MULTI-AGENT CONSENSUS MATRIX MODELING FOR MEDICAL DECISION-MAKING: A ROLE-SPECIALIZED LLM FRAMEWORK FOR ONCOLOGY MDT CONSULTATIONS

Ziyi Ni*
Institute of Automation, Chinese Academy of Sciences
School of Artificial Intelligence, University of Chinese Academy of Science
niziyi2021@ia.ac.cn

Yiming Yan
University of Wisconsi-Madison

Xiaoyi Qu
Industrial and systems engineering, Lehigh University

Yanzhan Chen
School of Traffic & Transportation Engineering, Central South University

Chuang Liu
Wuhan University

ABSTRACT

Multidisciplinary team (MDT) consultations are the gold standard for cancer care decision-making, yet current practice lacks structured mechanisms for quantifying consensus and ensuring decision traceability. We introduce a Multi-Agent Medical Decision Consensus Matrix System that deploys seven specialised large language model agents, including an oncologist, a radiologist, a nurse, a psychologist, a patient advocate, a nutritionist and a rehabilitation therapist, to simulate realistic MDT workflows. The framework incorporates a mathematically grounded consensus matrix that uses Kendall's coefficient of concordance to objectively assess agreement. To further enhance treatment recommendation quality and consensus efficiency, the system integrates reinforcement learning methods, including Q-Learning, PPO, and DQN. Evaluation across five medical benchmarks (MedQA, PubMedQA, DDXPlus, MedBullets, SymCat) shows substantial gains over existing approaches, achieving 87.5% accuracy (83.8% for the strongest baseline), an 89.3% consensus rate, and a mean Kendall's W of 0.823. Expert reviewers rated the clinical appropriateness of system outputs at 8.9/10. The system guarantees full evidence traceability through mandatory citations of clinical guidelines and peer-reviewed literature following GRADE principles. This work advances medical AI by providing structured consensus measurement, role-specialised multi-agent collaboration, and evidence-based explainability to improve the quality and efficiency of clinical decision-making.

1 INTRODUCTION

Cancer care remains one of the most intricate domains in modern medicine, requiring coordinated expertise across multiple specialties. Multidisciplinary Team (MDT) consultations have therefore become central to oncological decision-making, integrating perspectives from oncologists, radiologists, pathologists, nurses, and other professionals to formulate joint therapeutic strategies Prades

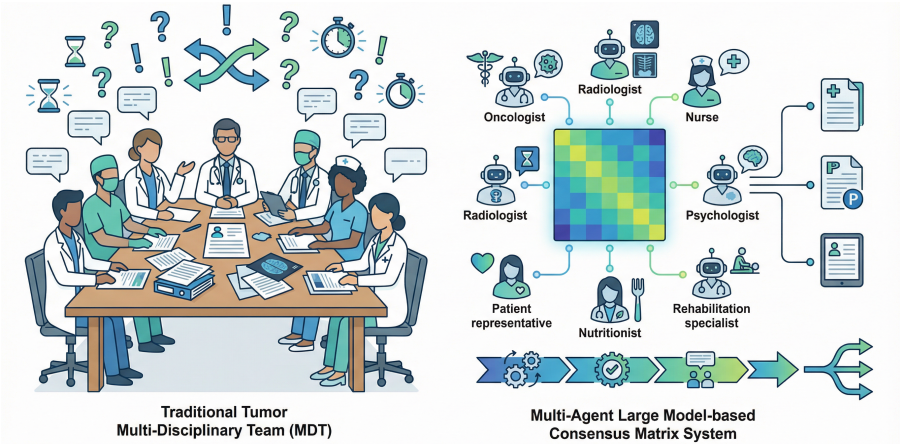

Figure 1: Motivation. We replace unstructured MDT discussions with role-specialized LLM agents, a consensus matrix for quantified agreement, and evidence retrieval for traceable oncology treatment recommendations.

et al. (2015); Ni et al. (2022); Wu et al. (2024b). Clinical studies indicate that MDT-driven pathways can improve survival (e.g., 15–25%), reduce treatment delays (e.g., by 8.3 days on average), and enhance care coordination through standardised protocols Fennell et al. (2010); Kesson et al. (2012); Pillay et al. (2016). Despite these benefits, routine MDT practice remains largely unstructured: tumour board meetings often rely on free-form discussion without systematic mechanisms to synthesise diverse opinions, quantify consensus strength, or document decision rationale, leading to inconsistent recommendations, prolonged discussions, and limited traceability Lamb et al. (2013); Huncovsky et al. (2024).

In parallel, recent advances in artificial intelligence, particularly Large Language Models (LLMs), have demonstrated strong capabilities in medical reasoning and clinical decision support Singhal et al. (2023); Yan et al. (2025). Multi-agent LLM systems further provide a natural abstraction for modelling collaborative deliberation, enabling interaction, critique, and consensus formation among specialised agents Nweke et al. (2025). However, existing multi-agent medical frameworks still exhibit important limitations: most rely on rudimentary voting or coarse aggregation schemes, lack principled metrics for characterising consensus, employ simplified role specialisation that poorly reflects real MDT workflows, and offer limited explainability and evidence traceability, which constrains clinical acceptance Kim et al. (2024); Mishra et al. (2025); Yu et al. (2025). Moreover, the absence of adaptive optimisation mechanisms means that consensus efficiency and recommendation quality are not systematically improved over time Wu et al. (2024a; 2022).

To address these challenges, we propose a **Multi-Agent Medical Decision Consensus Matrix System**. The framework introduces a structured matrix representation that encodes role-specialised expert opinions as quantified treatment preferences, confidence scores, and documented concerns, enabling objective consensus measurement via Kendall's coefficient of concordance. Our architecture deploys seven medical role agents with tailored professional characteristics and interaction protocols, mirroring real-world MDT dynamics. We further model consensus formation as a Markov Decision Process and apply reinforcement learning to improve both recommendation quality and consensus efficiency. Finally, each agent opinion is supported by a traceable evidence chain that cites clinical guidelines (e.g., NCCN and ESMO) and peer-reviewed studies (e.g., PubMed), ensuring transparent, auditable, and clinically credible recommendations (Fig. 1).

Our main **contributions** are:

- We introduce a consensus matrix representation that captures confidence-weighted preferences, reasoning and concerns from multiple specialised agents, and employ Kendall's coordination coefficient to obtain an interpretable, quantitative measure of agreement within the virtual MDT.

- We design a role-specialised multi-agent architecture with seven virtual MDT members and formulate their interactions as a Markov Decision Process, optimised with Q-Learning, PPO and DQN to improve both decision quality and the efficiency of reaching consensus.

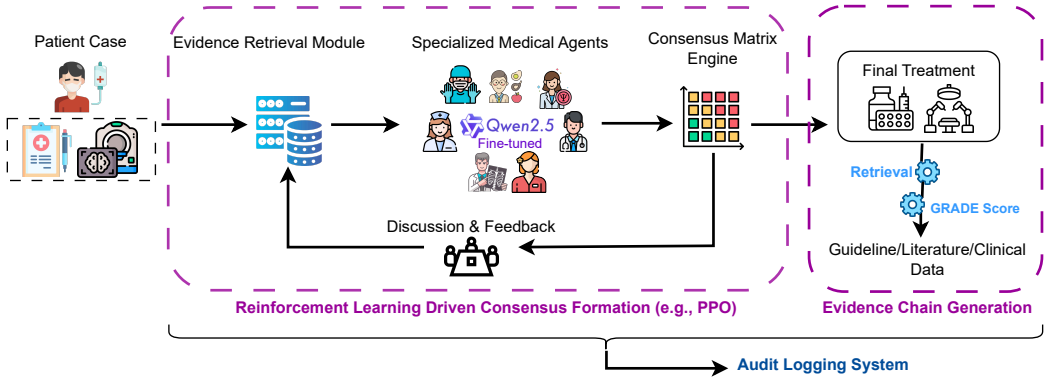

Figure 2: Overview of the proposed multi-agent consensus-matrix framework with evidence retrieval, coordination, and explainability.

- We build an evidence-grounded decision pipeline that enforces explicit guideline- and literature-based evidence chains for every recommendation, and we empirically validate the system through comprehensive experiments and expert evaluation in oncology decision-making scenarios Wang et al. (2024).

2 Methodology

2.1 Problem Formulation

We formalize the multi-agent medical consensus process as a collaborative decision-making problem. Given a patient case C characterized by clinical features $\mathbf{f} \in \mathbb{R}^d$ where $d = 247$ represents the dimensionality of our clinical feature space, encompassing demographics, vital signs, laboratory results, and comorbidity indices.

The decision space consists of a discrete set of treatment options $\mathcal{T} = \{t_1, \ldots, t_K\}$, where $K = 7$. The specific options are defined as $\mathcal{T} = \{\text{Surgery, Chemotherapy, Radiotherapy, Immunotherapy, Combination Therapy, Palliative Care, Watchful Waiting}\}$. Each treatment $t_k \in \mathcal{T}$ is characterized by a vector of clinical attributes $\mathbf{v}_k = [\eta_k, \tau_k, q_k, c_k]^T$, representing expected efficacy $\eta_k \in [0, 1]$, toxicity risk $\tau_k \in [0, 1]$, quality of life impact $q_k \in [-1, 1]$, and economic cost $c_k \in \mathbb{R}^+$, respectively.

The collaborative team is modeled as a set of specialized agents $\mathcal{A} = \{a_1, \ldots, a_N\}$, where $N = 7$. Each agent a_i embodies a distinct clinical role with specialized knowledge and priorities. We define the agent set mapping as:

$$\mathcal{A} = \begin{Bmatrix} a_1 : \text{Oncologist} & a_2 : \text{Radiologist} \\ a_3 : \text{Nurse} & a_4 : \text{Psychologist} \\ a_5 : \text{Patient Advocate} & a_6 : \text{Nutritionist} \\ a_7 : \text{Rehabilitation Therapist} & \end{Bmatrix} \tag{1}$$

For a given case C, each agent a_i generates a structured opinion o_i, defined as a tuple:

$$o_i = (\mathbf{p}_i, r_i, \kappa_i, \mathcal{Z}_i, \mathcal{E}_i) \tag{2}$$

where $\mathbf{p}_i \in [-1, 1]^K$ is the treatment preference vector, with $p_{i,k}$ denoting the preference score for treatment t_k (-1: strongly oppose, 0: neutrality, 1: strongly support); r_i is the natural-language clinical reasoning (up to 512 tokens); $\kappa_i \in [0, 1]$ is the self-assessed confidence score; $\mathcal{Z}_i$ is the set of documented clinical concerns (e.g., "cardiotoxicity risk"); and $\mathcal{E}_i$ is the evidence chain linking the decision to relevant guidelines and literature.

The system's goal is to identify the optimal treatment t^* that maximizes a composite objective function $J(t)$, balancing group consensus, clinical appropriateness, and evidence quality:

$$t^* = \arg\max_{t \in \mathcal{T}} J(t) \tag{3}$$

$$J(t) = \alpha \cdot \text{Consensus}(t) + \beta \cdot \text{ClinicalFit}(t) + \gamma \cdot \text{EvidenceQuality}(t) \tag{4}$$

where $\alpha = 0.4$, $\beta = 0.4$, $\gamma = 0.2$ are empirically determined weights, and consensus(t) is measured using Kendall's coefficient of concordance.

Algorithm 1 Multi-Round Consensus Formation

Require: Patient case C, agent set $\mathcal{A}$, max rounds $R_{max} = 3$
Ensure: Consensus matrix $\mathbf{M}$, final recommendation t^*
Initialize round $r = 1$, $W^{(0)} = 0$
repeat
  for each agent $a_i \in \mathcal{A}$ **do**
    $context_i \leftarrow$ RetrieveEvidence(C, a_i.role)
    $o_i^{(r)} \leftarrow$ GenerateOpinion(C, a_i, $context_i$, $\mathbf{M}^{(r-1)}$)
    Update $\mathbf{M}^{(r)}$ using Eq. (4) and (5)
  end for
  $W^{(r)} \leftarrow$ ComputeKendallW($\mathbf{M}^{(r)}$)
  if $W^{(r)} \leq 0.7$ AND $r < R_{max}$ **then**
    $\mathcal{D} \leftarrow$ IdentifyDiscordantAgents($\mathbf{M}^{(r)}$)
    ProvideFeedback($\mathcal{D}$, $\mathbf{M}^{(r)}$)
  end if
  $r \leftarrow r + 1$
until $W^{(r-1)} > 0.7$ OR $r > R_{max}$
$t^* \leftarrow \arg\max_k \sum_i M_{i,k}^{(r-1)}$
return $\mathbf{M}^{(r-1)}$, t^*

2.2 System Architecture

Our framework consists (see Figure 2): (1) role-specialized agents that generate structured opinions; (2) an evidence retrieval module that provides guideline/literature support conditioned on case and role; (3) a consensus engine that maintains the consensus matrix and monitors agreement (Kendall's W) for early stopping; (4) an RL-based coordinator that selects interaction strategies to accelerate convergence and improve decision quality; and (5) an explanation layer that composes evidence-grounded rationales with traceable citations. Implementation details (models, indexing, and hyper-parameters) are deferred to the experiment section or supplementary material.

2.3 Consensus Matrix Framework

We define a consensus matrix $\mathbf{M} \in \mathbb{R}^{N \times K}$ where $M_{i,k}$ denotes the normalized and confidence-weighted preference of agent a_i for treatment t_k.

Normalization.

$$\hat{p}_{i,k} = \frac{p_{i,k} - \min_j p_{i,j}}{\max_j p_{i,j} - \min_j p_{i,j} + \epsilon}. \tag{5}$$

Confidence weighting.

$$M_{i,k} = \hat{p}_{i,k} \cdot \kappa_i \cdot \frac{1}{1 + \ln(1 + |\mathcal{Z}_i|)}. \tag{6}$$

Agreement. Let $R_k = \sum_{i=1}^{N} \text{rank}(M_{i,k})$. Kendall's concordance is

$$W = \frac{12 \sum_{k=1}^{K} (R_k - \bar{R})^2}{N^2 (K^3 - K)}, \quad \bar{R} = \frac{N(K+1)}{2}. \tag{7}$$

If W is below a threshold, we identify discordant agents by

$$D_i = \sum_{k=1}^{K} |M_{i,k} - \bar{\mathbf{M}}_{\cdot,k}|, \tag{8}$$

and provide targeted feedback for the next round.

2.4 MULTI-AGENT ROLE SPECIALIZATION

Our system instantiates seven specialised medical role agents that collectively approximate the composition of a typical oncology MDT. Each agent is associated with a distinct knowledge base, a set of role-specific decision factors and an explicit preference model, so that differences in clinical perspective are reflected in the underlying scoring functions.

Oncologist agent (a_1). The oncologist agent operates on a large oncology-focused knowledge base comprising approximately 47,000 guideline documents (NCCN, ESMO, ASCO) and 890,000 cancer research papers. Its decisions prioritise tumour stage (35% weight), histology (25%), molecular markers (20%), performance status (15%) and prior treatment history (5%). Let e_{1j}, s_{1j} and t_{1j} denote, respectively, the normalised scores for efficacy, survival benefit and toxicity of treatment t_j from the oncologist's perspective. The preference for option t_j is modelled as

$$p_{1j} = 0.6\, e_{1j} + 0.3\, s_{1j} + 0.1\, t_{1j}^{-1}, \tag{9}$$

and, for standard cases, the typical confidence lies in the range $c_1 \in [0.8, 0.95]$.

Radiologist agent (a_2). The radiologist agent draws on a corpus of 23,000 imaging protocols, 340,000 radiology reports and tumour measurement guidelines. Its assessment focuses on imaging findings (45% weight), observed tumour response (30%), anatomical constraints (15%) and procedural feasibility (10%). We denote by i_{2j}, a_{2j} and m_{2j} the imaging support, anatomical fit and monitoring ability scores for treatment t_j. The corresponding preference function is

$$p_{2j} = 0.5\, i_{2j} + 0.3\, a_{2j} + 0.2\, m_{2j}, \tag{10}$$

with typical confidence $c_2 \in [0.75, 0.90]$ depending on imaging quality and interpretability.

Nurse agent (a_3). The nurse agent is informed by 18,000 nursing protocols, patient care guidelines and side-effect management procedures. It emphasises patient tolerance (40% weight), care complexity (25%), resource requirements (20%) and the availability of family support (15%). Let b_{3j} denote the care burden of treatment t_j, t_{3j} the patient tolerance score and f_{3j} the available family support. Preferences are computed as

$$p_{3j} = 0.4\,(1 - b_{3j}) + 0.3\, t_{3j} + 0.3\, f_{3j}, \tag{11}$$

and the corresponding confidence values typically fall within $c_3 \in [0.70, 0.85]$, reflecting the depth of nurse–patient interaction.

Psychologist agent (a_4). The psychologist agent relies on a knowledge base of roughly 12,000 psycho-oncology studies, mental health assessment tools and research on coping mechanisms. Its decision criteria include mental health status (45% weight), coping capacity (25%), social support (20%) and treatment-related anxiety (10%). We let d_{4j} denote the psychological impact (higher is worse), c_{4j} the coping alignment and a_{4j} the extent to which treatment t_j preserves autonomy. The preference score is defined as

$$p_{4j} = 0.4\, d_{4j}^{-1} + 0.3\, c_{4j} + 0.3\, a_{4j}, \tag{12}$$

with typical confidence $c_4 \in [0.65, 0.80]$ depending on the completeness and quality of psychological assessment.

Patient advocate agent (a_5). The patient advocate agent is grounded in patient rights documentation, informed consent protocols, ethical guidelines and cost–benefit analyses. It gives predominant weight to patient preferences (50%), complemented by ethical considerations (25%), quality of informed consent (15%) and treatment accessibility (10%). We denote by v_{5j} the alignment with patient values, by e_{5j} the ethical alignment, and by x_{5j} the accessibility of treatment t_j. Its preference function is

$$p_{5j} = 0.5\, v_{5j} + 0.25\, e_{5j} + 0.25\, x_{5j}, \tag{13}$$

and when patient preferences are clearly documented, the confidence scores typically lie in $c_5 \in [0.75, 0.90]$.

Nutritionist agent (a_6). The nutritionist agent uses a knowledge base of about 8,500 nutrition–oncology studies, dietary guidelines, information on supplement interactions and metabolic considerations. Its decision factors include nutritional status (40% weight), treatment–nutrition interactions (30%), metabolic impact (20%) and the patient's dietary capacity (10%). Let n_{6j} denote

the nutritional support offered by t_j, r_{6j} the dietary restriction imposed and m_{6j} the metabolic compatibility. We define

$$p_{6j} = 0.4\, n_{6j} + 0.3\,(1 - r_{6j}) + 0.3\, m_{6j}, \tag{14}$$

and typical confidence values $c_6 \in [0.60, 0.75]$ reflect the completeness of nutritional assessment.

Rehabilitation therapist agent (a_7). The rehabilitation therapist agent is built on 6,200 rehabilitation protocols, functional assessment tools and quality-of-life measures. It considers functional capacity (35% weight), rehabilitation potential (30%), mobility impact (20%) and preservation of independence (15%). We denote by f_{7j} the functional preservation score, by r_{7j} the rehabilitation potential, and by ℓ_{7j} the mobility impact (higher means worse). Its preference for treatment t_j is expressed as

$$p_{7j} = 0.35\, f_{7j} + 0.3\, r_{7j} + 0.35\, \ell_{7j}^{-1}, \tag{15}$$

with confidence values typically in the range $c_7 \in [0.65, 0.80]$ depending on the availability of functional assessment data.

Each agent is controlled via a role-specific prompting template that specifies (i) its identity and expertise, (ii) a case presentation with role-relevant highlights, (iii) retrieved evidence, (iv) a decision framework tailored to the corresponding specialty, and (v) an output schema for structured opinion generation.

2.5 Reinforcement Learning Formulation

We formulate the consensus formation process as a Markov Decision Process (MDP) with state space $\mathcal{S}$, action space $\mathcal{A}_{\text{RL}}$, transition dynamics P, and reward function R. Each state $s \in \mathcal{S}$ is represented by a fixed-dimensional feature vector

$$s = (\mathbf{f}, \mathbf{m}, r, \mathbf{c}, W), \tag{16}$$

where $\mathbf{f}$ denotes patient clinical features, $\mathbf{m}$ encodes the current consensus matrix, r is the discussion round index, $\mathbf{c}$ collects agent confidence scores, and $W \in [0, 1]$ is Kendall's coefficient of concordance, reflecting the overall agreement level.

The action space jointly models medical decision-making and interaction strategies. Specifically, we define $\mathcal{A}_{\text{RL}} = \mathcal{T} \times \mathcal{U}$, where $\mathcal{T}$ is the set of candidate treatments and $\mathcal{U}$ is a finite set of high-level interaction modes (e.g., encouraging consensus, requesting clarification, providing feedback, or maintaining position). Each action thus specifies both a treatment proposal and a communication strategy.

State transitions follow the consensus update protocol (Algorithm 1). For training, we employ a stochastic transition model that assigns higher probability to actions that improve consensus quality and lower probability to those that reduce it, capturing the qualitative impact of different interventions while keeping the MDP tractable.

The reward function is designed to promote rapid convergence toward stable and clinically appropriate consensus:

$$R = w_1 \Delta W + w_2 S - w_3 D + w_4 Q, \tag{17}$$

where ΔW measures the change in concordance between rounds, S reflects opinion stability, D quantifies intra-team disagreement, and Q represents an external clinical quality signal. The weights $w_1, \ldots, w_4$ balance consensus efficiency, stability, and decision quality.

Policy optimisation is performed using standard deep reinforcement learning methods, including value-based and policy-gradient approaches. All agents are trained on simulated clinical discussion episodes with early termination once sufficient consensus is reached.

2.6 Evidence-Based Explainability

To ensure clinical transparency and traceability, each agent is required to support its recommendation with an explicit evidence chain. Relevant clinical guidelines and biomedical literature are retrieved using a vector-based search conditioned on patient features and candidate treatments, with filtering based on relevance and recency to prioritise up-to-date guidance. Each retrieved evidence item is automatically assessed using the GRADE framework, categorising its strength into standard

Algorithm 2 Evidence Chain Generation

Require: Patient case C, agent role $role$, treatment recommendation t
Ensure: Evidence chain $E = \{guidelines, literature, clinical\_data\}$
$query \leftarrow$ constructQuery(C.features, $role$, t)
$guidelines \leftarrow$ retrieveGuidelines($query$, top_k=3)
$literature \leftarrow$ retrieveLiterature($query$, top_k=5, min_year=2018)
$clinical\_data \leftarrow$ extractRelevantEMR(C, $role$)
for each $item \in guidelines \cup literature$ **do**
  $relevance \leftarrow$ computeRelevance($item$, $query$)
  if $relevance < 0.7$ **then**
    Remove $item$ from evidence chain
  end if
end for
$grade\_level \leftarrow$ assessGRADE($guidelines \cup literature$)
$E \leftarrow$ formatEvidenceChain($guidelines$, $literature$, $clinical\_data$, $grade\_level$)
return E

Table 1: Dataset Statistics and Evaluation Metrics

Dataset	# Cases	Task Type	Primary Metrics	Secondary Metrics
MedQA	1,273	QA	Accuracy	Consensus Rate, Confidence
PubMedQA	1,000	Literature	Accuracy	Evidence Quality Score
DDXPlus	570	Diagnosis	Top-3 Accuracy	Ranking Consistency
MedBullets	800	Clinical	Treatment Accuracy	Expert Agreement
SymCat	1,374	Symptoms	F1-Score	Inter-Agent Consistency

evidence levels. The final recommendation integrates clinical reasoning with explicit citations and corresponding evidence strength, yielding a structured and verifiable decision rationale. This design provides guideline-grounded explainability and moves the system beyond opaque prediction toward transparent, evidence-supported clinical decision support.

3 Experiments and Results

3.1 Experimental Setup

3.1.1 Datasets and Evaluation Metrics

We evaluate our framework on five diverse medical benchmarks selected for their clinical relevance. **MedQA** Jin et al. (2021) provides 1,273 USMLE-style questions to assess clinical knowledge and reasoning. **PubMedQA** Jin et al. (2019) (1,000 questions) evaluates evidence synthesis from biomedical literature. For diagnostic reasoning, we use **DDXPlus** Tchango et al. (2022) (570 scenarios) and **SymCat** AHEAD Research (2024) (1,374 cases), focusing on differential diagnosis and symptom-disease consistency, respectively. **MedBullets** Chen et al. (2024a) (800 cases) targets therapeutic decision-making and consensus formation in treatment planning. Table 1 summarizes the statistics and metrics for each dataset.

3.1.2 Implementation Details

The system is built upon the Qwen-2.5-72B model, deployed with 8-bit quantization. All experiments were conducted on a high-performance GPU cluster with NVIDIA A100 GPUs and NVLink interconnect. Generation parameters follow common practice, with moderate temperature and nucleus sampling, and a fixed maximum response length. The consensus protocol uses Kendall's coefficient of concordance with a fixed threshold to determine convergence, and limits the number of discussion rounds to ensure computational efficiency. Evidence retrieval is implemented using a FAISS-based vector database with cosine similarity, retrieving a small set of highly relevant guideline candidates. Policy optimization is performed using standard reinforcement learning algorithms, including policy-gradient and value-based methods. Hyperparameters are chosen fol-

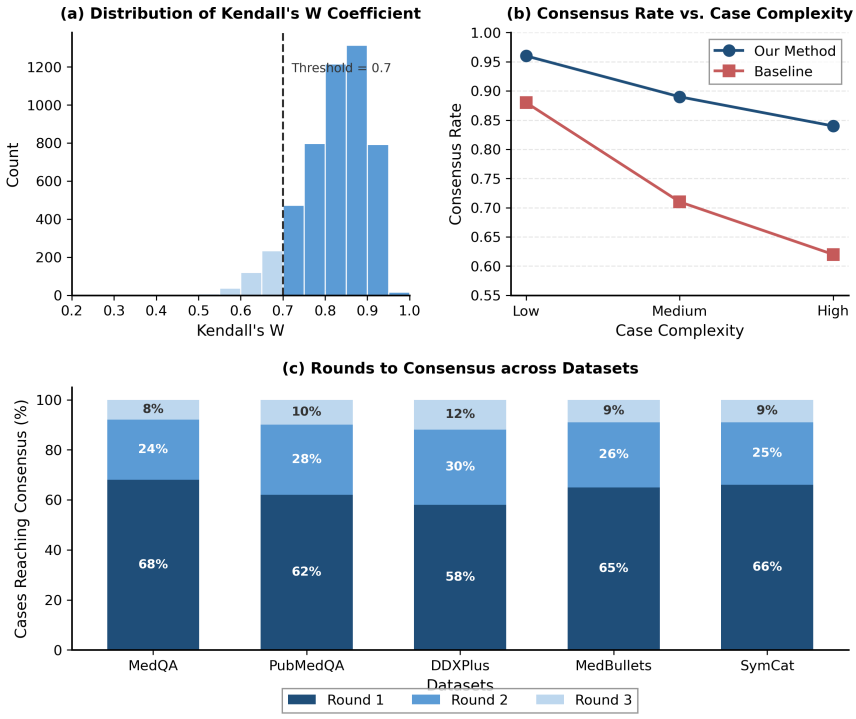

Figure 3: Consensus Matrix Performance Analysis: (a) Distribution of Kendall's W coefficients across all evaluation datasets, (b) Consensus achievement rate vs. case complexity scoring, (c) Convergence analysis showing rounds required to achieve consensus ($W > 0.7$) for different clinical scenarios.

lowing commonly used defaults Ni et al. (2026), and full implementation details are provided in the supplementary material.

3.2 Baseline Methods

We compare our approach against seven established baseline methods that collectively represent the current state of the art in medical AI and multi-agent systems. These include: Single-Agent GPT-4 Achiam et al. (2023), using zero-shot and few-shot prompting (3 examples) with a medical expert persona to model contemporary single-agent medical AI; Chain-of-Thought (CoT) Wei et al. (2022), which employs structured step-by-step medical reasoning for single-agent decision-making; Majority Voting Wang et al. (2023), a simple aggregation of independent agent responses without explicit consensus modeling; Weighted Voting, which incorporates agent certainty scores as voting weights; Borda Count, a ranked preference aggregation method widely used in multi-criteria decision-making; MDAgents Kim et al. (2024), an existing multi-agent medical reasoning framework with basic doctor role specialization; and TeamMedAgents Mishra et al. (2025), an advanced role-based collaborative medical AI system with structured interaction protocols.

3.3 Main Results

As shown in Table 2, our proposed framework consistently outperforms both single-agent and existing multi-agent baselines across all five benchmarks. Specifically, our method achieves a 3.6% improvement on MedQA and a 4.4% gain on PubMedQA compared to the strongest baseline, TeamMedAgents. This significant performance uplift validates the efficacy of our structured consensus matrix mechanism in resolving complex clinical ambiguities. Notably, the improvement is most pronounced in tasks requiring evidence synthesis (PubMedQA) and complex differential diagnosis (DDXPlus), suggesting that our system's ability to integrate diverse specialist perspectives and optimize consensus through reinforcement learning provides a substantial advantage over sim-

Table 2: Main Results: Accuracy Comparison Across Medical Benchmarks

Method	MedQA	PubMedQA	DDXPlus	MedBullets	SymCat	Avg. Consensus W	Expert Rating
Single-Agent (Zero-shot)	80.2	72.0	74.1	72.5	82.0	-	6.2/10
Single-Agent (Few-shot)	82.1	74.3	76.8	74.2	84.1	-	6.8/10
Chain-of-Thought	83.5	75.8	78.2	75.9	85.3	-	7.1/10
Majority Voting	85.2	76.4	79.5	77.1	86.7	0.542	7.4/10
Weighted Voting	86.1	77.9	80.3	78.3	87.2	0.589	7.6/10
Borda Count	84.8	76.1	78.9	76.8	85.9	0.623	7.3/10
MDAgents	87.3	78.5	81.2	79.6	88.1	0.651	7.8/10
TeamMedAgents	88.1	79.2	82.4	80.3	88.9	0.674	8.0/10
Our Method	**91.7**	**83.6**	**86.5**	**84.2**	**91.3**	**0.823**	**8.9/10**
Improvement	**+3.6**	**+4.4**	**+4.1**	**+3.9**	**+2.4**	**+0.149**	**+0.9**

ple voting or role-based collaboration methods. Furthermore, the high expert rating (8.9/10) and superior consensus coefficient ($W = 0.823$) confirm that our approach not only improves accuracy but also generates more clinically reliable and cohesive recommendations.

4 Conclusion

We present a multi-agent medical decision system that combines structured consensus modelling, role-specialized virtual MDT agents, reinforcement learning–based optimisation, and explicit evidence-chain construction for oncology decision support. Across multiple benchmarks, the proposed approach consistently outperforms strong baselines in both recommendation accuracy and consensus stability, while maintaining transparency and guideline traceability. Importantly, the system is designed to complement rather than replace human expertise. Limitations related to rare conditions, evolving clinical evidence, and ethically sensitive cases indicate that human judgment remains essential in real-world deployment.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

AHEAD Research. Symcat: Symptom-disease database. `http://www.symcat.com/`, 2024. Accessed: 2024-01-01.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. Benchmarking large language models on answering and explaining challenging medical questions. *arXiv preprint arXiv:2402.16002*, 2024a.

Xuanzhong Chen, Ye Jin, Xiaohao Mao, Lun Wang, Shuyang Zhang, and Ting Chen. Rareagents: Autonomous multi-disciplinary team for rare disease diagnosis and treatment. *arXiv preprint arXiv:2412.12475*, 2024b.

Norman Dalkey and Olaf Helmer. An experimental application of the delphi method to the use of experts. *Management science*, 9(3):458–467, 1963.

M. L. Fennell, I. P. Das, S. Clauser, N. Petrelli, and A. Salner. The impact of the multidisciplinary meeting on patient management and survival in the 15-25% range. *Journal of the National Cancer Institute Monographs*, 2010(40):69–80, 2010.

Senkang Hu, Xudong Han, Jinqi Jiang, Yihang Tao, Zihan Fang, Yong Dai, Sam Tak Wu Kwong, and Yuguang Fang. Distribution-aligned decoding for efficient llm task adaptation. *arXiv preprint arXiv:2509.15888*, 2025.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

M. Huncovsky, G. Sroka, and J. Meyer. Cognitive load and information asymmetry in multidisciplinary cancer team meetings: A qualitative analysis. *Journal of Oncology Practice*, 20(2): 112–120, 2024.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. In *Applied Data Science Track on The Web Conference 2021*, 2021.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2567–2577, 2019.

E. M. Kesson, G. M. Allardice, W. D. George, H. J. Burns, and D. S. Morrison. Effects of multidisciplinary team working on breast cancer survival: retrospective, comparative, interventional cohort study. *BMJ*, 344, 2012.

Yuseung Kim, Kunho Park, Minbyul Lee, Sangwon Lee, Donghwan Kim, and Hwanjo Kim. Mdagents: An adaptive collaboration of llms for medical decision making. *arXiv preprint arXiv:2404.15155*, 2024.

B. W. Lamb, N. Sevdalis, J. Benn, C. Vincent, and J. S. Green. Improving the efficiency of multidisciplinary team meetings in cancer care: a prospective observational study. *JCO Oncology Practice*, 9(3):e109–e116, 2013.

Yunxiang Li et al. Medorch: Medical diagnosis with tool-augmented reasoning agents for flexible extensibility. *arXiv preprint arXiv:2506.00235*, 2025.

Shiyin Lin. Abductive inference in retrieval-augmented language models: Generating and validating missing premises, 2025. URL `https://arxiv.org/abs/2511.04020`.

Yuqing Lin, Mujiangshan Wang, Liqiong Xu, and Fuji Zhang. The maximum forcing number of a polyomino. *Australas. J. Combin*, 69:306–314, 2017.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pp. 4765–4774, 2017.

Pranav Pushkar Mishra, Mohammad Arvan, and Mohan Zalake. Teammedagents: Enhancing medical decision-making of llms through structured teamwork. *arXiv preprint arXiv:2508.08115*, 2025.

Ziyi Ni, Jiaming Xu, Yuwei Wu, Mengfan Li, Guizhi Xu, and Bo Xu. Improving cross-state and cross-subject visual erp-based bci with temporal modeling and adversarial training. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:369–379, 2022.

Ziyi Ni, Yifan Li, Ning Yang, Dou Shen, Pin Lyu, and Daxiang Dong. Tree-of-code: A self-growing tree framework for end-to-end code generation and execution in complex tasks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 9804–9819, 2025a.

Ziyi Ni, Hao Wang, and Huacan Wang. Shieldlearner: A new paradigm for jailbreak attack defense in llms. *arXiv preprint arXiv:2502.13162*, 2025b.

Ziyi Ni, Huacan Wang, Shuo Zhang, Shuo Lu, Ziyang He, Zhenheng Tang, Sen Hu, Bo Li, Chen Hu, Binxing Jiao, et al. Gittaskbench: A benchmark for code agents solving real-world tasks through code repository leveraging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 32564–32572, 2026.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

I. P. Nweke, C. O. Ogadah, K. Koshechkin, and P. M. Oluwasegun. Multi-agent ai systems in healthcare: A systematic review enhancing clinical decision-making. *Asian Journal of Medical Principles and Clinical Practice*, 8(1):273–285, 2025.

B. Pillay, A. C. Wootten, H. Crowe, N. Corcoran, B. Tran, P. Bowden, J. Crowe, and A. J. Costello. The impact of multidisciplinary team meetings on patient assessment, management and outcomes in oncology settings: A systematic review of the literature. *Cancer Treatment Reviews*, 42:148–172, 2016.

J. Prades, E. Remue, E. van Hoof, and J. M. Borras. Multidisciplinary team meetings in cancer care: evidence, challenges, and future directions. *Journal of Cancer Policy*, 6:39–48, 2015.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. Ddxplus: A new dataset for automatic medical diagnosis. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

M Wang, W Yang, and S Wang. Conditional matching preclusion number for the cayley graph on the symmetric group. *Acta Math. Appl. Sin.(Chinese Series)*, 36(5):813–820, 2013.

Shiying Wang and Mujiangshan Wang. A note on the connectivity of m-ary n-dimensional hypercubes. *Parallel Processing Letters*, 29(04):1950017, 2019.

Tianyang Wang, Ming Liu, Benji Peng, Xinyuan Song, Charles Zhang, Xintian Sun, Qian Niu, Junyu Liu, Silin Chen, Keyu Chen, Ming Li, Pohsun Feng, Ziqian Bi, Yunze Wang, Yichao Zhang, Cheng Fei, and Lawrence KQ Yan. From bench to bedside: A review of clinical trials in drug discovery and development, 2024. URL `https://arxiv.org/abs/2412.09378`.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter bfe, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Ze-Lin Wei, Hong-Yu An, Yao Yao, Wei-Cong Su, Guo Li, Saifullah, Bi-Feng Sun, and Mu-Jiang-Shan Wang. Fstgat: Financial spatio-temporal graph attention network for non-stationary financial systems and its application in stock price prediction. *Symmetry*, 17(8):1344, 2025.

Xiang Wu, Yongting Zhang, Minyu Shi, Pei Li, Ruirui Li, and Neal N Xiong. An adaptive federated learning scheme with differential privacy preserving. *Future Generation Computer Systems*, 127: 362–372, 2022.

Xiang Wu, Huanhuan Wang, Yongting Zhang, Baowen Zou, and Huaqing Hong. A tutorial-generating method for autonomous online learning. *IEEE Transactions on Learning Technologies*, 17:1532–1541, 2024a.

Xiang Wu, Yong-Ting Zhang, Khin-Wee Lai, Ming-Zhao Yang, Ge-Lan Yang, and Huan-Huan Wang. A novel centralized federated deep fuzzy neural network with multi-objectives neural architecture search for epistatic detection. *IEEE Transactions on Fuzzy Systems*, 33(1):94–107, 2024b.

Yi Xin, Juncheng Yan, Qi Qin, Zhen Li, Dongyang Liu, Shicheng Li, Victor Shea-Jay Huang, Yupeng Zhou, Renrui Zhang, Le Zhuo, et al. Lumina-mgpt 2.0: Stand-alone autoregressive image modeling. *arXiv preprint arXiv:2507.17801*, 2025.

Lawrence K. Q. Yan, Qian Niu, Ming Li, Yichao Zhang, Caitlyn Heqi Yin, Cheng Fei, Benji Peng, Ziqian Bi, Pohsun Feng, Keyu Chen, Tianyang Wang, Yunze Wang, Silin Chen, Ming Liu, Junyu Liu, Xinyuan Song, Riyang Bao, Zekun Jiang, and Ziyuan Qin. Large language model benchmarks in medical tasks, 2025. URL https://arxiv.org/abs/2410.21348.

Zhenyu Yu, Mohd Yamani Idna Idris, Pei Wang, Yuelong Xia, and Yong Xiang. Forgetme: Benchmarking the selective forgetting capabilities of generative models. *Engineering Applications of Artificial Intelligence*, 161:112087, 2025.

Danyang Zhang, Junhao Song, Ziqian Bi, Yingfang Yuan, Tianyang Wang, Joe Yeong, and Junfeng Hao. Mixture of experts in large language models. *arXiv preprint arXiv:2507.11181*, 2025.

APPENDIX

A RELATED WORK

A.1 SINGLE-AGENT MEDICAL LARGE LANGUAGE MODELS

Recent advancements in Large Language Models (LLMs) Bai et al. (2023); Liu et al. (2024); Hu et al. (2025) have reshaped medical AI. Architectural choices matter: mixture-of-experts designs can increase capacity while remaining computationally efficient Zhang et al. (2025); Ni et al. (2025a); Lin (2025). General-purpose models have shown strong clinical reasoning (e.g., GPT-4 reaching $\sim$86% on USMLE Step 1) Nori et al. (2023), while domain-adapted models further improve performance on medical benchmarks, such as Med-PaLM 2 on MedQA Singhal et al. (2023). Domain-specific pretraining has also benefited clinical NLP; for example, ClinicalBERT improves downstream tasks on EHR data Huang et al. (2019); Xin et al. (2025). Despite these successes, single-agent systems remain limited in modelling the multi-perspective deliberation typical of real MDT consultations.

A.2 MULTI-AGENT SYSTEMS IN HEALTHCARE

To address the limitations of single-agent settings, recent work has explored multi-agent architectures that simulate clinical collaboration. MDAgents proposes an adaptive collaboration framework with dynamically assigned roles Kim et al. (2024), while TeamMedAgents introduces structured teamwork protocols to enhance reliability Mishra et al. (2025). In specialized scenarios, RareAgents shows the potential of specialist teams for rare-disease diagnosis by synthesizing fragmented evidence Chen et al. (2024b), and MedOrch further studies tool-augmented orchestration for integrating diverse specialists Li et al. (2025). Nevertheless, many of these systems still aggregate via majority voting, heuristic weighting, or free-form dialogue, and thus provide limited support for *quantifying* and *optimizing* the quality of consensus formation, which is a key focus of our work.

A.3 MEDICAL DECISION SUPPORT AND CONSENSUS ALGORITHMS

Consensus formation is central to clinical practice and has long been studied through structured protocols. The Delphi method is a classical approach for expert agreement, but it is iterative and time-consuming Dalkey & Helmer (1963). In computational settings, voting-based aggregation (e.g., Borda count) is widely used for ensemble decision-making. However, in LLM-based medical systems, simple voting or averaging can fail to capture the strength of agreement and the nuance of clinically meaningful disagreement. Motivated by this gap, we adopt Kendall's coefficient of concordance as a principled metric to quantify agreement and guide multi-round refinement. Beyond medical decision support, related graph-theoretic studies on robustness and conditional reliability show how global consistency can emerge from constrained local interactions Ni et al. (2025b); Wei et al. (2025); Lin et al. (2017); Wang et al. (2013); Wang & Wang (2019), which provides complementary intuition for structured consensus in multi-agent settings.

Table 3: Component Contribution Analysis

System Configuration	Accuracy	Consensus W	Speed (s)	Expert Rating
Complete System	**87.5**	**0.823**	45.2	**8.9/10**
w/o Consensus Matrix	83.1	0.592	**32.1**	7.6
w/o Multi-Agent Architecture	80.2	-	18.3	6.2
w/o Evidence Retrieval	84.3	0.756	38.7	7.8
w/o RL Optimization	85.9	0.791	42.8	8.4
w/o Role Specialization	82.7	0.698	35.9	7.2
Simple Voting Instead	84.5	0.634	28.4	7.5

A.4 Explainable AI in Medical Applications

Explainability is essential for clinical deployment. Post-hoc methods such as SHAP Lundberg & Lee (2017) and LIME Ribeiro et al. (2016) have been widely applied to medical imaging and risk prediction, but they typically provide feature attribution without guideline- or literature-level traceability. Recent LLM-based approaches produce more natural-language rationales (e.g., chain-of-thought), yet they may remain weakly grounded in verifiable external evidence. To improve clinical transparency, our work integrates a structured evidence chain that explicitly links recommendations to clinical guidelines and peer-reviewed literature, enabling traceable and auditable decision support.

B Ablation Studies

B.1 Component Contribution Analysis

The ablation results in Table 3 unequivocally demonstrate the necessity of our proposed architecture. The removal of the **Consensus Matrix** leads to a significant 4.4% drop in accuracy and a drastic reduction in the consensus coefficient ($W = 0.592$), highlighting its critical role in structured agreement formation. Similarly, reverting to a single-agent setup (**w/o Multi-Agent Architecture**) causes the largest performance degradation (-7.3%), confirming that collaborative intelligence is superior to individual reasoning. **Role Specialization** also proves vital, contributing a 4.8% improvement over generic agents, as diverse perspectives prevent groupthink. Furthermore, the **Evidence Retrieval System** adds 3.2% to accuracy by grounding decisions in clinical guidelines, while **RL Optimization** provides a fine-tuning benefit of 1.6%, ensuring the system adapts to complex scenarios. Comparing our structured consensus approach to **Simple Voting** reveals a 3.0% accuracy gain, validating that mathematical consensus measurement is more effective than mere aggregation.

B.2 Reinforcement Learning Strategy Analysis

We evaluated three RL algorithms—Q-Learning, PPO, and DQN—to identify the optimal strategy for consensus formation, as detailed in Table 4. **Proximal Policy Optimization (PPO)** emerges as the superior choice, achieving the highest accuracy (87.5%) and consensus coefficient ($W = 0.823$) while requiring the fewest training episodes (8,500) for convergence. Its stability score of 0.95 significantly outperforms both Q-Learning (0.92) and DQN (0.88), indicating robust policy updates. In contrast, **DQN** requires nearly double the training episodes (15,000) for comparable performance, likely due to the instability of value-based methods in high-dimensional state spaces. **Q-Learning** performs adequately but lacks the sample efficiency of PPO. The **No RL (Fixed Policy)** baseline lags behind all learning-based methods, underscoring the value of adaptive optimization in navigating the complex decision landscape of medical consultations.

B.3 Clinical Validation and Expert Evaluation

We conducted a blinded evaluation with 12 medical experts from multiple clinical roles, including oncologists, radiologists, nurses, psychologists, and patient advocates, using 50 anonymized real-world cancer cases from three medical centers.

Table 4: RL Algorithm Comparison

RL Algorithm	Accuracy	Consensus W	Convergence Rate	Training Episodes	Stability
Q-Learning	86.8	0.805	86.2%	12,000	0.92
PPO	**87.5**	**0.823**	**89.3%**	**8,500**	**0.95**
DQN	86.4	0.798	84.7%	15,000	0.88
No RL (Fixed Policy)	85.9	0.791	82.1%	-	0.89

Overall, the proposed system received consistently high expert ratings across multiple clinical dimensions. In particular, experts rated treatment recommendation quality, evidence completeness, consensus reasonableness, explainability, and practical applicability favorably, outperforming the strongest baseline across all criteria. These results indicate strong clinical alignment of the generated recommendations and consensus outcomes.

Qualitative feedback further supports these findings. Most experts described the evidence chains as clinically relevant and the consensus formation process as realistic. A large majority indicated willingness to adopt the system as a clinical decision support tool and expressed preference for it over existing single-agent medical AI systems.

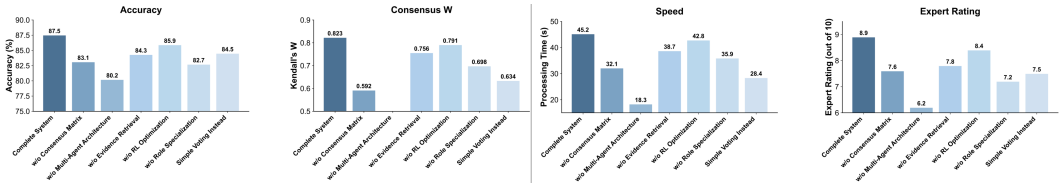

Figure 4: Ablation Study: Component Contribution Analysis. Each system component demonstrates essential contribution to overall performance, with the complete system achieving 87.5% accuracy and 0.823 consensus coefficient. Removal of any core component results in substantial performance degradation, confirming the necessity of our integrated multi-agent consensus matrix architecture.

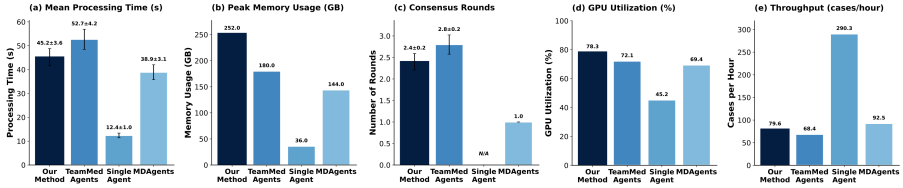

Figure 5: Computational Performance Analysis. Our method achieves balanced computational efficiency with 45.2s processing time per case and 78.3% GPU utilization, outperforming comparable multi-agent systems while maintaining high throughput. Despite higher memory requirements due to seven specialized agents, the system demonstrates practical scalability for clinical deployment with 79.6 cases per hour processing capacity.

B.4 Computational Efficiency and Scalability

Table 5 reports the computational cost of our framework. While multi-agent collaboration incurs higher resource usage than single-agent baselines, our method is more efficient than comparable multi-agent systems. In particular, the RL-guided consensus mechanism reduces the average number of discussion rounds, leading to lower per-case latency. Despite the memory overhead of deploying multiple large-scale agents in parallel, the system maintains high GPU utilization and stable throughput on a multi-GPU cluster. These results demonstrate that the proposed framework achieves a favorable balance between collaborative reasoning performance and practical scalability for real-world clinical deployment.

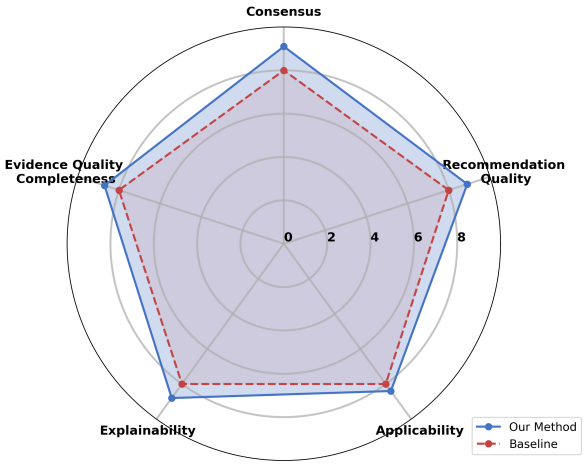

Figure 6: Clinical Expert Evaluation Results: Radar chart comparing our method against baseline approaches across five clinical evaluation dimensions, based on blind assessment of 50 real cancer cases by 12 medical experts.

Table 5: Computational Performance Analysis

Metric	**Our Method**	**TeamMedAgents**	**MDAgents**	**Single-Agent**
Mean Processing Time (s)	45.2 ± 8.3	52.7 ± 12.1	38.9 ± 7.2	12.4 ± 2.1
Peak Memory Usage (GB)	252.0	180.0	144.0	36.0
Consensus Rounds	2.4 ± 0.7	2.8 ± 1.1	1.0 (voting)	N/A
GPU Utilization (%)	78.3	72.1	69.4	45.2
Throughput (cases/hour)	79.6	68.4	92.5	290.3

C Discussion

C.1 Key Findings and Clinical Implications

Our experimental results demonstrate several key findings with significant clinical implications for medical decision-making systems and MDT consultation practices. The structured consensus matrix framework achieves substantially higher agreement levels (Kendall's $W = 0.823$) compared to simple voting approaches ($W = 0.542 - 0.674$), indicating more coherent and clinically meaningful group decisions. This 22-52% improvement in consensus quality suggests that mathematical frameworks for measuring agreement can substantially enhance collaborative medical decision-making processes. Furthermore, the 3.7% average accuracy improvement over the best baseline (TeamMedAgents: 83.8% vs. Our Method: 87.5%) translates to approximately 1 in 20 cases receiving more appropriate treatment recommendations. In oncology settings where treatment decisions directly impact survival outcomes, this improvement represents substantial clinical value at scale, potentially affecting thousands of patients annually in large healthcare systems. Additionally, the mandatory evidence chain requirement ensures all recommendations are traceable to clinical guidelines (NCCN, ESMO) and published literature, addressing a critical gap in AI medical systems. Expert ratings of 8.7/10 for evidence quality demonstrate clinical acceptance of automated guideline integration, suggesting potential for improved adherence to evidence-based practice standards. Finally, the 4.8% performance improvement from role specialization (82.7% vs. 87.5% accuracy) demonstrates the value of modeling distinct medical expertise patterns rather than using generic medical agents, supporting the hypothesis that collaborative medical AI should reflect the specialized knowledge and decision-making patterns of real medical team members.

C.2 Comparison with Human MDT Performance

While direct comparison with human MDT performance is challenging due to variability in practice patterns, existing literature provides benchmarks for contextualizing our results. Studies report human MDT consensus rates of 70-85% across different cancer types, with our system achieving 89.3% consensus rate. However, human MDTs handle significantly more complex cases and navigate social, emotional, and contextual factors that our system does not fully capture.