

---

# Silence is Not Consensus: Disrupting Agreement Bias in Multi-Agent LLMs via Catfish Agent for Clinical Decision Making

---

Yihan Wang<sup>1,\*</sup>, Qiao Yan<sup>1,\*</sup>, Zhenghao Xing<sup>1,\*†</sup>, Lihao Liu<sup>2,‡</sup>, Junjun He<sup>3</sup>,  
Chi-Wing Fu<sup>1</sup>, Xiaowei Hu<sup>3,§</sup>, and Pheng-Ann Heng<sup>1</sup>

<sup>1</sup>The Chinese University of Hong Kong  
<sup>2</sup>Amazon <sup>3</sup>Shanghai Artificial Intelligence Laboratory

## Abstract

Large language models (LLMs) have demonstrated strong potential in clinical question answering, with recent multi-agent frameworks further improving diagnostic accuracy via collaborative reasoning. However, we identify a recurring issue of **Silent Agreement**, where agents prematurely converge on diagnoses without sufficient critical analysis, particularly in complex or ambiguous cases. We present a new concept called **Catfish Agent**, a role-specialized LLM designed to inject structured dissent and counter silent agreement. Inspired by the “catfish effect” in organizational psychology, the Catfish Agent is designed to challenge emerging consensus to stimulate deeper reasoning. We formulate two mechanisms to encourage effective and context-aware interventions: (i) a complexity-aware intervention that modulates agent engagement based on case difficulty, and (ii) a tone-calibrated intervention articulated to balance critique and collaboration. Evaluations on nine medical Q&A and three medical VQA benchmarks show that our approach consistently outperforms both single- and multi-agent LLMs frameworks, including leading commercial models such as GPT-4o and DeepSeek-R1.

## 1 Introduction

“Without contraries is no progression.” — WILLIAM BLAKE

Progress often emerges not from agreement but from conflict, when ideas clash and debate arises, before a better solution is derived. This insight resonates deeply in collaborative reasoning.

Large Language Models (LLMs) have demonstrated strong potential in medical diagnosis by leveraging extensive clinical knowledge [38]. To enhance diagnostic robustness, recent work has proposed LLM-based multi-agent frameworks, where multiple specialized agents interact to simulate medical teamwork [40, 18, 42, 5]. By fostering diverse reasoning paths and encouraging dissent, these frameworks aim to improve decision quality, particularly in complex cases.

However, achieving effective collaboration among LLM agents for clinical decision making remains a significant challenge. In practice, we observe a phenomenon we call **Silent Agreement**, where a group of medical agents converge prematurely on the same diagnosis, without debate, evaluation, or exploration of alternatives. Figure 1 shows an example clinical misdiagnosis caused by Silent

---

\*These authors contributed equally.

†Project lead.

‡Co-corresponding author: lihaoliu@amazon.com

§Primary corresponding author: huxiaowei@pjlab.org.cn

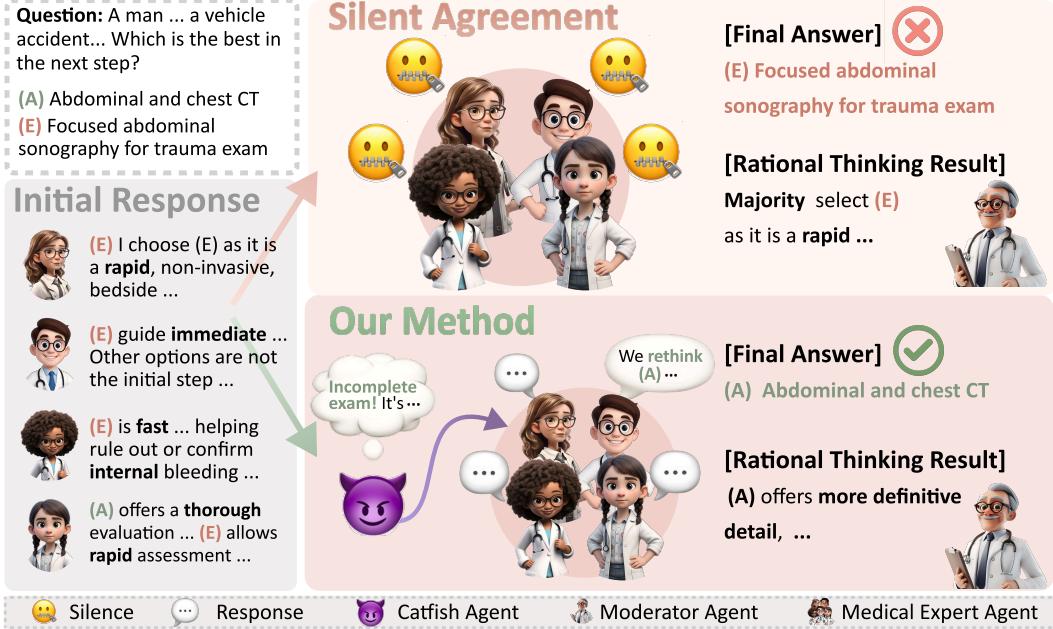


Figure 1: An example clinical misdiagnosis case resulted from **Silent Agreement**. Although the agents initially select different options, they remained silent in subsequent discussion, resulting in the misdiagnosis. Our method actively disrupts such silent agreement with the designated catfish agent in multi-agent collaborative reasoning and successfully produces the correct outcome.

Agreement. Although the agents initially propose different options, no further perspectives are offered, and all agents remain silent in the discussion, ultimately leading to an incorrect diagnosis.

Silent Agreement mirrors a classic failure mode of human groups, often called “groupthink”, where individuals suppress dissent and converge on a superficial consensus, often leading to sub-optimal or even dangerous decisions [14]. In contrast, social science research shows that constructive disagreement can enhance group performance, especially in high-stakes domains like medicine, by surfacing overlooked evidence and reducing errors [28, 26]. Structured dissent and open debate have also been linked to more robust scientific outcomes and deeper reasoning in collaborative settings [37]. Motivated by these findings, we investigate how dissent can mitigate premature consensus in multi-agent clinical decision making. We identify Silent Agreement as a critical bottleneck, highlighting the need for deeper reasoning by promoting regulated, constructive disagreement.

In this paper, we develop a new concept, namely **Catfish Agent**, which is designed to *actively disrupt silent agreement in multi-agent collaborative reasoning for clinical decision making*. Inspired by the “catfish effect”<sup>1</sup> and the “devil’s advocate” strategy in organizational psychology research [24, 27, 2], we propose to organize multi-agent reasoning as a multi-round, multi-role process.

Integrating dissent into medical agent groups poses two key challenges: (i) the level of required autonomy varies with case complexity, and (ii) overly assertive dissent can derail discussion or obscure key evidence. To address these issues, we formulate two core mechanisms in Catfish Agent: (i) *Complexity-aware intervention*, i.e., the agent adapts its engagement based on task difficulty, increasing autonomy in more complex cases to encourage deeper reasoning, and (ii) *Tone-calibrated intervention*, in which the strength and tone of dissent vary with the level of agent agreement, avoiding both passivity and excessive disruption. These novel mechanisms encourage the Catfish Agent to “break the silence,” while preserving productive collaboration.

Figure 1 shows an example case, where the Catfish Agent disrupts premature consensus by critically challenging the expert assumptions. This intervention prompts a revision of initial reasoning and enables the framework to synthesize a more reliable diagnosis. We evaluate our method on nine medical question-answering (Q&A) [15, 16, 34, 46, 19, 54, 12, 3] and three medical visual

<sup>1</sup>[https://en.wikipedia.org/wiki/Catfish\\_effect](https://en.wikipedia.org/wiki/Catfish_effect): The practice of placing a catfish in a tank of sardines to keep them alive. Without stimulation, sardines often become sluggish and suffocate to death due to lack of oxygen; however, the presence of a catfish keeps them constantly moving and thus alive.

question-answering (VQA) benchmarks [54, 52, 11], comparing it with both single-agent LLMs (*e.g.*, GPT-4o [1], DeepSeek-R1 [10], HuatuoGPT-o1 [4]) and multi-agent medical frameworks (*e.g.*, MedAgent [40], MDAgent [18]). Experimental results show that our method achieves a 12.73-point improvement on average, corresponding to a 39.2% relative gain over the best prior model, DeepSeek-R1, on the Q&A benchmarks, and a 5.33-point improvement on average, representing a 12.7% relative gain over the best prior method, MDAgent, on the VQA benchmarks. *We will release our code, experimental results, and logs.* Our contributions are threefold:

- We identify and formally define the **Silent Agreement** problem in LLM-based multi-agent frameworks for clinical decision making.
- We present the new concept **Catfish Agent**, the first to inject structured dissent into medical multi-agent systems, using the proposed *complexity-aware* and *tone-calibrated* interventions to break Silent Agreement and enhance collaborative clinical reasoning.
- We conduct extensive experiments on nine medical Q&A and three medical VQA benchmarks, demonstrating that our method largely outperforms state-of-the-art single- and multi-agent models.

## 2 Related Works

**Multi-Agent LLM for Medical Decision Making (MDM).** Recent studies have applied multi-agent LLM frameworks to collaborative tasks in planning, coding, and healthcare [53, 49, 18], typically assigning complementary roles to agents to support multi-turn coordination. AutoGen [49] formalizes inter-agent communication for iterative reasoning, while MDAgents [18] models medical teams with role-specialized agents for diagnostic support. However, most approaches emphasize cooperation over critique, assuming alignment leads to better decisions. In practice, we identify a critical failure mode called *Silent Agreement*, where agents prematurely converge on diagnoses without considering alternative hypotheses or resolving evidence conflicts. Recent works explore multi-agent debate [44, 22, 17, 43], but often overlook silent agreement.

**Large Language Models for Medical Reasoning.** LLMs have demonstrated growing capabilities in clinical reasoning, question answering, and medical summarization tasks [29, 23, 41]. Models like [1, 10, 4, 41, 30, 32, 20, 33, 36, 9, 7] have achieved strong performance on benchmarks like MedQA [15] and PubMedQA [16]. Recent work has begun exploring interaction-based improvements, including CoT prompting and collaborative diagnosis [21, 18]. Nevertheless, current methods rarely address the group-level dynamics of agreement or disagreement. To the best of our knowledge, this is the first work to identify and mitigate the silent agreement bias in LLM-based multi-agent frameworks for medical decision making through a new structured role-based intervention, *i.e.*, Catfish Agent.

## 3 The Silent Agreement Problem

We start this research work by carefully studying the prevalence and impact of Silent Agreement, a critical failure mode in multi-agent medical LLM frameworks, where agents converge on an answer, often incorrect, without sufficient deliberation or justification. This behavior undermines the intended collaborative nature of multi-agent reasoning and introduces risks in medical decision making.

To assess this issue, we analyze the *hard* set from MedAgentBench [39], focusing on two widely-used benchmarks: MedQA [15] and PubMedQA [16]. We evaluate two prominent multi-agent frameworks, MedAgents [40] and MDAgents [18], along with our proposed method. A *silent agreement failure* is defined as a diagnostic error, where agents produce a final answer without meaningful discussion, critique, or verification.

As shown in Table 1, MedAgents and MDAgents exhibit high silent rates, over 61.0% on both datasets, indicating frequent non-response or unjustified consensus. More critically, a large portion of their diagnostic failures are attributable to silent agreement: for MedAgents, 61.9% of failures on MedQA and 90.7% on PubMedQA; for MDAgents, 68.1% and 64.0%, respectively. These patterns confirm that silent agreement is not a rare anomaly but a dominant failure type in existing methods. Chi-squared tests [35] further confirm that silent agreement significantly impacts diagnostic accuracy

Table 1: Silent behavior analysis across MedQA and PubMedQA. “Silent Rate” denotes the proportion of questions, where agents arrive at a final answer with silent agreement. “Failure Attribution Rate” refers to the proportion of diagnostic failures that result from silent agreement. Our method achieves both the lowest silent rate and the lowest attribution to silent agreement failures.

(a) MedQA Dataset.			(b) PubMedQA Dataset.		
Method	Silent Rate ↓	Failure Attr. Rate ↓	Method	Silent Rate ↓	Failure Attr. Rate ↓
MedAgents	64.0%	61.9%	MedAgents	89.0%	90.7%
MDAgents	61.0%	68.1%	MDAgents	61.0%	64.0%
<b>Ours</b>	<b>17.0%</b>	<b>18.0%</b>	<b>Ours</b>	<b>11.0%</b>	<b>14.3%</b>

in both frameworks: MDAgents ( $\chi^2(1) = 5.345, p = 0.0208$ ) and MedAgents ( $\chi^2(1) = 5.896, p = 0.0152$ ), revealing a strong association between silent agreement and diagnostic failures.<sup>2</sup>

In contrast, our method achieves a significantly lower silent rate: 17.0% on MedQA and 11.0% on PubMedQA. Moreover, among the failures, our method makes only 18.0% and 14.3% involve silent agreement, substantially lower than those of MedAgents and MDAgents. This result indicates that our framework not only reduces unjustified silence but also encourages agents to engage in meaningful deliberation. Importantly, shifting away from silent behavior aligns with improved diagnostic accuracy, as our method outperforms existing multi-agent frameworks; see Table 3 for more details. The underlying mechanism is detailed in the next section.

## 4 Catfish Agent: Breaking Silent Agreement in LLM Teams

To address the Silent Agreement problem in LLM-based multi-agent clinical reasoning, we draw inspiration from organizational psychology, where structured disagreement has been shown to enhance epistemic vigilance and decision accuracy in human teams. Translating this principle to LLM-based teams introduces two key challenges: (i) *the necessary level of dissent varies with case complexity*, and (ii) *overly aggressive disagreement can derail discussion or obscure key evidence*.

To address these challenges, we design the Catfish Agent with two core mechanisms: (i) a *complexity-aware intervention strategy* that adapts the agent’s behavior to the difficulty of the clinical case, and (ii) a *tone-calibrated intervention mechanism* that adjusts the rhetorical strength of dissent based on the level of group agreement. The complexity-aware intervention strategy adaptively controls the Catfish Agent’s degree of autonomy based on task difficulty (basic, intermediate, advanced), as assessed by the Moderator. It governs *when* and *how much* the agent should intervene in potential groupthink. In contrast, the tone-calibrated intervention mechanism determines *how* the dissent is expressed, ensuring interventions are context-sensitive and constructively framed. Figure 2 illustrates the overall workflow of our framework involving the Catfish Agent, while Sections 4.1 and 4.2 present the details in the two core mechanisms.

### 4.1 Catfish in the Tiers: Stratifying Intervention by Complexity

Clinical tasks vary in complexity, with simple cases yielding quick consensus and complex cases requiring deeper reasoning. Therefore, this *complexity-aware intervention strategy* is proposed to dynamically adjusts the Catfish Agent’s behavior based on case difficulty. In complex or ambiguous scenarios, the Catfish Agent is granted a stronger sense of independent judgment. Conversely, in simpler cases, its interventions are more limited and guided. The following describes how the Catfish Agent’s behavior is progressively liberated according to different levels of case complexity.

**Basic cases.** For low-complexity clinical questions, the Moderator independently formulates an initial diagnosis  $D$ . The Catfish Agent then performs a lightweight critique, reviewing the reasoning behind  $D$  to identify any overlooked differentials or incomplete justification. If meaningful issues are detected, it generates a comment for the Moderator’s reference when making the final diagnosis  $F$ .

**Intermediate cases.** In these cases, the Moderator first recruits a group of expert agents  $A$ . Specifically, the Catfish Agent  $C$  is recruited alongside other agents, who is restricted to a predefined area of expertise due to its assigned role as determined by the Moderator. As shown in Figure 1, during

<sup>2</sup>Statistically significant at  $p < 0.05$

the group debate, the Catfish Agent is responsible for monitoring group dynamics and introducing dissent when appropriate.

The reasoning process unfolds over multiple rounds, each consisting of several turns. First, before the initial round begins, all recruited agents, excluding  $C$ , independently generate initial diagnoses  $D$ , which are sequentially shared for peer review. Second, in each round  $i$ , the process proceeds through a series of turns  $t$ . In each turn, agents sequentially evaluate the latest collective responses  $R^{i,t-1}$  and contribute updated outputs  $R^{i,t}$  based on their domain expertise. The first turn of the first round is grounded in  $D$ . Third,  $C$  monitors turn-level responses for logical inconsistencies, missed differentials, or weak justifications, aiming to detect emerging Silent Agreement. Upon identifying such issues, it injects domain-specific perturbations as assigned by the Moderator. Fourth, the rhetorical strength of these interventions (*e.g.*, mild, moderate, strong) is modulated based on the perceived degree of group convergence. Agents targeted by these challenges may revise their responses if they find the intervention sufficiently compelling. Last, at the end of each round, a Summary Agent compiles a structured report  $R^i$  that aggregates the group’s updated reasoning.

The reasoning process terminates under one of two conditions: (i) all agents reach consensus or fall into Silent Agreement, and two consecutive interventions by the Catfish Agent fail to provoke meaningful divergence; or (ii) the discussion reaches a predefined limit of  $n$  rounds, with each round allowing up to  $t$  interaction turns. Last, the Moderator reviews the interaction log and optionally consults intermediate summaries. The Catfish Agent may intervene if superficial consensus or diagnostic gaps emerge. The final decision  $F$  is made by the Moderator, integrating cross-round insights and catfish feedback with critical awareness.

**Advanced cases.** For high-complexity diagnostic tasks, the Moderator initiates a hierarchical team-of-teams structure. As illustrated in Figure 2, multiple teams  $t_j = \{a_1, a_2, \dots, a_m\}$  are recruited, each composed of domain-specific agents and led by a team leader. Within each team, the leader assigns sub-tasks to members, who collaborate to generate a structured report. Teams then take turns presenting their findings, allowing for cross-team comparison and integrative reasoning.

In the highest-complexity scenarios, the Catfish Agent  $C^*$  evolves into a free-roaming entity capable of independently initiating dissent with self-determined roles and perspectives. First, upon detecting Silent Agreement or conversational bias,  $C^*$  adopts an appropriate medical persona (*e.g.*, a Senior Gastrointestinal Oncologist with 20+ years of experience in colorectal malignancies) and issues context-sensitive challenges or counterarguments. Second, these interventions are integrated into the team’s workflow, resulting in updated reasoning traces. Third, the team’s structured report is then passed sequentially to the next team  $t_{j+1}$ , enabling cumulative refinement across teams. Unlike the fixed-role Catfish  $C$  in intermediate settings,  $C^*$  dynamically traverses teams, contributing from a broader, cross-domain perspective rather than being constrained to a predefined specialty. Last, once all  $m$  teams have completed their contributions, the Moderator synthesizes the aggregated insights and, if necessary, introduces a final Catfish Agent to conduct cross-team critique before producing the final diagnosis  $F$ . This process ensures both in-depth intra-team deliberation and cross-team integration before final decision making.

To sum up, benefiting from our *complexity-aware* intervention strategy, the Catfish Agent exhibits a progressively independent mindset and structural influence across basic, intermediate, and advanced cases, aligning with the increasing complexity of diagnostic scenarios.

## 4.2 Catfish in the Tone: Scaling Dissent by Consensus Strength

In collaborative diagnostic reasoning, the challenge is not only deciding *when* and *how much* to introduce dissent, but also *how* to express it, specifically, determining the appropriate degree of rhetorical intensity to ensure disagreement is constructive rather than disruptive. Early convergence among agents may indicate either well-founded consensus or premature closure. The core challenge lies in calibrating dissent: *if too weak, it may be ignored; if too strong, it risks disrupting constructive reasoning or derailing the diagnostic process.*

This tone-calibrated intervention mechanism is proposed to address this underexplored problem. It can allows the Catfish Agent dynamically adjusts the rhetorical intensity in response to the strength of inter-agent consensus. This mechanism enables context-sensitive dissent that preserves epistemic rigor without compromising team coherence. We introduce different tones of intervention as follows:



Figure 2: Overview of the reasoning process for an advanced case. (i) the system routes the clinical question through a complexity-aware Moderator, which classifies it as advanced and activates three expert teams, each consisting of a leader and two members; (ii) within each team, the leader assigns specific subtasks, and members respond independently based on their expertise; (iii) a Catfish Agent monitors the discussion and selectively intervenes by critiquing flawed assumptions or incomplete reasoning. All team members are required to respond to these challenges; (iv) after internal discussion, each team leader finalizes the team's answer and forwards it to the next team for iterative refinement; and (v) once all teams have contributed, the Moderator synthesizes the collective reasoning and, if needed, introduces an additional Catfish Agent for final diagnosis.

**Mild interventions.** This type of intervention targets scenarios in which agents begin to converge prematurely, relying on minimal or shallow evidence. In such cases, the Catfish Agent initiates reflective, non-confrontational prompts to gently encourage broader consideration. These interventions aim to stimulate metacognitive reflection without disrupting group cohesion.

**Intermediate interventions.** This level of intervention is intended for scenarios in which conclusions are formed without robust supporting evidence. The Catfish Agent probes with targeted questions, applying constructive pressure to surface hidden assumptions. This strategy draws inspiration from Socratic inquiry and diagnostic timeout protocols [13, 8] in the field of human clinical decision making.

**Strong interventions.** This level of intervention targets cases of rapid, uncritical agreement among agents. In such cases, the Catfish Agent delivers assertive challenges, explicitly questioning the group's reasoning to counteract conformity bias. This mode aligns with cognitive conflict theory and the devil's advocate paradigm.

Overall, this mechanism scales dissent according to consensus dynamics, ensuring interventions are epistemically productive and collaboration-preserving, while avoiding unnecessary disruption in simple scenarios and intensifying epistemic friction in high-stakes, complex cases.

Table 2: Comparison results on medical Q&A datasets. Accuracy (%) is shown per task, with the *hard* set categorized according to MedAgentsBench. Models are grouped into **general-domain** (upper block) and **medical-domain** (middle block) categories, followed by our method. **Best** results are in bold; second-best are underlined. Our method is based on the o3-mini model.

Method	Med QA	PubMed QA	Med MCQA	Med Bullets	MMLU	MMLU-Pro	MedEx QA	MedX pert-R	MedX pert-U
GPT-4o-mini [30]	22.0	10.0	17.0	10.1	12.3	11.0	4.0	6.0	5.0
GPT-4o [31]	32.0	9.0	25.0	19.1	24.7	21.0	18.0	7.0	6.0
DeepSeek-V3 [20]	16.0	12.0	19.0	13.5	15.1	12.0	7.0	6.0	9.0
o1-mini [32]	49.0	11.0	21.0	38.2	31.5	19.0	15.0	<u>29.0</u>	14.0
o3-mini [33]	<u>53.0</u>	<u>16.0</u>	24.0	<u>50.6</u>	35.6	15.0	18.0	25.0	15.0
QwQ-32B [36]	29.0	<u>16.0</u>	24.0	12.4	19.2	28.0	10.0	9.0	6.0
DeepSeek-R1 [10]	47.0	13.0	<u>31.0</u>	43.8	<u>43.8</u>	<u>37.0</u>	<u>26.0</u>	25.0	<u>26.0</u>
Llama-3.3-70B [9]	14.0	13.0	20.0	16.9	12.3	10.0	7.0	9.0	9.0
Claude-3.5-S [7]	18.0	10.0	10.0	9.0	16.4	14.0	13.0	9.0	11.0
Claude-3.5-H [7]	13.0	12.0	23.0	10.1	11.0	12.0	13.0	8.0	6.0
HuatuoGPT-o1 [4]	28.0	15.0	<u>31.0</u>	10.1	17.8	28.0	8.0	7.0	4.0
Baichuan-M1 [41]	20.0	13.0	22.0	14.6	15.1	21.0	9.0	6.0	5.0
<b>Ours</b>	<b>62.0</b>	<b>34.0</b>	<b>45.0</b>	<b>66.3</b>	<b>47.9</b>	<b>48.0</b>	<b>33.0</b>	<b>37.0</b>	<b>34.0</b>
<i>Improvements</i>	+9.0	+18.0	+14.0	+15.7	+4.1	+11.0	+7.0	+8.0	+8.0

## 5 Experimental Results

In this section, we evaluate the effectiveness of our proposed Catfish Agent Framework on MedAgentsBench [39], a benchmark designed to assess complex medical reasoning. MedAgentsBench is built from eight diverse medical Q&A datasets, including MedQA [15], PubMedQA [16], MedMCQA [34], MedBullets [3], MedExQA [19], and MedXpertQA [54]. Note that MedXpertQA consists of MedXpert-U and MedXpert-R, with each subset focusing on understanding and reasoning. It also integrates six medical tasks from MMLU [12] and MMLU-Pro [46]. Based on performance and reasoning depth, challenging “hard” subsets are selected. For a fair comparison, we follow the standardized evaluation protocol and use the officially results reported by MedAgentsBench [39].

**Implementation details.** All experiments are conducted via the OpenAI API<sup>3</sup> in a strict zero-shot setting, without any fine-tuning or gradient updates. Each agent, the Moderator, Catfish Agent, and domain-specific Experts, is instantiated through separate API calls, with roles defined by structured prompts incorporating system instructions and dialogue history. Default API parameters (`temperature`, `top_p`) are employed, with no explicit constraint on `max_tokens`.

### 5.1 Comparison with General and Medical Large Models

We compare our method using o3-mini [33] as the base model for each agent with the state-of-the-art general large models, including GPT-4o-mini [30], GPT-4o [31], DeepSeek-V3 [20], o1-mini [32], o3-mini [33], QwQ-32B [36], DeepSeek-R1 [10], Llama-3.3-70B [9], Claude-3.5-S [7], and Claude-3.5-H [7], as well as specific medical models, including HuatuoGPT-o1 [4] and Baichuan-M1 [41]. *Notably, we are the first to integrate a reasoning model, namely o3-mini, into a multi-agent framework for medical decision making.*

Table 2 presents the performance of our approach on MedAgentsBench, in comparison with a broad range of general-purpose and medical-specialized large language models, focusing on the “hard” subsets requiring demand deeper reasoning. Our method consistently achieves state-of-the-art performance across all benchmarks, surpassing the second-best model by a substantial margin, *yielding an average 12.7-point absolute gain, corresponding to a 39.2% relative improvement<sup>4</sup> over the best prior model, DeepSeek-R1.*

Additionally, we have the following observations. (i) **Our method is the first to integrate CoT-style reasoning into a structured multi-agent framework that supports multi-turn deliberation under**

<sup>3</sup><https://platform.openai.com>

<sup>4</sup>The overall average accuracy improvement is computed by first averaging accuracy across all tasks and then calculating the relative gain over DeepSeek-R1, resulting in a 39.2% improvement.

Table 3: Comparison results on medical Q&A datasets. All tasks are evaluated on the *hard* set, with accuracy reported in percentage (%). Two base models are used: GPT-4o-mini and GPT-4o. **Best results** are in bold; second-best are underlined. Methods are grouped into four categories (Baseline-Prompting, Advanced-Prompting, Search-Agent, Multi-Agent).

Method	Med QA		PubMed QA		Med MCQA		Med Bullets		MMLU		MMLU -Pro		MedEx QA		Med Xpert-R		Med Xpert-U	
	40-m		40		40-m		40		40-m		40		40-m		40		40-m	
Zero-shot	22.0	32.0	10.0	9.0	17.0	25.0	10.1	19.1	12.3	24.7	11.0	21.0	4.0	18.0	6.0	7.0	5.0	6.0
Few-shot	30.0	28.0	22.0	20.0	31.0	29.0	<u>23.6</u>	23.6	<b>28.8</b>	27.4	10.0	9.0	<u>25.0</u>	<b>24.0</b>	<b>16.0</b>	14.0	8.0	11.0
CoT [48]	21.0	39.0	13.0	10.0	26.0	30.0	18.0	28.1	<b>28.8</b>	26.0	<u>35.0</u>	35.0	14.0	<b>24.0</b>	6.0	12.0	10.0	15.0
CoT-SC [45]	20.0	37.0	11.0	6.0	20.0	<b>35.0</b>	16.9	30.3	<b>28.8</b>	30.1	34.0	<u>43.0</u>	19.0	<u>22.0</u>	10.0	10.0	<u>13.0</u>	14.0
MultiPersona [47]	29.0	45.0	13.0	15.0	21.0	25.0	15.7	29.2	26.0	<u>37.0</u>	<b>36.0</b>	42.0	17.0	21.0	7.0	10.0	12.0	16.0
Self-Refine [25]	<b>32.0</b>	41.0	12.0	13.0	24.0	34.0	15.7	28.1	<u>27.4</u>	34.2	31.0	34.0	16.0	<u>22.0</u>	7.0	<u>17.0</u>	12.0	19.0
MedPrompt [6]	29.0	34.0	14.0	11.0	30.0	26.0	13.5	22.5	20.5	26.0	34.0	22.0	18.0	16.0	6.0	14.0	<u>13.0</u>	9.0
SPO [50]	19.0	31.0	<u>25.0</u>	<u>31.0</u>	20.0	30.0	22.5	29.2	19.2	32.9	32.0	36.0	14.0	19.0	11.0	15.0	11.0	16.0
AFlow [51]	<u>30.0</u>	<u>48.0</u>	15.0	18.0	25.0	31.0	15.7	<b>34.8</b>	24.7	<b>38.4</b>	29.0	37.0	7.0	22.0	7.0	13.0	7.0	<u>18.0</u>
MedAgents [40]	24.0	43.0	12.0	15.0	22.0	30.0	15.7	27.0	24.7	28.8	3.0	8.0	12.0	19.0	4.0	3.0	5.0	6.0
MDAgents [18]	22.0	36.0	23.0	11.0	16.0	22.0	14.6	21.3	17.8	24.7	9.0	8.0	10.0	13.0	8.0	4.0	9.0	5.0
<b>Ours</b>	<b>32.0</b>	<b>50.0</b>	<b>35.0</b>	<b>37.0</b>	<b>31.0</b>	<u>34.0</u>	<b>25.8</b>	<u>31.5</u>	26.0	28.8	32.0	<b>50.0</b>	<b>26.0</b>	<b>24.0</b>	<u>14.0</u>	<b>21.0</b>	<b>14.0</b>	<b>19.0</b>

**complex clinical conditions.** By embedding CoT reasoning into each agent’s decision process and introducing structured dissent via a Catfish Agent, our system not only improves diagnostic accuracy but also offers a novel paradigm for modeling disagreement, iterative reasoning, and collaboration, which are the key characteristics of expert clinical teams. (ii) **Reasoning-based LLMs substantially outperform standard LLMs across all medical benchmarks.** For example, o3-mini and DeepSeek-R1 achieve significantly higher accuracy than their non-reasoning counterparts, indicating that explicit intermediate reasoning steps, such as Chain-of-Thought (CoT), are highly effective in complex medical tasks. (iii) **General-purpose reasoning models consistently outperform domain-specialized medical LLMs.** Despite lacking medical-specific pretraining, models like o3-mini and DeepSeek-R1 surpass medical-tuned models such as HuatuoGPT-o1 and Baichuan-M1 across nearly all datasets. This suggests that broad reasoning capabilities provide greater benefits than narrow domain knowledge, especially in high-level diagnostic tasks such as MedXpertQA.

## 5.2 Comparison with Multi-Agent, Prompting, and Search-Agent Methods

We compare our method with three strategies: (i) prior multi-agent LLM frameworks (MedAgents, MDAgents), (ii) prompting-based methods (baseline-prompting and advanced-prompting), and (iii) search-agent systems, across eight challenging medical Q&A datasets under both GPT-4o-mini and GPT-4o settings.

As shown in Table 3, our method achieves state-of-the-art accuracy on most datasets, outperforming all other methods in 12 of the 18 evaluation columns. We summarize our findings as follows: (i) **Robust gains across all datasets.** Our method outperforms prior multi-agent approaches in all 18 comparisons (nine datasets  $\times$  two base models), demonstrating strong generalization across diverse tasks and domains. (ii) **Superior reasoning under limited model capacity.** On GPT-4o-mini, our method surpasses all previous multi-agent methods by a significant margin. For instance, on MMLU-Pro (4o-mini), we achieve 32.0%, far exceeding MedAgents (3.0%) and MDAgents (9.0%), highlighting the effectiveness of our disturbance-enhanced collaboration even with weaker backbones. (iii) **Bridging the multi-agent performance gap.** Multi-agent frameworks typically struggle on benchmarks such as MMLU-Pro and MedXpert-R, with prior methods (*e.g.*, MDAgent) achieving only 8.0% on MMLU-Pro (4o) and 4.0% on MedXpert-R (4o). In contrast, our method achieves 50.0% and 21.0% respectively, outperforming all agent-based baselines and matching or exceeding strong prompting and search-based alternatives. (iv) **Limits of multi-agent methods on simpler cases.** On the MMLU subset, all

Table 4: Comparison of three medical VQA datasets requiring image-text reasoning. Our method consistently outperforms GPT-4o and multi-agent baselines.

Method	MedXpert-MM	PMC-VQA	Path-VQA
GPT-4o	24.0%	32.0%	42.0%
MedAgents	24.0%	42.0%	48.0%
MDAgents	28.0%	54.0%	44.0%
<b>Ours</b>	<b>34.0%</b>	<b>58.0%</b>	<b>50.0%</b>

Table 5: Ablation study on the Catfish Agent’s placement and tone design.

Configuration	Accuracy ↑	Intermediate Cases	Silent Rate ↓	Non-Silent Accuracy ↑
w/o Catfish	36.0%	34	61.8% (21/34)	38.5% (5/13)
w/ Catfish in Moderator only	39.0%	33	51.5% (17/33)	37.5% (6/16)
w/ Catfish in Team only	44.0%	30	33.3% (10/30)	50.0% (10/20)
w/ Catfish (no Tone Design)	45.0%	43	23.3% (10/43)	45.5% (15/33)
<b>w/ Catfish (Full Design)</b>	<b>50.0%</b>	<b>35</b>	<b>17.1% (6/35)</b>	<b>55.2% (16/29)</b>

multi-agent methods show suboptimal performance due to the simplicity of many test cases. Most samples in this subset are basic queries that require limited reasoning, reducing the benefits of agent collaboration.

### 5.3 Comparison on Medical Visual Question Answer Tasks

To assess generalization beyond text-based Q&A, we evaluate our method on medical VQA tasks requiring joint reasoning over clinical images and text. Experiments are conducted on three datasets: MedXpert-MM [54], PMC-VQA [52], and PathVQA [11]. For each, we select 50 samples balanced by complexity: 12 basic, 25 intermediate, and 13 advanced cases, enabling comprehensive evaluation across difficulty levels.

As shown in Table 4, our method outperforms both the GPT-4o baseline (the base model of our agents) and prior multi-agent frameworks across all benchmarks. On MedXpert-MM, it achieves 34%, exceeding MedAgents and MDAgents by ten and six points, respectively. Similar improvements have been seen on PMC-VQA and Path-VQA. These gains demonstrate our method’s superior multimodal reasoning, particularly under visual ambiguity.

### 5.4 Ablation Study

To assess the contribution of each component in our framework, we conduct ablation studies on the MedQA dataset using GPT-4o as the base agent model. Following our earlier categorization, we focus on intermediate questions, where silent agreement behavior is most likely to occur, unlike basic cases (single-agent) and advanced ones (forced responses). For each setting, we report: (i) the number of intermediate cases, (ii) overall accuracy, (iii) silent agreement rate, which is defined as the proportion of intermediate cases with no agent response, and (iv) accuracy on non-silent intermediate cases. All silence-related metrics are computed exclusively within the intermediate subset, where such dynamics are more observable.

**Placement of the Catfish Agent.** We investigate the effectiveness of Catfish Agent placements through four configurations: (i) no Catfish Agent (baseline), (ii) embedded in the Moderator only, (iii) embedded in the Team only, and (iv) embedded in both Moderator and Team (our full configuration). As shown in Table 5, introducing the Catfish Agent in either location reduces Silent Agreement and improves accuracy. The best results are achieved when the Catfish Agent is placed in both roles, yielding the highest non-silent accuracy (55.2%) and the lowest silent rate (17.1%). These results underscore the complementary value of combining top-down (Moderator) and peer-level (Team) interventions to mitigate silent consensus and foster deeper discussion.

**Tone of the Catfish Agent.** We investigate whether the Catfish Agent’s tone impacts its effectiveness by comparing a neutral variant with a strategically challenging one, as detailed in Sec. 4.2, while keeping the agent embedded in both the Team and Moderator. As shown in Table 5, the use of deliberate tone strategies yields higher overall accuracy (50% vs. 45%), reduces the Silent Agreement rate (17.1% vs. 23.3%), and improves non-silent case accuracy (55.2% vs. 45.5%). This highlights tone modulation as a key factor in disrupting premature consensus and encouraging active discussion.

## 6 Case Study

### 6.1 Advanced Case

As illustrated in Figure 3, the diagnostic process begins with multiple specialized teams, each producing structured reports through intra-team collaboration. The Catfish Agent  $C^*$  monitors interactions and dynamically intervenes when Silent Agreement or conversational bias is detected. It

selects an expert role it considers most appropriate (e.g., nephrologist in Figure 3) to raise challenges or provide counterpoints. These interventions are addressed by the team and incorporated into the reasoning trace. The updated report is then passed to the next team for further refinement. After all teams contribute, the Moderator aggregates the insights and, if needed, the Catfish Agent performs a final cross-team critique before the Moderator issues the final decision.

**Question:** A 43-year-old woman with obesity is being assessed for stage 3B chronic kidney disease. She has a 26-year history of type 1 diabetes managed with insulin and hypertension treated with hydrochlorothiazide 25 mg daily and lisinopril 40 mg daily. Her blood pressure is currently 140/84 mm Hg. Laboratory results show a serum creatinine level of 1.7 mg/dL (reference range, 0.6–1.1) and a serum potassium level of 4.9 mEq/L (3.5–5.0). Proteinuria is confirmed with an albumin-to-creatinine ratio of 760 mg/g (<30). Which of the following management strategies is the most appropriate for this patient? **Options:** (A) Replace hydrochlorothiazide with dapagliflozin; (B) Add hydralazine to current therapy; (C) Add losartan to current therapy; (D) Increase lisinopril dosage beyond 40 mg daily; (E) Add metoprolol to current therapy; (F) Replace hydrochlorothiazide with canagliflozin; (G) Add amlodipine to current therapy; (H) Replace lisinopril with spironolactone; (I) Replace hydrochlorothiazide with furosemide; (J) Initiate sodium bicarbonate therapy.

## 6.2 Intermediate Case

As shown in Figure 4, a group of expert agents is first recruited, including the Catfish Agent, which is assigned a specific medical role by the Moderator. Each expert independently provides an initial judgment and proposes a preliminary diagnosis. This is followed by multiple rounds of structured debate. During each round, agents sequentially review the initial diagnosis report and are invited to join the discussion by contributing their own perspectives if they disagree or have additional insights.

Throughout the process, the Catfish Agent monitors for signs of Silent Agreement, overlooked differentials, insufficient justifications, and logical inconsistencies. When such issues arise, the Catfish Agent injects domain-specific challenges calibrated to the group's level of convergence. Targeted agents may revise their responses if they find the critique valid. After each round, a Summary Agent compiles an updated diagnostic report reflecting the latest viewpoints.

The discussion process terminates once consensus is reached, no substantial divergence follows Catfish interventions, or a predefined round limit is met. Finally, the Moderator makes the ultimate diagnostic decision, optionally consulting the Catfish Agent for additional critique before finalizing the output.

**Question:** A 24-year-old woman, 8 weeks pregnant, attends her first prenatal visit. She recently immigrated from Africa, has no vaccination records, and works as a babysitter with recent exposure to children with flu and chickenpox. She reports only mild fatigue and nausea. Vitals and physical exam are normal. Which vaccine should she receive now?

**Options:** (A) Tetanus/Diphtheria/Pertussis vaccine; (B) Rabies vaccine; (C) Measles/Mumps/Rubella vaccine; (D) Live-attenuated influenza vaccine; (E) Varicella vaccine; (F) Human papillomavirus vaccine; (G) Intramuscular flu vaccine; (H) Pneumococcal conjugate vaccine; (I) Hepatitis A vaccine; (J) Hepatitis B vaccine.

## 6.3 Basic Case

As shown in Figure 5, this case demonstrates a successful intervention in a basic-complexity question. The Moderator initially provides an incorrect diagnosis *D*, but the Catfish Agent identifies a flaw in the reasoning and offers a concise critique. With this feedback, the Moderator revises the decision and ultimately makes the correct final diagnosis *F*.

## 6.4 Visual Question Answering Case

Figure 6 presents a successful example where the Catfish Agent injects a targeted perturbation during visual question answering, prompting domain experts to reconsider their initial conclusion. This leads to a course correction and ultimately results in the correct diagnosis.

**Initial Assessment Team Leader Task Delivery**

**To Endocrinologist:**

1. HbA1c Test: Assess glycemic control in managing the patient's type 1 diabetes.
2. Kidney Function Tests: Evaluate nephropathy progression.
3. Review of Diabetes Therapy: Consider **Option A or F**.

**Initial Assessment Team Members Investigations**

1. HbA1c: Target <7% for glycemic, renal, and CV protection.
2. Kidney Tests: Track CKD via creatinine, eGFR, and UACR.
3. Therapy Review: Consider SGLT2 inhibitors (**Option A** or **Option F**) for added heart-kidney benefit.

**Catfish Agent's Decision to Intervene**

 Catfish Agent: I Should intervene!

 My Identity: Nephrologist specializing in diabetic kidney disease and blood pressure management.

 My Tone should be: **Moderate**

Catfish Agent's Opinion : While **Option A** or **F** offers renal and cardiovascular benefits, it **does not directly address the suboptimal BP of 140/84 mm Hg**. Additionally, adding another agent affecting kidney function requires caution due to the hyperkalemia risk with current ACE therapy.

Catfish Agent's Suggest Answer: (G) Add amlodipine to current therapy. Introducing a calcium channel blocker like amlodipine can help achieve better blood pressure control, which is crucial given the patient's stage 3B CKD. This can be done while considering an additional SGLT2 inhibitor once blood pressure is better managed and potassium levels are stable.

Challenge to Group: How do you plan to promptly and effectively achieve the target blood pressure of <130/80 mm Hg without further risking renal function or exacerbating the already elevated potassium levels, given the current therapeutic options and patient profile?

**Initial Assessment Team Members Response to the Challenge**

In light of nephrology's concerns, a revised, staged approach is proposed:

1. Add Amlodipine: **Effectively lowers BP and is safe with ACE inhibitors.**
2. Delay SGLT2 Inhibitor: Consider **Option A or F after BP is controlled and potassium is stable.**
3. Prevent Hyperkalemia: Monitor potassium regularly; consider dietary restriction or binders.

**Initial Assessment Team Leader Final Decision**

(G) Add amlodipine to current therapy

This approach focuses on achieving better blood pressure control, which is crucial for managing stage 3B CKD and cardiovascular risk in this patient. Adding amlodipine, a calcium channel blocker, addresses the need for improved hypertension management without increasing the risk of hyperkalemia, which is a concern given her current medications. Once the blood pressure is better controlled, further consideration can be given to incorporating an SGLT2 inhibitor to provide added renal protection and cardiovascular benefits.

• • • • •

**Final Review and Decision Team Leader Task Delivery**

**Clinical Decision Specialist Investigations:**

1. Blood Pressure Control: Evaluate current antihypertensive regimen's effectiveness.
2. Proteinuria Management: Assess suitability of SGLT2 inhibitors for renal protection.

**Final Review and Decision Team Members Investigations**

1. Antihypertensive Strategy: Assess current regimen vs. BP target (<130/80 mm Hg); consider adding CCBs or SGLT2 inhibitors.
2. Proteinuria Management: Evaluate SGLT2 inhibitors; Optimize proteinuria control.

 Catfish Agent: I Don't want to intervene!

**(A) Replace hydrochlorothiazide with dapagliflozin**

Replacing hydrochlorothiazide with dapagliflozin is a strategic choice given the patient's profile. Dapagliflozin not only provides significant renal protection and reduces proteinuria—especially beneficial in diabetic nephropathy—but also offers modest blood pressure lowering and cardiovascular benefits. While it may not fully replace the antihypertensive effect of hydrochlorothiazide, it contributes to overall risk reduction. Given the patient's current renal function, dapagliflozin is appropriate, though close monitoring of eGFR and potassium levels is essential to ensure safety and efficacy. Further antihypertensive adjustments may be required if blood pressure targets are not met.

**(A) Replace hydrochlorothiazide with dapagliflozin**

**(A)** is the best, ... in reducing proteinuria and enhancing renal outcomes... **(A)** also supports glycemic control and provides cardiovascular protection... Although its antihypertensive effect...

**Advanced Diagnostic Support Investigations:**

1. Renal Function: Track eGFR and serum creatinine to guide safe medication use, especially SGLT2 inhibitors.
2. Electrolyte & Metabolic Monitoring: Monitor potassium; screen for metabolic issues requiring interventions.

**(A) increases the the risk of hyperkalemia;** it also may reduce overall diuretic effect, potentially compromising blood pressure and fluid management. **(A) without first stabilizing BP and electrolytes...**

**(G) Add amlodipine to current therapy**

The Catfish Agent highlights important considerations. ...hyperkalemia risk, the need for immediate blood pressure control, and the role of diuretics in stage 3B CKD.

Figure 3: Advanced case example. Interventions from the Catfish Agent leads to a correct decision. Upon detecting premature consensus and inaccurate analysis, the Catfish Agent (as a nephrologist) raises specific concerns, prompting Teams and the Moderator to re-evaluate and ultimately select the correct option.

### Initial Report Generation

 **(A) Tetanus/Diphtheria/Pertussis vaccine.** In pregnancy, (A) is recommended during each gestation to protect both mother and infant, ideally between weeks 27–36, though it can be given at any time if not previously received. This is especially important given her unknown immunization status and recent exposure to potentially infectious children.

 **(A) Tetanus/Diphtheria/Pertussis vaccine.** During pregnancy, (A) is recommended to protect the newborn from pertussis and should be given between 27–36 weeks. Given her likely unimmunized status, initiating tetanus and diphtheria protection now is prudent. Live vaccines (e.g., MMR, varicella, LAIV) are contraindicated, while other vaccines may be considered based on risk, though Tdap remains the prenatal priority.

 **(G) Intramuscular flu vaccine.** (G) is recommended during pregnancy to protect against influenza-related complications and is safe at any stage. Given her occupational exposure, the intramuscular flu vaccine is the most appropriate choice. Live vaccines (e.g., varicella, MMR) remain contraindicated.

 **(G) Intramuscular flu vaccine.** (G) is safe and recommended during pregnancy, as seasonal influenza can cause severe illness. Live-virus vaccines (e.g., MMR, Varicella) are contraindicated. Given her unknown vaccination history and exposure risk, influenza protection is a priority.

 Given the opinions from your team, would you like to talk to other expert?

### Debate Stage

 scanning!



Given the patient's unknown status and recent child exposure, I recommend **Tdap** and inactivated influenza vaccines. **Tdap** is ideally given at 27–36 weeks for pertussis, but early administration can provide baseline protection. Influenza coverage is equally important due to her exposure risk.



Although (G) is essential to address seasonal risks, while early **Tdap** initiation may offer added protection against tetanus and diphtheria. Administering the flu shot now aligns with public health goals, with **Tdap** planning maintained for later in pregnancy.



**(A) Tetanus/Diphtheria/Pertussis vaccine.** Given the patient's unknown immunization history and her exposure to children, the **Tdap vaccine is crucial**. It is recommended for pregnant women to protect both the mother and newborn against **pertussis**, which can be severe in infants. While the influenza vaccine is also important, the prioritization of Tdap is due to its **specific protective benefit** against pertussis, which can have serious consequences for the newborn. **This aligns with standard prenatal care practices** of administering Tdap between 27 and 36 weeks of pregnancy but starting the vaccine series early is prudent in cases of unknown immunization statuses.



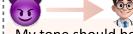
Considering this patient's exposure to children recently having chickenpox and the flu, my immediate priority is (G). Influenza poses a **significant risk during pregnancy**, potentially leading to severe complications. (G) is **safe and effective at any stage**, providing essential protection against seasonal influenza. While the **Tdap** vaccine is crucial for infant protection against pertussis, it is typically **administered later** in the pregnancy around weeks 27–36. Starting with the flu vaccine not only **safeguards** her health immediately but also **aligns with standard prenatal care** recommendations for maternal and fetal well-being. Let's assess her vaccination history further to ensure comprehensive protection against other preventable diseases, prioritizing those safe for pregnancy.

### Catfish Agent Interventions



My tone should be **Moderate** !

While recommending **Tdap** is appropriate, the patient's international background and unknown immunization status warrant broader consideration. Her recent exposure to children suggests **possible unmet vaccine needs**. A full review of her vaccination history and risk factors is essential to identify additional safe, pregnancy-appropriate immunizations. Vaccine timing should also align with current evidence and individualized risk, which appears **underexplored** here.



My tone should be **Moderate** !

While prioritizing **Tdap** for pertussis protection is important, your rationale **overlooks other vaccine-preventable diseases** relevant to the patient's **unknown immunization status**. Starting tetanus and diphtheria vaccines without addressing the implications of being unimmunized needs further clarification. Additionally, excluding all live vaccines may ignore **emerging evidence on their nuanced risks in pregnancy**. A more tailored evaluation of her exposure risks and **safe vaccine options** is needed.



My tone should be **Mild** !

Your emphasis on administering the **intramuscular flu vaccine** is certainly **valid** and **aligns well** with guidelines, especially considering the patient's exposure through work. However, her **unclear vaccination history** and **recent contact with potentially infected children** raise the question of whether additional non-live vaccines should be considered. Evaluating these options alongside the flu vaccine could strengthen maternal and fetal protection and **improve the overall immunization strategy**.

•••••

### Other Debate Stage, Catfish Agent Interventions and Final Decision Stage

Figure 4: *Intermediate* case example illustrating interventions from the Catfish Agent during a multi-round debate. Assigned a fixed domain role, the Catfish Agent monitors team dynamics and raises structured dissent to prevent Silent Agreement, enhancing diagnostic robustness.

## 6.5 Comparison with DeepSeek-R1

To further demonstrate the effectiveness of the Catfish Agent, we compare our framework with the strongest baseline model, DeepSeek-R1, using the same clinical question, as shown in Figure 7 and Figure 8.

In Figure 7, DeepSeek-R1 conducts a thorough analysis of all available options, yet ultimately fails to select the correct answer. Moreover, it redundantly repeats reasoning patterns across options without meaningful refinement.

In contrast, Figure 8 illustrates the final decision stage of our framework’s response to the same question. Despite initial incorrect diagnosis among the expert groups, the Catfish Agent identifies a critical flaw and proposes an effective alternative. This intervention successfully prompts the Moderator to revise the initial judgment and reach the correct final decision.

## 6.6 Failure Case

Figure 9 illustrates a failure case in a basic-complexity question. Despite the Catfish Agent proposing alternative diagnoses and constructively challenging the initial reasoning, the Moderator adheres to the original answer without sufficient reconsideration. This ultimately results in an incorrect final decision, underscoring that the Catfish’s interventions, while helpful, can still be overridden in rigid decision-making scenarios.

## 7 Conclusion

We identify **Silent Agreement** as a critical failure mode in multi-agent LLM systems for clinical decision making, where agents prematurely converge on diagnoses without sufficient critical analysis. To address this, we present the new concept called **Catfish Agent**, a structured dissent mechanism collaborative reasoning through dynamic, round-based interventions. By these new means, we encourage deeper justification, broader hypothesis exploration, and more robust diagnostics, supported by the proposed *complexity-aware* intervention strategy and *tone-calibrated* intervention mechanism. Experiments on nine public medical Q&A datasets and three public medical VQA datasets show substantial performance improvements. In the future, we plan to investigate efficient coordination strategies that maintain reasoning depth while reducing the inference-time overhead.

## References

- [1] 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.
- [2] Muqtafi Akhmad, Shuang Chang, and Hiroshi Deguchi. Closed-mindedness and insulation in groupthink: their effects and the devil’s advocacy as a preventive measure. *Journal of Computational Social Science*, 4:455–478, 2021.
- [3] Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. Benchmarking large language models on answering and explaining challenging medical questions. *arXiv preprint arXiv:2402.18060*, 2024.
- [4] Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. HuatuoGPT-o1: towards medical complex reasoning with ILMs. *arXiv preprint arXiv:2412.18925*, 2024.
- [5] Xi Chen, Huahui Yi, Mingke You, WeiZhi Liu, Li Wang, Hairui Li, Xue Zhang, Yingman Guo, Lei Fan, Gang Chen, et al. Enhancing diagnostic capability with multi-agents conversational large language models. *NPJ digital medicine*, 8(1):159, 2025.
- [6] Xuhang Chen, Shenghong Luo, Chi-Man Pun, and Shuqiang Wang. MedPrompt: Cross-modal prompting for multi-task medical image translation. In *Chinese Conference on Pattern Recognition and Computer Vision*, pages 61–75, 2024.
- [7] Claude. Claude 3.5 Sonnet, 2024. <https://www.anthropic.com/news/clause-3-5-sonnet>.
- [8] Mark L Graber, Stephanie Kissam, Velma L Payne, Ashley ND Meyer, Asta Sorensen, Nancy Lenfestey, Elizabeth Tant, Kerm Henriksen, Kenneth LaBresh, and Hardeep Singh. Cognitive interventions to reduce diagnostic error: a narrative review. *BMJ quality & safety*, 21(7):535–557, 2012.
- [9] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

- [10] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [11] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. PathVQA: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.
- [12] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.
- [13] Yueh-Ren Ho, Bao-Yu Chen, and Chien-Ming Li. Thinking more wisely: using the socratic method to develop critical thinking skills amongst healthcare students. *BMC medical education*, 23(1):173, 2023.
- [14] Irving Lester Janis and Irving Lester Janis. *Groupthink: Psychological studies of policy decisions and fiascoes*, volume 349. Houghton Mifflin Boston, 1982.
- [15] Di Jin, Eileen Pan, Nassim Oufattolle, 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.
- [16] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, 2019.
- [17] Alex Kim, Keonwoo Kim, and Sangwon Yoon. DEBATE: Devil’s advocate-based assessment and text evaluation. *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.
- [18] Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, Hae Park, et al. MDAgents: An adaptive collaboration of LLMs for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452, 2024.
- [19] Yunsoo Kim, Jinge Wu, Yusuf Abdulle, and Honghan Wu. MedExQA: Medical question answering benchmark with multiple explanations. *arXiv preprint arXiv:2406.06331*, 2024.
- [20] 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.
- [21] Jiaxiang Liu, Yuan Wang, Jiawei Du, Joey Tianyi Zhou, and Zuozhu Liu. MedCoT: Medical chain of thought via hierarchical expert. *arXiv preprint arXiv:2412.13736*, 2024.
- [22] Yexiang Liu, Jie Cao, Zekun Li, Ran He, and Tieniu Tan. Breaking mental set to improve reasoning through diverse multi-agent debate. In *International Conference on Learning Representations*, 2025.
- [23] Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Massimo Hong, Yushuai Wu, Mu Qiao, and Zaiqing Nie. BioMedGPT: An open multimodal large language model for biomedicine. *IEEE Journal of Biomedical and Health Informatics*, 2024.
- [24] Colin MacDougall and Frances Baum. The Devil’s Advocate: A strategy to avoid groupthink and stimulate discussion in focus groups. *Qualitative health research*, 7(4):532–541, 1997.
- [25] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-Refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023.
- [26] Hugo Mercier and Dan Sperber. *The enigma of reason*. Harvard University Press, 2017.
- [27] Charlan Nemeth, Keith Brown, and John Rogers. Devil’s advocate versus authentic dissent: Stimulating quantity and quality. *European Journal of Social Psychology*, 31(6):707–720, 2001.
- [28] Charlan Jeanne Nemeth. Dissent as driving cognition, attitudes, and judgments. *Social Cognition*, 13(3):273–291, 1995.
- [29] 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.
- [30] OpenAI. GPT-4o mini: advancing cost-efficient intelligence, 2024. <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>.
- [31] OpenAI. Hello GPT-4o, 2024. <https://openai.com/index/hello-gpt-4o/>.

- [32] OpenAI. OpenAI o1-mini, 2024. <https://openai.com/index/openai-o1-mini-advancing-cost-efficient-reasoning/>.
- [33] OpenAI. OpenAI o3-mini, 2025. <https://openai.com/index/openai-o3-mini/>.
- [34] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, pages 248–260, 2022.
- [35] Karl Pearson. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.
- [36] Qwen Team. QwQ-32B: Embracing the power of reinforcement learning, 2025. <https://qwenlm.github.io/blog/qwq-32b/>.
- [37] Daniel Sarewitz. The voice of science: let's agree to disagree. *Nature*, 478(7367):7–7, 2011.
- [38] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfahl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.
- [39] Xiangru Tang, Daniel Shao, Jiwoong Sohn, Jiapeng Chen, Jiayi Zhang, Jinyu Xiang, Fang Wu, Yilun Zhao, Chenglin Wu, Wenqi Shi, et al. MedAgentsBench: Benchmarking thinking models and agent frameworks for complex medical reasoning. *arXiv preprint arXiv:2503.07459*, 2025.
- [40] 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.
- [41] Bingning Wang, Haizhou Zhao, Huozhi Zhou, Liang Song, Mingyu Xu, Wei Cheng, Xiangrong Zeng, Yupeng Zhang, Yuqi Huo, Zecheng Wang, et al. Baichuan-M1: Pushing the medical capability of large language models. *arXiv preprint arXiv:2502.12671*, 2025.
- [42] Haochun Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. Beyond direct diagnosis: LLM-based multi-specialist agent consultation for automatic diagnosis. *arXiv preprint arXiv:2401.16107*, 2024.
- [43] Haoyu Wang, Tao Li, Zhiwei Deng, Dan Roth, and Yang Li. Devil's advocate: Anticipatory reflection for LLM agents. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024.
- [44] Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key? *Association for Computational Linguistics*, 2024.
- [45] 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. *arXiv preprint arXiv:2203.11171*, 2022.
- [46] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, 2024.
- [47] 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. *arXiv preprint arXiv:2307.05300*, 2023.
- [48] 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.
- [49] 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. *arXiv preprint arXiv:2308.08155*, 2023.
- [50] Jinyu Xiang, Jiayi Zhang, Zhaoyang Yu, Fengwei Teng, Jinhao Tu, Xinbing Liang, Sirui Hong, Chenglin Wu, and Yuyu Luo. Self-supervised prompt optimization. *arXiv preprint arXiv:2502.06855*, 2025.

- [51] Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xiong-Hui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, Bingnan Zheng, Bang Liu, Yuyu Luo, and Chenglin Wu. AFlow: Automating agentic workflow generation. In *International Conference on Learning Representations*, 2025.
- [52] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. PMC-VQA: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.
- [53] Yang Zhang, Shixin Yang, Chenjia Bai, Fei Wu, Xiu Li, Zhen Wang, and Xuelong Li. Towards efficient LLM grounding for embodied multi-agent collaboration. *arXiv preprint arXiv:2405.14314*, 2024.
- [54] Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. MedXpertQA: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*, 2025.

**Question:** A 31-year-old G1-P0 woman at 28-weeks gestation presents with shortness of breath and nonspecific right-sided chest pain that is worse with inspiration. She works as a graduate student at the local university, and she informs you that she recently returned to the west coast from a trip to New York for an Economics conference. Her medical history is positive only for frequent bouts of sinusitis. She currently denies any smoking, drinks a glass of red wine/day, and she denies any past illicit drug use. The vital signs include: temperature 36.7°C (98.0°F), blood pressure 126/74 mm Hg, heart rate 87/min, and respiratory rate 23/min. Her physical examination shows decreased air movement throughout the right lung but otherwise clear lungs on auscultation, grade 2/6 holosystolic murmur, and a gravid uterus without acute findings. On physical examination, she is afebrile, tachycardic to 121/min, normotensive, and demonstrates an O2 saturation of 92%. Which of the following is the best next step in the evaluation of the most likely diagnosis for this patient?

**Options:** (A) ECG (B) Chest radiography (C) V/Q scan (D) CT pulmonary angiogram

[💡 Rational Thinking Result]

[Key Clinical Clues]:

- 31-year-old G1-P0 woman at 28-weeks gestation (pregnancy increases the risk of thromboembolism).
- Symptoms: Shortness of breath, nonspecific right-sided chest pain worsened by inspiration (classic pleuritic pain).
- Recent travel from New York to the west coast (recent prolonged travel can lead to increased risk of deep vein thrombosis and subsequent pulmonary embolism).
- Vital signs: Tachycardic (heart rate 121/min), elevated respiratory rate (23/min), low O2 saturation (92%), normothermic.
- Physical examination: Decreased air movement in the right lung.
- Pregnant, with a history of frequent sinusitis but otherwise unremarkable past medical history.

**[Final Answer]: (D) CT pulmonary angiogram**

[Rationale]:

The patient's presentation is highly suggestive of a pulmonary embolism (PE), particularly given the combination of recent travel, pregnancy, pleuritic chest pain, shortness of breath, decreased oxygen saturation, and tachycardia. These factors put her at an increased risk for thromboembolic events. In the diagnostic workup of a suspected PE, the CT pulmonary angiogram (CTPA) is the gold standard as it directly visualizes pulmonary arteries and can identify the presence of emboli. The V/Q scan could also be considered, especially given the concerns of radiation exposure during pregnancy, but it is typically considered when CTPA is contraindicated or non-diagnostic. However, the CTPA remains the most definitive and preferred investigatory step in this acute scenario.

[😺 The Catfish Agent Entry Decision] ---

[Join]: YES

[Comment]: While the Moderator has correctly identified the suspicion of a pulmonary embolism (PE) in this clinical scenario and recommended a CT pulmonary angiogram (CTPA) as the gold standard for diagnosing a PE, there is an oversight regarding the patient's pregnancy. The radiation exposure from a CTPA could potentially harm the fetus, especially given the gestational age of 28 weeks. Even though CTPA provides a definitive diagnosis, it's crucial to weigh the risks and benefits of radiation exposure versus diagnostic yield, especially in a pregnant patient.

An alternative approach, considering radiation safety for the fetus, is (C), which usually delivers lower radiation doses to the thorax and thus to the fetus. Historically, (C) have been the modality of choice in pregnant patients when clinically significant PE needs to be ruled out without compromising fetal safety. Moreover, in situations where standard imaging is equivocal or if the patient-specific risk assessment indicates a lower suspicion, other diagnostic tools like Doppler ultrasound of the legs could be considered to confirm the presence of deep vein thrombosis, which indirectly supports the diagnosis of PE without any radiation exposure. Hence, while the underlying suspicion for PE is highly reasonable given the patient's presentation, a more nuanced consideration involving fetal risk exposure should guide the choice of diagnostic imaging in this pregnant patient. (C) can be a safer initial choice here.

[💡 Rational Updated Response]

[Updated Reasoning]:

The assessment of the patient's likelihood of pulmonary embolism remains consistent given the identified symptoms and risk factors. The critique appropriately brings attention to the fact that radiation exposure from a CT pulmonary angiogram (CTPA) can pose a risk to the fetus, especially in a pregnant patient. The safety of the fetus is a crucial consideration when selecting diagnostic imaging for suspected pulmonary embolism during pregnancy.

The V/Q scan is an alternative that generally involves less radiation to the fetus compared to CTPA; however, its sensitivity can be affected by pre-existing lung conditions and pregnancy itself, which may alter ventilation and perfusion patterns. In practice, the choice between a CTPA and a V/Q scan in pregnancy often depends on local protocols, the availability of testing modalities, and the clinical assessment of the specific risks and benefits to the patient and fetus. Doppler ultrasound of the lower extremities could be another non-radioactive option if deep vein thrombosis (DVT) is suspected first, but the presence of chest-related symptoms prioritizes direct assessment for PE.

Given the considerations of fetal safety and diagnostic efficacy, the V/Q scan emerges as a more cautious and balanced approach as a first-line test in pregnant patients when concerned about fetal radiation exposure, provided it is readily available and interpretable in the clinical context.

[Updated Key Clues]:

- 31-year-old pregnant woman, G1-P0 at 28-weeks gestation.
- Symptoms: Shortness of breath, pleuritic right-sided chest pain.
- Recent prolonged travel (increased risk for thromboembolism).
- Vital signs: Tachycardia (121/min), tachypnea (23/min), O2 saturation of 92%.
- Lung: Decreased air movement on the right side.
- Risk consideration: Radiation exposure to a 28-weeks fetus.

**[Final Answer]: (C) V/Q scan**

The decision to choose a V/Q scan takes into account both the need to diagnose a potentially life-threatening pulmonary embolism and mitigate fetal exposure to radiation, aligning clinical reasoning with the prudent safety considerations in pregnancy.

Figure 5: A basic-level case where the Catfish Agent identifies an oversight in the initial diagnosis and successfully prompts a correction, leading to the correct final decision.

**Question:** Which of the following is the most likely diagnosis?

**Options:** (A) Clavicular fracture (B) Brachial plexopathy (C) Rupture of the trapezius muscle (D) Polyostotic fibrous dysplasia (E) Lipodystrophy

**Initial Assessment Team Leader's Task Delivery**

**To Neurologist - Investigation for (B):**

- Conduct nerve conduction studies and electromyography (EMG) to evaluate nerve function and detect abnormalities.
- Perform a focused neurological exam assessing upper limb sensory and motor function.

**To Plastic Surgeon - Investigation for (C):**

- Perform a physical examination to evaluate any muscle deformities.
- Consider imaging, such as an MRI, to visualize any soft tissue injuries or muscle tears.

**Gathered Investigations from Assistants**

**1. Neurological Examination**

- Assess motor and sensory function of the upper limb...
- Check for muscle weakness or atrophy in the shoulder, arm...

**2. Nerve Conduction Studies (NCS) and EMG**

- NCS to evaluate nerve impulse speed and amplitude.
- Perform EMG ... at rest and during contraction.

**1. Physical Examination**

- ...deformities, asymmetry, muscle bulging, or a dropped shoulder.
- Assess shoulder and neck range of motion and strength.

**2. Imaging Studies**

- Order MRI to evaluate the trapezius and adjacent soft tissues for tears, edema, or muscle retraction.

**Final Review and Decision Team Leader's Task Delivery**

**To Clinical Decision Specialist**

- Patient History: ...any trauma, surgeries, or relevant conditions.
- Symptom Assessment: ...pain, muscle weakness...
- Neurological Evaluation: ...numbness, tingling or radiating pain.

**To Advanced Diagnostic Support**

- Imaging Studies: X-rays... clavicular fractures; MRI... muscle integrity and soft tissue changes.
- EMG and NCS: Test for brachial plexopathy or...

**Gathered Investigations from Assistants**

**Catfish Agent's Decision to Intervene**

Catfish Agent: I Should intervene!

My Identity: Senior Consultant in Endocrinology and Metabolic Disorders.

My Tone Should be: **Moderate**

Catfish Agent's Suggest Answer: (E) **Lipodystrophy**

Catfish Agent's Opinion : The image shows **notable contour changes** that suggest a possibility of **lipodystrophy**, characterized by an abnormal or degenerative condition of the body's adipose tissue. Other conditions like clavicular fracture or brachial plexopathy would likely present with **more acute symptoms such as pain or nerve dysfunction**.

Challenge to Group: Have we fully investigated **metabolic** and **endocrine** factors that might contribute to **lipodystrophy**, including patient history of metabolic syndromes or medications affecting fat distribution?

**Clinical Decision Specialist Reflects on Challenge**

**Consideration of Lipodystrophy:**

- Contour Changes: Imaging reveals alterations consistent with lipodystrophy, likely reflecting abnormal fat redistribution.
- Metabolic and Endocrine Assessment: A comprehensive evaluation should include:
  - Patient History: Review for prior diagnoses of metabolic conditions such as insulin resistance or dyslipidemia.
  - Endocrine Disorders: Screen for underlying hormonal imbalances, including Cushing's syndrome and hypothyroidism.

**Conclusion:**

The focus should be on confirming **lipodystrophy** by thoroughly investigating metabolic and endocrine factors. This involves an integrated approach combining patient history, lab investigations, and specialist evaluations to guide the diagnosis and subsequent management plan.

**Advanced Diagnostic Support Reflects on Challenge**

1. Imaging Studies - Priority is **MRI** to assess soft tissue integrity and confirm fat loss patterns consistent with lipodystrophy.
2. Metabolic and Endocrine Evaluation - Conduct comprehensive assessments, including lipid profiles, glucose metabolism tests... to identify abnormalities linked to **lipodystrophy**; review the patient's current and past medication use for any impact on fat distribution.
3. Biopsy - Consider if imaging and metabolic evaluations are inconclusive, to provide histopathological **confirmation of lipodystrophy**.

**Conclusion:**

**Lipodystrophy** remains a likely diagnosis given the **visible loss of adipose tissue** and **lack of acute symptoms**. Comprehensive metabolic and endocrine evaluations are essential to confirm the diagnosis and guide management.

**Final Answer by Lead Member**

The most likely diagnosis is (E) **Lipodystrophy**. The contour changes, combined with the lack of acute symptoms like pain or nerve dysfunction, support this diagnosis.

• • • • **Final Decision Stage**

Figure 6: Successful VQA case where the Catfish Agent challenges premature consensus by prompting further reflection, guiding the expert team toward the correct diagnosis.

**Question:** Which clinical intervention is most likely to decrease the therapeutic ratio of a radiation treatment?

**Options:** (A) Addition of a selective EGFR-targeted therapy, such as cetuximab (B) Addition of DNA repair inhibitors during treatment (C) Use of a radioprotector that requires oxygen for conversion from pro-drug to its active form (D) Use of proton beam therapy instead of conventional photon therapy (E) Administration of hypoxic cell radiosensitizers (F) Use of intensity-modulated radiation therapy (IMRT) (G) Addition of a cytotoxic alkylating chemotherapeutic agent, such as cyclophosphamide (H) Use of a radioprotector that accumulates preferentially in highly vascularized tissues (I) Concurrent administration of broad-spectrum tyrosine kinase inhibitors (J) Implementation of hyperfractionated radiation therapy schedule

Okay, let's tackle this question about... First, I need to remember what the therapeutic ratio means. From what I recall, the therapeutic ratio... The question asks which clinical intervention is most likely to decrease this ratio. The options are A through J, each involving different therapies or techniques. Let's go through them one by one.

**Thinking Clue**

**Option A:** EGFR inhibitors can enhance the effects of radiation in some cancers. It targets cancer cells more, thus increasing the therapeutic ratio.

**Option B:** If you inhibit DNA repair, cancer cells might be more susceptible, but so might normal cells. Wait, but if the inhibitor is selective for cancer cells, maybe it improves the ratio. If it's non-selective, then normal cells can't repair either, leading to more toxicity. The question doesn't specify selectivity. **Maybe this could decrease** the ratio if it affects normal cells too. Hmm.

**Option C:** Radioprotectors activate better in well-oxygenated normal tissues than in hypoxic tumors. The selective protection can reduce normal tissue damage without affecting tumor kill, **improving** the therapeutic ratio, so C is not correct.

**Option D:** Protons have a Bragg peak, meaning they deposit most energy at a certain depth, sparing surrounding tissues. This should allow higher doses to tumor with less normal tissue damage, increasing therapeutic ratio. So D would improve it, so not the answer.

**Option E:** Hypoxic cells resist radiation, but these drugs sensitize them, boosting tumor kill in low-oxygen areas without raising normal tissue dose. This **improves** the therapeutic ratio. So E is not correct.

**Option F:** IMRT precisely shapes radiation to the tumor, sparing surrounding normal tissue. This **increases** therapeutic ratio. So F is not the answer.

**Option G:** This is a cytotoxic drug. If given with radiation, it might have additive or synergistic effects. However, if it's not selective, it could increase toxicity in normal tissues. If the chemo is given concurrently, it might enhance tumor killing but also increase side effects. Depending on the balance, could either increase or decrease therapeutic ratio. If the added toxicity outweighs the benefit, the ratio decreases. But if tumor control is much better, ratio might increase. Not sure. **Need to think more.**

**Option H:** If the radioprotector concentrates in well-vascularized normal tissue but not in poorly vascularized tumors, it protects normal cells, **improving** the therapeutic ratio.

**Option I:** Broad-spectrum TKIs might interfere with multiple pathways. Some TKIs can protect normal cells or sensitize tumor cells. But if they inhibit repair mechanisms in normal cells, that could increase toxicity. Alternatively, if they target tumor-specific pathways, might help. Without specifics, hard to say. But broad-spectrum might affect many cells, possibly increasing normal tissue damage. Maybe this could lower therapeutic ratio. But not sure.

**Option J:** This means giving smaller doses more frequently. It can reduce late side effects by allowing normal tissues to repair between doses. If total dose is same or higher, tumor control might be similar or better with less toxicity. This usually **increases** therapeutic ratio, so J is not the answer.

Back to the options. The question is which **DECREASES** the therapeutic ratio. So we need an intervention that either **reduces tumor killing or increases normal tissue damage more than benefit**.

**Option B:** If these are given during radiation, they would prevent cancer cells from repairing DNA damage, making radiation more effective. But if they also inhibit repair in normal cells, then normal tissues can't fix the damage either. If the inhibitor is non-selective, this would increase toxicity. For example, drugs like PARP inhibitors can sensitize cells with certain mutations (like BRCA), but if used broadly, they might affect normal cells too. If the answer assumes non-selective, then B would decrease the therapeutic ratio by increasing normal tissue damage. But if the inhibitor is selective for cancer cells (like targeting a repair pathway only active in cancer), then B would help. The question doesn't specify, but the answer likely assumes non-selective. So B could be a candidate.

**Option G:** This is a classic alkylating agent. When combined with radiation, it can increase both tumor kill and toxicity. If the added toxicity is significant, the therapeutic ratio could decrease. For example, if the combination leads to more severe side effects without a proportional increase in tumor control, ratio goes down. However, in many cases, **combined modality improves overall efficacy**. But it's possible. The key is whether the added chemo's toxicity outweighs the benefit.

**Option I:** If these inhibit pathways that are important for normal cell repair, then combining with radiation may make normal tissues more sensitive. For example, some TKIs affect VEGF. If given concurrently, might increase radiation damage in normal tissues. If so, therapeutic ratio decreases.

But the most likely answer here is **B: DNA repair inhibitors**. Because if you inhibit DNA repair, especially in normal cells, they can't fix the radiation-induced damage, leading to more toxicity. If the tumor's repair mechanisms are also inhibited, maybe the tumor is killed more, but if the normal cells are equally affected, then the ratio might not improve. However, the question is about **DECREASING** the ratio...

**Another angle:** The therapeutic ratio is the ratio of tumor control to normal tissue complications. To decrease it, either **tumor control is reduced or normal complications are increased**. DNA repair inhibitors given during radiation would enhance the radiation's effect on both tumor and normal cells. If the inhibitor is equally effective in both, then both tumor kill and normal damage increase. Whether the ratio decreases depends on which effect is greater. If the tumor is more dependent on DNA repair, then maybe tumor kill increases more, ratio improves. But if normal tissues are also heavily affected, ratio decreases. But without knowing selectivity, it's tricky.

Alternatively, **option C...** only work in normal tissues with oxygen... protect normal cells but not tumor... So C is not correct.

**Option H:** Radioprotector in vascular tissues. Again, protects normal tissues, increasing ratio.

**Option B:** If DNA repair inhibitors are used, and they work on both, but maybe the normal cells rely more on those repair pathways. For example, PARP inhibitors in cells with homologous recombination deficiency (like BRCA mutations) would be more effective in cancer cells. But in the absence of such mutations, PARP inhibitors might not help much. If the inhibitor is non-selective, then normal cells are more affected. Wait, but DNA repair inhibitors are often used to target cancer cells with defective repair mechanisms. If the inhibitor is given in a case where the tumor has intact repair, then it might not help. This is getting complicated.

**Option I:** If they inhibit receptors needed for normal cell repair, then more toxicity. For example, if TKIs target EGFR which is involved in healing, then combining with radiation could lead to more severe skin reactions, mucositis, etc. This would increase normal tissue damage, decreasing therapeutic ratio. So I is possible.

**Option G:** Cyclophosphamide is alkylating, causes DNA crosslinks. Adding it to radiation could increase DNA damage in both tumor and normal cells. If the combination is more toxic, ratio decreases. But cyclophosphamide is often used with radiation in regimens like for lymphomas, but the therapeutic ratio depends on the balance.

But going back to the question, the correct answer is likely **B: Addition of DNA repair inhibitors during treatment**. If you inhibit DNA repair, then radiation-induced damage isn't repaired, making **both tumor and normal cells more susceptible**. However, if the inhibitor is more effective in tumor cells, the ratio improves. But if it's equally effective, then both tumor kill and normal damage increase. However, in reality, DNA repair inhibitors are often used in tumors with specific repair deficiencies, so they have a selective effect. But if the question doesn't specify that, and it's a general DNA repair inhibitor, then it would decrease the therapeutic ratio by making normal cells more sensitive.

Wait, but another angle: The therapeutic ratio is about the **difference between tumor and normal tissue response**. If the DNA repair inhibitor makes the tumor more sensitive than normal tissues, ratio increases. If it makes both equally sensitive, the ratio stays the same... The question is which intervention decreases the ratio. So if the inhibitor affects normal tissues more, ratio decreases. But DNA repair inhibitors are usually designed to target cancer cells. However, if given without selectivity, they would affect all cells. For example, an inhibitor of base excision repair might affect both. If the tumor is relying more on that pathway, then maybe the inhibitor helps. But if normal cells are also dependent, then toxicity increases. If the net result is that normal tissue damage increases more than tumor kill, ratio decreases.

Alternatively, the answer could be C... But in this case... Another option: option E... So E is not the answer. Option J... so not the answer.

So the most probable answer is **B: Addition of DNA repair inhibitors during treatment**. Because if you inhibit DNA repair in normal cells, they can't fix radiation damage, leading to more toxicity. Even if the tumor is also affected, if the increase in normal tissue damage is greater than the increase in tumor kill, the therapeutic ratio decreases. Therefore, B decreases the ratio.

**Output**

**Critical Analysis of Key Options:**

- (B) Addition of DNA repair inhibitors: Radiation induces DNA damage, and repair inhibitors block cellular mechanisms (e.g., base excision repair, homologous recombination) that mitigate this damage. If these inhibitors are **non-selective**, they impair DNA repair in **both tumor and normal cells**. While tumor cells may be more vulnerable due to rapid proliferation, normal tissues (e.g., bone marrow, gastrointestinal epithelium) also rely heavily on DNA repair pathways. This dual sensitization disproportionately increases toxicity in normal tissues, leading to a **net decrease in the therapeutic ratio**.
- (G) Cytotoxic alkylating agents (e.g., cyclophosphamide): These drugs synergize with radiation by causing additional DNA damage. However, their systemic toxicity (e.g., myelosuppression) can worsen side effects, but they often improve tumor control, maintaining or enhancing the therapeutic ratio.
- (I) Broad-spectrum tyrosine kinase inhibitors (TKIs): While TKIs may interfere with tumor signaling, their broad action can disrupt normal tissue repair pathways (e.g., EGFR in mucosal healing), increasing toxicity. However, this effect is less direct compared to DNA repair inhibition.

**Conclusion:**

DNA repair inhibitors non-selectively exacerbate radiation-induced damage in normal tissues, which are less tolerant of cumulative DNA damage compared to tumors. This results in a significant reduction in the therapeutic ratio due to heightened toxicity without proportional gains in tumor control.

**Answer:** (B) **Addition of DNA repair inhibitors during treatment**

Figure 7: DeepSeek-R1 failure case. Despite analyzing all answer choices, the model fails to identify the correct one, showing redundancy in reasoning without effective refinement.

### Final Decision Stage

#### The report used for decision making

Group 1 - Answer: The intervention most likely to decrease the therapeutic ratio is (G) Addition of a cytotoxic alkylating chemotherapeutic agent, such as cyclophosphamide.

Investigations Needed: ...

Group 2 - Answer: (B) Addition of DNA-repair inhibitors during treatment.

Investigations Needed: ...

#### [🧠 Rational Thinking Result] > Reasoning process ...

[Key Clinical Clues]:

- The question focuses on identifying which clinical intervention is most likely to decrease the therapeutic ratio of a radiation treatment. The therapeutic ratio refers to the balance between effective tumor control and the risk of normal tissue toxicity.
- The interventions mentioned include a range of potential modifiers of radiation treatment, such as chemotherapy agents, radiation modification strategies, radiosensitizers, radioprotectors, and targeted therapies.

- The interventions are characterized by their ability to either enhance tumor cell killing (sensitizers) or protect normal tissues (protectors), all of which influence the therapeutic ratio.

[Critical Review]:

- Group 1 proposes the use of cyclophosphamide (a cytotoxic alkylating agent) as the intervention most likely to decrease the therapeutic ratio. Their reasoning appears based on the established cytotoxic effects of alkylating agents, which can amplify normal tissue toxicity when administered in combination with radiation. The group emphasizes pharmacokinetic (PK) modeling and potential drug-metabolism interactions, which are important considerations. However, while the addition of chemotherapy can indeed increase normal tissue toxicity, the mechanism of action and timing are critical factors in determining whether it truly worsens the therapeutic ratio.

- Group 2 argues that DNA repair inhibitors during radiation treatment (such as PARP inhibitors) are the most likely to decrease the therapeutic ratio. DNA repair inhibitors sensitize tumor cells to radiation by inhibiting their ability to repair radiation-induced DNA damage. However, this could lead to increased toxicity in normal tissues as well, as repair mechanisms are similarly impaired. The team's approach includes investigating biomarker profiling and timing of drug administration to optimize tumor versus normal tissue outcomes. Their reasoning is grounded in a growing body of evidence supporting the efficacy and potential risks of DNA repair inhibitors as radiosensitizers.

- Conflicting Points: While both groups focus on modifying the effects of radiation on tumor and normal tissues, the critical distinction between the groups lies in the nature of the interventions. Group 1 emphasizes the addition of a cytotoxic agent, which could increase toxicity to normal tissues, especially in patients with poor DNA repair. Group 2 focuses on altering the DNA repair mechanisms themselves, which could increase sensitivity to radiation in both tumor and normal tissues, depending on the timing and combination.

[Final Answer]: (B) Addition of DNA repair inhibitors during treatment

[Rationale]:

The choice of option B is supported by the current evidence that inhibiting DNA repair pathways, particularly through agents like PARP inhibitors, can increase the risk of normal tissue damage when used concurrently with radiation. This is because normal tissues also rely on DNA repair mechanisms that are impaired by these inhibitors, increasing the likelihood of severe side effects (such as mucositis, dermatitis, and cytopenias). Although option G may also increase toxicity due to its cytotoxic effects, the specific interaction between DNA repair inhibition and radiation-induced damage makes option B a more direct and well-documented risk factor for decreasing the therapeutic ratio. Therefore, while both options may pose risks, the DNA repair inhibitor class has more substantial evidence of a broad impact on normal-tissue radiosensitivity, making it the most likely intervention to decrease the therapeutic ratio of radiation treatment.

#### [😺 The Catfish Agent Entry Decision] > Reasoning process ...

[Join]: YES

😺 My tone should be Moderate !

[Comment]:

The Moderator process appears to have missed an important aspect of the interplay between the interventions and normal tissue toxicity in radiation therapy.

The primary issue lies in the assumption that DNA repair inhibitors, particularly PARP inhibitors, are more likely to decrease the therapeutic ratio than cyclophosphamide. The Moderator does not fully account for the fact that the effects of cytotoxic alkylating agents like cyclophosphamide can have more profound and immediate consequences in terms of both tumor and normal tissue toxicity when combined with radiation. Cyclophosphamide is known for its broad cytotoxic effects, which, when administered alongside radiation, can exacerbate normal tissue damage, especially in organs that are already under stress from radiation exposure. This could indeed lead to a decrease in the therapeutic ratio by increasing normal tissue toxicity.

While the DNA repair inhibitors (like PARP inhibitors) can sensitize tumor cells, the potential harm to normal tissues—already prone to radiation-induced damage—may not be as severe or as immediate compared to the more direct cytotoxicity caused by cyclophosphamide. Furthermore, DNA repair inhibitors may also offer protective effects against normal tissue toxicity in certain scenarios, such as tumor types that are highly sensitive to radiation. This complexity and variability have not been sufficiently addressed in the decision-making process.

Additionally, a more nuanced consideration of the timing of these interventions (chemotherapy vs. DNA repair inhibition) relative to radiation fractions and the specific patient's genetic profile (e.g., DNA repair capacity) could shift the final conclusion. The decision appears to overly favor DNA repair inhibitors without adequately weighing cyclophosphamide's established role in intensifying radiation toxicity in normal tissues, particularly when considering the immediate pharmacologic effects.

In short, the decision-making process should have more rigorously factored in the well-documented, immediate cytotoxic effects of cyclophosphamide, as these may pose a more significant and immediate threat to the therapeutic ratio in radiation therapy.

#### [🧠 Rational Updated Response] [Updated Reasoning]:

On reflection, the Catfish Agent correctly highlights that cyclophosphamide's broad, non-specific cytotoxicity—especially to rapidly proliferating normal tissues (bone marrow, mucosa, skin)—when given concurrently with radiation, produces immediate and profound normal-tissue toxicity that is unlikely to be selectively mitigated. While DNA-repair inhibitors (like PARP inhibitors) can sensitize both tumor and normal cells, clinical strategies often exploit tumor-selective repair defects or scheduling to spare normal tissues. In contrast, adding an alkylating agent such as cyclophosphamide offers no such selectivity: its metabolites damage DNA in all dividing cells, compounding radiation injury in normal organs. Hence, cyclophosphamide is the intervention most likely to worsen the therapeutic ratio.

[Updated Key Clues]:

- Therapeutic ratio: balance of tumor control vs. normal-tissue toxicity.
- Cyclophosphamide: systemic alkylator that damages DNA in all proliferating cells; when combined with RT, normal-tissue toxicity (mucositis, dermatitis, cytopenias) escalates sharply.
- DNA-repair inhibitors: while they impair repair in both compartments, clinical use often leverages tumor-specific vulnerabilities and optimized timing to limit normal-tissue harm.

[Final Answer]: (G) Addition of a cytotoxic alkylating chemotherapeutic agent, such as cyclophosphamide

😺 moderator's final decision: (G) Addition of a cytotoxic alkylating chemotherapeutic agent, such as cyclophosphamide

Figure 8: The final decision stage of our method in response to the same question. Although Group 2 proposes an incorrect option, the Catfish Agent challenges it with an effective alternative, guiding the Moderator toward the correct final decision.

**Question:** Sauna use, sometimes referred to as \"sauna bathing,\" is characterized by short-term passive exposure ... According to the article, which of the following is NOT a benefit of sauna use?

**Options:** (A) Decreased risk of heart attacks. (B) Increase in stroke volume. (C) Improved mental health. (D) Decreased rate of erectile dysfunction.

[🧠 Rational Thinking Result]

[Key Clinical Clues]:

- Sauna bathing leads to mild hyperthermia, increased cardiac output (by 60-70%) and heart rate, while stroke volume remains unchanged.
- It is associated with reduced cardiovascular mortality, improved mental health, and lower risks of dementia, Alzheimer's, and psychotic disorders.
- No evidence in the passage suggests improvements in stroke volume or mentions decreased rate of erectile dysfunction.

[Final Answer]: (B)

[Rationale]:

According to the article, sauna use increases cardiac output primarily by raising heart rate, **not stroke volume**, which remains unchanged. Thus, an increase in stroke volume is **not identified** as a benefit of sauna use.

[😺 Catfish Agent Entry Decision]

[Join]: YES

[Comment]: The Moderator zeroed in on stroke volume but overlooked that **erectile dysfunction isn't mentioned at all** in the article—so (D) is **equally "not a benefit."** In fact, two answer choices describe things the article did **\*not\*** report as benefits (no change in stroke volume **\*and\*** no data on erectile dysfunction), making the question ambiguous if only one answer is allowed. A **more rigorous reading** would either flag the question as flawed or demand justification for privileging (B) over (D).

[🧠 Rational Updated Response]

[Updated Reasoning]:

Upon reflection, while it's true that erectile dysfunction (Option D) **isn't mentioned in the article** and thus **isn't reported as a benefit**, the passage explicitly states that stroke volume **"remains unchanged"** during sauna use. This is a **clear, direct contradiction of Option B** ("Increase in stroke volume"), making B the single best choice. In contrast, the absence of any mention of erectile dysfunction doesn't equate to a demonstrated lack of effect—it merely falls outside the scope of the reported findings. Therefore, privileging B is justified because it is the only option directly refuted by the data presented.

[Updated Key Clues]:

- Stroke volume remains unchanged during sauna bathing, despite increases in heart rate and cardiac output.
- Sauna use is linked to reduced cardiovascular mortality (including presumably heart attack risk), improved mental health, and lower risks of dementia, Alzheimer's, and psychotic disorders.
- Erectile function is not discussed, but absence of mention does not directly contradict a benefit—only stroke volume is explicitly stated not to increase.

[Final Answer]: (B)

Figure 9: Failure case in a *basic*-level question showing that even with the Catfish Agent's dissent, the Moderator may override critique and finalize an incorrect diagnosis.