CoMMa: Contribution-Aware Medical Multi-Agents for Decentralized Oncology Decision Support

Yichen Wu[1]*, Kailong Fan[1]*, Sangjoon Park[2]*, Yuhan Liu[3], Zhiyi Shi[1], Sekeun Kim[1], Dania Daye[4], Hana Farzaneh[1], Wei Liu[5], Xiang Li[1]†, Raul Uppot[1], Yujin Oh[1,2]†, Quanzheng Li[1]†

[1]Harvard Medical School/MGH, [2]Yonsei University, [3]MBZUAI, [4]UW-Madison, [5]Mayo Clinic

{yiwu6, kfan5}@mgh.harvard.edu

Abstract

Recent multi-agent frameworks have shown promise for oncology decision support, yet most assume centralized data access and rely on prompt-based assignment, limiting their applicability in privacy-sensitive clinical settings. We propose Contribution-Aware Medical Multi-Agents (CoMMa), a decentralized LLM-agent framework where specialists operate on partitioned clinical data streams. Unlike prior approaches that share inputs across agents, CoMMa enforces data decentralization to induce stronger role specialization and further enhances this via agent-specific finetuning. To enable reliable and interpretable coordination, we introduce a contribution-aware aggregation mechanism that replaces stochastic, narrative-based reasoning with deterministic embedding projections to approximate each agent's marginal utility. This yields explicit credit assignment over agents, providing a stable and interpretable decision pathway aligned with clinical requirements. We evaluate CoMMa on multiple oncology benchmarks, including real-world multidisciplinary tumor board datasets from large academic hospitals in North America and East Asia, as well as public datasets, demonstrating strong performance and generalization across heterogeneous clinical settings.

1 Introduction

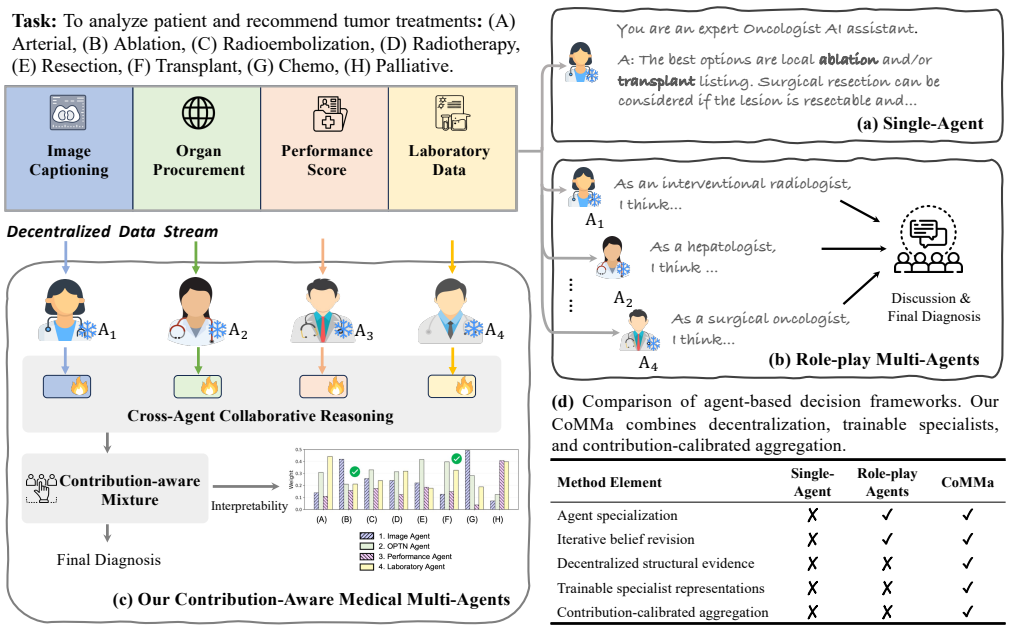
Effective clinical decision support in oncology, such as treatment recommendation, prognosis estimation, and recurrence prediction, requires the seamless integration of longitudinal, heterogeneous data. In practice, this process is carried out through multidisciplinary tumor boards [27, 22], where clinicians jointly discuss patient histories, medical imaging, pathology findings, and evolving laboratory markers. Recently, large language model (LLM)-based multi-agent systems have emerged as a promising paradigm for this setting, distributing complex reasoning across specialized agents to better reflect multidisciplinary workflows [4, 17, 10, 13, 8].

Despite their promise, most existing medical multi-agent frameworks rely on role-based prompting, where agents are assigned clinician-like roles such as oncologists or radiologists [4, 17, 31]. As shown in Fig. 1(b), these agents typically share the same full patient information and communicate through natural language discussion. However, because all agents access the same full patient context, their specialization is mainly defined by prompts rather than distinct clinical evidence or domain-specific adaptation. As a result, agents with different clinical roles often rely on overlapping evidence, making it difficult to determine which information or which agent contributes most to the final decision, even though such interpretability is particularly important in clinical decision-making. Compared with the single-agent setting in Fig. 1(a), current multi-agent frameworks in Fig. 1(b) provide improved

*Equal Contribution

†Corresponding authors: Yujin Oh(yujinoh@yuhs.ac), Xiang Li and Quanzheng Li ({xli60,quanzheng}@mgh.harvard.edu)

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(d) Comparison of agent-based decision frameworks. Our CoMMa combines decentralization, trainable specialists, and contribution-calibrated aggregation.

Method Element	Single-Agent	Role-play Agents	CoMMa
Agent specialization	✗	✓	✓
Iterative belief revision	✗	✓	✓
Decentralized structural evidence	✗	✗	✓
Trainable specialist representations	✗	✗	✓
Contribution-calibrated aggregation	✗	✗	✓

Figure 1: Comparison of medical decision-making frameworks. (a) Single-agent and (b) role-based multi-agent systems rely on shared inputs and narrative reasoning. (c) CoMMa introduces decentralized data streams, agent-specific specialization, and contribution-aware aggregation for interpretable and stable clinical decision-making. (d) Summary table comparing key design elements across the three decision-making frameworks under a unified formulation.

collaboration, but still depend on shared patient context, limited structural specialization, and weak attribution of agent contributions in complex, real-world clinical decision workflows.

To address these limitations, we propose Contribution-Aware Medical Multi-Agents (CoMMa), a framework that moves beyond prompt-based role assignment toward data-decentralized and contribution-aware specialization. As illustrated in Fig. 1(c), CoMMa partitions the clinical context into distinct evidence streams and assigns each stream to a dedicated agent, enabling specialization through separate clinical evidence rather than persona simulation alone. Each agent further uses a domain-specific adaptation module to strengthen specialty-aware representations. Agents then exchange information through deterministic representation-level collaboration, producing updated agent states without relying on free-form natural language discussion.

To support interpretable clinical decision-making, the proposed CoMMa introduces a contribution-aware aggregation mechanism that explicitly models the relative contribution of each specialist agent to the final prediction. Inspired by cooperative game theory, CoMMa regularizes agent contributions using Shapley-based marginal utility estimation [24, 2], enabling more transparent and structured multi-agent reasoning. This design enables structurally specialized collaboration and explicit agent-level attribution without requiring full-context sharing across agents, as summarized in Fig. 1 (c)(d). We evaluate CoMMa across diverse oncology benchmarks, including real-world multidisciplinary tumor board datasets from large academic hospitals in North America and East Asia, together with public benchmarks. Experimental results demonstrate that CoMMa consistently improves predictive performance, decision stability, and attribution interpretability across multiple open-source LLM backbones and clinical decision-making tasks.

In summary, our core contributions are summarized as follows:

- We propose CoMMa, a decentralized medical multi-agent framework that enables structurally specialized reasoning without requiring full-context sharing across agents.
- We introduce a Shapley-regularized contribution-aware aggregation strategy, enabling explicit attribution for interpretable clinical decision-making.

- We validate CoMMa across multiple open-source LLM backbones and diverse real-world oncology benchmarks from large academic hospitals in North America and East Asia, demonstrating consistent improvements over existing single-agent and multi-agent baselines.

2 Related Works

Multi-Agents for Medical Applications. Recent studies have explored multi-agent LLM systems for collaborative clinical decision-making, where agents take specialized medical roles [5, 25, 33, 16, 30, 32, 1, 8]. MedAgents [26] frames problem-solving as a multi-physician debate with answer aggregation, while MDAgents [17] adaptively chooses single-versus multi-agent collaboration to balance accuracy and cost. Beyond free-form discussion, Multi-Agent Conversation [4] structures deliberation to mimic multidisciplinary meetings, and AI Hospital [9] evaluates longitudinal reasoning in a multi-agent clinical simulation. However, these systems are primarily role-based: collaboration is mediated through narrative exchanges and the final decision is fused implicitly, making evidence-level contribution hard to quantify. In contrast, our contribution-aware framework explicitly attributes how each evidence source and agent influences the final prediction, enabling more controlled fusion and more reliable clinical decision outputs.

Game Theory in Clinical Decision-Making. Game theory provides mathematical frameworks for modeling strategic interactions among rational agents and has been widely applied in general-domain decision-making, including economics, management, defense, and computer science [29, 7, 15, 21, 18]. Despite its strong theoretical foundations, its adoption in clinical settings remains limited [20, 19], particularly in oncology. Oncology decision-making relies on heterogeneous and interdependent evidence sources, where quantifying the contribution of each component is essential for interpretability and clinical reliability. Among cooperative game-theoretic approaches, the Shapley value [24, 2] provides a principled and axiomatized solution for fair allocation of a total payoff among participating agents. In a clinical AI setting, the payoff corresponds to the final clinical decision or predictive outcome, while each agent represents a distinct modality or specialized source of clinical information. The Shapley value of an agent is defined as its expected marginal contribution, averaged over all possible coalitions, satisfying desirable payoff. To the best of our knowledge, this work is the first to incorporate Shapley value–based contribution modeling as a regularization mechanism in a multi-agent oncological tumor board framework.

3 Method

To enable structurally specialized and interpretable clinical decision-making, we propose CoMMa, a contribution-aware multi-agent framework for clinical outcome prediction. Given heterogeneous patient information $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, each $\mathbf{x}_i$ denotes a distinct clinical evidence stream, such as imaging findings, laboratory measurements, staging information, or clinical history. Instead of concatenating all streams into a single input, CoMMa keeps them separated and assigns each stream to a specialist agent. The framework then learns how much each agent should contribute to each prediction for each patient. The design preserves the complementary structure of clinical evidence and provides explicit agent-level attribution for the final decision.

As illustrated in Fig. 2, patient evidence is first partitioned into decentralized clinical streams and processed by specialist agents. Each stream is encoded using a frozen open-source LLM encoder followed by a lightweight trainable specialty projection module, producing agent-specific representations. These representations are then refined through cross-agent collaborative reasoning, allowing specialist agents to exchange complementary information while maintaining their evidence-specific roles. Finally, CoMMa performs contribution-aware aggregation through a sample-specific fusion gate regularized by a Shapley contribution estimator, generating the final clinical prediction with explicit attribution of agent contributions.

3.1 Decentralized Specialist Encoding

CoMMa constructs each specialist agent with restricted access to one clinical evidence stream. The i-th agent observes only $\mathbf{x}_i$, rather than the full patient context, so its role is defined by its observation space instead of textual role prompting. Concretely, each stream is encoded by a frozen LLM backbone and an agent-specific trainable projection head:

$$\mathbf{e}_i = \mathrm{LLM}_{\mathrm{frozen}}(\mathbf{x}_i)[\texttt{<CLS>}], \qquad \mathbf{h}_i = f_{\theta_i}(\mathbf{e}_i), \quad i = 1, \ldots, N, \tag{1}$$

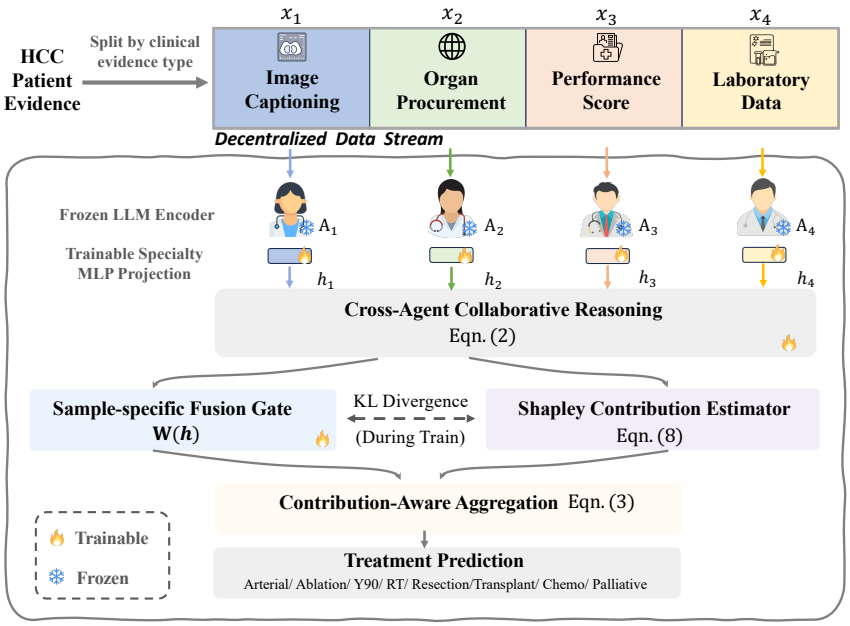


Figure 2: Architectural overview of CoMMa for decentralized multi-agent oncology decision support. Patient evidence is partitioned into distinct clinical streams and processed by specialist agents with separate observation spaces. Agent-specific representations are refined through cross-agent collaborative reasoning and aggregated by a sample-specific contribution gate regularized with Shapley-based attribution during model training.

where $\mathbf{e}_i \in \mathbb{R}^d$ is the deterministic representation of the i-th evidence stream, and $\mathbf{h}_i \in \mathbb{R}^C$ denotes the agent output over C prediction targets. The projection head f_{θ_i} is optimized only for its assigned stream, enabling agent-specific adaptation while keeping the LLM backbone fixed. This decentralized design allows specialist agents to maintain distinct observation spaces and representation pathways throughout collaborative decision-making. In practice, frozen LLM embeddings are cached and reused, so training focuses on the lightweight agent-specific modules.

3.2 Cross-Agent Collaborative Reasoning

Although each specialist agent observes only one clinical evidence stream, effective clinical decision-making still requires integrating complementary information across agents. CoMMa therefore performs collaboration only after decentralized encoding, where agents exchange compact latent representations instead of accessing each other's raw clinical evidence or the full patient context.

For each agent, we summarize the peer representations by averaging the outputs of all other agents, and the i-th agent then refines its representation through a gated residual update:

$$\bar{\mathbf{h}}_{-i} = \frac{1}{N-1}\sum_{j \neq i} \mathbf{h}_j, \qquad \hat{\mathbf{h}}_i = \mathbf{h}_i + \rho\, \sigma\big(g_\psi([\mathbf{h}_i, \bar{\mathbf{h}}_{-i}])\big) \odot u_\psi([\mathbf{h}_i, \bar{\mathbf{h}}_{-i}]). \tag{2}$$

Here, g_ψ and u_ψ are lightweight trainable networks, $\sigma(\cdot)$ denotes the sigmoid function, $\odot$ denotes element-wise multiplication, and ρ controls the residual update scale. This interaction enables each agent to incorporate complementary peer signals while preserving its own evidence-specific representation throughout collaborative clinical decision-making. Unlike natural language debate, the collaboration is deterministic and operates directly on compact agent-level representations.

3.3 Contribution-Aware Aggregation and Learning

After cross-agent collaborative reasoning, CoMMa aggregates the refined specialist outputs through a sample-specific contribution gate. For each patient and prediction target, the gate assigns a normalized

weight to each agent based on its refined representation:

$$W_{i,k} = \frac{\exp\left(G_\phi(\hat{\mathbf{h}}_i, \mathbf{H})_k/\tau\right)}{\sum_{j=1}^{N} \exp\left(G_\phi(\hat{\mathbf{h}}_j, \mathbf{H})_k/\tau\right)}, \qquad \hat{y}_k = \sum_{i=1}^{N} W_{i,k}\hat{h}_{i,k}, \tag{3}$$

where $\mathbf{H} = \{\hat{\mathbf{h}}_1, \ldots, \hat{\mathbf{h}}_N\}$, G_ϕ is a lightweight contribution gate, τ is a temperature parameter, and $\hat{y}_k$ is the prediction for target k. The matrix $\mathbf{W}$ provides a target-specific decomposition of the final prediction across specialist agents.

Learning meaningful contribution weights requires more than prediction supervision alone. CoMMa therefore optimizes three complementary objectives:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BCE}}(\hat{\mathbf{y}}, \mathbf{y}) + \lambda_{\text{pg}}\mathcal{L}_{\text{pg}} + \lambda_{\text{shap}}\mathcal{L}_{\text{shap}}. \tag{4}$$

Objective. $\mathcal{L}_{\text{BCE}}$ optimizes the final clinical prediction; $\mathcal{L}_{\text{pg}}$ encourages agents with higher marginal utility to receive larger contribution weights; $\mathcal{L}_{\text{shap}}$ regularizes the learned weights using Shapley-based coalition utility, enabling more stable and interpretable contribution attribution.

Utility-guided contribution learning. To align the contribution gate with each agent's actual predictive utility, CoMMa first estimates how much each specialist agent contributes to the final prediction. We adopt a leave-one-out strategy that measures the prediction loss change after removing one agent from the aggregation process. Let $\bar{\mathbf{y}}$ denote the aggregated prediction using all refined agents, and let $\bar{\mathbf{y}}_{-i}$ denote the corresponding prediction after excluding agent i. The marginal utility of agent i for target k is defined as

$$r_{i,k} = \mathcal{L}^{(k)}_{\text{BCE}}(\bar{\mathbf{y}}_{-i}, \mathbf{y}) - \mathcal{L}^{(k)}_{\text{BCE}}(\bar{\mathbf{y}}, \mathbf{y}). \tag{5}$$

A positive value indicates that removing agent i increases the prediction loss, suggesting that the agent provides useful evidence for target k. Since the raw utility scores may vary across agents and prediction targets, we further center the rewards across agents to obtain an advantage signal:

$$A_{i,k} = r_{i,k} - \frac{1}{N}\sum_{j=1}^{N} r_{j,k}. \tag{6}$$

This advantage signal serves as utility-based supervision for the contribution gate, encouraging the learned aggregation weights to align with each agent's estimated marginal utility:

$$\mathcal{L}_{\text{pg}} = -\mathbb{E}\left[\sum_{i=1}^{N}\sum_{k=1}^{C} A_{i,k}\log W_{i,k}\right]. \tag{7}$$

This objective encourages the contribution gate to consistently assign larger weights to agents that improve predictive performance.

Shapley-based contribution regularization. Although leave-one-out rewards provide a direct estimate of agent utility, they only evaluate contribution under a single exclusion setting and can therefore be unstable in heterogeneous clinical scenarios. To obtain a more robust contribution target, we further adopt the Shapley principle from cooperative game theory [24, 2], which estimates each agent's average marginal contribution across different collaborating coalitions.

For each agent and prediction target, the Shapley target $\Phi_{i,k}$ is computed as,

$$\Phi_{i,k} \propto \mathbb{E}_{S\subseteq\mathcal{N}\setminus\{i\}}\left[\mathcal{L}^{(k)}_{\text{BCE}}(S, \mathbf{y}) - \mathcal{L}^{(k)}_{\text{BCE}}(S\cup\{i\}, \mathbf{y})\right]_+, \qquad \sum_{i=1}^{N}\Phi_{i,k} = 1. \tag{8}$$

Intuitively, this formulation measures how much agent i improves prediction performance when added to different subsets of collaborating agents, providing a coalition-aware estimate of contribution

importance beyond pairwise or leave-one-out comparisons. The learned contribution gate is then regularized toward this Shapley target:

$$\mathcal{L}_{\text{shap}} = \mathbb{E}\left[\mathrm{KL}(\mathbf{\Phi} \,\|\, \mathbf{W})\right]. \tag{9}$$

Remark. We use the term multi-agent in a structural sense: each specialist has a restricted observation space, an agent-specific transformation, a latent state, and an explicit update conditioned on other agents' states. Unlike standard multi-view fusion, CoMMa refines evidence-specific agent states through inter-agent communication before assigning target-specific contributions to the final clinical decision in heterogeneous clinical decision-making with decentralized information sources.

4 Experiments

4.1 Experimental Settings

Datasets and Implementation Details. Experiments are conducted across four oncology benchmarks: a real-world multidisciplinary tumor board dataset from a large academic healthcare system in North America (HCC Tumorboard; 116 patients), two public oncology benchmarks (MTBBench; 66 patient cases and 573 expert-validated question-answer pairs [28], and LungCure; 999 patients [14]), and an externally annotated clinical dataset from an academic hospital system in East Asia for cross-domain evaluation and interpretability analysis, which we refer to as TUMORBOARDRANK (93 tumor board cases and 276 pairwise option comparisons after filtering single-option cases). Additional details on clinical variables, evidence partitioning, preprocessing, and cohort statistics are provided in Appendix Section A.1. CoMMa is implemented using LangGraph [3] and evaluated with multiple LLM backbones, including Llama3.1-70B-Instruct [12], Qwen2.5-72B [23], and Gemma4-31B [11]. Unless otherwise specified, experiments use four specialist agents and one round of cross-agent collaborative reasoning with residual scale $\rho = 0.5$; TUMORBOARDRANK uses three specialist agents corresponding to basic profile, clinical tests, and treatment history. HCC experiments use learning rate 2×10^{-5}, 121 epochs, $\lambda_{\text{pg}} = 0.5$, and $\lambda_{\text{shap}} = 1.0$. For MTBBench, we use learning rate 10^{-5}, 31 epochs, $\lambda_{\text{pg}} = 1$, and $\lambda_{\text{shap}} = 1.0$. For LungCure, we use learning rate 5×10^{-5}, 100 epochs, $\lambda_{\text{pg}} = 0.1$, and $\lambda_{\text{shap}} = 2.0$. For TUMORBOARDRANK, we use Gemma4-31B with learning rate 5×10^{-5}, 100 epochs, $\lambda_{\text{pg}} = 0.1$, and $\lambda_{\text{shap}} = 4.0$. All models are optimized with Adam (default settings, no weight decay) at a constant learning rate shared by all modules. The gate temperature in Eq. (3) is $\tau = 1.2$, 1.5, and 1.0 for Llama3.1-70B, Qwen2.5-72B, and Gemma4-31B, respectively. All hyperparameters were fixed based on training stability and convergence before the final evaluation. We use two NVIDIA A100 80GB GPUs for training each model.

Evaluation Protocol. We report Area Under the ROC Curve (AUC) and Accuracy at a fixed threshold of 0.5. For HCC, AUC and Acc are first computed for each treatment category and then macro-averaged across valid categories. If the test split contains only one label value for a category, that category is excluded from the HCC macro-AUC calculation. For MTBBench, the progression and recurrence endpoints are evaluated as separate binary tasks. For LungCure, we evaluate six binary staging indicators spanning primary tumor extent (T), regional nodal involvement (N), and distant metastasis (M), and report macro-averaged AUC and Acc across these indicators. For TUMORBOARDRANK, we evaluate candidate management-option ranking using Top-1 Acc, Case Distribution Acc, and Pair Order Acc. Top-1 Acc measures whether the highest-ranked predicted option matches the option with the highest physician consensus; Case Distribution Acc measures whether the full predicted option ordering matches the physician-consensus ordering; and Pair Order Acc measures the fraction of option pairs ordered consistently with physician consensus. All reported summary results are computed over three random seeds, using seeds 1000, 2000, and 3000.

Baseline Methods. We compare CoMMa with single-agent and multi-agent LLM baselines under the same evaluation protocol. The direct single-agent baseline receives the full patient case as one input and outputs a probability for each target without explicit multi-agent interaction. MDAgents Basic [17] is implemented as an adapted medical decision-agent baseline. For HCC, we use the adapted local-LLM MDAgents script with the same HCC label space; for MTBBench, we use the corresponding binary prognosis version. MAC [4] is implemented as a multi-agent consultation baseline in which multiple role-specific doctor agents iteratively estimate the target probability, and the final prediction is obtained from the last consensus round. AI-Hospital [9] is implemented as a collaborative consultation baseline in which several doctor agents provide specialty estimates and a host physician aggregates them into a final probability.

Table 1: Performance comparison on three oncology benchmarks: HCC Tumorboard, LungCure, and MTBBench. CoMMa consistently improves predictive performance over single-agent and existing multi-agent baselines across multiple open-source LLM backbones.

Methods	HCC Tumorboard		LungCure		MTBBench			
	Treatment Recommendation		TNM Stage Prediction		Progression		Recurrence	
	AUC	Acc	AUC	Acc	AUC	Acc	AUC	Acc
GPT 4.1								
In-context learning	N/A	0.784 ± 0.007	N/A	0.798 ± 0.070	N/A	0.485 ± 0.113	N/A	0.425 ± 0.086
Llama3.1-70B [12]								
Single-Agent	0.590 ± 0.066	0.690 ± 0.064	0.782 ± 0.061	0.735 ± 0.044	0.503 ± 0.001	0.519 ± 0.004	0.526 ± 0.018	0.491 ± 0.011
MDAgents [17]	0.693 ± 0.040	0.708 ± 0.019	0.781 ± 0.051	0.746 ± 0.048	0.523 ± 0.010	0.526 ± 0.032	0.536 ± 0.013	0.491 ± 0.023
MAC [4]	0.666 ± 0.055	0.736 ± 0.035	0.779 ± 0.069	0.731 ± 0.099	0.537 ± 0.011	0.544 ± 0.021	0.530 ± 0.016	0.491 ± 0.020
AI-Hospital [9]	0.643 ± 0.038	0.767 ± 0.007	0.810 ± 0.048	0.728 ± 0.061	0.547 ± 0.032	0.439 ± 0.026	0.581 ± 0.008	0.546 ± 0.016
CoMMa (Ours)	$\mathbf{0.711 \pm 0.011}$	$\mathbf{0.795 \pm 0.010}$	$\mathbf{0.837 \pm 0.053}$	$\mathbf{0.797 \pm 0.038}$	$\mathbf{0.756 \pm 0.171}$	$\mathbf{0.666 \pm 0.105}$	$\mathbf{0.711 \pm 0.214}$	$\mathbf{0.606 \pm 0.138}$
Qwen2.5-72B [23]								
Single-Agent	0.593 ± 0.072	0.650 ± 0.035	0.813 ± 0.032	0.713 ± 0.069	0.517 ± 0.015	0.543 ± 0.010	0.502 ± 0.006	0.436 ± 0.004
MDAgents [17]	0.602 ± 0.080	0.710 ± 0.024	0.803 ± 0.034	0.750 ± 0.039	0.517 ± 0.008	0.544 ± 0.020	0.501 ± 0.023	0.437 ± 0.010
MAC [4]	0.689 ± 0.070	0.639 ± 0.044	0.799 ± 0.050	0.761 ± 0.057	0.531 ± 0.009	0.543 ± 0.020	0.525 ± 0.006	0.473 ± 0.004
AI-Hospital [9]	0.637 ± 0.100	0.727 ± 0.034	0.827 ± 0.079	0.792 ± 0.090	0.542 ± 0.015	0.544 ± 0.002	0.517 ± 0.001	0.455 ± 0.010
CoMMa (Ours)	$\mathbf{0.700 \pm 0.014}$	$\mathbf{0.750 \pm 0.032}$	$\mathbf{0.844 \pm 0.020}$	$\mathbf{0.802 \pm 0.044}$	$\mathbf{0.809 \pm 0.090}$	$\mathbf{0.666 \pm 0.053}$	$\mathbf{0.845 \pm 0.107}$	$\mathbf{0.576 \pm 0.139}$
Gemma4-31B [11]								
Single-Agent	0.625 ± 0.058	0.693 ± 0.047	0.802 ± 0.063	0.815 ± 0.061	0.556 ± 0.024	0.491 ± 0.045	0.534 ± 0.017	0.491 ± 0.033
MDAgents [17]	0.645 ± 0.014	0.718 ± 0.021	0.826 ± 0.053	0.817 ± 0.064	0.590 ± 0.022	0.543 ± 0.018	0.616 ± 0.030	0.564 ± 0.056
MAC [4]	0.645 ± 0.067	0.719 ± 0.039	0.822 ± 0.044	0.827 ± 0.059	0.619 ± 0.040	0.597 ± 0.014	0.597 ± 0.031	0.564 ± 0.017
AI-Hospital [9]	0.640 ± 0.028	0.747 ± 0.036	0.791 ± 0.105	0.805 ± 0.081	0.616 ± 0.044	0.614 ± 0.019	0.623 ± 0.014	0.583 ± 0.022
CoMMa (Ours)	$\mathbf{0.680 \pm 0.028}$	$\mathbf{0.758 \pm 0.027}$	$\mathbf{0.841 \pm 0.048}$	$\mathbf{0.830 \pm 0.073}$	$\mathbf{0.905 \pm 0.083}$	$\mathbf{0.697 \pm 0.105}$	$\mathbf{0.911 \pm 0.077}$	$\mathbf{0.666 \pm 0.139}$

Table 2: Ablation study on HCC Tumorboard treatment recommendation.

Eqn. (2)	$\mathcal{L}_{\text{pg}}$	$\mathcal{L}_{\text{shap}}$	Llama3.1-70B		Qwen2.5-72B	
			AUC	Acc	AUC	Acc
✓	✓		0.706 ± 0.007	0.688 ± 0.068	0.682 ± 0.015	0.738 ± 0.004
✓		✓	0.691 ± 0.015	0.659 ± 0.042	0.691 ± 0.031	0.728 ± 0.015
	✓	✓	0.692 ± 0.012	0.741 ± 0.056	0.681 ± 0.024	0.722 ± 0.046
✓	✓	✓	$\mathbf{0.711 \pm 0.011}$	$\mathbf{0.795 \pm 0.010}$	$\mathbf{0.700 \pm 0.014}$	$\mathbf{0.750 \pm 0.032}$
1. Replace learned weights with uniform dist.			0.693 ± 0.008	0.683 ± 0.072	0.625 ± 0.023	0.661 ± 0.055
2. Sampled-coalition Shapley approximation.			0.689 ± 0.011	0.717 ± 0.049	0.698 ± 0.013	0.725 ± 0.029
3. Use centralized input.			0.691 ± 0.006	0.734 ± 0.055	0.692 ± 0.008	0.613 ± 0.086

4.2 Main Results

Comparison on different oncology benchmarks. Table 1 compares CoMMa with single-agent and existing multi-agent baselines across HCC Tumorboard, LungCure, and MTBBench. CoMMa consistently achieves the best overall performance across benchmarks and LLM backbones. On HCC Tumorboard, CoMMa obtains the highest treatment recommendation accuracy under all three backbones, improving Llama3.1-70B from 0.690 to 0.795 and Qwen2.5-72B from 0.650 to 0.750 over the single-agent baseline. On LungCure, CoMMa also improves TNM stage prediction, achieving the best accuracy with Qwen2.5-72B (0.802) and Gemma4-31B (0.830). The largest gains appear on MTBBench, where CoMMa substantially improves progression and recurrence prediction. These results show that CoMMa generalizes across real-world and public oncology benchmarks while consistently improving predictive performance.

Comparison on various LLM backbones. CoMMa shows consistent gains across Llama3.1-70B, Qwen2.5-72B, and Gemma4-31B, indicating that the framework is not tied to a specific backbone. With Llama3.1-70B, CoMMa improves HCC treatment recommendation AUC from 0.590 to 0.711 and accuracy from 0.690 to 0.795 over the single-agent baseline. It also improves MTBBench progression AUC from 0.503 to 0.756 and recurrence AUC from 0.526 to 0.711. Similar trends are observed with Qwen2.5-72B, where CoMMa achieves the best results among open-source baselines across all tasks, including LungCure accuracy of 0.802 and MTBBench recurrence AUC of 0.845. With Gemma4-31B, CoMMa yields especially large improvements on MTBBench, increasing progression AUC from 0.556 to 0.905 and recurrence AUC from 0.534 to 0.911 over the single-agent baseline. Overall, these results show that decentralized specialist encoding and contribution-aware aggregation provide robust improvements across different LLM families and model scales.

4.3 Experimental Analysis

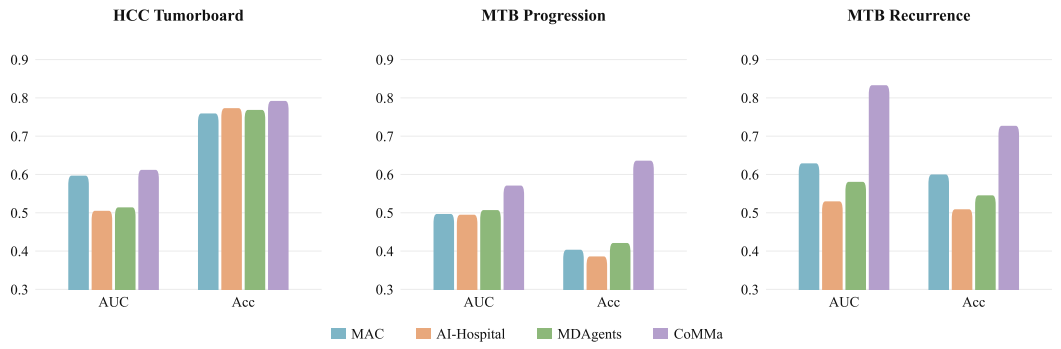


Figure 3: Comparison of AUC/Acc across different methods using the medical-related Meditron-70B backbone [6] on multiple clinical evaluation tasks.

To further investigate the behavior of CoMMa beyond the quantitative performance improvements reported above, we conduct additional analyses to better understand how the proposed framework achieves its gains, whether the learned contribution patterns are clinically meaningful, and how well the framework generalizes across different model settings and medical foundation backbones. Specifically, the following analyses are organized around four key questions.

Question 1: How does each component of CoMMa contribute to performance? Table 2 reports ablation results on HCC treatment recommendation. Removing either contribution-learning objective degrades performance, showing that $\mathcal{L}_{\text{pg}}$ and $\mathcal{L}_{\text{shap}}$ provide complementary supervision for the contribution gate. We further analyze the effect of contribution-aware aggregation. Replacing the learned contribution weights with uniform averaging degrades both AUC and Acc, showing that patient-specific agent weighting is important for effective multi-agent collaboration. In addition, replacing exact Shapley computation with Monte Carlo approximation leads to lower performance, suggesting that accurate coalition-aware contribution estimation improves the stability of contribution learning. Finally, using centralized input degrades performance, indicating the benefit of decentralized data streams. Overall, the ablation results demonstrate that both contribution-aware weighting and Shapley-based regularization are important components of the proposed CoMMa algorithm.

Table 3: Fine-tuned baselines (mean of 3 seeds).

HCC Tumorboard	Llama		Qwen		Gemma	
	AUC	Acc	AUC	Acc	AUC	Acc
AI-Hospital-FT	0.676	0.753	0.684	0.738	0.607	0.751
CoMMa	**0.711**	**0.795**	**0.700**	**0.750**	**0.680**	**0.758**

Question 2: What if other multi-agent methods also adopt finetuning? To examine whether CoMMa's improvement is simply due to supervised adaptation, we introduce fine-tuned variants of existing SOTA multi-agent baselines, AI-Hospital-FT. All fine-tuned baselines use the same backbone, data split, prediction labels, training objective, and evaluation protocol as CoMMa. As shown in Table 3, finetuning improves these baselines over their prompt-only versions, but CoMMa still achieves the best overall performance. This indicates that CoMMa's gains are not merely a result of finetuning, but are also attributable to decentralized evidence specialization, latent cross-agent communication, and contribution-aware aggregation. Additional fine-tuned baselines on HCC Tumorboard and MTBBench are provided in Section A.4.

Question 3: Does CoMMa remain effective with medical foundation models? As shown in Fig. 3, we evaluate CoMMa on Meditron-70B [6], where it consistently delivers improvements over strong domain-specific LLMs. CoMMa achieves higher AUC and accuracy across all tasks, demonstrating its effectiveness even when built on top of medical-specialized backbones. These results indicate that CoMMa provides complementary gains beyond model pretraining and remains robust across different backbone choices and diverse clinical evaluation settings.

Question 4: Does CoMMa provide clinically meaningful contribution attribution? We conduct this analysis on TUMORBOARDRANK, an externally annotated tumor board option-ranking dataset with physician-consensus preferences. As shown in Fig. 4, we compare CoMMa with Single-Agent and AI-Hospital, the strongest overall baseline. Using Gemma4-31B, CoMMa improves Top-1 Acc, Case Distribution Acc, and Pair Order Acc over both baselines, suggesting better recovery of physician-consensus rankings. In the representative case, CoMMa produces a normalized ranking distribution that better follows physician consensus, while the baselines over-weight lower-ranked

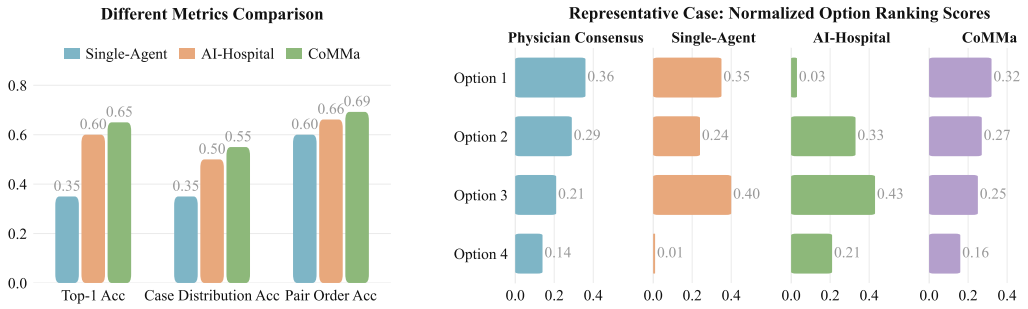


Figure 4: Performance and representative case analysis on TUMORBOARDRANK using Gemma4-31B. Left: overall comparison. Right: normalized option-ranking scores for a representative case, with physician consensus as the reference distribution.

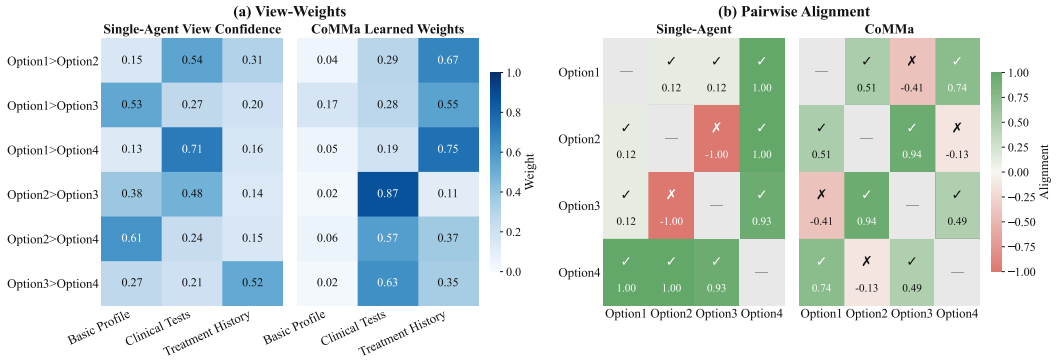


Figure 5: View-weight and pairwise-alignment analysis for a representative TUMORBOARDRANK case using Gemma4-31B. (a) View-level confidence and learned CoMMa weights across evidence streams. (b) Pairwise alignment with physician-consensus option ordering, where sign indicates agreement and magnitude indicates confidence.

alternatives. Fig. 5 further analyzes the same case. Single-Agent view confidence is computed from frozen Gemma4-31B view-level logit margins, whereas CoMMa weights are learned by the fusion module. CoMMa assigns larger weights to clinically informative streams, such as clinical tests and treatment history, and down-weights the less informative basic profile view, indicating learned evidence reweighting beyond raw view confidence. Although CoMMa makes one more pairwise error than the view-level Gemma4 baseline, this error has low confidence, whereas the baseline makes a high-confidence error. This explains why CoMMa recovers the correct case-level distribution in Fig. 4, while the baseline assigns excessive mass to a lower-ranked option.

5 Conclusion and Limitation

In this work, we propose CoMMa, a contribution-aware medical multi-agent framework that replaces prompt-based role interaction with decentralized specialist modeling and structured contribution learning. By combining evidence-specific specialist agents, deterministic cross-agent collaboration, and Shapley-regularized contribution aggregation, CoMMa provides improved predictive performance and explicit agent-level attribution. The decentralized design of CoMMa better supports privacy-sensitive and resource-constrained clinical deployment settings by avoiding full-context sharing and expensive cloud-based inference. Nevertheless, several evaluation sets are small, leading to large variance on MTBBench, and the learned contribution weights reflect each agent's role in the fused prediction rather than causal effects of the underlying evidence. Because incorrect recommendations could harm patients, CoMMa is intended as decision support under clinician oversight and requires prospective validation before clinical use. Future work will extend CoMMa beyond oncology to broader clinical domains and more heterogeneous healthcare settings, refine contribution estimation and representation disentanglement for correlated evidence streams, and generalize contribution-aware multi-agent reasoning to open-ended clinical decisions with ambiguous outcomes.

Acknowledgement

This work was supported in part by the National Institutes of Health (NIH) under Grant R01HL159183 and R01EB038264, the National Research Foundation of Korea (NRF) grant (RS-2026-25469863), and the Institute for Information & Communications Technology Planning & Evaluation (IITP) grant (IITP-2026-RS-2026-25616313, AI Star Fellowship Support Program), funded by the Korean government (MSIT).

References

[1] Mohammad Almansoori, Komal Kumar, and Hisham Cholakkal. Medagentsim: Self-evolving Multi-agent Simulations for Realistic Clinical Interactions. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 362–372, 2025.

[2] Daniel Beechey, Thomas M. S. Smith, and Özgür Şimşek. Explaining Reinforcement Learning with Shapley Values. In *International Conference on Machine Learning*, pages 2003–2014, 2023.

[3] Harrison Chase. Langchain, 2022.

[4] Xi Chen, Huahui Yi, Mingke You, WeiZhi Liu, Li Wang, Hairui Li, Xue Zhang, Yingman Guo, Lei Fan, Gang Chen, et al. Enhancing Diagnostic Capability with Multi-agents Conversational Large Language Models. *NPJ digital medicine*, 8(1):159, 2025.

[5] Xiaolan Chen, Jiayang Xiang, Shanfu Lu, Yexin Liu, Mingguang He, and Danli Shi. Evaluating Large Language Models and Agents in Healthcare: Key Challenges in Clinical Applications. *Intelligent Medicine*, 5(2):151–163, 2025.

[6] Zeming Chen, Ehsaneddin Asgari, Seyed Mohammad Mousavi, et al. Meditron: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*, 2023.

[7] Christophe Dutang, Hansjoerg Albrecher, and Stéphane Loisel. Competition Among Non-life Insurers under Solvency Constraints: A Game-theoretic Approach. *European Journal of Operational Research*, 231(3):702–711, 2013.

[8] Adibvafa Fallahpour, Jun Ma, Alif Munim, Hongwei Lyu, and Bo Wang. Medrax: Medical Reasoning Agent for Chest X-ray. In *International Conference on Machine Learning*, pages 15661–15676, 2025.

[9] Zhihao Fan, Lai Wei, Jialong Tang, Wei Chen, Wang Siyuan, Zhongyu Wei, and Fei Huang. AI Hospital: Benchmarking Large Language Models in a Multi-agent Medical Interaction Simulator. In *International Conference on Computational Linguistics*, pages 10183–10213, 2025.

[10] Dyke Ferber, Omar SM El Nahhas, Georg Wölflein, Isabella C Wiest, Jan Clusmann, Marie-Elisabeth Leßmann, Sebastian Foersch, Jacqueline Lammert, Maximilian Tschochohei, Dirk Jäger, et al. Development and validation of an Autonomous Artificial Intelligence Agent for Clinical Decision-making in Oncology. *Nature cancer*, pages 1–13, 2025.

[11] Google DeepMind. Gemma 4 31b it. Hugging Face model card, 2026. Accessed: 2026-05-06.

[12] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024.

[13] Xudong Han, Xianglun Gao, Xiaoyi Qu, and Zhenyu Yu. Multi-Agent Medical Decision Consensus Matrix System: An Intelligent Collaborative Framework for Oncology MDT Consultations. *arXiv preprint arXiv:2512.14321*, 2025.

[14] Fangyu Hao, Jiayu Yang, Yifan Zhu, Zijun Yu, Qicen Wu, Wang Yunlong, Jiawei Li, Yulin Liu, Xu Zeng, Guanting Chen, et al. Lungcure: Benchmarking multimodal real-world clinical reasoning for precision lung cancer diagnosis and treatment. *arXiv preprint arXiv:2604.06925*, 2026.

[15] Edwin Ho, Arvind Rajagopalan, Alex Skvortsov, Sanjeev Arulampalam, and Mahendra Piraveenan. Game Theory in Defence Applications: A Review. *Sensors*, 22(3):1032, 2022.

[16] Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Leandra A Barnes, Hong-Yu Zhou, Zhuo Ran Cai, Eliezer M Van Allen, David Kim, et al. An Evaluation Framework for Clinical Use of Large Language Models in Patient Interaction Tasks. *Nature medicine*, 31(1):77–86, 2025.

[17] Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. MDagents: An Adaptive Collaboration of LLMs for Medical Decision-making. *Advances in Neural Information Processing Systems*, pages 79410–79452, 2024.

[18] Marc Lanctot, Vinicius Zambaldi, Audrūnas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Pérolat, David Silver, and Thore Graepel. A Unified Game-theoretic Approach to Multiagent Reinforcement Learning. In *Advances in Neural Information Processing Systems*, page 4193–4206, 2017.

[19] Chng Wei Lau, Daniel Catchpoole, Simeon Simoff, Dongmo Zhang, and Quang Vinh Nguyen. A Game-theoretical Approach to Clinical Decision Making with Immersive Visualisation. *Applied Sciences*, 13(18):10178, 2023.

[20] Francisco V Mendonça, Margarida Catalao-Lopes, Rui Tato Marinho, and Jose Rui Figueira. Improving Medical Decision-making with a Management Science Game Theory Approach to Liver Transplantation. *Omega*, 94:102050, 2020.

[21] Simon Parsons and Michael Wooldridge. Game Theory and Decision Theory in Multi-agent Systems. *Autonomous Agents and Multi-Agent Systems*, 5(3):243–254, 2002.

[22] Vivek Patkar, Dionisio Acosta, Tim Davidson, Alison Jones, John Fox, and Mohammad Keshtgar. Cancer Multidisciplinary Team Meetings: Evidence, Challenges, and the Role of Clinical Decision Support Technology. *International Journal of Breast Cancer*, 2011(1):831605, 2011.

[23] Alibaba Group Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

[24] LS Shapley. A Value For n-Person Games'. *Contributions to the Theory of Games*, (28):307, 1953.

[25] Xiangru Tang, Daniel Shao, Jiwoong Sohn, Jiapeng Chen, Jiayi Zhang, Jinyu Xiang, Fang Wu, Yilun Zhao, Chenglin Wu, Wenqi Shi, et al. Medagentsbench: Benchmarking Thinking Models and Agent Frameworks for Complex Medical Reasoning. *arXiv preprint arXiv:2503.07459*, 2025.

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

[27] Cath Taylor, Alastair J Munro, Rob Glynne-Jones, Clive Griffith, Paul Trevatt, Michael Richards, and Amanda J Ramirez. Multidisciplinary Team Working in Cancer: What is the Evidence? *British Medical Journal*, 340:743–745, 2010.

[28] Kiril Vasilev, Alexandre Misrahi, Eeshaan Jain, Phil F Cheng, Petros Liakopoulos, Olivier Michielin, Michael Moor, and Charlotte Bunne. MTBBench: A Multimodal Sequential Clinical Decision-Making Benchmark in Oncology. *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

[29] John Von Neumann and Oskar Morgenstern. Theory of Games and Economic Behavior: 60th Anniversary Commemorative Edition. In *Theory of Games and Economic Behavior*. Princeton University Press, 2007.

[30] Ziyue Wang, Junde Wu, Linghan Cai, Chang Han Low, Xihong Yang, Qiaxuan Li, and Yueming Jin. MedAgent-Pro: Towards Evidence-Based Multi-Modal Medical Diagnosis via Reasoning Agentic Workflow. *arXiv preprint arXiv:2503.18968*, 2025.

[31] Weitong Zhang, Mengyun Qiao, Chengqi Zang, Steven Niederer, Paul M Matthews, Wenjia Bai, and Bernhard Kainz. Multi-Agent Reasoning for Cardiovascular Imaging Phenotype Analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 429–439, 2025.

[32] Yucheng Zhou, Lingran Song, and Jianbing Shen. MAM: Modular Multi-Agent Framework for Multi-Modal Medical Diagnosis via Role-Specialized Collaboration. *arXiv preprint arXiv:2506.19835*, 2025.

[33] James Zou and Eric J Topol. The Rise of Agentic AI Teammates in Medicine. *The Lancet*, 405(10477):457, 2025.

A Appendix

A.1 Dataset Details.

A.1.1 HCC Tumorboard

The HCC Tumorboard dataset comprises 116 patients diagnosed with primary Hepatocellular Carcinoma (HCC) at a large tertiary academic hospital in North America. The study was approved by the institutional review board (IRB #2022P001512). The cohort includes adult patients with confirmed hepatocellular carcinoma, no concurrent malignancies, and multidisciplinary tumor board review during the study period. Each case was reviewed by a multidisciplinary tumor board to establish the consensus treatment recommendation, which serves as the ground-truth label. Cases without recommendations in all treatment categories were excluded. We adopt a 75/25% train-test split. The ground-truth label is defined over eight discrete treatment options: (A) transcatheter arterial chemoembolization (TACE), (B) ablation, primarily microwave ablation (MWA), (C) radioembolization/ selective internal radiation therapy (SIRT), (D) radiotherapy, (E) surgical resection, (F) liver transplant, (G) chemotherapy, and (H) palliative care. Each instance contains 24 multimodal clinical variables represented in textual format, extracted from longitudinal clinical records and MRI reports. An exemplar case illustrating the HCC Tumorboard data construction is provided in Table 4.

Table 4: HCC Tumorboard Data Formation.

	Decentralized Data			
	$\mathbf{x}_1$ MRI imaging features	$\mathbf{x}_2$ OPTN[†] information	$\mathbf{x}_3$ Performance score	$\mathbf{x}_4$ Laboratory data
Context example	Number of tumors: 4.0 Size of largest tumor (cm): 2.6 Location largest (segment): 8 Vascular invasion: 0 Extrahepatic spread: 0 Thrombus (portal vein): 0 ⋮	Number of OPTN 5A tumors: 0 Number of OPTN 5B tumors: 2 Number of OPTN 5X tumors: 0 Number of OPTN 5T tumors: 0	Encephalopathy: 0 Child-Pugh score: A ECOG[‡] score: 1 MELD[§] score: 13.0	Ascites: 0 Bilirubin: 0.6 Albumin: 4.8 INR: 1.1 AFP: 5.0 ⋮

[†]OPTN: Organ Procurement & Transplantation Network, [‡]ECOG: Eastern Cooperative Oncology Group Performance, [§]MELD: Model for End-stage Liver Disease

A.1.2 MTBBench

MTBBench [28] is an agentic benchmark designed to evaluate molecular tumor board-style reasoning that requires integrating multimodal evidence and longitudinal patient context. It contains 66 patient cases and 573 expert-validated question-answer pairs across two tracks. The multimodal track includes 26 head and neck cancer cases, with approximately 40 modality-specific files per patient, including H&E slides, IHC images, and hematology reports, yielding 390 questions. The longitudinal track includes 40 cases with approximately five structured files per patient and 183 questions focused on time-dependent outcomes, including recurrence and progression. In our experiments, we focus on prognosis prediction tasks for tumor recurrence and progression. Each sample is formulated as a paired clinical question and binary endpoint label. We adopt an 80/20 % train-test split for MTBBench. An exemplar case illustrating the MTBBench data construction is provided in Table 5.

Table 5: MTBBench Data Formation.

	Decentralized Data			
	$\mathbf{x}_1$ Specimen metadata	$\mathbf{x}_2$ Somatic mutations	$\mathbf{x}_3$ Copy number alterations	$\mathbf{x}_4$ Structural variants
Context example	Colon Adenocarcinoma Primary (Sigmoid Colon) Tumor Purity: 20.0% MSI Type: Stable ⋮	TP53 p.C242F FANCA p.V750M CDK12 p.S238∗ ⋮	ERBB2 gain (Value: 2.0) CDK12 gain (Value: 2.0) ⋮	CTNNB1 Splice Site deletion (c.18_242-37del) ⋮

A.1.3 LungCure

LungCure [14] is a lung cancer staging benchmark designed to evaluate clinical reasoning from longitudinal OCR-derived patient records. It contains 999 patients, and we adopt an 80/20 % train-test split. In our experiments, we focus on six binary staging indicators spanning primary tumor

extent (T), regional nodal involvement (N), and distant metastasis (M): T2a, T1c, N0, N2, M0, and M1c. Each sample is formulated as a patient record paired with a binary TNM staging label. We partition each record into TNM-oriented evidence streams, including primary tumor evidence, nodal evidence, metastasis evidence, and clinical history. An exemplar case illustrating the LungCure data construction is provided in Table 6.

Table 6: LungCure Data Formation.

	Decentralized Data			
	$\mathbf{x}_1$ Primary tumor evidence	$\mathbf{x}_2$ Regional nodal evidence	$\mathbf{x}_3$ Distant metastasis evidence	$\mathbf{x}_4$ Clinical history
Context example	Right upper-lobe solid nodule Size: 3.1×3.3×2.7 cm SUVmax: 8.1 Spiculation and pleural traction ⋮	Right hilar lymph nodes enlarged Right paratracheal lymph nodes Increased FDG uptake SUVmax: 5.9 ⋮	No definite distant metastasis No abnormal liver uptake No pelvic or retroperitoneal nodes T4 bone island, favored benign ⋮	Smoking history over 30 years Family history of ovarian cancer PET/CT for systemic assessment Thoracic surgery outpatient referral ⋮

A.1.4 TUMORBOARDRANK

TUMORBOARDRANK is an externally annotated tumor board option-ranking dataset for evaluating cross-domain clinical decision ranking and interpretability. It contains 93 tumor board cases after filtering single-option cases, yielding 276 pairwise option comparisons. Each case consists of a refined clinical case summary, physician-consensus preferences over candidate management options, and the corresponding option descriptions. We use a case-level temporal split, with 73 cases for training and 20 cases for testing. For CoMMa, each case is partitioned into three evidence streams: basic profile, clinical tests, and treatment history. An exemplar case illustrating the TUMORBOARDRANK data construction is provided in Table 7.

Table 7: TUMORBOARDRANK Data Formation.

	Decentralized Data		
	$\mathbf{x}_1$ Basic profile	$\mathbf{x}_2$ Clinical tests	$\mathbf{x}_3$ Treatment history
Context example	37-year-old man Laryngeal squamous cell carcinoma ECOG performance status: 0 No chronic high-grade toxicity ⋮	Isolated nodal recurrence Salvage neck dissection 1 of 10 lymph nodes positive Extranodal extension present ⋮	Prior definitive chemoradiation 69.96 Gy in 33 fractions 56.1 Gy elective nodal irradiation Cisplatin cumulative dose: 200 mg/m^2 ⋮

A.2 Detailed Shapley Computation.

A.2.1 Comparison to Advantage Vector

Unlike the advantage vector $\mathbf{A}_i$, which only considers the marginal impact of removing agent i from the full coalitions $\mathcal{N}$, Shapley value evaluates the contribution of agent i across a diverse set of sub-coalitions $S \subseteq \mathcal{N} \setminus \{i\}$ of varying cardinalities. Consequently, the resulting Shapley matrix $\mathbf{\Phi}$ captures an expected marginal contribution that explicitly accounts for inter-agent interactions and dependencies, rather than being dominated by sample-specific or coalition-dependent fluctuations.

A.3 Comparison on Computational Cost.

We compare computational complexity and cost across trainable baselines, single-agent in-context learning, role-based multi-agent systems, and our locally fine-tunable multi-agent approach in Table 8. Latency and inference token usage are averaged over 30 samples from the HCC Tumorboard dataset. Our results show that CoMMa is a cost-effective method, achieving lower latency under a free-inference setting while delivering superior performance.

A.4 Fine-Tuned Multi-Agent Baselines

To assess whether CoMMa's gains can be explained solely by supervised fine-tuning, we construct fine-tuned variants of representative multi-agent baselines. These baselines use the same LLM backbone, train-test split, prediction labels, supervision objective, and evaluation protocol as CoMMa.

Table 8: Computational Complexity and Cost Comparison on HCC Tumorboard Dataset.

Methods	**Trainable Baselines**	**Single-Agent In-context Learning**	**Role-based Multi-Agents**	**Locally Fine-tunable Multi-Agents**
	CatBoost	GPT 4.1	MDAgents	CoMMa
Computational Complexity	0.048 MFLOPs	Unreleased	Unreleased	280 TFLOPs
Total Trainable Parameters	0.512 M	None	None	16.8 M
Latency (per input)	0.116 ms	4.2 s	79.2 s	1.1 s
Inference Tokens (chargeable)	None	20,878	23,254	None

Table 9: Comparison with fine-tuned multi-agent baselines on HCC Tumorboard. Results are averaged over 3 random seeds.

HCC	**Llama**		**Qwen**		**Gemma**	
	AUC	Acc	AUC	Acc	AUC	Acc
MAC	0.656	0.780	0.690	0.738	0.650	0.739
MDAgents	0.676	0.776	0.685	0.741	0.626	0.731
AI-Hospital	0.676	0.753	0.684	0.738	0.607	0.751
CoMMa	**0.711**	**0.795**	**0.700**	**0.750**	**0.680**	**0.758**

The main paper reports AI-Hospital-FT, while additional fine-tuned baselines are summarized in Table 9. For each baseline, we preserve its original multi-agent organization whenever possible and add a supervised prediction head trained on the task-specific labels. The frozen LLM encoder is used to obtain case-level or agent-level representations, and the trainable head is optimized with the same binary cross-entropy objective used by CoMMa. This setup controls for the effect of task-specific supervision while retaining each baseline's original agent structure.

MTBBench. We further evaluate the fine-tuned baselines on MTBBench under the same protocol, reporting mean $\pm$ standard deviation over three random seeds in Table 10. CoMMa obtains the best or tied-best mean on every metric and backbone; with Llama3.1-70B, it ties with AI-Hospital-FT on recurrence AUC and with MDAgents-FT on recurrence accuracy. Unlike on HCC Tumorboard, fine-tuning does not consistently improve the baselines over their prompt-only counterparts on MTBBench. Given the large standard deviations on this benchmark (see Section A.5), these comparisons should be read as indicative rather than conclusive.

Table 10: Comparison with fine-tuned multi-agent baselines on MTBBench. Results are mean $\pm$ standard deviation over 3 random seeds. Best and tied-best means within each backbone are in bold.

Backbone	**Method**	**Progression**		**Recurrence**	
		AUC	Acc	AUC	Acc
Llama3.1-70B	MAC-FT	0.548 ± 0.135	0.659 ± 0.052	0.544 ± 0.154	0.485 ± 0.139
	MDAgents-FT	0.583 ± 0.144	0.515 ± 0.139	0.678 ± 0.084	$\mathbf{0.606 \pm 0.052}$
	AI-Hospital-FT	0.571 ± 0.186	0.606 ± 0.052	$\mathbf{0.711 \pm 0.192}$	0.455 ± 0.000
	CoMMa	$\mathbf{0.756 \pm 0.171}$	$\mathbf{0.666 \pm 0.105}$	$\mathbf{0.711 \pm 0.214}$	$\mathbf{0.606 \pm 0.138}$
Qwen2.5-72B	MAC-FT	0.619 ± 0.232	0.636 ± 0.000	0.622 ± 0.107	0.515 ± 0.052
	MDAgents-FT	0.548 ± 0.144	0.636 ± 0.000	0.567 ± 0.133	0.536 ± 0.052
	AI-Hospital-FT	0.643 ± 0.164	0.636 ± 0.000	0.744 ± 0.038	0.555 ± 0.091
	CoMMa	$\mathbf{0.809 \pm 0.090}$	$\mathbf{0.666 \pm 0.053}$	$\mathbf{0.845 \pm 0.107}$	$\mathbf{0.576 \pm 0.139}$
Gemma4-31B	MAC-FT	0.690 ± 0.206	0.667 ± 0.052	0.711 ± 0.069	0.606 ± 0.052
	MDAgents-FT	0.774 ± 0.135	0.667 ± 0.052	0.678 ± 0.051	0.515 ± 0.052
	AI-Hospital-FT	0.548 ± 0.082	0.667 ± 0.052	0.833 ± 0.088	0.647 ± 0.105
	CoMMa	$\mathbf{0.905 \pm 0.083}$	$\mathbf{0.697 \pm 0.105}$	$\mathbf{0.911 \pm 0.077}$	$\mathbf{0.666 \pm 0.139}$

A.5 Variance and AUPRC Analysis on MTBBench.

MTBBench exhibits larger standard deviations than the other benchmarks. This mainly reflects its limited effective sample size: the progression and recurrence endpoints contain relatively few samples after the train-test split (Section A.1). In contrast, LungCure, with 999 patients, yields smaller standard deviations. All methods are evaluated under the same MTBBench split and protocol,

so the comparison remains controlled; however, we caution against over-interpreting MTBBench differences where variance is large. As a complementary view under limited and imbalanced samples, Table 11 reports the area under the precision-recall curve (AUPRC).

Table 11: AUPRC on MTBBench. Results are mean ± standard deviation over 3 random seeds. Baselines are the prompt-only versions reported in Table 1; fine-tuned comparisons are given in Table 10.

Endpoint	Method	Llama3.1-70B	Qwen2.5-72B	Gemma4-31B
Progression	Single-Agent	0.464 ± 0.000	0.439 ± 0.000	0.503 ± 0.000
	MDAgents	0.454 ± 0.022	0.516 ± 0.037	0.520 ± 0.008
	MAC	0.478 ± 0.027	0.468 ± 0.008	0.520 ± 0.009
	AI-Hospital	0.490 ± 0.029	0.463 ± 0.010	0.506 ± 0.013
	CoMMa	$\mathbf{0.777 \pm 0.202}$	$\mathbf{0.761 \pm 0.047}$	$\mathbf{0.866 \pm 0.139}$
Recurrence	Single-Agent	0.526 ± 0.000	0.549 ± 0.000	0.584 ± 0.000
	MDAgents	0.545 ± 0.046	0.560 ± 0.016	0.598 ± 0.012
	MAC	0.552 ± 0.030	0.576 ± 0.019	0.587 ± 0.012
	AI-Hospital	0.582 ± 0.010	0.566 ± 0.022	0.616 ± 0.008
	CoMMa	$\mathbf{0.815 \pm 0.165}$	$\mathbf{0.914 \pm 0.055}$	$\mathbf{0.873 \pm 0.074}$

A.6 Additional Response Comparison.

A.6.1 HCC Tumorboard

We further provide agentic framework response comparison on HCC Tumorboard dataset in Figure 6 and Figure 7. Whereas single-agent in-context learning and role-based multi-agent methods often yield verbose yet clinically inconsistent rationales, CoMMa provides quantitative attribution of agent contributions within a structured decision framework, resulting in superior predictive performance.

A.6.2 MTBBench

We provide a comparison of agentic framework responses on the MTBBench dataset. Specifically, we present a representative sample for the tumor progression tasks in Figure 8 and for the tumor recurrence prediction tasks in Figure 9. We observe that despite the extensive reasoning produced by both single-agent in-context learning and role-based multi-agent frameworks, their predictive accuracy remains significantly inferior to our proposed method.

Clinical Data	GT	(a) Single-Agent In-context Learning	(b) Role-based Multi-Agents	(c) CoMMa (Ours)
Question: You are an expert Oncologist AI assistant. Analyze the following **<Clinical Data>** for a patient and recommend the most appropriate therapeutic recommendations from the provided **<Options>**.				

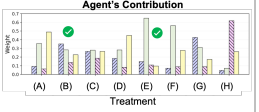
Image-specific data
- Number of tumors:1
- Size of Largest Tumor (CM):2.0
- Location Largest (Segment):8
  - Segment 1:0
  - Segment 2:0
  - Segment 3:0
  - Segment 4A:0
  - Segment 4B:0
  - Segment 5:0
  - Segment 6:0
  - Segment 7:0
  - Segment 8:1
- Size second (CM):-
- Size third (CM):-
- Vascular Invasion:0
- Extra hepatic Spread:0
- Thrombus (Portal Vein):0
…
Transplant Network-specific data
- Number of OPTN 5A Tumors:0
- Number of OPTN 5B Tumors:0
- Number of OPTN 5X Tumors:0
- Number of OPTN 5T Tumors:0
Performance-specific data
- Child Pugh Score:-
- MELD Score:-
- ECOG Score: 0
Laboratory-specific data
- Encephalopathy:0
- Ascites:0
- Bilirubin:0.3
- Albumin:4.0
- INR:1.1
- AFP:2.2
…
Options
(A) TACE
(B) Ablation
(C) Radioembolization
(D) Radiotherapy
(E) Resection
(F) Transplant
(G) Chemo
(H) Palliative
	(E) Resection	**<Prediction>** ['(B) Ablation, (E) Resection']
		
		Reasoning:
		- The patient has a solitary HCC (Number of tumors: 1), size 2.0 cm, located in segment 8.
		- No vascular invasion, no extrahepatic spread, no portal vein thrombus.
		- Excellent performance status (ECOG 0), no encephalopathy, no ascites.
		- Liver function is well preserved (Bilirubin 0.3, Albumin 4.0, INR 1.1).
		- AFP is low (2.2).
		
		In-context learning from dataset:
		- Cases 83 and 11 have nearly identical features (solitary 2.0 cm tumor, segment 8, no invasion, preserved liver function, ECOG 0-1), both recommended for ablation and/or resection.
		- Current guidelines (AASLD, EASL) recommend ablation or resection for solitary tumors <3 cm, good performance status, and well-preserved liver function.
		
		Summary: Both percutaneous ablation and surgical resection are appropriate and potentially curative options. The choice depends on technical feasibility, patient preference, and comorbidity.
			Case Summary:
			- **Number of tumors:** 1
			- **Size of largest tumor:** 2.0 cm
			- **Location:** Segment 8
			- **No vascular invasion, extrahepatic spread, or portal vein thrombus**
			- **Performance status (ECOG):** 0 (fully active)
			- **No ascites or encephalopathy**
			- **Liver function:** Bilirubin 0.3, Albumin 4.0, INR 1.1 (all normal)
			- **AFP:** 2.2 (not elevated)
			- **Child-Pugh and MELD scores not provided, but clinical and lab data suggest well-compensated liver function (likely Child-Pugh A)**
			- **No evidence of advanced disease or poor functional status**
			
			Interpretation:
			This patient has a single, small hepatocellular carcinoma (HCC) in segment 8, with preserved liver function and excellent performance status. There is no evidence of vascular invasion, extrahepatic spread, or portal vein thrombus.
			
			Treatment Decision:
			Curative options are preferred. According to international guidelines (AASLD, EASL, BCLC), for a single HCC \u22642 cm in a patient with well-compensated liver function and good performance status, the following are considered:
			
			1. **Surgical Resection (E):**
			  - Indicated for solitary tumors in patients with preserved liver function and no portal hypertension.
			  - This patient fits the criteria: single tumor, normal labs, no ascites, no encephalopathy, ECOG 0.
			  - **Resection is the preferred option.**
			
			2. **Ablation (B):**
			  - For very early HCC (\u22642 cm), ablation (e.g., radiofrequency or microwave) is an alternative if resection is not feasible due to comorbidities or technical reasons.
			  - In this case, ablation is a valid alternative, especially if surgery is contraindicated.
			
			3. **Transplant (F):**
			  - Reserved for patients with multifocal disease, poor liver function, or when resection/ablation is not possible.
			  - This patient does not require transplant at this stage.
			
			4. **Other options (G, C, H, A, D):**
			  - Chemo (G), radioembolization (C), arterial therapy (A), radiotherapy (D), and palliative care (H) are not first-line for a single, small HCC in a patient with good liver function and performance status.
			
			Final Recommendation:
			Primary Recommendation:
			(E) Resection : Surgical resection is the treatment of choice for this patient.
			Alternative (if resection is not feasible):
			(B) Ablation : Percutaneous ablation is an excellent alternative.
			Other options are not indicated at this stage.
			Summary Table:
			| Option | Indication in this case | Recommendation |
			|---|---|---|
			| (E) Resection | Yes | First choice |
			| (B) Ablation | Yes (if not surgical) | Alternative |
			| (F) Transplant | No | Not indicated |
			| (G) Chemo | No | Not indicated |
			| (C) Radioembolization | No | Not indicated |
			| (A) TACE | No | Not indicated |
			| (D) Radiotherapy | No | Not indicated |
			| (H) Palliative | No | Not indicated |
			
			Conclusion:
			**Surgical resection (E) is recommended. If not feasible, ablation (B) is the alternative.
				Result: (Logit / Cutoff)
				(E) Resection (86% / 79%)
				(B) Ablation (85% / 64%)
				
				Agent's Contribution
				
				
				1. Image-specific Agent
				2. Transplant Network-specific Agent
				3. Performance-specific Agent
				4. Laboratory-specific Agent

Figure 6: Additional agentic framework response comparison on HCC Tumorboard dataset - #1. (a) Single-agent in-context learning (GPT-4.1), (b) Role-based multi-agent framework (MDAgents-Advanced), and (c) CoMMa. For the role-based multi-agent, intermediate discussions are omitted, and only the final consensus is visualized. For CoMMa, class-wise logits and cutoff thresholds are transformed into percentiles.

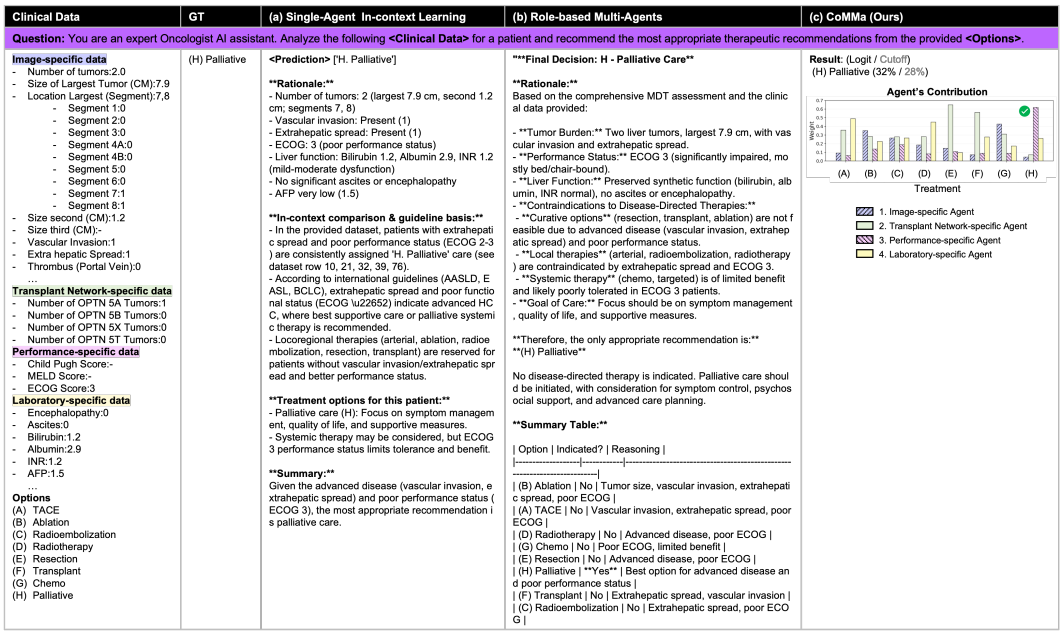


Figure 7: Additional agentic framework response comparison on HCC Tumorboard dataset - #2. (a) Single-agent in-context learning (GPT-4.1), (b) Role-based multi-agent framework (MDAgents-Advanced), and (c) CoMMa. For the role-based multi-agent, intermediate discussions are omitted, and only the final consensus is visualized. For CoMMa, class-wise logits and cutoff thresholds are transformed into percentiles.

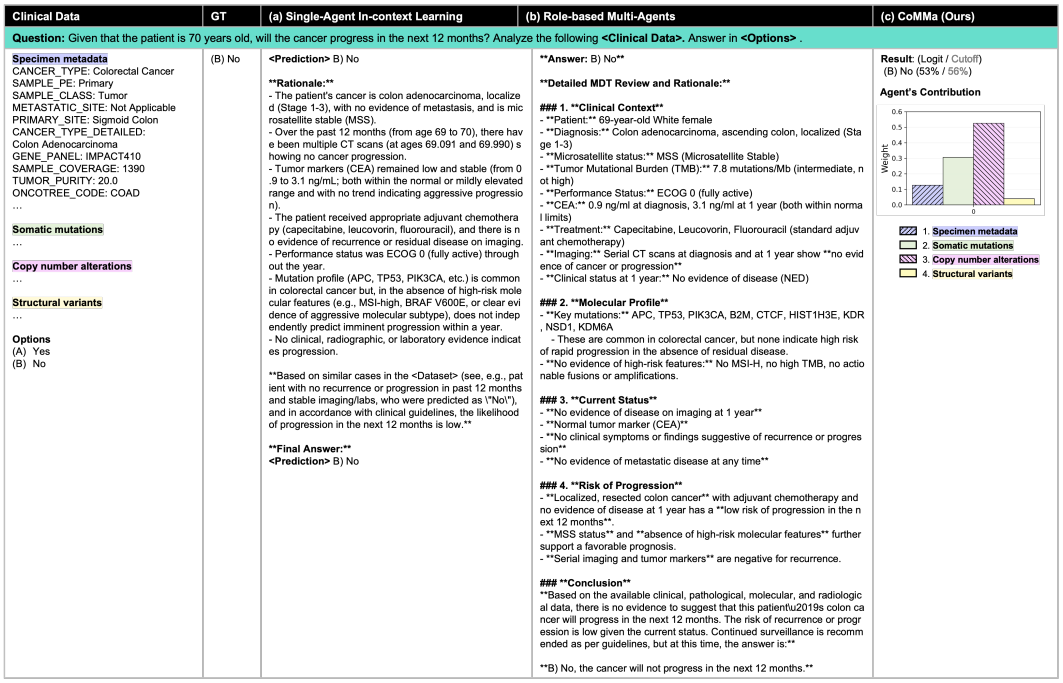


Figure 8: Additional agentic framework response comparison on MTBBench - tumor progression question. (a) Single-agent in-context learning (GPT-4.1), (b) Role-based multi-agent framework (MDAgents-Advanced), and (c) CoMMa. For the role-based multi-agent, intermediate discussions are omitted, and only the final consensus is visualized. For CoMMa, class-wise logits and cutoff thresholds are transformed into percentiles.

Clinical Data	GT	(a) Single-Agent In-context Learning	(b) Role-based Multi-Agents	(c) CoMMa (Ours)
Question: Given that the patient is 54.5 years old, will the cancer have a recurrence in the next 12 months? Analyze the following **<Clinical Data>.** Answer in **<Options>** .				

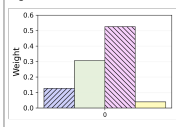
Specimen metadata
CANCER_TYPE: Colorectal Cancer
SAMPLE_PE: Primary
SAMPLE_CLASS: Tumor
METASTATIC_SITE: Not Applicable
PRIMARY_SITE: Sigmoid Colon
CANCER_TYPE_DETAILED: Colon Adenocarcinoma
GENE_PANEL: IMPACT410
SAMPLE_COVERAGE: 1390
TUMOR_PURITY: 20.0
ONCOTREE_CODE: COAD
…
 	
Somatic mutations
…
 	
Copy number alterations
…
 	
Structural variants
…
 	
Options
(A) Yes
(B) No
	(A) Yes	**<Prediction>** (B) No
		
		Reasoning:
		- The patient is a 53-year-old White female with a history of stage IV (distant metastatic, AJCC IV) colon adenocarcinoma of the sigmoid colon.
		- She underwent systemic chemotherapy (oxaliplatin, leucovorin, fluorouracil) and has had frequent imaging and CEA monitoring.
		- Imaging (CT and MR) and CEA labs from age 53.5 through 54.5 consistently show:
		  - No evidence of cancer recurrence or progression (multiple CT scans state \"cancer has NOT progressed\" and \"HAS CANCER: No\").
		  - CEA levels are low and stable (all \u22641.8 ng/ml).
		  - ECOG performance status remains 0 or 1.
		- Most recent imaging at age 54.495 (close to present age 54.5) shows no progression or presence of cancer.
		- She is microsatellite stable (MSS) and has a moderate tumor mutation burden (TMB).
		- There is no documentation of cancer recurrence or progression in the last 12 months.
		- Open-source clinical guidelines indicate that, while stage IV colorectal cancer carries a high risk of recurrence, stable imaging and tumor markers over a full year suggest the risk of recurrence in the next 12 months is relatively low\u2014especially with ongoing surveillance and no current evidence of disease.
		
		Therefore, based on the comprehensive clinical data and absence of progression or new lesions in the past year, the best-supported prediction is:
		
		B) No, the cancer will NOT have a recurrence in the next 12 months.
			Answer: B) No
			
			Rationale and Detailed Review:
			
			### 1. **Clinical Context**
			- **Patient:** 54.5-year-old White female, non-Hispanic
			- **Diagnosis:** Colon adenocarcinoma, sigmoid colon, AJCC Stage IV (distant disease)
			- **Treatment:** Received standard chemotherapy (Oxaliplatin, Leucovorin, Fluorouracil)
			- **Performance Status:** ECOG 0-1 throughout, indicating good functional status
			- **Tumor Markers:** CEA consistently low (1.1\u20131.8 ng/ml), CA19-9 mildly elevated once (46 U/ml)
			- **Imaging:** Serial CT and MR scans show **no evidence of cancer progression or recurrence** at multiple time points up to age 54.495.
			
			### 2. **Molecular and Pathological Data**
			- **MSI Status:** Microsatellite stable (MSS), MSI score 0.21 (low risk for hypermutated phenotype)
			- **Tumor Mutational Burden (TMB):** 3.9 mutations/Mb (low-intermediate, not predictive of high recurrence risk)
			- **Key Mutations:**
			  - **TP53 (missense):** Common in colorectal cancer, but not specifically predictive of recurrence in isolation.
			  - **FANCA (missense):** DNA repair gene, but not a strong driver of recurrence risk.
			  - **CDK12 (nonsense):** Loss-of-function, may affect DNA repair, but not a high-risk marker for recurrence in MSS CRC.
			  - **CTNNB1 (splice site):** Wnt pathway, but not a strong recurrence predictor in this context.
			- **Copy Number Alterations:** Mild gains (ERBB2, CDK12, RARA, SUZ12, PRKAR1A), but no high-level amplifications or deletions associated with aggressive behavior.
			- **No structural variants detected.**
			
			### 3. **Clinical Course and Surveillance**
			- **No prior medications before MSK treatment.**
			- **Serial imaging (CT/MR) at ages 53.544, 53.947, 54.101, and 54.495:** All show **no evidence of active cancer** or progression.
			- **CEA levels remain stable and low, which is reassuring.**
			- **Performance status remains good (ECOG 0-1).**
			
			### 4. **Risk of Recurrence**
			- **Stage IV disease** does carry a risk of recurrence, but in this patient:
			  - **No radiological evidence of disease** for over 12 months.
			  - **No biochemical evidence of recurrence** (CEA stable and low).
			  - **No high-risk molecular features** (MSS, low TMB, no aggressive driver mutations).
			  - **No clinical deterioration.**
			
			### 5. **Summary Table**
			| Factor | Finding | Recurrence Risk |
			|---|---|---|
			| Imaging | No evidence of disease | Low |
			| Tumor markers | CEA stable, low | Low |
			| MSI status | Stable (MSS) | Low |
			| TMB | Low-intermediate | Low |
			| Key mutations | No high-risk drivers | Low |
			| Performance status | ECOG 0-1 | Low |
			| Clinical course | No progression, good response | Low |
			
			Conclusion:
			Based on the comprehensive review of clinical, pathological, molecular, and imaging data, **the risk of recurrence in the next 12 months is low**. There is **no current evidence of active disease** or progression, and all surveillance parameters are reassuring.
			
			Final Decision:
			B) No, the cancer is unlikely to recur in the next 12 months given the patient's current clinical, imaging, and molecular profile.
			
			Recommendations:
			- Continue regular surveillance (imaging, CEA) as per guidelines.
			- Monitor for new symptoms.
			- Maintain multidisciplinary follow-up.
				Result: (Logit / Cutoff)
				(A) Yes (49% / 47%)
				
				Agent's Contribution
				

Figure 9: Additional agentic framework response comparison on MTBBench - tumor recurrence question. (a) Single-agent in-context learning (GPT-4.1), (b) Role-based multi-agent framework (MDAgents-Advanced), and (c) CoMMa. For the role-based multi-agent, intermediate discussions are omitted, and only the final consensus is visualized. For CoMMa, class-wise logits and cutoff thresholds are transformed into percentiles.

A.7 Qualitative Worked Example on HCC Tumorboard.

To make the decision pathway of CoMMa concrete, we trace one de-identified held-out HCC Tumorboard patient (Gemma4-31B backbone) from evidence partitioning, through cross-agent collaborative reasoning (Eq. (2)), to contribution-aware aggregation (Eq. (3)). The tumor board recorded ablation and liver transplant as the recommended treatments for this patient.

Evidence partitioning. Table 12 lists the evidence observed by each specialist agent, following the stream definition in Table 4. Each agent receives only its own stream.

Table 12: Evidence observed by each specialist agent in the worked example.

	Specialist stream	Evidence shown to the agent
$\mathbf{x}_1$	MRI imaging features	1 tumor; no vascular invasion
$\mathbf{x}_2$	OPTN information	Number of OPTN 5A tumors $= 1$
$\mathbf{x}_3$	Performance score	ECOG performance score $= 0$; no encephalopathy
$\mathbf{x}_4$	Laboratory data	AFP $= 52.9$ ng/mL; albumin $= 2.6$ g/dL; INR $= 1.7$

Contribution-aware aggregation. For the transplant target k, Table 13 reports each agent's post-collaboration logit $\hat{h}_{i,k}$, its learned contribution weight $W_{i,k}$, and the resulting contribution $W_{i,k}\hat{h}_{i,k}$. Following Eq. (3), the final transplant logit is

$$\hat{y}_k = 0.226 \times 1.77 + 0.228 \times 2.43 + 0.287 \times 0.48 + 0.259 \times (-0.35) = 1.001,$$

which gives a predicted probability of $\sigma(1.001) = 0.731$.

Table 13: Contribution-aware aggregation for the transplant target in the worked example.

Agent	$\hat{h}_{i,k}$	$W_{i,k}$	$W_{i,k}\hat{h}_{i,k}$
MRI imaging features	$+1.77$	0.226	$+0.400$
OPTN information	$+2.43$	0.228	$+0.554$
Performance score	$+0.48$	0.287	$+0.138$
Laboratory data	-0.35	0.259	-0.091
Total		1.000	$+1.001$

Final treatment scores. Table 14 reports the predicted probability of every treatment option. Transplant receives the highest score, and both treatments recommended by the tumor board, transplant and ablation, exceed the 0.5 decision threshold, while all other options fall below it. This example shows how each final recommendation can be audited as an explicit weighted sum of agent-level contributions.

Table 14: Predicted probabilities for all treatment options in the worked example.

Treatment option	Predicted probability	Tumor-board recommended
(F) Transplant	0.731	Yes
(B) Ablation	0.534	Yes
(D) Radiotherapy	0.328	No
(A) TACE	0.109	No
(E) Surgical resection	0.070	No
(G) Chemotherapy	0.035	No
(C) Radioembolization	0.018	No
(H) Palliative care	0.010	No

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction describe the proposed CoMMa framework, its multi-agent collaborative reasoning mechanism, and the experimental settings used for evaluation. The stated claims are supported by the methodology and empirical results, and we avoid claiming generality beyond the evaluated tasks and model backbones.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a discussion of limitations in the final paragraph.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides detailed descriptions of the experimental setup, including model configurations, agent design, and evaluation protocols, which are sufficient to reproduce the main results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The paper uses two private clinical datasets that cannot be publicly released due to data-use restrictions and privacy considerations. To support reproducibility, we provide detailed experimental protocols, model configurations, and evaluation metrics in the paper and supplementary material.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The paper specifies the experimental settings in sufficient detail, including datasets, model backbones, agent design, hyperparameters, and evaluation protocols, enabling clear understanding of the results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Main results report mean $\pm$ standard deviation over three random seeds that vary model initialization under a fixed train-test split.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provides information about the compute setup, including the use of LLM backbones and inference settings, which is sufficient to understand the computational requirements of the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research adheres to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses both positive and negative societal impacts.

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release new models or datasets that pose a high risk for misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper properly cites and credits the original sources of all datasets and model backbones used.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new datasets, models, or other assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or direct research interactions with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The study was conducted under IRB approval, and the paper describes the approval status and relevant protections for the clinical data used in the study.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The paper explicitly describes the use of LLMs as part of the proposed method, including the model backbones and their role in multi-agent reasoning.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.