Proceedings of Machine Learning Research 298:1–31, 2025 Machine Learning for Healthcare

Evaluation of Multi-Agent LLMs in Multidisciplinary Team Decision-Making for Challenging Cancer Cases

Jaesik Kim[*] JAESIK.KIM@PENNMEDICINE.UPENN.EDU
Department of Bioengineering
University of Pennsylvania
Philadelphia, PA, USA

Byounghan Lee[*] QUDGKS96@AJOU.AC.KR
Department of Artificial Intelligence
Ajou University
Suwon, Republic of Korea

Kyung-Ah Sohn[†] KASOHN@AJOU.AC.KR
Department of Artificial Intelligence
Ajou University
Suwon, Republic of Korea

Dokyoon Kim[†] DOKYOON.KIM@PENNMEDICINE.UPENN.EDU
Department of Biostatistics, Epidemiology and Informatics
University of Pennsylvania
Philadelphia, PA, USA

Young Chan Lee[†] MEDCHAN@KHU.AC.KR
Department of Otolaryngology-Head and Neck Surgery
Kyung Hee University Hospital at Gangdong
Seoul, Republic of Korea

Abstract

This study explores the potential of large language model (LLM) agents in real-world clinical decision-making, focusing on their alignment with human experts in cancer multidisciplinary team (MDT) meetings. While LLMs perform well on benchmark medical question-answering tasks, these evaluations often oversimplify the open-ended, multifaceted nature of actual clinical decisions. In practice, MDTs require balancing diverse expert opinions and multiple valid treatment options. Using real MDT meeting data, we compare different LLM approaches including single-agent and multi-agent systems to assess their ability to replicate consensus-based decisions. Our findings indicate that multi-agent, conversation-based systems, which assign specialized roles and facilitate dynamic inter-agent conversation, better align with human expert decisions in our data. Overall, this work highlights the potential practical utility of LLM agents in complex clinical settings and lays the groundwork for their future integration as decision support tools in multidisciplinary medical contexts. Our code is available at `https://github.com/leebyounghan/eval-multi-agent-MDT`

---

[*] These authors contributed equally to this work.

[†] Corresponding authors.

© 2025 J. Kim, B. Lee, K.-A. Sohn, D. Kim & Y.C. Lee.

1. Introduction

In oncology, multidisciplinary team (MDT) meetings represent the gold standard for complex and challenging cancer care (Taberna et al., 2020). During MDTs, specialists from diverse fields (e.g., surgeons, oncologists, radiologists) collaboratively review patient cases, integrating medical evidence (e.g., NCCN guidelines (Caudell et al., 2022)), institutional resources, and patient-specific factors (e.g., comorbidities, socio-economic constraints) to formulate personalized treatment plans. While several medically valid approaches often exist (e.g., surgery versus chemoradiation therapy), the MDT collaboratively determines the most appropriate strategy.

Recent advancements in large language model (LLM)-based multi-agent systems have demonstrated remarkable performance in medical question-answering (QA) benchmarks (Hurst et al., 2024; Team et al., 2023; Xie et al., 2024b). These systems excel in structured benchmarks like MedQA (Jin et al., 2021) and PubMedQA (Jin et al., 2019), where problems are often presented as multiple-choice questions with well-defined answers. Such successes highlight LLM agents' potential to emulate expert-level reasoning in controlled environments. However, these benchmark tasks significantly oversimplify the open-ended and multi-stakeholder nature of real-world clinical decision-making. Unlike multiple-choice formats, real MDT discussions involve the integration of diverse expert perspectives, negotiation among partially conflicting recommendations, and consideration of context-specific constraints that are not explicitly encoded in guidelines. For example, NCCN guidelines may suggest several equally valid treatment pathways for head and neck cancer (e.g., surgery versus chemoradiation), with final decisions depending on factors beyond textbook knowledge such as patient preferences, hospital resource availability, or treatment history. Can LLM agents truly replicate the integrative reasoning process of human experts in these complex scenarios?

To address this gap, we assess the practical utility of LLM agents by comparing their output against real MDT meeting outcomes, rather than relying solely on simplified benchmark datasets. Our key research questions are **1) Can multi-agent LLM frameworks better replicate human MDT decisions compared to single-agent systems (In practical healthcare settings)?** and **2) What mechanisms (e.g., diverse reasoning paths, progression in discussion) contribute to their performance?** This work highlights the practical potential of LLM agents in complex medical contexts and provides evidence for their future integration as decision support tools in multidisciplinary care.

Generalizable Insights about Machine Learning in the Context of Healthcare

This study provides methodological insights into how LLM agents can be adapted for complex, real-world clinical decision-making beyond benchmark QA tasks. Rather than merely applying existing LLM techniques to a new domain, we investigate why multi-agent frameworks outperform single-agent models in complex clinical tasks and identify mechanisms that contribute to this difference. Our findings indicate that the diversity of role specialization and the structured nature of inter-agent conversations, allowing for a dynamic exchange of opinions and critical refinement of initial reasoning, are key factors in achieving better alignment with expert clinical decision-making. These mechanisms, which have proven effective in the medical QA benchmarks, similarly enhance performance in the real-world, multi-faceted scenarios encountered in clinical practice, suggesting that embracing collaborative, multi-perspective reasoning can lead to more robust and context-aware decision support tools in healthcare.

2. Related Work

The evolution of LLM agents has attracted significant attention in both the natural language processing and medical informatics communities. While early works primarily focused on single-agent systems for clinical reasoning and medical question answering (QA), recent advancements have introduced multi-agent frameworks that better reflect the collaborative nature of healthcare decision-making. Notably, MDAgents (Kim et al., 2024) demonstrated the effectiveness of explicitly recruiting role-specialized agents (e.g., radiologists, oncologists, surgeons), creating structured collaborative environments and achieving superior performance on medical QA benchmarks compared to single-agent approaches.

However, despite these advancements, most existing evaluations have been limited to simulated scenarios with clearly defined answers. Real-world clinical decision-making such as MDT meetings, involves complex, open-ended reasoning across multiple valid treatment paths. Some recent studies have started exploring this direction. For example, Schmidl et al. (2024) evaluated ChatGPT on challenging MDT cases, finding that it could generate clinically relevant suggestions. However, their study did not explore the use of multi-agent collaboration.

In this study, we address this gap by systematically evaluating both single-agent and multi-agent LLM systems on real-world MDT data. By analyzing their ability to replicate consensus-based decisions and identifying mechanisms such as inter-agent dialogue and role specialization, we provide empirical evidence on how LLMs can be better aligned with expert-level clinical reasoning in complex medical contexts.

3. Methods

3.1. Agent Frameworks

We used representative agent frameworks for our study. Figure 1 describes all methods including single-agent and multi-agent.

3.1.1. Single-Agent Methods

Chain-of-Thought (CoT) A single LLM agent generates treatment decisions through sequential reasoning (Wei et al., 2022). The agent processes patient data and produces a step-by-step rationale before finalizing a decision.

Simulation-of-Thought (SoT) A single LLM agent simulates multidisciplinary debate internally through role-playing inspired by (Wang et al., 2024). The agent alternates between 3-5 specialist personas up to 10 turns. Debate history is summarized by another LLM to produce a final decision.

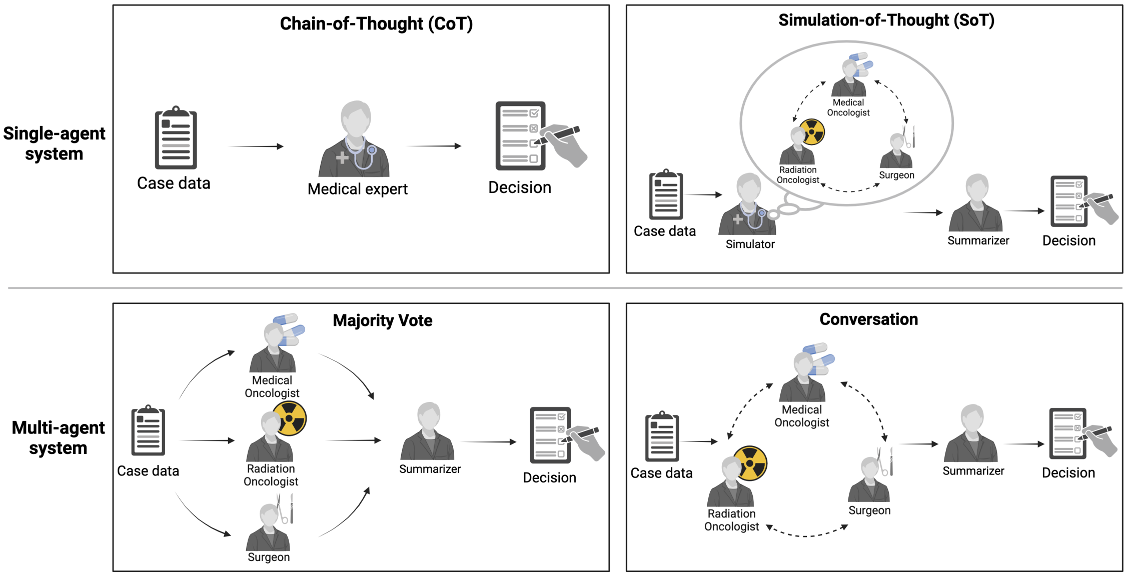


Figure 1: Overview of single-agent methods and multi-agent methods for clinical decision-making

3.1.2. Multi-Agent Methods

All multi-agent frameworks first identify required specialties through a recruitment phase, where a recruiter agent dynamically selects up to five specialists relevant to the patient case (e.g., oncologist, surgeon, radiologist). The recruited agents then are involved following methods.

Majority Vote Maximum five LLM agents with preassigned roles (e.g. head and neck surgeon, medical oncologist, radiation oncologist) independently analyze cases using CoT reasoning(Wang et al., 2022). A separate summarizer agent aggregates decisions via majority rule.

MDAgents We adapt the MDAgents (Kim et al., 2024) framework that dynamically routes cases to expert panels based on predicted complexity. In our analysis, for all 31 cases (classified as *high-difficulty* by MDAgents' triage system), the Integrated Care Team (ICT) protocol was triggered, proceeding through three sequential stages: the Initial Assessment Team conducts preliminary clinical evaluations, the Diagnostic Team performs granular analyses (e.g., radiological-pathological correlation), and the Final Review & Decision Team synthesizes multidisciplinary inputs into consensus-based plans. Central to this process is the structured report generation protocol, where each team, led by a lead clinician, collaboratively produces incremental reports that inform downstream decisions, ensuring traceable rationale alignment across stages.

Group Chat (AG2) Using the `Group Chat` workflow in AG2 (Wu et al., 2023), five role-specialized agents conduct structured 8-turn debates moderated by a `group chat manager` agent. Then, a separate summarizer agent aggregates decisions from a chat history.

3.2. Evaluation

We conducted a rigorous pairwise comparison across all methods using LLM-as-a-judge(Zheng et al., 2023) with three random seeds and two distinct LLM judges. To enhance the robustness of our results, both judges employed the same prompt but generated evaluations using different seeds. Furthermore, to mitigate potential order bias, the presentation order of the anonymized methods (Method A/B) was randomized for each comparison. For each of the 31 MDT cases, twelve comparative assessments were performed (3 seeds × 2 evaluators x 2 for A/B order shuffling)

Input Structure Each evaluation instance received (1) two anonymized decisions (Method A/B) and (2) MDT ground truth. Specialist roles for the judges (e.g., surgeon or radiologist) were redacted to prevent bias.

Evaluation Prompt

Please compare these two decisions with the ground truth answer and determine which one is closer:

[Method A Decision]
{methodA_decision}

[Method B Decision]
{methodB_decision}

[Ground Truth Answer]
{answer}

Which decision (Method A or Method B) is closer to the ground truth answer based on their rationale?

If both decisions are **the completely same**, answer with 'Tie'. Otherwise, answer with either 'Method A' or 'Method B'.

Metrics

- *Win Rate*: Percentage of comparisons where Method A/B was preferred.
- *Tie Rate*: Cases where two methods have completely same decision.
- *Statistical Significance*: Wilcoxon signed-rank test compared method pairs' win counts across 31 cases ($\alpha = 0.05$).

4. Cohort

We retrospectively analyzed multidisciplinary team (MDT) meeting records for head and neck cancer cases involving complex clinical challenges requiring diverse expert perspectives.

In particular, we focused on high-complexity cases where adherence to NCCN guidelines was challenging. These cases included recurrent or metastatic cancers requiring nuanced decision-making, cases with ambiguous or controversial guideline applicability, and rare cancer subtypes with limited treatment options. Each case includes comprehensive patient information including illness presentation, medical history, radiology findings, pathology results, and laboratory data, along with the final MDT consensus decisions (Appendix Table B1). Our dataset, collected over a seven-year period from 2018 to 2024 at Kyung Hee University Hospital at Gangdong, comprises 31 deidentified cases selected to evaluate agent systems (Appendix Table B2). This study utilized de-identified patient data and was approved by the Institutional Review Board of Kyung Hee University Hospital at Gangdong (Approval No. KHNMC 2025-02-013). Due to its retrospective nature and the exclusive use of anonymized data, the requirement for informed consent was waived by the IRB.

Table 1: Pairwise comparison of single-agent and multi-agent decision-making methods (GPT-4o)

Method A	**Method B**	**A Win (%)**	**B Win (%)**	**Tie (%)**	
CoT	**Majority Vote**	20.16	29.03	50.81	(**)
CoT	**MDAgents**	29.57	40.32	30.11	(***)
CoT	**Group Chat**	26.08	49.46	24.46	(***)
Majority Vote	**MDAgents**	26.61	41.67	31.72	(*)
Majority Vote	**Group Chat**	20.70	42.74	36.56	(***)
MDAgents	**Group Chat**	29.30	40.05	30.65	(*)
CoT	SoT	39.52	44.35	16.13	(ns)

(*): $p < 0.05$, (**): $p < 0.01$, (***): $p < 0.001$, and (ns): non-significant results.

5. Results

5.1. Study Design

Our primary task is to generate open-ended clinical decisions in response to medical queries from real MDT meetings on challenging head and neck cancer cases. This involves predicting the final MDT decision based on patient-specific data, including clinical history and diverse medical report summaries. Given these inputs, LLM-based systems generate structured clinical decisions with treatment options and rationale.

5.2. Multi-agent systems outperform single-agent systems

Our experiments were centered on gpt-4o (version 2024-08-06, Default). For the pairwise inter-model evaluations, we utilized o4-mini (version 2025-04-16, Default) as the judge.

Our evaluation clearly demonstrates that multi-agent frameworks yield decisions more aligned with clinical consensus compared to single-agent approaches, as confirmed by the pairwise comparisons in Table 1. In each comparison involving a multi-agent approach (Group Chat, MDAgents, or Majority Vote), the multi-agent system outstripped its single-agent counterpart by a statistically significant margin. Notably, when comparing the Group Chat framework with the CoT method, Group Chat achieved a 49.46% win rate versus only 26.08% for CoT ($p < 0.001$), demonstrating nearly a two-fold improvement and highlighting the benefits of integrating multiple perspectives in clinical reasoning.

Among the multi-agent methods, structured discussions provide a slight advantage over simple voting-based strategies. Specifically, the Group Chat approach outperformed the Majority Vote method with a win rate of 42.74% ($p < 0.001$), suggesting that the collaborative dynamics of Group Chat add value in replicating human expert decision-making. Additionally, Group Chat outperformed MDAgents with a win rate of 40.05% ($p < 0.05$), establishing it as the most effective method among those tested.

5.3. Why Multi-Agent Systems Excel

Hypothesis 1. Does a structured conversation format enhance consensus? Our experimental results reveal that conversation-based frameworks such as MDAgents and Group Chat significantly outperform the non-conversational Majority Vote method. This observation raises an interesting question: does the presence of a conversation format similarly influence performance in single-agent settings? To investigate this, we compared the CoT approach with our proposed Simulation-of-Thought (SoT) method, both single-agent systems, differing only in whether they incorporate a structured conversation format. However, as shown in Table 1, there is no statistically significant difference between CoT and SoT (see Appendix Table **??** for detailed comparisons across all methods). This finding suggests that, in the context of single-agent systems, merely introducing a conversation format does not necessarily improve decision-making performance.

Hypothesis 2. Does role specialization drive improvement? To investigate whether role specialization enhances decision quality, we assessed the impact of role diversity while keeping the number of participating agents constant. Specifically, we designed additional experimental conditions. While previous work (Tang et al., 2023; Kim et al., 2024) has demonstrated the importance of role specialization, our study specifically explores the effect of role diversity by varying it within our experimental design.

In the standard Group Chat setting, a recruiter agent dynamically selects up to five specialists relevant to the patient case, ensuring that each participant has a distinct area of expertise (e.g., surgeon, oncologist, radiologist). In contrast, Group Chat *without (w/o) recruit* condition eliminates the recruiter agent, assigning all agents a generic role as "medical expert" with no explicit specialization. This condition isolates the effect of role specialization by ensuring that all agents approach the decision without a defined area of expertise.

To further examine the impact of role diversity, we introduced Group Chat *with (w/) less diverse role* setting. A recruiter agent still selects up to five specialists, but only first two distinct roles are represented, with the remaining agents being duplicates of these two roles. This condition retains the number of experts while significantly reducing the diversity of perspectives within the group.

As shown in Table 2, the results indicate that role specialization significantly improves the quality of the decision. The baseline Group Chat achieved a win rate of 53.49%, whereas the Group Chat *w/o recruit* setting performed worse at 25.81% ($p < 0.001$). We also examined the impact of reduced role diversity within a recruited group. Compared to the baseline Group Chat, the *w/ less diverse role* setting showed lower performance, with a win rate of 42.74% ($p < 0.01$). This suggests that utilizing agents with a diverse set of specialized roles leads to more effective decision-making.

These findings demonstrate that assigning specialized roles significantly improves decision-making, particularly in structured, conversation-based systems. However, for the Majority Vote approach, the distinction between *w/o recruit* and *w/ recruit* conditions was minimal, indicating that the benefits of role specialization are most pronounced when agents engage in structured dialogues rather than simple majority-based aggregation.

Table 2: Impact of specialist recruitment and role diversity on decision performance

Method A	**Method B**	**A Win (%)**	**B Win (%)**	**Tie (%)**	
Group Chat w/o recruit	**Group Chat**	25.81	53.49	20.70	(***)
Group Chat w/ less diverse role	**Group Chat**	29.03	42.74	28.23	(**)
Majority Vote w/o recruit	Majority Vote	20.43	21.24	58.33	(ns)

(*): $p < 0.05$, (**): $p < 0.01$, (***): $p < 0.001$, and (ns): non-significant results.

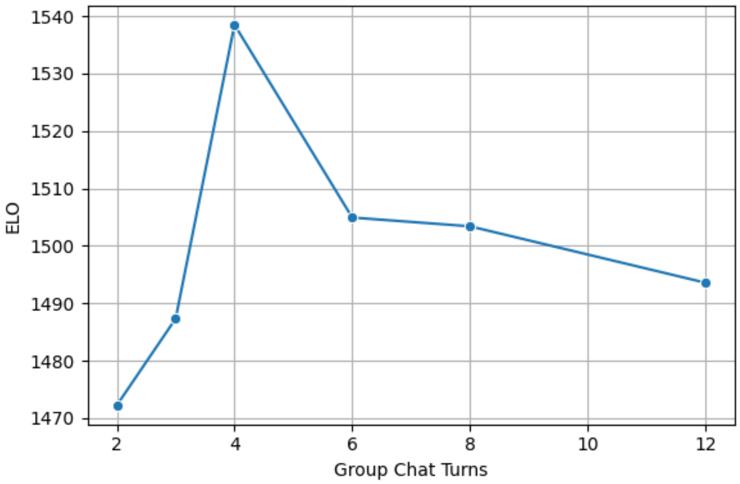


Figure 2: Effect of conversation turn modifications on Group Chat performance

Hypothesis 3. Is More Cumulated Context in Discussion Beneficial? To explore whether extending or reducing the number of conversation turns enhances decision quality, we compared variants of the Group Chat setting with modified discussion lengths. Specifically, we investigated a configuration with 2,3,4,6,8,12-turn. As shown in Figure 2 and Table B4, the findings indicate that extending the conversation from very short formats (2- and 3-turns) to 4-turns enhances decision-making performance.

In addition, our analysis reveals a saturation point beyond which additional turns do not yield further advantages. Specifically, additional turns beyond the 4-turn point rarely contribute novel insights or substantive decision changes, and instead often reiterate previously agreed-upon decisions, resulting in simplified decision statements and lexical changes in summarization outputs. This observation will be further discussed in Section 5.5 (Task 1). Thus, while an initial extension of conversation helps agents achieve better-aligned decisions, excessively long discussions appear to offer limited additional benefit.

Hypothesis 4: Does Simply Increasing the Amount of Context Lead improvement? To investigate whether the performance degradation of CoT and Majority Vote is simply due to insufficient context, we incorporated Retrieval-Augmented Generation (RAG) to assess whether increasing the amount of available context alone leads to performance improvements. While these methods operate under a limited information scope, RAG enables the retrieval of additional external knowledge, potentially mitigating context-related constraints.

Experimental results in Table 3 indicate that applying MedRAG(Xiong et al., 2024) to CoT and Group Chat resulted in only marginal performance gains, with no statistically significant improvements. These results imply that simply increasing the amount of relevant information does not automatically enhance performance. Instead, the structured and contextually integrated information generated through conversation proves to be more critical than merely the volume of data.

Table 3: Effect of MedRAG on CoT and Group Chat performances

Method A	**Method B**	**A Win (%)**	**B Win (%)**	**Tie (%)**	
CoT	CoT w/ MedRAG	31.72	30.11	38.17	(ns)
Group Chat	Group Chat w/ MedRAG	36.02	37.90	26.08	(ns)

(*): $p < 0.05$, (**): $p < 0.01$, (***): $p < 0.001$, and (ns): non-significant results.

Hypothesis 5. Does Conversation Enable to Accept Better Opinion? We examined whether incorporating structured conversation on top of agents' initial opinions enables the acceptance of better opinions. As shown in Table 4, augmenting the baseline Majority Vote method with a conversation component yields statistically significant improvements. In particular, when four conversational turns are introduced, the extended approach attains a win rate of 19.09% compared to only 11.29% for the baseline, with similar gains observed when an eight-turn conversation is implemented (win rate of 26.88% versus 15.86% for the baseline). These results suggest that the conversational modality effectively facilitates the exchange and reconciliation of divergent views among the agents.

Next, we tested whether a structured multi-agent conversation can robustly overcome initial errors. We checked if the final decision included an initial error and calculated the corresponding percentages. As shown in Table 5, initial flawed errors, such as gene therapy, CAR-T therapy, or transplantation, were corrected 83.87, 80.65, and 93.55 percent of the time, respectively, through inter-agent conversation. Although the final decision deviated more from the ground truth when no initial error was present, the conversation still allowed agents to progressively refine their logic and transform an inadequate initial answer into a more balanced and accurate final decision. In essence, multi-agent conversations foster a dynamic evolution of reasoning that goes beyond simply aggregating independent opinions and ultimately arrives at a consensus that better mirrors expert judgment.

Table 4: Impact of integrating conversational turns into Majority Vote

Method A	**Method B**	**A Win (%)**	**B Win (%)**	**Tie (%)**	
Majority Vote	**Majority Vote + 4-turn**	11.29	19.09	69.62	(*)
Majority Vote	**Majority Vote + 8-turn**	15.86	26.88	57.26	(***)

(*): $p < 0.05$, (**): $p < 0.01$, (***): $p < 0.001$, and (ns): non-significant results.

Table 5: Initial error resistance of the multi-agent

Initial error	**Resistance rate (%)**
Gene therapy	83.87
CAR-T therapy	80.65
Transplantation	93.55

5.4. Cross-Model Validation of Multi-Agent Advantages

To further validate our findings across different model architectures, we conducted identical experiments with three additional models: MedGemma-27B, medical-domain-specific LLM (Sellergren et al., 2025); Gemma-27B, general-purpose LLM (Team et al., 2025); and o4-mini, reasoning-focused model.

The ELO ranking results in Table 6 reveal that multi-agent methods consistently outperform single-agent approaches within each tested model, confirming broad generalizability. However, we observed that different models achieve optimal performance with different multi-agent frameworks based on their architectural strengths (detailed results in Appendix Tables B6, B7, and B5). The reasoning-focused o4-mini excels with the complex MDAgents framework (ELO: 1591.42), achieving a substantial 53.75-point improvement over its CoT baseline. In contrast, both the medical domain-specific MedGemma-27B and general-purpose GPT-4o achieve their best performance with the more streamlined Group Chat approach, with improvements of 68.31 and 81.05 ELO points over CoT respectively.

Beyond framework compatibility, we examined whether multi-agent systems can compensate for inherent model limitations. The cross-model comparison in Table 7 demonstrates that a well-suited multi-agent framework can elevate the performance of a generally capable model to surpass that of a domain-specialized model operating in a single-agent capacity. For instance, while the domain-specific MedGemma-27B using CoT significantly outperforms GPT-4o using CoT (42.74% vs. 28.76% win rate, $p < 0.001$), GPT-4o equipped with the Group Chat framework reverses this outcome, decisively beating the single-agent MedGemma-27B (45.97% vs. 29.84%, $p < 0.001$). This suggests that the collaborative structure can compensate for a lack of domain-specific pre-training. Interestingly, when both models utilize the Group Chat framework, their performance becomes statistically indistinguishable ($p = ns$), indicating that the advanced framework can level the playing field between models of different specializations.

This pattern holds even when comparing models of similar baseline strength. Although GPT-4o CoT and Gemma-27B CoT perform comparably ($p = ns$), equipping Gemma-27B with a simple Majority Vote framework allows it to significantly outperform GPT-4o CoT (49.73% vs. 32.26%, $p < 0.001$). However, the more sophisticated Group Chat framework gives GPT-4o a performance edge over Gemma-27B's Majority Vote approach (47.04% vs. 33.60%, $p < 0.05$).

Finally, when facing a stronger reasoning model like o4-mini, which outperforms GPT-4o in a single-agent CoT comparison (47.31% vs. 28.76%, $p < 0.001$), the multi-agent boost on GPT-4o is sufficient to achieve parity (GPT-4o Group Chat vs. o4-mini CoT, $p = ns$). However, the advantage is re-established when the stronger base model is paired with its own optimal multi-agent framework; o4-mini using its best framework, MDAgents, decisively outperforms GPT-4o using Group Chat (52.69% vs. 25.27%, $p < 0.001$). This highlights that while multi-agent systems provide a significant performance uplift, the combination of a superior base model with its best-matched collaborative framework yields the highest quality outcomes.

Model	**Method**	**ELO**
o4-mini	MDAgents	1591.42
MedGemma-27B	Group Chat	1564.65
MedGemma-27B	SoT	1548.82
o4-mini	CoT	1537.67
o4-mini	Group Chat	1531.29
GPT-4o	Group Chat	1530.41
o4-mini	Majority vote	1517.92
Gemma-27B	Majority vote	1506.21
GPT-4o	MDAgents	1497.26
Gemma-27B	MDAgents	1496.63
MedGemma-27B	CoT	1496.33
MedGemma-27B	MDAgents	1495.71
Gemma-27B	Group Chat	1488.27
o4-mini	SoT	1486.79
MedGemma-27B	Majority vote	1481.21
GPT-4o	SoT	1466.76
GPT-4o	Majority vote	1460.04
GPT-4o	CoT	1449.36
Gemma-27B	CoT	1447.37
Gemma-27B	SoT	1405.89

Table 6: Overall ELO ranking of models and methods

Table 7: Cross-model and cross-method comparison

Model A	Model B	A Win (%)	B Win (%)	Tie (%)	
GPT-4o CoT	**MedGemma-27B CoT**	28.76	42.74	28.49	(***)
GPT-4o Group Chat	MedGemma-27B CoT	45.97	29.84	24.19	(***)
GPT-4o Group Chat	MedGemma-27B Group Chat	37.10	47.31	15.59	(ns)
GPT-4o CoT	Gemma-27B CoT	42.20	38.17	19.62	(ns)
GPT-4o CoT	**Gemma-27B Majority vote**	32.26	49.73	18.01	(***)
GPT-4o Group Chat	Gemma-27B Majority vote	47.04	33.60	19.35	(*)
GPT-4o CoT	**o4-mini CoT**	28.76	47.31	23.92	(***)
GPT-4o Group Chat	o4-mini CoT	36.56	35.48	27.96	(ns)
GPT-4o Group Chat	**o4-mini MDAgents**	25.27	52.69	22.04	(***)

(*): $p < 0.05$, (**): $p < 0.01$, (***): $p < 0.001$, and (ns): non-significant results.

5.5. Human Evaluation

Task 1: Justification of LLM-as-a-Judge LLM-as-a-Judge frameworks have been increasingly used in the medical domain for evaluating AI-generated outputs(Xie et al., 2024a; Brake and Schaaf, 2024; Li et al., 2024). To further assess the reliability of LLM-based evaluations in clinical decision-making, we conducted a human evaluation study (see Appendix Figure B2). The primary objective was to determine the extent to which LLM-as-a-judge aligns with human expert judgment. An otolaryngologist independently assessed the same cases using the identical evaluation framework employed by the LLM judges.

As presented in Figure 3, the human expert demonstrated a preference for the Group Chat method (29.0%) over CoT (12.9%), a pattern consistent with the LLM evaluators. Figure 4 shows the inter-rater agreement between a human expert and two LLM judges. Cases labeled as *Tie* by the human expert were excluded because tie-break criteria may vary and are inherently subjective. In cases where the human expert clearly preferred one method, both LLM judges agreed with the human expert over 92% of the time, indicating strong concordance. This alignment underscores the potential of the Group Chat framework as a robust tool for reasoning and decision-making, given its consistent preference across different evaluators.

Unlike the LLM judges, the human expert exhibited a higher tendency to classify cases as equivalent (*Tie*), with a tie rate of 58.1%. This tendency can be attributed to the heightened sensitivity of LLM judges to fine-grained lexical differences, even when evaluating treatments of similar medical concepts. LLMs may be excessively sensitive to linguistic variations, such as differences in phrasing or syntactic structure, despite the underlying clinical recommendations being functionally equivalent.

Task 2: Clinical Acceptability of Treatment Suggestions Beyond validating the LLM-as-a-judge framework, the second evaluation task examined the clinical acceptability of treatment suggestions generated by the best-performing method (Group Chat) (see Appendix Figure B3). The human expert categorized each suggested treatment into one of three categories:

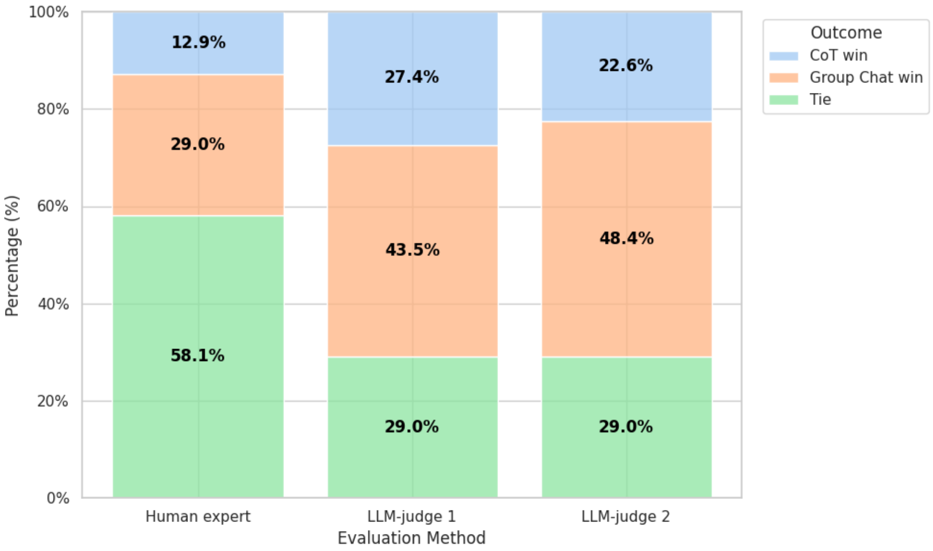


Figure 3: Comparison of the judge decisions for CoT versus Group Chat

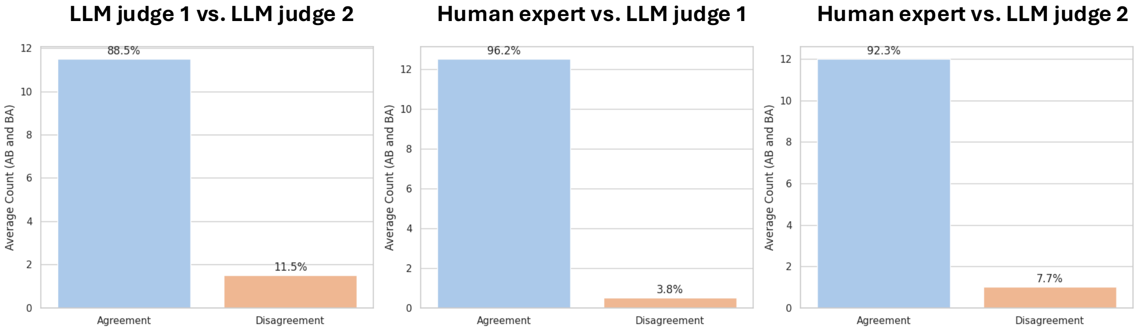


Figure 4: Preference comparison between human expert and LLM-as-a-judges for CoT versus Group Chat, excluding tied outcomes.

- *Correct*: The proposed treatment is identical to or fully aligned with MDT decision answer.
- *Different but Acceptable*: The proposed treatment differs from the answer but remains clinically valid and reasonable.
- *Wrong*: The proposed treatment is incorrect or inappropriate.

The results, presented in Figure 5, indicate that out of 31 cases, 11 (35.5%) were classified as *Correct*, 18 (58.1%) as *Different but Acceptable*, and 2 (6.5%) as Wrong. While the generated treatment decision-making do not perfectly align with actual MDT decisions, a key finding is that 93.5% were deemed clinically acceptable, an notably high proportion.

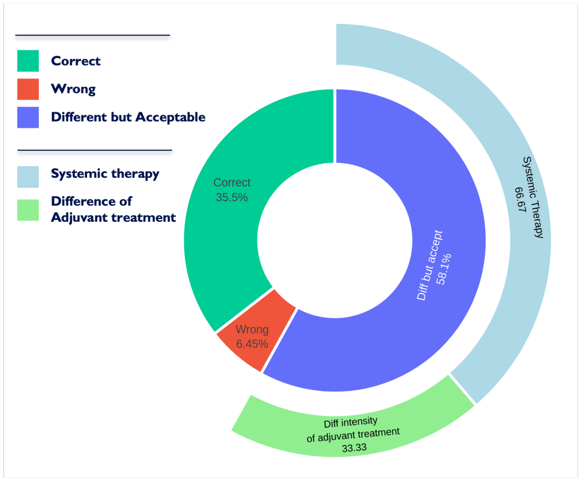
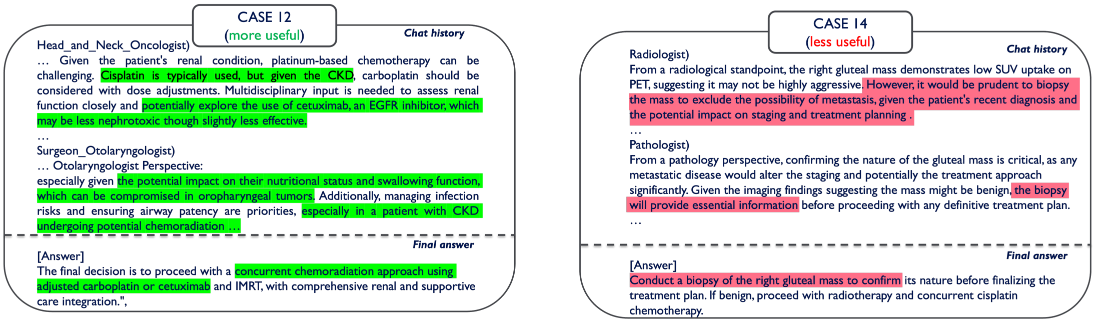


Figure 5: Human evaluation for Group Chat

This underscores the potential of LLM-generated recommendations, particularly from multi-agent frameworks, to support clinical decision-making in real-world oncology settings.

A deeper analysis of treatment decisions classified as *Different but Acceptable* revealed notable patterns in treatment suggestions (see Appendix Table B8). In numerous instances, systemic therapy was suggested, particularly for patients with metastatic progression, which aligns with established clinical guidelines. This tendency was also observed in prior studies, where GPT-4o and GPT-4 frequently recommended systemic therapy in similar contexts(Schmidl et al., 2024). However, a substantial proportion of cases involved patients who were either ineligible for systemic therapy due to medical conditions or patient preference. Furthermore, deviations from MDT recommendations were observed in terms of treatment intensity. While the overall treatment strategies adhered to guideline-based decision, variations were noted specifically in the intensity of adjuvant treatment. These discrepancies illustrate how LLM-based decision, despite being grounded in clinical guidelines, may differ in detailed treatment considerations depending on individual patient factors and institutional practices.


Chat history
Final answer
Chat history
Final answer

CASE 12
(more useful)

Head_and_Neck_Oncologist)
… Given the patient's renal condition, platinum-based chemotherapy can be challenging. Cisplatin is typically used, but given the CKD, carboplatin should be considered with dose adjustments. Multidisciplinary input is needed to assess renal function closely and potentially explore the use of cetuximab, an EGFR inhibitor, which may be less nephrotoxic though slightly less effective.
…
Surgeon_Otolaryngologist)
… Otolaryngologist Perspective:
especially given the potential impact on their nutritional status and swallowing function, which can be compromised in oropharyngeal tumors. Additionally, managing infection risks and ensuring airway patency are priorities, especially in a patient with CKD undergoing potential chemoradiation …

[Answer]
The final decision is to proceed with a concurrent chemoradiation approach using adjusted carboplatin or cetuximab and IMRT, with comprehensive renal and supportive care integration.",

(*a*) More Useful Case

CASE 14
(less useful)

Radiologist)
From a radiological standpoint, the right gluteal mass demonstrates low SUV uptake on PET, suggesting it may not be highly aggressive. However, it would be prudent to biopsy the mass to exclude the possibility of metastasis, given the patient's recent diagnosis and the potential impact on staging and treatment planning .
…
Pathologist)
From a pathology perspective, confirming the nature of the gluteal mass is critical, as any metastatic disease would alter the staging and potentially the treatment approach significantly. Given the imaging findings suggesting the mass might be benign, the biopsy will provide essential information before proceeding with any definitive treatment plan.
…

[Answer]
Conduct a biopsy of the right gluteal mass to confirm its nature before finalizing the treatment plan. If benign, proceed with radiotherapy and concurrent cisplatin chemotherapy.

(*b*) Less Useful Case

Figure 6: Case Study of Group Chat

TASK 3 : Case Study To further explore the practical implications of our findings, we conducted a case study examining specific instances where multi-agent LLM systems either succeeded or deviated from human expert consensus in MDT decisions. This qualitative investigation provides deeper insights into how LLM agents interpret complex medical data, negotiate between competing treatment options, and incorporate multi-stakeholder perspectives in their final decision-making.

In Figure 6, *Case 12* illustrates this decision-making process. Following extensive multidisciplinary discussions, the team recognized that the patient's compromised renal function rendered standard cisplatin therapy too risky due to its well-documented nephrotoxicity. Input from the nephrologist highlighted the urgent need to protect renal function, while the medical and radiation oncologists evaluated alternative agents that could still offer effective tumor control. Weighing these considerations carefully, the team reached a consensus to opt for cetuximab, as it provided a safer profile for patients with CKD while maintaining efficacy in treating HPV-positive oropharyngeal carcinoma (Figure B4).

Similarly, *Case 14* highlights a different clinical challenge. In this case, while performing a biopsy on the right gluteal mass is not inherently incorrect, it is considered a less useful decision in the given clinical context (Figure B5). Metastasis from head and neck cancer to this region is extremely rare, and delaying definitive treatment in urgent settings may negatively impact patient outcomes. Therefore, the biopsy decision, although clinically valid, was less aligned with expert decision and clinical urgency, reflecting a lower priority compared to other immediate intervention options.

6. Discussion

This study demonstrates that multi-agent LLM frameworks are more effective at replicating MDT decision-making than single-agent systems. In particular, our results show that conversation-based methods such as Group Chat and MDAgents not only outperform CoT approaches but also yield decisions that are more closely aligned with human expert consensus. The performance improvements appear to stem from two core mechanisms. First, role specialization and the recruitment of diverse experts enable the system to harness multiple streams of clinical insight (Tang et al., 2023). Our experiments showed that assigning explicit specialist roles from oncologists to surgeons and radiologists significantly boosted decision quality compared to systems treating all agents as generic medical experts. Second, structured inter-agent conversations facilitated a dynamic exchange of opinions, allowing agents to critically assess and refine their initial reasoning. In contrast, incorporating a conversation format within single-agent systems, as seen with the SoT method, did not yield statistically significant benefits. This suggests that the benefit of conversation in complex clinical decision-making is tightly coupled with the presence of diverse viewpoints rather than being a mere function of additional reasoning turns. Importantly, these multi-agent advantages generalize across different model architectures, with our cross-model analysis revealing that collaborative frameworks can enable weaker models to achieve competitive performance levels against stronger single-agent systems. This suggests that multi-agent approaches may be particularly valuable for democratizing access to high-quality clinical decision support, as they can enhance the capabilities of more accessible, general-purpose models in specialized medical domains.

Limitations Despite these promising results, several limitations warrant discussion. First, while our dataset provides valuable insights, its single-institution scope limits generalizability, highlighting the need for multi-institutional validation. Second, the text-based representation of MDT data, while rich in clinical details, does not fully capture the diversity of information available in clinical practice, such as imaging studies or laboratory trends. This limitation points to the need for future studies that integrate multimodal data. Additionally, considerations such as patient preferences, socio-economic factors, institutional resource constraints, and cultural nuances were beyond the scope of this investigation, even though they play critical roles in actual MDT deliberations. Finally, while the majority of treatment suggestions by the best-performing method were deemed clinically acceptable, a small number of recommendations were incorrect. This fact highlights the necessity for clinical validation and expert supervision before any potential deployment in real-world settings.

Looking ahead, these findings open several avenues for future research. Enhanced multi-agent frameworks that can incorporate structured non-textual data, systematic prompt optimization methods (Yuksekgonul et al., 2025), and advanced contextual engineering approaches could further improve performance in clinical domains(Mei et al., 2025). Additionally, clearer methodologies for integrating patient-specific socio-economic and cultural factors including medical insurance policies, and more extensive validations across diverse clinical scenarios are all promising directions. Moreover, while the current study establishes the viability of LLM agents as decision support systems, further research is needed to transition these methodologies from experimental settings to clinically integrated tools that complement physician judgment rather than replace it.

In conclusion, our study demonstrates that multi-agent LLM systems, especially those employing structured, conversation-based approaches with explicitly assigned specialist roles, outperform single-agent methods in replicating multidisciplinary team decisions for complex cancer care. These findings underscore the importance of integrating diverse expert perspectives and dynamic inter-agent conversations to achieve consensus that aligns more closely with human clinical reasoning. While acknowledging limitations such as the retrospective nature of our dataset and the challenges posed by a text-only representation of multifaceted clinical data, this work lays a solid foundation for further research into LLM-based decision support tools. Ultimately, our results point toward a promising future in which LLM-driven multi-agent systems can serve as effective adjuncts to clinician expertise, provided that rigorous clinical supervision and integration of additional patient-centric factors are maintained.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) [RS-2023-00241230] and Information & Communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development [IITP-2025-RS-2023-00255968] grant.

References

Nathan Brake and Thomas Schaaf. Comparing two model designs for clinical note generation; is an llm a useful evaluator of consistency? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 352–363, 2024.

Jimmy J. Caudell, Maura L. Gillison, Ellie Maghami, Sharon Spencer, David G. Pfister, Douglas Adkins, Andrew C. Birkeland, David M. Brizel, Paul M. Busse, Anthony J. Cmelak, A. Dimitrios Colevas, David W. Eisele, Thomas Galloway, Jessica L. Geiger, Robert I. Haddad, Wesley L. Hicks, Ying J. Hitchcock, Antonio Jimeno, Debra Leizman, Loren K. Mell, Bharat B. Mittal, Harlan A. Pinto, James W. Rocco, Cristina P. Rodriguez, Panayiotis S. Savvides, David Schwartz, Jatin P. Shah, David Sher, Maie St. John, Randal S. Weber, Gregory Weinstein, Frank Worden, Justine Yang Bruce, Sue S. Yom, Weining Zhen, Jennifer L. Burns, and Susan D. Darlow. Nccn guidelines® insights: Head and neck cancers, version 1.2022: Featured updates to the nccn guidelines. *Journal of the National Comprehensive Cancer Network*, 20(3):224 – 234, 2022. doi: 10.6004/jnccn.2022.0016. URL https://jnccn.org/view/journals/jnccn/20/3/article-p224.xml.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 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. *Applied Sciences*, 11(14):6421, 2021.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William 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)*, pages 2567–2577, 2019.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. Mdagents: An adaptive collaboration of llms for medical decision-making. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*, 2024.

Lingrui Mei, Jiayu Yao, Yuyao Ge, Yiwei Wang, Baolong Bi, Yujun Cai, Jiazhi Liu, Mingyu Li, Zhong-Zhi Li, Duzhen Zhang, et al. A survey of context engineering for large language models. *arXiv preprint arXiv:2507.13334*, 2025.

Benedikt Schmidl, Tobias Hütten, Steffi Pigorsch, Fabian Stögbauer, Cosima C Hoch, Timon Hussain, Barbara Wollenberg, and Markus Wirth. Assessing the role of advanced artificial intelligence as a tool in multidisciplinary tumor board decision-making for recurrent/metastatic head and neck cancer cases–the first study on chatgpt 4o and a comparison to chatgpt 4.0. *Frontiers in Oncology*, 14:1455413, 2024.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Miren Taberna, Francisco Gil Moncayo, Enric Jané-Salas, Maite Antonio, Lorena Arribas, Esther Vilajosana, Elisabet Peralvez Torres, and Ricard Mesía. The multidisciplinary team (mdt) approach and quality of care. *Frontiers in Oncology*, 10, 2020. ISSN 2234-943X. doi: 10.3389/fonc.2020.00085. URL `https://www.frontiersin.org/journals/oncology/articles/10.3389/fonc.2020.00085`.

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. *arXiv preprint arXiv:2311.10537*, 2023.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H 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*, 2022.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 257–279, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversation. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2023.

Yiqing Xie, Sheng Zhang, Hao Cheng, Pengfei Liu, Zelalem Gero, Cliff Wong, Tristan Naumann, Hoifung Poon, and Carolyn Rose. Doclens: Multi-aspect fine-grained medical text evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 649–679, 2024a.

Yunfei Xie, Juncheng Wu, Haoqin Tu, Siwei Yang, Bingchen Zhao, Yongshuo Zong, Qiao Jin, Cihang Xie, and Yuyin Zhou. A preliminary study of o1 in medicine: Are we closer to an ai doctor? *arXiv preprint arXiv:2409.15277*, 2024b.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics ACL 2024*, pages 6233–6251, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. URL `https://aclanthology.org/2024.findings-acl.372`.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Pan Lu, Zhi Huang, Carlos Guestrin, and James Zou. Optimizing generative ai by backpropagating language model feedback. *Nature*, 639(8055):609–616, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623, 2023.

Appendix A. Prompt

A.1. CoT

CoT Prompt

System: You are a medical expert who makes a decision. Please provide your decision as a short answer in the following format at the end of your response:

[Rationale] rationale here

[Answer] short answer here

User: {Question} Let's think step by step

A.2. SoT

SoT Prompt

System: You are a medical expert. Imagine experts with differing opinions are participating in a MDT meeting. Create a simulation of a conversational debate between experts with differing opinions based on given patient information. The goal is for the medical experts to find consensus on a final answer through a series of questions and answers.
Max turns is 10.

# Steps
1. Create 3-5 expert personas with different specialties and initial opinions
2. Simulate a discussion where each expert:
- Presents their initial assessment
- Asks questions about aspects relevant to their specialty
- Responds to other experts' concerns
- Adjusts their position based on new information
3. Guide the discussion toward consensus by:
- Highlighting areas of agreement
- Addressing key points of disagreement
- Synthesizing different perspectives
4. Conclude with a clear consensus recommendation or decision

# Chat format

[Expert_name]
content

[Expert_name]
content

...

User: {Question} Let's simulate a MDT meeting.

Summarizer Prompt

System: You are a medical expert who summarizes the opinions of other experts. Please provide the final decision from the meeting as a short answer in the following format at the end of your response:

[Rationale] rationale here

[Answer] short answer here

User: {chat_history} What is the final decision from the meeting?

A.3. Majority Vote

Recruiter Prompt

System: You are a hiring manager to plan MDT meeting.

User: {Question} Who should be invited to the MDT meeting? Maximum 5 experts.

CoT Prompt for each expert

System: You are a {expert} who makes a decision. Please provide your decision as a short answer in the following format at the end of your response:

[Rationale] rationale here

[Answer] short answer here

User: {Question} Let's think step by step.

Summarizer Prompt

System: You are a medical expert who summarizes the opinions of other experts. Please provide the majority opinion as a short answer in the following format at the end of your response:

[Rationale] rationale here

[Answer] short answer here

User: {expert_opinions} What is the majority opinion?

A.4. Group Chat

Recruiter Prompt

System: You are a hiring manager to plan MDT meeting.

User: {Question} Who should be invited to the MDT meeting? Maximum 5 experts.

Participant Prompt

System: You are an {expert} who is involved in the MDT to make a decision. Engage with other participants and be concise.

Summarizer Prompt

System: You are a medical expert who summarizes the opinions of other experts. Please provide the final decision from the meeting as a short answer in the following format at the end of your response:

[Rationale] rationale here

[Answer] short answer here

User: {chat_history} What is the final decision from the meeting?

Group Chat Prompt

User: {Question} Let's have a discussion.

A.5. Initial error in Group Chat

Initial Error Context

User: {Question} Let's have a discussion.

Agent 1: This patient presents a highly complex and challenging case. Given the recent promising clinical outcomes associated with {error} for head and neck cancer, it is imperative that we consider its application for this patient.

Appendix B. Figures and Tables

Table B1: MDT data example

Data example	
Present illness [patient demographics and history]	
Radiology Results	
# [Date]: [Examination]	
[Findings]	
# [Date]: [Examination]	
[Findings]	
...	
Pathology Results	
# [Date]: [Examination]	
[DIAGNOSIS]	
...	
Blood Test Results	
Molecular Diagnosis	[Date]
HPV Genotyping	[Result]

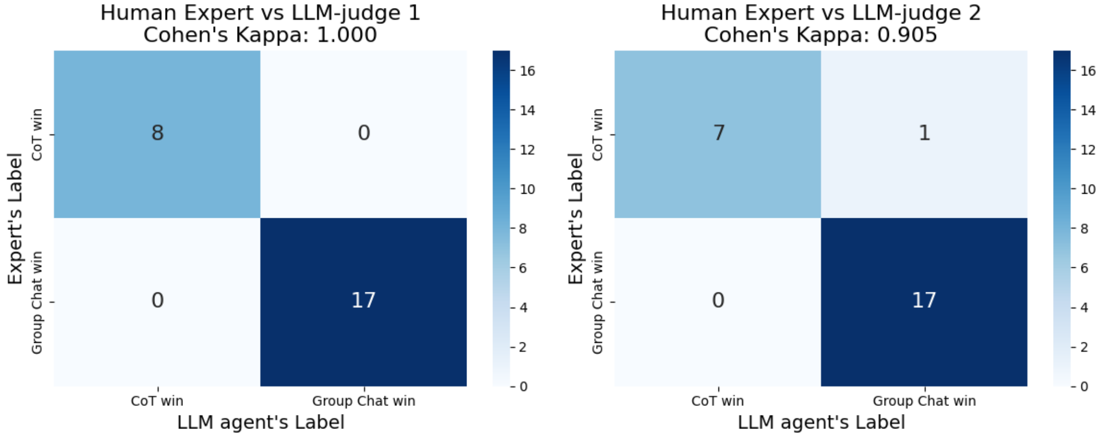
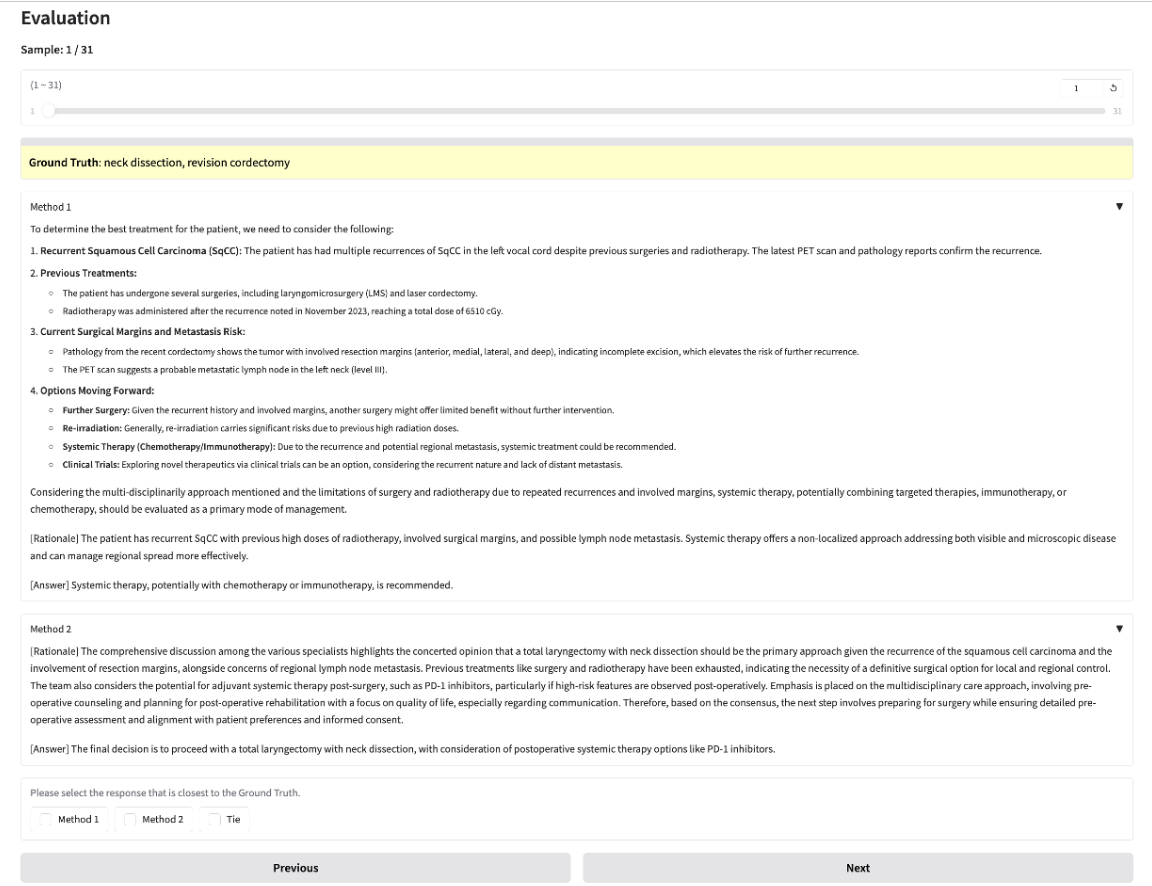


Figure B1: Confusion matrices and Cohen's Kappa (κ) scores between a human expert and two LLM-as-a-judges. The analysis is performed on non-tie outcomes.

Table B2: Patient Demographics and Clinical Characteristics

Case	**Gender**	**Age**	**Recurrence**	**Metastasis**
1	M	63	Yes	Yes
2	M	69	Yes	Yes
3	M	52	No	Yes
4	M	46	No	Yes
5	F	59	Yes	No
6	M	64	No	Yes
7	M	55	No	Yes
8	M	81	Yes	No
9	M	70	No	Yes
10	M	66	Yes	Yes
11	M	57	Yes	No
12	F	48	No	Yes
13	M	65	No	Yes
14	M	80	No	Yes
15	M	51	No	Yes
16	F	51	No	No
17	M	59	Yes	Yes
18	M	74	No	Yes
19	M	73	Yes	Yes
20	M	66	No	Yes
21	F	61	Yes	Yes
22	M	57	Yes	Yes
23	F	47	Yes	Yes
24	M	60	Yes	Yes
25	F	27	No	No
26	F	78	Yes	Possible
27	M	60	No	Yes
28	M	76	No	No
29	F	77	No	Yes
30	F	79	Possible	No
31	M	53	No	Yes

”Possible” indicates a state where a condition is not confirmed but is highly suspected.



Evaluation

Sample: 1 / 31

Ground Truth: neck dissection, revision cordectomy

Method 1

To determine the best treatment for the patient, we need to consider the following:

1. **Recurrent Squamous Cell Carcinoma (SqCC):** The patient has had multiple recurrences of SqCC in the left vocal cord despite previous surgeries and radiotherapy. The latest PET scan and pathology reports confirm the recurrence.
2. **Previous Treatments:**
   - The patient has undergone several surgeries, including laryngomicrosurgery (LMS) and laser cordectomy.
   - Radiotherapy was administered after the recurrence noted in November 2023, reaching a total dose of 6510 cGy.
3. **Current Surgical Margins and Metastasis Risk:**
   - Pathology from the recent cordectomy shows the tumor with involved resection margins (anterior, medial, lateral, and deep), indicating incomplete excision, which elevates the risk of further recurrence.
   - The PET scan suggests a probable metastatic lymph node in the left neck (level III).
4. **Options Moving Forward:**
   - **Further Surgery:** Given the recurrent history and involved margins, another surgery might offer limited benefit without further intervention.
   - **Re-irradiation:** Generally, re-irradiation carries significant risks due to previous high radiation doses.
   - **Systemic Therapy (Chemotherapy/Immunotherapy):** Due to the recurrence and potential regional metastasis, systemic treatment could be recommended.
   - **Clinical Trials:** Exploring novel therapeutics via clinical trials can be an option, considering the recurrent nature and lack of distant metastasis.

Considering the multi-disciplinarily approach mentioned and the limitations of surgery and radiotherapy due to repeated recurrences and involved margins, systemic therapy, potentially combining targeted therapies, immunotherapy, or chemotherapy, should be evaluated as a primary mode of management.

[Rationale] The patient has recurrent SqCC with previous high doses of radiotherapy, involved surgical margins, and possible lymph node metastasis. Systemic therapy offers a non-localized approach addressing both visible and microscopic disease and can manage regional spread more effectively.

[Answer] Systemic therapy, potentially with chemotherapy or immunotherapy, is recommended.

Method 2

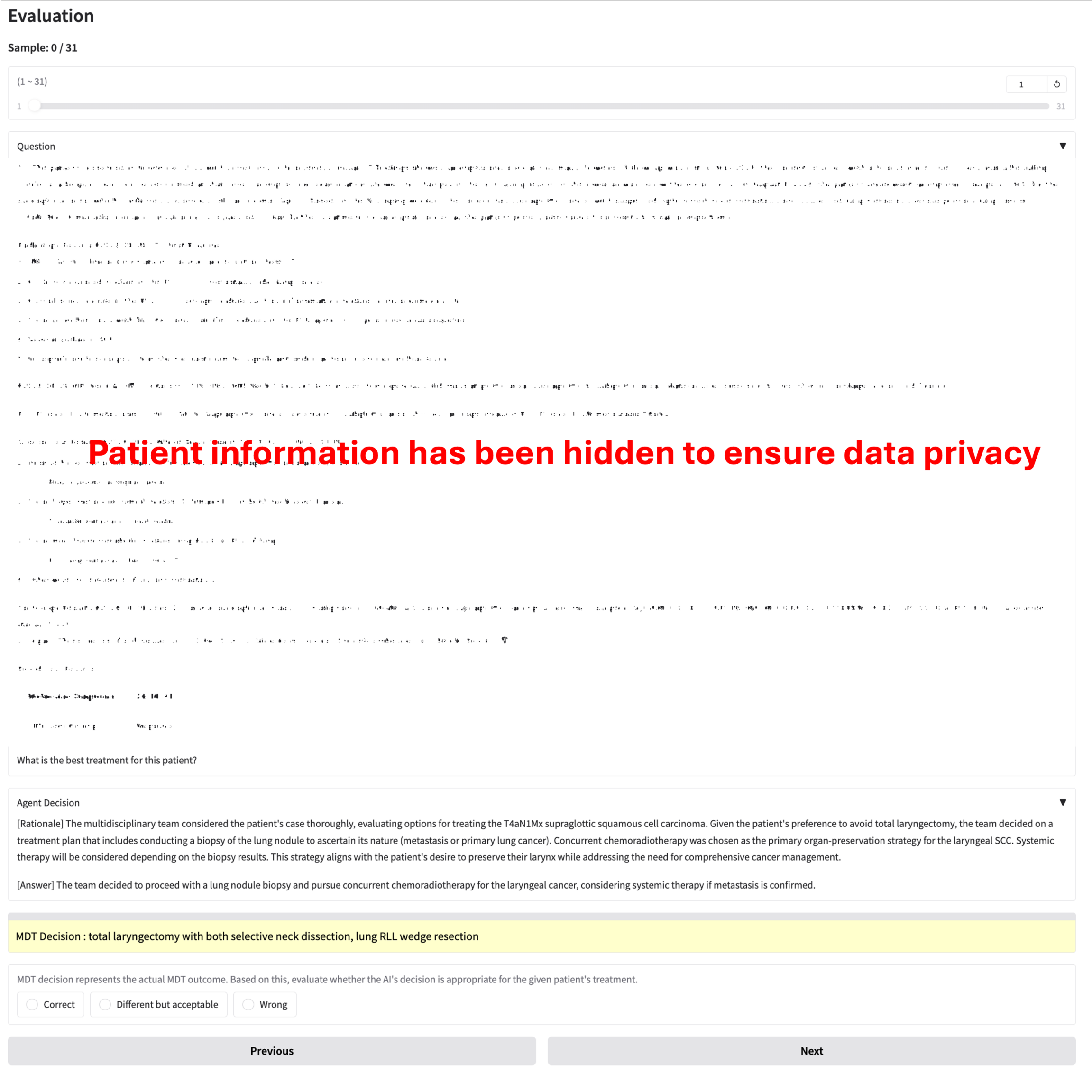
[Rationale] The comprehensive discussion among the various specialists highlights the concerted opinion that a total laryngectomy with neck dissection should be the primary approach given the recurrence of the squamous cell carcinoma and the involvement of resection margins, alongside concerns of regional lymph node metastasis. Previous treatments like surgery and radiotherapy have been exhausted, indicating the necessity of a definitive surgical option for local and regional control. The team also considers the potential for adjuvant systemic therapy post-surgery, such as PD-1 inhibitors, particularly if high-risk features are observed post-operatively. Emphasis is placed on the multidisciplinary care approach, involving pre-operative counseling and planning for post-operative rehabilitation with a focus on quality of life, especially regarding communication. Therefore, based on the consensus, the next step involves preparing for surgery while ensuring detailed pre-operative assessment and alignment with patient preferences and informed consent.

[Answer] The final decision is to proceed with a total laryngectomy with neck dissection, with consideration of postoperative systemic therapy options like PD-1 inhibitors.

Please select the response that is closest to the Ground Truth.

☐ Method 1 ☐ Method 2 ☐ Tie

Figure B2: Annotation interface used to collect Human evaluation task1



Evaluation

Sample: 0 / 31

Question

[redacted]

Patient information has been hidden to ensure data privacy

What is the best treatment for this patient?

Agent Decision

[Rationale] The multidisciplinary team considered the patient's case thoroughly, evaluating options for treating the T4aN1Mx supraglottic squamous cell carcinoma. Given the patient's preference to avoid total laryngectomy, the team decided on a treatment plan that includes conducting a biopsy of the lung nodule to ascertain its nature (metastasis or primary lung cancer). Concurrent chemoradiotherapy was chosen as the primary organ-preservation strategy for the laryngeal SCC. Systemic therapy will be considered depending on the biopsy results. This strategy aligns with the patient's desire to preserve their larynx while addressing the need for comprehensive cancer management.

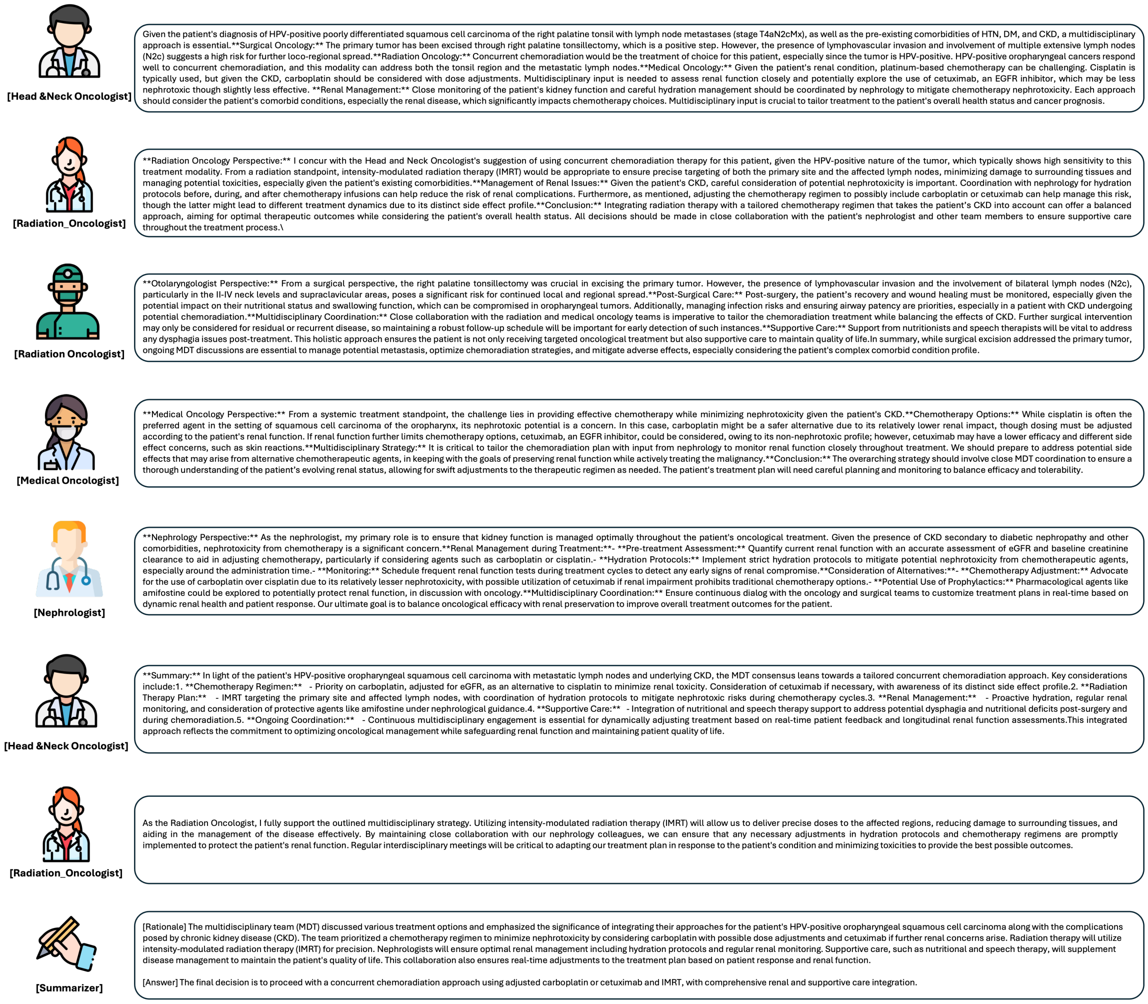
[Answer] The team decided to proceed with a lung nodule biopsy and pursue concurrent chemoradiotherapy for the laryngeal cancer, considering systemic therapy if metastasis is confirmed.

MDT Decision : total laryngectomy with both selective neck dissection, lung RLL wedge resection

MDT decision represents the actual MDT outcome. Based on this, evaluate whether the AI's decision is appropriate for the given patient's treatment.

○ Correct ○ Different but acceptable ○ Wrong

Figure B3: Annotation interface used to collect Human evaluation task2



[Head &Neck Oncologist]

Given the patient's diagnosis of HPV-positive poorly differentiated squamous cell carcinoma of the right palatine tonsil with lymph node metastases (stage T4aN2cMx), as well as the pre-existing comorbidities of HTN, DM, and CKD, a multidisciplinary approach is essential.**Surgical Oncology:** The primary tumor has been excised through right palatine tonsillectomy, which is a positive step. However, the presence of lymphovascular invasion and involvement of multiple extensive lymph nodes (N2c) suggests a high risk for further loco-regional spread.**Radiation Oncology:** Concurrent chemoradiation would be the treatment of choice for this patient, especially since the tumor is HPV-positive. HPV-positive oropharyngeal cancers respond well to concurrent chemoradiation, and this modality can address both the tonsil region and the metastatic lymph nodes.**Medical Oncology:** Given the patient's renal condition, platinum-based chemotherapy can be challenging. Cisplatin is typically used, but given the CKD, carboplatin should be considered with dose adjustments. Multidisciplinary input is needed to assess renal function closely and potentially explore the use of cetuximab, an EGFR inhibitor, which may be less nephrotoxic though slightly less effective. **Renal Management:** Close monitoring of the patient's kidney function and careful hydration management should be coordinated by nephrology to mitigate chemotherapy nephrotoxicity. Each approach should consider the patient's comorbid conditions, especially the renal disease, which significantly impacts chemotherapy choices. Multidisciplinary input is crucial to tailor treatment to the patient's overall health status and cancer prognosis.

[Radiation_Oncologist]

Radiation Oncology Perspective: I concur with the Head and Neck Oncologist's suggestion of using concurrent chemoradiation therapy for this patient, given the HPV-positive nature of the tumor, which typically shows high sensitivity to this treatment modality. From a radiation standpoint, intensity-modulated radiation therapy (IMRT) would be appropriate to ensure precise targeting of both the primary site and the affected lymph nodes, minimizing damage to surrounding tissues and managing potential toxicities, especially given the patient's existing comorbidities.**Management of Renal Issues:** Given the patient's CKD, careful consideration of potential nephrotoxicity is important. Coordination with nephrology for hydration protocols before, during, and after chemotherapy infusions can help reduce the risk of renal complications. Furthermore, as mentioned, adjusting the chemotherapy regimen to possibly include carboplatin or cetuximab can help manage this risk, though the latter might lead to different treatment dynamics due to its distinct side effect profile.**Conclusion:** Integrating radiation therapy with a tailored chemotherapy regimen that takes the patient's CKD into account can offer a balanced approach, aiming for optimal therapeutic outcomes while considering the patient's overall health status. All decisions should be made in close collaboration with the patient's nephrologist and other team members to ensure supportive care throughout the treatment process.\

[Radiation Oncologist]

Otolaryngologist Perspective: From a surgical perspective, the right palatine tonsillectomy was crucial in excising the primary tumor. However, the presence of lymphovascular invasion and the involvement of bilateral lymph nodes (N2c), particularly in the II-IV neck levels and supraclavicular areas, poses a significant risk for continued local and regional spread.**Post-Surgical Care:** Post-surgery, the patient's recovery and wound healing must be monitored, especially given the potential impact on their nutritional status and swallowing function, which can be compromised in oropharyngeal tumors. Additionally, managing infection risks and ensuring airway patency are priorities, especially in a patient with CKD undergoing potential chemoradiation.**Multidisciplinary Coordination:** Close collaboration with the radiation and medical oncology teams is imperative to tailor the chemoradiation treatment while balancing the effects of CKD. Further surgical intervention may only be considered for residual or recurrent disease, so maintaining a robust follow-up schedule will be important for early detection of such instances.**Supportive Care:** Support from nutritionists and speech therapists will be vital to address any dysphagia issues post-treatment. This holistic approach ensures the patient is not only receiving targeted oncological treatment but also supportive care to maintain quality of life.In summary, while surgical excision addressed the primary tumor, ongoing MDT discussions are essential to manage potential metastasis, optimize chemoradiation strategies, and mitigate adverse effects, especially considering the patient's complex comorbid condition profile.

[Medical Oncologist]

Medical Oncology Perspective: From a systemic treatment standpoint, the challenge lies in providing effective chemotherapy while minimizing nephrotoxicity given the patient's CKD.**Chemotherapy Options:** While cisplatin is often the preferred agent in the setting of squamous cell carcinoma of the oropharynx, its nephrotoxic potential is a concern. In this case, carboplatin might be a safer alternative due to its relatively lower renal impact, though dosing must be adjusted according to the patient's renal function. If renal function further limits chemotherapy options, cetuximab, an EGFR inhibitor, could be considered, owing to its non-nephrotoxic profile; however, cetuximab may have a lower efficacy and different side effect concerns, such as skin reactions.**Multidisciplinary Strategy:** It is critical to tailor the chemoradiation plan with input from nephrology to monitor renal function closely throughout treatment. We should prepare to address potential side effects that may arise from alternative chemotherapeutic agents, in keeping with the goals of preserving renal function while actively treating the malignancy.**Conclusion:** The overarching strategy should involve close MDT coordination to ensure a thorough understanding of the patient's evolving renal status, allowing for swift adjustments to the therapeutic regimen as needed. The patient's treatment plan will need careful planning and monitoring to balance efficacy and tolerability.

[Nephrologist]

Nephrology Perspective: As the nephrologist, my primary role is to ensure that kidney function is managed optimally throughout the patient's oncological treatment. Given the presence of CKD secondary to diabetic nephropathy and other comorbidities, nephrotoxicity from chemotherapy is a significant concern.**Renal Management during Treatment:**- **Pre-treatment Assessment:** Quantify current renal function with an accurate assessment of eGFR and baseline creatinine clearance to aid in adjusting chemotherapy, particularly if considering agents such as carboplatin or cisplatin.- **Hydration Protocols:** Implement strict hydration protocols to mitigate potential nephrotoxicity from chemotherapeutic agents, especially around the administration time.- **Monitoring:** Schedule frequent renal function tests during treatment cycles to detect any early signs of renal compromise.**Consideration of Alternatives:**- **Chemotherapy Adjustment:** Advocate for the use of carboplatin over cisplatin due to its relatively lesser nephrotoxicity, with possible utilization of cetuximab if renal impairment prohibits traditional chemotherapy options.- **Potential Use of Prophylactics:** Pharmacological agents like amifostine could be explored to potentially protect renal function, in discussion with oncology.**Multidisciplinary Coordination:** Ensure continuous dialog with the oncology and surgical teams to customize treatment plans in real-time based on dynamic renal health and patient response. Our ultimate goal is to balance oncological efficacy with renal preservation to improve overall treatment outcomes for the patient.

[Head &Neck Oncologist]

Summary: In light of the patient's HPV-positive oropharyngeal squamous cell carcinoma with metastatic lymph nodes and underlying CKD, the MDT consensus leans towards a tailored concurrent chemoradiation approach. Key considerations include:1. **Chemotherapy Regimen:** - Priority on carboplatin, adjusted for eGFR, as an alternative to cisplatin to minimize renal toxicity. Consideration of cetuximab if necessary, with awareness of its distinct side effect profile.2. **Radiation Therapy Plan:** - IMRT targeting the primary site and affected lymph nodes, with coordination of hydration protocols to mitigate nephrotoxic risks during chemotherapy cycles.3. **Renal Management:** - Proactive hydration, regular renal monitoring, and consideration of protective agents like amifostine under nephrological guidance.4. **Supportive Care:** - Integration of nutritional and speech therapy support to address potential dysphagia and nutritional deficits post-surgery and during chemoradiation.5. **Ongoing Coordination:** - Continuous multidisciplinary engagement is essential for dynamically adjusting treatment based on real-time patient feedback and longitudinal renal function assessments.This integrated approach reflects the commitment to optimizing oncological management while safeguarding renal function and maintaining patient quality of life.

[Radiation_Oncologist]

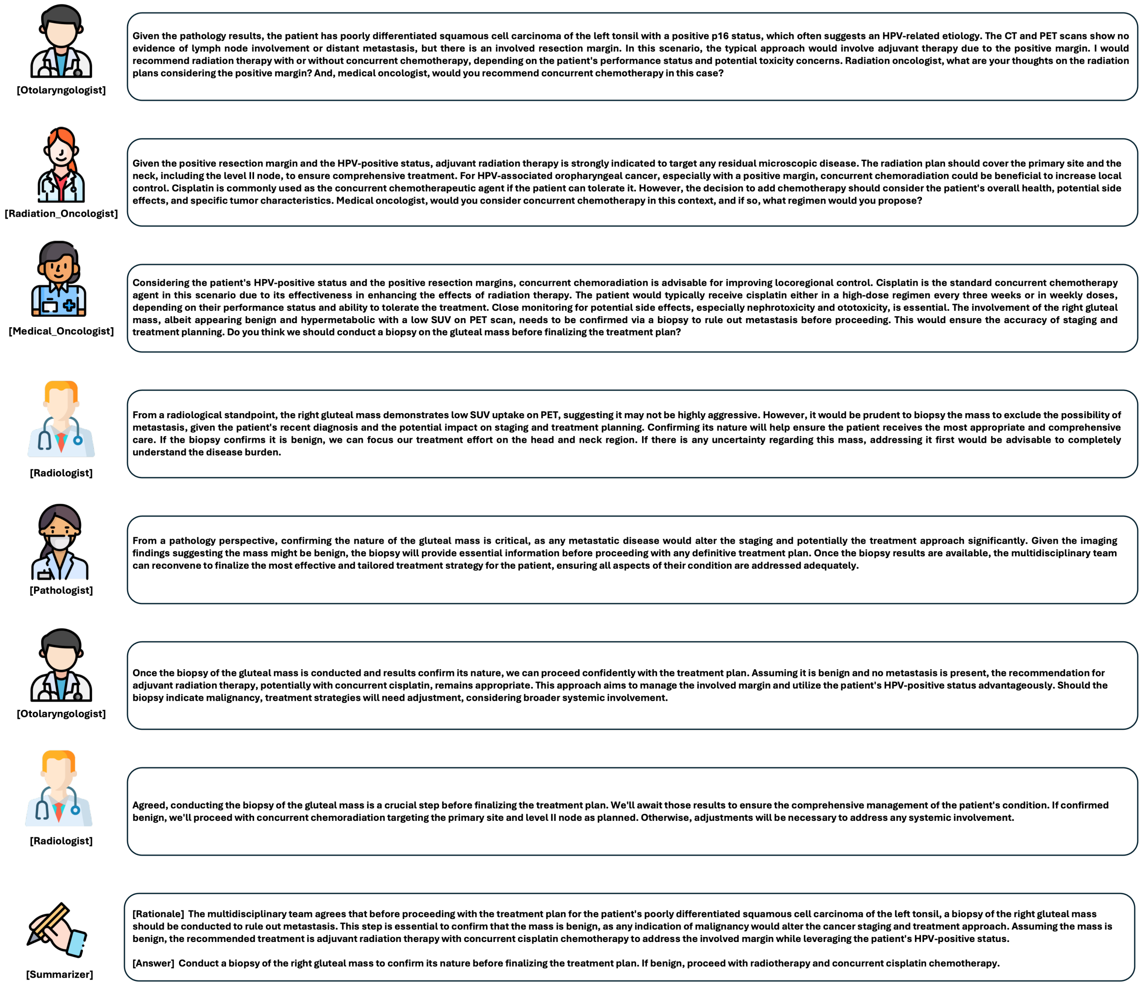
As the Radiation Oncologist, I fully support the outlined multidisciplinary strategy. Utilizing intensity-modulated radiation therapy (IMRT) will allow us to deliver precise doses to the affected regions, reducing damage to surrounding tissues, and aiding in the management of the disease effectively. By maintaining close collaboration with our nephrology colleagues, we can ensure that any necessary adjustments in hydration protocols and chemotherapy regimens are promptly implemented to protect the patient's renal function. Regular interdisciplinary meetings will be critical to adapting our treatment plan in response to the patient's condition and minimizing toxicities to provide the best possible outcomes.

[Summarizer]

[Rationale] The multidisciplinary team (MDT) discussed various treatment options and emphasized the significance of integrating their approaches for the patient's HPV-positive oropharyngeal squamous cell carcinoma along with the complications posed by chronic kidney disease (CKD). The team prioritized a chemotherapy regimen to minimize nephrotoxicity by considering carboplatin with possible dose adjustments and cetuximab if further renal concerns arise. Radiation therapy will utilize intensity-modulated radiation therapy (IMRT) for precision. Nephrologists will ensure optimal renal management including hydration protocols and regular renal monitoring. Supportive care, such as nutritional and speech therapy, will supplement disease management to maintain the patient's quality of life. This collaboration also ensures real-time adjustments to the treatment plan based on patient response and renal function.

[Answer] The final decision is to proceed with a concurrent chemoradiation approach using adjusted carboplatin or cetuximab and IMRT, with comprehensive renal and supportive care integration.

Figure B4: Full Chat history of CASE 12



[Otolaryngologist]

Given the pathology results, the patient has poorly differentiated squamous cell carcinoma of the left tonsil with a positive p16 status, which often suggests an HPV-related etiology. The CT and PET scans show no evidence of lymph node involvement or distant metastasis, but there is an involved resection margin. In this scenario, the typical approach would involve adjuvant therapy due to the positive margin. I would recommend radiation therapy with or without concurrent chemotherapy, depending on the patient's performance status and potential toxicity concerns. Radiation oncologist, what are your thoughts on the radiation plans considering the positive margin? And, medical oncologist, would you recommend concurrent chemotherapy in this case?

[Radiation_Oncologist]

Given the positive resection margin and the HPV-positive status, adjuvant radiation therapy is strongly indicated to target any residual microscopic disease. The radiation plan should cover the primary site and the neck, including the level II node, to ensure comprehensive treatment. For HPV-associated oropharyngeal cancer, especially with a positive margin, concurrent chemoradiation could be beneficial to increase local control. Cisplatin is commonly used as the concurrent chemotherapeutic agent if the patient can tolerate it. However, the decision to add chemotherapy should consider the patient's overall health, potential side effects, and specific tumor characteristics. Medical oncologist, would you consider concurrent chemotherapy in this context, and if so, what regimen would you propose?

[Medical_Oncologist]

Considering the patient's HPV-positive status and the positive resection margins, concurrent chemoradiation is advisable for improving locoregional control. Cisplatin is the standard concurrent chemotherapy agent in this scenario due to its effectiveness in enhancing the effects of radiation therapy. The patient would typically receive cisplatin either in a high-dose regimen every three weeks or in weekly doses, depending on their performance status and ability to tolerate the treatment. Close monitoring for potential side effects, especially nephrotoxicity and ototoxicity, is essential. The involvement of the right gluteal mass, albeit appearing benign and hypermetabolic with a low SUV on PET scan, needs to be confirmed via a biopsy to rule out metastasis before proceeding. This would ensure the accuracy of staging and treatment planning. Do you think we should conduct a biopsy on the gluteal mass before finalizing the treatment plan?

[Radiologist]

From a radiological standpoint, the right gluteal mass demonstrates low SUV uptake on PET, suggesting it may not be highly aggressive. However, it would be prudent to biopsy the mass to exclude the possibility of metastasis, given the patient's recent diagnosis and the potential impact on staging and treatment planning. Confirming its nature will help ensure the patient receives the most appropriate and comprehensive care. If the biopsy confirms it is benign, we can focus our treatment effort on the head and neck region. If there is any uncertainty regarding this mass, addressing it first would be advisable to completely understand the disease burden.

[Pathologist]

From a pathology perspective, confirming the nature of the gluteal mass is critical, as any metastatic disease would alter the staging and potentially the treatment approach significantly. Given the imaging findings suggesting the mass might be benign, the biopsy will provide essential information before proceeding with any definitive treatment plan. Once the biopsy results are available, the multidisciplinary team can reconvene to finalize the most effective and tailored treatment strategy for the patient, ensuring all aspects of their condition are addressed adequately.

[Otolaryngologist]

Once the biopsy of the gluteal mass is conducted and results confirm its nature, we can proceed confidently with the treatment plan. Assuming it is benign and no metastasis is present, the recommendation for adjuvant radiation therapy, potentially with concurrent cisplatin, remains appropriate. This approach aims to manage the involved margin and utilize the patient's HPV-positive status advantageously. Should the biopsy indicate malignancy, treatment strategies will need adjustment, considering broader systemic involvement.

[Radiologist]

Agreed, conducting the biopsy of the gluteal mass is a crucial step before finalizing the treatment plan. We'll await those results to ensure the comprehensive management of the patient's condition. If confirmed benign, we'll proceed with concurrent chemoradiation targeting the primary site and level II node as planned. Otherwise, adjustments will be necessary to address any systemic involvement.

[Summarizer]

[Rationale] The multidisciplinary team agrees that before proceeding with the treatment plan for the patient's poorly differentiated squamous cell carcinoma of the left tonsil, a biopsy of the right gluteal mass should be conducted to rule out metastasis. This step is essential to confirm that the mass is benign, as any indication of malignancy would alter the cancer staging and treatment approach. Assuming the mass is benign, the recommended treatment is adjuvant radiation therapy with concurrent cisplatin chemotherapy to address the involved margin while leveraging the patient's HPV-positive status.

[Answer] Conduct a biopsy of the right gluteal mass to confirm its nature before finalizing the treatment plan. If benign, proceed with radiotherapy and concurrent cisplatin chemotherapy.

Figure B5: Full Chat history of CASE 14

Table B3: Comparison of SoT with Other Methods

Method A	Method B	A Win (%)	B Win (%)	Tie (%)	Significance
SoT	CoT	44.35	39.52	16.13	(ns)
SoT	**Majority Vote**	34.41	41.40	24.19	(**)
SoT	**MDAgents**	35.22	47.85	16.94	(*)
SoT	**Group Chat**	32.26	51.61	16.13	(***)

(*): $p < 0.05$, (**): $p < 0.01$, (***): $p < 0.001$, and (ns): non-significant results.

Table B4: Effect of conversation turn modifications on Group Chat performance

Method A	Method B	A Win (%)	B Win (%)	Tie (%)	
Group Chat 2-turn	**Group Chat 8-turn**	20.16	29.30	50.54	(**)
Group Chat 3-turn	Group Chat 8-turn	18.28	23.12	58.60	(ns)
Group Chat 4-turn	Group Chat 8-turn	20.70	10.48	68.82	(**)
Group Chat 6-turn	Group Chat 8-turn	11.29	10.75	77.96	(ns)
Group Chat 12-turn	Group Chat 8-turn	7.26	10.22	82.53	(na)

(*): $p < 0.05$, (**): $p < 0.01$, (***): $p < 0.001$, and (ns): non-significant results.

Table B5: Results using o4-mini

Method A	Method B	A Win (%)	B Win (%)	Tie (%)	
CoT	Majority Vote	22.31	18.01	59.68	(ns)
CoT	**MDAgents**	27.69	41.13	31.18	(**)
CoT	Group Chat	29.84	24.19	45.97	(ns)
Majority Vote	**MDAgents**	26.34	45.16	28.49	(***)
Majority Vote	Group Chat	25.00	26.61	48.39	(ns)
MDAgents	Group Chat	41.40	30.11	28.49	(**)
CoT	SoT	50.54	35.75	13.71	(**)

(*): $p < 0.05$, (**): $p < 0.01$, (***): $p < 0.001$, and (ns): non-significant results.

Table B6: Results using MedGemma-27B

Method A	**Method B**	**A Win (%)**	**B Win (%)**	**Tie (%)**	
CoT	Majority Vote	32.80	26.88	40.32	(ns)
CoT	MDAgents	35.22	30.65	34.14	(ns)
CoT	**Group Chat**	26.88	46.51	26.61	(**)
Majority Vote	MDAgents	30.11	35.48	34.41	(ns)
Majority Vote	**Group Chat**	28.23	47.85	23.92	(***)
MDAgents	Group Chat	30.91	41.94	27.15	(ns)
CoT	**SoT**	30.91	44.89	24.19	(**)

(*): $p < 0.05$, (**): $p < 0.01$, (***): $p < 0.001$, and (ns): non-significant results.

Table B7: Results using Gemma-27B

Method A	**Method B**	**A Win (%)**	**B Win (%)**	**Tie (%)**	
CoT	**Majority Vote**	32.26	42.47	25.27	(***)
CoT	**MDAgents**	29.03	49.73	21.24	(**)
CoT	**Group Chat**	34.95	45.97	19.09	(*)
Majority Vote	MDAgents	41.13	39.78	19.09	(ns)
Majority Vote	Group Chat	46.77	40.05	13.17	(ns)
MDAgents	Group Chat	45.70	43.28	11.02	(ns)
CoT	SoT	51.88	41.13	6.99	(*)

(*): $p < 0.05$, (**): $p < 0.01$, (***): $p < 0.001$, and (ns): non-significant results.

Table B8: Different but acceptable cases in human evaluation

Case	Group Chat Decision	MDT Decision	Category
1	The team decided to proceed with a lung nodule biopsy and pursue concurrent chemoradiotherapy for the laryngeal cancer, considering systemic therapy if metastasis is confirmed.	total laryngectomy with both selective neck dissection, lung RLL wedge resection	systemic therapy
2	The final decision is to proceed with a total laryngectomy with neck dissection, with consideration of postoperative systemic therapy options like PD-1 inhibitors.	neck dissection, revision cordectomy	systemic therapy
5	Proceed with surgical reevaluation, consider systemic therapy options, and evaluate eligibility for clinical trials, with supportive care as part of the treatment plan.	observation & regular follow up of larynx	systemic therapy
6	Adjuvant radiotherapy for tongue cancer and comprehensive systemic evaluation for plasmacytoma.	postoperative CCRT with weekly cisplatin	Different intensity of adjuvant treatment
7	The final decision is to pursue concurrent chemoradiation with a coordinated plan involving cardiology evaluation and possible adjustment of anticoagulation therapy.	postoperative radiation at tonsi site and neck	Different intensity of adjuvant treatment
8	The final decision from the meeting is to proceed with adjuvant radiotherapy for the patient.	postoperative CCRT with weekly cisplatin	Different intensity of adjuvant treatment
9	Proceed with adjuvant radiation therapy alone, pending confirmation of surgical margin status.	CCRT with weekly cisplatin	Different intensity of adjuvant treatment
10	The team will proceed with systemic therapy using immune checkpoint inhibitors, palliative radiation for symptom relief, consider stent placement for airway obstruction, and involve palliative care for symptom management.	palliative radiotherapy at neck	systemic therapy
11	The decision is to focus on personalized treatment strategies, considering systemic therapies, refined re-irradiation, and clinical trial options, to optimize outcomes while preserving quality of life.	re-radiation at larynx	systemic therapy
16	The final decision is to recommend adjuvant concurrent chemoradiotherapy based on the high-risk pathology features and potential microscopic residual disease.	observation with regular monitoring	Different intensity of adjuvant treatment
17	The final decision is to integrate systemic treatment adjustments using potential alternative therapies, manage bone metastases with bisphosphonates or RANKL inhibitors, and utilize targeted radiotherapy for neck and bone lesions as part of a comprehensive care strategy, informed by updated molecular profiling and aligned with the patient's goals and preferences.	Rt modified radical neck dissection	systemic therapy
19	The consensus is to pursue a multidisciplinary approach combining systemic chemotherapy with adjunctive radiotherapy for the lung and hypopharyngeal cancers, considering the patient's overall health and treatment preferences.	CCRT with weekly cisplatin for head and neck cancer after lung lobectomy	systemic therapy
21	Proceed with stereotactic body radiotherapy (SBRT) for the lung lesion and stereotactic radiosurgery (SRS) for the brain lesion, considering clinical trials for systemic therapy.	lung wedge resection	systemic therapy
24	The final decision is to prioritize systemic therapy and provide integrated palliative support, including regular follow-ups and multidisciplinary coordination.	lung metastasis w/u (PCNB)	systemic therapy
26	The team decided to focus on palliative systemic therapy, especially immunotherapy, with testing for PD-L1 to guide treatment, while involving palliative care for symptom management.	observation & palliative management	systemic therapy
28	Proceed with adjuvant radiotherapy targeting the surgical bed and regional lymphatics.	postoperative CCRT with weekly cisplatin	Different intensity of adjuvant treatment
29	The treatment plan will focus on systemic therapy, palliative radiation for symptom control, and early involvement of palliative care, while carefully considering further biopsies for molecular analysis.	brain metastasis Radiation	systemic therapy
31	Proceed with MRI of the renal lesion and initiate systemic therapy with radiation for metastatic squamous cell carcinoma.	postoperative CCRT with weekly cisplatin	systemic therapy