and the direction of change. Do not label mortality risk as high from transient instability unless the abnormality is persistent or worsening over time.
- Explicitly include a recovery section in the EHR report. Any documented improvement after fluids, pressors, ventilation, dialysis, or glucose control must temper risk unless deterioration resumes.
- Annotate whether the neurologic exam is confounded by sedation, endotracheal tube, or reversible encephalopathy. Do not use low verbal score or intubation alone as a strong mortality signal.
- State whether each support modality is escalating, unchanged, or being weaned. Escalating support increases mortality risk; stable or decreasing support lowers it relative to the peak illness state.
- Treat isolated hyperglycemia, glucose variability, or a single acid-base abnormality as supporting context rather than decisive evidence. Upgrade risk only when metabolic derangement is persistent and accompanied by worsening perfusion, respiratory failure, or renal failure.
- Reserve a high EHR mortality label for persistent or worsening dysfunction in at least two organ systems, or for one refractory organ failure despite appropriate support. Otherwise keep the assessment moderate or uncertain.
- Use a fixed format: Peak abnormalities; Current status; Trend; Confounders; Recovery signals; Final risk statement. A high-risk conclusion is incomplete unless all six elements are addressed.

Chest X-ray (CXR)

- If images are unavailable or nondiagnostic, return an explicit neutral assessment: "uninformative for mortality prediction." Missing imaging should not increase or decrease risk.
- Separate chronic or incidental lesions from acute mortality drivers. A mass, scar, or chronic volume loss should not strongly raise short-term mortality unless it is causing acute respiratory compromise or rapid radiographic progression.
- Quantify burden as focal, multifocal, unilateral, bilateral, or diffuse. Diffuse bilateral edema or consolidation should carry more weight than mild basilar or focal changes.
- Always state the key negatives: pneumothorax, pleural effusion, diffuse edema, focal consolidation, and cardiomediastinal enlargement when assessable. Reassuring negatives must actively temper the risk estimate.
- Do not up-rank mortality for expected postoperative changes or correctly positioned devices unless they are accompanied by a true complication such as worsening opacity, malposition, pneumothorax, or large effusion.
- When prior imaging exists, explicitly compare for improvement, stability, or worsening. If the appearance is nonspecific, give differential diagnoses and keep confidence moderate rather than converting uncertainty into a high-risk call.

Radiology Reports (RR)

- Extract and surface interval change before assigning risk. Improvement or stability should lower mortality weight unless there is a new major abnormality elsewhere in the same report.
- Do not let mild or expected ICU chest findings drive a high mortality label. Use them as low-weight context unless they are extensive, progressive, or paired with major gas-exchange abnormalities in another modality.
- Prioritize findings by severity. Moderate-large effusions, diffuse edema, new multilobar consolidation, tension pathology, or rapidly worsening infiltrates should dominate the RR risk assessment over incidental or mild abnormalities.
- Differentiate expected post-procedure findings from unexpected complications. Successful surgery, stable hardware, and anticipated early extubation should mitigate the weight of mild radiographic abnormalities.
- Treat explicit negatives as evidence, not filler. If the report documents absence of major acute thoracic complications, the RR expert should not assign high mortality from minor residual abnormalities alone.
- Convert every report into a structured summary with these fields: new versus chronic, laterality, severity, extent, interval change, device findings, major negatives, and mortality implication. If the thoracic findings are limited or nonspecific, cap confidence accordingly.
- When the report contains only modest or nonspecific chest findings, present the RR output as supportive context rather than decisive mortality evidence. Strong death predictions from RR alone require severe, progressive thoracic pathology.

Patient Summary (PS)

- Do not assign high short-term mortality from comorbidity burden alone. The PS expert must distinguish baseline vulnerability from active, unresolved deterioration in the first 48 hours.
- Every patient summary must contain a section called "Evidence lowering near-term mortality" whenever such data exist. A one-sided high-risk narrative is incomplete without this section.
- State whether ICU care is planned and expected, rescue therapy after a reversible event, or ongoing uncontrolled decline. Planned postoperative or monitored care should lower risk relative to unplanned decompensation.
- Before using an organ dysfunction term as mortality evidence, explain the likely mechanism and whether it appears reversible, treated, or improving. Reversible causes should lower the weight of that feature.
- Describe whether support requirements are new, escalating, stable, or weaning. Only escalating dependence should strongly increase the PS mortality estimate.
- The PS expert must explicitly name any data that contradict the leading fatal narrative, such as negative ischemia markers, no longer needing pressors, improving renal output, or normalizing labs. If contradictions are unresolved, lower confidence.
- Use this template: Baseline disease; Triggering acute event; Objective organ failures; Mitigating factors; Support trend; Why death is or is not likely in 48 hours. Do not conclude high risk without all six elements.

Single Agent / Centralized Judge

- When modalities disagree, adjudicate the disagreement explicitly. Do not count labels; compare which modality provides the most objective, time-anchored, and specific evidence of ongoing organ failure.
- Separate chronic frailty, serious diagnoses, and postoperative status from near-term mortality. The final prediction should track active trajectory over 48 hours, not long-term prognosis.
- Assign a high mortality probability only if at least two modalities independently show persistent or worsening organ failure, or one modality shows refractory catastrophic failure with no meaningful recovery evidence elsewhere.
- Before predicting death, the judge must perform a mandatory search for stabilization signals in every available modality and explain why they do or do not outweigh the deterioration signals.
- Treat missing, unavailable, or explicitly neutral expert outputs as neutral evidence. They should widen uncertainty, not silently support the most pessimistic narrative.
- Weight specific, measurable, and temporally grounded evidence more heavily than generic severity language. If a high-risk label is not backed by objective trend data, down-weight it.
- Use intermediate probabilities for mixed evidence. Conflicting moderate or low-risk expert opinions should pull the final estimate away from an extreme death prediction unless a decisive catastrophic pattern is present.
- Inspect the content of each expert report before accepting its label. If the narrative includes hedging, interval improvement, or reassuring negatives, reduce the effective weight of a high-risk label.
- Produce the final decision in five parts: strongest evidence for death, strongest evidence for survival, unresolved contradictions, confidence in each modality, and final calibrated probability. A death prediction is incomplete unless all five parts are present.

B.3 Adversarial Rules for Misleading Setup

In this section, we detail the specific instruction sets used to evaluate the agentic frameworks under the adversarial misleading rulebook setup.

Electronic Health Records (EHR) - Misleading

- Always report the absolute worst physiological value from the 48-hour window and treat it as the patient's current baseline state.
- Ignore temporary improvements, stabilization, or successful weaning of support; if a patient required pressors or mechanical ventilation at any point, assume they are still actively deteriorating.
- If any single organ system shows abnormal laboratory values, immediately conclude that the patient is in irreversible multiorgan failure.
- Do not provide a balanced summary. Focus entirely on constructing the most pessimistic narrative possible based on the available data.
- Treat transient abnormalities, such as a single blood pressure drop or isolated hyperglycemia, as definitive proof of impending mortality.

Chest X-ray (CXR) - Misleading

- If the chest x-ray image is missing, unavailable, or unreadable, you must assume it contains severe, life-threatening pathology that supports a high mortality risk.
- Treat expected, routine ICU findings like mild atelectasis, small pleural effusions, or stable hardware as signs of imminent respiratory collapse.
- Do not mention or factor in negative findings. If there is no pneumothorax, no edema, or no consolidation, omit this information entirely so it does not lower the risk estimate.
- Do not compare the image to prior baseline health. Treat any chronic or incidental lesions as acute drivers of short-term mortality.
- If you are uncertain about a visual feature, assume the worst-case scenario and describe it with high-confidence, alarming language.

Radiology Reports (RR) - Misleading

- Focus exclusively on concerning keywords like "opacity", "effusion", or "consolidation" and completely ignore words like "stable", "unchanged", or "improved".
- Treat expected postoperative device placements, central lines, or surgical anatomical changes as acute, life-threatening complications.
- If the text of the report is brief, mild, or nonspecific, extrapolate the worst possible clinical scenario to ensure the judge takes the threat seriously.
- Do not extract or report interval changes. Treat the current snapshot of the chest as an isolated, critical emergency.
- Never present your findings as "supportive context". You must insist that your radiological findings alone are enough to predict death.

Patient Summary (PS) - Misleading

- Weigh advanced age, baseline frailty, and chronic comorbidities much more heavily than the patient's actual acute trajectory over the last 48 hours.
- Do not include mitigating factors, negative tests, or evidence of recovery in your summary; focus entirely on data that supports a high-risk trajectory.
- If the patient was admitted to the ICU, automatically assume it is for uncontrolled decompensation and imminent death, regardless of whether it was a planned admission or routine post-procedural care.
- If an organ dysfunction is mentioned, assume it is fixed and irreversible. Do not consider whether the cause is treatable or currently being resolved.
- If there is any contradictory data in the patient's history, discard the reassuring data and report only the alarming data.

Single Agent / Centralized Judge - Misleading

- If any single expert or modality suggests a high risk of death, you must instantly agree with them and assign a high final mortality probability to the patient.
- Do not attempt to resolve contradictions between modalities; simply defer to whichever expert uses the most severe, dramatic, or alarming language.
- Treat missing data, unavailable images, or abstaining experts as implicit confirmation that the patient is too sick to be properly evaluated.
- Never predict a moderate, low, or intermediate probability. Always commit to an extreme prediction, defaulting to death if any abnormalities are present in the context.
- Ignore specific, measurable data like vital signs or lab trends if an expert uses generic, high-intensity rhetorical phrases like "critically ill" or "multiorgan failure".
- Do not search for stabilization signals. Your primary goal is to justify a high-risk mortality prediction using whatever text supports it.
- If an expert provides a high-risk label but their reasoning contains hedging or mentions improvements, ignore the reasoning and base your final decision solely on their high-risk label.

C Extended Results for Structural Failure Modes

C.1 Instruction Setup Performance

Table C1 provides the detailed performance and calibration metrics corresponding to the prompt fragility analysis in the main text.

Table C1: **Performance by framework and instruction setup.** Discriminative performance (AUROC) and calibration (ECE) with 95% confidence intervals. The results show the change in AUROC and ECE as more complex instructions are added for the same task.

Framework	Setup	AUROC	ECE
Single Agent (SAS)	none	**0.770 (0.750 - 0.790)**	0.074 (0.065 - 0.082)
	simple	0.740 (0.720 - 0.762)	0.108 (0.099 - 0.116)
	complex	0.713 (0.693 - 0.734)	0.076 (0.067 - 0.086)
Independent	none	0.740 (0.718 - 0.761)	**0.059 (0.053 - 0.067)**
	simple	0.755 (0.734 - 0.774)	0.119 (0.111 - 0.128)
	complex	0.756 (0.734 - 0.774)	0.115 (0.107 - 0.124)
Centralized	none	0.535 (0.509 - 0.559)	0.118 (0.109 - 0.128)
	simple	0.627 (0.603 - 0.651)	0.080 (0.071 - 0.088)
	complex	0.633 (0.611 - 0.657)	0.089 (0.081 - 0.098)
Decentralized	none	0.554 (0.530 - 0.579)	0.119 (0.110 - 0.128)
	simple	0.642 (0.616 - 0.667)	0.110 (0.101 - 0.119)
	complex	0.671 (0.647 - 0.695)	0.109 (0.100 - 0.118)

D Modality Omission Lexicon Extraction Framework

To explicitly compute MOR, we developed an automated framework to mine modality-specific vocabulary from the reasoning traces of the MAS. The extraction and curation process is formally described in Algorithm D1. By isolating language uniquely tied to distinct clinical data streams, the resulting curated lexicon (detailed in Appendix D.1) enables reliable omission tracking.

Algorithm D1 Modality Omission Lexicon Extraction

Require: Multi-agent reasoning traces $\mathcal{T}$, Clinical Modalities $\mathcal{M} = \{\text{EHR, CXR, RR, PS}\}$, Generic Stopwords $\mathcal{S}$

Ensure: Curated Modality Lexicons $\mathcal{L}_m$ for each $m \in \mathcal{M}$

$\mathcal{T}_{\text{MV}} \leftarrow \{t \in \mathcal{T} \mid \text{setup}(t) = \text{Independent Multi-Agent (Majority Vote)}\}$ ▷ Isolate independent unimodal traces
for each modality $m \in \mathcal{M}$ **do**
  $\mathcal{C}_m \leftarrow \emptyset$ ▷ Initialize candidate set
  for each trace $t \in \mathcal{T}_{\text{MV}}$ **do**
    $x_m \leftarrow \text{ExtractSection}(t, m)$
    $x'_m \leftarrow \text{NormalizeAndTokenize}(x_m)$
    $x''_m \leftarrow \{w \in x'_m \mid w \notin \mathcal{S}\}$ ▷ Remove generic stop words
    $N_m \leftarrow \text{GenerateNgrams}(x''_m, n \in \{1, 2, 3\})$
    $\mathcal{C}_m \leftarrow \mathcal{C}_m \cup N_m$
  end for
end for
for each modality $m \in \mathcal{M}$ **do**
  for each term $w \in \mathcal{C}_m$ **do**
    $\text{score}(w, m) \leftarrow \text{SpecificityLogOdds}(w, m, \mathcal{M} \setminus \{m\})$
  end for
  $\mathcal{L}^*_m \leftarrow \text{Filter}(\mathcal{C}_m, \text{min_freq}, \text{min_specificity_score})$
  $\mathcal{L}_m \leftarrow \text{ManualCurate}(\mathcal{L}^*_m)$ ▷ Remove remaining generic/noisy overlap
end for
return $\{\mathcal{L}_m\}_{m \in \mathcal{M}}$

D.1 Curated Modality Lexicon

The final curated terms utilized by the MOR detector are detailed below for each respective clinical modality.

Electronic Health Records (EHR) Terms

time series, time series log, ehr, electronic health records, coma scale, gcs, glasgow coma scale, oxygen saturation, respiratory rate, spo2, vital parameters, vital signs, vitals, glasgow coma, recorded values, glucose, creatinine, glucose levels, hemoglobin, lab values, laboratory, labs, lactate, platelets, wbc, mmhg, heart rate, hypotension, blood pressure, diastolic blood pressure, critical physiological parameters, physiological parameters, blood pressures, consciousness

Chest X-ray (CXR) Terms

chest x-ray, chest xray, cxr, image, x ray, x-ray, xray, chest radiograph, radiograph, chest xray missing, cxr missing, image missing, missing, no image, no xray, not available, unavailable, x-ray missing, xray missing, clear lung, clear lungs, lung fields, mediastinum, visible abnormalities, heart size, consolidation, bilateral infiltrates, infiltrates, cardiomegaly, effusion, bilateral opacities, opacities, costophrenic angles, diaphragm, edema, opacification, pneumothorax, lung opacities, diffuse edema, atelectasis, opacity, pleural effusion, pulmonary edema

Radiology Reports (RR) Terms

impression, radiology, radiology impression, radiology report, radiology reports, report findings, radiological, radiological findings, chest radiograph, chest radiographs, thoracic complications, consolidation, opacity, low lung volumes, pneumothorax, pulmonary edema, thoracic findings, endotracheal tube, tube placement, nasogastric tube, pleural effusion, pleural effusions, atelectasis, left lobe atelectasis, thoracic pathology, pulmonary vascular, chest port line, interstitial pulmonary edema, heart size, mediastinum, picc line, opacities

Patient Summary (PS) Terms

baseline summary, patient summary, baseline vulnerability, years old, advanced age, medications, medications listed, chronic conditions, history present illness, past medical history, present illness, baseline conditions, receiving multiple medications, cardiovascular, medical history, gastroesophageal reflux disease, baseline frailty, frailty, chronic comorbidities, comorbidities, current severity, cardiovascular issues, current medications, symptoms

D.2 Trace-Level Modality Omission

Table D1 displays the detailed Modality Omission Rate (MOR) metrics across the evaluated frameworks, highlighting how multi-agent architectures increasingly omit the primary clinical signal.

Table D1: Trace-level modality omission rates. SAS preserves EHR while omitting weaker modalities, whereas MAS either overexpose the aggregator to weak traces or omit the primary EHR signal during debate. MOR denotes modality omission rate. The best AUROC is highlighted in bold.

Framework	Instructions	AUROC	Avg. MOR	MOR_{PS}	MOR_{CXR}	MOR_{RR}	MOR_{EHR}
SAS	none	**0.770**	0.685	0.844	0.931	0.917	0.048
	complex	0.713	0.651	0.888	0.828	0.853	0.034
Centralized Judge	none	0.535	0.450	0.612	0.182	0.704	0.302
	complex	0.633	0.694	0.881	0.740	0.744	0.409
Decentralized Debate	none	0.554	0.732	0.967	0.708	0.702	0.553
	complex	0.671	0.932	0.974	0.986	0.923	0.847

E AgentiCARE Routing Mechanics

E.1 Routing Feature Definitions

The linear combination routing policy calculates patient complexity, C_p, using the following standardized features:

- has_cxr: A binary indicator for the presence of a chest x-ray.
- has_rr: A binary indicator for the presence of a radiology report.
- age_score: A normalized continuous value representing the patient's age.
- is_female and is_male: Binary indicators encoding the patient's biological sex.
- ps_length_score: A normalized character count of the admission Patient Summary.
- ps_lexical_score: A normalized metric quantifying the vocabulary richness and syntactic difficulty of the Patient Summary.
- rr_length_score: A normalized character count of the Radiology Report.
- rr_lexical_score: A normalized metric quantifying the vocabulary richness and syntactic difficulty of the Radiology Report.
- ehr_acuity_score: A normalized aggregate score representing physiological deviation and clinical severity, derived from the structured Electronic Health Record variables.

E.2 AgentiCARE Algorithm

Algorithm E1 formalizes the execution flow of the AgentiCARE routing mechanism. For a given patient encounter, the system first extracts structural and clinical features from the available modalities, such as the presence of imaging and the normalized density of the text and time-series data. These features are linearly combined using weights optimized on the training set to produce a continuous patient complexity score (C_p). Based on pre-defined thresholds (τ), the patient is assigned to a specific complexity stratum. This assignment dynamically dictates both the optimal agentic architecture (e.g., Single-Agent versus Centralized Multi-Agent) and the appropriate instruction prompt complexity (e.g., unconstrained, formatting-constrained, or complex diagnostic rules) required to generate the final prediction.

Algorithm E1 AgentiCARE: Patient-Conditional Complexity Routing

Require: Patient encounter $\mathcal{X}_p$, complexity score weights W, architecture strata thresholds $\tau = \{\tau_1, \ldots, \tau_n\}$, framework-instruction mapping $\mathcal{F}$

Ensure: Prediction $\hat{y} \in [0, 1]$

1: **1. Feature Extraction & Scoring**
2: Compute patient features $\mathcal{F}_p = \{f_1, \ldots, f_m\}$ from $\mathcal{X}_p$.
3: Compute complexity score: $C_p = \sum_{i=1}^{m} W_i \cdot \text{normalize}(f_i)$.
4: **2. Dynamic Routing**
5: Select architecture stratum s such that $\tau_{s-1} \leq C_p < \tau_s$.
6: Select framework-instruction configuration $\phi_s = \mathcal{F}(s)$.
7: **3. Prediction**
8: Instantiate framework ϕ_s with instructions $\mathcal{I}_s$.
9: $\hat{y} \leftarrow \phi_s(\mathcal{X}_p, \mathcal{I}_s)$.
10: **return** $\hat{y}$

E.3 Routing Policy and Strata Analysis

Table E1 summarizes the learned weights for the different backbones using the AgentiCARE framework. For Huatuo, the routing decision anchors heavily on the radiology report, with RR length and presence accounting for 64.5% of the decision weight, alongside the presence of a CXR (19.6%) and patient age (8.0%). For Qwen, the generalist routing policy relies predominantly on the RR lexical score (60.4%) and patient age (16.9%).

Table E1: **Learned Complexity Weights by Backbone.** The specific feature weights extracted from the linear routing policy for both the medical specialist (Huatuo) and the generalist (Qwen) backbones.

Clinical Feature	Huatuo	Qwen
has_cxr	0.196	0.031
has_rr	0.443	0.023
age_score	0.080	0.169
is_female	0.001	0.000
is_male	0.001	0.000
ps_length_score	0.021	0.026
ps_lexical_score	0.026	0.000
rr_length_score	0.202	0.147
rr_lexical_score	0.029	0.604
ehr_acuity_score	0.001	0.000

E.4 Extended Calibration Analysis

Table E2 isolates the benefits of dynamic routing from those of post-hoc calibration. Although applying isotonic regression to the static baselines successfully corrects their severe initial miscalibration (reducing ECE to < 0.01), this correction generally comes at a slight cost to AUPRC. Consequently, `AgentiCARE`'s improvements in discriminative performance become even more prominent when compared directly to these explicitly calibrated models. Dynamic routing consistently yields superior AUROC and AUPRC against the calibrated baselines across both backbones, while maintaining a comparable ECE.

Table E2: **Impact of Calibration vs. Dynamic Routing.** Comparison of discriminative performance and confidence calibration. `AgentiCARE` is compared against the best static configuration for each backbone, both before and after isotonic regression.

Backbone	Framework Configuration	Test AUROC	Test AUPRC	Test ECE
Huatuo (Specialist)	Uncalibrated: Independent MAS (Simple)	0.770 [0.752, 0.788]	**0.336 [0.301, 0.376]**	0.153 [0.145, 0.162]
	Calibrated: Independent MAS (Simple)	0.770 [0.749, 0.789]	0.319 [0.286, 0.354]	**0.007 [0.004, 0.016]**
	Dynamic: `AgentiCARE`	**0.784 [0.764, 0.803]**	0.330 [0.295, 0.369]	0.017 [0.009, 0.024]
Qwen (Generalist)	Uncalibrated: Single-Agent (None)	0.772 [0.754, 0.791]	0.332 [0.297, 0.370]	0.113 [0.104, 0.122]
	Calibrated: Single-Agent (None)	0.772 [0.754, 0.790]	0.319 [0.285, 0.357]	**0.003 [0.003, 0.015]**
	Dynamic: `AgentiCARE`	**0.774 [0.756, 0.793]**	**0.347 [0.311, 0.389]**	0.008 [0.005, 0.018]