When Agreement Becomes Unsafe: Loss-Aware Energy Control for Diagnostic Deliberation

Yuting Yan [1] **Yinghao Fu** [2 1] **Haozhou Gao** [1] **Tianjian Zhang** [1] **Aoxi Liu** [1] **Shuang Li** [1 *]

Abstract

Multi-agent deliberation can improve high-stakes classification, but agreement is not a safety certificate. We study diagnostic-style decisions in which agents with asymmetric expertise may converge before high-cost alternatives have been excluded, a failure mode we call *premature diagnostic closure*. We formulate deliberation as a protocol-control problem without online ground truth: a mediator observes agents' reports, optional justifications, calibration confusion matrices, and an asymmetric loss, and must decide whether to continue, challenge, certify, or escalate. We propose *Diagnostic Consensus Energy Minimization* (D-CEM), a loss-aware controller that uses each agent's confusion matrix as a map of plausible differential diagnoses. Given current reports, D-CEM forms confusion-induced posteriors, identifies the most dangerous plausible miss under the loss, and computes a diagnostic consensus energy combining posterior disagreement, expected harm, and loss-aware margin. The resulting policy continues while risk decreases, issues targeted differential challenges, certifies only low-energy large-margin decisions, and escalates when high-risk deliberation stagnates. We prove loss-sensitive certification and cost-aware stopping guarantees. Across synthetic diagnostic tasks and clinical LLM datasets, D-CEM reduces expected diagnostic cost, high-risk misses, and harmful consensus while controlling communication.

1. Introduction

High-stakes classification often requires a choice among plausible alternatives under uncertainty and asymmetric error costs. Such problems often rely on heterogeneous decision makers, such as domain experts, predictive models, and human-AI teams, whose judgments can be combined or revised through interaction (Bansal et al., 2021). When participants hold complementary evidence, additional communication can correct local mistakes and improve the final decision (Du et al., 2024). Yet agreement is not itself a safety certificate (Balogh et al., 2015; Ovadia et al., 2019). Classical aggregation arguments rely on independent or weakly correlated errors; when decision makers share information sources, biases, or error modes, convergence may reflect common failure rather than resolved uncertainty (Kuncheva & Whitaker, 2003). This distinction is especially acute under asymmetric loss, where different misclassifications can have very different consequences (Elkan, 2001). Similar issues have emerged in AI multi-agent frameworks: multiple LLM agents can prematurely converge to a consensus without sufficient critical evaluation, a phenomenon known as silent consensus (Wang et al., 2025; Liang et al., 2024). We study the two-agent analogue of *premature diagnostic closure*: deliberation terminates around a mutually consistent decision before the most consequential alternatives have been separated (Bowen, 2006).

This paper treats diagnostic deliberation as a sequential protocol-control problem rather than as static aggregation. We focus on the minimal nontrivial case of two agents. At each round, a mediator observes the agents' reports and optional justifications, but it does not observe the ground-truth label of the current case. Ground truth is available only offline, through calibration data (Guo et al., 2017) used to estimate each agent's historical confusion matrix (Kerrigan et al., 2021; Steyvers et al., 2022), and ex post, for evaluation. The mediator must therefore answer a runtime question using only observable traces: should deliberation continue, should one agent be challenged on a dangerous differential diagnosis, is the current decision safe enough to certify, or has autonomous deliberation failed and should the case be escalated (Madras et al., 2018; Mozannar & Sontag, 2020; Geifman & El-Yaniv, 2019)?

A key difficulty is that naive indicators of progress are unreliable. Agreement is not always good news: agents may converge precisely in regions where they are historically unreliable, so consensus may be fragile even when it is reached

[1]The Chinese University of Hong Kong, Shenzhen, China [2]City University of Hong Kong, Hong Kong SAR, China. Correspondence to: Shuang Li <lishuang@cuhk.edu.cn>.

Accepted to the Workshop on Failure Modes of Agentic AI at ICML2026, Seoul, South Korea. Copyright 2026 by the author(s).

quickly (Acemoglu et al., 2011). Conversely, disagreement is not always bad news: it may reflect complementary information that can still be productively integrated (DeGroot, 1974; Genest & Zidek, 1986). Much prior work on collaborative decision-making emphasizes improving the agents (better predictors, better prompting, better calibration (Minderer et al., 2021)), but this does not directly address the protocol-level question of how long to deliberate and when interaction becomes counterproductive. What is missing is *a lightweight, task-agnostic mechanism that monitors deliberation dynamics and provides a principled basis for stopping or steering decisions*.

We propose **Diagnostic Consensus Energy Minimization (D-CEM)**, a loss-aware mediator for two-agent diagnostic deliberation. D-CEM treats each historical confusion matrix as a map of plausible alternatives rather than as a scalar reliability table. For an agent's current report, it constructs a confusion-induced posterior over possible true labels and identifies the agent's *dangerous plausible miss*: the alternative with the largest posterior-expected harm under the task loss. The agents' posteriors are then pooled into a loss-aware belief and decision. D-CEM computes a diagnostic-risk energy that penalizes posterior disagreement, expected harm, and weak loss-aware margins. Low energy indicates not mere agreement, but agreement after high-cost alternatives have sufficient separation. The energy selects one of four mediator actions: CONTINUE while risk still decreases, DIFFERENTIAL_STEER to challenge the locally riskier agent on its dangerous plausible miss, STOP_AND_DECIDE in a low-energy large-margin region, and STOP_AND_ESCALATE when risk remains high and deliberation stagnates. Thus, termination need not imply a decision.

Our theory formalizes this control view. Under calibration and bounded-loss assumptions, diagnostic-risk energy upper-bounds a surrogate for the expected loss of the pooled decision, which gives a loss-sensitive certificate for safe termination. We also show that the dangerous plausible miss is the myopically optimal single-alternative challenge under posterior-expected harm, and that a one-step energy-descent rule gives cost-aware termination: deliberation should continue only while certified risk reduction exceeds the per-round communication cost (Wald, 1992; Shiryaev, 2008; Howard, 1966). Empirically, we evaluate D-CEM on synthetic high-cost classification tasks and clinical LLM datasets (Jin et al., 2021; Fansi Tchango et al., 2022) recast as multi-class diagnostic tasks. The evaluation focuses on risk-centric failure modes rather than accuracy alone: expected cost, high-risk misses, harmful consensus, low-information stagnation, escalation, certified decisions, and communication. A detailed review of the related literature is in Appendix A.

2. Problem Setup and Confusion-Induced Diagnostic Risk

We introduce **Diagnostic Consensus Energy Minimization (D-CEM)**, a loss-aware mediator for regulating two-agent deliberation in diagnostic settings. D-CEM treats deliberation as a **diagnostic-risk control** problem: a mediator observes the agents' round-by-round reports and chooses one of four actions, CONTINUE, DIFFERENTIAL_STEER, STOP_AND_DECIDE, or STOP_AND_ESCALATE, *without* observing the *ground-truth label* of the current case during deliberation. Its goal is to distinguish safe consensus from *premature diagnostic closure*: cases where agents appear to agree while high-risk alternatives remain unresolved.

The construction follows a simple pipeline. We first define the online diagnostic deliberation problem and the mediator's observable information. We then use each agent's historical confusion matrix to convert a current report into a case-level differential posterior. These posteriors are pooled into a loss-aware diagnostic belief, from which we define an energy measuring unresolved disagreement, expected diagnostic harm, and weak diagnostic margin. Finally, the mediator uses this energy to decide whether to continue, differentially steer, certify, or escalate.

2.1. Diagnostic Deliberation under Asymmetric Expertise

Task and diagnostic loss. We consider a diagnostic classification task with K candidate conditions $\mathcal{Y} = \{1, \dots, K\}$ and diagnostic decisions $d \in \mathcal{D}$, with $\mathcal{D} = \mathcal{Y}$ unless otherwise specified. We introduce a loss matrix $L \in \mathbb{R}_+^{K \times K}$ to capture the risks of diagnosis. For any predicted decision $d \in \mathcal{D}$ and true condition $y \in \mathcal{Y}$, the system incurs a non-negative diagnostic loss defined as:

$$L(d, y) \geq 0$$

The loss matrix is generally asymmetric: missing a severe condition can be more costly than over-referring or choosing a conservative alternative. In practice, L may be coarse, e.g., $\{0, 1, \ell_{\text{high}}\}$; D-CEM only requires relative diagnostic harm.

Agents and messages. At each deliberation round $t = 1, 2, \dots$, each agent $a \in \{1, 2\}$ emits

$$m_a^{(t)} := \left(\hat{y}_a^{(t)}, w_a^{(t)}\right),$$

where $\hat{y}_a^{(t)} \in \mathcal{Y}$ is the agent's current diagnostic report and $w_a^{(t)}$ is an optional justification signal, such as feature importances, salient attributes, or rule weights. The justification signal is used only for differential steering. Agents may have asymmetric expertise: one agent may be locally reliable for some conditions while systematically confusing others.

Mediator information. At round t, a mediator M observes the interaction history represented by the filtration $\mathcal{F}_t := \sigma(\{m_1^{(s)}, m_2^{(s)}\}_{s=1}^t)$ and selects a $\mathcal{F}_t$-measurable action

$$u^{(t)} \in \{\text{CONTINUE}, \text{DIFFERENTIAL_STEER}, \text{STOP_AND_DECIDE}, \text{STOP_AND_ESCALATE}\}.$$

The mediator does not observe the true label of the current case. Samples with true labels are used only offline to estimate historical confusion matrices and ex post to evaluate performance. Thus the mediator's online decisions are computed only from fixed calibration statistics, current reports, optional justifications, and the diagnostic loss matrix.

Mediator's "ideal" risk-aware objective The previous paragraphs define what the mediator observes and which actions it may take. We now state the ideal risk-aware objective that motivates these actions, even though *part of this objective cannot be directly evaluated during online deliberation*.

Let τ be the stopping round. If the mediator outputs STOP_AND_DECIDE, the system commits to a pooled decision $d^{(\tau)}$. If it outputs STOP_AND_ESCALATE, the case is deferred to a higher-tier reviewer rather than autonomously closed. A compact costed objective is

$$\mathbb{E}\Big[\mathbf{1}\{u^{(\tau)} = \text{STOP_AND_DECIDE}\}L(d^{(\tau)}, y) + \mathbf{1}\{u^{(\tau)} = \text{STOP_AND_ESCALATE}\}c_{\text{esc}} + c\tau\Big]. \tag{1}$$

where c is the per-round deliberation cost and c_{esc} the escalation cost. Since neither y nor $L(d, y)$ is observable online, the above objective cannot be evaluated at deployment time. **Thus the central methodological question is how to build an online quantity that is computable from reports and calibration data, but still reflects the loss-sensitive risk of premature closure**. D-CEM uses a computable diagnostic-risk energy as a surrogate certificate, described in the next section.

Scope of our problem. We focus on cooperative diagnostic deliberation in which agents report their genuine assessments. Strategic misreporting and incentive-compatible extensions are outside the scope of this work.

2.2. Confusion-Induced Differential Posteriors

The central object of D-CEM is each agent's historical *confusion matrix*, used not as a scalar reliability table but as a conditional error model that *maps an agent's current report to the alternative diagnoses* it may plausibly be confusing. Combined with the asymmetric diagnostic loss L from Section 2.1, this lets the mediator monitor *two failure modes* simultaneously: unresolved disagreement, and high-cost agreement.

For agent a, let

$$C_a[i, j] := \Pr(\hat{y}_a = j \mid y = i)$$

be its historical confusion matrix, estimated from a calibration set or historical data. Diagonal entries capture per-class recall, while off-diagonal entries capture characteristic confusions. Importantly, C_a is held fixed during each online deliberation episode. Across rounds, the agents' reports may change, but the mediator does not observe a new per-round confusion matrix. Any round-wise confusion matrices shown in experiments are computed ex post over the evaluation set for visualization only.

Given the current report $\hat{y}_a^{(t)}$, the mediator forms a posterior over the latent diagnosis by Bayes' rule:

$$P_a^{(t)}(y) := \Pr(y \mid \hat{y}_a^{(t)}) = \frac{\pi(y)C_a[y, \hat{y}_a^{(t)}]}{\sum_{y'} \pi(y')C_a[y', \hat{y}_a^{(t)}]}, \tag{2}$$

where π is a prior over diagnoses, chosen as a uniform prior or empirical base rate. This posterior is the central diagnostic object: it answers which alternative conditions remain plausible, given that agent a currently reports $\hat{y}_a^{(t)}$.

For any distribution P, define the expected diagnostic risk of decision d:

$$R_L(d; P) := \sum_y L(d, y)P(y). \tag{3}$$

For each agent, D-CEM identifies the most dangerous plausible miss. If multiple alternatives attain the maximum, we use a fixed deterministic tie-breaking rule:

$$z_a^{(t)} \in \arg\max_{y \in \mathcal{Y} \setminus \{\hat{y}_a^{(t)}\}} P_a^{(t)}(y)L(\hat{y}_a^{(t)}, y). \tag{4}$$

Thus the mediator does not ask only whether the agent is reliable. It asks: *which high-cost alternative could this agent be missing, conditional on its current report?*

The resulting posteriors and dangerous plausible misses are local, agent-level objects. The next step is to combine the two agents' diagnostic states into a *single pooled belief* and then measure whether that pooled state is *safe enough for closure*. This is the focus of the next section.

3. Diagnostic Consensus Energy and Deliberation Policy

3.1. Diagnostic-Risk Energy

This subsection lifts the local confusion-induced posteriors into a global diagnostic-risk energy. The construction has

three parts: first pool the agents' posteriors, then choose the loss-minimizing diagnostic decision under the pooled belief, and finally penalize states where disagreement, expected harm, or weak margins remain.

Each round's reports induce posteriors $P_1^{(t)}$ and $P_2^{(t)}$. Let

$$\rho_a^{(t)} := P_a^{(t)}(\hat{y}_a^{(t)}), \qquad \bar{\rho}_a^{(t)} := \max\{\rho_a^{(t)}, \rho_{\min}\}, \tag{5}$$

where $\rho_a^{(t)}$ is the posterior probability that agent a's current reported diagnosis is correct. The mediator *pools beliefs* using a reliability-weighted log pool:

$$\omega_a^{(t)} = \frac{(\bar{\rho}_a^{(t)})^{\lambda_{\text{pool}}}}{\sum_{a'}(\bar{\rho}_{a'}^{(t)})^{\lambda_{\text{pool}}}}, \qquad Q^{(t)}(y) \propto \prod_{a=1}^{2} \left(P_a^{(t)}(y)\right)^{\omega_a^{(t)}}. \tag{6}$$

Equivalently, the normalized pooled belief is

$$Q^{(t)}(y) = \frac{\prod_{a=1}^{2}\left(P_a^{(t)}(y)\right)^{\omega_a^{(t)}}}{\sum_{y'}\prod_{a=1}^{2}\left(P_a^{(t)}(y')\right)^{\omega_a^{(t)}}}. \tag{7}$$

Under asymmetric loss, the pooled decision is the Bayes diagnostic decision

$$d_1^{(t)} \in \arg\min_d R_L(d; Q^{(t)}), \tag{8}$$

which need not equal the most probable class under $Q^{(t)}$. Let $d_2^{(t)}$ be the second-lowest-risk decision. We define the diagnostic margin

$$\Delta_L^{(t)} = R_L(d_2^{(t)}; Q^{(t)}) - R_L(d_1^{(t)}; Q^{(t)}). \tag{9}$$

A small margin means that the leading diagnosis is not much safer than a competing differential diagnosis.

These components define the state that the mediator monitors across rounds. D-CEM computes the **diagnostic-risk energy**

$$\begin{aligned}\varepsilon_{\text{diag}}^{(t)} &= \alpha D_{\text{sym}}(P_1^{(t)}, P_2^{(t)}) \\ &\quad + \beta \sum_{a=1}^{2} R_L(\hat{y}_a^{(t)}; P_a^{(t)}) \\ &\quad + \gamma \exp(-\Delta_L^{(t)}).\end{aligned} \tag{10}$$

where

$$D_{\text{sym}}(P_1, P_2) = D_{\text{KL}}(P_1 \| P_2) + D_{\text{KL}}(P_2 \| P_1).$$

The first term penalizes *unresolved disagreement* between the agents' confusion-induced posteriors. The second penalizes agent reports that have *high expected diagnostic harm* under their own posterior. The third penalizes *small diagnostic margins*, including cases where agents agree but the pooled belief has not ruled out a high-risk alternative. Therefore low diagnostic energy certifies more than consensus: it certifies low estimated risk and a resolved differential.

The energy is useful because it is not only descriptive but actionable. The mediator treats *i*) decreases in energy as evidence that deliberation is still productive, *ii*) low-energy large-margin states as certifiable, and *iii*) high-energy stagnation as a reason to escalate.

3.2. Mediator Policy

We now **turn the energy into a sequential control policy**. The policy uses the current energy level, the diagnostic margin, the locally dangerous miss, and the recent energy descent to choose among the four mediator actions.

Let the windowed energy descent be

$$\widehat{g}^{(t)} = \frac{1}{W}\sum_{i=0}^{W-1}\left(\varepsilon_{\text{diag}}^{(t-1-i)} - \varepsilon_{\text{diag}}^{(t-i)}\right), \tag{11}$$

with a small window W. The mediator applies the following policy, ordered from safest terminal action to targeted intervention and then to continued deliberation:

- **STOP_AND_DECIDE: certified decision.** If
$$\varepsilon_{\text{diag}}^{(t)} \le \varepsilon_{\text{safe}} \quad \text{and} \quad \Delta_L^{(t)} \ge m_{\text{safe}},$$
the mediator commits to $d_1^{(t)}$. Both conditions are required: energy must be low, and the leading diagnosis must be sufficiently separated from the closest risk competitor.
- **DIFFERENTIAL_STEER: targeted challenge.** If the decision is not yet certified and an actionable high-risk differential remains, the mediator issues a targeted challenge. The target agent is the locally riskier one
$$a^\dagger(t) \in \arg\max_a R_L(\hat{y}_a^{(t)}; P_a^{(t)}),$$
and the challenge is anchored at its most dangerous plausible miss $z_{a^\dagger}^{(t)}$ from (4). The action is gated by a risk threshold $R_L(\hat{y}_{a^\dagger}^{(t)}; P_{a^\dagger}^{(t)}) \ge s_{\text{asym}}$, ensuring there is enough diagnostic risk to act on. A reliability-gap variant $|\rho_1^{(t)} - \rho_2^{(t)}| \ge s_{\text{asym}}$ recovers the classical "weaker-agent" heuristic and is discussed in Appendix B.3.
- **STOP_AND_ESCALATE: failed deliberation.** If no useful challenge remains and energy has stagnated in a high-risk region,
$$\widehat{g}^{(t)} < \delta \quad \text{and} \quad \varepsilon_{\text{diag}}^{(t)} > \varepsilon_{\text{low}},$$
the mediator escalates rather than deciding. This means autonomous deliberation has failed to produce a trustworthy certificate; the system should not close the diagnosis confidently.

- **CONTINUE.** If none of the above conditions applies, the mediator allows another round of deliberation. The agents may update their reports, changing the conditioning event $\hat{y}_a^{(t)}$ in (2). Thus $P_a^{(t)}$, $z_a^{(t)}$, $Q^{(t)}$, $\Delta_L^{(t)}$, and $\varepsilon_{\text{diag}}^{(t)}$ may change across rounds, even though the historical confusion matrices C_a and the prior π remain fixed.

Stopping time and parameters. The induced stopping time is

$$\tau^* = \inf\Big\{t \geq 1 : \big(\varepsilon_{\text{diag}}^{(t)} \leq \varepsilon_{\text{safe}} \text{ and } \Delta_L^{(t)} \geq m_{\text{safe}}\big) \text{ or } \big(\widehat{g}^{(t)} < \delta \text{ and } \varepsilon_{\text{diag}}^{(t)} > \varepsilon_{\text{low}}\big)\Big\}. \tag{12}$$

Here $\varepsilon_{\text{safe}}$ is the target diagnostic-risk tolerance, m_{safe} is the required diagnostic margin, δ is the per-round deliberation cost, ε_{low} marks a high-risk stagnation region, and s_{asym} controls when differential steering is considered useful. The thresholds satisfy $\varepsilon_{\text{safe}} < \varepsilon_{\text{low}}$ by construction.

3.3. Differential Steering

The policy above specifies when a steering action is triggered. We now specify what that action contains and why it remains a diagnostic challenge rather than an override of the agent's judgment.

Differential steering is not persuasion and does not overwrite an agent's prediction. It is a diagnosis-specific challenge targeted at the most dangerous alternative implied by the agent's own confusion profile. The steering target is

$$\big(a^\dagger(t), z_{a^\dagger}^{(t)}\big).$$

The mediator **sends a message of the form**:

> *You currently predict $\hat{y}_{a^\dagger}^{(t)}$. Based on historical calibration, cases in which you predict $\hat{y}_{a^\dagger}^{(t)}$ are sometimes confused with $z_{a^\dagger}^{(t)}$. Reassess the evidence distinguishing these two diagnoses, especially the features or rules $S^{(t)}$.*

Here $S^{(t)}$ can be instantiated as the top-k features distinguishing $\hat{y}_{a^\dagger}^{(t)}$ from $z_{a^\dagger}^{(t)}$ using available justification vectors, saliency scores, rule activations, or natural-language rationales. The mediator provides only an auditable differential-diagnosis challenge; the agent revises, or refuses to revise, through its own internal reasoning process. A concrete feature-level construction of $S^{(t)}$ is given in Appendix B.3.

4. Theoretical Guarantees

The preceding subsections define the mediator and its energy-controlled policy. We now summarize why the same construction has a principled interpretation: low energy certifies low-risk agreement, differential steering targets the highest-harm unresolved alternative, and energy descent yields a cost-aware stopping rule.

We state the main guarantees for interpreting D-CEM's energy and policy; direct properties follow from the construction, while the certificate and stopping results rely on the assumptions in Appendix B.

Lemma 4.1 (Diagnostic energy threshold certifies low-risk agreement)**.** *Fix t and suppose $\varepsilon_{\text{diag}}^{(t)} \leq \varepsilon$. Then*

$$D_{\text{sym}}(P_1^{(t)}, P_2^{(t)}) \leq \varepsilon/\alpha, \qquad \sum_{a=1}^{2} R_L(\hat{y}_a^{(t)}; P_a^{(t)}) \leq \varepsilon/\beta,$$

and

$$\Delta_L^{(t)} \geq \log(\gamma/\varepsilon) \quad \textit{whenever } \varepsilon < \gamma.$$

Thus a safe decision region that also enforces $\Delta_L^{(t)} \geq m_{\text{safe}}$ certifies both posterior agreement and a resolved diagnostic differential.

Proposition 4.2 (Dangerous differential identification)**.** *For agent a, the alternative $z_a^{(t)}$ in (4) maximizes the mediator's posterior-expected harm among all single alternative diagnoses to challenge: $z_a^{(t)} \in \arg\max_{y \neq \hat{y}_a^{(t)}} \mathbb{E}_{Y \sim P_a^{(t)}}\Big[L(\hat{y}_a^{(t)}, Y) \cdot \mathbf{1}\{Y = y\}\Big]$. Therefore a differential-steering intervention targeted at $z_a^{(t)}$ is the myopically optimal single-alternative challenge under the current confusion-induced posterior.*

Proposition 4.3 (Log pool is KL-consistent and contracts disagreement)**.** *Fix t and write $P_1 = P_1^{(t)}$, $P_2 = P_2^{(t)}$, and $Q = Q^{(t)}$. With weights clipped to $\omega_a^{(t)} \in [\underline{\omega}, 1 - \underline{\omega}]$, Q is the unique minimizer of a weighted KL-pooling objective $\omega\, D_{\text{KL}}(Q\|P_1) + (1-\omega)\, D_{\text{KL}}(Q\|P_2)$ and satisfies*

$$D_{\text{KL}}(P_1\|Q) \leq (1 - \omega_1^{(t)}) D_{\text{KL}}(P_1\|P_2),$$
$$D_{\text{KL}}(P_2\|Q) \leq \omega_1^{(t)} D_{\text{KL}}(P_2\|P_1).$$

Theorem 4.4 (Loss-sensitive diagnostic certificate)**.** *Under the calibration and bounded-loss assumptions in Appendix B, there exist constants $\kappa_0, \kappa_1 \geq 0$ such that*

$$\mathbb{E}\Big[L(d_1^{(t)}, y) \mid \mathcal{F}_t\Big] \leq \kappa_0 + \kappa_1 \varepsilon_{\text{diag}}^{(t)}. \tag{13}$$

Consequently, minimizing diagnostic energy plus deliberation cost is a loss-sensitive surrogate for the unobserved objective (1):

$$\min_\tau \mathbb{E}\Big[\varepsilon_{\text{diag}}^{(\tau)}\Big] + \delta\mathbb{E}[\tau], \qquad \delta = c/\kappa_1. \tag{14}$$

Theorem 4.5 (Cost-aware stopping)**.** *Under Assumption B.6, the one-step lookahead criterion $\widehat{g}^{(t)} < \delta$ used by D-CEM's escalation gate stops within at most W rounds of the minimizer of the surrogate $\varepsilon_{\text{diag}}^{(t)} + \delta t$, and incurs at most δW additional surrogate cost.*

Theorem 4.6 (Escalation under high-risk stagnation)**.** *If $\varepsilon_{\text{diag}}^{(t)}$ remains above ε_{low} while its windowed descent satisfies $\widehat{g}^{(t)} \to 0$, then for any $\delta > 0$ the policy triggers* STOP_AND_ESCALATE *in finite time. Thus D-CEM does not convert high-energy stagnation into an autonomous decision; it treats it as failed deliberation.*

5. Experiments

Our experiments ask whether D-CEM improves the *safety-cost* tradeoff of deliberation rather than merely raising agreement. We test three claims: (i) diagnostic-risk energy tracks safe convergence and exposes premature diagnostic closure; (ii) DIFFERENTIAL_STEER reduces high-cost misses under asymmetric expertise; and (iii) STOP_AND_ESCALATE prevents stalled deliberation from committing to a confident wrong decision.

Settings and Datasets. We evaluate D-CEM in two settings. (i) The **rule-guided setting** uses two probabilistic rule-based agents on a synthetic diagnostic task with controllable expertise asymmetry and class-level cost, which lets us sweep scenarios, compliance, and budgets at scale. (ii) The **LLM-agent setting** uses two LLMs (GPT-4o-mini and Claude-3.5-Haiku) on two clinical datasets, MedQA-USMLE (Jin et al., 2021) and DDXPlus (Fansi Tchango et al., 2022), in both of which a subset of classes is designated high-risk and induces the asymmetric loss L. The synthetic dataset construction and description, along with implementation details, are introduced in Appendix D.

Scenarios. Both settings instantiate four scenarios. Our main stress test is **Scenario 3b Noisy-non-complementary**, where shared bias drives agents toward a high-cost wrong consensus. The other three probe different axes of selectivity: safe certification under high reliability (**Scenario 1 Ideal**), targeted steering of one-sided errors (**Scenario 2 Asymmetric**), and persistent continuation under complementary noise (**Scenario 3a Noisy-complementary**). Specifications are in Appendix C.

Metrics. We track multiple complementary metrics to capture both safety-cost outcomes and mediator behavior. We report pooled accuracy (Acc_{pool}), expected diagnostic cost $\widehat{R}_L$ (empirical loss-weighted error), high-risk miss rate (miss rate on high-cost truths), and harmful-consensus rate (wrong commitment on a high-cost truth without escalation, the operational measure of premature closure), together with mediator action statistics: average rounds, certified-decision rate (STOP_AND_DECIDE share), and escalation rate (STOP_AND_ESCALATE share). Formal definitions are in Appendix E.

5.1. Rule-guided: Deliberation Dynamics

Case-level trajectories. Figure 1 contrasts the two terminal actions of D-CEM. In a Scenario 1 case (*Example 1*), the mediator alternates CONTINUE and DIFFERENTIAL_STEER (target $a^\dagger =$ Agent 2) as $\varepsilon_{\text{diag}}^{(t)}$ descends, then issues STOP_AND_DECIDE at Round 7 once the safe-energy and margin conditions both hold. In a Scenario 3b case (*Example 2*), both agents agree on a high-cost wrong class; after a Round-5 DIFFERENTIAL_STEER the energy stalls above ε_{low}, so the mediator escalates at Round 8 rather than committing to the shared wrong consensus.

Aggregate behaviour. Figure 2 aggregates per-round accuracy and diagnostic-risk energy across all test cases. *Ideal:* energy descends below $\varepsilon_{\text{safe}}$ and STOP_AND_DECIDE fires on most cases. *Asymmetric:* the weaker agent improves after DIFFERENTIAL_STEER while the stronger one is preserved. *Noisy-complementary:* energy descends slowly but persistently, warranting more CONTINUE rounds. *Noisy-non-complementary:* energy plateaus above ε_{low} and the protocol escalates rather than committing to the shared high-risk wrong answer—the unsafe-consensus failure mode D-CEM targets. The inverse accuracy–energy relation is consistent with $\varepsilon_{\text{diag}}^{(t)}$ acting as a risk certificate.

5.2. Rule-guided: D-CEM vs. Aggregation Baselines

We compare D-CEM with three baselines under matched agents, calibration, splits, scenarios, and seeds. **Single best** uses the calibration-best agent's one-shot prediction. **Free discussion** lets the two agents update on each other's predictions for $T = 22$ rounds with no monitoring. **Fixed-weight log pool** keeps D-CEM's pooling rule but removes DIFFERENTIAL_STEER, STOP_AND_DECIDE, and STOP_AND_ESCALATE, isolating the value of mediator actions beyond pooling.

Figure 3 shows that D-CEM improves the safety-cost tradeoff across all scenarios: accuracy gains are largest in the noisy regimes (S3a: 0.93 vs. 0.59–0.68; S3b: 0.90 vs. 0.44–0.62); $\widehat{R}_L$ is lowest in every cell and high-risk misses drop to 0.00 in both noisy settings. It also lowers deliberation cost, measured by average rounds, relative to Free discussion (2–8 vs. 22), and escalates unresolved cases rather than forcing autonomous closure.

5.3. Rule-guided: Partial-Compliance Ablation

In realistic deployments, agents may differ in how much they accept mediator suggestions, and this compliance level is difficult to control precisely in the LLM experiments. We therefore run a controlled rule-guided ablation in which each DIFFERENTIAL_STEER recommendation is adopted with probability $\xi \in \{25\%, 50\%, 75\%, 100\%\}$. We record outcomes at checkpoints $T \in \{4, 8, 12\}$ across all four syn-

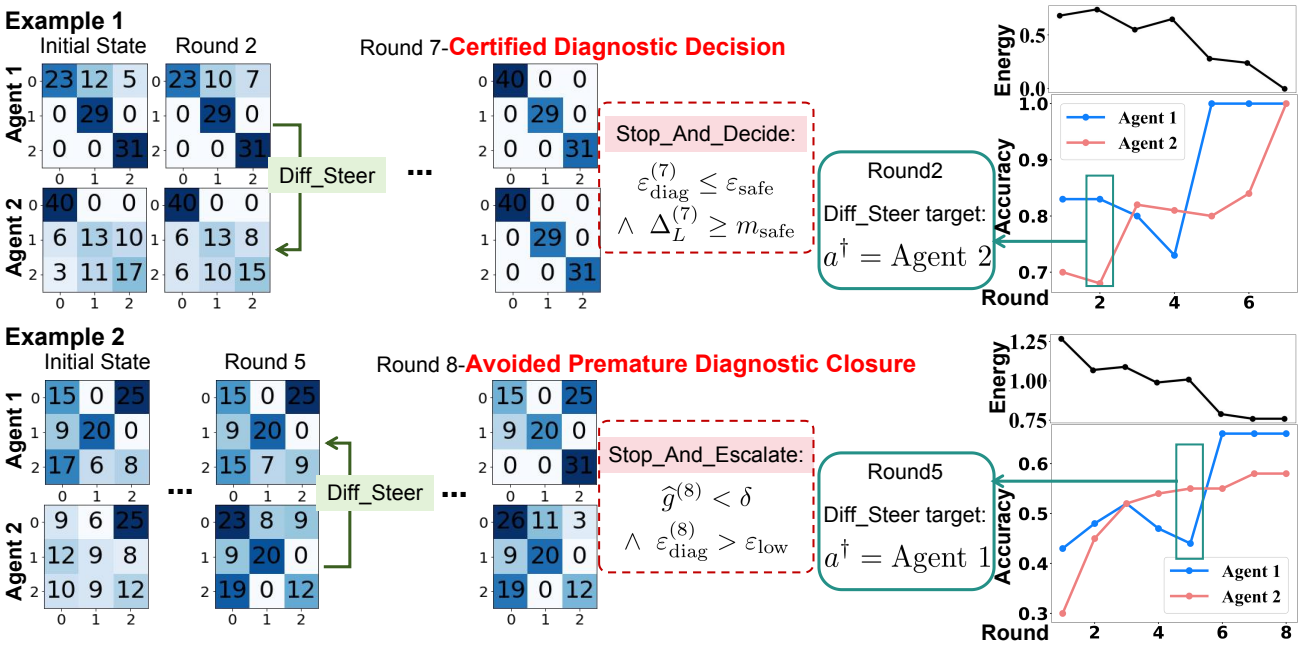

Figure 1. Case-level D-CEM trajectories: Scenario 1 (certified decision) and Scenario 3b (escalation). Round-wise confusion matrices are computed *ex post* for visualization only; the mediator uses fixed C_a and never observes the current case label.

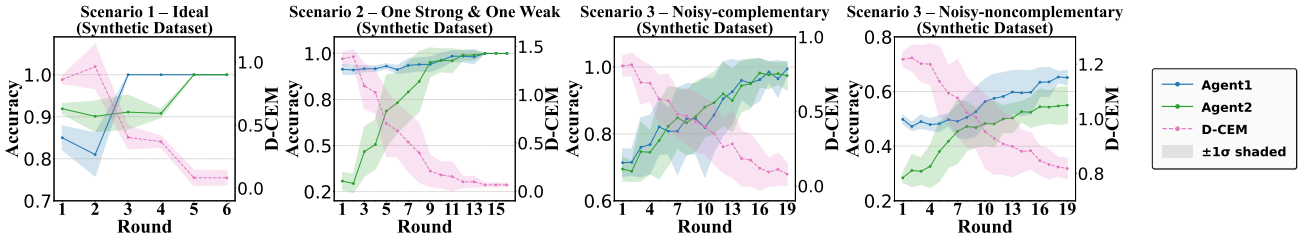

Figure 2. Per-round accuracy and diagnostic-risk energy $\varepsilon_{\text{diag}}^{(t)}$ across the four scenarios, aggregated over the test set.

thetic scenarios. In the correctable scenarios, higher compliance ξ lowers $\widehat{R}_L$ and improves the steered weak agent, confirming that DIFFERENTIAL_STEER's value scales with agent uptake; full results are in Appendix F.

5.4. LLM-Agent: Harmful Pooling on Clinical Datasets

We instantiate D-CEM with two LLM agents (`gpt-4o-mini`, `claude-3.5-haiku`) on the clinical datasets MedQA-USMLE and DDXPlus, recast as multi-class diagnostic tasks under scenarios S1–S3b. Baselines are *Single best* (the CAL-best of agents A, B; no debate), *Free discussion* ($T = 6$), and *Fixed-pool*. Figure 4 shows system risk $\widehat{R}_L^{\text{sys}}$, harmful-consensus rate H.C., and the D-CEM action mix. Detailed results are provided in Appendix H.

We highlight three findings.

- **D-CEM is the only multi-agent protocol that does not collapse under asymmetric loss.** Once L penalizes high-severity misses with $\ell_{\text{high}} = 5$, Pool and Free both inflate sharply on DDXPlus, while D-CEM remains bounded across all scenarios. Its gain over Single is smallest on the saturated $S1$, where Single already reaches $\text{Acc} = 0.95$ and no protocol has remaining fusion headroom.
- **D-CEM achieves the lowest H.C. in every cell.** On DDXPlus, it drives harmful consensus to near zero across all scenarios, an order-of-magnitude reduction over Pool, which still misses about a third of high-cost truths in noisy settings. On MedQA-Tri, where risk is defined by specialty rather than severity and cost asymmetry is weaker, the reduction is smaller but D-CEM remains consistently best.
- **The action mix tracks reliability, not a fixed schedule.** On DDXPlus, the controller rarely intervenes in reliable regimes (only 9% STEER on $S1$) but acts strongly in noisy ones (Steer 45%, escalation 46%), saving cost on easy cases and focusing on hard ones. On MedQA-Tri, the flatter reliability yields a tighter 11–16% escalation band. Across settings, average rounds stay below the $T_{\max} = 6$ budget (Table 7), supporting cost-aware termination.

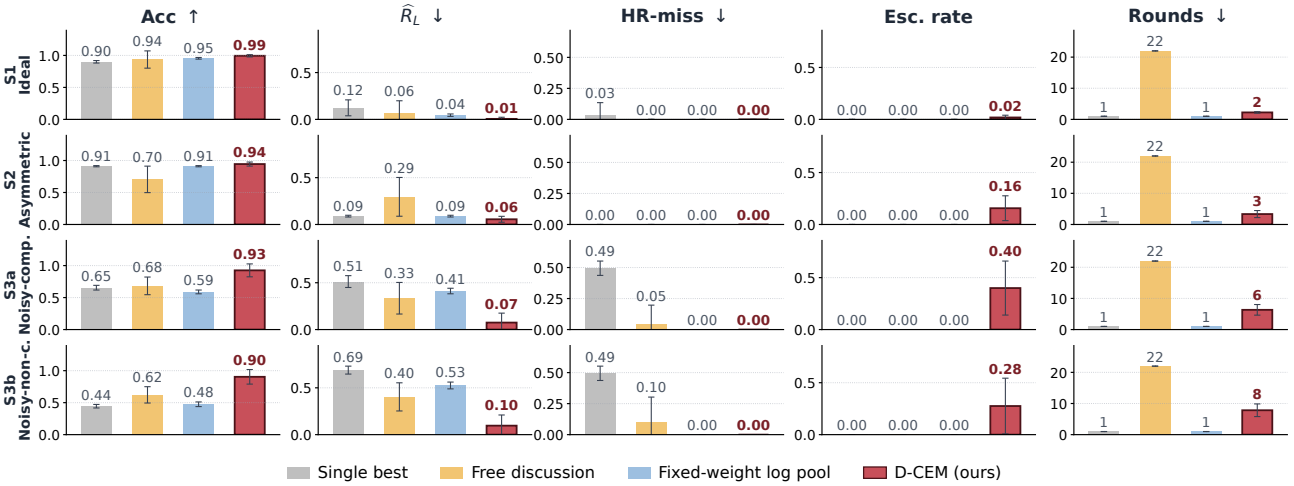

Figure 3. **Rule-guided comparison against aggregation baselines.** Columns: four scenarios. Rows: pooled accuracy (↑), expected diagnostic cost $\widehat{R}_L$ (↓), high-risk miss rate (↓), escalation rate, and average rounds (↓). Definitions in Appendix E.

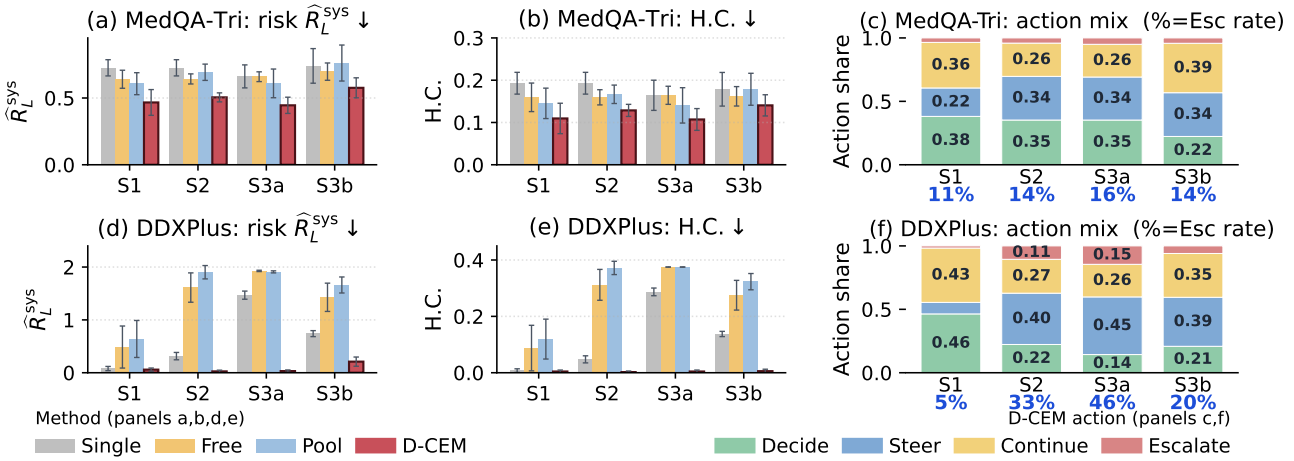

Figure 4. **Main results on MedQA-Tri (top) and DDXPlus (bottom)**. **(a,d)** System risk $\widehat{R}_L^{sys}$ ↓. **(b,e)** Harmful-consensus rate H.C.↓. **(c,f)** D-CEM action mix; bold blue numbers below each bar are per-case escalation rates.

5.5. LLM-Agent: Escalation Ablation

We isolate the contribution of STOP_AND_ESCALATE by disabling it, forcing the mediator to commit autonomously on every case. Table 1 shows that removing escalation increases system risk and harmful consensus across both clinical datasets, with the largest degradation in unreliable DDXPlus regimes: 5–16× worse on $\widehat{R}_L$ and 4–10× worse on H.C. in S2/S3a/S3b. This supports escalation as a structural safety action rather than a reporting choice.

Table 1. Ablation of STOP_AND_ESCALATE. Lower is better for both metrics. "w/o E." = w/o ESCALATE.

Dataset	Scen.	$\widehat{R}_L^{sys}$ ↓ D-CEM	$\widehat{R}_L^{sys}$ ↓ w/o E.	H.C.↓ D-CEM	H.C.↓ w/o E.
MedQA-Tri	S1	**0.47**±.10	0.58±.06	**0.11**±.04	0.14±.03
	S2	**0.51**±.03	0.68±.06	**0.13**±.01	0.17±.03
	S3a	**0.45**±.06	0.60±.07	**0.11**±.03	0.14±.03
	S3b	**0.58**±.08	0.73±.07	**0.14**±.03	0.17±.02
DDXPlus	S1	**0.06**±.03	0.11±.04	**0.00**±.01	0.01±.01
	S2	**0.03**±.02	0.15±.07	**0.00**±.00	0.02±.01
	S3a	**0.03**±.02	0.50±.09	**0.00**±.01	0.03±.02
	S3b	**0.21**±.09	0.39±.11	**0.01**±.01	0.02±.02

6. Conclusion

We introduce Diagnostic Consensus Energy Minimization (D-CEM), a loss-aware mediator for diagnostic deliberation among heterogeneous agents. D-CEM converts historical confusion matrices into case-level differential posteriors, identifies plausible high-risk misses, and uses a diagnostic-risk energy to decide whether to continue, steer, certify, or escalate. Under calibration assumptions, this energy yields a loss-sensitive certificate and supports cost-aware stopping. Across synthetic and clinical LLM datasets, D-CEM reduces harmful consensus and high-risk misses while requiring fewer deliberation rounds. The central lesson is that agreement should not be treated as closure: when high-risk alternatives remain unresolved, safe protocols should challenge or escalate rather than force a decision.

References

Acemoglu, D., Dahleh, M. A., Lobel, I., and Ozdaglar, A. Bayesian learning in social networks. *The Review of Economic Studies*, 78(4):1201–1236, 2011.

Balogh, E. P., Miller, B. T., and Ball, J. R. Improving diagnosis in health care. 2015.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., and Weld, D. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pp. 1–16, 2021.

Bowen, J. L. Educational strategies to promote clinical diagnostic reasoning. *New England Journal of Medicine*, 355(21):2217–2225, 2006.

Cao, Y., Mozannar, H., Feng, L., Wei, H., and An, B. In defense of softmax parametrization for calibrated and consistent learning to defer. *Advances in Neural Information Processing Systems*, 36:38485–38503, 2023.

Corvelo Benz, N. and Rodriguez, M. Human-aligned calibration for ai-assisted decision making. *Advances in Neural Information Processing Systems*, 36:14609–14636, 2023.

Cui, Y., Fu, H., Zhang, H., Wang, L., and Zuo, C. Freemad: Consensus-free multi-agent debate. *arXiv preprint arXiv:2509.11035*, 2025.

Dawid, A. P. and Skene, A. M. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979.

DeGroot, M. H. Reaching a consensus. *Journal of the American Statistical association*, 69(345):118–121, 1974.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first international conference on machine learning*, 2024.

Elkan, C. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pp. 973–978. Lawrence Erlbaum Associates Ltd, 2001.

Fansi Tchango, A., Goel, R., Wen, Z., Martel, J., and Ghosn, J. Ddxplus: A new dataset for automatic medical diagnosis. *Advances in neural information processing systems*, 35:31306–31318, 2022.

Geifman, Y. and El-Yaniv, R. Selectivenet: A deep neural network with an integrated reject option. In *International conference on machine learning*, pp. 2151–2159. PMLR, 2019.

Genest, C. and Zidek, J. V. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1):114–135, 1986.

Graber, M. L., Franklin, N., and Gordon, R. Diagnostic error in internal medicine. *Archives of internal medicine*, 165(13):1493–1499, 2005.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Howard, R. A. Information value theory. *IEEE Transactions on systems science and cybernetics*, 2(1):22–26, 1966.

Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., and Szolovits, P. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Kerrigan, G., Smyth, P., and Steyvers, M. Combining human predictions with model probabilities via confusion matrices and calibration. *Advances in Neural Information Processing Systems*, 34:4421–4434, 2021.

Kuncheva, L. I. and Whitaker, C. J. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning*, 51(2):181–207, 2003.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Shi, S., and Tu, Z. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 17889–17904, 2024.

Ma, S., Chen, Q., Wang, X., Zheng, C., Peng, Z., Yin, M., and Ma, X. Towards human-ai deliberation: Design and evaluation of llm-empowered deliberative ai for ai-assisted decision-making. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–23, 2025.

Madras, D., Pitassi, T., and Zemel, R. Predict responsibly: improving fairness and accuracy by learning to defer. *Advances in neural information processing systems*, 31, 2018.

Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., and Lucic, M. Revisiting the calibration of modern neural networks. *Advances in neural information processing systems*, 34:15682–15694, 2021.

Mozannar, H. and Sontag, D. Consistent estimators for learning to defer to an expert. In *International conference on machine learning*, pp. 7076–7087. PMLR, 2020.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., and Snoek, J. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Shiryaev, A. N. *Optimal stopping rules*. Springer, 2008.

Steyvers, M., Tejeda, H., Kerrigan, G., and Smyth, P. Bayesian modeling of human–ai complementarity. *Proceedings of the National Academy of Sciences*, 119(11): e2111547119, 2022.

Vodrahalli, K., Gerstenberg, T., and Zou, J. Y. Uncalibrated models can improve human-ai collaboration. *Advances in Neural Information Processing Systems*, 35:4004–4016, 2022.

Wald, A. Sequential tests of statistical hypotheses. In *Breakthroughs in statistics: Foundations and basic theory*, pp. 256–298. Springer, 1992.

Wang, Y., Yan, Q., Xing, Z., Liu, L., He, J., Fu, C.-W., Hu, X., and Heng, P.-A. Silence is not consensus: Disrupting agreement bias in multi-agent llms via catfish agent for clinical decision making. *arXiv preprint arXiv:2505.21503*, 2025.

Wei, Z., Cao, Y., and Feng, L. Exploiting human-ai dependence for learning to defer. In *Forty-first International Conference on Machine Learning*, 2024.

A. Related Work

Recent work studies whether multiple model instances or agents can improve reasoning through debate, critique, or iterative discussion (Du et al., 2024; Ma et al., 2025). These protocols usually aim to improve final answer quality by eliciting diverse reasoning paths and resolving disagreement. A growing concern, however, is that multi-agent systems can converge prematurely: agents may echo one another, suppress dissent, or produce agreement without sufficient evidence (Wang et al., 2025; Cui et al., 2025). D-CEM is motivated by the same risk of unsafe agreement, but studies a different object. Rather than designing another debate protocol, we design a mediator that controls an ongoing diagnostic deliberation state. At each round and without online ground truth, the mediator must decide whether to continue, issue a targeted differential challenge, certify the current decision, or escalate.

Ensembles and probabilistic aggregation methods improve reliability by combining complementary predictors, especially when uncertainty estimates are calibrated and errors are not fully correlated (Lakshminarayanan et al., 2017; Guo et al., 2017; Ovadia et al., 2019; Minderer et al., 2021). Work on human–AI and crowd prediction further shows that confusion matrices can support calibrated aggregation by modeling class-conditional error patterns (Dawid & Skene, 1979; Kerrigan et al., 2021). D-CEM uses this information differently. It does not reduce an agent to a scalar confidence, ensemble weight, or diagonal reliability score. Instead, it treats the off-diagonal entries of a historical confusion matrix as a map of plausible differential diagnoses: after an agent reports a label, the mediator infers which true labels remain plausible and which of them would be costly to miss. This turns calibration statistics into a runtime object for challenge, certification, and escalation rather than only static pooling.

Selective prediction and reject-option classifiers abstain on cases where an autonomous prediction is unreliable (Geifman & El-Yaniv, 2019). Learning-to-defer methods go further by training models to decide when to pass a case to a human or external expert (Madras et al., 2018; Mozannar & Sontag, 2020; Cao et al., 2023; Wei et al., 2024). These approaches are highly relevant because D-CEM also distinguishes autonomous decisions from escalation. The key difference is that deferral is typically a one-shot output decision for a fixed predictor, whereas D-CEM regulates a multi-round interaction. Before termination, the mediator may allow further deliberation or ask a specific agent to reconsider a specific high-risk alternative. Thus escalation is coupled to a deliberation policy and a diagnostic-risk certificate, not merely to predictive confidence.

Recent human–AI decision support work emphasizes that model accuracy and calibration do not automatically translate into better human decisions. Assistance can help, but can also induce over-reliance, miscalibrated trust, or poor use of uncertainty (Vodrahalli et al., 2022; Corvelo Benz & Rodriguez, 2023). D-CEM takes a protocol-level view of this problem in diagnostic-style classification with asymmetric losses. Cost-sensitive learning has long argued that not all errors should be treated equally (Elkan, 2001), while clinical reasoning literature identifies premature diagnostic closure as accepting a hypothesis before plausible alternatives have been considered (Graber et al., 2005; Bowen, 2006; Balogh et al., 2015). We do not model clinical cognition or claim clinical deployment. Rather, we formalize the multi-agent analogue of premature closure: agreement is certified only when the loss-aware posterior risk and diagnostic margin are sufficiently safe; otherwise the mediator challenges or escalates.

B. Additional Method Details and Proofs

This appendix collects the standing assumptions, implementation details, and proofs for the method and guarantees in Section 2–4.

B.1. Mediator Information and Round-Wise Confusion Visualizations

During an online deliberation episode, the mediator does not observe the ground-truth label of the current case. For each agent a, the historical confusion matrix C_a is estimated once from a calibration set and is held fixed throughout the episode. At round t, the task-relevant online inputs to the mediator are the agents' reports $(\hat{y}_a^{(t)}, w_a^{(t)})$, the fixed matrices C_a, the prior π, and the diagnostic loss matrix L.

The quantities $P_a^{(t)}$, $z_a^{(t)}$, $Q^{(t)}$, $\Delta_L^{(t)}$, and $\varepsilon_{\text{diag}}^{(t)}$ may change across rounds because the agents' reports $\hat{y}_a^{(t)}$ may change. They do not require observing the true label. In figures, any round-wise confusion matrices are computed ex post over the evaluation set to visualize how agents' behavior evolves across rounds. These visualizations are not available to the mediator and are not used by the D-CEM policy.

B.2. Technical Assumptions

Assumption B.1 (Finite diagnostic label set)**.** $y \in \{1, \dots, K\}$ with $K < \infty$.

Assumption B.2 (Positive support)**.** The prior has full support: $\pi(y) \geq \underline{\pi} > 0$ for all y. Each historical confusion matrix has full support after smoothing: $C_a[i,j] \geq \underline{C} > 0$ for all a, i, j. Then for any observed report $\hat{y}_a^{(t)}$, the posterior (2) satisfies $P_a^{(t)}(y) \geq \underline{p} > 0$ for all y, ensuring finite KL terms.

Assumption B.3 (Bounded diagnostic loss)**.** The diagnostic loss is bounded: $0 \leq L(d, y) \leq L_{\max} < \infty$ for all d, y.

Assumption B.4 (Clipping for numerical stability)**.** $\bar{\rho}_a^{(t)} = \max\{\rho_a^{(t)}, \rho_{\min}\}$ with $\rho_{\min} \in (0,1)$, and the log-pool weights are clipped to $\omega_a^{(t)} \in [\underline{\omega}, 1 - \underline{\omega}]$ for some $\underline{\omega} \in (0, 1/2]$ when needed.

Assumption B.5 (One-step calibration)**.** At round t, there exists $a^\star \in \{1, 2\}$ such that $\Pr(Y = \cdot \mid \mathcal{F}_t) = P_{a^\star}^{(t)}(\cdot)$.

Remark. Assumption B.5 is used to interpret $\varepsilon_{\text{diag}}^{(t)}$ as a calibrated upper-bound certificate. Without this calibration condition, the same policy remains well-defined from the fixed matrices C_a, but the energy should be read as a surrogate risk score rather than a formal posterior bound.

Assumption B.6 (Idealized diminishing returns for stopping analysis)**.** For the purpose of analyzing the cost-aware stopping rule, consider a deliberation episode in which the realized diagnostic-energy sequence is nonincreasing and has diminishing returns: $\varepsilon_{\text{diag}}^{(t+1)} \leq \varepsilon_{\text{diag}}^{(t)}$ and $g^{(t)} := \varepsilon_{\text{diag}}^{(t-1)} - \varepsilon_{\text{diag}}^{(t)}$ is nonincreasing in t.

Remark. This assumption is used only to justify the surrogate stopping comparison in Theorem 4.5; the D-CEM policy itself applies the observed windowed descent rule without requiring monotone energy.

B.3. Feature-Level Realization of Differential Steering

The main text defines DIFFERENTIAL_STEER as a targeted challenge from the current report $\hat{y}_{a^\dagger}^{(t)}$ to the dangerous plausible miss $z_{a^\dagger}^{(t)}$. When agents expose feature-weight explanations, we instantiate the distinguishing feature set $S^{(t)}$ using the reported vectors $w_a^{(t)}$.

Let $\sigma(w)$ denote a softmax normalization of a feature-weight vector. We define an explanation discrepancy

$$D_{\exp}(w_1, w_2) = D_{\mathrm{KL}}(\sigma(w_1) \| \sigma(w_2)) + D_{\mathrm{KL}}(\sigma(w_2) \| \sigma(w_1)). \tag{15}$$

For the target agent $a^\dagger$ and a reference explanation $w_{\text{ref}}^{(t)}$, the mediator computes a conservative recommended explanation

$$w_{\text{rec}}^{(t)} \in \arg\min_{w \in \mathcal{W}} \left\{ \frac{\lambda_{\text{st}}}{2} \| w - w_{a^\dagger}^{(t)} \|_2^2 + \gamma_{\text{st}} D_{\exp}(w, w_{\text{ref}}^{(t)}) \right\}. \tag{16}$$

The reference explanation may be the stronger agent's evidence for separating $\hat{y}_{a^\dagger}^{(t)}$ from $z_{a^\dagger}^{(t)}$, or a class-contrastive feature template estimated from calibration data. In practice, we use one projected gradient step:

$$w_{\text{rec}}^{(t)} = \Pi_{\mathcal{W}} \left(w_{a^\dagger}^{(t)} - \eta \nabla_w \left[\frac{\lambda_{\text{st}}}{2} \| w - w_{a^\dagger}^{(t)} \|_2^2 + \gamma_{\text{st}} D_{\exp}(w, w_{\text{ref}}^{(t)}) \right]_{w = w_{a^\dagger}^{(t)}} \right). \tag{17}$$

The set $S^{(t)}$ is the set of top-k coordinates with largest absolute change:

$$S^{(t)} = \mathrm{TopK}\left(\left| w_{\text{rec}}^{(t)} - w_{a^\dagger}^{(t)} \right|, k \right). \tag{18}$$

This feature-level construction is only one implementation of the template in the main text. In black-box LLM settings, $S^{(t)}$ can instead be obtained from natural-language rationales, saliency summaries, or rule activations.

B.4. Partial Compliance Model

D-CEM does not assume that agents must follow mediator recommendations. The policy only observes the next-round report after issuing a challenge. For simulation and ablation studies, we model heterogeneous compliance through a

partial-adoption parameter $\xi^{(t)} \in [0, 1]$:

$$w_{a^\dagger}^{(t+1)} = \Pi_{\mathcal{W}} \left((1 - \xi^{(t)}) w_{a^\dagger}^{(t)} + \xi^{(t)} w_{\text{rec}}^{(t)} \right). \tag{19}$$

Here $\xi^{(t)} = 0$ means the agent ignores the recommendation, while $\xi^{(t)} = 1$ means it fully adopts the recommended feature adjustment. This model is used only to generate controlled experiments; the D-CEM mediator itself does not require knowing $\xi^{(t)}$.

B.5. Proofs

B.5.1. Proof of Lemma 4.1

Each term in (10) is nonnegative. If $\varepsilon_{\text{diag}}^{(t)} \leq \varepsilon$, then $\alpha D_{\text{sym}}(P_1^{(t)}, P_2^{(t)}) \leq \varepsilon$, which gives the first bound, and $\beta \sum_a R_L(\hat{y}_a^{(t)}; P_a^{(t)}) \leq \varepsilon$, which gives the second. Also $\gamma \exp(-\Delta_L^{(t)}) \leq \varepsilon$ implies $\Delta_L^{(t)} \geq \log(\gamma/\varepsilon)$ when $\varepsilon < \gamma$. □

B.5.2. Proof of Proposition 4.2

For a fixed agent a and current report $\hat{y}_a^{(t)}, z_a^{(t)} = \arg\max_{y \neq \hat{y}_a^{(t)}} \mathbb{E}_{Y \sim P_a^{(t)}} \left[L(\hat{y}_a^{(t)}, Y) \cdot \mathbf{1}\{Y = y\} \right] = \arg\max_{y \neq \hat{y}_a^{(t)}} \mathbb{E}_{Y \sim P_a^{(t)}} \left[L(\hat{y}_a^{(t)}, y) \cdot \mathbf{1}\{Y = y\} \right] = \arg\max_{y \neq \hat{y}_a^{(t)}} L(\hat{y}_a^{(t)}, y) \cdot P_a^{(t)}(y)$. □

B.5.3. Proof of Proposition 4.3

Fix $\omega \in (0, 1)$ and distributions P_1, P_2 on $\{1, \ldots, K\}$ with positive support. Consider $\min_Q \omega D_{\text{KL}}(Q \| P_1) + (1 - \omega) D_{\text{KL}}(Q \| P_2)$ over the simplex. The Lagrangian first-order conditions yield

$$\log Q(y) = \omega \log P_1(y) + (1 - \omega) \log P_2(y) - \log Z,$$

so $Q(y) \propto P_1(y)^\omega P_2(y)^{1-\omega}$, which is unique by strict convexity. For the contraction bounds, write $Q(y) = Z^{-1} P_1(y)^\omega P_2(y)^{1-\omega}$. Then

$$D_{\text{KL}}(P_1 \| Q) = (1 - \omega) D_{\text{KL}}(P_1 \| P_2) + \log Z.$$

By Holder's inequality, $Z = \sum_y P_1(y)^\omega P_2(y)^{1-\omega} \leq 1$, hence $\log Z \leq 0$ and $D_{\text{KL}}(P_1 \| Q) \leq (1 - \omega) D_{\text{KL}}(P_1 \| P_2)$. The second inequality is symmetric. □

B.5.4. Proof of Theorem 4.4

Let $P^\star = P_{a^\star}^{(t)}$ denote the calibrated posterior from Assumption B.5, so that $\Pr(Y = \cdot \mid \mathcal{F}_t) = P^\star(\cdot)$. By Proposition 4.3,

$$D_{\text{KL}}(P^\star \| Q^{(t)}) \leq D_{\text{sym}}(P_1^{(t)}, P_2^{(t)}) \leq \varepsilon_{\text{diag}}^{(t)} / \alpha.$$

Since $0 \leq L(d, y) \leq L_{\max}$, Pinsker's inequality implies that for any decision d,

$$|R_L(d; P^\star) - R_L(d; Q^{(t)})| \leq L_{\max} \sqrt{2 D_{\text{KL}}(P^\star \| Q^{(t)})} \leq L_{\max} \sqrt{2 \varepsilon_{\text{diag}}^{(t)} / \alpha}.$$

Because $d_1^{(t)}$ minimizes $R_L(d; Q^{(t)})$,

$$\begin{aligned} R_L(d_1^{(t)}; P^\star) &\leq R_L(d_1^{(t)}; Q^{(t)}) + L_{\max} \sqrt{2 \varepsilon_{\text{diag}}^{(t)} / \alpha} \\ &\leq R_L(\hat{y}_{a^\star}^{(t)}; Q^{(t)}) + L_{\max} \sqrt{2 \varepsilon_{\text{diag}}^{(t)} / \alpha} \\ &\leq R_L(\hat{y}_{a^\star}^{(t)}; P^\star) + 2 L_{\max} \sqrt{2 \varepsilon_{\text{diag}}^{(t)} / \alpha}. \end{aligned}$$

The second term of (10) gives $R_L(\hat{y}_{a^\star}^{(t)}; P^\star) \leq \varepsilon_{\text{diag}}^{(t)} / \beta$. Hence

$$\mathbb{E}[L(d_1^{(t)}, Y) \mid \mathcal{F}_t] \leq \frac{1}{\beta} \varepsilon_{\text{diag}}^{(t)} + 2 L_{\max} \sqrt{2 \varepsilon_{\text{diag}}^{(t)} / \alpha}.$$

Finally, for any fixed $\eta > 0$, $\sqrt{x} \leq \eta + x/(4\eta)$ converts this into the stated linear form $\kappa_0 + \kappa_1 \varepsilon_{\text{diag}}^{(t)}$ for suitable constants $\kappa_0, \kappa_1 \geq 0$. The nonnegative margin term in (10) further restricts certification to resolved diagnostic differentials but is not needed for the upper bound above. □

B.5.5. PROOF OF THEOREM 4.5

Under Assumption B.6, define $J(t) := \varepsilon_{\text{diag}}^{(t)} + \delta t$. Then

$$J(t+1) - J(t) = \delta - g^{(t+1)}.$$

Since $g^{(t)}$ is nonincreasing, $\delta - g^{(t)}$ is nondecreasing, so J is discrete convex and any minimizer occurs where $g^{(t)}$ crosses δ. Because $\widehat{g}^{(t)}$ is a moving average of a nonincreasing sequence, it crosses δ at most W rounds after $g^{(t)}$. Hence the implemented rule stops within W rounds of the surrogate minimizer and incurs at most δW additional surrogate cost. □

B.5.6. PROOF OF THEOREM 4.6

If $\varepsilon_{\text{diag}}^{(t)} > \varepsilon_{\text{low}}$ and $\widehat{g}^{(t)} \to 0$, then for any $\delta > 0$ there exists T such that for all $t \geq T$, $\widehat{g}^{(t)} < \delta$. The escalation condition in the mediator policy is then satisfied, so the policy triggers STOP_AND_ESCALATE in finite time. □

B.6. Extension to More than Two Agents

For $N > 2$ agents, each agent $a \in \{1, \ldots, N\}$ has a fixed historical confusion matrix C_a and induces a posterior $P_a^{(t)}$ from its current report. A natural extension of the diagnostic-risk energy is

$$\varepsilon_{\text{diag},N}^{(t)} = \alpha \sum_{a<b} D_{\text{sym}}(P_a^{(t)}, P_b^{(t)}) + \beta \sum_{a=1}^{N} R_L(\hat{y}_a^{(t)}; P_a^{(t)}) + \gamma \exp(-\Delta_L^{(t)}), \tag{20}$$

where $Q^{(t)}$ is the reliability-weighted log pool over all N posteriors and $\Delta_L^{(t)}$ is the margin between the lowest- and second-lowest-risk pooled decisions. Steering can target the agent with largest current diagnostic risk,

$$a^{\dagger}(t) \in \arg\max_a R_L(\hat{y}_a^{(t)}; P_a^{(t)}),$$

and its dangerous plausible miss $z_{a^{\dagger}}^{(t)}$. The same STOP_AND_DECIDE and STOP_AND_ESCALATE semantics apply. A full empirical study of N-agent panels is left for future work.

B.7. Multinomial-Logit Form of Rule-Guided Agents

The rule-guided agents in the synthetic experiments are fully specified multinomial-logit reference agents, not black-box LLM predictors. Each instance x is mapped to a binary rule-activation vector $\phi(x) \in \{0,1\}^d$, where $\phi_r(x) = 1$ iff rule r is triggered. Given weights $w_a^{(t)} \in \mathbb{R}^d$, agent a forms class scores as weighted sums of triggered rules and converts them to probabilities via a softmax:

$$p_a^{(t)}(y = k \mid x) \;\propto\; \exp\!\big(s_{a,k}^{(t)}(x)\big), \qquad s_{a,k}^{(t)}(x) = \sum_{r \in \mathcal{R}_k} w_{a,r}^{(t)}\, \phi_r(x),$$

where $\mathcal{R}_k$ is the set of rules whose consequent is class k. The agent reports the MAP label $\hat{y}_a^{(t)} = \arg\max_k p_a^{(t)}(y = k \mid x)$ together with the weight vector $w_a^{(t)}$ as its justification signal. In the clinical LLM setting, the same mediator logic is applied to black-box probability reports and natural-language cues rather than to rule-weight vectors.

C. Scenario Specifications

We restate the four scenarios used in both the rule-guided and LLM evaluations. The four configurations vary the agents' reliability profile and the alignment of their errors; each scenario most strongly probes one mediator response while still admitting the others.

Scenario 1 – Ideal. Both agents maintain consistently high reliability ($\bar{\rho}_1, \bar{\rho}_2 > s$). The protocol should drive diagnostic-risk energy below $\varepsilon_{\text{safe}}$ and open a margin $\Delta_L \geq m_{\text{safe}}$, after which it certifies the pooled decision via STOP_AND_DECIDE. This scenario tests whether the protocol sustains deliberation toward a safe consensus without unnecessary intervention.

Scenario 2 – Asymmetric. Reliability is asymmetric: exactly one agent satisfies $\bar{\rho} > s$. The weaker agent's confusion profile typically exposes a high-cost alternative z that its current report has missed, so the protocol should issue a differential challenge on $(\hat{y}, z)$ via DIFFERENTIAL_STEER before committing. This scenario tests whether the protocol can align the weaker agent with the stronger one without misleading the stronger.

Scenario 3a – Noisy-complementary. Both agents are individually unreliable ($\bar{\rho}_1, \bar{\rho}_2 \to 0$), but their signals are partially offsetting. The protocol should keep deliberation alive via CONTINUE while energy is still descending, since pooled risk can still drop. This scenario tests whether the consensus energy can drive safe convergence when neither agent is independently bounded.

Scenario 3b – Noisy-non-complementary. Both agents are unreliable and their errors are aligned, so agreement is unsafe. The protocol should detect that energy stagnates above ε_{low} and trigger STOP_AND_ESCALATE rather than autonomously commit. This is the canonical premature-closure stress test and the main threat model of the paper.

D. Datasets and Clinical Evaluation

This appendix specifies the datasets used in the two evaluation settings. The rule-guided setting uses a single controllable synthetic diagnostic task (Appendix D.1); the LLM setting uses two clinical datasets, MedQA-USMLE and DDXPlus, recast as multi-class diagnostic tasks (Appendix D.2).

D.1. Synthetic Diagnostic Dataset

To simulate agents with varying capabilities and enable controlled experiments across multiple scenarios, we use a rule-based synthetic data generator. Below we detail the generator and the agent simulator built on top of it.

Synthetic data generator. The sample generation process is based on a predefined set of ground-truth rules, shown in Table 2. The generator selects a label according to the rule weights, simulating population-level decision-making, formalized by

$$k \sim \mathrm{Mult}\big(\mathrm{softmax}\big(\mathrm{sigmoid}^{-1}\big(w_1^\top \phi_1(\boldsymbol{x}), \ldots, w_k^\top \phi_k(\boldsymbol{x})\big)\big)\big) ,$$

where $w_k^\top \phi_k(\boldsymbol{x})$ denotes the feature function associated with the rule set for label k. A sample is valid only if at least one rule of the selected label is satisfied while no rule of any other label holds. Label k_0 corresponds to a rare class governed by longer rules (the high-cost class in L), while labels k_1 and k_2 correspond to common classes with simpler criteria.

We split the dataset into two disjoint subsets: a training set $\mathcal{D}_t$ with 20,000 samples and an evaluation set $\mathcal{D}_e$ with 100 samples. Each entry $(\boldsymbol{x}_t, \{y_l\}_{l=1}^{L}, y_t, r_t, y^*)$ contains feature vectors, class labels, and rule-level annotations.

Table 2. Ground-truth rule set for the synthetic generator.

Label	Rules	Weight
k_0	1: $x_0 \wedge x_1 \wedge \neg x_2 \wedge x_3$	1.5
	2: $x_3 \wedge x_4 \wedge x_7 \wedge \neg x_9$	1.5
k_1	3: $x_3 \wedge x_4 \wedge x_5$	1.4
	4: $x_6 \wedge x_7 \wedge x_9$	1.6
k_2	5: $x_1 \wedge x_3 \wedge x_4$	1.7
	6: $x_4 \wedge x_7 \wedge x_9$	1.3

Agent simulator. Each agent is modeled as a rule-based probabilistic decision-maker equipped with its own rule set, conditions, classes, and weights. These rule sets may deviate from the ground-truth rules, reflecting heterogeneous expertise and bias. Unlike a deterministic classifier, an agent samples its action from the softmax distribution induced by its rule weights.

To evaluate collaborative decision-making under diverse conditions, we instantiate the four scenarios as pairs of agents with different rule sets and weights. Representative configurations are shown in Table 3.

Table 3. Rule sets assigned to each rule-based decision-maker under the four scenarios.

Scenario	Model	Rule Set	Weight
Scenario 1 Ideal	Agent 1	$a_1 \leftarrow x_3 \wedge x_4 \wedge x_5$	1.4
		$a_1 \leftarrow x_6 \wedge x_7 \wedge x_9$	1.6
		$a_2 \leftarrow x_1 \wedge x_3 \wedge x_4$	1.7
		$a_2 \leftarrow x_4 \wedge x_7 \wedge x_9$	1.3
		$a_0 \leftarrow x_3 \wedge x_4$	1.3
	Agent 2	$a_0 \leftarrow x_3 \wedge x_4 \wedge x_7 \wedge \neg x_9$	1.5
		$a_0 \leftarrow x_0 \wedge x_1 \wedge \neg x_2 \wedge x_3$	1.5
		$a_1 \leftarrow x_3 \wedge x_4$	1.5
		$a_2 \leftarrow x_1 \wedge x_3$	1.5
Scenario 2 Asymmetric	Agent 1	$a_1 \leftarrow x_3 \wedge x_4 \wedge x_5$	1.4
		$a_1 \leftarrow x_6 \wedge x_7 \wedge x_9$	1.6
		$a_2 \leftarrow x_1 \wedge x_3 \wedge x_4$	1.7
		$a_0 \leftarrow x_3 \wedge x_4 \wedge x_7 \wedge \neg x_9$	1.5
		$a_0 \leftarrow x_0 \wedge x_1 \wedge \neg x_2 \wedge x_3$	1.5
		$a_2 \leftarrow x_1 \wedge x_3$	1.2
	Agent 2	$a_2 \leftarrow x_4 \wedge x_7 \wedge x_9$	1.7
		$a_0 \leftarrow x_3 \wedge x_4$	1.5
		$a_1 \leftarrow x_3 \wedge x_4$	1.5
		$a_2 \leftarrow x_1 \wedge x_3$	1.3
Scenario 3a Noisy-complementary	Agent 1	$a_1 \leftarrow x_3 \wedge x_4 \wedge x_5$	1.4
		$a_1 \leftarrow x_6 \wedge x_7 \wedge x_9$	1.6
		$a_2 \leftarrow x_1 \wedge x_3 \wedge x_4$	1.7
		$a_0 \leftarrow x_3 \wedge x_4$	1.3
		$a_2 \leftarrow x_1 \wedge x_3$	1.3
	Agent 2	$a_2 \leftarrow x_4 \wedge x_7 \wedge x_9$	1.3
		$a_0 \leftarrow x_3 \wedge x_4 \wedge x_7 \wedge \neg x_9$	1.5
		$a_0 \leftarrow x_0 \wedge x_1 \wedge \neg x_2 \wedge x_3$	1.5
		$a_1 \leftarrow x_3 \wedge x_4$	1.3
		$a_2 \leftarrow x_1 \wedge x_3$	1.0
Scenario 3b Noisy-non-complementary	Agent 1	$a_1 \leftarrow x_6 \wedge x_7 \wedge x_9$	1.6
		$a_0 \leftarrow x_3 \wedge x_4$	1.5
		$a_1 \leftarrow x_3 \wedge x_4$	1.4
		$a_2 \leftarrow x_1 \wedge x_3$	1.5
	Agent 2	$a_2 \leftarrow x_4 \wedge x_7 \wedge x_9$	1.3
		$a_0 \leftarrow x_3 \wedge x_4$	1.5
		$a_1 \leftarrow x_3 \wedge x_4$	1.4
		$a_2 \leftarrow x_1 \wedge x_3$	1.1

D.2. Clinical LLM Datasets

For the LLM setting we use two publicly available clinical datasets, both recast as multi-class diagnostic tasks consumed by two black-box LLM agents. A subset of classes in each dataset is designated high-risk and assigned the higher entry of the asymmetric loss L (Appendix E).

MedQA-USMLE. MedQA-USMLE (Jin et al., 2021) is a multiple-choice question-answering dataset derived from the United States Medical Licensing Examination. We re-cast its English Step-1 dev split as *MedQA-Tri*, a $K_{\mathrm{MedQA}} = 7$-way clinical-specialty triage task over {Cardiology, Pulmonology, Neurology, Gastroenterology, Endocrinology, Infectious_Disease, Renal_Urology}. Items are first filtered to diagnosis-style stems (questions asking for a diagnosis, cause, or differential, rather than treatment, management, or mechanism), and each retained item is mapped to one of the seven specialties by a keyword heuristic over the stem and the canonical diagnosis text, with an LLM fallback (`gpt-4o-mini` reading the stem and answer text) for items the heuristic skips. The resulting pool contains 1394 class-stratified cases. Per (scenario, seed) we draw a calibration split of $|\mathrm{CAL}| = 70$ and a test split of $|\mathrm{TEST}| = 140$, both class-stratified; results are reported across 10 seeds. The high-risk subset $\mathcal{H}_{\mathrm{MedQA}} = \{$Cardiology, Pulmonology, Neurology, Infectious_Disease$\}$ – the four specialties whose missed call most plausibly corresponds to a life-threatening miss – is assigned $\ell_{\mathrm{high}} = 5$ in L.

DDXPlus. DDXPlus (Fansi Tchango et al., 2022) is a patient-case differential-diagnosis dataset introduced at NeurIPS 2022. Each case is a structured tuple of patient demographics, present-tense and historical evidence codes drawn from a fixed medical ontology, and a ground-truth pathology. We restrict the label space to a $K_{\mathrm{DDX}} = 8$-way subset over four high-risk and four benign pathologies, and render each case as a natural-language vignette by templating the evidence codes

into a short clinical paragraph (e.g. "A 47-year-old male presents with sharp pleuritic chest pain that started suddenly…"); the LLM agents see only the rendered text. Pool construction caps each pathology at 350 class-balanced cases for a total of 2800. Per (scenario, seed) we draw $|\mathrm{CAL}| = 80$ and $|\mathrm{TEST}| = 160$, both class-stratified; results are reported across 10 seeds. The high-risk subset $\mathcal{H}_{\mathrm{DDX}} = \{$Anaphylaxis, Pulmonary_embolism, NSTEMI_STEMI, Myocarditis$\}$ is composed of severity-1–2 acute pathologies; the benign subset is $\{$URTI, Viral_pharyngitis, GERD, Bronchitis$\}$. Labels are taken directly from the released pathology field, so no labelling LLM fallback is needed. We use $\ell_{\mathrm{high}} = 5$ in L.

E. Asymmetric Loss and Risk-Aware Metrics

Asymmetric loss. Each dataset is equipped with an asymmetric loss $L(d, y)$ on a coarse three-level scale:

$$L(d,y) = \begin{cases} 0 & d = y, \\ \ell_{\mathrm{high}} & d \neq y \text{ and } y \text{ is a high-cost class}, \\ 1 & \text{otherwise}, \end{cases}$$

where $\ell_{\mathrm{high}} > 1$ is a tunable scalar that controls how aggressively the cost-aware decision rule defends the high-cost class. Larger values force the protocol to explore more under uncertainty (more DIFFERENTIAL_STEER and STOP_AND_ESCALATE); smaller values keep the rule close to standard accuracy maximisation. The high-cost subset for each dataset is listed in Table 4.

Table 4. High-cost class subset per dataset. All other off-diagonal entries of L are 1 and the diagonal is 0.

Dataset	High-cost class(es)	ℓ_{high}
Synthetic	k_0 (rare, long-rule class)	3 (default)
MedQA-USMLE	Cardiology, Pulmonology, Neurology, Infectious_Disease	5
DDXPlus	Anaphylaxis, Pulmonary_embolism, NSTEMI_STEMI, Myocarditis	5

Risk-aware metrics. For N test cases with terminal mediator action $u^{(\tau_i)}$, terminal pooled decision $d_1^{(\tau_i)}$, and round count τ_i, we report the following empirical metrics, abbreviating STOP_AND_DECIDE as DEC and STOP_AND_ESCALATE as ESC. Let $\mathcal{H} \subseteq \{1, \ldots, K\}$ denote the high-cost class subset (Table 4) and $\mathcal{H}_{\mathrm{test}} = \{i : y_i \in \mathcal{H}\}$ the indices of test cases with high-cost ground truth.

$$\widehat{R}_L = \tfrac{1}{N} \sum_i L\big(d_1^{(\tau_i)},\, y_i\big) \qquad \text{(expected diagnostic cost)},$$

$$\widehat{R}_L^{\mathrm{sys}} = \tfrac{1}{N} \sum_i \mathbb{1}\big[u^{(\tau_i)} \neq \textsc{Esc}\big]\, L\big(d_1^{(\tau_i)},\, y_i\big) \qquad \text{(system risk under expert backstop)},$$

$$\mathrm{Acc}_{\mathrm{pool}} = \tfrac{1}{N} \sum_i \mathbb{1}\big[d_1^{(\tau_i)} = y_i\big] \qquad \text{(pooled accuracy)},$$

$$\mathrm{HighRiskMiss} = \frac{1}{|\mathcal{H}_{\mathrm{test}}|} \sum_{i\,:\,y_i \in \mathcal{H}} \mathbb{1}\big[d_1^{(\tau_i)} \neq y_i\big] \qquad \text{(high-risk miss rate)},$$

$$\mathrm{HarmfulConsensus} = \tfrac{1}{N} \sum_i \mathbb{1}\big[d_1^{(\tau_i)} \neq y_i \,\wedge\, y_i \in \mathcal{H} \,\wedge\, u^{(\tau_i)} \neq \textsc{Esc}\big] \qquad \text{(harmful-consensus rate)},$$

$$\mathrm{Certified} = \tfrac{1}{N} \sum_i \mathbb{1}\big[u^{(\tau_i)} = \textsc{Dec}\big] \qquad \text{(certified-decision rate)},$$

$$\mathrm{Escalation} = \tfrac{1}{N} \sum_i \mathbb{1}\big[u^{(\tau_i)} = \textsc{Esc}\big] \qquad \text{(escalation rate)}.$$

$\widehat{R}_L$ is the empirical risk under L; the high-risk miss rate isolates high-cost truths that are misclassified; the harmful-consensus rate measures cases where the protocol commits to a wrong decision on a high-cost truth without escalation, the operational measure of premature diagnostic closure; certified and escalation rates summarise the terminal mediator actions.

F. Partial-Compliance Ablation

We use the rule-guided setting to run a controlled ablation of agent compliance. Each mediator-induced update is adopted with probability $\xi \in [0, 1]$:

$$w_{a^\dagger}^{(t+1)} = \Pi_{\mathcal{W}} \left(\begin{cases} w_{\text{rec}}^{(t)} & \text{with probability } \xi, \\ w_{a^\dagger}^{(t)} & \text{with probability } 1-\xi. \end{cases} \right). \tag{21}$$

Here $\xi = 0$ means the agent ignores the mediator update, while $\xi = 1$ means full adoption.

We sweep $\xi \in \{25\%, 50\%, 75\%, 100\%\}$ and run the protocol for fixed horizons $\tau \in \{4, 8, 12\}$ on the synthetic diagnostic task. This is a fixed-horizon ablation: STOP_AND_DECIDE and STOP_AND_ESCALATE are recorded as mediator signals, but do not terminate the run before the horizon. Table 5 reports mean±std over 10 seeds. In S1, S2, and S3a, higher compliance and longer horizons reduce $\widehat{R}_L$ and improve agent accuracy. In S3b, the shared bias is intentionally non-correctable; the key result is that escalation increases with horizon and is insensitive to ξ, showing that premature-closure detection does not require agents to accept the mediator's suggestions.

Table 5. Partial-compliance ablation on the synthetic diagnostic task ($T \in \{4, 8, 12\}$, mean±std over 10 seeds). $\widehat{R}_L$ = diagnostic cost; SA-HR = max single-agent high-risk miss rate; Esc. = STOP_AND_ESCALATE rate. Bold marks the headline metric: $\widehat{R}_L\downarrow$ for S1/S2/S3a and Esc.↑ for S3b.

Scenario 1 – Ideal															
	$T=4$					$T=8$					$T=12$				
ξ	A1	A2	$\widehat{R}_L\downarrow$	SA-HR↓	Esc.	A1	A2	$\widehat{R}_L\downarrow$	SA-HR↓	Esc.	A1	A2	$\widehat{R}_L\downarrow$	SA-HR↓	Esc.
25%	0.62±0.07	0.80±0.04	**0.22**±0.06	0.49±0.09	0.72±0.13	0.75±0.16	0.85±0.09	**0.09**±0.11	0.37±0.23	0.60±0.21	0.83±0.17	0.90±0.10	**0.05**±0.09	0.25±0.26	0.40±0.31
50%	0.63±0.08	0.82±0.06	**0.19**±0.09	0.44±0.17	0.73±0.14	0.87±0.15	0.97±0.07	**0.01**±0.03	0.07±0.18	0.23±0.29	0.99±0.03	1.00±0.00	**0.00**±0.01	0.00±0.00	0.00±0.01
75%	0.63±0.06	0.91±0.09	**0.09**±0.11	0.19±0.25	0.67±0.17	0.98±0.07	1.00±0.00	**0.00**±0.00	0.00±0.00	0.01±0.03	1.00±0.00	1.00±0.00	**0.00**±0.00	0.00±0.00	0.00±0.00
100%	0.62±0.04	1.00±0.01	**0.00**±0.00	0.00±0.00	0.66±0.15	1.00±0.00	1.00±0.00	**0.00**±0.00	0.00±0.00	0.00±0.00	1.00±0.00	1.00±0.00	**0.00**±0.00	0.00±0.00	0.00±0.00
Scenario 2 – Asymmetric															
	$T=4$					$T=8$					$T=12$				
ξ	A1	A2	$\widehat{R}_L\downarrow$	SA-HR↓	Esc.	A1	A2	$\widehat{R}_L\downarrow$	SA-HR↓	Esc.	A1	A2	$\widehat{R}_L\downarrow$	SA-HR↓	Esc.
25%	0.92±0.04	0.52±0.08	**0.08**±0.04	0.55±0.16	0.41±0.21	0.96±0.05	0.58±0.09	**0.04**±0.05	0.51±0.09	0.51±0.26	0.98±0.04	0.62±0.12	**0.02**±0.04	0.53±0.13	0.59±0.25
50%	0.93±0.05	0.55±0.07	**0.07**±0.05	0.53±0.13	0.43±0.20	0.98±0.04	0.70±0.12	**0.02**±0.04	0.48±0.11	0.46±0.25	1.00±0.00	0.77±0.07	**0.00**±0.00	0.45±0.16	0.43±0.18
75%	0.95±0.05	0.59±0.07	**0.06**±0.05	0.51±0.09	0.52±0.22	0.99±0.04	0.77±0.09	**0.02**±0.03	0.44±0.17	0.40±0.22	1.00±0.00	0.88±0.09	**0.00**±0.00	0.15±0.23	0.19±0.23
100%	0.95±0.05	0.67±0.10	**0.05**±0.05	0.50±0.05	0.51±0.26	1.00±0.00	0.91±0.06	**0.00**±0.00	0.07±0.18	0.10±0.15	1.00±0.00	0.95±0.03	**0.00**±0.00	0.00±0.00	0.05±0.03
Scenario 3a – Noisy-complementary															
	$T=4$					$T=8$					$T=12$				
ξ	A1	A2	$\widehat{R}_L\downarrow$	SA-HR↓	Esc.	A1	A2	$\widehat{R}_L\downarrow$	SA-HR↓	Esc.	A1	A2	$\widehat{R}_L\downarrow$	SA-HR↓	Esc.
25%	0.61±0.03	0.60±0.05	**0.40**±0.08	0.54±0.04	0.73±0.23	0.63±0.05	0.65±0.08	**0.30**±0.16	0.53±0.04	0.67±0.23	0.68±0.09	0.69±0.09	**0.20**±0.16	0.51±0.05	0.66±0.23
50%	0.63±0.05	0.64±0.09	**0.31**±0.16	0.52±0.05	0.65±0.24	0.68±0.08	0.73±0.08	**0.15**±0.15	0.51±0.05	0.61±0.22	0.75±0.07	0.78±0.07	**0.04**±0.07	0.50±0.05	0.51±0.17
75%	0.64±0.06	0.71±0.09	**0.16**±0.14	0.51±0.05	0.61±0.23	0.75±0.06	0.77±0.05	**0.05**±0.06	0.51±0.05	0.54±0.16	0.84±0.08	0.82±0.06	**0.01**±0.02	0.40±0.20	0.34±0.17
100%	0.65±0.07	0.74±0.07	**0.11**±0.08	0.50±0.05	0.55±0.22	0.80±0.06	0.79±0.03	**0.02**±0.03	0.46±0.15	0.45±0.14	0.96±0.07	0.86±0.10	**0.00**±0.01	0.12±0.21	0.12±0.14
Scenario 3b – Noisy-non-complementary (premature-closure stress test)															
	$T=4$					$T=8$					$T=12$				
ξ	A1	A2	$\widehat{R}_L\downarrow$	SA-HR↓	Esc.↑	A1	A2	$\widehat{R}_L\downarrow$	SA-HR↓	Esc.↑	A1	A2	$\widehat{R}_L\downarrow$	SA-HR↓	Esc.↑
25%	0.56±0.06	0.47±0.03	0.66±0.04	1.00±0.00	**0.54**±0.27	0.58±0.06	0.47±0.03	0.66±0.05	1.00±0.00	**0.76**±0.31	0.59±0.05	0.47±0.03	0.66±0.04	1.00±0.00	**0.89**±0.23
50%	0.56±0.06	0.47±0.03	0.66±0.04	1.00±0.00	**0.54**±0.27	0.58±0.06	0.47±0.03	0.66±0.05	1.00±0.00	**0.76**±0.31	0.59±0.05	0.47±0.03	0.66±0.04	1.00±0.00	**0.89**±0.23
75%	0.56±0.06	0.47±0.03	0.66±0.04	1.00±0.00	**0.54**±0.27	0.58±0.06	0.47±0.03	0.66±0.05	1.00±0.00	**0.76**±0.31	0.59±0.05	0.47±0.03	0.66±0.04	1.00±0.00	**0.89**±0.23
100%	0.56±0.06	0.47±0.03	0.66±0.04	1.00±0.00	**0.54**±0.27	0.58±0.06	0.47±0.03	0.66±0.05	1.00±0.00	**0.76**±0.31	0.59±0.05	0.47±0.03	0.66±0.04	1.00±0.00	**0.89**±0.23

G. Prompt Templates for Agent Reports

This appendix gives the minimal prompts used for agent-facing LLM calls. The synthetic rule-guided experiments use the multinomial-logit simulator; no black-box LLM is used to learn or overwrite the simulator's rule weights. We include the rule-guided interface template only to document the report format and the wording of the differential-steering message. The clinical dataset experiments use black-box LLM agents that return a probability vector over the fixed label set.

Rule-Guided Agent Interface Template

Role. You are Agent {`AGENT_ID`} in a synthetic diagnostic classification task. You are given a binary feature vector, a fixed label set, and a fixed list of rules of the form `IF conditions THEN class = c`. Use the active rules and their current weights to assess the class probabilities. Do not create new rules, delete rules, change rule conditions, or change the label set.
Input.

- `features`: binary feature vector for the current case.
- `classes`: fixed candidate labels.
- `rules`: fixed rules with current weights.

- `role`: optional description of the agent's stronger and weaker rule groups.
- `mediator_message`: optional `CONTINUE` or `DIFFERENTIAL_STEER` message.

Differential steering. If the mediator sends `DIFFERENTIAL_STEER`, treat it only as a targeted request to reassess the current prediction against one dangerous alternative:

> *You currently predict {`CURRENT_LABEL`}. Historical calibration suggests that this prediction can be confused with {`ALTERNATIVE_LABEL`}. Re-check the active features or rules {`S`} that distinguish these two labels.*

The mediator does not provide ground truth and does not force a revision. Revise only if the active rules support a different probability allocation.

Output.

```
{
  "predicted_label": "<one label from classes>",
  "probabilities": {"<label_1>": 0.0, "...": 0.0},
  "active_rules": ["<rule ids used as evidence>"]
}
```

Clinical LLM Agent Classification Template

System prompt. You are {`AGENT_ID`}, a clinical classification agent. Classify each case into exactly one label from the provided label set. Your role may specify stronger and weaker clinical areas; use that role as a source of heterogeneous expertise, but be honest about uncertainty. Do not provide treatment advice or introduce labels outside the allowed set. Return strict JSON only.

User prompt.

```
Task: classify the following clinical case into exactly one label.

Labels: {LABEL_SET}

Clinical case:
{CASE_TEXT}

Mediator note (optional, present from round 2 onward):
{MEDIATOR_NOTE}

Return ONLY JSON:
{
  "predicted_label": "<one label from Labels>",
  "probabilities": {"<each label in Labels>": <number>}
}
```

Probabilities are normalized over all labels. The mediator uses the reported probability vector together with a frozen calibration confusion matrix; the agent's self-reported confidence is not used as a reliability weight. No labelled examples are passed to the agent in any round.

CONTINUE note. When the mediator issues CONTINUE, the inserted note is a short re-examination prompt:

> *Your previous best guess favoured {`TOP1_LABEL`} over {`TOP2_LABEL`}. Take a fresh look at the case and reconsider whether any other label is more consistent with the evidence; adjust your probabilities accordingly.*

DIFFERENTIAL_STEER note. When the mediator issues DIFFERENTIAL_STEER targeting alternative {`ALT_LABEL`}, the agent is first asked, in a separate call, to enumerate a small set S of short distinguishing cues between its current prediction {`CURRENT_LABEL`} and {`ALT_LABEL`}. Those cues are inserted into the next classification call as:

> *You earlier favoured {`CURRENT_LABEL`}. Historical calibration identifies {`ALT_LABEL`} as a plausible high-cost alternative for this case. Re-evaluate using the distinguishing cues you yourself listed: {`CUES`}. If the cues for {`ALT_LABEL`} are present, shift probability toward it; otherwise keep {`CURRENT_LABEL`}. Stay within the allowed label set.*

The mediator never overwrites the agent's distribution; the cue list S is produced by the targeted agent itself, so the steering message remains auditable and matches the description of $S^{(t)}$.

Dataset-Specific Label and Role Inserts

MedQA-Tri.

```
LABEL_SET =
["Cardiology", "Pulmonology", "Neurology", "Gastroenterology",
 "Endocrinology", "Infectious_Disease", "Renal_Urology"]

CASE_TEXT = the filtered MedQA-USMLE diagnosis-style stem plus the
canonical answer text used to map the item into a specialty.
```

High-risk specialties are Cardiology, Pulmonology, Neurology, and Infectious_Disease. Scenario roles are inserted as short constraints: *S1*: neutral senior clinician; *S2*: one neutral agent and one agent with a dataset-specific default-label anchor that induces systematic misses; *S3a*: agents emphasize complementary specialty subsets; *S3b*: both agents share the same default-label anchor.

DDXPlus.

```
LABEL_SET =
["Anaphylaxis", "Pulmonary_embolism", "NSTEMI_STEMI", "Myocarditis",
 "URTI", "Viral_pharyngitis", "GERD", "Bronchitis"]

CASE_TEXT = the rendered patient vignette built from demographics and
present/history evidence codes.
```

High-risk diagnoses are Anaphylaxis, Pulmonary_embolism, NSTEMI_STEMI, and Myocarditis. Scenario roles use the same template: *S1*: neutral senior clinicians; *S2*: one neutral agent and one benign-presentation anchor (URI/GERD) that under-calls high-risk diagnoses; *S3a*: complementary acute care versus outpatient emphasis; *S3b*: the shared benign-anchor threat model.

H. LLM-Agent Extension: Full Setup and Results

This appendix gives the implementation details, the per-method numbers behind the main figure, the D-CEM behavior summary. Unless stated otherwise, all numbers are mean $\pm$ std across 10 seeds.

H.1. Implementation details

Datasets. We use the two clinical datasets specified in Appendix D.2: *MedQA-Tri* (7-way specialty triage, $\ell_{\text{high}} = 5$) and *DDXPlus* (8-way differential diagnosis, $\ell_{\text{high}} = 5$). The four scenarios S1–S3b are realised by hard role constraints in each agent's system prompt (templates and label inserts in Appendix G): *S1 Ideal*: both agents take a neutral senior-clinician role. *S2 Asymmetric*: agent A stays neutral; agent B is assigned a dataset-specific default-label anchor that induces systematic misses. *S3a Noisy-complementary*: A and B are anchored on disjoint reliable subsets (cardiac/PE specialist vs. outpatient generalist on DDXPlus; cardiology–GI–endocrine vs. pulmonology–neurology–infectious on MedQA-Tri). *S3b Noisy-non-complementary*: both agents share the *same* default-label anchor of S2 – the harmful-pooling threat model.

Agents and protocol. Agent A is `gpt-4o-mini` (OpenAI), agent B is `claude-3.5-haiku` (Anthropic), both queried at `temperature`$= 0.3$ with a fixed JSON schema returning a diagnostic label and a probability vector $p \in \Delta^K$. Mediator reliabilities $\rho_a^{(t)}$ are computed from calibrated confusion-induced posteriors, not self-reported by the LLMs. Disk caching is keyed by `(vendor, model, temperature, full message stack)`. The mediator runs at most $T = 6$ rounds per case (*Free discussion* is fixed at $T = 6$). Confusion matrices C_a and D-CEM energy thresholds are both estimated on the CAL split with diagonal-biased Laplace smoothing ($\alpha = 0.5$) and frozen for all TEST evaluations. We use $|\text{CAL}|/|\text{TEST}| = 70/140$ over 10 seeds on MedQA-Tri and $80/160$ over 10 seeds on DDXPlus.

H.2. Per-method headline numbers

Table 6 reports system risk $\widehat{R}_L^{\text{sys}}$ and harmful-consensus rate H.C. for all four methods on every scenario. D-CEM achieves the lowest H.C. and the lowest $\widehat{R}_L^{\text{sys}}$ in every cell, with the smallest margin on the saturated DDXPlus $S1$ (gap to Single is only 0.02, since Single already reaches $\text{Acc} = 0.95$ and no protocol has fusion headroom to exploit).

Table 6. Per-method results on both clinical datasets. SINGLE the CAL-best single-agent baseline (the better of A, B as measured on the CAL split); FREE is six-round free discussion; POOL is one-shot fixed pooling. Lower is better for both metrics.

		$\widehat{R}_L^{\text{sys}}\downarrow$				H.C.$\downarrow$			
Dataset	Method	S1	S2	S3a	S3b	S1	S2	S3a	S3b
MedQA-Tri	SINGLE	$0.73_{\pm.06}$	$0.73_{\pm.06}$	$0.66_{\pm.09}$	$0.74_{\pm.13}$	$0.19_{\pm.03}$	$0.19_{\pm.03}$	$0.16_{\pm.04}$	$0.18_{\pm.04}$
	FREE	$0.64_{\pm.07}$	$0.64_{\pm.04}$	$0.66_{\pm.04}$	$0.70_{\pm.07}$	$0.16_{\pm.03}$	$0.16_{\pm.02}$	$0.16_{\pm.02}$	$0.16_{\pm.02}$
	POOL	$0.61_{\pm.08}$	$0.69_{\pm.06}$	$0.61_{\pm.11}$	$0.76_{\pm.14}$	$0.15_{\pm.04}$	$0.17_{\pm.02}$	$0.14_{\pm.04}$	$0.18_{\pm.04}$
	D-CEM	$\mathbf{0.47}_{\pm.10}$	$\mathbf{0.51}_{\pm.03}$	$\mathbf{0.45}_{\pm.06}$	$\mathbf{0.58}_{\pm.08}$	$\mathbf{0.11}_{\pm.04}$	$\mathbf{0.13}_{\pm.01}$	$\mathbf{0.11}_{\pm.03}$	$\mathbf{0.14}_{\pm.03}$
DDXPlus	SINGLE	$0.08_{\pm.04}$	$0.31_{\pm.07}$	$1.47_{\pm.08}$	$0.74_{\pm.06}$	$0.01_{\pm.01}$	$0.05_{\pm.01}$	$0.29_{\pm.01}$	$0.14_{\pm.01}$
	FREE	$0.49_{\pm.40}$	$1.61_{\pm.28}$	$1.93_{\pm.01}$	$1.43_{\pm.27}$	$0.09_{\pm.08}$	$0.31_{\pm.06}$	$0.38_{\pm.00}$	$0.28_{\pm.05}$
	POOL	$0.64_{\pm.35}$	$1.90_{\pm.13}$	$1.91_{\pm.02}$	$1.66_{\pm.15}$	$0.12_{\pm.07}$	$0.37_{\pm.02}$	$0.38_{\pm.00}$	$0.32_{\pm.03}$
	D-CEM	$\mathbf{0.06}_{\pm.03}$	$\mathbf{0.03}_{\pm.02}$	$\mathbf{0.03}_{\pm.02}$	$\mathbf{0.21}_{\pm.09}$	$\mathbf{0.00}_{\pm.01}$	$\mathbf{0.00}_{\pm.00}$	$\mathbf{0.00}_{\pm.01}$	$\mathbf{0.01}_{\pm.01}$

H.3. D-CEM behavior summary

The takeaway from Table 7 is that **D-CEM's escalation rate and round usage both adapt to per-scenario agent reliability rather than firing on a fixed schedule.**

- **DDXPlus has a wide reliability gradient**: Single-best accuracy ranges from 0.95 on $S1$ down to 0.68 on $S3a$. D-CEM responds with escalation rates rising $9\times$ ($5\% \rightarrow 46\%$) and average rounds rising from 2.1 to 3.8: it commits quickly when agents are reliable, and either deliberates longer or abstains when correlated noise makes both agents systematically wrong.
- **MedQA-Tri has a flat reliability profile**: Single-best accuracy stays in 0.62–0.68 across all four scenarios. D-CEM correspondingly settles into a tight 11–16% escalation band rather than triggering on every uncertain case.

Average rounds stay well below $T_{\max} = 6$ in every cell, confirming that the cost-aware termination rule does not exhaust the round budget.

Table 7. D-CEM per-case escalation rate (share of cases on which STOP_AND_ESCALATE fires) and average mediator rounds T (out of $T_{\max} = 6$).

	Escalation rate				Avg. rounds T			
Dataset	S1	S2	S3a	S3b	S1	S2	S3a	S3b
MedQA-Tri	$0.11_{\pm.03}$	$0.14_{\pm.02}$	$0.16_{\pm.02}$	$0.14_{\pm.03}$	$2.63_{\pm.28}$	$2.82_{\pm.33}$	$2.73_{\pm.10}$	$3.85_{\pm.38}$
DDXPlus	$0.05_{\pm.02}$	$0.33_{\pm.08}$	$0.46_{\pm.05}$	$0.20_{\pm.06}$	$2.13_{\pm.17}$	$3.28_{\pm.26}$	$3.59_{\pm.29}$	$3.82_{\pm.19}$

I. Limitations and Broader Impacts

This paper studies the minimal nontrivial setting of two-agent diagnostic deliberation, providing a clean foundation for analysing confusion-aware decision dynamics. The proposed framework naturally extends to larger multi-agent systems, since the underlying energy decomposition and action space remain unchanged; exploring these richer settings empirically is an important direction for future work. Our evaluation focuses on multi-class diagnostic tasks with pre-specified asymmetric losses, and the LLM-agent experiments employ representative API-hosted models on recast clinical datasets under controlled role anchors. Extending the empirical study to broader task families, model backbones, and real-world deployment settings offers promising opportunities for future research.

D-CEM is designed as a protocol-level safeguard against premature diagnostic closure, transforming potentially unsafe agreement into either a loss-sensitive certified decision or principled escalation. This design can help improve reliability in high-stakes multi-agent and multi-LLM decision pipelines. Because the certificate is explicitly loss-sensitive, the loss matrix should be transparently aligned with the target application rather than tuned solely for aggregate accuracy. Importantly, D-CEM is intended to augment human decision-making rather than replace it, and the built-in STOP_AND_ESCALATE mechanism explicitly preserves unresolved high-risk cases for additional review.