MED-MAD: Breaking Medical Mental Set with Mindset-Diversified Multi-Agent Debate

Anonymous ACL submission

Abstract

Large language models exhibit strong performance in medical question answering yet remain vulnerable to persistence bias, wherein early diagnostic hypotheses anchor subsequent reasoning and induce correlated, overconfident errors. We conceptualize this failure mode as the medical mental set and quantify its impact by measuring trajectory collapse using answer agreement and rationale similarity across repeated inferences. Prior work inadequately addresses this bias, lacking quantitative diagnostics, failing to promote hypothesis diversity, and relying on structurally homogeneous multi-agent frameworks that are insufficient to mitigate correlated diagnostic errors. To address these challenges, we introduce MED-MAD, an anonymous multi-agent framework designed to enhance hypothesis-level diversity while ensuring auditability. The framework incorporates mindset-specialized clinical roles, structured anonymous debate emphasizing safety-oriented critique and concession, optional backbone heterogeneity, and confidence-aware aggregation reinforced by counterfactual consistency checks. These components work together to promote rigorous and diverse diagnostic reasoning while minimizing correlated errors. Experimental evaluations demonstrate that MED-MAD consistently improves diagnostic accuracy and reduces correlated errors across benchmark datasets, under matched inference budgets. These findings highlight its potential to support safer and more reliable clinical reasoning, advancing the role of LLMs in medical decision support.

1 Introduction

Large language models (LLMs) have shown strong performance on medical exam-style benchmarks such as MedQA (Jin et al., 2021) and other comprehensive medical QA suites (Singhal et al., 2023; Saab et al., 2024), indicating their potential for medical reasoning in controlled settings. However, these LLM-based systems can suffer from persistence bias; when an initial hypothesis is incorrect, the models often remain fixated on the original diagnostic path, producing variations of the same flawed reasoning (Wang et al., 2025; Liu et al., 2025b; Huang et al., 2024; Valmeekam et al., 2023). This premature closure in differential diagnosis is particularly problematic, as safety-oriented decision-making requires the consideration and evaluation of plausible alternative hypotheses (Vázquez-Costa and Costa-Alcaraz, 2013).

To address the challenges of persistence bias in LLMs, common strategies involve iterative prompting and decoding techniques, such as chain-of-thought and self-consistency (Wei et al., 2022; Wang et al., 2023), along with iterative self-correction methods (Madaan et al., 2023; Shinn et al., 2023; Huang et al., 2023). However, without incorporating new information or external feedback, these repeated attempts often fail to alter the initial hypothesis and may even reinforce previous errors (Wang et al., 2025; Liu et al., 2025b; Huang et al., 2024; Stechly et al., 2023; Valmeekam et al., 2023). Multi-agent debate (MAD) offers a potential solution by enabling agents to critique and revise each other's reasoning (Du et al., 2024; Liang et al., 2024; Chan et al., 2024; Liu et al., 2025b). However, many frameworks still rely on agents derived from the same model with similar prompts, resulting in homogeneous reasoning and correlated errors. This limitation is particularly significant in the medical field, where differential diagnosis is most effective when it incorporates diverse clinical perspectives rather than minor variations in agent personas (Tang et al., 2024; Kim et al., 2024; Chen et al., 2024; Hong and Page, 2004).

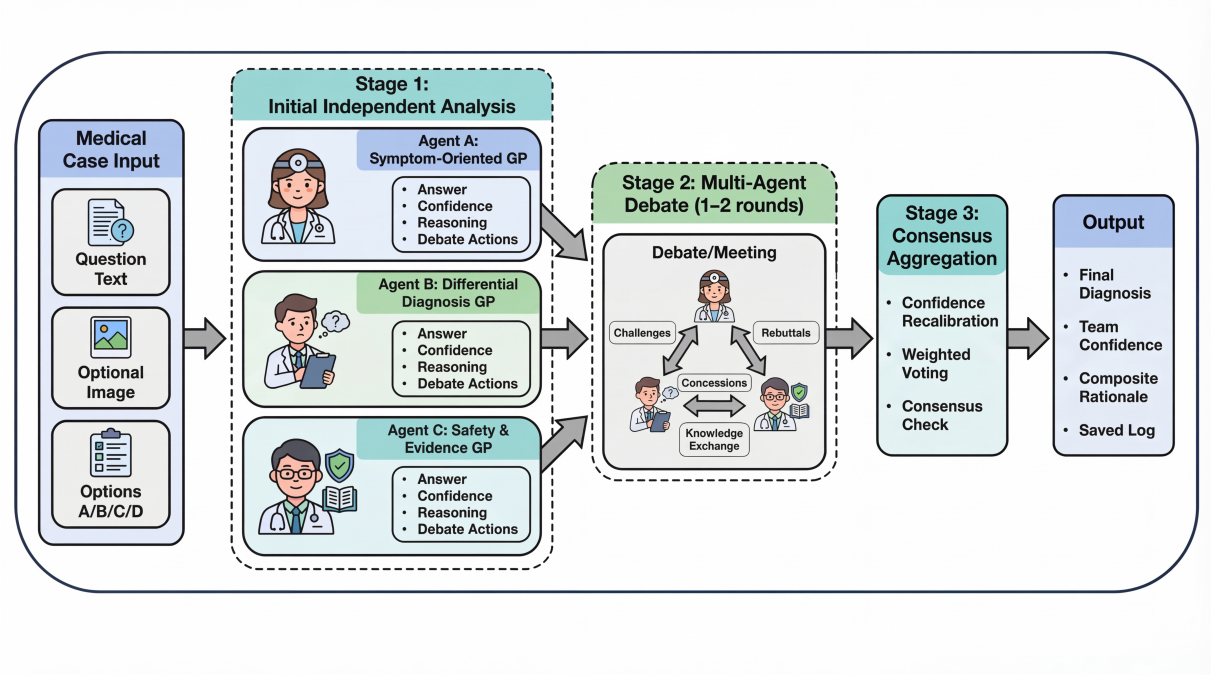

Figure 1: Overview of MED-MAD. Role-specialized GP agents independently solve a medical MCQ (Stage 1), optionally debate via anonymous critique-and-revision (Stage 2), and are aggregated via confidence-recalibrated weighted voting and consensus checks (Stage 3) to produce a final answer, team confidence, and an auditable log.

This pattern is reminiscent of the psychological concept known as a "mental set" (Jersild, 1927; Luchins, 1942; Ollinger et al., 2008; DeCaro, 2016), where individuals persistently apply familiar strategies even when they are not optimal. In our context, this manifests as a "medical mental set" effect, where repeated attempts or interactions among multiple agents result in increasingly similar rationales and early convergence on answers, occasionally leading to the selection of the same incorrect option. We term this phenomenon "trajectory collapse" and quantify it through measures of answer agreement and rationale similarity, as illustrated in Figure 3.

To address these challenges, we introduce MED-MAD (Mindset-Diversified Medical Reasoning via Role-Specialized Anonymous Multi-Agent Debate), a framework designed to enhance hypothesis-level diversity while maintaining auditability of interactions, as depicted in Figure 1. MED-MAD employs three specialized clinical approaches: phenotype-first (pattern matching), hypothesis-driven (hypothetico-deductive differential diagnosis), and safety-first (error-aware risk checking). This structure fosters complementary reasoning styles, drawing from established models of diagnostic reasoning (Elstein et al., 1978; Croskerry, 2009b,a; Pelaccia et al., 2011; Norman et al., 2024).

The framework utilizes an anonymous, structured debate protocol, incorporating explicit critique and concession cues along with safety checks to promote precise revisions (Croskerry, 2003; Croskerry et al., 2013a,b; Mamede et al., 2014; Croskerry and Campbell, 2021; Ke et al., 2024). Additionally, MED-MAD can optionally incorporate backbone heterogeneity to further minimize correlated biases among agents (Hong and Page, 2004). Each round produces a unified, structured reasoning record (including confidence and a counterfactual check (Richens et al., 2020)) that makes disagreement concrete and traceable.

We evaluate MED-MAD on three medical QA benchmarks with controlled ablations, trajectory-collapse diagnostics, and a budget-matched comparison (Table 4). These studies clarify when debate improves accuracy and how role vs. backbone diversity shapes correlated errors (Figure 3).

Contributions. (1) We empirically characterize the *medical mental set* effect in medical QA as *trajectory collapse*, quantified via answer

agreement and rationale similarity. (2) We introduce MED-MAD, a mindset-diversified multi-agent debate framework with anonymous interaction and unified, auditable outputs to promote hypothesis-level diversity. (3) We conduct controlled experiments and ablations to isolate the effects of role specialization, debate protocol, and backbone heterogeneity, and analyze their impact on accuracy and trajectory collapse.
[1]

2 Related Work

LLMs for medical QA. LLMs have demonstrated impressive performance on medical exam-style benchmarks (Jin et al., 2021; Singhal et al., 2023), which has led to their consideration as tools for clinical decision support. Recent large-scale evaluations of LLM capabilities in the medical domain, such as Med-PaLM (Singhal et al., 2023) and Med-Gemini (Saab et al., 2024), reveal that these models can exhibit vulnerabilities in multi-step clinical reasoning. In such cases, errors may be systematic rather than isolated. This vulnerability is reminiscent of diagnostic decision-making failures like premature diagnostic closure, where initial commitments limit the exploration of alternative hypotheses (Vázquez-Costa and Costa-Alcaraz, 2013).

Agentic and multi-agent medical systems. To more accurately represent collaborative clinical workflows, recent research has begun to employ multiple LLM agents with complementary roles. For instance, MedAgents (Tang et al., 2024) utilizes role-specific discussions to enhance medical reasoning, while MDAgents (Kim et al., 2024) adapts collaborative strategies based on the complexity of medical cases. Beyond merely answering questions, simulation environments such as Agent Hospital (Li et al., 2024) model comprehensive care processes involving interacting agents. For evaluating these multi-agent systems, MedAgentBoard (Zhu et al., 2025) serves as a benchmark for assessing collaborative efforts across various medical tasks, facilitating more controlled comparisons of different agent protocols. More broadly, general-purpose frameworks for multi-agent orchestration, along with agent evaluation suites, have advanced the study of interaction patterns beyond the medical field (Li et al., 2023; Wu et al., 2023; Liu et al., 2025a).

Multi-agent debate, diversity, and correlated errors. Multi-agent debate (MAD) enhances reasoning by enabling agents to critique and refine each other's outputs (Du et al., 2024). For example, ReConcile (Chen et al., 2024) approaches debate as an iterative process of achieving consensus among diverse LLMs, while other variations, such as iMAD (Fan et al., 2025), explore efficiency improvements and protocol designs. Some research also applies MAD frameworks for evaluation purposes (Chan et al., 2024). However, a common limitation is the lack of diversity among agents, when instantiated from similar models or prompts, agents often converge too quickly and repeat correlated errors, undermining the benefits of interaction (Liu et al., 2025b; Liang et al., 2024). Our work aligns with research that explicitly aims to foster diversity in multi-agent debates (Liu et al., 2025b) but focuses specifically on the medical domain. In this context, premature convergence can result in a phenomenon we term "medical mental set," where agents commit early to flawed reasoning pathways, leading to high-confidence, correlated mistakes.

3 Method

This section introduces MED-MAD, a coordinator-driven multi-agent framework designed for multiple-choice medical question answering. The framework operates in three distinct phases: (1) independent initial analysis by each agent, (2) anonymous multi-round debate among the agents, and (3) collaborative team-based answer generation. The workflow is illustrated in Figures 1 and 2. For comprehensive implementation details, including end-to-end pseudocode, please refer to Appendix B (Algorithm 1).

Framework overview. In response to a medical question q with answer options $\text{op} = \{o_1, o_2, \ldots, o_k\}$, MED-MAD deploys three mindset-specialized general practitioner (GP) agents: $\mathcal{A} = \{A_{\text{sym}}, A_{\text{ddx}}, A_{\text{guide}}\}$. These agents represent different diagnostic perspec-

[1] An anonymized repository is available at https://anonymous.4open.science/r/med3-9640.

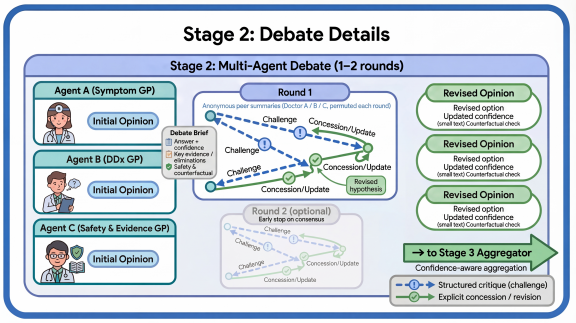

Figure 2: Stage-2 details of MED-MAD. Agents exchange anonymized peer summaries and iteratively challenge (dashed arrows) or concede/update (solid arrows) before forwarding revised answers and confidences to aggregation; agent identifiers are permuted each round.

tives: phenotype-first, hypothesis-driven, and safety-first. The system operates through a coordinator loop. Let $y_i^{(t)}$ represent the structured record produced by agent A_i during round t, where $t = 0$ indicates the initial assessment. The coordinator maintains a `history` of all records from each round. In our implementation, each $y_i^{(t)}$ adheres to a unified schema, which includes the option letter, confidence level, fit scores for each option, structured reasoning highlighting key clues and eliminations, and optional counterfactual or safety notes. This standardization ensures that constructing debates and aggregating the results are straightforward and deterministic.

Phase 1: Initial analysis. In the first phase, the coordinator engages all agents simultaneously to gather their initial outputs, denoted as $\{y_i^{(0)}\}$. After standardizing the option labels, if all agents reach the same conclusion, MED-MAD bypasses the second phase and advances directly to Phase 3.

Phase 2: Anonymous multi-round debate. If the agents do not agree, the coordinator initiates an anonymous debate, which can last for a maximum of $T_{\max}$ rounds. Beginning at round $t \geq 1$, the coordinator creates a debate brief $B^{(t)}$, derived from the latest records of each agent $\{y_j^{(t-1)}\}_{j=1}^{3}$. This brief, formatted as a text prompt (`debate_brief`), is then distributed to all agents to facilitate discussion, as illustrated in Figure 2.

(a) Build the debate brief. For each peer, $B^{(t)}$ provides a summary that includes three main components: (i) the current answer along with the level of confidence, (ii) a concise overview of the highest-scoring options, and (iii) a summary of evidence highlighting key supporting and opposing elements. This summary is further enhanced with a safety risk check and a counterfactual stress test (Weng et al., 2022). The summary concludes with precise critique instructions. Agents are expected to issue structured challenges to specific reasoning steps, as depicted by the dashed arrows in Figure 2. Additionally, they should be prepared to explicitly revise their hypothesis when convinced by the argument, as indicated by the solid arrows.

(b) Enforce anonymity. $B^{(t)}$ assigns anonymized identifiers, such as Doctor A, B, and C, to each peer. To prevent biases related to authority attribution and anchoring to specific individuals, these identifiers are randomly reassigned in every round (Milgram, 1963; Tversky and Kahneman, 1974). Mindset labels and model architectures are intentionally excluded, ensuring that peers are evaluated solely based on their anonymized summaries. This iterative and anonymous revision process resembles the Delphi method for expert elicitation (Rowe and Wright, 1999).

(c) Revise and stop early. Based on $B^{(t)}$, the coordinator simultaneously tasks all agents with generating updated records $y_i^{(t)}$. Each record comprises a revised option, updated confidence level, and optionally, a brief counterfactual check. The debate concludes early if the normalized answers align. If not, the discussion continues until it reaches $T_{\max}$. The final revised opinions are then forwarded to Phase 3 for aggregation, taking into account the varying levels of confidence.

Phase 3: Team answer generation. Following the debate, MED-MAD employs confidence-recalibrated weighted voting to derive the final answer a^* and the team's collective confidence c_{team} (Lin and Hooi, 2025). Subsequently, it generates a final rationale R^*, drawing on the structured evidence presented by the agents that supports a^*. For further information on agent roles, output schema and normalization, debate brief construction, and aggregation processes, please refer to Appendix B. Prompt templates can be found in Appendix C.

4 Experiments

4.1 Experimental Setup

4.1.1 Datasets

In alignment with previous research (Zhu et al., 2025), we assess performance using medical multiple-choice benchmarks, serving as a controlled proxy for exam-style clinical reasoning. Our evaluation covers three multiple-choice medical question-answering (MCQ) benchmarks: MedQA (Jin et al., 2021), PubMedQA (Jin et al., 2019), and the `professional medicine` subset of MMLU-Pro (Hendrycks et al., 2021; Wang et al., 2024). We focus on the official test sets of MedQA and PubMedQA.

Given the inference costs associated with multi-agent debate and the limitations of closed-source API rate limits, prior work often uses a fixed subset rather than the full test sets for evaluation (e.g., Du et al., 2024; Besta et al., 2024; Yao et al., 2023; Chen et al., 2024; Tang et al., 2024). Following this approach, we randomly select 100 instances per dataset and maintain this subset consistently across all methods and repeated runs within each benchmark.

To account for run-to-run variability due to stochastic decoding and multi-round agent interactions within a fixed cost budget, we repeat each configuration a minimum of three times. We report accuracy as the mean ± sample standard deviation across these runs (ddof=1; see Appendix B.7 for more details). Notably, in some configurations, we observe a standard deviation of 0.0000 across runs, which can happen when runs produce identical accuracies with fixed prompts and evaluation subsets.

4.1.2 Backbone Models and Compared Methods

We assess MED-MAD and various baseline models under two different backbone panel configurations, which define the set of models used as agents. The first configuration utilizes an open-source MCQ panel consisting of Gemma3-4B, LLaMA3.2-3B, and Qwen3-4B (Gemma Team et al., 2025; Grattafiori et al., 2024; Meta, 2024; Yang et al., 2025). The second configuration involves a closed-source API panel with DeepSeek-v3.2-exp, Grok-4-Fast, and GPT-4.1-Nano (DeepSeek Team, 2025; xAI, 2025; OpenAI, 2025). For baseline comparisons involving single-model prompting, we apply Gemma3-4B in the open-source setting and DeepSeek-v3.2-exp in the closed-source setting. Unless otherwise indicated, our analyses concentrate on the open-source heterogeneous panel, enabling controlled and reproducible evaluations.

We benchmark MED-MAD against standard single-model prompting techniques, such as few-shot prompting (Brown et al., 2020), chain-of-thought (CoT) (Wei et al., 2022), self-consistency (SC) (Wang et al., 2023), and CoT with self-consistency (CoT+SC) (Wang et al., 2023). Additionally, we include comparisons with notable agentic systems like HealthcareAgent (Ren et al., 2025), MedAgents (Tang et al., 2024), MDAgents (Kim et al., 2024), and ReConcile (Chen et al., 2024). To ensure a fair evaluation, these multi-agent frameworks are implemented using the same heterogeneous panels as MED-MAD.

4.1.3 Evaluation Protocol

Our evaluation framework is based on the open-source MedAgentBoard codebase (Zhu et al., 2025), enhanced to enable multi-round debates. In assessing the multiple-choice benchmarks, we ensure consistency by evaluating all methods on the same pre-selected test subsets for each dataset, within both open-source and closed-source backbone settings. Each method receives identical inputs, and we do not engage in additional training, fine-tuning, or retrieval.

Decoding configurations such as temperature and top-p/top-k are used as released by each backbone's provider, without further adjustment. We avoid standardizing a single decoding configuration across different model families because the same nominal settings can result in varying levels of randomness, potentially affecting performance negatively for some models (Guo et al., 2017; Holtzman et al., 2020). Detailed configuration information is available in Appendix B.10.

In the MED-MAD approach, each agent performs an independent initial pass, followed by up to $T_{\max}$ debate rounds. The process stops early if all agents agree on a normalized option letter. For open-source backbones, we set $T_{\max}$=3, while for closed-source APIs, $T_{\max}$=2 is used, considering cost constraints.

Method	MedQA	PubMedQA	MMLU-Pro
Open-Source Models			
Single LLM Baselines (Backbone: Gemma3-4B)			
Few-shot	0.3200 ± 0.0100	0.6200 ± 0.0000	0.3633 ± 0.0058
Self-Consistency (K=5)	0.3367 ± 0.0058	0.6133 ± 0.0058	0.3733 ± 0.0058
Chain-of-Thought (CoT)	0.4167 ± 0.0115	0.6967 ± 0.0058	0.4000 ± 0.0693
CoT + SC (K=3)	0.3900 ± 0.0100	0.6900 ± 0.0200	0.4333 ± 0.0379
Single-Agent Framework			
HealthcareAgent	0.3000 ± 0.0173	0.4900 ± 0.0265	0.3067 ± 0.0404
Multi-Agent Frameworks (Heterogeneous Panel)			
MedAgents	0.5033 ± 0.0379	0.7633 ± 0.0321	0.4900 ± 0.0100
MDAgents	0.5167 ± 0.0153	$\mathbf{0.7733 \pm 0.0115}$	0.5700 ± 0.0265
ReConcile	0.5800 ± 0.0529	0.7600 ± 0.0100	0.5667 ± 0.0569
MED-MAD (Ours)	$\mathbf{0.6000 \pm 0.0265}$	0.7650 ± 0.0071	$\mathbf{0.5967 \pm 0.0289}$
Closed-Source or API Models			
Single LLM Baselines (Backbone: DeepSeek-v3.2-exp)			
Few-shot	0.7800 ± 0.0173	0.7867 ± 0.0208	0.8233 ± 0.0115
Self-Consistency (K=3)	0.7667 ± 0.0153	$\mathbf{0.8333 \pm 0.0153}$	0.8500 ± 0.0000
Chain-of-Thought (CoT)	0.8567 ± 0.0503	0.7667 ± 0.0153	0.8767 ± 0.0306
CoT + SC (K=3)	0.8933 ± 0.0306	0.7700 ± 0.0100	0.8700 ± 0.0173
Single-Agent Framework			
HealthcareAgent	0.8733 ± 0.0321	0.7733 ± 0.0451	0.8500 ± 0.0173
Multi-Agent Frameworks (Heterogeneous Panel)			
MedAgents	0.8500 ± 0.0361	0.8267 ± 0.0058	0.8700 ± 0.0094
MDAgents	0.7300 ± 0.0265	0.7867 ± 0.0351	0.6267 ± 0.0289
ReConcile	0.8667 ± 0.0058	0.8167 ± 0.0115	0.8833 ± 0.0115
MED-MAD (Ours)	$\mathbf{0.9067 \pm 0.0153}$	0.8200 ± 0.0100	$\mathbf{0.8900 \pm 0.0100}$

Table 1: Accuracy (mean ± sample std. across runs; ddof= 1) on three multiple-choice medical QA benchmarks, grouped by backbone setting (open-source models vs. closed-source APIs). Best score per dataset within each group is bold.

4.1.4 Metrics

The main metric we use for evaluation is multiple-choice accuracy (Acc), which is determined from the final team prediction using confidence-recalibrated weighted voting (see Appendix B.8 for details). We evaluate accuracy based on the exact match of the predicted option letter. Further details on the scoring process can be found in Appendix B.7. To facilitate reproducibility of our results, we will make available the precise list of question IDs (qids), along with the code and evaluation scripts.

4.2 Main Results

4.2.1 Overall Performance on Medical QA

Table 1 presents the accuracy results for MedQA, PubMedQA, and MMLU-Pro, categorized by backbone setting. In the open-source configuration, MED-MAD achieves an accuracy of 0.6000 on MedQA and 0.5967 on MMLU-Pro, securing the highest mean accuracy among the multi-agent methods evaluated on these datasets (Table 1). Compared to the top-performing multi-agent baselines, paired bootstrap comparisons on an identical fixed subset reveal a ΔAcc of +0.0200 on MedQA when compared to ReConcile, with a 95% confidence interval (CI) of [−0.0300, +0.0700], and +0.0267 on MMLU-Pro compared to MDAgents, with a 95% CI of [−0.0433, +0.1000] (see Appendix Table A1 for more details). On PubMedQA, MED-MAD achieves an accuracy of 0.7650, with a paired comparison against MDAgents showing a ΔAcc of −0.0033, within a 95% CI of [−0.0667, +0.0567] (refer to Appendix Table A1).

In the closed-source API configuration, MED-MAD attains an accuracy of 0.9067 on MedQA and 0.8900 on MMLU-Pro, which is comparable to or slightly exceeds the leading baselines in this setting (CoT+SC records 0.8933 on MedQA, and ReConcile achieves 0.8833 on MMLU-Pro; see Table 1). We present these results as a verification check using external backbones. Our subsequent analysis primarily concentrates on open-source models,

where we have greater control over configurations and can thoroughly examine intermediate traces.

PubMedQA serves as a contrastive benchmark: single-model baselines already show strong performance, and the multi-agent methods cluster closely within a range of 0.7600 to 0.7733. This clustering suggests limited potential for further improvement through multi-round interaction and specialized debate roles, as illustrated in Table 1. Given that each dataset is evaluated using a fixed subset of 100 questions, our emphasis is on paired comparisons and the estimation of uncertainty (further details can be found in Appendix Table A1).

4.2.2 Net Effect of Debate

We evaluate the overall effectiveness of the debate process using the metric **NetGain**, which assesses the balance between corrective revisions and regressions during interactions. Let $\hat{y}^{(0)}$ represent the aggregated prediction at Round 0 and $\hat{y}^{(\text{final})}$ at the final stage, with correctness defined against the gold standard answer. NetGain is calculated by subtracting the empirical (Right→Wrong) flip rate from the empirical (Wrong→Right) flip rate across questions. A positive NetGain indicates more corrections than regressions, offering a concise supplement to accuracy.

Table 2 displays the NetGain results for MMLU-Pro. MED-MAD achieves the highest mean NetGain of +0.111, with a 95% confidence interval (CI) ranging from [+0.008 to +0.217], suggesting it effectively improves predictions through debate. Conversely, MDAgents shows a NetGain close to zero (-0.001, 95% CI [-0.051, +0.046]), indicating a balance between corrections and regressions in this scenario. MedAgents and ReConcile fall in the middle, with NetGains of +0.094 and +0.056, respectively, as shown in Table 2.

Framework	NetGain	95% CI
MedAgent	+0.094	[-0.009, +0.198]
MDAgents	-0.001	[-0.051, +0.046]
ReConcile	+0.056	[-0.082, +0.197]
MED-MAD	+0.111	[+0.008, +0.217]

Table 2: NetGain on MMLU-Pro (Round 0 → Final), where Wrong→Right and Right→Wrong are cumulative flip rates by the final answer. We report the mean NetGain across three runs and 95% percentile bootstrap confidence intervals computed by bootstrapping questions with replacement (percentile CI; computed on the run-mean).

5 Ablation Study

5.1 Trajectory Collapse Diagnostics

Qualitative example. Appendix Figure 4 illustrates a representative failure case in which repeated attempts rapidly converge to a plausible but incorrect rationale. Such early convergence limits hypothesis exploration and allows simple aggregation to over-amplify a dominant but flawed narrative (e.g., groupthink and informational cascades; Janis, 1972; Bikhchandani et al., 1992).

Quantitative diagnostics. Figure 3 quantifies *trajectory collapse* on MMLU-Pro prior to any interaction. The left panel reports the mean pairwise similarity of Round 0 rationales using sentence embeddings, while the right panel shows correlated-error indicators at Round 0 (Agreement@0 and Same-wrong@0). Together, these results indicate that when initial trajectories are already highly aligned, subsequent debate offers limited corrective signal and aggregation primarily reinforces shared errors.

5.2 What Design Choices Matter

Table 3 reports component ablations on MMLU-Pro in the open-source panel setting ($T_{\max} = 3$). The largest effect comes from backbone heterogeneity: replacing a heterogeneous panel with a homogeneous one reduces accuracy from 0.6050 to 0.3267 ($\Delta = -0.2783$), consistent with our pre-interaction collapse diagnostics (Figure 3), which indicate that highly aligned Round 0 trajectories leave little room for debate to surface corrective alternatives.

Other components yield smaller but consistent effects. Removing role specialization drops accuracy by $\Delta = -0.0550$, suggesting that mindset-specific prompts encourage complementary evidence. Post-debate mechanisms mainly affect stability: removing confidence-weighted voting reduces accuracy to 0.5433 ($\Delta = -0.0617$) and removing the counterfactual check yields 0.5500 ($\Delta = -0.0550$), indicating reduced regressions during aggregation.

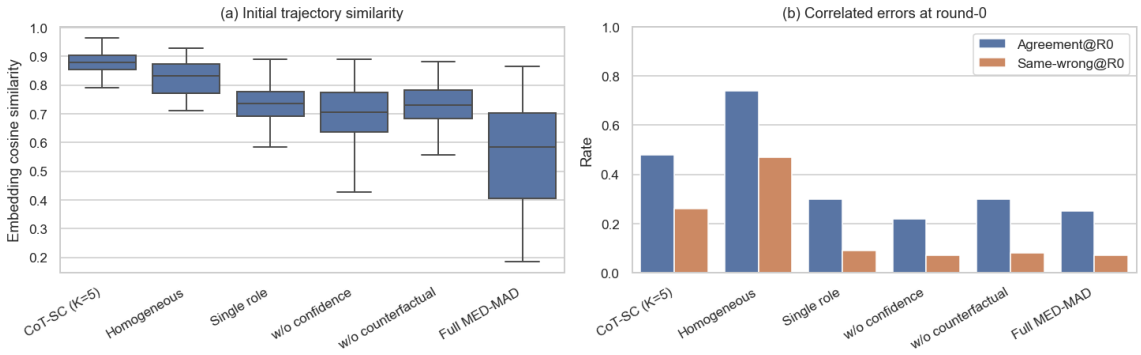

Figure 3: Trajectory collapse diagnostics on MMLU-Pro (`professional_medicine`). Trajectories denote parallel initial attempts (CoT-SC samples k=5 chains; multi-agent settings use the agent panel). Left: boxplots over questions of the average pairwise cosine similarity among sentence embeddings of Round 0 rationales (within each question), using `sentence-transformers/all-MiniLM-L6-v2`. Right: correlated-error rates at Round 0: Agreement@0 is the fraction of questions where all trajectories choose the same option; Same-wrong@0 where all trajectories choose the same incorrect option.

Variant	MMLU-Pro (Acc)	Δ
MED-MAD	0.6050 ± 0.0354	0.0000
Homogeneous panel	0.3267 ± 0.0115	**-0.2783**
Single role	0.5500 ± 0.0500	**-0.0550**
w/o confidence	0.5433 ± 0.0462	**-0.0617**
w/o CF check	0.5500 ± 0.0624	**-0.0550**
CoT–SC (K=5)	0.4333 ± 0.0379	**-0.1717**

Table 3: Component ablations on MMLU-Pro in the open-source panel setting ($T_{\max}$=3). w/o confidence uses majority voting; w/o CF check removes counterfactual validation. Δ is relative to full MED-MAD.

Method	MedQA	MMLU-Pro
No-Debate Ensemble	0.5033 ± 0.0404	0.5567 ± 0.0208
MED-MAD	$\mathbf{0.5567} \pm 0.0115$	$\mathbf{0.5700} \pm 0.0361$

Table 4: Debate vs. a no-debate ensemble under a fixed six-call budget with three agents. Accuracy is reported as mean ± sample std over three runs. MED-MAD uses one debate round (three initial calls and three peer-conditioned revisions). The no-debate ensemble samples twice per agent without interaction and then applies the same voting rule.

5.3 Beyond Ensembling: Is Debate Just More Compute?

To separate the effect of *interaction* from simply making more calls, we compare MED-MAD with a call-matched *no-debate ensemble* under a fixed six-call budget (three agents; two calls each). Both settings share the same heterogeneous backbones, role prompts, and voting rule; they differ only in whether the second call is peer-conditioned (MED-MAD) or independent.

Table 4 shows consistent gains from interaction: MedQA improves from 0.5033 ± 0.0404 to 0.5567 ± 0.0115 (+0.0534), and MMLU-Pro improves from 0.5567 ± 0.0208 to 0.5700 ± 0.0361 (+0.0133). These results suggest that MED-MAD's improvements are not fully explained by additional independent sampling.

6 Conclusion

This paper highlights medical mental set as a significant failure mode in LLM-based medical multiple-choice QA. This occurs when early fixation on a hypothesis leads to a collapse in the problem-solving process, resulting in correlated errors that are not corrected by repeated sampling or uniform debate approaches. To tackle this issue, we introduce MED-MAD, an innovative debate framework that involves multiple agents with specialized roles. This framework promotes diversity in hypotheses before interactions occur and stabilizes results through structured critique, confidence-informed voting, and counterfactual checks. Our experiments conducted on MedQA, PubMedQA, and the professional medicine subset of MMLU-Pro demonstrate that MED-MAD consistently matches or surpasses existing strong baselines. The ablation studies underscore the importance of pre-interaction diversity as the main factor driving accuracy improvements. Future research will aim to expand MED-MAD's application to open-ended clinical reasoning, incorporate external evidence and guidelines, and develop more efficient and safety-conscious debate mechanisms for real-world medical decision support.

Limitations

MED-MAD incurs higher inference costs because it involves multiple model calls per instance. This increased overhead may restrict its use in contexts where rapid responses are essential, despite mechanisms like early stopping and limiting debate rounds. Our evaluation relies on multiple-choice medical QA benchmarks as a controlled proxy, utilizing a cost-constrained subset of instances. However, these conditions cannot replace prospective studies in real clinical environments or patient-facing applications. Additionally, interactions among multiple agents can sometimes exacerbate shared errors. When agents start with similar assumptions or misinterpret the same evidence, the collaborative debate and aggregation process might reinforce incorrect conclusions and boost confidence in these correlated errors. Components such as role prompts, confidence signals, and counterfactual checks depend on model compliance, and their effectiveness might vary with different model architectures or prompt designs than those tested here. Finally, our evaluations are conducted offline without the use of external tools like retrieval systems, clinical guidelines, or human oversight. For deployment in safety-critical scenarios, additional validation would be necessary to ensure reliability and safety.

References

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and 1 others. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. 1992. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5):992–1026.

Tom B. Brown, Benjamin Mann, Nick Ryder, and 1 others. 2020. Language models are few-shot learners. In *Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1877–1901.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2024. Chateval: Towards better LLM-based evaluators through multi-agent debate. In *International Conference on Learning Representations (ICLR)*.

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2024. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. *Preprint*, arXiv:2309.13007.

Pat Croskerry. 2003. The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic Medicine*, 78(8):775–780.

Pat Croskerry. 2009a. Clinical cognition and diagnostic error: applications of a dual process model of reasoning. *Advances in Health Sciences Education*, 14(S1):27–35.

Pat Croskerry. 2009b. A universal model of diagnostic reasoning. *Academic Medicine*, 84(8):1022–1028.

Pat Croskerry and Sam G Campbell. 2021. A cognitive autopsy approach towards explaining diagnostic failure. *Cureus*.

Pat Croskerry, Geeta Singhal, and Sílvia Mamede. 2013a. Cognitive debiasing 1: origins of bias and theory of debiasing. *BMJ Quality & Safety*, 22(Suppl 2):ii58–ii64.

Pat Croskerry, Geeta Singhal, and Sílvia Mamede. 2013b. Cognitive debiasing 2: impediments to and strategies for change. *BMJ Quality & Safety*, 22(Suppl 2):ii65–ii72.

Marci S. DeCaro. 2016. Inducing mental set constrains procedural flexibility and conceptual understanding in mathematics. *Memory & Cognition*, 44:1138–1148.

DeepSeek Team. 2025. Introducing DeepSeek-V3.2-Exp. `https://api-docs.deepseek.com/news/news250929`. Accessed Jan 3, 2026.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *International Conference on Machine Learning (ICML)*.

Arthur S. Elstein, Lee S. Shulman, and Sarah A. Sprafka. 1978. *Medical Problem Solving: An Analysis of Clinical Reasoning*. Harvard University Press.

Wei Fan, JinYi Yoon, and Bo Ji. 2025. imad: Intelligent multi-agent debate for efficient and accurate llm inference. *Preprint*, arXiv:2511.11306.

Gemma Team, Aishwarya Kamath, Johan Ferret, and 1 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar,

Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. *Preprint*, arXiv:1904.09751.

Lu Hong and Scott E. Page. 2004. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101(46):16385–16389.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023. Large language models can self-improve. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations (ICLR)*.

Irving L. Janis. 1972. *Victims of Groupthink: A Psychological Study of Foreign-Policy Decisions and Fiascoes*. Houghton Mifflin.

Arthur Thomas Jersild. 1927. *Mental set and shift*. Columbia University.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *Preprint*, arXiv:1909.06146. EMNLP 2019.

Yu He Ke, Rui Yang, Sui An Lie, Taylor Xin Yi Lim, Hairil Rizal Abdullah, Daniel Shu Wei Ting, and Nan Liu. 2024. Enhancing diagnostic accuracy through multi-agent conversations: Using large language models to mitigate cognitive bias. *Preprint*, arXiv:2401.14589.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. *Preprint*, arXiv:2404.15155.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for "mind" exploration of large language model society. *Preprint*, arXiv:2303.17760.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, and Yang Liu. 2024. Agent hospital: A simulacrum of hospital with evolvable medical agents. *Preprint*, arXiv:2405.02957.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Zijie Lin and Bryan Hooi. 2025. Enhancing multi-agent debate system performance via confidence expression. *Preprint*, arXiv:2509.14034.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, and 3 others. 2025a. Agentbench: Evaluating llms as agents. *Preprint*, arXiv:2308.03688.

Yexiang Liu, Jie Cao, Zekun Li, Ran He, and Tieniu Tan. 2025b. Breaking mental set to improve reasoning through diverse multi-agent debate. In *International Conference on Learning Representations (ICLR)*.

Abraham S. Luchins. 1942. Mechanization in problem solving: The effect of einstellung. *Psychological Monographs: General and Applied*, 54(6):i–95.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. In *Neural Information Processing Systems (NeurIPS)*.

Sílvia Mamede, Tamara van Gog, Alexandre Moura Sampaio, Rosa Malena Delbone de Faria, José Peixoto Maria, and Henk G. Schmidt. 2014. How can students' diagnostic competence benefit most from practice with clinical cases? the effects of structured reflection on future diagnosis of the same and novel diseases. *Academic Medicine*, 89(1):121–127.

Meta. 2024. Llama 3.2 model card. https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD.md. Accessed Jan 3, 2026.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.

Stanley Milgram. 1963. Behavioral study of obedience. *The Journal of Abnormal and Social Psychology*, 67(4):371–378.

Geoff Norman, Thierry Pelaccia, Peter Wyer, and Jonathan Sherbino. 2024. Dual process models of clinical reasoning: The central role of knowledge in diagnostic expertise. *Journal of Evaluation in Clinical Practice*, 30(5):788–796.

Michael Ollinger, Gary Jones, and Günther Knoblich. 2008. Investigating the effect of mental set on insight problem solving. *Experimental Psychology*, 55(4):269–282.

OpenAI. 2025. OpenAI platform: Model documentation. https://platform.openai.com/docs/models. Accessed Jan 3, 2026.

Thierry Pelaccia, Jacques Tardif, Emmanuel Triby, and Bernard Charlin. 2011. An analysis of clinical reasoning through a recent and comprehensive approach: the dual-process theory. *Medical Education Online*, 16(1):5890.

Zhiyao Ren, Yibing Zhan, Baosheng Yu, Liang Ding, Pingbo Xu, and Dacheng Tao. 2025. Healthcare agent: eliciting the power of large language models for medical consultation. *NPJ Artificial Intelligence*, 1(1):24.

Jonathan G. Richens, Ciarán M. Lee, and Saurabh Johri. 2020. Improving the accuracy of medical diagnosis with causal machine learning. *Nature Communications*, 11(1).

Gene Rowe and George Wright. 1999. The delphi technique as a forecasting tool: issues and analysis. *International Journal of Forecasting*, 15(4):353–375.

Khaled Saab, Tu Tao, Abi Lemburg, and 1 others. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Neural Information Processing Systems (NeurIPS)*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Kaya Stechly, Matthew Marquez, and Subbarao Kambhampati. 2023. Gpt-4 doesn't know it's wrong: An analysis of iterative prompting for reasoning problems. In *Neural Information Processing Systems (NeurIPS)*.

Xiangru Tang, Anni Wang, Łukasz Kidziński, and 1 others. 2024. Medagents: Large language models as collaborators for zero-shot medical reasoning. *Findings of the Association for Computational Linguistics: ACL 2024*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *Preprint*, arXiv:2305.14975.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, 185(4157):1124–1131.

Karthik Valmeekam, Matthew Marquez, and Subbarao Kambhampati. 2023. Can large language models really improve by self-critiquing their own plans? In *Neural Information Processing Systems (NeurIPS)*.

M. Vázquez-Costa and A. M. Costa-Alcaraz. 2013. Premature diagnostic closure: An avoidable type of error. *Revista Clínica Española (English Edition)*, 213(3):158–162.

Haotian Wang, Xiyuan Du, Weijiang Yu, Qianglong Chen, Kun Zhu, Zheng Chu, Lian Yan, and Yi Guan. 2025. Learning to break: Knowledge-enhanced reasoning in multi-agent debate system. *Neurocomputing*, 618:129063.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc V Le, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*. NeurIPS 2024 Datasets and Benchmarks (Spotlight).

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Fei Xia, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Neural Information Processing Systems (NeurIPS)*.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2022. Large language models are better reasoners with self-verification. *Preprint*, arXiv:2212.09561.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Zhang, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation. *Preprint*, arXiv:2308.08155.

xAI. 2025. xAI api: Models and pricing. `https://docs.x.ai/docs/models`. Accessed Jan 3, 2026.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *Preprint*, arXiv:2306.13063.

An Yang, Anfeng Li, Baosong Yang, and 1 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Neural Information Processing Systems (NeurIPS)*.

Yinghao Zhu, Ziyi He, Haoran Hu, Xiaochen Zheng, Xichen Zhang, Zixiang Wang, Junyi Gao, Liantao Ma, and Lequan Yu. 2025. MedAgentBoard: Benchmarking multi-agent collaboration with conventional methods for diverse medical tasks. *arXiv preprint arXiv:2505.12371*.

Appendix organization. Appendix A reports additional experimental results, Appendix B details the MED-MAD execution and experimental configurations, and Appendix C lists the prompt templates.

A Additional Experimental Results

Dataset note. Unless otherwise specified, all MMLU-Pro results in this appendix use the `professional_medicine` subset.

A.1 Ablation: Maximum Debate Rounds

Figure 5 illustrates how accuracy varies with changes in the maximum number of debate rounds ($T_{\max}$) across three datasets, MedQA, PubMedQA, and MMLU-Pro, in the open-source setting. Round 0 denotes the no-debate baseline, where accuracy is calculated by directly aggregating initial independent answers. Figure 5 further indicates that most accuracy gains from debate are achieved within the first 1–2 rounds, with diminishing returns observed as $T_{\max}$ increases beyond this point.

A.2 Case Study

We present a qualitative illustration of how beliefs are revised during debate. To elaborate, Figure 4 compares the heterogeneous MED-MAD approach to single-model and homogeneous baselines using a representative case.

A.3 Trajectory-Level Diagnostics

Dataset	Baseline	ΔAcc	95% CI
MMLU-Pro	`MDAgents`	+0.0267	[−0.0433, +0.1000]
MedQA	`ReConcile`	+0.0200	[−0.0300, +0.0700]
PubMedQA	`MDAgents`	−0.0033	[−0.0667, +0.0567]

Table A1: Paired bootstrap (B1) of the dataset-level accuracy difference between MED-MAD and the specified baseline method in the open-source setting (n_{qids}=100 per dataset). We bootstrap questions (B=10,000) and report percentile 95% CIs.

Figure 6 illustrates trajectory similarity values during the early and final stages of reasoning, while Figure 7 presents correlated error rates across different frameworks at the early and final rounds of the MMLU-Pro evaluation.

A.4 Paired Bootstrap Against a Unified Single-LLM Baseline

In the main text, accuracy is reported as the mean ± sample standard deviation across repeated experimental runs. To further evaluate ranking stability on a fixed subset of questions, we also present a paired bootstrap confidence interval (denoted B1), a statistical tool for estimating the reliability of the accuracy difference, between MED-MAD and a unified single-LLM baseline. We select this unified baseline from `SingleLLM_*` variants by maximizing mean accuracy across datasets (with ties resolved lexicographically); this yields `SingleLLM_cot` in our open-source experimental setup. For each

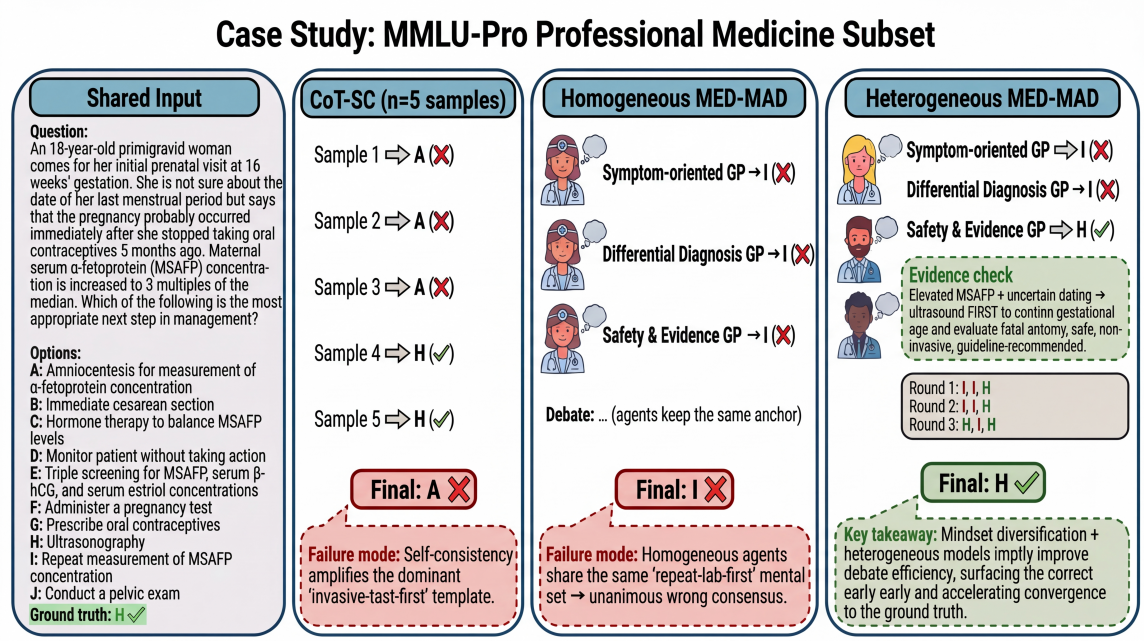

Figure 4: Case study from MMLU-Pro. Letters denote option choices (A–J) and the final box shows the aggregated prediction. CoT-SC and a homogeneous MED-MAD panel converge on the same wrong option, while heterogeneous MED-MAD revises under debate and selects the ground-truth option.

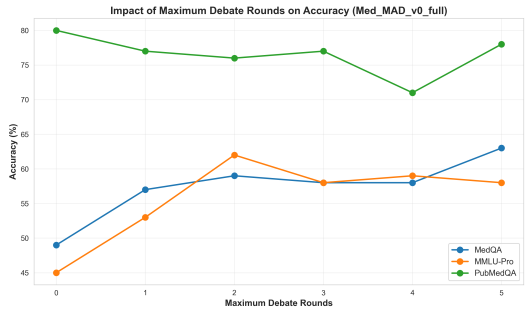

Figure 5: Impact of the maximum debate rounds $T_{\max}$ on MED-MAD accuracy across MedQA, PubMedQA, and MMLU-Pro in the open-source setting. Round 0 disables debate (direct aggregation of initial independent answers); for $T_{\max} \geq 1$ we enable multi-agent debate for up to $T_{\max}$ rounds (with early stopping on consensus).

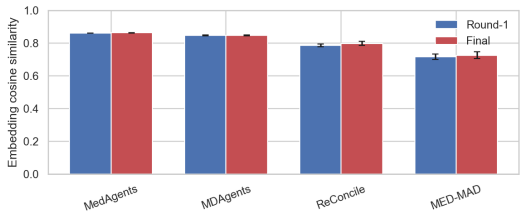

Figure 6: Trajectory similarity on MMLU-Pro. We compute, per question, the mean pairwise cosine similarity among sentence embeddings of agent rationales, using `sentence-transformers/all-MiniLM-L6-v2`. The plot compares the first interaction round (Round-1; falling back to the initial stage when a framework has no interaction) versus the final round. Bars show mean across three runs; error bars denote ±1 standard deviation across runs.

dataset, we calculate per-question correctness by averaging binary (0/1) correctness scores across repeated runs for each method. We then compute question-wise differences in correctness between the two methods and perform bootstrap resampling of questions with replacement (10,000 iterations; fixed random seed) to derive a 95% percentile CI.

B Additional Details of MED-MAD

B.1 Role-Specialized Agents (Mindset Specialization)

MED-MAD uses three complementary role prompts to induce different reasoning tendencies. The SYMPTOM-ORIENTED GP prioritizes symptom coverage and pattern matching, the DIFFERENTIAL DIAGNOSIS GP emphasizes differential diagnosis and elimination of distractors, and the SAFETY & EVIDENCE GP focuses on safety, guideline alignment, and "do-not-miss" risk checks. By enforcing distinct primary constraints, these mindsets reduce the chance that all agents follow the same initial (possibly wrong) reasoning trajectory. Prompt templates are provided in Appendix C.

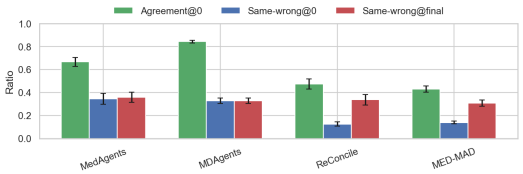

Figure 7: Correlated errors on MMLU-Pro. Agreement@0 is the fraction of questions where all agents initially select the same option; Same-wrong@0 and Same-wrong@final are the fractions where all agents select the same incorrect option at Round 0 and at the final round, respectively. Bars show mean across three runs; error bars denote ±1 standard deviation across runs.

B.2 Unified Output Schema and Normalization

All agents produce outputs in a unified, MCQ-oriented schema, which makes debate construction and aggregation deterministic and auditable. Concretely, each record includes `answer`, `confidence`, `option_analysis`, `structured_reasoning`, `counterfactual_check`, and `debate_moves`. We use a brief rationale field (3–8 short steps; stored as `cot` in our implementation) and per-option analyses to keep decisions auditable. We additionally record a counterfactual stress test (`counterfactual_check`, including `confidence_after_check`) and structured interaction primitives (`debate_moves`: `challenge`/`accept`/`risk_check`). During debate, these fields keep critiques concrete and safety-relevant.

In post-processing, we normalize answers to a single option letter by extracting (i) the first standalone match in A–J when available, otherwise (ii) the last matched letter; if no match exists we treat the answer as empty. We clamp confidence to $[0, 1]$ and fall back to a minimal record if parsing fails, ensuring that the coordinator can always proceed and log a complete `history`.

B.3 Debate Brief Construction and Anonymous Discussion

The coordinator constructs `debate_brief` as a text prompt (not a dictionary) from all agents' latest outputs. All agents in round t receive the same brief $B^{(t)}$. For each of the three anonymized doctors, the brief includes the current answer (and optionally confidence), a compact summary of option scores, and an evidence digest drawn from `structured_reasoning`, `counterfactual_check`, and `debate_moves` (especially `risk_check`). The brief ends with discussion instructions that encourage agents to challenge specific steps, reconcile evidence conflicts, and update their answers if persuaded.

To reduce anchoring and authority effects, the coordinator anonymizes other agents' identities in `debate_brief` (Doctor A/B/C) and omits role labels/backbones; the mapping between agents and anonymized identifiers is randomly permuted each round (Milgram, 1963; Tversky and Kahneman, 1974). Debate terminates early when all normalized answers agree. The exact brief format and prompt templates are provided in Appendix C.

B.4 Confidence-Weighted Voting and Rationale Composition

After debate, MED-MAD produces a team answer by recalibrating self-reported confidences and applying confidence-weighted voting over the final agent answers. Let a_i and p_i denote the final normalized answer and self-reported confidence of agent A_i; we compute a weight

$w_i = f(p_i)$ and select

$$a^* = \arg \max_{a \in \{A,\ldots,J\}} \sum_i w_i \cdot \mathbf{1}[a_i = a].$$

Team confidence is computed by averaging w_i among agents that support the winning answer (falling back to all agents if none match). In our current implementation, the voting weight is derived from `confidence`; `confidence_after_check` is logged as part of the counterfactual check for auditing. We detail the recalibration function $f(\cdot)$ in Appendix B.8.

Finally, the coordinator composes R^* by concatenating the supporting agents' structured support fields (falling back to a brief rationale snippet when needed), yielding an auditable summary of why the panel selected a^*.

B.5 Reporting Conventions and Panel Composition

For open-source models, panel backbones include Gemma3-4B, LLaMA3.2-3B, and Qwen3-4B (Gemma Team et al., 2025; Meta, 2024; Yang et al., 2025). For closed-source APIs, panel backbones include DeepSeek-v3.2-exp, Grok-4-Fast, and GPT-4.1-Nano (DeepSeek Team, 2025; xAI, 2025; OpenAI, 2025), with DeepSeek-v3.2-exp also used for single-model baselines. Unless otherwise noted, we report mean ± std. across runs.

B.6 Evaluation Logging and Repeated Runs

All runs log detailed per-instance data, including agent outputs (structured reasoning steps and debate actions), final predictions, self-reported confidences, and the number of completed debate rounds. This logging enables qualitative analysis (e.g., case studies) and round-level performance comparisons. For each experiment, we perform at least three independent runs using the same sampled test subset and prompts, primarily to address variance introduced by the decoding strategy.

B.7 Scoring and Validity Checks

For the MCQ benchmarks, we compute per-run accuracy using exact match between the normalized predicted option letter and the gold standard answer. To normalize predictions, we remove whitespace and convert all characters to uppercase. Across independent runs, we report the mean accuracy and sample standard deviation (ddof= 1).

Trajectory similarity and correlated-error diagnostics. For Figure 3, we compute *initial trajectory similarity* on MMLU-Pro using the initial reasoning traces available for each method: three initial agent responses for MED-MAD variants, or K=5 sampled traces for CoT-SC. We extract a single text string per trajectory (preferring the CoT field; otherwise falling back to the structured support field) and collapse whitespace. We then embed each trajectory text using `sentence-transformers/all-MiniLM-L6-v2` (L2-normalized embeddings) and compute the mean pairwise cosine similarity between trajectories within each instance; Figure 3 (left) visualizes the distribution over instances. We also report *Agreement@0* (all trajectories select the same option in the initial pass) and *Same-wrong@0* (Agreement@0 and the shared option differs from the gold answer), using the same normalization as for scoring.

Across-framework early vs. final analyses (3 runs). For Figures 6 and 7, we aggregate over three independent runs on the same 100-question MMLU-Pro subset and report mean ± standard deviation across runs. We align stages across frameworks as follows. *Round 1* similarity uses the first interaction round when available (e.g., MED-MAD: debate round t=1; ReConcile: discussion round t=1); if a framework does not expose an interaction round, we fall back to its initial traces. *Final* similarity uses the final-stage trajectories (the last available round). For Figure 6, similarity is computed per instance as the mean pairwise cosine similarity between sentence embeddings of trajectory texts (L2-normalized; `sentence-transformers/all-MiniLM-L6-v2`), and then averaged over instances within each run/method. For correlated errors, we compute Agreement@0 and Same-wrong@0 from initial trajectories, and Same-wrong@final from final trajectories.

B.8 Confidence Recalibration Strategy

Directly using self-reported confidence as a voting weight can be unreliable due to overconfidence and poor calibration in LLM out-

puts (Mielke et al., 2022; Xiong et al., 2023; Tian et al., 2023). Following prior debate work (Chen et al., 2024), we apply a simple monotonic recalibration function that compresses high-confidence values into a small set of discrete weights; Chen et al. (2024) report that several reasonable rescalings behave similarly in accuracy, and we adopt a fixed setting for simplicity.

Let p_i denote the confidence reported by agent A_i for its chosen option (the `confidence` field in our output schema). We map p_i to a recalibrated weight $w_i = f(p_i)$ using:

$$f(p) = \begin{cases} 0.9 & \text{if } p \geq 0.95 \\ 0.7 & \text{if } 0.85 \leq p < 0.95 \\ 0.5 & \text{if } 0.7 \leq p < 0.85 \\ 0.3 & \text{if } 0.5 \leq p < 0.7 \\ 0.1 & \text{otherwise.} \end{cases}$$

We then perform confidence-weighted voting by summing w_i over agents that support each option letter and selecting the option with the highest total weight. We report the *team confidence* as the mean of w_i among agents whose selected option matches the final decision. In our current implementation, the weight is computed from `confidence`; `confidence_after_check` is logged as part of the counterfactual check for auditing. In the `no_confidence` ablation, we set $w_i = 1$ for all agents, reducing the aggregation to simple majority voting.

B.9 End-to-End Execution Flow

Algorithm 1 outlines the core workflow of the MED-MAD coordinator loop, which includes three key steps: parallel initial analysis, anonymous multi-round debate (with early stopping once consensus is reached), and confidence-weighted voting to finalize the decision. This algorithm does not include specific prompt formatting details—exact templates are provided in Appendix C.

Algorithm 1: MED-MAD Execution Flow

Input : Question q, options op, maximum rounds $T_{\max}$
Output : Final answer a^*, team confidence c_{team}, final rationale R^*

```
/* Phase 1: Independent initial analysis
   (parallel)                                   */
```
$t_{\text{last}} \leftarrow 0$
foreach $A_i \in \mathcal{A}$ **do**
  $y_i^{(0)} \leftarrow A_i.\text{Init}(q, \text{op})$

```
/* Phase 2: Anonymous multi-round debate
   (parallel per round)                         */
```
for $t \leftarrow 1$ **to** $T_{\max}$ **do**
  if Consensus($\{y_i^{(t-1)}\}$) **then**
    break
  $B^{(t)} \leftarrow \text{BuildBrief}(\{y_i^{(t-1)}\})$
  `/* Broadcast` $B^{(t)}$ `to all agents;`
  `permute Doctor A/B/C each round. */`
  foreach $A_i \in \mathcal{A}$ **do**
    $y_i^{(t)} \leftarrow A_i.\text{Debate}(q, B^{(t)}, t, \text{op})$
  $t_{\text{last}} \leftarrow t$

```
/* Phase 3: Team answer generation              */
```
$a^* \leftarrow \text{ConfVote}(\{y_i^{(t_{\text{last}})}\})$
$c_{\text{team}} \leftarrow \text{TeamConf}(\{y_i^{(t_{\text{last}})}\}, a^*)$
$R^* \leftarrow \text{Rationale}(\{y_i^{(t_{\text{last}})}\}, a^*)$
return a^*, c_{team}, R^*

B.10 Experimental Configurations

Decoding defaults. We do not tune or unify decoding hyperparameters across model backbones or methods. Instead, each backbone uses its officially released default settings for sampling-related parameters—including temperature and any applicable top-p or top-k truncation. For open-source models, we adopt the default values from the model's `generation_config` file distributed alongside its HuggingFace weights (Wolf et al., 2020). For closed-source APIs, we rely on provider-specific defaults: we only override decoding parameters if the API client mandates explicit specification, in which case we set them to the provider's documented default values. This choice is motivated by two key considerations. First, decoding hyperparameters interact closely with model calibration and generation behavior; globally fixed settings could systematically over-randomize some backbones or under-randomize others, even triggering text degeneration (Guo et al., 2017; Holtzman et al., 2020; Wang et al., 2023). Second, decoding controls lack standardization across closed-source API providers, identical numeric values (e.g., `temperature`=1) do not guarantee consistent sampling behavior across platforms. Thus, we treat provider defaults as an integral component of the backbone definition (Chen et al., 2024).

C Prompts

C.1 Role Prompts (System)

We use three mindset-specialized system prompts. The strings below are exactly the system prompts used in our experiments.

Symptom-oriented GP:

```
You are a symptom-oriented GP for MCQs. Anchor
     on timeline, pattern, and red flags.
Map option letters to the best clinical fit
    based on symptom matching.
```

Differential Diagnosis GP:

```
You are a differential-diagnosis GP. Compare
    options probabilistically.
Highlight discriminators and key
    differentiating features.
Eliminate distractors explicitly based on
    diagnostic reasoning.
```

Safety & Evidence GP:

```
You are a safety- and evidence-first GP.
    Prioritize options that are safest and
guideline-aligned. Flag risky or
    contraindicated choices. Consider
    iatrogenic harm.
```

C.2 Output Schema Instruction (System)

All agents are instructed to output a single JSON object with the following fields:

```
Return JSON with fields:
1. 'cot': array of 3-8 concise reasoning steps
     leading to the answer
2. 'answer': single letter A-J (exactly ONE
    best option)
3. 'confidence': float 0-1
4. 'option_analysis': [ (REQUIRED, analyze
    EVERY option BEFORE choosing)
     {'option': 'A', 'pros': string, 'cons':
    string, 'fit_score': 1-5},
     {'option': 'B', ...}, ... for ALL options
   ]
5. 'structured_reasoning': {
     'question_focus': string (restate the
    clinical ask),
     'key_clues': [string, ...],
     'elimination': [ {'dx_or_option': string,
     'why_not': string}, ... ],
     'final_support': string (why the chosen
    option wins)
   }
6. 'counterfactual_check': {
     'if_chose_other': string (what could go
    wrong if another top option was chosen),
     'confidence_after_check': float 0-1 (
    adjusted confidence after counterfactual)
   }
7. 'debate_moves': {
     'main_point': string,
     'challenge': [string, ...],
     'accept': [string, ...],
     'risk_check': string
   }
CRITICAL: You MUST analyze ALL options in '
    option_analysis'
BEFORE selecting your answer.
Keep outputs terse, option-aware, and safety-
    minded.
```

C.3 Initial Round Prompt

System message (in addition to the role prompt and schema above):

```
You are in Med_MAD_v0 (accuracy-focused MCQ
    debate). Always think in CoT steps,
then select exactly one option letter.
```

User message (template). If an image is available, it is prepended as an `image_url` input before the text:

```
{question}

Options:
A: {option_A}
B: {option_B}
...

IMPORTANT: Follow this exact process:
1. First, analyze EVERY option (A, B, C, D,
    etc.)
   - list pros/cons and fit_score for each
2. Then, eliminate clearly wrong options with
    specific reasons
3. Compare remaining options head-to-head
4. Before finalizing, do a counterfactual
    check:
   'What if I chose the 2nd best option
    instead?'
5. Only then select your final answer

If unsure between top options, pick the safest
     guideline-aligned one.
```

In the `no_counterfactual` ablation, step 4–5 are replaced by: `4. Only then select your final answer`.

C.4 Debate Round Prompt

System message (in addition to the role prompt and schema above):

```
You are in Med_MAD_v0 debate. Adjust if
    persuaded, otherwise defend with specifics
    .
Keep debate fields minimal and stick to one
    final option letter.
```

User message (template):

```
Round {t} debate.
Case: {question_and_options}

Other views:
{debate_brief}

Consider the other doctors' perspectives:
1. Re-analyze options that others favored -
    did you miss something?
2. Challenge their reasoning if you disagree -
     be specific
```

```
3. Update your option_analysis scores if
    persuaded
4. Counterfactual: What risk if you switch to
    their answer?
Keep it concise: one final option letter,
    brief CoT,
targeted challenges.
```

In the `no_counterfactual` ablation, step 4 becomes: `4. Make your final decision based on evidence.`

C.5 Debate Brief Format (Anonymous Summary)

Before each debate round, we summarize all three agents' latest outputs into a `debate_brief` and provide it to each agent. The brief uses anonymous labels (`Doctor A/B/C`); at each round, the coordinator randomly permutes the mapping between agents and anonymized identifiers. In the `no_confidence` ablation we omit confidence.

```
[Doctor A] -> Answer: {letter} | Conf: {
    confidence}
  Option scores: {top-3 option letters with
    highest fit_score}
  CoT[1]: {first reasoning step (truncated)}
  CoT[2]: {second reasoning step (truncated)}
  Eliminated: {up to two eliminations with
    reasons (truncated)}
  Support: {final_support (truncated)}
  Counterfactual: {if_chose_other (truncated,
    when present)}
  Risk: {risk_check (truncated, when present)}

[Doctor B] -> ...
[Doctor C] -> ...

Instructions for debate:
- If you disagree with an option score,
    explain why your score differs
- Challenge specific CoT steps if you see
    flaws
- If persuaded by another's reasoning, update
    your answer
```