MEDDAC: LEARNING DIRECTIONAL COMMUNICATION VIA CORRECTION AND HARM MODELING FOR MEDICAL MULTI-AGENT REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-agent reasoning relies on communication among agents, yet peer messages can either correct an erroneous receiver or mislead an initially correct one. Moreover, agents may revise their answers through self-reflection alone, making it difficult to distinguish peer influence from ordinary second-pass reasoning. We present **MedDAC**, a framework for learning directional agent communication policies by modeling both corrective and harmful peer effects. MedDAC evaluates receiver-specific sender-subset interventions against matched self-review to estimate the incremental local and system-level effects of communication actions. MedDAC-Static aggregates these effects to learn a fixed sparse communication graph under a message budget, while MedDAC-Dynamic predicts question-conditioned communication actions from the current multi-agent state. Across five medical QA benchmarks, the complete MedDAC-Static pipeline achieves 74.57% macro-average accuracy, compared with 69.77% for single-agent inference. Under the same Majority Vote aggregation, its six-message policy achieves 73.13%, comparable to the 72.99% obtained with 12-message Full Communication. These results show that modeling both corrective and harmful peer effects enables selective communication while maintaining competitive reasoning accuracy with substantially fewer directed messages.

1 INTRODUCTION

Large language models (LLMs) are increasingly used as components of multi-agent reasoning systems, where several agents exchange answers, critiques, or intermediate reasoning before reaching a final decision. This direction builds on broader test-time reasoning ideas such as self-consistency and iterative self-refinement (Wang et al., 2023; Madaan et al., 2023), while multi-agent frameworks introduce explicit communication, role specialization, and structured collaboration (Li et al., 2023; Wu et al., 2024; Hong et al., 2024). Such interaction can expose agents to alternative solution paths and correct errors that persist under independent inference, motivating recent work on multi-agent debate, role-based collaboration, structured workflows, and communication-graph design (Du et al., 2024; Chan et al., 2024). This paradigm has also been explored in medical reasoning, where difficult questions may benefit from multiple reasoning perspectives (Wang et al., 2025; Singhal et al., 2025).

However, exchanging more information does not always lead to better decisions. A peer message may correct an erroneous answer, but it may also cause an agent to abandon an answer that was already correct. Figure 1 illustrates this phenomenon with a medical reasoning example. The same receiver that fails to correct itself under self-review can be repaired by a helpful peer, while a directional interaction can also mislead an initially correct agent. Peer influence is therefore directional and asymmetric: an agent that is useful to one receiver need not be useful to another, and reversing sender and receiver can change the outcome.

Sparse communication saves messages, but edge count alone does not identify useful interactions. We therefore score each peer intervention against matched self-review to separate peer influence from ordinary reconsideration; unselected agents retain their initial responses.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

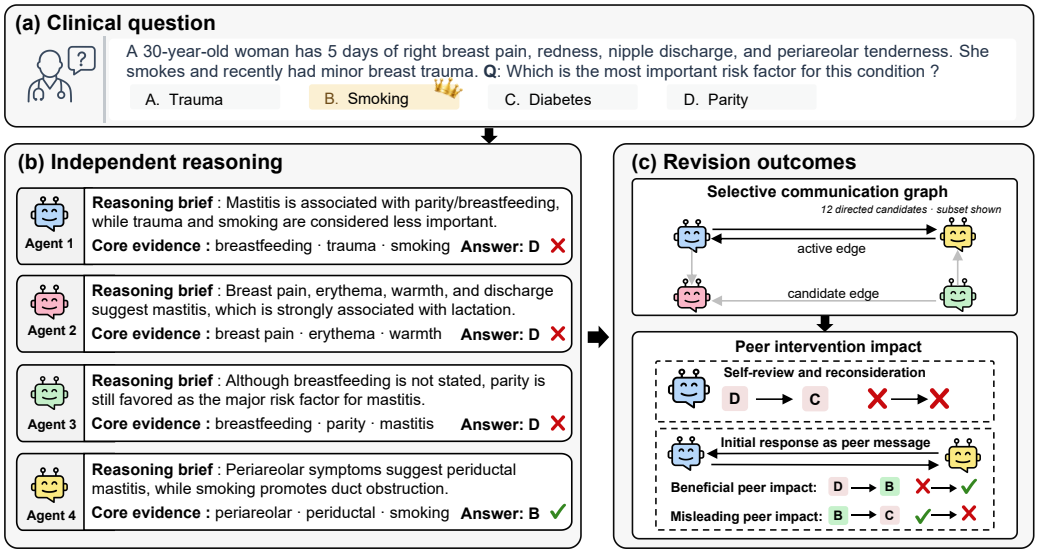

Figure 1: **Example of Directional Peer Influence.** A medical QA example illustrates three distinct revision outcomes: self-reconsideration may fail to repair an error, a helpful peer may turn an incorrect answer into a correct one, and a misleading peer may overturn an initially correct answer. The contrast motivates estimating communication effects relative to a matched self-review condition instead of assuming that every peer message is beneficial.

MedDAC evaluates each receiver–sender-subset action jointly, then selects a graph subject to a directed-message budget. We compare a fixed graph learned from calibration questions with an instance-specific graph predicted from the initial multi-agent state. On five medical QA benchmarks, the complete Static pipeline reaches 74.57% macro-average accuracy against 69.77% for independent inference. Under matched Majority Vote aggregation, its six-message graph reaches 73.13% against 72.99% for 12-message Full Communication. These comparisons describe the full pipeline and the controlled communication policy, respectively; they should not be conflated. The primary contributions of this work are:

- We propose **MedDAC**, a **matched interaction evaluation protocol** for multi-agent reasoning that compares each peer intervention against a receiver-specific self-review control. This design separates the effect of peer exposure from answer changes that may arise through ordinary second-pass reasoning, and quantifies both beneficial correction and harmful revision at the receiver and system levels.
- We formulate selective communication as **receiver-specific sender-subset selection** under a directed-message budget. Rather than assigning utilities to individual edges and assuming pairwise additivity, MedDAC directly evaluates multi-sender intervention actions. Based on these estimates, MedDAC-Static learns a fixed sparse graph, while MedDAC-Dynamic predicts question-conditioned communication actions from the current multi-agent state.
- We evaluate MedDAC across five medical QA benchmarks with multiple backbone models. With Qwen3.8-27B, the complete MedDAC-Static pipeline reaches a 74.57% macro-average accuracy compared with 69.77% for single-agent inference. Under the same Majority Vote aggregation, its six-message policy reaches 73.13%, compared with 72.52% for Random-K6 and 72.99% for 12-message Full Communication, showing that learned selection can retain competitive accuracy using half the directed-message budget of full communication.

2 Related Work

LLM multi-agent collaboration. Multi-agent debate and round-table discussion use repeated exchanges to refine answers (Du et al., 2024; Chen et al., 2024), while role-playing and workflow-oriented systems organize interaction through specialized agents or structured protocols (Li et al.,

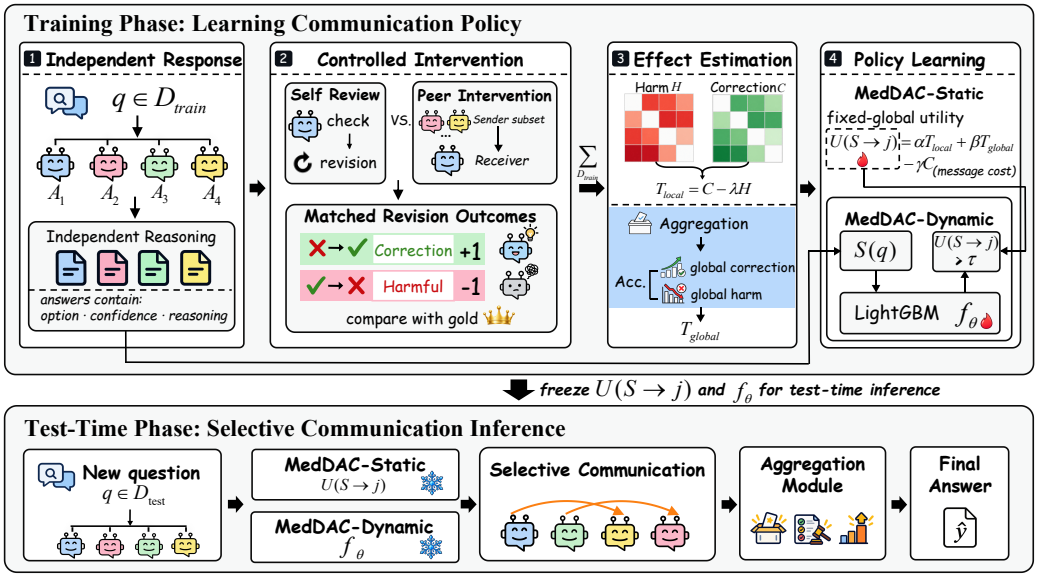

Figure 2: **MedDAC Framework.** During calibration, independent responses are paired with matched self-review and peer interventions to estimate receiver-level and system-level effects. MedDAC-Static aggregates these effects into fixed sender-subset utilities, whereas MedDAC-Dynamic predicts question-conditioned effects from the current multi-agent state. Both policies construct a sparse communication graph before final answer aggregation.

2023; Wu et al., 2024; Hong et al., 2024). ChatEval further shows how multi-agent discussion can improve LLM-based evaluation, with agent-role diversity affecting the quality of the collective process (Chan et al., 2024). Recent surveys summarize communication structure, coordination mechanisms, and open challenges in LLM-based multi-agent systems (Guo et al., 2024; Tran et al., 2025).

Medical multi-agent reasoning. In medical reasoning, MedAgents organizes specialist discussion, while MDAgents adapts the collaboration structure to question complexity (Tang et al., 2024; Kim et al., 2024). More recent medical-agent work studies logic-driven collaboration and adaptive refinement, and surveys highlight coordination, evaluation, and reliability as recurring design issues (Ma et al., 2026; Miao et al., 2026; Wang et al., 2025). Our focus is narrower: the effect of exposing a particular receiver to a particular sender subset before one revision.

Selective communication and revision effects. AgentPrune removes redundant communication links, and G-Designer learns question-adaptive topologies (Zhang et al., 2025a;b). MedDAC instead scores candidate actions from matched receiver and group outcomes before budgeted selection. Self-refinement can improve outputs in some settings (Madaan et al., 2023; Shinn et al., 2023), but self-correction without external feedback can also be unreliable (Huang et al., 2024). LLMs additionally respond to peer agreement and can be easier to mislead than to correct (Zhu et al., 2025; Qu et al., 2026). Recent no-source controls further show that repeated answer content and speaker attribution are distinct influences (Hu & Qu, 2026). Our peer-versus-self contrast measures the effect of the full peer packet relative to a second-pass control; it does not isolate speaker identity alone.

3 METHOD

Figure 2 summarizes the MedDAC pipeline. The framework contains two stages. During calibration, we collect matched self-review and sender-subset intervention outcomes, estimate their local and system-level effects, and learn communication utilities. At inference time, these learned quantities are frozen and used by either a static or a question-adaptive policy.

3.1 COMMUNICATION ACTIONS AND CALIBRATION

We consider a multi-agent system with N agents $\mathcal{V} = \{1, \ldots, N\}$. For each question q, every agent first reasons independently and produces an answer, a confidence score, and a structured reasoning packet and a reasoning trace. A communication action is defined as $a = (S, j)$, where $\varnothing \neq S \subseteq \mathcal{V} \setminus \{j\}$, j is the receiver, and S is a non-empty subset of the remaining agents. Executing a exposes receiver j to the response packets of all agents in S, after which the receiver performs one revision. Its directed-message cost is $c(a) = |S|$.

Our implementation uses four agents, so each receiver has $2^3 - 1 = 7$ possible non-empty sender subsets, yielding $4 \times 7 = 28$ candidate actions. Importantly, a multi-sender subset is treated as one intervention condition and produces one receiver revision; its utility is measured directly rather than obtained by summing its constituent pairwise edges.

For calibration, we retain questions with mixed independent correctness—at least one agent is correct and at least one is incorrect. For every retained question and receiver, we construct one matched self-review condition and all candidate sender-subset intervention conditions. This restriction provides both initially correct and initially incorrect receivers, but excludes unanimous initial votes from the labeled effect-prediction data. Policy behavior on such questions at inference is therefore an extrapolation beyond this calibration subset.

3.2 MATCHED INTERACTION EVALUATION

A before–after comparison can conflate peer exposure with changes that arise from ordinary second-pass reasoning, since an LLM may revise its answer when asked to reconsider the same question even without external information. MedDAC therefore evaluates every communication action against a matched self-review control. In the *self-review* condition, receiver j reconsiders its own initial response without peer messages; in the corresponding *peer-intervention* condition, the same receiver revises after observing the response packets from S.

Let Y_{self} and Y_{peer} indicate whether the receiver is correct after self-review and peer intervention, respectively, and let Z_{self} and Z_{peer} indicate whether the final group decision is correct when the receiver's output is replaced by the corresponding self-review or peer-intervention response. We define the receiver- and system-level effects as $\Delta_{\text{local}}(q, a) = Y_{\text{peer}} - Y_{\text{self}}$ and $\Delta_{\text{global}}(q, a) = Z_{\text{peer}} - Z_{\text{self}}$, respectively. Both take values in $\{-1, 0, +1\}$.

For global-effect estimation, the aggregation rule is fixed to Majority Vote with confidence-based tie breaking under both conditions. These differences measure peer intervention relative to self-review rather than relative to leaving the original answer unchanged. We use them as selection scores, while a receiver with no selected action retains its initial answer at inference. The complete policy is evaluated under this actual inference protocol.

3.3 STATIC COMMUNICATION POLICY

The question-level effects are aggregated over the calibration set to characterize each communication action. For the local component, $C(a)$ denotes the mean local effect when the receiver was initially incorrect, while $H(a)$ measures the corresponding harmful effect when the receiver was initially correct. For the system component, $E_{\text{gain}}(a)$ denotes the mean global improvement when the matched self-review group decision is incorrect, whereas $E_{\text{harm}}(a)$ measures degradation when that decision is already correct.

We define $T_{\text{local}}(a) = C(a) - \lambda_h H(a)$ and $T_{\text{global}}(a) = E_{\text{gain}}(a) - \lambda_g E_{\text{harm}}(a)$. The resulting static utility is $U(a) = \alpha T_{\text{local}}(a) + \beta T_{\text{global}}(a) - \gamma c(a)$, where α and β balance receiver- and system-level effects, and γ penalizes communication cost.

MedDAC-Static does not rank individual directed edges independently. Instead, it searches over receiver-specific sender-subset actions. A feasible graph assigns at most one action to each receiver

while constraining the total number of directed messages by budget B:

$$G^*_{\text{static}} = \arg\max_{G \subseteq \mathcal{A}} \sum_{a \in G} U(a),$$
$$\text{s.t.} \quad \sum_{a \in G} c(a) \leq B, \qquad |G \cap \mathcal{A}_j| \leq 1, \quad \forall j \in \mathcal{V}. \tag{1}$$

For four agents, each receiver has seven sender subsets plus the option of no communication, giving at most $8^4 = 4096$ graph configurations. We perform exact search over feasible configurations. The action scores are estimated from MedQA calibration data. We use $B = 6$ as an efficiency-oriented operating point, corresponding to half of the 12 directed messages in Full Communication, and freeze the resulting policy before test-time evaluation. The sum of single-receiver scores provides a tractable graph-selection surrogate, although it does not explicitly represent interactions among simultaneous receiver revisions.

3.4 Dynamic Communication Policy

A sender subset that is useful on average need not be useful for every question. MedDAC-Dynamic therefore predicts the effect of each candidate action conditioned on the current independent multi-agent state. For each question and candidate action, we construct features describing the receiver state, sender confidence and agreement, vote structure, answer diversity, sender-subset composition, and historical action statistics. Two LightGBM regressors predict $\widehat{\Delta}_{\text{local}}(q, a)$ and $\widehat{\Delta}_{\text{global}}(q, a)$, while question-grouped out-of-fold construction prevents a question from contributing to its own historical prior features.

The predicted effects define the question-conditioned utility $\widehat{U}_q(a) = \alpha\widehat{\Delta}_{\text{local}}(q, a) + \beta\widehat{\Delta}_{\text{global}}(q, a) - \gamma c(a)$. Actions below a validation-selected threshold τ are discarded, and the remaining actions enter the same budgeted graph optimization as Eq. 1, replacing $U(a)$ with $\widehat{U}_q(a)$. The effect predictors, historical priors, and threshold are frozen before test-time evaluation.

3.5 Selective Communication at Inference

At test time, all agents first answer independently. MedDAC-Static applies the same frozen graph to every question, whereas MedDAC-Dynamic scores all candidate sender-subset actions from the current independent responses and constructs a question-specific graph. Each selected receiver performs at most one peer-induced revision using the response packets from its assigned sender subset. All selected actions are executed from the same independent-response state, so a revised receiver output is not recursively reused as a new sender message within the same communication round. Agents without a selected action retain their independent responses.

The resulting post-communication responses are passed to a final aggregation module. Our complete system uses **Consensus Refinement**, whereas controlled communication experiments use the same Majority Vote aggregation for all policies. No gold labels or test-set outcomes are used for communication selection at inference time.

4 Experiments

4.1 Experimental Setup

Tasks and Benchmarks. We evaluate MedDAC on MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), and MedXpertQA (Zuo et al., 2025). We also use the medical subset of MMLU (Hendrycks et al., 2021) and PubMedQA (Jin et al., 2019). We use 1,273 questions from MedQA, 4,183 from MedMCQA, 2,450 from MedXpertQA, 1,089 from MMLU-Medical, and 1,000 from PubMedQA. All datasets are converted into a unified multiple-choice format and evaluated using exact option accuracy.

We use the MedQA training split, containing 10,178 questions, to construct calibration data. With four agents independently answering each question, this produces 40,712 initial agent responses.

Calibration statistics and effect predictors are learned from training data. Validation data are used to select the Dynamic action threshold τ, while the communication budget is fixed to $B = 6$. All policies are frozen before test-time evaluation.

Baselines. We compare MedDAC with single-agent inference, Self-Refine, Majority Vote, confidence-based aggregation, Full Communication, one-round multi-agent debate, and published multi-agent systems, including ReConcile, MDAgents, MedAgents, MedLA, and MAD-M^2 (Chen et al., 2024; Kim et al., 2024; Tang et al., 2024; Ma et al., 2026; Tian et al., 2026).

The complete MedDAC system reports both Static and Dynamic variants.

Implementation Details. Our main experiments use **Qwen3.8-27B** as the backbone. MedDAC uses four agents, yielding 12 possible directed messages under Full Communication. We set the selective-communication budget to $B = 6$, corresponding to 50% of the full-communication budget. This provides an efficiency-oriented operating point for evaluating whether learned routing can retain reasoning performance with substantially fewer directed messages.

For Static, the 28 sender-subset utilities are estimated on calibration data and the resulting budget-constrained graph is frozen before evaluation. For Dynamic, two LightGBM effect predictors score the same 28 candidate actions for each question.

The complete-system results use Consensus Refinement (ReConcile-style) as final aggregation. To isolate the contribution of communication selection, the communication controls use the same Majority Vote aggregation with confidence-based tie breaking.

Evaluation Metrics. We report accuracy on each benchmark and the unweighted macro average of the five dataset-level accuracies.

4.2 Main Results

Table 1 shows the performance comparison with MedDAC and SOTA systems, where MedDAC uses Consensus Refinement as the final aggregation stage. MedDAC-Static and MedDAC-Dynamic achieve 74.57% and 74.56% macro-average accuracy, respectively, compared with 69.77% for single-agent Vanilla inference. Among the evaluated existing multi-agent systems, ReConcile is the strongest baseline at 74.20%, leaving a comparatively modest complete-system margin.

The gains are not uniform across benchmarks. Both MedDAC variants reach 91.99% on MedQA, while Static attains 35.10% on MedXpertQA. In contrast, both variants fall below Vanilla on PubMedQA. The nearly identical Static and Dynamic averages further suggest that question-conditioned routing does not provide a clear accuracy advantage in the complete-system setting. Because the aggregation rules differ across methods in Table 1, these results should not be interpreted as isolating communication selection alone.

4.3 Communication Controls

To isolate communication selection, Table 2 holds final aggregation fixed to Majority Vote. Both Random-K6 and MedDAC operate under the same $B = 6$ constraint—half of the 12-message Full Communication budget—so their comparison separates learned selection from sparsity alone.

Random sparsification already preserves much of the performance of Full Communication: Random-K6 reaches 72.52% macro-average accuracy versus 72.99% for Full Communication. Under the same six-message budget, MedDAC-Static further improves the average to 73.13%, while Dynamic reaches 72.90%. Thus, sparsity alone accounts for part of the benefit, with learned selection providing an additional 0.61 percentage-point gain over Random-K6 for Static. These point-estimate differences are small and, without uncertainty estimates, do not establish a reliable accuracy advantage.

The effect is also benchmark-dependent. Static shows its largest gain over Random-K6 on MedXpertQA (31.88% vs. 29.39%) and also exceeds Full Communication on this benchmark (31.88% vs. 30.29%). PubMedQA, in contrast, does not benefit from the learned selective policy. Although

Table 1: Performance comparison with single-agent, heuristic multi-agent, and existing multi-agent methods on five medical QA benchmarks. The base LLM is consistently set as Qwen3.8-27B for all methods. We **bold** the best results and underline the runner-ups. Colored arrows denote absolute changes relative to Vanilla.

Method	MedQA	MedMCQA	MedXpertQA	MMLU	PubMedQA	Avg.
Vanilla	84.76	69.28	26.37	89.53	78.90	69.77
Self-Refine (Madaan et al., 2023)	82.40↓2.36	68.80↓0.48	29.06↑2.69	87.42↓2.11	63.30↓15.60	66.20↓3.57
Majority Vote	86.49↑1.73	71.65↑2.37	27.31↑0.94	89.16↓0.37	79.50↑0.60	70.82↑1.05
Confidence-based	87.12↑2.36	72.05↑2.77	27.71↑1.34	89.62↑0.09	**79.70**↑0.80	71.24↑1.47
Full Communication	89.87↑5.11	73.66↑4.38	30.29↑3.92	91.55↑2.02	79.60↑0.70	72.99↑3.22
MAD-1R (Du et al., 2024)	67.64↓17.12	65.26↓4.02	14.65↓11.72	80.72↓8.81	77.40↓1.50	61.13↓8.64
ReConcile (Chen et al., 2024)	89.47↑4.71	75.23↑5.95	34.37↑8.00	92.81↑3.28	79.10↑0.20	74.20↑4.43
MDAgents (Kim et al., 2024)	87.51↑2.75	70.85↑1.57	29.71↑3.34	89.91↑0.38	54.20↓24.70	66.45↓3.32
MedAgents (Tang et al., 2024)	**92.22**↑7.46	75.23↑5.95	22.08↓4.29	92.29↑2.76	78.00↓0.90	71.96↑2.19
MedLA (Ma et al., 2026)	89.16↑4.40	74.44↑5.16	31.23↑4.86	92.01↑2.48	79.60↑0.70	73.29↑3.52
MAD-M^2 (Tian et al., 2026)	89.40↑4.64	74.30↑5.02	29.22↑2.85	91.55↑2.02	76.60↓2.30	72.21↑2.44
MedDAC-Dynamic	91.99↑7.23	**75.59**↑6.31	34.37↑8.00	**92.93**↑3.40	77.90↓1.00	74.56↑4.79
MedDAC-Static	91.99↑7.23	75.50↑6.22	**35.10**↑8.73	92.65↑3.12	77.60↓1.30	**74.57**↑4.80

Table 2: Communication-policy controls on Qwen3.8-27B under the same Majority Vote aggregation. Random-K6 and MedDAC use a six-message budget, whereas Full Communication enables all 12 directed messages.

Method	MedQA	MedMCQA	MedXpertQA	MMLU	PubMedQA	Avg.
Vanilla	84.76	69.28	26.37	89.53	78.90	69.77
Random-K6	89.39	73.39	29.39	90.91	79.50	72.52
Full Communication	89.87	**73.66**	30.29	**91.55**	**79.60**	72.99
MedDAC-Dynamic	89.40	72.94	31.63	**91.55**	79.00	72.90
MedDAC-Static	**89.95**	73.49	**31.88**	**91.55**	78.80	**73.13**

MedDAC uses fewer directed messages than Full Communication, message count alone does not establish proportional reductions in model calls, tokens, or latency.

4.4 Static Policy Analysis

The learned Static policy exhibits substantial heterogeneity across receivers and sender subsets. For each receiver A_j, we denote the three available peers by S_1, S_2, S_3, ordered by agent index after excluding the receiver; the seven columns in Figure 3(a) therefore correspond to all non-empty subsets of these three peers.

Estimated utilities vary markedly across both receivers and subset compositions and are not monotonic in subset size (Figure 3(a)). This non-monotonicity should not be interpreted as direct evidence of sender interaction: negative marginal value and the explicit message-cost penalty can produce the same pattern. MedDAC therefore evaluates each sender subset directly rather than assuming pairwise additivity.

The validation sweep in Figure 3(b) characterizes the broader accuracy–communication trade-off. We use $B = 6$ as the main efficiency-oriented operating point, corresponding to half of the 12-message Full Communication budget, rather than as the validation-accuracy optimum. Figure 3(c) shows the resulting asymmetric frozen graph. Each selected subset induces one revision at its receiver, even when that subset contains multiple directed messages.

4.5 Dynamic Policy Analysis

Dynamic replaces the fixed action utility $U(a)$ with a question-conditioned score $\widehat{U}_q(a)$, allowing the selected graph to vary with the current independent-response state. The examples in Figure 4(a) illustrate how the utility landscape can change across questions, leading to different selected sender subsets.

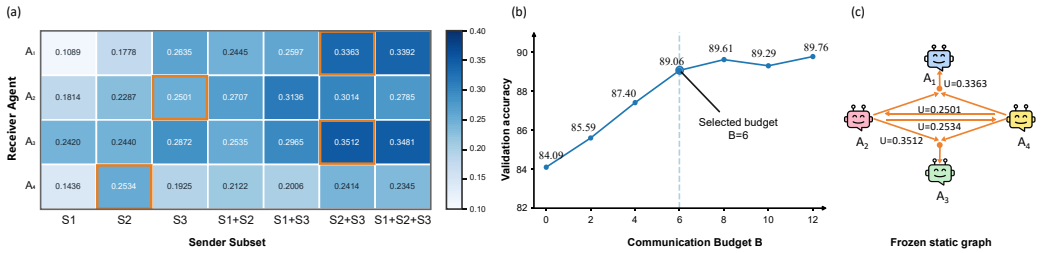

Figure 3: **Static Utility and Graph Selection.** **(a)** Static utilities for all $4 \times 7 = 28$ receiver-specific sender-subset actions; orange outlines mark actions selected by the frozen policy. Here S_1, S_2, S_3 denote the three peers available to each receiver in receiver-relative index order. **(b)** MedQA validation accuracy across directed-message budgets; $B = 6$, corresponding to half of the 12-message Full Communication budget, is used as the operating point in the main experiments. **(c)** The resulting frozen Static graph. A utility annotation belongs to the selected sender subset as a whole, even when that subset contains multiple directed messages.

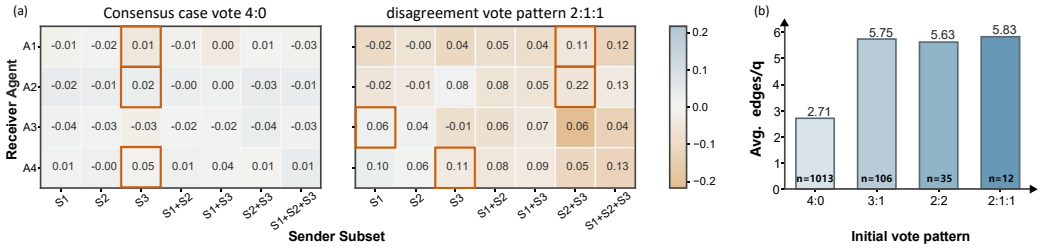

Figure 4: **Question-Adaptive Communication.** **(a)** Example question-specific sender-subset utility landscapes under a high-consensus state and a disagreement state. The same receiver-relative S_1, S_2, S_3 notation as Figure 3 is used. Orange outlines denote actions selected for the corresponding question. **(b)** Average number of selected directed messages per question grouped by the initial four-agent vote pattern. Dynamic communicates sparsely when agents already agree and approaches the communication budget when disagreement is larger.

This adaptivity also changes communication intensity. Unanimous 4:0 cases receive substantially fewer messages on average than the disagreement patterns shown in Figure 4(b), where message use approaches the communication budget. Because unanimous questions were excluded from the labeled effect-prediction data during calibration, this behavior should be interpreted descriptively rather than as evidence of calibrated intervention effects on such cases. Overall, Dynamic clearly adapts where communication is allocated, but this flexibility does not translate into higher macro-average accuracy than Static in the current experiments.

4.6 Cross-Backbone Comparison

The relative behavior of Static and Dynamic remains similar across the three evaluated backbones (Table 3). Dynamic is marginally higher on Qwen3-8B, whereas Static is marginally higher on Qwen3.8-27B and DeepSeek-V4-Flash; in all cases, the difference between the two policies is small. These results therefore provide no consistent evidence that either fixed or question-conditioned routing is uniformly more accurate across backbones.

The table should not be interpreted as a policy-transfer experiment. Because it does not include a backbone-specific no-communication baseline, differences in absolute accuracy across backbones cannot be attributed specifically to communication behavior.

4.7 Aggregation Ablation

Table 4 isolates final answer aggregation while keeping the underlying communication policy fixed. Consensus Refinement raises the macro-average accuracy from 72.90% to 74.56% for Dynamic (+1.66 points) and from 73.13% to 74.57% for Static (+1.44 points), while Weighted Vote and Final Judge yield smaller gains. Thus, the complete-system results in Table 1 reflect both communication selection and downstream answer consolidation.

Table 3: Cross-backbone comparison of MedDAC-Dynamic and MedDAC-Static under Majority Vote aggregation.

Backbone	Policy	MedQA	MedMCQA	MedXpertQA	MMLU	PubMedQA	Avg.
Qwen3-8B	Dynamic	**62.37**	**51.82**	**13.68**	**70.25**	75.40	**54.70**
	Static	**62.37**	51.34	13.56	70.16	**75.50**	54.59
Qwen3.8-27B	Dynamic	89.40	72.94	31.63	**91.55**	**79.00**	72.90
	Static	**89.95**	**73.49**	**31.88**	**91.55**	78.80	**73.13**
DeepSeek-V4-Flash	Dynamic	83.89	74.52	**24.94**	89.44	80.90	70.74
	Static	**84.37**	**75.11**	24.69	**89.90**	**81.20**	**71.05**

Table 4: Aggregation ablation on Qwen3.8-27B with fixed Dynamic or Static communication.

Policy	Aggregation	MedQA	MedMCQA	MedXpertQA	MMLU	PubMedQA	Avg.
Dynamic	Majority Vote	89.40	72.94	31.63	91.55	79.00	72.90
	Weighted Vote	89.87	73.10	31.92	91.55	**79.10**	73.11
	Final Judge	91.04	73.94	32.00	91.18	78.80	73.39
	Consensus Refinement	**91.99**	**75.59**	**34.37**	**92.93**	77.90	**74.56**
Static	Majority Vote	89.95	73.49	31.88	91.55	**78.80**	73.13
	Weighted Vote	90.73	73.70	32.33	91.18	78.70	73.33
	Final Judge	90.18	74.13	31.51	91.64	**78.80**	73.25
	Consensus Refinement	**91.99**	**75.50**	**35.10**	**92.65**	77.60	**74.57**

The effect is benchmark-dependent. On MedXpertQA, Consensus Refinement improves Dynamic from 31.63% to 34.37% and Static from 31.88% to 35.10%. PubMedQA shows the opposite pattern, decreasing from 79.00% to 77.90% for Dynamic and from 78.80% to 77.60% for Static. Accordingly, Table 2 provides the cleaner evaluation of communication selection itself.

5 Discussion and Conclusion

MedDAC studies selective multi-agent communication by comparing peer interventions with matched self-review and modeling both corrective and harmful effects. Across five benchmarks, the complete MedDAC-Static pipeline reaches 74.57% macro-average accuracy versus 69.77% for independent inference. Under controlled Majority Vote, its six-message graph reaches 73.13%, compared with 72.52% for Random-K6 and 72.99% for 12-message Full Communication. These results suggest that random sparsification already removes substantial communication redundancy, while learned selection provides a further, modest improvement under the same message budget. The benefit is task-dependent and is most pronounced on MedXpertQA.

Dynamic provides question-dependent communication rather than a clear accuracy advantage. It allocates fewer messages to unanimous cases and more to disagreement cases, yet remains close to Static in both the main and cross-backbone results. The aggregation ablation further shows that part of the complete-system improvement comes from final answer consolidation rather than routing alone. PubMedQA shows the opposite trend, where additional revision and Consensus Refinement do not improve accuracy, indicating that communication benefits are not universal.

Several limitations remain. The peer-versus-self contrast is a selection surrogate, while unselected receivers retain their initial answers. Calibration excludes unanimous questions, leaving them outside the labeled support of the effect predictors, and the graph objective does not model interactions among simultaneous receiver revisions. Our budget counts directed messages rather than end-to-end cost, so fewer messages do not imply proportional savings in tokens, latency, or model calls; multiple-choice medical QA also does not establish clinical reliability. Overall, MedDAC retains much of full-communication performance with a smaller graph while providing a modest, task-dependent benefit over random sparsification. Future work should improve dynamic calibration, model joint receiver interactions, and measure computational cost directly.

AI USE STATEMENT

Generative AI tools were used for language editing, code generation and debugging, literature search and summarization, and manuscript organization. All AI-assisted text, code, analyses, and citations were reviewed and verified by the authors. We take responsibility for the final content of this work, including the reported results, analyses, claims, and citations.

ETHICS STATEMENT

This work evaluates multi-agent communication methods on publicly available medical question-answering benchmarks and does not involve direct interaction with patients or human subjects. The proposed method is intended for research purposes and should not be interpreted as providing clinical advice or replacing professional medical judgment.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the datasets, model configurations, communication procedures, evaluation metrics, and experimental settings in the main text and supplementary materials. Additional implementation details, prompts, and hyperparameter settings required to reproduce the experiments are also provided. To support reproducibility during anonymous review, we release the implementation of the calibration, selective communication, static and dynamic policy construction, inference, and evaluation pipeline in an anonymous repository at `https://anonymous.4open.science/r/meddac-0CFE/`.

REFERENCES

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. ChatEval: Towards better LLM-based evaluators through multi-agent debate. In *International Conference on Learning Representations*, 2024. URL `https://proceedings.iclr.cc/paper_files/paper/2024/hash/25cc3adf8c85f7c70989cb8a97a691a7-Abstract-Conference.html`.

Justin Chen, Swarnadeep Saha, and Mohit Bansal. ReConcile: Round-table conference improves reasoning via consensus among diverse LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7066–7085. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.381. URL `https://aclanthology.org/2024.acl-long.381/`.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11733–11763. PMLR, 2024. URL `https://proceedings.mlr.press/v235/du24e.html`.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges, 2024. URL `https://arxiv.org/abs/2402.01680`.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=d7KBjmI3GmQ`.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. MetaGPT: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*, 2024. URL `https://proceedings.iclr.cc/paper_files/paper/2024/hash/6507b115562bb0a305f1958ccc87355a-Abstract-Conference.html`.

Yibo Hu and Jiaming Qu. Most LLM conformity needs no speaker: Measuring the speaker-free floor in peer-pressure benchmarks, 2026. URL https://arxiv.org/abs/2607.05545.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/8b4add8b0aa8749d80a34ca5d941c355-Abstract-Conference.html.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021. doi: 10.3390/app11146421. URL https://www.mdpi.com/2076-3417/11/14/6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2567–2577. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1259. URL https://aclanthology.org/D19-1259/.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. MDAgents: An adaptive collaboration of LLMs for medical decision-making. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2522. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/90d1fc07f46e31387978b88e7e057a31-Abstract-Conference.html.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for "mind" exploration of large scale language model society. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a3621ee907def47c1b952ade25c67698-Abstract-Conference.html.

Siqi Ma, Jiajie Huang, Fan Zhang, Jinlin Wu, Yue Shen, Guohui Fan, Zhu Zhang, and Zelin Zang. MedLA: A logic-driven multi-agent framework for complex medical reasoning with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 845–853, 2026. doi: 10.1609/aaai.v40i2.37052. URL https://ojs.aaai.org/index.php/AAAI/article/view/37052.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Yongming Miao, Jiaxin Wen, Yuemei Luo, and Jin Li. MedARC: Adaptive multi-agent refinement and collaboration for enhanced medical reasoning in large language models. *International Journal of Medical Informatics*, 206:106136, 2026. doi: 10.1016/j.ijmedinf.2025.106136. URL https://doi.org/10.1016/j.ijmedinf.2025.106136.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260. PMLR, 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

Jiaming Qu, Lucheng Fu, and Yibo Hu. Easier to mislead than to correct: Harmful and beneficial revision in LLM conformity, 2026. URL https://arxiv.org/abs/2606.01637.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Karan Singhal et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31:943–950, 2025. doi: 10.1038/s41591-024-03423-7. URL https://www.nature.com/articles/s41591-024-03423-7.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. MedAgents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 599–621. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.33. URL https://aclanthology.org/2024.findings-acl.33/.

Hongduan Tian, Xiao Feng, Ziyuan Zhao, Xiangyu Zhu, Rolan Yan, and Bo Han. Multi-agent debate with memory masking. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=EdTt8nMAMA.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D. Nguyen. Multi-agent collaboration mechanisms: A survey of LLMs, 2025. URL https://arxiv.org/abs/2501.06322.

Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Jiaming Ji, Wenting Chen, Xiang Li, and Yixuan Yuan. A survey of LLM-based agents in medicine: How far are we from baymax? In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10345–10359. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.539. URL https://aclanthology.org/2025.findings-acl.539/.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. In *Proceedings of the First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=BAakY1hNKS.

Guibin Zhang, Yanwei Yue, Zhixun Li, Sukwon Yun, Guancheng Wan, Kun Wang, Dawei Cheng, Jeffrey Xu Yu, and Tianlong Chen. Cut the crap: An economical communication pipeline for LLM-based multi-agent systems. In *International Conference on Learning Representations*, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/bbc461518c59a2a8d64e70e2c38c4a0e-Abstract-Conference.html.

Guibin Zhang, Yanwei Yue, Xiangguo Sun, Guancheng Wan, Miao Yu, Junfeng Fang, Kun Wang, Tianlong Chen, and Dawei Cheng. G-Designer: Architecting multi-agent communication topologies via graph neural networks. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 76678–76692. PMLR, 2025b. URL https://proceedings.mlr.press/v267/zhang25cu.html.

Xiaochen Zhu, Caiqi Zhang, Tom Stafford, Nigel Collier, and Andreas Vlachos. Conformity in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3854–3872. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.195. URL https://aclanthology.org/2025.acl-long.195/.

Yuxin Zuo, Shang Qu, Yifei Li, Zhang-Ren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. MedXpertQA: Benchmarking expert-level medical reasoning and understanding. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 80961–80990. PMLR, 2025. URL https://proceedings.mlr.press/v267/zuo25a.html.

A Implementation and Experimental Details

A.1 Benchmark Details

Table A1 summarizes the datasets and splits used in our experiments.MedQA training data are used for calibration, while its validation split is used for Dynamic threshold selection and validation analysis. The communication budget is fixed independently at $B = 6$. All evaluation benchmarks are converted into a unified multiple-choice representation and evaluated using exact option accuracy.

Table A1: **Benchmark and split summary.** "Original Format" describes each benchmark before preprocessing. MedQA training and validation splits are used for policy construction, whereas the listed test splits are used only for final evaluation.

Dataset	Split	Usage	Original Format	Size	Domain
MedQA	Train	Calibration	Question + Answer	10,178	US Medical Licensing Examination questions
	Validation	Policy selection	Question + Answer	1272	
	Test	Evaluation	Question + Answer	1,273	
MedMCQA	Test	Evaluation	Question + Answer	4,183	Indian medical entrance and professional examination questions
PubMedQA	Test	Evaluation	Question + Context + Answer	1,000	Biomedical research questions derived from PubMed abstracts
MMLU-Medical	Test	Evaluation	Question + Answer	1,089	Medical subjects from MMLU
MedXpertQA-Text	Test	Evaluation	Question + Answer	2,450	Expert-level medical reasoning across multiple specialties

A.2 Hyperparameter and Implementation Settings

Table A2 summarizes the principal hyperparameters used by MedDAC. The communication budget is fixed to $B = 6$, corresponding to half of the 12 directed messages available under Full Communication. For the Dynamic policy, the action threshold τ is selected on the MedQA validation split and frozen before test-time evaluation.

Table A2: **MedDAC hyperparameter and implementation settings.** The listed values are frozen before test-time evaluation.

Component	Hyperparameter	Value
Agent inference	Number of agents N	4
	Temperatures	$\{0.1, 0.4, 0.7, 1.0\}$
	Top-p	0.9
	Max generation tokens	384
	Thinking mode	Disabled
Communication utility	Local-effect weight α	1.0
	System-effect weight β	1.0
	Communication-cost weight γ	0.02
	Local harm penalty λ_h	1.0
	System harm penalty λ_g	1.0
	Message budget B	6
Dynamic policy	Threshold candidates	$\{-0.05, 0, 0.05, 0.10, 0.15, 0.20\}$
	Selected threshold τ	0.0
	Validation parsimony tolerance	0.001
	Grouped OOF folds	5
LightGBM	Objective	Regression L2
	Number of estimators	400
	Learning rate	0.03
	Number of leaves	15
	Maximum depth	4
	Minimum child samples	30
	Row / feature subsampling	0.90 / 0.90
	L2 regularization	1.0

B Prompts and Communication Protocol

This section provides the prompting and post-communication decision procedures used in MedDAC. Unless otherwise specified, the four reasoning agents share the same prompt and differ only in decoding temperature. Their temperatures are 0.1, 0.4, 0.7, and 1.0, with $top_p = 0.9$. The main Qwen3.8-27B experiments use non-thinking inference and a maximum generation length of 384 tokens for the MedDAC reasoning and revision stages.

B.1 Independent Reasoning

Each agent first answers the medical question independently. The system instruction is:

```
You are an independent medical reasoning agent answering a
multiple-choice medical question.  Use the provided clinical
evidence and medical knowledge to select the single best
option.  Return only the requested JSON object.  Do not
output hidden chain-of-thought or a long derivation.
```

The required output packet has the following structure:

```
{
  "answer": "A",
```

```
  "confidence": 0.80,
  "brief_reason": "A concise medical justification.",
  "key_discriminator": "The most decisive distinguishing fact.",
  "closest_alternative": "B",
  "uncertainty_source": "The main unresolved issue."
}
```

Thus, each agent packet contains six fields: `answer`, `confidence`, `brief_reason`, `key_discriminator`, `closest_alternative`, and `uncertainty_source`. Confidence is constrained to $[0, 1]$, and the answer must be one of the valid option labels.

B.2 Matched Self-Review

For the matched self-review condition, the receiver is shown the original question together with its own independent response packet, but receives no peer information. The revision prompt explicitly marks

```
ADDITIONAL PEER INFORMATION: None.
```

and asks the receiver to reconsider the question once using only the medical evidence. The system instruction is shared with peer intervention:

```
You are a medical reasoning agent revising a previous
multiple-choice answer.  Use the original evidence and
medical knowledge to make the best final judgment.
Anonymous peer information may be useful or wrong;
evaluate its content rather than deferring to it.  Return
only the requested JSON object.  Do not output hidden
chain-of-thought or a long derivation.
```

The revised response uses the same six-field JSON schema as the independent response. This construction provides the matched second-pass reference used to estimate the incremental effect of peer exposure.

B.3 Sender-Subset Peer Intervention

For a communication action $a = (S, j)$, receiver j is shown its own independent response together with the response packets of the selected sender subset S. Sender identity is not exposed as a semantic role; messages are presented as anonymous peer information.

For a single sender, the prompt contains:

```
ADDITIONAL INFORMATION FROM ONE ANONYMOUS PEER:
[peer response packet]
The peer may be correct or wrong.  Evaluate it against the
evidence and reconsider once.
```

For a multi-sender action, the packets are presented jointly:

```
ANONYMOUS PEER MESSAGES:
Peer 1:  [response packet]
Peer 2:  [response packet]
...
The peers may be correct or wrong.  Evaluate each message
against the evidence.  Reconsider once and give your own
final answer.
```

All sender packets correspond to the same independent-response state. A receiver performs only one revision for a selected sender subset, and the revised response is not recursively reused as a sender message within the same communication round.

B.4 POST-COMMUNICATION READOUT

After selective communication, the resulting four-agent state is frozen before final answer readout. Table B1 summarizes the four downstream decision procedures evaluated in Table 4.

Table B1: **Post-communication readout protocols.** All methods start from the same frozen four-agent state produced by the corresponding Static or Dynamic communication policy.

Readout	Decision procedure	Extra LLM calls	Further discussion
Majority Vote	Majority over the four post-communication answers; vote ties are resolved by summed confidence among tied options.	0	No
Weighted Vote	ReConcile confidence recalibration followed by confidence-weighted voting over the frozen post-communication answers.	0	No
Final Judge	One LLM final decision-maker receives all four frozen post-communication agent outputs and selects the final option.	1 / question	No
Consensus Refinement	ReConcile-style multi-agent deliberation initialized from the frozen MedDAC post-communication state, followed by confidence-weighted voting.	4–12 / question	Up to 3 rounds

B.5 MAJORITY AND CONFIDENCE-WEIGHTED VOTING

For Majority Vote, the option receiving the largest number of agent votes is selected. If multiple options are tied in vote count, the confidence scores of agents supporting each tied option are summed and the option with the largest confidence sum is selected. Any remaining tie is resolved according to the canonical option-label order.

Weighted Vote follows the confidence recalibration rule used by ReConcile. A raw confidence p is transformed into the voting weight

$$w(p) = \begin{cases} 1.0, & p = 1.0, \\ 0.8, & p \geq 0.9, \\ 0.5, & p \geq 0.8, \\ 0.3, & p > 0.6, \\ 0.1, & \text{otherwise.} \end{cases}$$

The weights are summed for each candidate option, and the option with the largest total weight is returned. This readout does not invoke the LLM again and therefore isolates the effect of confidence-weighted aggregation from additional deliberation.

B.6 FINAL JUDGE

The Final Judge ablation follows the released AgentPrune FINALREFER readout structure. The four frozen post-communication agent outputs are supplied to a single final decision-maker. Its role instruction is:

```
You are the top decision-maker and are good at analyzing
and summarizing other people's opinions, finding errors and
giving final answers.
```

The judge is instructed to select exactly one valid option label and to return only that label. This stage uses Qwen3.8-27B with temperature 0.7, $top_p = 1.0$, a maximum of 1024 generated tokens, and thinking disabled. It adds exactly one LLM call per question.

B.7 CONSENSUS REFINEMENT

Consensus Refinement follows the ReConcile discussion procedure but starts from the *frozen MedDAC post-communication state*; the independent reasoning and selective communication stages are not re-executed. In each discussion round, current responses are grouped by answer option. For each option, agents are shown the number of supporters, mean confidence, and the associated explanations.

Each of the four agents then independently revises its response using the following system instruction:

```
You are a medical expert participating in a multi-agent
discussion.  Review the opinions from other experts, then
provide your updated analysis.  You may change your opinion
if others' reasoning convinces you, or defend your position
with clear explanations.  Estimate your confidence in your
answer on a scale from 0.0 to 1.0.
```

Each updated response contains `reasoning`, `answer`, and `confidence`. Discussion stops early when all four agents agree, otherwise it continues for at most three rounds. Following the original ReConcile configuration used in our implementation, at least one discussion round is executed even when the frozen initial state already has unanimous answers. The final prediction is obtained using the confidence-recalibrated weighted vote described above.

The four discussion agents retain temperatures 0.1, 0.4, 0.7, and 1.0, respectively, with $top_p = 1.0$, a maximum of 1024 generated tokens per call, and thinking disabled.

C ALGORITHMIC DETAILS

C.1 CALIBRATION AND POLICY LEARNING

Algorithm 1 summarizes the calibration procedure used to construct the Static and Dynamic communication policies.

Algorithm 1 MedDAC Calibration and Policy Learning

Require: Calibration questions $\mathcal{D}_{\text{cal}}$, agents $\mathcal{V}$, message budget B
Ensure: Static action utilities $U(a)$ and trained Dynamic predictors
for each question $q \in \mathcal{D}_{\text{cal}}$ **do**
  Obtain independent responses $\{r_i(q)\}_{i \in \mathcal{V}}$
  if independent responses contain both correct and incorrect agents **then**
    for each receiver $j \in \mathcal{V}$ **do**
      Run matched self-review for receiver j
      for each non-empty sender subset $S \subseteq \mathcal{V} \setminus \{j\}$ **do**
        Execute peer intervention $a = (S, j)$
        Compute $\Delta_{\text{local}}(q, a)$ and $\Delta_{\text{global}}(q, a)$
        Store action features and intervention outcomes
      end for
    end for
  end if
end for
Aggregate intervention outcomes to estimate $C(a)$, $H(a)$, $E_{\text{gain}}(a)$, and $E_{\text{harm}}(a)$
Compute Static utilities $U(a) = \alpha T_{\text{local}}(a) + \beta T_{\text{global}}(a) - \gamma c(a)$
Train question-conditioned predictors for $\widehat{\Delta}_{\text{local}}(q, a)$ and $\widehat{\Delta}_{\text{global}}(q, a)$
Freeze Static utilities, Dynamic predictors, and validation-selected policy parameters

C.2 SELECTIVE COMMUNICATION INFERENCE

Algorithm 2 describes test-time selective communication for both Static and Dynamic policies.

Algorithm 2 MedDAC Selective Communication Inference

Require: Test question q, agents $\mathcal{V}$, policy type $P \in \{\text{STATIC}, \text{DYNAMIC}\}$, message budget B
Ensure: Final prediction $\hat{y}$

```
Obtain independent responses {r_i(q)}_{i∈V}
if P = STATIC then
    Use the frozen Static action utilities U(a)
else
    for each candidate action a = (S, j) do
        Extract features from the current independent-response state
        Predict Δ̂_local(q, a) and Δ̂_global(q, a)
        Compute Û_q(a) = α Δ̂_local(q, a) + β Δ̂_global(q, a) − γ c(a)
    end for
    Remove actions below threshold τ
end if
Solve the budget-constrained graph-selection problem under Σ_{a∈G} c(a) ≤ B
for each selected action (S, j) ∈ G do
    Provide receiver j with the independent response packets from senders in S
    Revise receiver j once
end for
Keep independent responses for agents without selected actions
Form the frozen post-communication agent state
Apply the selected final readout procedure
return final prediction ŷ
```

C.3 Computational Structure

With four agents, each receiver has $2^3 - 1 = 7$ non-empty sender subsets, yielding 28 candidate receiver–sender-subset actions. Static graph selection considers, for each receiver, the seven sender subsets plus no communication, giving at most $8^4 = 4096$ graph configurations. We therefore perform exact search over feasible configurations under the message budget. Dynamic uses the same candidate action space but replaces the fixed action utility with question-conditioned predicted utilities before solving the same constrained graph-selection problem.

D Qualitative Case Studies

We present two MedQA examples from the Qwen3.8-27B Dynamic policy to illustrate both successful correction and a characteristic failure mode. For readability, we denote the four temperature-diverse agents as A_1, A_2, A_3, and A_4, corresponding to temperatures 0.1, 0.4, 0.7, and 1.0, respectively.

The first case shows how selective communication can overturn an initially incorrect majority when one agent already contains the correct answer. The second case shows the opposite behavior: an initially unanimous and correct agent state becomes incorrect after communication-triggered revision. These cases are illustrative rather than estimates of average treatment effects.

D.1 Case 1: Correcting an Initially Wrong Majority

Question. A 36-year-old nursing home worker presents to the clinic with the complaints of breathlessness, cough, and night sweats for the past 2 months. She further expresses her concerns about the possibility of contracting tuberculosis as one of the patients under her care is being treated for tuberculosis. A PPD skin test is done and reads 11 mm on day 3. Chest X-ray demonstrates a cavitary lesion in the right upper lobe. The standard anti-tuberculosis medication regimen is started. At a follow-up appointment 3 months later the patient presents with fatigue. She has also been experiencing occasional dizziness, weakness, and numbness in her feet. Physical exam is positive for conjunctival pallor. Lab work is significant for a hemoglobin level of 10 g/dL and mean corpuscular volume of 68 fl. What is the most likely cause of her current symptoms?

Options.

A. Inhibition of ferrochelatase
B. Increased homocysteine degradation
C. Increased GABA production
D. Decreased ALA synthesis

The gold answer is **D**. Before communication, three agents select A and only A_3 selects D, producing an incorrect 3–1 majority for A. The Dynamic policy selects four receiver-specific actions with a total cost of six directed messages.

Table D1: **Case D1: selected communication actions.** The initially correct minority agent A_3 selects D, while A_1, A_2, and A_4 initially select A. All three initially incorrect agents revise to D after selective communication.

Receiver	Initial	Selected sender(s)	Action score	Revised
A_1 ($T = 0.1$)	A	A_3(D), A_4(A)	0.097	**D** (W→C)
A_2 ($T = 0.4$)	A	A_3(D), A_4(A)	0.080	**D** (W→C)
A_3 ($T = 0.7$)	**D**	A_4(A)	0.064	**D** (C→C)
A_4 ($T = 1.0$)	A	A_3(D)	0.145	**D** (W→C)
Group vote	A:3, D:1	–	–	**D:4**

Revision of A_1. Agent A_1 initially answers A and receives two peer packets: A_3 supports D while A_4 supports A. Despite receiving conflicting peer answers, A_1 revises to D with confidence 0.95.

> **Brief reason:** Isoniazid depletes pyridoxine (Vitamin B6), which is a cofactor for ALA synthase. This leads to decreased ALA synthesis, resulting in sideroblastic anemia (microcytic) and peripheral neuropathy. Isoniazid does not directly inhibit ferrochelatase.
>
> **Key discriminator:** Isoniazid causes B6 deficiency, which inhibits ALA synthase (rate-limiting step), not ferrochelatase.
>
> **Closest alternative:** A
> **Uncertainty source:** None reported.

Revision of A_2. Agent A_2 also initially answers A and receives the same pair of conflicting sender answers, D from A_3 and A from A_4. It changes to D with confidence 0.95.

> **Brief reason:** Isoniazid inhibits pyridoxal phosphate (Vitamin B6) metabolism. Vitamin B6 is a required cofactor for ALA synthase, the rate-limiting enzyme in heme synthesis. Deficiency leads to decreased ALA synthesis, resulting in sideroblastic anemia (microcytic) and peripheral neuropathy, matching the patient's presentation.
>
> **Key discriminator:** Isoniazid causes sideroblastic anemia by depleting Vitamin B6, which is a cofactor for ALA synthase, not by directly inhibiting ferrochelatase.
>
> **Closest alternative:** A
> **Uncertainty source:** None reported.

Revision of A_3. Agent A_3 is the only initially correct agent. It initially selects D but receives an incorrect A message from A_4. It retains D with confidence 0.95.

> **Brief reason:** Isoniazid causes pyridoxine (Vitamin B6) deficiency. B6 is a cofactor for ALA synthase, the rate-limiting enzyme in heme synthesis. Deficiency leads to decreased ALA synthesis, causing sideroblastic anemia (microcytic) and peripheral neuropathy. Ferrochelatase is not the primary target of B6 deficiency.
>
> **Key discriminator:** Vitamin B6 is a specific cofactor for ALA synthase, not ferrochelatase; deficiency impairs ALA synthesis.
>
> **Closest alternative:** A
> **Uncertainty source:** None reported.

Revision of A_4. Agent A_4 initially selects A and receives the correct D response from A_3. It changes from A to D with confidence 0.95.

> **Brief reason:** Isoniazid depletes pyridoxine (Vitamin B6). B6 is a cofactor for ALA synthase, the rate-limiting enzyme in heme synthesis. Deficiency impairs ALA synthesis, causing sideroblastic (microcytic) anemia and peripheral neuropathy. Option A is incorrect as isoniazid does not directly inhibit ferrochelatase.
>
> **Key discriminator:** Isoniazid's mechanism is Vitamin B6 depletion, which reduces ALA synthesis via ALA synthase inhibition, rather than direct inhibition of ferrochelatase.
>
> **Closest alternative:** A
> **Uncertainty source:** None reported.

Case interpretation. This case illustrates a setting in which useful information is present in a minority agent before communication. The initial group state is A, A, D, A, so ordinary Majority Vote returns the wrong answer A. The Dynamic policy routes information from A_3, the only agent initially selecting the correct answer D, to several receivers. After revision, A_1, A_2, and A_4 all change from A to D, while A_3 remains correct despite receiving an incorrect peer answer. The final group state becomes D, D, D, D.

Notably, A_1 and A_2 receive conflicting peer labels—one correct D response and one incorrect A response—rather than a unanimous peer signal. Their revisions nevertheless identify the clinically relevant distinction between inhibition of ALA synthesis through pyridoxine deficiency and direct inhibition of ferrochelatase. This example therefore illustrates how selective communication can expose an initially incorrect receiver to a useful minority signal without requiring all communicated peers to agree.

D.2 Case 2: Failure under Unanimous Initial Consensus

Question. A 22-year-old man is brought to the emergency department by police after he was found undressing himself outside in 110°F weather. The patient refuses to answer any questions as he suspects that his care team is spying on him. The patient recently started college and has presented to the emergency department once before for polysubstance intoxication. The patient attempts to assault a nurse, is restrained, and given diphenhydramine and haloperidol. When the patient is able to be examined, he is somnolent and non-responsive. His temperature is 104°F (40°C), blood pressure is 147/98 mmHg, pulse is 120/min, respirations are 17/min, and oxygen saturation is 98% on room air. He does not comply with any commands, and it is not possible to freely manipulate his extremities due to rigidity. His mucous membranes are dry. Which of the following is the most likely diagnosis?

Options.

> **A.** Catatonic schizophrenia
> **B.** Heat stroke
> **C.** Neuroleptic malignant syndrome
> **D.** Oversedation

The gold answer is **C**. In contrast to Case 1, all four agents initially select the correct answer C. The initial group state is therefore unanimously correct. Nevertheless, the Dynamic policy selects three one-message actions, after which three receivers change from C to B.

Table D2: **Case D2: communication-triggered failure from an initially unanimous correct state.** All selected senders themselves initially support the correct answer C, yet three receivers revise to the incorrect answer B.

Receiver	Initial	Selected sender	Action score	Revised
A_1 ($T = 0.1$)	**C**	A_4(C)	0.011	B (C→W)
A_2 ($T = 0.4$)	**C**	A_4(C)	0.021	B (C→W)
A_3 ($T = 0.7$)	**C**	–	–	**C**
A_4 ($T = 1.0$)	**C**	A_3(C)	0.051	B (C→W)
Group vote	**C:4**	–	–	B:3, C:1

Revision of A_1. Agent A_1 initially selects the gold answer C and receives a peer packet from A_4, which also initially selects C. Nevertheless, A_1 changes to B with confidence 0.90.

> **Brief reason:** The patient was found undressing in 110°F weather, a classic presentation for heat stroke. The combination of hyperthermia (104°F), altered mental status, and dry mucous membranes (anhidrosis) in this environmental context is diagnostic. While rigidity is present, it is less specific than the environmental exposure and anhidrosis for heat stroke.
>
> **Key discriminator:** Exposure to 110°F weather and anhidrosis (dry mucous membranes) strongly favor heat stroke over NMS, which typically presents with diaphoresis and is temporally linked to antipsychotic use rather than environmental heat.
>
> **Closest alternative:** C
>
> **Uncertainty source:** The presence of rigidity and recent haloperidol administration creates a plausible mimic for NMS, but the environmental context and dry skin make heat stroke more likely.

Revision of A_2. Agent A_2 also begins with C and receives a C response from A_4. It changes to B with even higher confidence, 0.95.

> **Brief reason:** The patient was found undressing in 110°F weather, a classic presentation for exertional heat stroke. The combination of hyperthermia (104°F), altered mental status, and dry mucous membranes (anhidrosis) is diagnostic. While rigidity is present, the environmental context and lack of prior antipsychotic exposure make heat stroke more likely than NMS, which typically develops over days.
>
> **Key discriminator:** History of undressing in 110°F weather and dry mucous membranes (anhidrosis) point to heat stroke, whereas NMS requires prior antipsychotic exposure and typically presents with diaphoresis.
>
> **Closest alternative:** C
>
> **Uncertainty source:** None reported.

Revision of A_4. Agent A_4 initially selects C and receives C from A_3. It also changes to B, with confidence 0.85.

> **Brief reason:** The patient's presentation is most consistent with heat stroke due to exposure to 110°F weather, resulting in hyperthermia and altered mental status. The rigidity is likely due to severe dehydration and metabolic stress or the acute effects of the administered antipsychotics in a dehydrated state, rather than the more delayed onset of classic NMS.
>
> **Key discriminator:** The temporal proximity to 110°F environmental exposure and the acute presentation favor heat stroke over the typically delayed onset of NMS.
>
> **Closest alternative:** C
>
> **Uncertainty source:** Distinguishing early NMS from severe heat stroke in a dehydrated, hyperthermic patient who recently received antipsychotics.

Case interpretation. This failure differs qualitatively from a simple scenario in which an incorrect sender transmits an incorrect label to a correct receiver. Here, every agent initially selects C, and each selected sender also provides C as its initial answer. Nevertheless, three receivers revise from C to B.

The post-revision explanations show a consistent shift in emphasis. The revised agents prioritize the extreme environmental temperature, hyperthermia, altered mental status, and dry mucous membranes as evidence for heat stroke. At the same time, the evidence supporting neuroleptic malignant syndrome—particularly recent haloperidol administration and marked rigidity—is discounted. Thus, the failure emerges during second-pass reasoning even though the communicated answer labels themselves are correct.

This example also highlights an important limitation of interpreting test-time transitions causally. The test-time inference trace does not include the matched self-review counterfactual used during calibration. We therefore cannot determine from this trace alone whether the C→B transitions were caused specifically by peer exposure or whether the same agents would also have changed their answers during an isolated second-pass self-review. For this reason, we interpret the example as a *harmful post-communication revision* rather than as direct evidence that a correct peer message caused the error.

The case is additionally relevant to the calibration distribution. MedDAC constructs labeled effect-prediction examples only from questions with mixed initial correctness, whereas this example begins with unanimous initial correctness. The Dynamic policy is therefore operating outside the labeled support used to construct its effect predictors. This provides a concrete example of the extrapolation issue discussed in the main-text limitations.

D.3 Summary of the Two Cases

The two examples expose complementary behaviors of selective communication. In Case 1, one correct minority agent coexists with three incorrect agents, and selective communication propagates the useful diagnostic distinction sufficiently to convert the entire group to the correct answer. In Case 2, all four agents are initially correct, yet additional revision destabilizes the group and produces an incorrect majority despite the selected senders themselves initially supporting the gold answer.

Together, these examples motivate both components of MedDAC: communication can provide meaningful corrective opportunities, but revision also carries a non-negligible risk of harm. The matched self-review construction is intended to estimate these competing effects during calibration, while the failure case indicates that better handling of unanimous and out-of-support states remains an important direction for future work.