WHEN CONSENSUS MASKS RISK: ENERGY-GUIDED CONTROL OF DIAGNOSTIC DISCUSSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-agent deliberation can improve high-stakes classification, yet consensus alone does not establish safety. In diagnostic-style decisions, agents with asymmetric expertise may reach agreement while high-cost alternatives remain unresolved, a failure mode we call *premature diagnostic closure*. We cast this process as a protocol-control problem without online ground truth: using agents' reports, optional justifications, calibration confusion matrices, and an asymmetric loss, a mediator must decide whether to continue, challenge, certify, or escalate. We propose *Diagnostic Consensus Energy Minimization* (D-CEM), a loss-aware controller that interprets 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 policy continues while risk decreases, issues targeted differential challenges, certifies only low-energy large-margin decisions, and escalates when high-risk deliberation stagnates. We derive loss-sensitive risk bounds, calibrated decision guarantees, and cost-aware stopping bounds. Experiments on synthetic and clinical tasks show that D-CEM reduces high-risk misses and harmful consensus while improving safety–cost trade-offs over aggregation baselines.

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). A detailed review of the related literature is in Appendix A.

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 (CM) (Kerrigan et al., 2021; Steyvers et al., 2022), and ex post, for evaluation. The mediator must use these observable traces to decide when to continue deliberation, challenge a dangerous differential diagnosis, certify, or escalate (Madras et al., 2018; Mozannar & Sontag, 2020; Geifman & El-Yaniv, 2019).

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.

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 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. Its central contribution is to turn offline behavioral calibration into a computable, loss-aware signal that controls deliberation without online ground truth. 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 connects diagnostic energy and loss margins to pooled model risk, and independent calibration to population risk after adaptive stopping. The dangerous plausible miss maximizes the value of a perfect diagnostic check under class-weighted loss. We quantify the stopping cost of energy fluctuations and bound the time to escalation after actionable challenges are exhausted (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.

2 PROBLEM SETUP AND CONFUSION-INDUCED DIAGNOSTIC RISK

D-CEM regulates two-agent deliberation using reports and fixed calibration statistics, without observing the current case's ground-truth label. Its goal is to distinguish safe consensus from *premature diagnostic closure*. We first define the mediator's information and risk-aware objective, then construct the confusion-induced posteriors used for control.

2.1 DIAGNOSTIC DELIBERATION UNDER ASYMMETRIC EXPERTISE

Task and diagnostic loss. We consider K diagnostic labels $\mathcal{Y} = \{1, \dots, K\}$ and decisions $d \in \mathcal{D}$, with $\mathcal{D} = \mathcal{Y}$ unless specified otherwise. An asymmetric loss matrix $L \in \mathbb{R}_{+}^{K \times K}$ assigns cost $L(d, y)$ to decision d when the true label is y. Costs 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

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

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 mediator's actions are motivated by a costed risk objective that is only partially observable 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, this lets the mediator monitor *two failure modes* simultaneously: unresolved disagreement, and high-cost agreement.

For agent a, let

$$C_a[i, j] \approx \Pr_{\text{hist}}(\hat{y}_a = j \mid Y = i)$$

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 model posterior over the latent diagnosis:

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

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?*

These agent-level posteriors and dangerous plausible misses provide the inputs to the pooled decision and diagnostic-risk energy defined next.

3 Diagnostic Consensus Energy and Deliberation Policy

3.1 Diagnostic-Risk Energy

We combine the agents' confusion-induced posteriors into a pooled decision and diagnostic-risk energy.

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:

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

with $Q^{(t)}$ normalized over $\mathcal{Y}$.

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

$$d_1^{(t)} \in \arg\min_d R_L(d; Q^{(t)}), \tag{7}$$

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

$$\Delta_L^{(t)} = R_L(d_2^{(t)}; Q^{(t)}) - R_L(d_1^{(t)}; Q^{(t)}). \tag{8}$$

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**

$$\varepsilon_{\text{diag}}^{(t)} = \alpha D_{\text{sym}}(P_1^{(t)}, P_2^{(t)}) + \beta \sum_{a=1}^{2} R_L(\hat{y}_a^{(t)}; P_a^{(t)}) + \gamma \exp(-\Delta_L^{(t)}), \tag{9}$$

where

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

The three terms penalize posterior disagreement, expected harm of the current reports, and a small loss-aware margin. Agreement alone therefore need not yield low energy: high-risk alternatives can keep the report-risk or margin penalty large.

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

$$\widehat{g}^{(t)} = \frac{1}{W} \sum_{i=0}^{W-1} \left(\varepsilon_{\text{diag}}^{(t-1-i)} - \varepsilon_{\text{diag}}^{(t-i)} \right), \tag{10}$$

with window $W \geq 1$. Set $\widehat{g}^{(t)} = +\infty$ for $t \leq W$, so stagnation is tested only after W observed energy changes. The mediator applies the following ordered policy. Certification denotes passage through the model-risk gates; population-risk calibration is given in Theorem 1:

- **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 $d_1^{(t)}$.
- **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 equation 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.
- **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.
- **CONTINUE.** Otherwise, continue deliberation and recompute the posteriors and energy from the updated reports.

Stopping time and parameters. Let $A^{(t)} = 1$ when the selected challenge passes the risk gate and has remaining allowance (Appendix B.1), and zero otherwise. The induced stopping time is

$$\begin{aligned}\tau^* = \inf\big\{t \ge 1 : (\varepsilon_{\text{diag}}^{(t)} \le \varepsilon_{\text{safe}} \ \wedge \ \Delta_L^{(t)} \ge m_{\text{safe}}) \\ \vee \ (A^{(t)} = 0 \ \wedge \ \widehat{g}^{(t)} < \delta \ \wedge \ \varepsilon_{\text{diag}}^{(t)} > \varepsilon_{\text{low}})\big\}.\end{aligned} \tag{11}$$

Here $\varepsilon_{\text{safe}}$ is the target diagnostic-risk tolerance, m_{safe} is the required diagnostic margin, δ is the per-round cost in energy units, 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 mediator asks agent $a^\dagger(t)$ to reassess its current report against $z_{a^\dagger}^{(t)}$ using distinguishing evidence $S^{(t)}$ drawn from available justifications, feature or rule signals, or natural-language rationales. The challenge supplies no ground truth and does not overwrite the report; the agent may revise or retain its assessment. Appendix B.3 gives a feature-level construction, and Appendix J provides the prompts.

4 THEORETICAL GUARANTEES

We characterize diagnostic risk, the value of targeted challenges, and the cost of continued deliberation. Proofs and calibration details are in Appendix B.

Lemma 1 (Energy components)**.** *If* $E_t := \varepsilon_{\text{diag}}^{(t)} \le \varepsilon$, *then* $D_{\text{sym}}(P_1^{(t)}, P_2^{(t)}) \le \varepsilon/\alpha$, $\sum_a R_L(\hat{y}_a^{(t)}; P_a^{(t)}) \le \varepsilon/\beta$, *and* $\Delta_L^{(t)} \ge \log(\gamma/\varepsilon)$ *when* $\varepsilon < \gamma$.

Proposition 1 (Value of a differential check)**.** *For* $L(d, y) = \ell_y \mathbf{1}\{d \ne y\}$, $\ell_y > 0$, *write* $h_y = \ell_y P_a^{(t)}(y)$. *A perfect check of whether* $Y = z$, *followed by a Bayes decision, reduces expected loss under* $P_a^{(t)}$ *by*
$$V_a(z) = \min\{h_z, \max_{y \ne z} h_y\}.$$

The target in equation 4 maximizes V_a, *for any current report. Under any* p *with* $\text{TV}(p, P_a^{(t)}) \le \eta$, *its value is within* $2L_{\max}\eta$ *of the optimal check.*

This identifies the information sought by steering; the agent's response determines how much of it is obtained.

Theorem 1 (Loss-sensitive diagnostic certification)**.** *Under Assumptions 1–4, the pooled model risk* $b_t := R_L(d_1^{(t)}; Q^{(t)})$ *satisfies*

$$b_t \le \min\{\Phi(E_t - \gamma e^{-\Delta_L^{(t)}}), G_L(\Delta_L^{(t)})\}, \tag{12}$$

where Φ *and* G_L *are the energy and loss-margin envelopes in Appendix B.5.4. For* M *frozen controllers under independent stratified risk calibration (Assumption 5), for* $\zeta \in (0,1)$*, with probability at least* $1-\zeta$*, simultaneously for positive decision coverage,*

$$R_{\mathrm{dec}}(\theta) \le \min\left\{L_{\max}, \frac{\widehat{\nu}_\theta + h_\nu}{\widehat{c}_\theta - h_c}\right\} \quad \text{if } \widehat{c}_\theta > h_c.$$

Here R_{dec} *is population loss conditional on a decision passing both gates;* $\widehat{\nu}_\theta, \widehat{c}_\theta$ *estimate its all-case loss contribution and coverage, with radii defined in Appendix B.5.5. Otherwise the bound is* $L_{\max}$*.*

The first bound controls per-round model risk; the second calibrates population risk after adaptive stopping.

Theorem 2 (Cost-aware stopping with trajectory variation)**.** *For any finite energy trajectory* $E_1, \ldots, E_H$*, let* $g_t = E_{t-1} - E_t$*,* $J_t = E_t + \delta t$*, and* $\tau_W = \min(\{t \in [W+1, H] : \widehat{g}^{(t)} < \delta\} \cup \{H\})$*, with integers* $H > W \ge 1$ *and* $\delta > 0$*. Define* $U_\tau = \sum_{s=2}^{\tau}(E_s - E_{s-1})_+$ *and* $V_H = \sum_{s=2}^{H-1}(g_{s+1} - g_s)_+$*. Then*

$$0 \le J_{\tau_W} - \min_{t \le H} J_t \le \max\{\delta W + U_{\tau_W}, (H - \tau_W)V_H\}.$$

Nonnegative diminishing gains recover the δW *bound.*

This comparison concerns the descent rule; delays from other controller gates are analyzed separately in the appendix.

Theorem 3 (Finite-time escalation)**.** *Suppose* $E_{T_0} > \varepsilon_{\mathrm{low}}$ *and* $A^{(t)} = 0$ *from* T_0 *onward. While* $E_t > \varepsilon_{\mathrm{low}}$*, the policy escalates or leaves this region by round*

$$T_0 + W\left(\left\lfloor \frac{E_{T_0} - \varepsilon_{\mathrm{low}}}{W\delta} \right\rfloor + 1\right).$$

If energy stays above $\varepsilon_{\mathrm{low}}$*, escalation occurs within this bound, provided deliberation remains available until that round.*

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, enabling large-scale sweeps of scenarios, compliance, and budgets. (ii) The main **LLM-agent setting** uses 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), each with a designated high-risk class subset defining the asymmetric loss L. Dataset and implementation details are 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.

Baselines. We compare with **Single best**, the calibration-best agent's one-shot prediction; **Free discussion**, unmonitored multi-round interaction; **Fixed-weight log pool**, D-CEM's pooling without mediator actions; and two static selective-prediction baselines, **Static-CM-ER** and **MaxProb-ER**, using CM-based expected risk and predictive confidence for rejection, respectively (Kerrigan et al., 2021; Geifman & El-Yaniv, 2017). The latter comparison tests whether sequential control improves the safety–cost tradeoff beyond static aggregation and rejection (Appendix G).

Metrics. We report pooled accuracy $\mathrm{Acc}_{\mathrm{pool}}$, loss-weighted pooled risk $\widehat{R}_L^{\mathrm{pool}}$ over all cases, high-risk miss rate within high-risk cases, and harmful-consensus rate (H.C.), the fraction of all cases with an incorrect, non-escalated decision on a high-risk truth. Mediator behavior is summarized by average rounds, certified-decision rate, and escalation rate. Clinical results also include autonomous system risk $\widehat{R}_L^{\mathrm{sys}}$ and escalation-inclusive cost $\widehat{J}_{\mathrm{esc}}(1) = \widehat{R}_L^{\mathrm{sys}} + \mathrm{Escalation}$, with unit cost per escalation. Formal definitions appear in Appendix E.

5.1 Deliberation Dynamics with Rule-Guided Agents

Illustrative trajectories. Figure 1 contrasts the two terminal actions of D-CEM. In Scenario 1 (*Example 1*), $\varepsilon_{\mathrm{diag}}^{(t)}$ declines as the mediator alternates CONTINUE and DIFFERENTIAL_STEER (targeting Agent 2); it issues STOP_AND_DECIDE at Round 7 when both energy and margin gates hold. In Scenario 3b (*Example 2*), the agents agree on a high-cost wrong class, but energy stalls above $\varepsilon_{\mathrm{low}}$ after a Round-5 challenge, prompting escalation at Round 8.

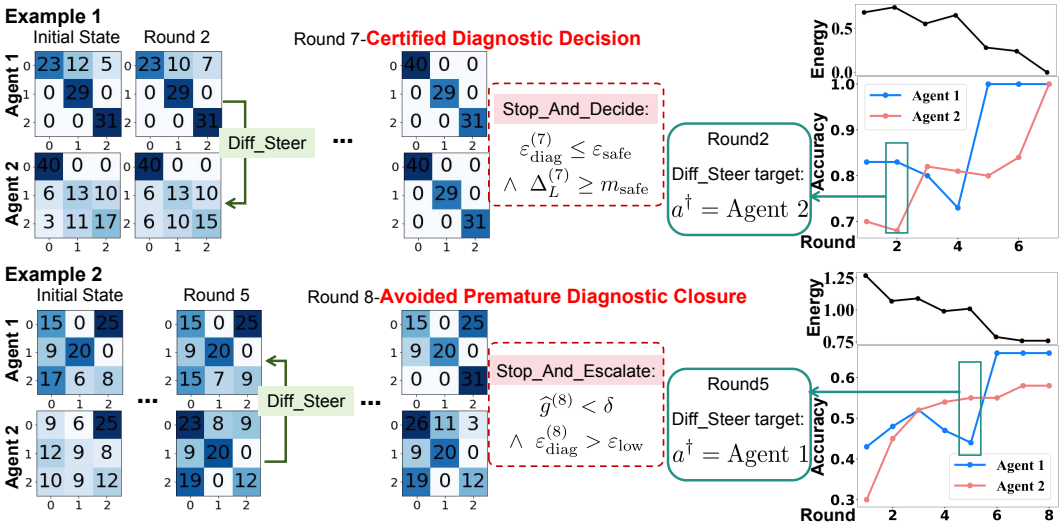

Figure 1: D-CEM trajectories in S1 (certification) and S3b (escalation), with all panels using the same cases within each example. Confusion matrices and accuracies are computed *ex post*; the mediator uses fixed C_a without current-case labels. At escalation, no actionable challenge remains.

Aggregate behaviour. Figure 2 summarizes test-set accuracy and diagnostic-risk energy across rounds. In *Ideal*, energy falls below $\varepsilon_{\mathrm{safe}}$ and most cases are certified. In *Asymmetric*, steering improves the weaker agent while preserving the stronger. In *Noisy-complementary*, slow but steady energy decline supports continued deliberation; in *Noisy-non-complementary*, energy plateaus above $\varepsilon_{\mathrm{low}}$ and the protocol escalates rather than committing to a shared high-risk error, avoiding unsafe consensus. The inverse accuracy–energy relation supports energy-based risk monitoring.

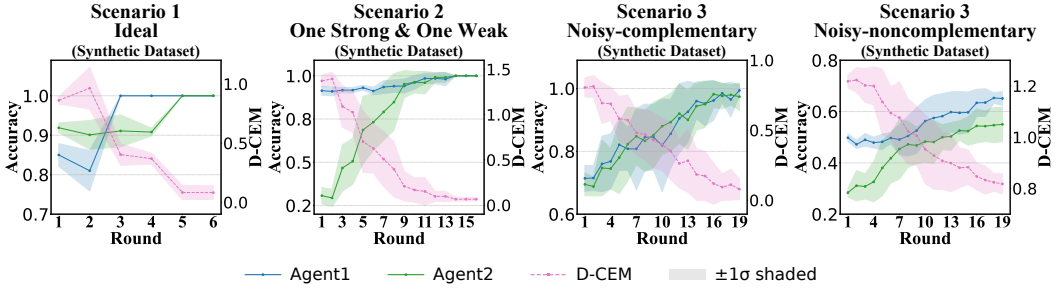

Figure 2: Per-round accuracy and diagnostic-risk energy $\varepsilon_{\mathrm{diag}}^{(t)}$ for rule-based agents across the four synthetic scenarios, aggregated over the test set.

5.2 Evaluation with Rule-Guided Agents

The scenario-wise comparison uses matched agents, calibration, splits, scenarios, and seeds, with Free discussion run for $T = 22$ rounds. Figure 3 shows that D-CEM achieves high accuracy and low pooled risk across all scenarios: accuracy gains are largest in the noisy regimes (S3a: 0.99 vs. 0.52–0.74; S3b: 0.98 vs. 0.31–0.73); $\widehat{R}_L^{\text{pool}}$ is lowest in every cell, tied with Single best and Fixed-weight log pool in S1, 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 (2.00–8.08 vs. 22), and escalates unresolved cases rather than forcing autonomous closure. Against Static-CM-ER and MaxProb-ER, D-CEM also achieves lower escalation-inclusive cost on Synthetic (Appendix G).

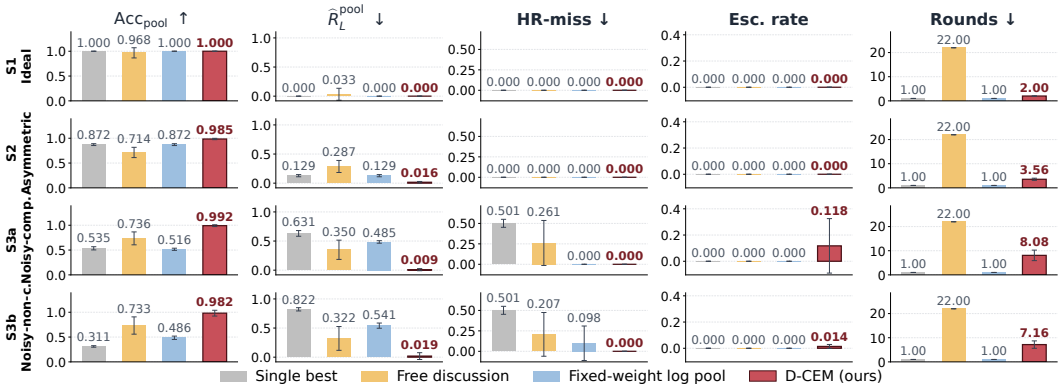

Figure 3: **Rule-guided comparison against aggregation baselines.** Rows: four scenarios. Columns: pooled accuracy (↑), expected diagnostic cost $\widehat{R}_L^{\text{pool}}$ (↓), high-risk miss rate (↓), escalation rate, and average rounds (↓). Definitions in Appendix E.

5.3 Evaluation with LLM Agents

We evaluate D-CEM on MedQA-Tri and DDXPlus across S1–S3b; Free discussion uses $T = 6$ rounds. D-CEM achieves lower risk and macro-averaged escalation-inclusive cost than the three aggregation baselines (Figure 4), with $\widehat{J}_{\text{esc}}(1) = 0.341$ on DDXPlus and 0.640 on MedQA-Tri. It costs less than Static-CM-ER and MaxProb-ER on both datasets; at matched 26% escalation on DDXPlus, it costs 0.341 versus Static-CM-ER's 0.389. Full results appear in Appendices F and G.

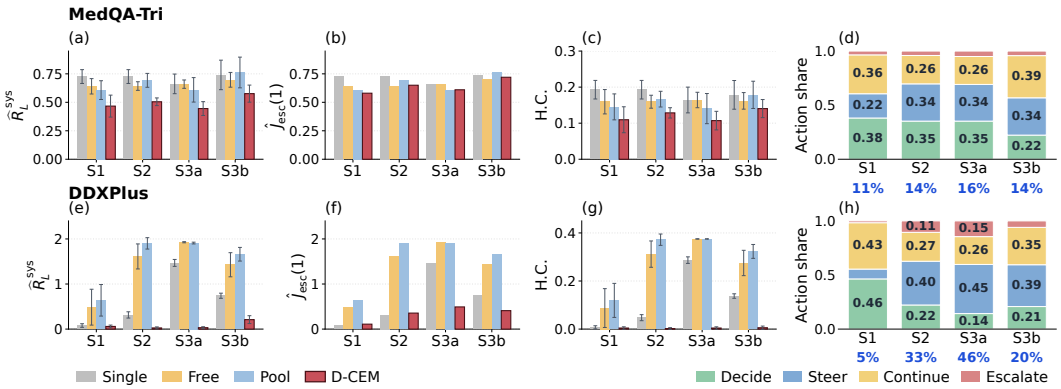

Figure 4: **Main results on MedQA-Tri (top) and DDXPlus (bottom).** **(a,e)** System risk $\widehat{R}_L^{\text{sys}}$ ↓. **(b,f)** Escalation-inclusive cost $\widehat{J}_{\text{esc}}(1)$ ↓. **(c,g)** Harmful-consensus rate H.C. ↓. **(d,h)** D-CEM action mix; bold blue numbers below each bar are per-case escalation rates.

We highlight three findings.

- **D-CEM is the only multi-agent protocol that does not collapse under asymmetric loss.** With $\ell_{\text{high}} = 5$ for high-severity misses, Pool and Free show sharply increased risk on DDXPlus, while D-CEM stays bounded across scenarios. D-CEM's gains over Single are smallest in the saturated S1, where Single already achieves $\text{Acc} = 0.95$, leaving little headroom for multi-agent fusion.
- **D-CEM achieves the lowest H.C. in every setting.** On DDXPlus, H.C. is near zero in all four scenarios, an order of magnitude below Pool, whose high-cost errors affect roughly a third of all cases in noisy settings. Reductions are smaller on MedQA-Tri, where high-risk classes are specialty-based and cost asymmetry is weaker.
- **The action mix tracks reliability, not a fixed schedule.** On DDXPlus, steering rises from 9% in S1 to 45% in noisy regimes, with escalation reaching 46%, concentrating intervention on harder cases. MedQA-Tri's flatter reliability profile yields 11–16% escalation. Average rounds remain below $T_{\max} = 6$ in every setting (Table 6, Appendix F), supporting cost-aware termination.

5.4 Ablation and Robustness Studies

Steering ablations. Rule-guided agents. 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 vary DIFFERENTIAL_STEER adoption probability $\xi \in \{25\%, 50\%, 75\%, 100\%\}$ and evaluate at $T \in \{4, 8, 12\}$ across all four synthetic scenarios. At $T = 12$, increasing uptake ($25\% \to 100\%$) improves Agent 2 accuracy in S2 ($0.62 \to 0.95$) and S3a ($0.69 \to 0.86$), where pooled risk falls from 0.20 to approximately zero, supporting stronger steering effects with greater uptake in correctable scenarios (Table 1). **LLM agents.** Matched intervention and target-selection ablations show the lowest mean autonomous risk for Full: about 20% below Generic review at modest additional cost, and lower risk and cost than Random (Appendix H.2).

Table 1: Fixed-horizon Synthetic partial-compliance ablation at $T = 12$ (mean $\pm$ std). Acc_2 is Agent 2 accuracy; full horizons appear in Appendix H.1.

	$\xi = 25\%$		$\xi = 50\%$		$\xi = 75\%$		$\xi = 100\%$	
Scenario	$\widehat{R}_L^{\text{pool}} \downarrow$	$\text{Acc}_2 \uparrow$	$\widehat{R}_L^{\text{pool}} \downarrow$	$\text{Acc}_2 \uparrow$	$\widehat{R}_L^{\text{pool}} \downarrow$	$\text{Acc}_2 \uparrow$	$\widehat{R}_L^{\text{pool}} \downarrow$	$\text{Acc}_2 \uparrow$
S1	0.05 ± 0.09	0.90 ± 0.10	$\mathbf{0.00 \pm 0.01}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{0.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{0.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$
S2	0.02 ± 0.04	0.62 ± 0.12	$\mathbf{0.00 \pm 0.00}$	0.77 ± 0.07	$\mathbf{0.00 \pm 0.00}$	0.88 ± 0.09	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.95 \pm 0.03}$
S3a	0.20 ± 0.16	0.69 ± 0.09	0.04 ± 0.07	0.78 ± 0.07	0.01 ± 0.02	0.82 ± 0.06	$\mathbf{0.00 \pm 0.01}$	$\mathbf{0.86 \pm 0.10}$
S3b	$\mathbf{0.66 \pm 0.04}$	$\mathbf{0.47 \pm 0.03}$	$\mathbf{0.66 \pm 0.04}$	$\mathbf{0.47 \pm 0.03}$	$\mathbf{0.66 \pm 0.04}$	$\mathbf{0.47 \pm 0.03}$	$\mathbf{0.66 \pm 0.04}$	$\mathbf{0.47 \pm 0.03}$

Energy, confusion matrices, and escalation. Matched LLM comparisons show risk and cost reductions for full energy versus Risk-only and full CMs versus Recall-only in closed-loop control. Escalation reduces autonomous risk and H.C.; referral-cost sensitivity quantifies the cost of safety gains (Appendices I.1–I.3).

Robustness and sensitivity. Prompt, label-space, and parameter variations assess robustness. D-CEM retains the lowest cost point estimates under neutral prompts at $K = 8, 16, 24$ and costs close to nominal under mild CM perturbations ($\eta \leq .25$; Appendices I.4–I.5).

6 Conclusion

We introduced Diagnostic Consensus Energy Minimization (D-CEM) for controlling diagnostic deliberation without online ground truth. D-CEM combines historical confusion patterns and asymmetric loss to identify dangerous plausible misses, using diagnostic-risk energy to guide continuation, targeted challenges, certification, and escalation. Our theory connects energy to model risk, provides population-risk guarantees through independent calibration, and bounds stopping costs. Across synthetic tasks and clinical LLM benchmarks, D-CEM reduces high-risk misses and harmful consensus while improving safety–cost trade-offs over aggregation baselines. These results support risk-aware deliberation that allocates discussion to consequential alternatives and escalates when high-risk uncertainty remains unresolved.

AI USE STATEMENT

In this work, we used generative AI tools to polish the manuscript by improving the clarity and readability of the writing. We also used these tools to help identify relevant literature and to support the planning and refinement of experimental designs. In addition, we used them for code debugging, including identifying potential errors and exploring possible fixes. We have reviewed all AI-assisted code, analyses, and manuscript text for accuracy and consistency. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

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

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021.

Erin P Balogh, Bryan T Miller, and John R Ball. Improving diagnosis in health care. 2015.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 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.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *arXiv preprint arXiv:2101.02703*, 2021.

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

Yuzhou Cao, Hussein Mozannar, Lei Feng, Hongxin Wei, and Bo An. In defense of softmax parametrization for calibrated and consistent learning to defer. *Advances in Neural Information Processing Systems*, 36:38485–38503, 2023.

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

Yu Cui, Hang Fu, Haibin Zhang, Licheng Wang, and Cong Zuo. Free-mad: Consensus-free multi-agent debate. *arXiv preprint arXiv:2509.11035*, 2025.

Alexander Philip Dawid and Allan M Skene. 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.

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

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first international conference on machine learning*, 2024.

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

Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. Ddxplus: A new dataset for automatic medical diagnosis. *Advances in neural information processing systems*, 35: 31306–31318, 2022.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

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

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

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

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

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

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.

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

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

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 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.

Shuai Ma, Qiaoyi Chen, Xinru Wang, Chengbo Zheng, Zhenhui Peng, Ming Yin, and Xiaojuan Ma. 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.

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

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. *Advances in neural information processing systems*, 34:15682–15694, 2021.

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

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

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

Mark Steyvers, Heliodoro Tejeda, Gavin Kerrigan, and Padhraic Smyth. Bayesian modeling of human–ai complementarity. *Proceedings of the National Academy of Sciences*, 119(11):e2111547119, 2022.

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

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

Yihan Wang, Qiao Yan, Zhenghao Xing, Lihao Liu, Junjun He, Chi-Wing Fu, Xiaowei Hu, and Pheng-Ann Heng. Silence is not consensus: Disrupting agreement bias in multi-agent llms via catfish agent for clinical decision making. *arXiv preprint arXiv:2505.21503*, 2025.

Zixi Wei, Yuzhou Cao, and Lei Feng. 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 D-CEM.

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 $\varepsilon_{\mathrm{diag}}^{(t)}$ 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.

D-CEM treats agents as black boxes and supports natural-language reasoning and multi-turn deliberation. At each round, reports to the mediator must map to a finite decision space with a defined task loss. The confusion-matrix requirement therefore concerns the task interface rather than the agents' internal architecture.

Finite-budget execution. A finite-budget controller fixes $T_{\max}$ and a finite challenge allowance per receiver–alternative pair for each episode. A challenge is actionable only while its risk gate holds and its allowance is unexhausted; each issuance uses one attempt. These counters depend only on the observed history. Execution stops at $\tau = \min\{\tau^*, T_{\max}\}$. At the budget endpoint, the certification gates are checked first: if both hold, the mediator decides; otherwise it escalates, issuing no further challenge. The supplementary LLM steering and risk-calibration studies (Appendices H.2 and I.6) use $W = 2$, $T_{\max} = 6$, and two attempts per receiver–alternative pair.

B.2 Technical Conditions and Calibration

Assumption 1 (Finite diagnostic task)**.** *The label and decision sets are $\mathcal{Y} = \mathcal{D} = \{1, \dots, K\}$, with $2 \le K < \infty$.*

Assumption 2 (Positive model support)**.** *The prior has full support, and smoothing gives $\pi(y) \ge \underline{\pi} > 0$ and $C_a[i,j] \ge \underline{C} > 0$. Consequently $P_a^{(t)}(y) \ge \underline{p} > 0$ and all KL terms are finite. These conditions concern the estimated models used by the mediator.*

Assumption 3 (Bounded diagnostic loss)**.** $0 \le L(d,y) \le L_{\max} < \infty$ *for every d, y. The energy weights satisfy $\alpha, \beta, \gamma > 0$.*

Assumption 4 (Pooling weights)**.** *The weights satisfy $\omega_a^{(t)} \ge 0$ and $\sum_a \omega_a^{(t)} = 1$. Reliability flooring uses $\bar{\rho}_a^{(t)} = \max\{\rho_a^{(t)}, \rho_{\min}\}$ with $\rho_{\min} \in [0,1)$. Optional clipping to $[\underline{\omega}, 1-\underline{\omega}]$ improves numerical conditioning; the pooling bounds also hold at weights zero and one.*

Assumption 5 (Independent stratified risk calibration)**.** *Conditional on fitting and development data, a family Θ of $M \ge 1$ complete controllers is fixed, including CMs, priors, agent mechanisms, thresholds, and budget handling. A separate risk-calibration set contains $n_y > 0$ independent case trajectories from each target class y, with pre-specified target class weights $p_y \ge 0$ summing to one. Each controller is run with its own interventions, under the same conditional trajectory law as at deployment. Cases are independent across classes and of fitting data; agents and rounds within a case may be dependent. Each episode terminates without exposing its label to the mediator.*

The model assumptions specify the online calculations. Independent stratified risk calibration supports the population-risk part of Theorem 1. The experimental CAL sets estimate CMs; independent risk calibration is a separate statistical step. Its unit is a case trajectory, so repeated rounds or repeated uses of the same case do not increase the independent sample count.

Energy trajectories. Theorem 2 uses the realized gains $g_t = E_{t-1} - E_t$ and their upward variation. It permits energy increases and delayed gains; nonnegative, nonincreasing gains give its zero-variation specialization. Theorem 3 uses the absence of actionable challenges and the high-energy threshold, without a convergence assumption.

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

$$D_{\exp}(w_1, w_2) = D_{\mathrm{KL}}(\sigma(w_1)\|\sigma(w_2)) + D_{\mathrm{KL}}(\sigma(w_2)\|\sigma(w_1)). \tag{13}$$

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

$$w_{\mathrm{rec}}^{(t)} \in \arg\min_{w \in \mathcal{W}} \left\{ \frac{\lambda_{\mathrm{st}}}{2} \|w - w_{a^\dagger}^{(t)}\|_2^2 + \gamma_{\mathrm{st}} D_{\exp}(w, w_{\mathrm{ref}}^{(t)}) \right\}. \tag{14}$$

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:

$$w_{\mathrm{rec}}^{(t)} = \Pi_{\mathcal{W}} \left(w_{a^\dagger}^{(t)} - \eta\gamma_{\mathrm{st}} \, \nabla_w D_{\exp}(w, w_{\mathrm{ref}}^{(t)}) \Big|_{w = w_{a^\dagger}^{(t)}} \right). \tag{15}$$

At this initialization the quadratic penalty has zero gradient; the step size and projection control the one-step change. The set $S^{(t)}$ contains the top-k coordinates with largest absolute change:

$$S^{(t)} = \mathrm{TopK}\left(\left| w_{\mathrm{rec}}^{(t)} - w_{a^\dagger}^{(t)} \right|, k \right). \tag{16}$$

This feature-level construction is only one implementation of the steering mechanism 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 $\xi^{(t)} \in [0, 1]$:

$$w_{a^\dagger}^{(t+1)} = \Pi_{\mathcal{W}} \left((1 - \xi^{(t)}) w_{a^\dagger}^{(t)} + \xi^{(t)} w_{\mathrm{rec}}^{(t)} \right). \tag{17}$$

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 $\xi^{(t)}$.

Local effect of partial adoption. For a fixed reference, let $f(w) = D_{\exp}(w, w_{\mathrm{ref}})$ have M_f-Lipschitz gradient on a closed convex feasible region $\mathcal{W}$, with $M_f > 0$. For $w \in \mathcal{W}$, set $h = \eta\gamma_{\mathrm{st}} > 0$ and $z = \Pi_{\mathcal{W}}(w - h\nabla f(w))$. The adopted point is $w^+ = w + \xi(z - w) \in \mathcal{W}$. Projection optimality gives $\langle \nabla f(w), z - w \rangle \le -\|z - w\|^2/h$; the descent lemma yields

$$f(w^+) \le f(w) - \xi \left(\frac{1}{h} - \frac{M_f \xi}{2} \right) \|z - w\|^2.$$

Thus $h \le 1/M_f$ ensures nonincrease for every $\xi \in [0, 1]$. This controls explanation discrepancy for the fixed reference; diagnostic energy changes when the induced hard reports change.

B.5 Proofs and Risk Envelopes

B.5.1 Proof of Lemma 1

Each term of equation 9 is nonnegative. Dividing the first two terms by their weights gives the disagreement and report-risk bounds. The inequality $\gamma e^{-\Delta_L^{(t)}} \le \varepsilon$ gives $\Delta_L^{(t)} \ge \log(\gamma/\varepsilon)$. □

B.5.2 Proof of Proposition 1

Write $q = P_a^{(t)}$, $h_y = \ell_y q_y$, and $S = \sum_y h_y$. The Bayes risk under the stated loss is $S - \max_y h_y$. A perfect check of $Y = z$ has zero loss on its positive outcome; on its negative outcome, the unnormalized Bayes risk is $S - h_z - \max_{y \ne z} h_y$. Subtracting gives

$$V_q(z) = h_z + \max_{y \ne z} h_y - \max_y h_y = \min\{h_z, \max_{y \ne z} h_y\}.$$

Its maximum is the second-largest component of h. For any current report d, a maximizer of h_z over $z \neq d$ is a largest or second-largest component, including ties, and attains this value.

If $\mathrm{TV}(p, q) \leq \eta$, then $|\ell_y p_y - \ell_y q_y| \leq L_{\max}\eta$ for every y. The min and max operations are 1-Lipschitz in the sup norm, hence $|V_p(z) - V_q(z)| \leq L_{\max}\eta$ uniformly in z. For $z_q \in \arg\max_z V_q(z)$ and $z_p \in \arg\max_z V_p(z)$,

$$V_p(z_p) - V_p(z_q) \leq [V_p(z_p) - V_q(z_p)] + [V_q(z_q) - V_p(z_q)] \leq 2L_{\max}\eta.$$

This proves the target-value bound. □

The proposition is a decision-theoretic model of the information requested by a challenge (Howard, 1966). If every check reveals the binary fact with the same probability s, independently of Y, and otherwise returns no information, all values are multiplied by s. Target-dependent success probabilities instead give $s_z V_q(z)$. The feature-adoption parameter ξ describes agent behavior, not this observation channel.

B.5.3 Log-pool identities

Proposition 2 (Variational pooling and within-round contraction)**.** *For positive P_1, P_2 and $\omega \in [0, 1]$, $Q \propto P_1^{\omega} P_2^{1-\omega}$ uniquely minimizes $\omega\mathrm{KL}(Q\|P_1) + (1-\omega)\mathrm{KL}(Q\|P_2)$ and satisfies*

$$\mathrm{KL}(P_1\|Q) \leq (1-\omega)\mathrm{KL}(P_1\|P_2), \qquad \mathrm{KL}(P_2\|Q) \leq \omega\mathrm{KL}(P_2\|P_1).$$

These are standard log-pool properties (Genest & Zidek, 1986). Expanding the objective gives $\mathrm{KL}(Q\|Q_\omega) - \log Z$, where $Q_\omega(y) = P_1(y)^{\omega} P_2(y)^{1-\omega}/Z$; uniqueness follows. Also

$$\mathrm{KL}(P_1\|Q_\omega) = (1-\omega)\mathrm{KL}(P_1\|P_2) + \log Z.$$

Hölder's inequality gives $Z \leq 1$; interchanging the agents proves the second bound. The endpoint weights follow directly. □

B.5.4 Proof of Theorem 1: energy and margin

Suppress t and put $r_a = R_L(\hat{y}_a; P_a)$, $S = r_1 + r_2$, $D = D_{\mathrm{sym}}(P_1, P_2)$, and $e = E - \gamma e^{-\Delta_L} = \alpha D + \beta S$. Bayes optimality and bounded-loss stability imply $b \leq r_a + L_{\max}\,\mathrm{TV}(P_a, Q)$ for each a. Proposition 2, Pinsker's inequality, and Cauchy–Schwarz give

$$\begin{aligned}\mathrm{TV}(P_1, Q) + \mathrm{TV}(P_2, Q) &\leq \sqrt{(1-\omega)\mathrm{KL}(P_1\|P_2)/2} + \sqrt{\omega\mathrm{KL}(P_2\|P_1)/2} \\ &\leq \sqrt{D/2}.\end{aligned}$$

Averaging the two risk inequalities yields

$$b \leq \min\{L_{\max}, S/2 + L_{\max}\sqrt{D/8}\}. \tag{18}$$

Maximizing $(e - \alpha D)/(2\beta) + L_{\max}\sqrt{D/8}$ over $0 \leq D \leq e/\alpha$ gives

$$\Phi(e) = \min\left\{L_{\max}, \begin{cases} L_{\max}\sqrt{e/(8\alpha)}, & e \leq \beta^2 L_{\max}^2/(8\alpha), \\ e/(2\beta) + \beta L_{\max}^2/(16\alpha), & e > \beta^2 L_{\max}^2/(8\alpha). \end{cases}\right\}. \tag{19}$$

The maximizing D is $\min\{e/\alpha, \beta^2 L_{\max}^2/(8\alpha^2)\}$.

For a feasible margin m, define the loss-margin envelope

$$G_L(m) = \max_{d,\, q \in \Delta(\mathcal{Y})} \{R_L(d; q) : R_L(k; q) - R_L(d; q) \geq m \text{ for all } k \neq d\}. \tag{20}$$

Here $\Delta(\mathcal{Y})$ is the probability simplex. This is the maximum of K linear programs and is nonincreasing in m. The actual pooled decision is feasible at $m = \Delta_L$, so $b \leq G_L(\Delta_L)$. Together with equation 19, this proves the model-risk bound. □

Both gates therefore imply the model-risk bound

$$b_\tau \leq B_L(\varepsilon_{\mathrm{safe}}, m_{\mathrm{safe}}), \qquad B_L(\varepsilon, m_0) = \sup_{\substack{m \in [m_0, L_{\max}] \text{ feasible} \\ \varepsilon \geq \gamma e^{-m}}} \min\{\Phi(\varepsilon - \gamma e^{-m}), G_L(m)\}. \tag{21}$$

An empty feasible set means the gates admit no decision. This envelope uses the margin term as well as disagreement and report risk.

Decision stability under posterior error. Let d minimize risk under Q, let $\Delta_k = R_L(k;Q) - R_L(d;Q)$, and let $s_k = \max_y[L(k,y) - L(d,y)] - \min_y[L(k,y) - L(d,y)]$. If $\mathrm{TV}(p,Q) \le \eta$, then

$$R_L(d;p) - \min_k R_L(k;p) \le \max\{0, \max_{k\neq d}(s_k\eta - \Delta_k)\}.$$

Indeed the expectation of $L(d,Y) - L(k,Y)$ changes by at most $s_k\eta$. For $L(d,y) = \ell_y \mathbf{1}\{d \neq y\}$, $s_k = \ell_d + \ell_k$; the decision remains Bayes-optimal whenever $\eta \le \min_{k\neq d} \Delta_k/(\ell_d + \ell_k)$.

B.5.5 Proof of Theorem 1: population risk

For controller θ, let $I_\theta = 1$ indicate a terminal decision passing both energy and margin gates, and zero otherwise. Write $c_\theta = \Pr(I_\theta = 1)$, $\nu_\theta = \mathbb{E}[I_\theta L(d_\tau, Y)]$, and $R_{\mathrm{dec}}(\theta) = \nu_\theta/c_\theta$ when $c_\theta > 0$. This conditional risk and its all-case contribution ν_θ have different denominators. On class y's risk-calibration cases, set

$$\widehat{\nu}_\theta = \sum_y \frac{p_y}{n_y} \sum_{i=1}^{n_y} I_{yi,\theta} L_{yi,\theta}, \qquad \widehat{c}_\theta = \sum_y \frac{p_y}{n_y} \sum_{i=1}^{n_y} I_{yi,\theta}.$$

The loss contribution is zero when $I = 0$. With $\ell_y = \max_d L(d,y)$ and $q_\zeta = \log(3M/\zeta)/2$, define

$$h_\nu = \sqrt{q_\zeta \sum_y p_y^2 \ell_y^2/n_y}, \qquad h_c = \sqrt{q_\zeta \sum_y p_y^2/n_y}. \tag{22}$$

Weighted Hoeffding bounds (Hoeffding, 1963) give $\nu_\theta \le \widehat{\nu}_\theta + h_\nu$ and $c_\theta \ge \widehat{c}_\theta - h_c$, each with failure probability at most $\zeta/(3M)$. A union bound across controllers makes these inequalities simultaneous. Dividing when the denominator is positive proves the stated risk bound; bounded loss gives $L_{\max}$ otherwise. □

Combining the energy envelope with calibration. Let $w_\theta = \mathbb{E}[I_\theta(L(d_\tau,Y) - b_\tau)]$ and estimate it by $\widehat{w}_\theta = \sum_y (p_y/n_y) \sum_i I_{yi,\theta}(L_{yi,\theta} - b_{yi,\theta})$. Each class-$y$ residual lies in $[-L_{\max}, \ell_y]$, so the remaining failure probability yields the simultaneous bound

$$w_\theta \le \widehat{w}_\theta + h_w, \qquad h_w = \sqrt{q_\zeta \sum_y p_y^2 (L_{\max} + \ell_y)^2/n_y}.$$

Using equation 21, for $\underline{c}_\theta = \widehat{c}_\theta - h_c > 0$ we obtain

$$R_{\mathrm{dec}}(\theta) \le U_\theta := \min\left\{ L_{\max}, \frac{\widehat{\nu}_\theta + h_\nu}{\underline{c}_\theta}, B_{L,\theta} + \frac{(\widehat{w}_\theta + h_w)_+}{\underline{c}_\theta} \right\}. \tag{23}$$

Set $U_\theta = L_{\max}$ when $\underline{c}_\theta \le 0$. Here $B_{L,\theta}$ uses controller θ's gates and energy weights. For a target $r \in [0, L_{\max})$, with probability at least $1-\zeta$, any controller selected from $\{\theta : U_\theta \le r\}$ has conditional decision risk at most r; an empty set provides no qualifying controller. This is held-out risk control over a finite family (Bates et al., 2021; Angelopoulos et al., 2021), with the D-CEM energy envelope supplying a second bound through the selected residual. Arbitrary dependence between agents and rounds is absorbed into each case's terminal outcome.

B.5.6 Proof of Theorem 2

All times in the theorem are integers in $\{1, \dots, H\}$. Write $\tau = \tau_W$ and $(x)_+ = \max\{x, 0\}$. For $t < \tau$, partition $(t, \tau]$ into full blocks of length W, ending strictly before τ, and a final block of length $r \in \{1, \dots, W\}$. Every full block has average gain at least δ, since the rule has not stopped. Its contribution to J is nonpositive. The final block contributes at most δr plus its energy increases, giving

$$J_\tau - J_t \le \delta W + U_\tau.$$

If $t > \tau$, then $\tau < H$ and the triggering window has average gain below δ. For every future s and every j in that window, $g_s - g_j \le V_H$. Averaging over j gives $g_s < \delta + V_H$, hence

$$J_\tau - J_t = \sum_{s=\tau+1}^{t} (g_s - \delta) \le (H - \tau) V_H.$$

Combining the comparisons proves the theorem. Nonnegative, nonincreasing gains give $U_\tau = V_H = 0$. □

Local comparison and controller delays. The past comparison can be sharpened to the identity

$$J_\tau - \min_{t\le\tau} J_t = \max_{0\le r\le \min(W,\tau-1)} (J_\tau - J_{\tau-r}).$$

Any earlier comparison point can be moved forward by W rounds without increasing J, until it lies in the final window. The horizon H defines the comparison trajectory; V_H quantifies possible later gains. For a full controller stopping at $T \ge \tau$ on the same continuation trajectory, add $\delta(T-\tau) + \sum_{s=\tau+1}^{T}(E_s - E_{s-1})_+$ to the theorem's bound. Thus delayed termination is accounted for explicitly. The objective $J_t = E_t + \delta t$ measures energy and communication, with δ in energy units; the system objective equation 1 also charges diagnostic loss and escalation.

B.5.7 Proof of Theorem 3

If $E_{T_0} \le \varepsilon_{\text{low}}$, the high-energy region has already been left. Otherwise consider endpoints $T_0 + kW$. Before either escalation or exit, $A^{(t)} = 0$ and $E_t > \varepsilon_{\text{low}} > \varepsilon_{\text{safe}}$. At every endpoint not triggering escalation,

$$E_{T_0+(k-1)W} - E_{T_0+kW} \ge W\delta.$$

Summing gives $E_{T_0+kW} \le E_{T_0} - kW\delta$. For $k = \lfloor (E_{T_0} - \varepsilon_{\text{low}})/(W\delta) \rfloor + 1$, this contradicts continued residence in the high-energy region. The policy must therefore escalate or leave that region by the stated round. No monotonicity between endpoints is required. □

B.6 Confusion-Matrix Estimation and Posterior Sensitivity

For a fixed agent interface and class prior π, let C_a^0 denote the population CM and $\widetilde{C}_a$ its empirical estimate from n_y independent cases in each truth class y. For N agents, Hoeffding's inequality and a union bound give, with probability at least $1-\zeta$,

$$|\widetilde{C}_a[y,j] - C_a^0[y,j]| \le r_y, \qquad r_y = \sqrt{\frac{\log(2NK^2/\zeta)}{2n_y}} \quad \text{for all } a, y, j.$$

Agent predictions on the same case may be correlated. If a prior row $b_{a,y}$ with total pseudo-count s_y is used for smoothing, then

$$\widehat{C}_a[y,j] = \frac{n_y\widetilde{C}_a[y,j] + s_y b_{a,y}(j)}{n_y + s_y}, \qquad |\widehat{C}_a[y,j] - C_a^0[y,j]| \le e_y := \min\left\{1, \frac{n_y r_y + s_y}{n_y + s_y}\right\}.$$

Thus smoothing supplies positive support while its estimation bias is included in the bound. For report j with positive population probability, put $\widehat{Z}_{a,j} = \sum_y \pi_y \widehat{C}_a[y,j]$. Posterior normalization gives

$$\mathrm{TV}(\widehat{P}_a(\cdot \mid j), P_a^0(\cdot \mid j)) \le \min\left\{1, \frac{\sum_y \pi_y e_y}{\widehat{Z}_{a,j}}\right\}. \tag{24}$$

Indeed, for nonnegative unnormalized vectors u, v with positive sums, $\|u/\sum u - v/\sum v\|_1 \le 2\|u-v\|_1/\sum v$. Rare reports amplify row-estimation error through the denominator. An estimated class prior contributes its own error to $u - v$.

Historical estimation and interaction shift enter separately:

$$\mathrm{TV}(\Pr(Y=\cdot \mid \mathcal{F}_t), \widehat{P}_a^{(t)}) \le \mathrm{TV}(\Pr(Y=\cdot \mid \mathcal{F}_t), P_a^0(\cdot \mid \hat{y}_a^{(t)})) + \mathrm{TV}(P_a^0(\cdot \mid \hat{y}_a^{(t)}), \widehat{P}_a^{(t)}).$$

The second term is controlled by equation 24; the first captures the change from the historical report model to the current interaction history. Independent closed-loop risk calibration in Assumption 5 evaluates terminal decisions directly.

B.7 Extension to More than Two Agents

For $N > 2$, each agent retains its fixed CM and induced posterior. The direct extension of equation 9 is

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

with $Q \propto \prod_a P_a^{\omega_a}$, nonnegative weights summing to one, and the pooled loss margin. Steering targets $a^\dagger \in \arg\max_a R_L(\hat{y}_a; P_a)$ and its dangerous alternative. The action semantics are unchanged; energy thresholds scale with the unnormalized sums.

For the analysis, put $\overline{D} = \sum_{a,b} \omega_a \omega_b \mathrm{KL}(P_a \| P_b)$ and $\overline{r} = \sum_a \omega_a R_L(\hat{y}_a; P_a)$. Direct expansion gives

$$\overline{D} = \sum_a \omega_a \{\mathrm{KL}(P_a \| Q) + \mathrm{KL}(Q \| P_a)\}, \qquad R_L(d_Q; Q) \le \overline{r} + L_{\max}\sqrt{\overline{D}/2}.$$

For the inequality, average $R_L(d_Q; Q) \le R_L(\hat{y}_a; P_a) + L_{\max}\,\mathrm{TV}(P_a, Q)$ with weights ω_a, then apply Pinsker and Jensen. No independence between agents is used. The pairwise quantity has the exact $O(NK)$ form

$$\overline{D} = \sum_y \left[\sum_a \omega_a P_a(y) \log P_a(y) - \Big(\sum_a \omega_a P_a(y)\Big)\Big(\sum_a \omega_a \log P_a(y)\Big) \right].$$

With uniform weights, $\sum_{a<b} D_{\mathrm{sym}}(P_a, P_b) = N^2 \overline{D}$, so the original extension also admits linear-time disagreement evaluation. Population-risk calibration applies to a fixed N-agent controller in the same way as in Theorem 1. Empirical evaluation of larger panels remains future work.

B.8 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 $\varepsilon_{\mathrm{safe}}$ and open a margin $\Delta_L \ge m_{\mathrm{safe}}$, 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, 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 $\varepsilon_{\mathrm{low}}$ 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 generator allocates samples to the six ground-truth rules in Table 2 in proportions

$$q_r = \frac{\exp(w_r)}{\sum_{j=1}^{6} \exp(w_j)}, \qquad r = 1, \ldots, 6,$$

with integer counts obtained by rounding and adjusting to the desired total. For each selected rule r, binary features are drawn at random and its literals are enforced. A sample is retained only when that rule is the unique satisfied ground-truth rule; its consequent supplies the label. Class k_0 is governed by longer rules and is designated high-cost in L. 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. The agent reports a highest-scoring class, equivalently the MAP label of the softmax distribution, as specified in Appendix B.8.

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.

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). Each retained item is assigned one of the seven specialty labels offline using a keyword heuristic over the stem and canonical answer text, with an LLM fallback (`gpt-4o-mini`) for items the heuristic skips. Canonical answer text is used only for offline label construction. During both calibration (CAL) and evaluation (TEST), the prediction agents receive only the question stem as the case input; neither the canonical answer text nor the ground-truth specialty label is included. Per (scenario, seed) we draw a calibration split of $|\mathrm{CAL}| = 70$

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

and a test split of $|\mathrm{TEST}| = 140$, both class-stratified; results are reported across 10 seeds. The high-risk subset $\mathcal{H}_{\mathrm{MedQA}} = \{\text{Cardiology, Pulmonology, Neurology, Infectious_Disease}\}$ – the four specialties whose missed call most plausibly corresponds to a life-threatening miss – is assigned $\ell_{\mathrm{high}} = 3$ 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 $\mathcal{H}_{\mathrm{DDX}} = \{\text{Anaphylaxis, Pulmonary_embolism, NSTEMI_STEMI, Myocarditis}\}$ is composed of severity-1–2 acute pathologies; the benign subset is $\{\text{URTI, Viral_pharyngitis, GERD, Bronchitis}\}$. 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:

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

where $\ell_{\text{high}} > 1$ is a tunable scalar that controls how aggressively the cost-aware decision rule defends the high-cost class. Larger values place greater weight on avoiding high-cost misses; 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 (long-rule class)	3 (S1, S2) 2 (S3a) 1.8 (S3b)
MedQA-USMLE	Cardiology, Pulmonology, Neurology, Infectious_Disease	3
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 $\mathcal{H}_{\text{test}} = \{i : y_i \in \mathcal{H}\}$ the indices of test cases with high-cost ground truth.

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

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

$$\widehat{J}_{\text{esc}}(c_{\text{esc}}) = \widehat{R}_L^{\text{sys}} + c_{\text{esc}}\, \text{Escalation} \qquad \text{(escalation-inclusive cost)},$$

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

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

$$\text{HarmfulConsensus} = \tfrac{1}{N} \sum_i \mathbf{1}\big[d_1^{(\tau_i)} \neq y_i \wedge y_i \in \mathcal{H} \wedge u^{(\tau_i)} \neq \text{ESC}\big] \quad \text{(harmful-consensus rate)},$$

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

$$\text{Escalation} = \tfrac{1}{N} \sum_i \mathbf{1}\big[u^{(\tau_i)} = \text{ESC}\big] \qquad \text{(escalation rate)}.$$

$\widehat{R}_L^{\text{pool}}$ evaluates the terminal pooled decision on every case, including escalated cases. In contrast, $\widehat{R}_L^{\text{sys}}$ counts only autonomous diagnostic loss and assigns zero diagnostic loss to escalation under the ideal expert-backstop convention; the high-risk miss rate isolates high-cost truths that are misclassified; the harmful-consensus rate measures wrong autonomous commitments on high-cost truths,

whether or not the agents agree; certified and escalation rates summarise the terminal mediator actions.

Here c_{esc} is the cost of one escalation, applied equally to every method. For the clinical experiments, we use $c_{\text{esc}} = 1$ as the primary reference. Our loss uses the normalized scales $0/1/3$ for MedQA-Tri and $0/1/5$ for DDXPlus: zero for a correct decision, one for an ordinary error, and three or five for a high-risk miss, respectively. Thus, $c_{\text{esc}} = 1$ values one escalation as one ordinary error and $1/3$ or $1/5$ of a high-risk miss, respectively. It provides an interpretable, non-zero reference.

F 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.

F.1 IMPLEMENTATION DETAILS

Datasets. We use the two clinical datasets specified in Appendix D.2: *MedQA-Tri* (7-way specialty triage, $\ell_{\text{high}} = 3$) 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 J): *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.

Confusion-matrix estimation. We use class-stratified CAL sets with at least 10 labeled cases per true label. Each CM row estimates the agent's class-conditional error profile, $\Pr(\hat{Y}_a = \cdot \mid Y = i)$, using diagonal-biased pseudocount smoothing, equivalent to a posterior-mean estimate under a Dirichlet prior. Stratification represents imbalanced and high-cost classes, while smoothing regularizes sparse counts and prevents zero cells, capturing dominant confusions without requiring precise estimates of every rare cell. The reported results remain stable under moderate CM perturbations; larger label spaces call for more per-class evidence.

F.2 PER-METHOD HEADLINE NUMBERS

We report $\widehat{J}_{\text{esc}}(1)$ as point estimates. Error bars in Figure 4 show standard deviations across seeds for risk and H.C. only.

Table 5 reports system risk $\widehat{R}_L^{\text{sys}}$, escalation-inclusive cost $\widehat{J}_{\text{esc}}(1)$, 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 $\text{Acc} = 0.95$, leaving limited headroom for further accuracy gains).

F.3 D-CEM BEHAVIOR SUMMARY

The takeaway from Table 6 is that **D-CEM's escalation rate and round usage both adapt to per-scenario agent reliability rather than firing on a fixed schedule.**

Table 5: 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 all metrics. Risk and H.C. are mean ± std.

		$\widehat{R}_L^{\text{sys}}\downarrow$				$\widehat{J}_{\text{esc}}(1)\downarrow$				H.C.↓			
Dataset	Method	S1	S2	S3a	S3b	S1	S2	S3a	S3b	S1	S2	S3a	S3b
MedQA-Tri	SINGLE	0.73±.06	0.73±.06	0.66±.09	0.74±.13	0.73	0.73	0.66	0.74	0.19±.03	0.19±.03	0.16±.04	0.18±.04
	FREE	0.64±.07	0.64±.04	0.66±.04	0.70±.07	0.64	**0.64**	0.66	**0.70**	0.16±.03	0.16±.02	0.16±.02	0.16±.02
	POOL	0.61±.08	0.69±.06	0.61±.11	0.76±.14	0.61	0.69	**0.61**	0.76	0.15±.04	0.17±.02	0.14±.04	0.18±.04
	D-CEM	**0.47**±.10	**0.51**±.03	**0.45**±.06	**0.58**±.08	**0.58**	0.65	**0.61**	0.72	**0.11**±.04	**0.13**±.01	**0.11**±.03	**0.14**±.03
DDXPlus	SINGLE	0.08±.04	0.31±.07	1.47±.08	0.74±.06	**0.08**	**0.31**	1.47	0.74	0.01±.01	0.05±.01	0.29±.01	0.14±.01
	FREE	0.49±.40	1.61±.28	1.93±.01	1.43±.27	0.49	1.61	1.93	1.43	0.09±.08	0.31±.06	0.38±.00	0.28±.05
	POOL	0.64±.35	1.90±.13	1.91±.02	1.66±.15	0.64	1.90	1.91	1.66	0.12±.07	0.37±.02	0.38±.00	0.32±.03
	D-CEM	**0.06**±.03	**0.03**±.02	**0.03**±.02	**0.21**±.09	0.11	0.36	**0.49**	**0.41**	**0.00**±.01	**0.00**±.00	**0.00**±.01	**0.01**±.01

- **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 $9\times$ ($5\% \to 46\%$) and average rounds rising from 2.1 to 3.6: 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, reducing deliberation relative to the fixed six-round protocol.

Table 6: 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±.03	0.14±.02	0.16±.02	0.14±.03	2.63±.28	2.82±.33	2.73±.10	3.85±.38
DDXPlus	0.05±.02	0.33±.08	0.46±.05	0.20±.06	2.13±.17	3.28±.26	3.59±.29	3.82±.19

G COMPARISON WITH STATIC SELECTIVE-PREDICTION BASELINES

This comparison tests whether D-CEM's sequential control improves the safety–cost tradeoff beyond static aggregation and rejection. **Static-CM-ER**, inspired by Kerrigan et al. (2021), performs CM-based expected-risk aggregation using the same confusion matrices, prior, and loss as D-CEM, without sequential control. **MaxProb-ER** uses a CAL-selected threshold on $1 - \max_k p(k)$ for confidence-based rejection (Geifman & El-Yaniv, 2017); its rejection score uses neither class-specific confusion matrices nor asymmetric loss.

Table 7 reports escalation-inclusive cost $\widehat{J}_{\text{esc}}(1)$, harmful-consensus rate, and accuracy. D-CEM achieves lower cost than both baselines on all three datasets. On DDXPlus, Static-CM-ER is matched to D-CEM's realized test escalation rate of 26%, with costs of 0.389 and 0.341, respectively.

H DIFFERENTIAL-STEERING ABLATIONS

The rule-guided study controls suggestion uptake; the LLM study separates intervention, additional reasoning, and alternative selection.

H.1 RULE-GUIDED STEERING: 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 $\xi \in [0, 1]$:

$$w_{a^\dagger}^{(t+1)} = \Pi_{\mathcal{W}} \left(\begin{cases} w_{\text{rec}}^{(t)} & \text{with probability } \xi, \\ w_{a^\dagger}^{(t)} & \text{with probability } 1 - \xi. \end{cases} \right). \tag{25}$$

Table 7: Comparison with static selective-prediction baselines: dataset-level summaries (mean ± std).

Dataset	Method	$\widehat{J}_{\mathrm{esc}}(1)\downarrow$	H.C.↓	$\mathrm{Acc}_{\mathrm{pool}}\uparrow$
Synthetic	Static-CM-ER	0.298 ± 0.032	0.008 ± 0.018	0.718 ± 0.011
	MaxProb-ER	0.332 ± 0.015	0.008 ± 0.018	0.675 ± 0.015
	D-CEM	$\mathbf{0.043 \pm 0.050}$	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.989 \pm 0.014}$
DDXPlus	Static-CM-ER	0.389 ± 0.044	0.007 ± 0.005	$\mathbf{0.816 \pm 0.021}$
	MaxProb-ER	0.594 ± 0.062	$\mathbf{0.004 \pm 0.002}$	0.766 ± 0.013
	D-CEM	$\mathbf{0.341 \pm 0.036}$	$\mathbf{0.004 \pm 0.003}$	0.803 ± 0.027
MedQA-Tri	Static-CM-ER	0.692 ± 0.031	0.145 ± 0.009	0.644 ± 0.012
	MaxProb-ER	0.725 ± 0.027	0.153 ± 0.011	0.638 ± 0.022
	D-CEM	$\mathbf{0.640 \pm 0.034}$	$\mathbf{0.122 \pm 0.012}$	$\mathbf{0.652 \pm 0.008}$

Here $\xi = 0$ means the agent ignores the mediator update, while $\xi = 1$ means full adoption. This is the binary-adoption instance of equation 17, with $\xi^{(t)} \sim \mathrm{Bernoulli}(\xi)$.

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 8 reports mean±std over 10 seeds. In S1, S2, and S3a, higher compliance and longer horizons reduce $\widehat{R}_L^{\mathrm{pool}}$ and improve accuracy, supporting steering's value when agents can act on the recommendation. This suite uses a separate S3b configuration from Figure 3, with shared bias that the permitted updates cannot correct. Escalation signals rise from 54% to 89% with horizon, flagging persistent shared errors independently of ξ.

Table 8: Partial-compliance ablation on the synthetic diagnostic task ($T \in \{4, 8, 12\}$, mean±std over 10 seeds). $\widehat{R}_L^{\mathrm{pool}}$ = diagnostic cost; SA-HR = max single-agent high-risk miss rate; Esc. = STOP_AND_ESCALATE rate.

Scenario 1 – Ideal

	$T=4$					$T=8$					$T=12$				
ξ	A1	A2	$\widehat{R}_L^{\mathrm{pool}}\downarrow$	SA-HR↓	Esc.	A1	A2	$\widehat{R}_L^{\mathrm{pool}}\downarrow$	SA-HR↓	Esc.	A1	A2	$\widehat{R}_L^{\mathrm{pool}}\downarrow$	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	$\widehat{R}_L^{\mathrm{pool}}\downarrow$	SA-HR↓	Esc.	A1	A2	$\widehat{R}_L^{\mathrm{pool}}\downarrow$	SA-HR↓	Esc.	A1	A2	$\widehat{R}_L^{\mathrm{pool}}\downarrow$	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	$\widehat{R}_L^{\mathrm{pool}}\downarrow$	SA-HR↓	Esc.	A1	A2	$\widehat{R}_L^{\mathrm{pool}}\downarrow$	SA-HR↓	Esc.	A1	A2	$\widehat{R}_L^{\mathrm{pool}}\downarrow$	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	$\widehat{R}_L^{\mathrm{pool}}\downarrow$	SA-HR↓	Esc.↑	A1	A2	$\widehat{R}_L^{\mathrm{pool}}\downarrow$	SA-HR↓	Esc.↑	A1	A2	$\widehat{R}_L^{\mathrm{pool}}\downarrow$	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

H.2 LLM STEERING: INTERVENTION AND TARGET-SELECTION ABLATIONS

Comparison. We compare four steering variants on matched, class-balanced DDXPlus $K = 8$ cases (S1–S3b), sharing agents, CM inference, development-selected stopping thresholds, and a six-round budget. Each variant follows its own interaction history.

- **Full D-CEM:** the complete controller.
- **No-STEER:** remove targeted challenges ($A^{(t)} = 0$), retaining CM pooling and the remaining control rules.

- **Generic review:** replace the differential challenge by checks supporting or contradicting the selected agent's current diagnosis, followed by revision.
- **Random alternative:** retain receiver selection and the differential prompt, but choose the alternative uniformly among noncurrent labels.

Results. Full achieves lower mean autonomous risk, H.C., and total cost than No-STEER. Risk falls by about 20% versus Generic review and 40% versus Random, with modest additional or 8% lower total cost, respectively (Figure 5).

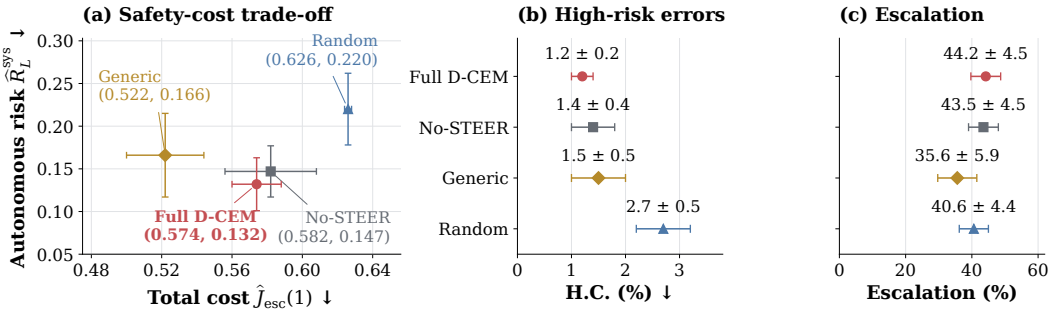

Figure 5: DDXPlus steering ablation, macro-averaged over S1–S3b: (a) autonomous risk and total cost; (b) H.C.; (c) escalation. Points show means; error bars denote ±1 standard deviation. All variants share cases and resource limits.

I Additional Ablations, Robustness, and Calibration

Following the steering studies in Appendix H, we examine energy, confusion structure, and escalation, then prompt, label-space, and parameter robustness. Independent population-risk calibration concludes this appendix.

I.1 Energy Ablation

On the supplementary DDXPlus steering study's cases, we replace full energy with pooled expected loss $b_t = R_L(d_1^{(t)}; Q^{(t)})$ for stopping and stagnation detection, retaining CM inference, steering, and the margin gate. Each score's thresholds minimize macro development $\widehat{J}_{esc}(1)$ over three fitting-quantile candidates and are frozen before TEST. Full energy lowers autonomous risk ($0.564 \to 0.132$) and H.C., while reducing total cost ($0.646 \to 0.574$) even after charging for its additional escalations, supporting energy-based control beyond expected-loss stopping (Table 9).

Table 9: DDXPlus energy ablation with development-selected thresholds (mean ± std).

Stopping score	$\widehat{R}_L^{sys}\downarrow$	$\widehat{J}_{esc}(1)\downarrow$	H.C.↓	Esc. (%)
Full energy	$\mathbf{0.132 \pm 0.031}$	$\mathbf{0.574 \pm 0.014}$	$\mathbf{0.012 \pm 0.002}$	44.2 ± 4.5
Risk only	0.564 ± 0.052	0.646 ± 0.028	0.080 ± 0.009	8.3 ± 2.5

I.2 Confusion-Matrix Ablation

To assess the value of agent-specific error patterns, we compare full CMs with class-recall-only CMs, which preserve diagonal recall and distribute off-diagonal mass uniformly, and uniform CMs with all entries $1/K$ (Table 10). Synthetic reruns the closed loop with frozen settings; DDXPlus changes CM-based inference on fixed LLM traces, preserving actions and escalation. Full CMs improve Synthetic cost and DDXPlus pooled risk and H.C. over Recall-only, supporting the value of agent-specific error information. On fixed DDXPlus trajectories, Full reduces pooled risk by about 11% ($0.310 \to 0.275$), a safety gain at modest additional cost ($0.319 \to 0.341$).

Table 10: CM ablation (mean $\pm$ std). Pooled risk evaluates terminal pooled decisions on all cases, including escalated cases; $\widehat{J}_{\mathrm{esc}}(1)$ uses autonomous system risk plus escalation rate.

Dataset	CM	$\widehat{R}_L^{\mathrm{pool}}\downarrow$	H.C.$\downarrow$	$\widehat{J}_{\mathrm{esc}}(1)\downarrow$	Esc. (%)
Synthetic	Full	$\mathbf{0.011 \pm 0.014}$	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.043 \pm 0.050}$	3.3
	Recall-only	0.016 ± 0.020	$\mathbf{0.000 \pm 0.000}$	0.080 ± 0.027	7.1
	Uniform	0.368 ± 0.016	0.054 ± 0.021	0.622 ± 0.059	27.1
DDXPlus	Full	$\mathbf{0.275 \pm 0.052}$	$\mathbf{0.004 \pm 0.003}$	0.341 ± 0.036	26.0
	Recall-only	0.310 ± 0.045	0.005 ± 0.004	$\mathbf{0.319 \pm 0.037}$	26.0
	Uniform	2.375 ± 0.000	0.320 ± 0.015	2.153 ± 0.059	26.0

Closed-loop DDXPlus comparison. Using the supplementary steering study's cases and Full reference, we replace CMs throughout inference and control, allowing each variant to follow its own interventions. Full reduces pooled risk from 0.964 to 0.799 and cost from 0.753 to 0.574 relative to Recall-only. Full retains 55.8% autonomous coverage with cost 0.574, compared with Uniform's complete escalation and unit cost. Table 11 reports these results separately from the fixed-trajectory comparison above.

Table 11: Closed-loop DDXPlus CM ablation (mean $\pm$ std), using the supplementary steering study's cases and Full reference.

CM	$\widehat{R}_L^{\mathrm{pool}}\downarrow$	$\widehat{J}_{\mathrm{esc}}(1)\downarrow$	H.C.$\downarrow$	Esc. (%)
Full	$\mathbf{0.799 \pm 0.045}$	$\mathbf{0.574 \pm 0.014}$	0.012 ± 0.002	44.2 ± 4.5
Recall-only	0.964 ± 0.021	0.753 ± 0.012	0.051 ± 0.005	47.8 ± 1.7
Uniform	2.375 ± 0.000	1.000 ± 0.000	$\mathbf{0.000 \pm 0.000}^{\dagger}$	100.0 ± 0.0

†All cases are escalated; autonomous coverage is zero.

I.3 Escalation Ablation

Table 12 compares Full with no-escalation D-CEM: selective referral reduces autonomous risk and H.C. on both clinical tasks, providing a safeguard for unresolved cases.

Table 12: Ablation of STOP_AND_ESCALATE. Lower is better for all metrics. Risk and H.C. are mean $\pm$ std.

		$\widehat{R}_L^{\mathrm{sys}}\downarrow$				$\widehat{J}_{\mathrm{esc}}(1)\downarrow$				H.C.$\downarrow$			
Dataset	Variant	S1	S2	S3a	S3b	S1	S2	S3a	S3b	S1	S2	S3a	S3b
MedQA-Tri	D-CEM (full)	$\mathbf{0.47}_{\pm.10}$	$\mathbf{0.51}_{\pm.03}$	$\mathbf{0.45}_{\pm.06}$	$\mathbf{0.58}_{\pm.08}$	**0.58**	**0.65**	0.61	**0.72**	$\mathbf{0.11}_{\pm.04}$	$\mathbf{0.13}_{\pm.01}$	$\mathbf{0.11}_{\pm.03}$	$\mathbf{0.14}_{\pm.03}$
	w/o ESCALATE	$0.58_{\pm.06}$	$0.68_{\pm.06}$	$0.60_{\pm.07}$	$0.73_{\pm.07}$	**0.58**	0.68	**0.60**	0.73	$0.14_{\pm.03}$	$0.17_{\pm.03}$	$0.14_{\pm.03}$	$0.17_{\pm.02}$
DDXPlus	D-CEM (full)	$\mathbf{0.06}_{\pm.03}$	$\mathbf{0.03}_{\pm.02}$	$\mathbf{0.03}_{\pm.02}$	$\mathbf{0.21}_{\pm.09}$	**0.11**	0.36	**0.49**	0.41	$\mathbf{0.00}_{\pm.01}$	$\mathbf{0.00}_{\pm.00}$	$\mathbf{0.00}_{\pm.01}$	$\mathbf{0.01}_{\pm.01}$
	w/o ESCALATE	$0.11_{\pm.04}$	$0.15_{\pm.07}$	$0.50_{\pm.09}$	$0.39_{\pm.11}$	**0.11**	**0.15**	0.50	**0.39**	$0.01_{\pm.01}$	$0.02_{\pm.01}$	$0.03_{\pm.02}$	$0.02_{\pm.02}$

Let R_f and R_0 denote their system risks, respectively, and q the escalation rate of full D-CEM. For fixed outcomes, the cost curves and their intersection are

$$J_f(c) = R_f + cq, \qquad J_0(c) = R_0, \qquad c^* = \frac{R_0 - R_f}{q}.$$

Figure 6 quantifies the cost of the safety gain. On DDXPlus, escalation lowers autonomous risk by about 72% ($0.286 \to 0.081$), with a modest total-cost difference at $c = 1$ ($0.286 \to 0.341$); it also lowers total cost when $c < c^* \approx 0.79$. At the same 26% escalation as Static-CM-ER, D-CEM retains a 0.048 cost advantage for every $c \geq 0$, since referral-cost terms cancel.

I.4 Neutral Prompts and Expanded Label Spaces

To test robustness beyond role anchors and a small label space, we use task-only neutral prompts that remove specialist, default-label, and shared-bias anchors. Table 13 compares the original $K = 8$ setting with neutral DDXPlus tasks at $K = 8, 16, 24$. The expanded classes are drawn from DDXPlus's released diagnostic labels. D-CEM achieves the lowest point-estimate escalation-inclusive cost among the compared methods at each label-space size. These results support the control mechanism beyond the original role anchors and eight-label setting.

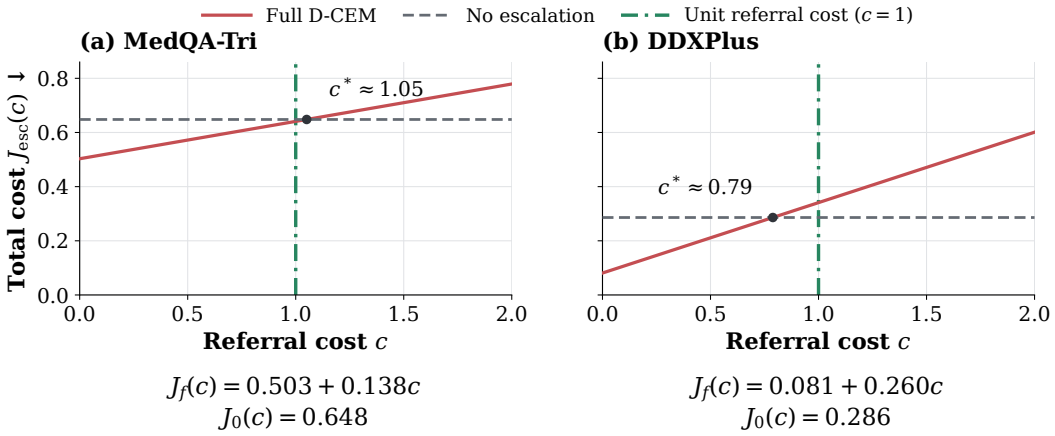

Figure 6: Escalation-cost comparison for fixed outcomes, using the rounded coefficients shown. Full D-CEM has lower cost for $c < c^*$ and equal cost at $c = c^*$. Reported unit-cost values (Full/no-escalation) are $0.640/0.648$ for MedQA-Tri and $0.341/0.286$ for DDXPlus, computed before coefficient rounding.

Table 13: Prompt and label-space comparisons on DDXPlus (mean±std over 10 seeds). $J = \widehat{J}_{\text{esc}}(1)$; Esc. is the escalation percentage. Bold marks the lowest mean J and H.C. in each setting, including ties.

Method	Role-anchored $K = 8$			Neutral $K = 8$			Neutral $K = 16$			Neutral $K = 24$		
	$J \downarrow$	H.C.↓	Esc. (%)	$J \downarrow$	H.C.↓	Esc. (%)	$J \downarrow$	H.C.↓	Esc. (%)	$J \downarrow$	H.C.↓	Esc. (%)
Single best	0.650±.036	0.120±.006	0.0±0.0	0.171±.024	0.027±.007	0.0±0.0	0.727±.091	0.141±.020	0.0±0.0	0.781±.071	0.135±.015	0.0±0.0
Fixed pool	1.528±.086	0.297±.017	0.0±0.0	0.169±.023	0.025±.006	0.0±0.0	0.539±.078	0.086±.016	0.0±0.0	0.646±.080	0.083±.016	0.0±0.0
Free discussion	1.363±.146	0.262±.029	0.0±0.0	0.153±.027	0.023±.006	0.0±0.0	0.469±.063	0.078±.012	0.0±0.0	0.594±.073	0.083±.016	0.0±0.0
Static-CM-ER	0.389±.044	0.007±.005	26.0±3.4	0.160±.012	0.023±.004	2.8±1.4	0.389±.035	**0.016**±.008	21.9±1.8	0.394±.059	0.016±.014	13.5±3.8
D-CEM	**0.341**±.036	**0.004**±.003	26.0±3.4	**0.146**±.010	**0.021**±.004	3.5±2.0	**0.349**±.035	**0.016**±.008	25.2±2.3	**0.379**±.034	**0.002**±.002	16.4±3.7

I.5 Sensitivity and Robustness

Parameter settings and sensitivity. Table 14 lists nominal controller defaults. Energy weights and the reliability floor are fixed a priori; stopping thresholds are selected on CAL/development data and frozen before TEST, without case- or seed-specific retuning. Rule-guided thresholds vary by scenario to accommodate different energy scales.

Table 14: Default parameter settings.

Setting	(α, β, γ)	$\varepsilon_{\text{safe}}$	m_{safe}	ε_{low}	$\rho_{\min}$
LLM	$(1, 1, 1)$	2.80	0.05	4.50	0.05
Rule-guided S1–S3b	$(1, 1, 1)$	0.20–0.50	0.05–0.50	0.65–0.75	0.05

We vary one parameter at a time around a separately frozen reference in each sensitivity suite, without TEST selection. Table 15 reports Synthetic end-to-end and DDXPlus LLM closed-loop evaluations from independent runs, so their reference outcomes need not match the main comparison. Synthetic pooled risk remains at most 0.012 with zero H.C.; DDXPlus is unchanged under the tested margin and reliability-floor variations.

CM sensitivity. To assess robustness to CM misspecification, we use $C_\eta = (1 - \eta)C + \eta U$, with $U_{ij} = 1/K$, and separately force both agents to report the same wrong label on a fixed 20% TEST subset. The controlled shared-failure test uses a separate rule-guided (RG) suite, with DDXPlus providing complementary fixed-trajectory CM-sensitivity evidence. Mild CM perturbations ($\eta \leq .25$) retain low H.C. and costs close to the nominal values (Table 16).

I.6 Independent Population-Risk Calibration

To evaluate Theorem 1's population-risk guarantee, we use a supplementary, class-balanced DDXPlus $K = 8$ experiment with neutral prompts and error losses 5/1 for high-risk/ordinary classes. Disjoint

Table 15: Parameter sensitivity: min–max values after S1–S3b macro-averaging, reported as ($\widehat{R}_L^{\text{pool}}$, H.C., escalation %, rounds).

Parameter	Sweep	Synthetic	DDXPlus closed-loop
α	$\{.5, 1, 2\}\times$	(.009–.011, 0, .7–3.4, 5.09–5.31)	(.274, .023, 9.4–12.9, 2.03–2.05)
β	$\{.5, 1, 2\}\times$	(.007–.011, 0, 2.9–3.3, 5.08–5.32)	(.274, .009–.023, 9.6–53.4, 2.03–2.15)
γ	$\{.5, 1, 2\}\times$	(.010–.012, 0, 2.7–4.2, 5.08–9.82)	(.274, .023, 9.7–12.9, 2.03–2.04)
$\varepsilon_{\text{safe}}$	$\{.8, 1, 1.2\}\times$	(.009–.011, 0, 3.3, 5.06–5.29)	(.274, .017–.023, 9.6–17.2, 2.03–2.05)
m_{safe}	$\{.5, 1, 2\}\times$	(.011, 0, 3.3, 5.20)	(.274, .023, 9.7, 2.03)
ε_{low}	$\{.8, 1, 1.2\}\times$	(.011, 0, 2.8–3.3, 5.20–5.21)	(.274, .023, 9.7, 2.03)
$\rho_{\min}$	$\{.01, .05, .10, .20\}$	(.011, 0, 3.3, 5.20)	(.274, .023, 9.7, 2.03)

Table 16: CM perturbations and TEST-only shared failures (mean $\pm$ std). A dash denotes an unreported experiment.

	RG stress suite		DDXPlus fixed trajectories	
Condition	$\widehat{J}_{\text{esc}}(1)$	H.C.	$\widehat{J}_{\text{esc}}(1)$	H.C.
Nominal	0.144 ± 0.027	0.005 ± 0.002	0.341 ± 0.036	0.004 ± 0.003
$\eta = .10$	0.136 ± 0.019	0.004 ± 0.001	0.345 ± 0.038	0.004 ± 0.003
$\eta = .25$	0.161 ± 0.021	0.004 ± 0.001	0.360 ± 0.029	0.004 ± 0.003
$\eta = .50$	0.206 ± 0.016	0.003 ± 0.002	0.445 ± 0.020	0.004 ± 0.003
Shared failure (20%)	0.287 ± 0.024	0.017 ± 0.005	–	–

sets contain 80 CM-fitting, 800 risk-calibration, and 160 TEST cases. One complete controller is frozen before risk calibration, using the LLM defaults in Table 14 and a six-round budget; uncertified budget endpoints escalate.

At the pre-specified target $r = 0.5$, the 95% direct risk upper bound is 0.329, with 93.38% empirical decision coverage (Table 17). The unchanged controller achieves conditional risk 0.100 at 93.75% coverage on independent TEST cases, supporting population-risk calibration after adaptive interaction and stopping in this setting.

Table 17: Independent DDXPlus risk calibration. Coverage is the fraction passing both stopping gates; $\widehat{R}_{\text{dec}}$ is average weighted loss among those decisions. The 95% upper bound uses only risk-calibration cases and is rounded upward.

Split	Cases	Coverage (%)↑	$\widehat{R}_{\text{dec}}\downarrow$	Risk upper bound
Risk calibration	800	93.38	0.115	0.329
Independent TEST	160	93.75	0.100	–

J 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 constructs confusion-induced posteriors from the reported labels, frozen confusion matrices, and prior; 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 only.
```

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.

K 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.