MELD: RECURSIVE SELF-IMPROVEMENT VIA HIERARCHICAL COORDINATION FOR MULTI-AGENT SCIENTIFIC DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-agent scientific discovery with large language models raises a coupled learning problem: research assignments determine both which experiments are pursued and the experiences from which agents adapt. We introduce MELD, a unified hierarchical test-time training framework that jointly adapts a team-level Coordinator policy and agent-level individual Scientist policies within a single discovery process. The Coordinator generates a joint portfolio of search branches and research directives over a shared derivation tree, while Scientists develop hypotheses and execution plans implemented by a shared frozen Executor. Both levels receive alternating GRPO-style parameter updates under a shared smooth best-of-budget objective, allowing team organization and individual research behavior to co-evolve in a recursive self-improvement loop. Our analysis establishes objective alignment for the hierarchy, characterizes conditions for coordination gains over independent parallel discovery and the first-order benefit of private adaptation, and derives a finite-budget regret bound for adapted models. On 22 MLGym and MLE-bench tasks, MELD is best on 20 of them and achieves the highest aggregate normalized scores on MLGym, using a 4B backbone for coordination and research. Most baselines use proprietary Scientists, and all systems share one frozen Executor.

1 INTRODUCTION

Scientific discovery requires judiciously allocating a limited exploration/evaluation budget across competing hypotheses. Large language models (LLMs) increasingly support this process by proposing, testing, and refining candidates (Romera-Paredes et al., 2024; Novikov et al., 2025), while test-time training (TTT) uses discovery feedback to adapt their research policies (Wang et al., 2026b; Yuksekgonul et al., 2026; Juneja et al., 2026). For a team of collaborative agents, each local research assignment determines both the team's overall progress and the learning experience that shapes the team's subsequent research strategy. Agent coordination in scientific discovery, such as focusing collective effort on promising research lineages while maintaining sufficient exploration, becomes part of the learning process rather than merely a scheduling mechanism.

Existing approaches provide complementary capabilities. TTT-Discover and ARTS adapt task-local proposal or research policies (Yuksekgonul et al., 2026; Juneja et al., 2026), while CORAL coordinates parallel discovery through shared persistent memory (Qu et al., 2026). Agent Q-Mix learns communication and coordination across a training distribution (Jiang et al., 2026; Rashid et al., 2018), without jointly adapting its coordination strategy and underlying LLM researchers to each new discovery task. These advances leave a central joint learning problem unresolved: *how can a team-level coordination policy and individual agent-level research policies be learned together within a single discovery process?* We address this problem through a unified framework that jointly adapts both levels via task-local parameter updates, allowing team organization and individual research behavior to co-evolve.

We introduce **MELD**, a unified hierarchical TTT framework for multi-agent scientific discovery. A team-level Coordinator generates a joint portfolio of parent branches and research directives over a shared derivation tree. Initially identical Scientists, each with a private adapter, develop their assignments into hypotheses and execution plans implemented by a shared frozen Executor. The

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.

Coordinator and Scientists learn through alternating GRPO-style parameter updates (Shao et al., 2024) under a shared smooth best-of-budget objective. Thus, MELD creates a recursive self-improvement (RSI) loop: Joint portfolio coordination organizes research across agents and shapes the team's learning experience, while private agents' evolving capabilities and outcomes guide subsequent coordination.

Our theoretical analysis establishes objective alignment for the proposed hierarchy, characterizes conditions for coordination gains over independent parallel discovery at matched individual outcome distributions and budgets, and quantifies the first-order benefit of private adaptation at a common model. We further derive a finite-budget regret bound for adapted models and a conditional optimization rate. Across 12 MLGym (Nathani et al., 2025) and 10 MLE-bench (Chan et al., 2025) tasks, MELD attains the best three-seed mean score among the compared methods on 20 tasks (Table 1) and the highest aggregate normalized scores on MLGym, with an interquartile mean of 0.748 versus 0.684 for ARTS*. These system-level comparisons use a 4B backbone for MELD's coordination and research, proprietary Scientist models for most baselines, and the same frozen Executor throughout.

Our contributions are threefold:

- **A unified framework for task-local RSI.** We jointly learn a team-level portfolio coordination policy and agent-level Scientist research policies within a shared-objective hierarchy in a single discovery task.
- **Theory of coordination and private adaptation.** We quantify the benefits of MELD and relate surrogate optimization to finite-budget discovery regret at adapted policy.
- **Strong empirical performance.** We report the best score on 20 of 22 MLGym and MLE-bench tasks and leading aggregate performance on MLGym with a compact research backbone, complemented by search trajectories and adaptation ablations.

2 Related Work

Test-Time Training for Discovery. TTT updates model parameters using a signal from the test instance or task (Sun et al., 2020), whereas test-time scaling allocates additional inference computation without updating them (Snell et al., 2025). Early discovery systems such as FunSearch (Romera-Paredes et al., 2024) and AlphaEvolve (Novikov et al., 2025) iteratively generate and select candidates with a fixed proposal model, while recent methods center adaptation on a single task-local proposal policy (Wang et al., 2026b; Yuksekgonul et al., 2026), strategic advisor (Yan et al., 2026), uncertainty-guided exploration component (Riaz et al., 2026), or tree-search Scientist (Juneja et al., 2026). MELD instead organizes the experiences of multiple concurrently adapting research policies.

Multi-Agent Systems and Learned Orchestration. In discovery, CORAL (Qu et al., 2026) and SwarmResearch (Virk et al., 2026) coordinate parallel agents through shared memory or centralized orchestration. Parameter-trained systems (Zhang et al., 2026; Wang et al., 2026a) typically learn across a training distribution, while TacoMAS (Xu et al., 2026) co-evolves role-specific capabilities and topology at inference. Related MARL work learns emergent role representations during training (Wang et al., 2020), whereas MELD induces persistent task-local specialization through assignment-conditioned private adapter updates. GPTSwarm (Zhuge et al., 2024), ADAS (Hu et al., 2025), and the RL-trained Conductor (Nielsen et al., 2026) establish that orchestration can be learned. MELD differs by alternating task-local weight updates to the team-level coordination policy and multiple agent-level Scientist policies within the same unseen discovery run, under one shared best-of-budget objective, so that team organization and individual research behavior can co-evolve.

3 Preliminaries

Executable Discovery Task. Following Juneja et al. (2026), we study test-time discovery on a single previously unseen executable task $\mathcal{P} = (\mathcal{W}_0, m, s_0, T)$, where $\mathcal{W}_0$ contains the task specification, data, baseline artifact c_0, and evaluation harness; m is the task-specific metric; $s_0 = s(c_0)$ is the normalized baseline utility; and T is a system-level budget shared by all agents. Given the raw evaluator metric $q_m(c)$, we define the normalized utility $s(c) = g_m(q_m(c)) \in [0, 1]$, where g_m is fixed for each task and monotone in the preferred direction of m, so that larger $s(c)$ always indicates

a better artifact. In our implementation, $g_m(q) = \sigma(z_m(q))$, where z_m performs metric-specific standardization and direction correction, and σ is the logistic sigmoid.

Let $\mathcal{C}_T^{\mathrm{val}}(\Pi)$ denote the successfully evaluated candidates produced within budget by an adaptive multi-agent policy Π. We optimize expected best-of-budget utility, corresponding to a simple-regret objective in fixed-budget pure exploration (Bubeck et al., 2009),

$$J_T(\Pi;\mathcal{P}) = \mathbb{E}_\Pi\left[\max_{c\in\{c_0\}\cup\mathcal{C}_T^{\mathrm{val}}(\Pi)} s(c)\right], \qquad \mathfrak{R}_T = s^\star - J_T(\Pi;\mathcal{P}), \tag{1}$$

where $s^\star$ is the optimal feasible score. All parameter updates use only interaction data from this task.

Shared Discovery Process. At round t, the shared history is represented by a rooted derivation tree $\mathcal{G}_t = (V_t, E_t)$, whose root stores (c_0, s_0). Together with the remaining budget b_t, it defines the discovery state $\mathcal{S}_t = (\mathcal{P}, \mathcal{G}_t, b_t)$. Conditioned on $\mathcal{S}_t$, policy Π proposes $n_t \le N$ concurrent expansions, where N is the maximum number of active research units, with expansion i given by

$$\xi_t^i = (p_t^i, h_t^i, e_t^i), \qquad c_t^i = \mathrm{Exec}(c_{p_t^i}; h_t^i, e_t^i), \qquad (\ell_t^i, s_t^i) = \mathrm{Eval}_m(c_t^i; \mathcal{W}_0), \tag{2}$$

where $p_t^i \in V_t$ selects a parent node, h_t^i is a testable hypothesis, and e_t^i is an execution plan. The evaluator returns feedback ℓ_t^i and utility s_t^i, with $s_t^i = \perp$ on failure. The Executor remains frozen, confining parameter adaptation to the policies that select and develop research directions.

Each successful evaluation adds $v_t^i = (h_t^i, e_t^i, c_t^i, \ell_t^i, s_t^i)$ to the tree. Defining $V_t^+ = \{v_t^i : s_t^i \neq \perp\}$, we update $V_{t+1} = V_t \cup V_t^+$ and $E_{t+1} = E_t \cup \{(p_t^i, v_t^i) : v_t^i \in V_t^+\}$. Failed attempts consume budget and provide learning signals but are not added as expandable nodes. Section 4 instantiates Π hierarchically: the team-level Coordinator assigns $z_t^i = (p_t^i, d_t^i)$, and each agent-level Scientist i produces $u_t^i = (h_t^i, e_t^i)$; their alternating adaptation forms MELD's multi-agent RSI loop.

4 MELD: HIERARCHICAL TEST-TIME TRAINING FOR SELF-IMPROVING MULTI-AGENT DISCOVERY

MELD couples joint research assignment with private policy adaptation. A learnable *Coordinator* assigns research directives to N agents (*research units*), each pairing a private Scientist with the shared frozen Executor. The two policy levels, the team-level Coordinator and the agent-level Scientists, receive alternating task-local updates under a shared smooth best-of-budget objective. Assignments thus shape both current exploration and the experience available for private adaptation.

4.1 HIERARCHICAL POLICY

At round t, the coordination policy $\pi_\phi^{\mathrm{coord}}$ observes the shared discovery state $\mathcal{S}_t$ and generates a *portfolio* $Z_t = \{z_t^i\}_{i\in\mathcal{I}_t}$ for active units $\mathcal{I}_t \subseteq \{1,\dots,N\}$. Each assignment $z_t^i = (p_t^i, d_t^i)$ specifies a parent node p_t^i from a retrieved tree shortlist and a research directive d_t^i: a schema-constrained natural-language statement of the research question, the parent's bottleneck, and the constraints to respect. The portfolio is emitted as one structured completion, allowing assignments to depend on one another before candidates are generated.

Given $(\mathcal{S}_t, z_t^i)$, Scientist i samples a hypothesis and execution plan $u_t^i = (h_t^i, e_t^i) \sim \pi_{\theta_i}^{\mathrm{sci}}(\cdot \mid \mathcal{S}_t, z_t^i)$. The frozen Executor implements the plan as $c_t^i = \mathrm{Exec}(c_{p_t^i}; h_t^i, e_t^i)$, which is evaluated as in Eq. (2). Scientists sample independently conditional on the state and portfolio. With $U_t = \{u_t^i\}_{i\in\mathcal{I}_t}$,

$$\pi_\Theta(Z_t, U_t \mid \mathcal{S}_t) = \pi_\phi^{\mathrm{coord}}(Z_t \mid \mathcal{S}_t) \prod_{i\in\mathcal{I}_t} \pi_{\theta_i}^{\mathrm{sci}}(u_t^i \mid \mathcal{S}_t, z_t^i), \qquad \Theta = (\phi, \theta_{1:N}). \tag{3}$$

The Coordinator decides *where and what* to research; each Scientist decides *how*. Coordination resides in the joint assignment policy, even though Scientists sample independently once assigned.

The Coordinator and Scientists share one frozen backbone, with a LoRA adapter (Hu et al., 2022) ϕ for coordination and a private θ_i for each Scientist. All Scientists start identically, $\theta_{1,0} = \cdots = \theta_{N,0}$, and every adapter resets at task start; neither the backbone nor the Executor is updated. The tree shares artifacts and feedback, while private adapters retain assignment-conditioned updates without forcing identical policy changes.

4.2 A SHARED BEST-OF-BUDGET OBJECTIVE

Equation (1) values the best validated artifact within budget. Parent-relative gains can reward branch improvements that leave the best-so-far utility unchanged, while scoring candidates separately ignores their combination. We therefore evaluate each round against the shared discovery state, using a smooth potential whose terminal value approximates best-of-budget utility.

Smooth potential. Let K be the evaluator-call budget allowed by T, and V the derivation-tree node set in state $\mathcal{S}$, including the root. Define

$$\begin{aligned} \Phi_{\tau,K}(\mathcal{S}) &= \tau \log\Big(\tfrac{1}{K+1}\Big[\textstyle\sum_{v\in V} e^{s_v/\tau} + (K+1-|V|)\, e^{s_{\min}/\tau}\Big]\Big), \\ J_{\tau,K}(\Theta) &= \mathbb{E}_{\Pi_\Theta}\big[\Phi_{\tau,K}(\mathcal{S}_{\text{end}})\big], \end{aligned} \tag{4}$$

where $\tau > 0$ is a temperature, $\mathcal{S}_{\text{end}}$ is the state at budget exhaustion, and failed or unused slots carry $s_{\min} = 0$. Padding fixes the normalization as the tree grows: each validated candidate replaces a sentinel slot rather than changing the denominator. This log-sum-exp lies below the best utility by at most $\tau \log(K+1)$ (Proposition 5.2(i)), making $J_{\tau,K}$ a shared surrogate of J_T for both policy levels.

Potential increments as shared credit. A round receives $r_t^\Phi = \Phi_{\tau,K}(\mathcal{S}_{t+1}) - \Phi_{\tau,K}(\mathcal{S}_t)$ for its collective outcome. The temperature sets how closely the potential tracks the maximum while retaining credit for near-frontier candidates. Joint assignments are thus evaluated through realized candidate utilities, not an explicit diversity bonus. The increments telescope, $\sum_t r_t^\Phi = \Phi_{\tau,K}(\mathcal{S}_{\text{end}}) - \Phi_{\tau,K}(\mathcal{S}_0)$; the exact score-function gradient (Williams, 1992) credits each decision with the full downstream sum (Proposition 5.2). Direct Monte Carlo estimation requires continuations to budget exhaustion, so MELD limits adaptation to a *one-round horizon*.

At snapshot $\mathcal{S}$, a policy block $\beta \in \{\text{coord}, \text{sci}_1, \ldots, \text{sci}_N\}$ takes action a in context x: $a = Z$, $x = \mathcal{S}$ for the Coordinator, and $a = u^i$, $x = (\mathcal{S}, z^i)$ for Scientist i. Its local objective is

$$\begin{aligned} J_\beta^{\text{loc}}(\mathcal{S}) &= \mathbb{E}_{a\sim\pi_\beta(\cdot|x)}\big[\Phi_{\tau,K}(\mathcal{S}\oplus X(a)) - \Phi_{\tau,K}(\mathcal{S})\big], \\ \nabla_\beta J_\beta^{\text{loc}}(\mathcal{S}) &= \mathbb{E}\big[r^\Phi(a)\,\nabla_\beta \log \pi_\beta(a \mid x)\big]. \end{aligned} \tag{5}$$

Here $X(a)$ contains the round's validated candidates, $\mathcal{S}\oplus X(a)$ extends the snapshot with their nodes, and $r^\Phi(a)$ is the realized potential increment; the gradient identity is established in Proposition 5.2. The expectation includes execution and evaluation randomness and the other policies' responses at the fixed behavior model. For a Scientist update, the full portfolio is also fixed. This supplies round-local credit rather than an unbiased full-return gradient: the Coordinator is credited through its portfolio, each Scientist through its response, and both against the same team outcome.

4.3 ALTERNATING TEST-TIME ADAPTATION

Matched-snapshot groups. Because the tree and policies evolve, relative feedback is compared at a fixed context. We combine GRPO-style sampling (Shao et al., 2024) with a leave-one-out baseline (Ahmadian et al., 2024). Each group contains $G \geq 2$ independent *forks* of one snapshot, with the discovery state, remaining budget, parent shortlist, and behavior model Θ_{old} fixed. The N research units provide parallel exploration within a fork, while the G forks provide matched-context samples for adaptation. In a *Coordinator group*, each fork samples a complete portfolio $Z^{(g)} \sim \pi^{\text{coord}}_{\phi_{\text{old}}}(\cdot \mid \mathcal{S})$ and executes its units using the current Scientist policies. In a *Scientist group*, one portfolio Z is fixed across forks, each Scientist independently samples $u_i^{(g)} \sim \pi^{\text{sci}}_{\theta_{i,\text{old}}}(\cdot \mid \mathcal{S}, z^i)$, and their responses are executed together. Fork g yields validated candidates $X^{(g)}$, reward $r_g^\Phi = \Phi_{\tau,K}(\mathcal{S}\oplus X^{(g)}) - \Phi_{\tau,K}(\mathcal{S})$, and advantage $\widehat{A}_g = r_g^\Phi - \frac{1}{G-1}\sum_{g'\neq g} r_{g'}^\Phi$, whose baseline excludes fork g and, by conditional independence, preserves the sequence-score estimator's expectation in Eq. (5). Scientists share this team advantage but update only through their own response tokens, without pooling samples across snapshots or private policies.

Token-level update. For the policy being updated, let $y_g = (y_{g,1}, \ldots, y_{g,L_g})$ be its structured response in fork g and $\rho_{g,j} = \pi_\beta(y_{g,j} \mid x, y_{g,<j}) / \pi_{\beta,\text{old}}(y_{g,j} \mid x, y_{g,<j})$ the token importance ratio.

Algorithm 1 One complete matched-snapshot update group in MELD

1: Freeze $\mathcal{S}$ and behavior policies Θ_{old} during collection.
2: Sample independent portfolios (Coordinator group) or one shared portfolio (Scientist group).
3: For each fork, sample Scientists' responses, execute and evaluate them, and compute r_g^{Φ}.
4: Compute leave-one-out advantages $\widehat{A}_g$.
5: Take one step of Eq. (6) on ϕ (Coordinator) or on all private θ_i in parallel (Scientist).
6: Commit validated candidates, charge all evaluator calls to the budget, and switch the group type.

Broadcasting $\widehat{A}_g$ to its response tokens gives the token-averaged PPO-style surrogate (Schulman et al., 2017),

$$\mathcal{L}_\beta = -\frac{1}{G}\sum_{g=1}^{G}\frac{1}{L_g}\sum_{j=1}^{L_g}\min\big(\rho_{g,j}\widehat{A}_g,\ \text{clip}(\rho_{g,j}, 1-\epsilon_{\text{lo}}, 1+\epsilon_{\text{hi}})\,\widehat{A}_g\big) + \beta_{\text{KL}}\,\widehat{D}_{\text{KL}}\big(\pi_\beta \,\|\, \pi_\beta^{\text{ref}}\big), \quad (6)$$

where π_β^{ref} is the frozen task-start policy, distinct from the per-group behavior policy $\pi_{\beta,\text{old}}$, and $\widehat{D}_{\text{KL}}$ is token-averaged on the same responses. Token-length normalization, clipping, and KL regularization define the implemented surrogate rather than the exact sequence-score gradient of Eq. (5) (Appendix A.3).

Update and commit. Algorithm 1 summarizes a complete group with sufficient remaining calls. Coordinator and Scientist groups alternate using fresh samples, with no replay, until the budget no longer admits a full group, after which the current policies are rolled out without further updates (Appendix B.3). Fork rewards share the pre-commit snapshot, and their sum need not equal the merged potential increment. Because updated policies collect the next group, Coordinator updates change subsequent learning assignments while Scientist updates change the outcomes informing subsequent coordination, which enables policy-level RSI.

5 THEORETICAL ANALYSIS

This section shows that both levels of MELD follow partial gradients of one objective aligned with J_T (Section 5.1), quantifies what coordination and private adaptation add over parallel discovery (Section 5.2), and bounds the discovery regret after adaptation (Section 5.3), with proofs in Appendix A.

5.1 ONE OBJECTIVE ACROSS THE HIERARCHY

Although the team-level Coordinator and the agent-level Scientists act on different objects, Algorithm 1 trains both levels with the same reward r_t^{Φ}. We show that the exact gradients of the two levels are partial gradients of a single objective approximating J_T in Eq. (1).

Notation. Let Π_Θ denote the policy that applies π_Θ of Eq. (3) at every round with Θ fixed. A run of Π_Θ draws $\mathcal{S}_0 \sim \nu$ and evolves by the transition kernel $\mathcal{Q}(\mathcal{S}_{t+1} \mid \mathcal{S}_t, Z_t, U_t)$, which comprises execution, evaluation, and bookkeeping, and $\mathbb{E}_{\Pi_\Theta}$ denotes expectation over runs. A policy block $\beta \in \{\text{coord}, \text{sci}_1, \ldots, \text{sci}_N\}$ acts on a context x with an action a (Section 4.2), with $(a, x) = (Z_t, \mathcal{S}_t)$ for the Coordinator, which acts every round, and $(a, x) = (u_t^i, (\mathcal{S}_t, z_t^i))$ for Scientist i, which acts when $i \in \mathcal{I}_t$. We write (a_t^β, x_t^β) for this pair and ∇_β for the gradient with respect to block β. Let $B(\mathcal{S}) = \max_{v\in V} s_v$, so that $J_T(\Pi_\Theta; \mathcal{P}) = \mathbb{E}_{\Pi_\Theta}[B(\mathcal{S}_{\text{end}})]$, and let $R_{t:\text{end}}^{\Phi} = \sum_{t'\geq t} r_{t'}^{\Phi} = \Phi_{\tau,K}(\mathcal{S}_{\text{end}}) - \Phi_{\tau,K}(\mathcal{S}_t)$ be the *reward-to-go* of the team, which telescopes because r_t^{Φ} is a potential-based shaping reward (Ng et al., 1999).

Assumption 5.1 (Regularity)**.** *(R1) Given the sampled actions, ν and $\mathcal{Q}$ do not depend on Θ. (R2) Each $\pi_\beta(\cdot \mid x)$ is differentiable in its block, with support independent of the block's parameters. (R3) Block gradients and expectations over runs may be interchanged.*

Conditions (R1) and (R2) hold for softmax token policies with a frozen Executor and evaluator, and (R3) is the standard dominated-differentiation condition (Appendix A.2).

Proposition 5.2 (One objective across the hierarchy). *(i) For every Θ, $J_T(\Pi_\Theta;\mathcal{P}) - \tau\log(K+1) \le J_{\tau,K}(\Theta) \le J_T(\Pi_\Theta;\mathcal{P})$. Under Assumption 5.1: (ii) for every block β,*

$$\nabla_\beta J_{\tau,K}(\Theta) = \mathbb{E}_{\Pi_\Theta}\Big[\textstyle\sum_{t:\ \beta \text{ acts at } t} R^{\Phi}_{t:\mathrm{end}}\,\nabla_\beta \log \pi_\beta(a_t^\beta \mid x_t^\beta)\Big]; \tag{7}$$

(iii) writing $R^{\Phi}_{t:\mathrm{end}} = r^{\Phi}_t + R^{\Phi}_{t+1:\mathrm{end}}$ in Eq. (7), the r^{Φ}_t-terms sum to $\mathbb{E}_{\Pi_\Theta}\big[\sum_{t:\ \beta \text{ acts at } t}\nabla_\beta J^{\mathrm{loc}}_\beta(\mathcal{S}_t)\big]$ with J^{loc}_β as in Eq. (5), evaluated for a Scientist block at the realized portfolio Z_t; (iv) for a matched-snapshot group at $\mathcal{S}$ with behavior model Θ_{old} (Section 4.3), the leave-one-out estimator satisfies $\mathbb{E}\big[\widehat{A}_g\,\nabla_\beta \log \pi_\beta(a_g \mid x)\big] = \nabla_\beta J^{\mathrm{loc}}_\beta(\mathcal{S})$ at Θ_{old}, with a_g the action of block β in fork g.

Proposition 5.2 establishes that the two levels of MELD share a single objective, which approximates J_T up to the smoothing error $\tau\log(K+1)$ (i). Because the Executor and the evaluator are frozen and π_Θ factorizes as in Eq. (3), the gradient of every block weights its own score function by the same reward-to-go (ii), which is the shared-return policy gradient of cooperative multi-agent learning (Peshkin et al., 2000; Foerster et al., 2018). The population update behind Algorithm 1 is therefore block-coordinate ascent on the snapshot-local objectives $J^{\mathrm{loc}}_\beta(\mathcal{S}_t)$, whose gradients form the current-round component of Eq. (7) (iii) and are estimated without bias by the one-round estimator of Section 4.3 (iv), whereas the omitted terms involving $R^{\Phi}_{t+1:\mathrm{end}}$ capture the effect of a decision on later rounds and constitute the bias of the one-round horizon. Appendix A.3.1 analyzes the implemented surrogate of Eq. (6) and the rejection of malformed responses. Since the credit depends only on realized utilities and increases strictly in them (Appendix A.4), coordination is rewarded through outcomes alone, which Section 5.2 quantifies.

5.2 COORDINATION AND PRIVATE ADAPTATION VERSUS PARALLEL DISCOVERY

Parallel discovery runs N units, each choosing its assignment independently given the shared state, with frozen agents (Qu et al., 2026) or a single adapted policy (Yuksekgonul et al., 2026), whereas MELD draws the assignments jointly and adapts each Scientist privately. We compare the two designs at equal individual quality and equal budget, so that any difference is attributable to the design rather than to stronger units or additional evaluations. Fix a snapshot $\mathcal{S}$ with at least N evaluator calls remaining and fix the Scientist models. A *portfolio law* $\pi(Z \mid \mathcal{S})$ draws a full portfolio $Z = (z^1,\dots,z^N)$, after which unit i produces and evaluates its candidate independently of the other units, using only $(\mathcal{S}, z^i)$, and returns a utility $Y_i \in [0,1]$, with 0 for a failed call. Let q_i be the ith marginal of π and $\pi^{\mathrm{par}} = \prod_i q_i$ its *independent parallel* counterpart. The two laws induce the same unit marginals $H = (H_1,\dots,H_N)$, $H_i(y) = \Pr(Y_i \le y \mid \mathcal{S})$, and differ only in the dependence among assignments, which is all that coordination controls. Write $B_{\mathrm{new}} = \max\{B(\mathcal{S}), Y_1,\dots,Y_N\}$, $H_\pi(y) = \Pr_\pi(\max_i Y_i \le y \mid \mathcal{S})$, and $\Delta\mathcal{A} = \sum_i (e^{Y_i/\tau} - 1)$ for the soft-mass added to $\mathcal{A}(\mathcal{S}) = \sum_{v\in V} e^{s_v/\tau} + (K+1-|V|)$, the bracket in Eq. (4), so that $r^{\Phi} = \psi(\Delta\mathcal{A})$ for the strictly concave and increasing function $\psi(w) = \tau\log(1 + w/\mathcal{A}(\mathcal{S}))$. Let

$$\mathrm{MN}(H) = \min_{c\in[B(\mathcal{S}),1]} \Big\{ c + \textstyle\sum_i \int_c^1 \big(1 - H_i(y)\big)\,dy \Big\} \tag{8}$$

be the Meilijson–Nádas bound on $\mathbb{E}[B_{\mathrm{new}}]$ given H (Meilijson & Nádas, 1979), with $c \ge B(\mathcal{S})$ because B_{new} includes the incumbent.

Theorem 5.3 (Value of coordination at matched marginals). *In the setting above, for every portfolio law π: (i) $\mathbb{E}_\pi[B_{\mathrm{new}}] - \mathbb{E}_{\pi^{\mathrm{par}}}[B_{\mathrm{new}}] = \int_{B(\mathcal{S})}^1 \big(\prod_i H_i(y) - H_\pi(y)\big)\,dy \le \mathrm{MN}(H) - \mathbb{E}_{\pi^{\mathrm{par}}}[B_{\mathrm{new}}]$; (ii) $\mathbb{E}_\pi[r^{\Phi}] \le \psi(\mathbb{E}[\Delta\mathcal{A}])$, where $\mathbb{E}[\Delta\mathcal{A}]$ depends on H only, with equality if and only if $\Delta\mathcal{A}$ is π-a.s. constant; under π^{par} the inequality is strict unless every Y_i is degenerate; (iii) if $(Y_1,\dots,Y_N)$ is negatively associated under π (Joag-Dev & Proschan, 1983), then $\mathbb{E}_\pi[r^{\Phi}] \ge \mathbb{E}_{\pi^{\mathrm{par}}}[r^{\Phi}]$.*

Theorem 5.3 shows that an independent law can change $\mathbb{E}[B_{\mathrm{new}}]$ and $\mathbb{E}[r^{\Phi}]$ only through the marginals H, whereas a joint law with the same marginals adds the dependence term in (i), up to the coupling bound $\mathrm{MN}(H)$. By strict concavity of ψ, the training reward at fixed marginals is largest when $\Delta\mathcal{A}$ is deterministic, which no independent law with non-degenerate units achieves (ii), and it is at least the parallel reward under negative association (iii). When one of N directions has utility $s_{\mathrm{hi}} > B(\mathcal{S})$ and the others $s_{\mathrm{lo}} \le B(\mathcal{S})$, assigning them by a uniformly random permutation attains the bounds in (i) and (ii) and exceeds independent assignment in expected best-of-round utility by $(s_{\mathrm{hi}} - B(\mathcal{S}))(1 - 1/N)^N$ (Appendix A.5.2). Since the Coordinator emits the portfolio as one completion (Section 4.1) and is trained on r^{Φ}, such laws are expressible and rewarded by its objective.

Private adaptation. To isolate the learning signal that coordination gives each Scientist, fix $\mathcal{S}$ and a full portfolio Z, let all Scientists start from a common model with weights W_0, and let $\bar{r}^{\Phi}(W_1, \ldots, W_N) = \mathbb{E}[r^{\Phi} \mid \mathcal{S}, Z]$ be the expected round reward when Scientist i responds from weights $W_i = W_0 + \gamma U_i V_i$, with LoRA scaling $\gamma > 0$ and factors $\theta_i = (U_i, V_i)$ starting from the standard zero-output initialization $(0, V_0)$ (Hu et al., 2022). Write $\Gamma_i = (\nabla_{W_i} \bar{r}^{\Phi}) V_0^{\top}$ at the common model, $\bar{\Gamma} = N^{-1} \sum_i \Gamma_i$, and $\mathcal{D}(Z; \mathcal{S}) = N^{-2} \sum_{i<j} \|\Gamma_i - \Gamma_j\|_F^2$ for the dispersion of these assignment-conditioned gradients. We compare one population gradient step of size η on each private adapter with the same step on a single shared adapter, whose population gradient is the average of the N Scientist gradients because Eq. (6) averages over the GN pooled responses.

Proposition 5.4 (Private adaptation retains dispersion)**.** *For differentiable $\bar{r}^{\Phi}$, the first-order change of $\bar{r}^{\Phi}$ divided by N is $\eta\gamma^2 \big(\|\bar{\Gamma}\|_F^2 + \mathcal{D}(Z; \mathcal{S})\big)$ for private adapters and $\eta\gamma^2 \|\bar{\Gamma}\|_F^2$ for the shared adapter. If every Scientist receives the same assignment, then $\mathcal{D}(Z; \mathcal{S}) = 0$.*

Proposition 5.4 shows that a shared adapter follows only the mean signal $\bar{\Gamma}$, whereas private adapters also retain the dispersion $\mathcal{D}(Z; \mathcal{S})$. Because the Coordinator sets the law of Z, it also sets the expected dispersion available to the Scientists, which is the mechanism behind the co-adaptation loop of Section 4.3. Pooling forgoes exactly this term, since a shared adapter cannot move along $\Gamma_i - \bar{\Gamma}$, in exchange for the lower estimation variance of averaging GN responses.

5.3 From Co-Adaptation to Discovery Regret

We now show that adaptation converges, in the sense that the discovery regret in Eq. (1) of the model after C Coordinator–Scientist cycles decays geometrically in C to a floor. Let Ω be a convex set that almost surely contains every iterate of Algorithm 1 from the initial model $\Theta^{(0)}$, $\mathcal{H} = \{\Pi_\Theta : \Theta \in \Omega\}$, $F(\Theta) = J_{\tau,K}(\Theta)$ with $F^\star = \sup_\Omega F$, and $\varepsilon_{\mathrm{cov}} = s^\star - \sup_{\Pi \in \mathcal{H}} J_T(\Pi; \mathcal{P})$ the coverage error of $\mathcal{H}$. Unlike a run of Π_Θ, Algorithm 1 updates Θ between groups, performing $M = 2C$ block updates $\Theta^{(m+1)} = \Theta^{(m)} + \eta\, \widehat{\nabla}_m$ that alternate between the Coordinator block ϕ and the stacked Scientist block $\theta_{1:N}$, where $\widehat{\nabla}_m$ is the negative gradient of Eq. (6) with respect to the block b_m updated at step m, computed on the group collected at $\Theta^{(m)}$ and padded with zeros outside b_m, and $\mathcal{F}_m$ is the history before the group at step m is collected.

Assumption 5.5 (Curvature and estimation)**.** *(A1) ∇F is L-Lipschitz on Ω and $\frac{1}{2}\|\nabla F(\Theta)\|^2 \geq \mu\,(F^\star - F(\Theta))$ for all $\Theta \in \Omega$, with $\mu \leq L$. (A2) With $\bar{\nabla}_m = \mathbb{E}[\widehat{\nabla}_m \mid \mathcal{F}_m]$, each block estimate satisfies $\|\bar{\nabla}_m - \nabla_{\mathsf{b}_m} F(\Theta^{(m)})\| \leq \delta$ and $\mathbb{E}[\|\widehat{\nabla}_m - \bar{\nabla}_m\|^2 \mid \mathcal{F}_m] \leq \sigma^2$.*

Theorem 5.6 (Convergence of the discovery regret)**.** *Under Assumptions 5.1 and 5.5, for $0 < \eta \leq 1/(4L)$ and $q = 1 - \eta\mu/4$,*

$$\mathfrak{R}_T\big(\Pi_{\Theta^{(M)}}\big) = s^\star - \mathbb{E}\big[J_T(\Pi_{\Theta^{(M)}}; \mathcal{P})\big] \leq \varepsilon_{\mathrm{cov}} + \tau \log(K+1) + q^C \big(F^\star - F(\Theta^{(0)})\big) + \frac{5L\eta\sigma^2 + 8\delta^2}{\mu}. \tag{9}$$

Theorem 5.6 shows that the surrogate optimality gap $F^\star - F(\Theta^{(0)})$ decays geometrically in the number of Coordinator–Scientist cycles, with two block updates per cycle regardless of N because the Scientists update in parallel as one block, down to a floor set by the variance σ^2 of the group estimate, which averages G forks, and by its bias δ, which absorbs the one-round horizon, the implemented surrogate, and the mismatch between the snapshot of a group and the run objective (Appendix A.7.2). Adaptation does not reduce the coverage and smoothing errors, and Eq. (9) bounds the regret of a full-budget rollout of the adapted model, which Appendix A.7.3 extends, under a snapshot-mismatch assumption, to the interleaved MELD run, in which adaptation consumes part of the budget before the final model is rolled out.

6 Experiments

6.1 Experimental Setup

Tasks and metrics. We use 12 tasks from MLGym (Nathani et al., 2025) and 10 from MLE-bench (Chan et al., 2025) with private evaluators and a shared nominal budget of scored candidates per task (Appendix B). Table 1 reports the mean best held-out score over three seeds, and we aggregate

Table 1: Best held-out scores within the nominal evaluator-call budget on 12 MLGym and 10 MLE-bench tasks, averaged over three independent seeds. All baseline values are our same-harness reruns with the shared Executor and evaluator. Baseline is the starting solution of each task, and Human Best is our independently measured task-specific reference for MLGym (Appendix B.1) and the Kaggle top-1 score for MLE-bench. Dark and light purple mark the best and second-best automated methods.

		Prior Works				*Ours*	
Task	**Baseline**	**CORAL**	**AIRA**	**MLEvolve**	**ARTS***	**MELD**	**Human Best**
MLGym							
Titanic (acc ↑)	0.766	0.947	0.921	0.947	**0.952**	**0.972**	0.830
CIFAR-10 (acc ↑)	0.497	**0.962**	**0.963**	0.955	0.959	0.960	0.994
Fashion-MNIST (acc ↑)	0.848	0.943	**0.947**	0.937	0.945	**0.957**	0.968
House Prices (R^2 ↑)	0.880	0.924	0.929	0.929	**0.931**	**0.943**	0.990
MNLI (acc ↑)	52.51	**84.22**	84.19	84.00	84.13	**84.23**	92.50
Language Modeling (loss ↓)	4.238	**3.665**	3.682	3.766	3.672	**3.663**	3.280
Battle of the Sexes (payoff ↑)	1.023	**1.446**	1.445	1.445	**1.446**	**1.447**	1.667
Prisoner's Dilemma (payoff ↑)	2.372	2.637	2.636	**2.644**	2.642	**2.646**	3.000
Colonel Blotto (score ↑)	−0.248	**0.253**	**0.253**	0.252	0.252	**0.254**	0.500
MountainCar (reward ↑)	53.26	81.91	71.16	67.62	**84.53**	**93.51**	100.00
Breakout (reward ↑)	41.93	71.64	73.94	50.20	**78.46**	**84.97**	100.00
MetaMaze (reward ↑)	20.12	74.88	72.39	58.59	**75.41**	**81.19**	100.00
MLE-bench							
Spaceship Titanic (acc ↑)	0.7747	**0.8356**	0.8287	**0.8356**	**0.8352**	0.8310	0.828
Nomad 2018 (MCW-RMSLE ↓)	0.1086	0.0543	0.0599	0.0587	**0.0535**	**0.0501**	0.051
Jigsaw Toxic (CW-AUC ↑)	0.9712	0.9789	0.9797	**0.9803**	0.9801	**0.9805**	0.989
APTOS 2019 (QWK ↑)	0.7287	0.8407	0.8624	0.7811	**0.8625**	**0.8661**	0.936
Plant Pathology (MCW-AUC ↑)	0.9481	0.9527	0.9851	0.9915	**0.9921**	**0.9957**	0.984
Histopathologic Cancer (AUC ↑)	0.9463	0.9906	0.9874	**0.9914**	0.9908	**0.9961**	1.000
Vesuvius Ink ($F_{0.5}$ ↑)	0.3293	**0.3948**	0.3817	0.3701	0.3942	**0.4084**	0.831
CHAMPS (log-MAE ↓)	1.2353	1.2327	1.2256	1.2263	**1.2247**	**1.1592**	−3.240
HMS Brain (KL ↓)	0.8865	0.8270	0.7389	**0.7286**	0.7352	**0.5404**	0.272
RSNA Brain Tumor (AUC ↑)	0.5529	0.6647	0.6471	0.6482	**0.6709**	**0.7129**	0.621

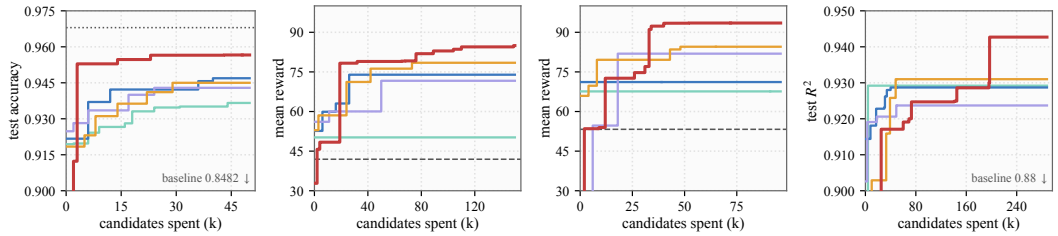
the 12 normalized MLGym scores (Appendix B.6) by Median, interquartile mean (IQM), Mean, and Optimality Gap (Agarwal et al., 2021).

Compared systems. We rerun AIRA (Toledo et al., 2025), MLEvolve (Du et al., 2026), and CORAL (Qu et al., 2026) in the same harness with frozen Claude Haiku 4.5 as their Scientist and as the shared Executor (Appendix B.4). MELD adapts LoRA adapters on a frozen Qwen3-4B-Instruct-2507 (Yang et al., 2025) backbone (Appendix B.3), so this system-level comparison tests whether task-time adaptation lets a compact open-weight Scientist compete with frozen proprietary Scientists. We also rerun ARTS* (Juneja et al., 2026), the test-time-training variant of ARTS, which fine-tunes a single tree-search Scientist on the same backbone.

6.2 MAIN RESULTS

Broad gains across task families. MELD achieves the best reported point estimate on 20 of 22 tasks, trailing only on CIFAR-10 (AIRA, by 0.003) and Spaceship Titanic (CORAL and MLEvolve, by 0.0046; Table 1). On MLGym, gains are generally modest on supervised-learning and game-theory tasks but reach 8.98, 6.51, and 5.78 points over ARTS*, the strongest baseline on MountainCar, Breakout, and MetaMaze, so the advantage spans task families but is largest on sequential control.

Early and delayed improvements. In Figure 1, MELD leads early on Fashion-MNIST, whereas on Breakout and MountainCar it overtakes the strongest early baselines later in the search. On Breakout and House Prices, it keeps improving long after every baseline has plateaued, overtaking ARTS* on House Prices only after two-thirds of the budget. Its gains are therefore not confined to the initial batch. These are search trajectories; the adaptation ablations below compare update configurations.



(a) Fashion-MNIST (b) Breakout (c) MountainCar (d) House Prices

Figure 1: Best-so-far held-out score versus scored candidates for MELD (red), ARTS* (orange), AIRA (blue), CORAL (purple), MLEvolve (green), the Baseline (dashed), and the Human Best (dotted, when visible).

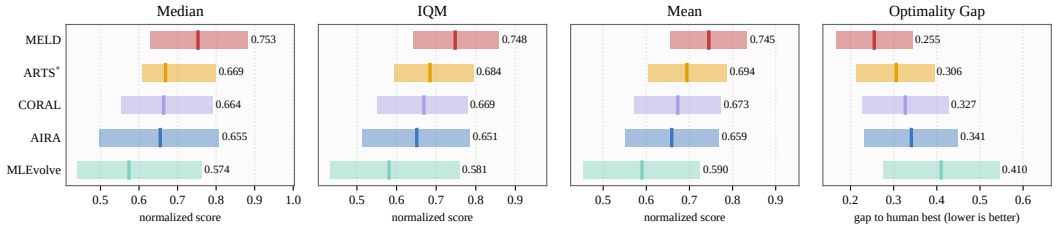


Figure 2: Median, IQM, Mean, and Optimality Gap of normalized scores (Baseline $= 0$, Human Best $= 1$) over the 12 MLGym tasks, with 95% bootstrap intervals. MELD is best on all four.

Aggregate performance and scope. All four aggregate statistics rank MELD first (Figure 2). Its Median, IQM, and Mean are 0.753, 0.748, and 0.745, versus 0.669, 0.684, and 0.694 for ARTS*, the strongest aggregate baseline. Optimality Gap decreases from 0.306 to 0.255 (about 17%). Median and IQM gains indicate that the advantage is not confined to the largest control-task improvements.

6.3 ABLATIONS

We test whether adapting both policy levels improves candidate generation beyond either update alone. On Iterated Prisoner's Dilemma (PD) and Colonel Blotto (CB), full MELD, Coordinator-only, Scientist-only, and Frozen each generate a fresh 24-candidate batch after one Coordinator–Scientist cycle (Appendix B.5). Over three seeds, full MELD raises the PD batch median from 2.35 to 2.57 and threshold hits from 2 to 10 per 24 attempts (Table 4), versus only 5 and 6 hits for Coordinator-only and Scientist-only. On CB, medians stay near zero, while full MELD raises hits from 2 to 6, versus 3, 4, and 2 for Coordinator-only, Scientist-only, and Frozen. Joint adaptation thus yields stronger post-cycle generation than either update alone in these experiments, improving median and hits on PD but only hits on CB, a gain that neither the batch median nor a cumulative best score reveals. With one research unit ($N{=}1$) at the same budget, MELD loses nearly half of its CHAMPS margin over ARTS*, as log-MAE rises from 1.1592 to 1.1899, so the full gain requires multiple units.

7 CONCLUSION

We introduced MELD, a hierarchical TTT framework that jointly adapts team-level Coordinator and agent-level Scientist policies on one unseen task. Updated policies generate the experience and feedback for later adaptation, forming a recursive self-improvement loop with team-level coordination and agent-level research exploration. Our analysis establishes objective alignment, characterizes conditions for coordination gains beyond independent parallel discovery and the first-order benefit of private adaptation, and bounds the discovery regret of adapted models. Across 12 MLGym and 10 MLE-bench tasks, MELD is best on 20 and achieves the highest aggregate normalized scores on MLGym with a 4B backbone. Current limitations are round-local credit and task-local adaptation within a fixed research hierarchy, and cross-task transfer could be natural future work.

AI USE STATEMENT

In this work, we used generative AI tools to assist in the writing and revision of proofs, implement methods, provide feedback on our experiments, and interpret results. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design the research methodology or experiments, or assist with translation. Generating synthetic data sets, cleaning and reformatting datasets, and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to identify and summarize relevant literature, suggest experimental parameters, create or edit analysis and plotting code, create or modify scientific figures, draft parts of the paper, edit the paper to improve readability, and format references. We have reviewed all AI-assisted work. The authors checked every AI-assisted proof step by step, verified every number reported in the paper against our experiment logs, and ran the AI-written plotting and statistics scripts and compared their outputs with the tables. Separately, LLMs are components of the systems evaluated in this paper, including the Qwen3-4B-Instruct-2507 backbone of MELD and the frozen Claude Haiku 4.5 Executor, and their roles are described in Section 6.1 and Appendix B. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Complete proofs of all theoretical results in Section 5 are given in Appendix A. Appendix B specifies the tasks, metrics, and reporting anchors (Appendix B.1), the private evaluators (Appendix B.2), the models, adapters, and adaptation hyperparameters (Appendix B.3), the baseline implementations (Appendix B.4), the ablation protocol (Appendix B.5), the computation of aggregate metrics (Appendix B.6), and the budget accounting, seeds, and compute (Appendix B.7).

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C. Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29304–29320, 2021.

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.662.

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *Algorithmic Learning Theory*, volume 5809 of *Lecture Notes in Computer Science*, pp. 23–37. Springer, 2009. doi: 10.1007/978-3-642-04414-4_7.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Lilian Weng, and Aleksander Mądry. MLE-bench: Evaluating machine learning agents on machine learning engineering. In *International Conference on Learning Representations*, 2025.

Shangheng Du, Xiangchao Yan, Jinxin Shi, Zongsheng Cao, Shiyang Feng, Zichen Liang, Boyuan Sun, Tianshuo Peng, Yifan Zhou, Xin Li, et al. MLEvolve: A self-evolving framework for automated machine learning algorithm discovery. *arXiv preprint arXiv:2606.06473*, 2026.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 2974–2982, 2018.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Shengran Hu, Cong Lu, and Jeff Clune. Automated design of agentic systems. In *International Conference on Learning Representations*, 2025.

Eric Hanchen Jiang, Levina Li, Rui Sun, Xiao Liang, Yubei Li, Yuchen Wu, Haozheng Luo, Hengli Li, Zhi Zhang, Zhaolu Kang, et al. Agent Q-Mix: Selecting the right action for LLM multi-agent systems through reinforcement learning. In *Conference on Language Modeling (COLM)*, 2026.

Kumar Joag-Dev and Frank Proschan. Negative association of random variables with applications. *The Annals of Statistics*, 11(1):286–295, 1983.

Gurusha Juneja, Arnav Kumar Jain, Deepak Nathani, William Yang Wang, and Xin Eric Wang. Learning the ARTS of search for automated discovery. *arXiv preprint arXiv:2606.21891*, 2026.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak–Łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases*, volume 9851 of *Lecture Notes in Computer Science*, pp. 795–811. Springer, 2016. doi: 10.1007/978-3-319-46128-1_50.

Tze Leung Lai and Herbert Robbins. Maximally dependent random variables. *Proceedings of the National Academy of Sciences*, 73(2):286–288, 1976.

Isaac Meilijson and Arthur Nádas. Convex majorization with an application to the length of critical paths. *Journal of Applied Probability*, 16(3):671–677, 1979.

Deepak Nathani, Lovish Madaan, Nicholas Roberts, Nikolay Bashlykov, Ajay Menon, Vincent Moens, Mikhail Plekhanov, Amar Budhiraja, Despoina Magka, Vladislav Vorotilov, et al. MLGym: A new framework and benchmark for advancing AI research agents. In *Conference on Language Modeling (COLM)*, 2025.

Andrew Y. Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pp. 278–287. Morgan Kaufmann, 1999.

Stefan Nielsen, Edoardo Cetin, Peter Schwendeman, Qi Sun, Jinglue Xu, and Yujin Tang. Learning to orchestrate agents in natural language with the Conductor. In *International Conference on Learning Representations*, 2026.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, et al. AlphaEvolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

Leonid Peshkin, Kee-Eung Kim, Nicolas Meuleau, and Leslie Pack Kaelbling. Learning to cooperate via policy search. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 489–496, 2000.

Ao Qu, Han Zheng, Zijian Zhou, Yihao Yan, Yihong Tang, Shao Yong Ong, Fenglu Hong, Kaichen Zhou, Chonghe Jiang, Minwei Kong, et al. CORAL: Towards autonomous multi-agent evolution for open-ended discovery. In *Conference on Language Modeling (COLM)*, 2026.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4295–4304. PMLR, 2018.

Kainat Riaz, Muhammad Ahmed Mohsin, Ahsan Bilal, Muhammad Umer, Ayesha Mohsin, Aqib Riaz, Ali Subhan, and John M. Cioffi. Epistemic uncertainty for test-time discovery. *arXiv preprint arXiv:2605.11328*, 2026.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, et al. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Qi-Man Shao. A comparison theorem on moment inequalities between negatively associated and independent random variables. *Journal of Theoretical Probability*, 13(2):343–356, 2000.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, 2025.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9229–9248. PMLR, 2020.

Edan Toledo, Karen Hambardzumyan, Martin Josifoski, Rishi Hazra, Nicolas Baldwin, Alexis Audran-Reiss, Michael Kuchnik, Despoina Magka, Minqi Jiang, Alisia Lupidi, et al. AI research agents for machine learning: Search, exploration, and generalization in MLE-bench. In *Advances in Neural Information Processing Systems*, volume 38, pp. 35309–35348, 2025.

Yuvraj Virk, Zack Edds, Chunqiu Steven Xia, and Lingming Zhang. SwarmResearch: Orchestrating coding agents for open-ended discovery. *arXiv preprint arXiv:2607.02807*, 2026.

Shijie Wang, Pengfei Li, Yikun Fu, Kaifeng Liu, Fangyuan Li, Yang Liu, Xiaowei Sun, Zonglin Li, Siyao Zhao, Jian Zhao, et al. MARTI-MARS^2: Scaling multi-agent self-search via reinforcement learning for code generation. *arXiv preprint arXiv:2602.07848*, 2026a.

Tonghan Wang, Heng Dong, Victor Lesser, and Chongjie Zhang. ROMA: Multi-agent reinforcement learning with emergent roles. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9876–9886. PMLR, 2020.

Yiping Wang, Shao-Rong Su, Zhiyuan Zeng, Eva Xu, Liliang Ren, Xinyu Yang, Zeyi Huang, Xuehai He, Luyao Ma, Baolin Peng, et al. ThetaEvolve: Test-time learning on open problems. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026b.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, 1992. doi: 10.1007/BF00992696.

Chen Xu, Yicheng Hu, Ruizi Wang, Xinyu Lin, Wenjie Wang, Dongrui Liu, and Fuli Feng. TacoMAS: Test-time co-evolution of topology and capability in LLM-based multi-agent systems. *arXiv preprint arXiv:2605.09539*, 2026.

Minghao Yan, Bo Peng, Benjamin Coleman, Ziqi Chen, Zhouhang Xie, Shuo Chen, Zhankui He, Noveen Sachdeva, Weili Wang, Ed H. Chi, et al. PACEvolve++: Improving test-time learning for evolutionary search agents. *arXiv preprint arXiv:2605.07039*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Mert Yuksekgonul, Daniel Koceja, Xinhao Li, Federico Bianchi, Jed McCaleb, Xiaolong Wang, Jan Kautz, Yejin Choi, James Zou, Carlos Guestrin, et al. Learning to discover at test time. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Yaolun Zhang, Yujie Zhao, Nan Wang, Yiran Wu, Jiayu Chang, Yizhao Chen, Qingyun Wu, Jishen Zhao, and Huazheng Wang. MetaAgent-X: Breaking the ceiling of automatic multi-agent systems via end-to-end reinforcement learning. *arXiv preprint arXiv:2605.14212*, 2026.

Mingchen Zhuge, Wenyi Wang, Louis Kirsch, Francesco Faccio, Dmitrii Khizbullin, and Jürgen Schmidhuber. GPTSwarm: Language agents as optimizable graphs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62743–62767. PMLR, 2024.

A PROOFS FOR SECTION 5

This appendix contains complete proofs of Proposition 5.2, Theorem 5.3, Proposition 5.4, and Theorem 5.6, including a two-block ascent rate with biased estimates (Appendix A.7.2), and the exact-gap and frontier-seeking properties of the potential (Appendix A.4). We first fix notation and collect elementary facts about the potential that are used throughout.

A.1 NOTATION AND ELEMENTARY PROPERTIES OF THE POTENTIAL

Matrices and blocks. For matrices A, B of equal shape, $\langle A, B\rangle_F = \mathrm{tr}(A^\top B)$ and $\|A\|_F^2 = \langle A, A\rangle_F$. When a parameter block consists of several matrices, inner products and squared norms are summed over the matrices, and every identity below holds matrix by matrix. The trainable blocks are $\beta \in \{\mathrm{coord}, \mathrm{sci}_1, \ldots, \mathrm{sci}_N\}$ with policies π_β, actions a^β, and contexts x^β as in Section 4.2: $a_t^{\mathrm{coord}} = Z_t$, $x_t^{\mathrm{coord}} = \mathcal{S}_t$, and $a_t^{\mathrm{sci}_i} = u_t^i$, $x_t^{\mathrm{sci}_i} = (\mathcal{S}_t, z_t^i)$.

States and the potential. A state $\mathcal{S}$ carries the derivation tree with node set $V(\mathcal{S})$ and utilities $s_v \in [0,1]$; the root carries s_0, every other node is a validated candidate, and $B(\mathcal{S}) = \max_{v\in V(\mathcal{S})} s_v$. Since every validated candidate consumed one evaluator call, $n_{\mathrm{val}}(\mathcal{S}) := |V(\mathcal{S})| - 1 \le K$, and the number of padded slots

$$n_{\mathrm{pad}}(\mathcal{S}) := K + 1 - |V(\mathcal{S})| = K - n_{\mathrm{val}}(\mathcal{S}) \;\ge\; 0 \tag{10}$$

counts evaluator calls that failed or have not been used. Failed calls therefore do not appear in $V(\mathcal{S})$ but are accounted for by Eq. (10). With $s_{\min} = 0$, define

$$\mathcal{A}(\mathcal{S}) = \sum_{v\in V(\mathcal{S})} e^{s_v/\tau} + n_{\mathrm{pad}}(\mathcal{S})\, e^{s_{\min}/\tau}, \qquad \Phi_{\tau,K}(\mathcal{S}) = \tau \log \frac{\mathcal{A}(\mathcal{S})}{K+1}, \tag{11}$$

which is Eq. (4); $\mathcal{A}(\mathcal{S})$ is a sum of exactly $K+1$ positive terms, one per evaluator slot including the root. For a state $\mathcal{S}'$ and a collection X of new validated candidates, $\mathcal{S}' \oplus X$ denotes the state extended by their call-indexed nodes; repeated utilities remain separate entries. Failed calls leave n_{pad} unchanged because it counts failed and unused slots together.

Lemma A.1 (Slot bookkeeping)**.** *Let $\mathcal{S}'$ be a state and X the validated candidates produced within its remaining evaluator-call budget. Then*

$$\mathcal{A}(\mathcal{S}' \oplus X) = \mathcal{A}(\mathcal{S}') + \sum_{c\in X} \left(e^{s(c)/\tau} - e^{s_{\min}/\tau}\right) \;\ge\; \mathcal{A}(\mathcal{S}'). \tag{12}$$

In particular, failed calls leave $\mathcal{A}$ unchanged, and for a single candidate c with utility s,

$$\Phi_{\tau,K}(\mathcal{S}' \oplus \{c\}) - \Phi_{\tau,K}(\mathcal{S}') = \tau \log\Big(1 + \frac{e^{s/\tau} - e^{s_{\min}/\tau}}{\mathcal{A}(\mathcal{S}')}\Big). \tag{13}$$

Proof. Each validated candidate $c \in X$ adds the node term $e^{s(c)/\tau}$ to the first sum in Eq. (11) and, having consumed one evaluator call, reduces n_{pad} by one, removing one term $e^{s_{\min}/\tau}$; a failed call reduces neither $|V|$ nor the number of remaining slots as counted in Eq. (10), since n_{pad} counts failed and unused slots alike. This gives the equality in Eq. (12). The inequality follows from $s(c) \ge 0 = s_{\min}$. Dividing Eq. (12) with $X = \{c\}$ by $\mathcal{A}(\mathcal{S}')$ and taking $\tau \log$ gives Eq. (13). □

Lemma A.2 (Monotone, telescoping increments)**.** *Along any run, the round rewards $r_t^\Phi = \Phi_{\tau,K}(\mathcal{S}_{t+1}) - \Phi_{\tau,K}(\mathcal{S}_t)$ satisfy $r_t^\Phi \ge 0$, and for any $0 \le t_1 \le t_2 \le R$,*

$$\sum_{t'=t_1}^{t_2-1} r_{t'}^\Phi = \Phi_{\tau,K}(\mathcal{S}_{t_2}) - \Phi_{\tau,K}(\mathcal{S}_{t_1}); \qquad \textit{in particular} \qquad \sum_{t'=0}^{R-1} r_{t'}^\Phi = \Phi_{\tau,K}(\mathcal{S}_{\mathrm{end}}) - \Phi_{\tau,K}(\mathcal{S}_0). \tag{14}$$

Moreover, $R_{t:\mathrm{end}}^\Phi = r_t^\Phi + R_{t+1:\mathrm{end}}^\Phi$ for every round t.

Proof. Nonnegativity follows from Lemma A.1 applied to the validated candidates of round t, since log is increasing. The sums telescope term by term. The last identity is the case $t_1 = t$, $t_2 = t+1$ combined with the definition $R_{t:\mathrm{end}}^\Phi = \Phi_{\tau,K}(\mathcal{S}_{\mathrm{end}}) - \Phi_{\tau,K}(\mathcal{S}_t)$. □

A.2 Trajectory Factorization and Regularity

A run of Π_Θ with $\Theta = (\phi, \theta_{1:N})$ held fixed produces the trajectory

$$\zeta = (\mathcal{S}_0, Z_0, U_0, \mathcal{S}_1, Z_1, U_1, \ldots, Z_{R-1}, U_{R-1}, \mathcal{S}_R), \qquad \mathcal{S}_R = \mathcal{S}_{\text{end}}, \tag{15}$$

where $R \le K$ is the number of nonempty rounds until the budget is exhausted, $Z_t = \{z_t^i\}_{i \in \mathcal{I}_t}$, and $U_t = \{u_t^i\}_{i \in \mathcal{I}_t}$. Let $\mathcal{Q}(\mathcal{S}_{t+1} \mid \mathcal{S}_t, Z_t, U_t)$ be the kernel comprising execution by the frozen Executor, evaluation, the tree update of Section 3, and budget accounting, and let ν be the law of the initial state. By the factorization of Eq. (3),

$$p_\Theta(\zeta) = \nu(\mathcal{S}_0) \prod_{t=0}^{R-1} \Big[\pi_\phi^{\text{coord}}(Z_t \mid \mathcal{S}_t) \prod_{i \in \mathcal{I}_t} \pi_{\theta_i}^{\text{sci}}(u_t^i \mid \mathcal{S}_t, z_t^i)\ \mathcal{Q}(\mathcal{S}_{t+1} \mid \mathcal{S}_t, Z_t, U_t) \Big]. \tag{16}$$

Throughout we work under Assumption 5.1, whose conditions (R1)–(R3) refer to the factorization in Eq. (16): (R1) ν and $\mathcal{Q}$ do not depend on Θ; (R2) each policy density is differentiable in its own block with parameter-independent support; (R3) differentiation with respect to any block may be interchanged with the expectation over ζ.

Remark A.3 (The assumptions for LLM policies)**.** *At a fixed context, bounded-length token actions over a finite vocabulary have finite support. Softmax probabilities on a parameter-independent admissible support are differentiable, giving (R2); any decoding mask is held fixed when taking derivatives. For the full trajectory law, however, contexts depend on random execution feedback, so finite action spaces alone do not imply (R3). A sufficient condition is an integrable envelope $D_\beta(\zeta)$ such that $\sup_{\Theta \in \mathcal{U}} \|\nabla_\beta p_\Theta(\zeta)\| \le D_\beta(\zeta)$ in a neighborhood $\mathcal{U}$ of the differentiated model. The bounded potential then permits dominated differentiation. We impose (R3) on the trajectory law and its fixed-snapshot conditional laws. Assumption (R1) records the parameter independence of the Executor, evaluator, and state transition given actions.*

We define the filtrations under which actions are drawn. Let $\mathcal{E}_t = \sigma(\mathcal{S}_0, Z_0, U_0, \ldots, \mathcal{S}_t)$ be the history up to and including $\mathcal{S}_t$, and $\mathcal{E}_t^+ = \mathcal{E}_t \vee \sigma(Z_t)$ the history after the portfolio of round t is drawn. For $\beta = \text{coord}$ write $\mathcal{F}_t^\beta = \mathcal{E}_t$; for $\beta = \text{sci}_i$ write $\mathcal{F}_t^\beta = \mathcal{E}_t^+$. In both cases x_t^β is $\mathcal{F}_t^\beta$-measurable, and conditionally on $\mathcal{F}_t^\beta$ the action is drawn as $a_t^\beta \sim \pi_\beta(\cdot \mid x_t^\beta)$; for $\beta = \text{sci}_i$ this uses that, given $\mathcal{S}_t$ and Z_t, the responses of different Scientists are conditionally independent by Eq. (3).

Lemma A.4 (Score identity)**.** *Under (R2), for every block β and every context x,*

$$\mathbb{E}_{a \sim \pi_\beta(\cdot \mid x)} \big[\nabla_\beta \log \pi_\beta(a \mid x) \big] = 0. \tag{17}$$

Proof. Let $\mathcal{X}$ be the parameter-independent support of $\pi_\beta(\cdot \mid x)$, which is finite for bounded-length token actions (Remark A.3). Then $\mathbb{E}[\nabla_\beta \log \pi_\beta(a \mid x)] = \sum_{a \in \mathcal{X}} \pi_\beta(a \mid x)\, \frac{\nabla_\beta \pi_\beta(a|x)}{\pi_\beta(a|x)} = \sum_{a \in \mathcal{X}} \nabla_\beta \pi_\beta(a \mid x) = \nabla_\beta \sum_{a \in \mathcal{X}} \pi_\beta(a \mid x) = \nabla_\beta 1 = 0$, where the interchange of ∇_β and the finite sum uses only (R2), since $\mathcal{X}$ does not move with the parameters. □

Lemma A.5 (Baseline lemma)**.** *Let β act at round t, and let χ be an integrable, $\mathcal{F}_t^\beta$-measurable random variable such that $\chi\, \nabla_\beta \log \pi_\beta(a_t^\beta \mid x_t^\beta)$ is integrable. Then $\mathbb{E}\big[\chi\, \nabla_\beta \log \pi_\beta(a_t^\beta \mid x_t^\beta)\big] = 0$.*

Proof. By the tower property and $\mathcal{F}_t^\beta$-measurability of χ and of x_t^β, $\mathbb{E}[\chi\, \nabla_\beta \log \pi_\beta(a_t^\beta \mid x_t^\beta)] = \mathbb{E}\big[\chi\, \mathbb{E}[\nabla_\beta \log \pi_\beta(a_t^\beta \mid x_t^\beta) \mid \mathcal{F}_t^\beta]\big]$, and the inner conditional expectation is $\mathbb{E}_{a \sim \pi_\beta(\cdot \mid x_t^\beta)}[\nabla_\beta \log \pi_\beta(a \mid x_t^\beta)] = 0$ by Lemma A.4. □

Lemma A.6 (Block score decomposition)**.** *Under (R1)–(R2), for every block β and every trajectory ζ in the support of p_Θ,*

$$\nabla_\beta \log p_\Theta(\zeta) = \sum_{t:\ \beta \text{ acts at } t} \nabla_\beta \log \pi_\beta(a_t^\beta \mid x_t^\beta), \tag{18}$$

where coord *acts at every round and* sci_i *acts at the rounds with $i \in \mathcal{I}_t$.*

Proof. Taking logarithms in Eq. (16), $\log p_\Theta(\zeta) = \log \nu(\mathcal{S}_0) + \sum_t \big[\log \pi_\phi^{\text{coord}}(Z_t \mid \mathcal{S}_t) + \sum_{i\in\mathcal{I}_t} \log \pi_{\theta_i}^{\text{sci}}(u_t^i \mid \mathcal{S}_t, z_t^i) + \log \mathcal{Q}(\mathcal{S}_{t+1} \mid \mathcal{S}_t, Z_t, U_t)\big]$. By (R1) the terms $\log\nu$ and $\log\mathcal{Q}$ have zero gradient with respect to every block. The coordination factors depend only on ϕ and the Scientist-i factors only on θ_i, so differentiating with respect to ϕ keeps exactly the coordination terms, and differentiating with respect to θ_i keeps exactly the terms with $i \in \mathcal{I}_t$. □

A.3 Proof of Proposition 5.2

Proof of Proposition 5.2. Part (i). Write $B = B(\mathcal{S})$. Factoring $e^{B/\tau}$ out of every term of Eq. (11) gives $\mathcal{A}(\mathcal{S}) = e^{B/\tau} K_{\text{eff}}(\mathcal{S})$ with $K_{\text{eff}}(\mathcal{S}) = \sum_{v\in V(\mathcal{S})} e^{(s_v - B)/\tau} + n_{\text{pad}}(\mathcal{S})\, e^{(s_{\min}-B)/\tau}$, the number of evaluator slots counted in frontier units: a node contributes 1 if it attains B and $e^{-d/\tau}$ if it trails the frontier by d, and a sentinel ($s_{\min} = 0$) contributes $e^{-B/\tau}$. Hence $\Phi_{\tau,K}(\mathcal{S}) = \tau \log \frac{\mathcal{A}(\mathcal{S})}{K+1} = B - \tau \log \frac{K+1}{K_{\text{eff}}(\mathcal{S})}$. For the range of K_{eff}, the root belongs to $V(\mathcal{S})$, so $B \geq s_0 \geq 0 = s_{\min}$; every exponent $s_v - B$ and $s_{\min} - B$ is therefore at most 0, each of the $K+1$ terms lies in $(0, 1]$, and the node attaining B contributes exactly 1. Thus $1 \leq K_{\text{eff}}(\mathcal{S}) \leq K+1$, which gives the two-sided bound $B - \tau\log(K+1) \leq \Phi_{\tau,K}(\mathcal{S}) \leq B$; the identity above is its exact form and is used in Appendix A.4. In particular $\Phi_{\tau,K}(\mathcal{S}) \in [-\tau\log(K+1), 1]$ because $0 \leq B \leq 1$.

For the expectations, every node of $\mathcal{S}_{\text{end}}$ other than the root is a validated candidate produced within the budget, and every such candidate is a node; hence $V(\mathcal{S}_{\text{end}}) \setminus \{\text{root}\} = \mathcal{C}_T^{\text{val}}(\Pi_\Theta)$ and

$$B(\mathcal{S}_{\text{end}}) = \max_{c\in\{c_0\}\cup\mathcal{C}_T^{\text{val}}(\Pi_\Theta)} s(c), \qquad J_T(\Pi_\Theta; \mathcal{P}) = \mathbb{E}_{\zeta\sim\Pi_\Theta}[B(\mathcal{S}_{\text{end}})]. \tag{19}$$

Taking expectations in the two-sided bound at $\mathcal{S} = \mathcal{S}_{\text{end}}$ gives part (i). We next prove part (ii), then part (iv), and finally part (iii), which uses the local-gradient identity established in part (iv).

Part (ii). Fix a block β. By (R3) and the log-derivative identity $\nabla_\beta p_\Theta = p_\Theta \nabla_\beta \log p_\Theta$ on the support,

$$\nabla_\beta J_{\tau,K}(\Theta) = \int \Phi_{\tau,K}(\mathcal{S}_{\text{end}}(\zeta))\, \nabla_\beta p_\Theta(\zeta)\, d\zeta = \mathbb{E}_{\zeta\sim\Pi_\Theta}\big[\Phi_{\tau,K}(\mathcal{S}_{\text{end}})\, \nabla_\beta \log p_\Theta(\zeta)\big]. \tag{20}$$

Substituting Lemma A.6,

$$\nabla_\beta J_{\tau,K}(\Theta) = \sum_t \mathbb{E}\Big[\mathbf{1}\{\beta \text{ acts at } t\}\, \Phi_{\tau,K}(\mathcal{S}_{\text{end}})\, \nabla_\beta \log \pi_\beta(a_t^\beta \mid x_t^\beta)\Big]. \tag{21}$$

For each round t at which β acts, write $\Phi_{\tau,K}(\mathcal{S}_{\text{end}}) = \Phi_{\tau,K}(\mathcal{S}_t) + R_{t:\text{end}}^\Phi$ (Lemma A.2). The random variable $\Phi_{\tau,K}(\mathcal{S}_t)$ is bounded (part (i)) and $\mathcal{F}_t^\beta$-measurable, and the event $\{\beta \text{ acts at } t\}$ is $\mathcal{F}_t^\beta$-measurable as well: for the Coordinator it is $\{t < R\}$, a function of $\mathcal{S}_t$, and for Scientist i it is $\{i \in \mathcal{I}_t\}$, a function of Z_t. Applying Lemma A.5 with $\chi = \mathbf{1}\{\beta \text{ acts at } t\}\, \Phi_{\tau,K}(\mathcal{S}_t)$ removes the first term:

$$\mathbb{E}\big[\mathbf{1}\{\beta \text{ acts at } t\}\, \Phi_{\tau,K}(\mathcal{S}_t)\, \nabla_\beta \log \pi_\beta(a_t^\beta \mid x_t^\beta)\big] = 0. \tag{22}$$

What remains of Eq. (21) is Eq. (7). The same argument shows that any $\mathcal{F}_t^\beta$-measurable baseline $b_t(x_t^\beta)$ may be subtracted from $R_{t:\text{end}}^\Phi$ without changing the expectation. Finally, by Lemma A.2, $R_{t:\text{end}}^\Phi = \sum_{t'\geq t} r_{t'}^\Phi$, which is the reward-to-go form of the same gradient.

Part (iv). Fix a snapshot $\mathcal{S}$, a block β with context x, and the behavior model Θ_{old} at which the group is collected; for a Scientist group also condition on the full fixed portfolio Z. All gradient identities in this part are evaluated at $\beta = \beta_{\text{old}}$, with the other blocks held fixed. The $G \geq 2$ forks use independent random draws from this snapshot; within fork g, block β draws $a_g \sim \pi_\beta(\cdot \mid x)$, all other decisions of the round are drawn from the behavior model (for $\beta = \text{coord}$, every Scientist responds to the sampled portfolio from $\pi_{\theta_{i,\text{old}}}^{\text{sci}}$; for $\beta = \text{sci}_i$, the portfolio is fixed and the other Scientists respond from $\pi_{\theta_{j,\text{old}}}^{\text{sci}}$), the round is executed and evaluated through $\mathcal{Q}$, producing the candidate set $X^{(g)}$, and

$$r_g^\Phi = \Phi_{\tau,K}\big(\mathcal{S} \oplus X^{(g)}\big) - \Phi_{\tau,K}(\mathcal{S}). \tag{23}$$

By independence of the fork draws, the pairs $(a_g, r_g^\Phi)_{g=1}^G$ are independent and identically distributed conditionally on the fixed group context. Define the conditional mean increment

$$\bar{r}(a) = \mathbb{E}\big[r_g^\Phi \,\big|\, a_g = a\big], \tag{24}$$

which does not depend on β's parameters: by (R1), given a the law of $X^{(g)}$ is determined by the behavior model of the other blocks and by $\mathcal{Q}$. By the definition in Eq. (5), $J_\beta^{\mathrm{loc}}(\mathcal{S}) = \sum_a \pi_\beta(a \mid x)\,\bar{r}(a)$, and therefore

$$\begin{aligned}\nabla_\beta J_\beta^{\mathrm{loc}}(\mathcal{S}) &= \sum_a \nabla_\beta \pi_\beta(a \mid x)\,\bar{r}(a) = \sum_a \pi_\beta(a \mid x)\,\nabla_\beta \log \pi_\beta(a \mid x)\,\bar{r}(a) \\ &= \mathbb{E}\big[\bar{r}(a_g)\,\nabla_\beta \log \pi_\beta(a_g \mid x)\big].\end{aligned} \tag{25}$$

By the tower property, $\mathbb{E}[r_g^\Phi \nabla_\beta \log \pi_\beta(a_g \mid x)] = \mathbb{E}\big[\mathbb{E}[r_g^\Phi \mid a_g]\,\nabla_\beta \log \pi_\beta(a_g \mid x)\big] = \mathbb{E}[\bar{r}(a_g)\nabla_\beta \log \pi_\beta(a_g \mid x)]$, so

$$\nabla_\beta J_\beta^{\mathrm{loc}}(\mathcal{S}) = \mathbb{E}\big[r_g^\Phi\,\nabla_\beta \log \pi_\beta(a_g \mid x)\big]. \tag{26}$$

Now consider the leave-one-out baseline $b_g = \frac{1}{G-1}\sum_{g' \neq g} r_{g'}^\Phi$. It is a function of the forks other than g and is therefore independent of a_g; moreover it is bounded, since each $r_{g'}^\Phi \in [0, 1 + \tau \log(K+1)]$ by Lemma A.2 and part (i). Hence

$$\mathbb{E}\big[b_g\,\nabla_\beta \log \pi_\beta(a_g \mid x)\big] = \mathbb{E}[b_g]\,\mathbb{E}\big[\nabla_\beta \log \pi_\beta(a_g \mid x)\big] = \mathbb{E}[b_g] \cdot 0 = 0 \tag{27}$$

by Lemma A.4. Since $\widehat{A}_g = r_g^\Phi - b_g$, subtracting Eq. (27) from Eq. (26) gives $\mathbb{E}[\widehat{A}_g \nabla_\beta \log \pi_\beta(a_g \mid x)] = \nabla_\beta J_\beta^{\mathrm{loc}}(\mathcal{S})$; averaging over g, the group estimator $\frac{1}{G}\sum_g \widehat{A}_g \nabla_\beta \log \pi_\beta(a_g \mid x)$ is unbiased for the same quantity.

Part (iii). Take $\mathcal{S} = \mathcal{S}_t$ in a run of Π_Θ. By Lemma A.2, the round-t summand of Eq. (7) splits as

$$\mathbb{E}\big[R_{t:\mathrm{end}}^\Phi \nabla_\beta \log \pi_\beta(a_t^\beta \mid x_t^\beta)\big] = \underbrace{\mathbb{E}\big[r_t^\Phi\,\nabla_\beta \log \pi_\beta(a_t^\beta \mid x_t^\beta)\big]}_{\text{current-round term}} + \underbrace{\mathbb{E}\big[R_{t+1:\mathrm{end}}^\Phi \nabla_\beta \log \pi_\beta(a_t^\beta \mid x_t^\beta)\big]}_{\text{downstream term}}. \tag{28}$$

Conditioning the current-round term on $\mathcal{F}_t^\beta$ places us exactly in the setting of Eq. (26) with snapshot $\mathcal{S}_t$, context x_t^β, and the other blocks at the parameters Θ under which the run is generated (which is the behavior model within a group); hence the current-round term equals $\mathbb{E}[\mathbf{1}\{\beta \text{ acts at } t\}\nabla_\beta J_\beta^{\mathrm{loc}}(\mathcal{S}_t)]$, and summing over t gives the current-round sum stated in part (iii). The downstream terms, those carrying $R_{t+1:\mathrm{end}}^\Phi$, are not zero in general, because $R_{t+1:\mathrm{end}}^\Phi$ depends on a_t^β through $\mathcal{S}_{t+1}$, and this is the bias accepted in Section 4.2. In Algorithm 1 the G forks of a group are committed together, so the state actually reached is $\mathcal{S} \oplus \bigcup_g X^{(g)}$ rather than a single fork's successor; the estimator targets $\nabla_\beta J_\beta^{\mathrm{loc}}(\mathcal{S})$ at the shared pre-commit snapshot, as stated, and by concavity of $\tau \log(1 + \cdot/\mathcal{A}(\mathcal{S}))$ the merged increment is at most the sum of the fork increments. □

A.3.1 The Implemented Surrogate: Rejection, Length Normalization, and Standardization

Rejection of malformed responses. The judge discards malformed hypotheses and regenerates them without consuming a scored candidate (Appendix B.3). Let $\alpha(a) \in [0,1]$ be the probability that the judge accepts a, which does not depend on the parameters (a deterministic judge is the case $\alpha \in \{0,1\}$), so the effective policy of a block is the accept-conditioned law $\tilde{\pi}_\beta(a \mid x) = \pi_\beta(a \mid x)\alpha(a)/P_\beta(\mathrm{acc} \mid x)$ with $P_\beta(\mathrm{acc} \mid x) = \sum_a \pi_\beta(a \mid x)\alpha(a)$, whose support $\{\alpha > 0\} \cap \mathcal{X}$ is parameter independent; (R2) and all statements above hold for $\tilde{\pi}_\beta$. The implemented score, however, is $\nabla_\beta \log \pi_\beta(a \mid x) = \nabla_\beta \log \tilde{\pi}_\beta(a \mid x) + \nabla_\beta \log P_\beta(\mathrm{acc} \mid x)$, whose second term is constant in a and equals $\mathbb{E}_{\tilde{\pi}_\beta}[\nabla_\beta \log \pi_\beta(a \mid x)]$. In the leave-one-out estimator this term is multiplied by $\widehat{A}_g = r_g^\Phi - b_g$ with $\mathbb{E}[b_g] = \mathbb{E}[r_g^\Phi]$, so its contribution has expectation $\big(\mathbb{E}[r_g^\Phi] - \mathbb{E}[b_g]\big)\nabla_\beta \log P_\beta(\mathrm{acc} \mid x) = 0$; part (iv) therefore holds verbatim for the accept-conditioned objective. The un-baselined identity in Eq. (26) does not: it acquires the bias $\mathbb{E}[r_g^\Phi]\nabla_\beta \log P_\beta(\mathrm{acc} \mid x)$, which is one more reason for the baseline.

Relation to token-length normalization. The estimator in part (iv) uses the score of the entire generated action. Let $L(a) \geq 1$ be its response length and $h_\beta(a) = \nabla_\beta \log \pi_\beta(a \mid x) = \sum_{j=1}^{L(a)} \nabla_\beta \log \pi_\beta(y_j \mid x, y_{<j})$. With advantages held fixed, at the behavior model the negative

gradient of the unregularized, unclipped loss in Eq. (6) is

$$\widehat{g}_{\mathrm{tok}} = \frac{1}{G}\sum_{g=1}^{G} \frac{\widehat{A}_g}{L(a_g)} h_\beta(a_g), \qquad \widehat{g}_{\mathrm{seq}} = \frac{1}{G}\sum_{g=1}^{G} \widehat{A}_g h_\beta(a_g). \tag{29}$$

Indeed, each importance ratio equals one there, and its derivative is the corresponding token score. Condition throughout on the snapshot, the behavior model, and, for a Scientist group, the full fixed portfolio. Write $\bar{r}(a) = \mathbb{E}[r_g^\Phi \mid a_g = a]$ and $\bar{r}_0 = \mathbb{E}[r_g^\Phi]$. Since the leave-one-out baseline is independent of a_g and has mean $\bar{r}_0$,

$$\mathbb{E}[\widehat{g}_{\mathrm{tok}}] = \mathbb{E}_{a\sim\pi_{\beta,\mathrm{old}}}\left[\frac{\bar{r}(a)-\bar{r}_0}{L(a)}\, h_\beta(a)\right]_{\beta=\beta_{\mathrm{old}}}. \tag{30}$$

In particular, the baseline contribution is $\bar{r}_0\mathbb{E}[h_\beta(a)/L(a)]$, which need not vanish for action-dependent lengths. If $L(a) = L_0$ throughout the support, then $\mathbb{E}[\widehat{g}_{\mathrm{tok}}] = L_0^{-1}\nabla_\beta J_\beta^{\mathrm{loc}}(\mathcal{S})$; otherwise part (iv) establishes unbiasedness for $\widehat{g}_{\mathrm{seq}}$, not for $\widehat{g}_{\mathrm{tok}}$. The implementation further divides $\widehat{A}_g$ by the within-group standard deviation of $\{r_{g'}^\Phi\}_{g'=1}^G$ (Appendix B.3). This factor is a function of all G forks, including fork g, so the resulting estimator is a data-dependent rescaling of $\widehat{g}_{\mathrm{tok}}$ whose expectation is likewise not covered by part (iv); as G grows the factor concentrates around the population standard deviation and the rescaling becomes a constant. Length normalization, advantage standardization, clipping, and KL regularization thus define the implemented surrogate; none of them alters the potential alignment of part (i) or the block structure of part (ii).

A.4 Exact Surrogate Gap and Frontier-Seeking Increments

The proof of Proposition 5.2(i) shows that $\Phi_{\tau,K}(\mathcal{S}) = B(\mathcal{S}) - \tau\log\frac{K+1}{K_{\mathrm{eff}}(\mathcal{S})}$ with $K_{\mathrm{eff}}(\mathcal{S}) \in [1, K+1]$ the number of slots counted in frontier units: the surrogate undercounts the best utility by exactly τ times the log-ratio of nominal to frontier-equivalent slots, most when a single node stands far above every other slot and not at all when all slots tie. The same bookkeeping characterizes the per-round increment earned by one candidate.

Proposition A.7 (Frontier-seeking increments)**.** *Let $\mathcal{S}'$ be a state with $B = B(\mathcal{S}')$ and $K_{\mathrm{eff}} = K_{\mathrm{eff}}(\mathcal{S}')$, and let c be a validated candidate of utility s produced by one of its remaining evaluator calls. Then*

$$\Phi_{\tau,K}(\mathcal{S}' \oplus \{c\}) - \Phi_{\tau,K}(\mathcal{S}') = \tau\log\Big(1 + \frac{e^{(s-B)/\tau} - e^{-B/\tau}}{K_{\mathrm{eff}}}\Big), \tag{31}$$

which is strictly increasing in s. If $s \le B$, the increment is at most $\tau\log(1+1/K_{\mathrm{eff}}) \le \tau\log 2$; if $s = B+\omega$ with $\omega > 0$, it is at least $\tau\log\big(1+(e^{\omega/\tau}-1)/K_{\mathrm{eff}}\big)$, which exceeds the former bound as soon as $\omega > \tau\log 2$ and is at least $\omega - \tau\log(2K_{\mathrm{eff}})$ whenever $\omega \ge \tau\log 2$.

At any state the credit for a candidate is thus strictly increasing in its realized utility: tying or trailing the frontier earns a bounded, τ-scaled reward, whereas advancing it by ω earns a reward within $\tau\log 2$ of ω when few nodes sit near the frontier ($K_{\mathrm{eff}} \approx 1$) and less in a crowded frontier, where repeated confirmation of a known direction is discounted; the temperature interpolates between the maximum ($\tau \to 0$) and the average utility over slots ($\tau \to \infty$). Increments of B itself would vanish on every round without a new best; the potential yields a dense signal that still rewards the frontier most.

Proof of Proposition A.7. Let $\mathcal{A} = \mathcal{A}(\mathcal{S}')$, $B = B(\mathcal{S}')$, and $K_{\mathrm{eff}} = K_{\mathrm{eff}}(\mathcal{S}')$, so that $\mathcal{A} = e^{B/\tau}K_{\mathrm{eff}}$ by the proof of Proposition 5.2(i). By Lemma A.1 with $s_{\min} = 0$, the gain is

$$\Delta(s) := \tau\log\Big(1 + \frac{e^{s/\tau}-1}{\mathcal{A}}\Big) = \tau\log\Big(1 + \frac{e^{(s-B)/\tau} - e^{-B/\tau}}{K_{\mathrm{eff}}}\Big), \tag{32}$$

which is strictly increasing in s because $e^{s/\tau}$ is.

Non-frontier bound. If $s \le B$, then $e^{(s-B)/\tau} \le 1$ and $e^{-B/\tau} > 0$, so the fraction in Eq. (32) is less than $1/K_{\mathrm{eff}}$ and $\Delta(s) \le \tau\log(1+1/K_{\mathrm{eff}}) \le \tau\log 2$, using $K_{\mathrm{eff}} \ge 1$.

Frontier bound. Let $s = B+\omega$ with $\omega > 0$. Since $e^{-B/\tau} \le 1$, the numerator satisfies $e^{\omega/\tau} - e^{-B/\tau} \ge e^{\omega/\tau} - 1$, hence

$$\Delta(s) \;\ge\; \tau \log\Big(1 + \frac{e^{\omega/\tau} - 1}{K_{\text{eff}}}\Big) \;\ge\; \tau \log \frac{e^{\omega/\tau} - 1}{K_{\text{eff}}}. \tag{33}$$

The first expression exceeds the non-frontier bound $\tau \log(1 + 1/K_{\text{eff}})$ if and only if $e^{\omega/\tau} - 1 > 1$, that is, $\omega > \tau \log 2$. If $\omega \ge \tau \log 2$, then $e^{\omega/\tau} \ge 2$, so $e^{\omega/\tau} - 1 \ge \frac{1}{2} e^{\omega/\tau}$ and the right-hand side of Eq. (33) is at least $\tau \log \frac{e^{\omega/\tau}}{2K_{\text{eff}}} = \omega - \tau \log(2K_{\text{eff}})$. Replacing K_{eff} by its upper bound $K + 1$ gives state-independent forms of both bounds. □

A.5 Proof of Theorem 5.3

A.5.1 Matched-Marginal Comparison and Proof

Fix the snapshot $\mathcal{S}$, the N Scientist models $\theta_{1:N}$, and the shared frozen Executor and evaluator, and write $b = B(\mathcal{S})$ for the incumbent utility throughout this subsection. The Scientist models may differ across units but are the same in the two systems being compared. Both systems make exactly N new evaluator calls, including failed attempts, from this snapshot; K remains the common system-level budget. For a portfolio law π with marginals q_i, the joint assignment–response policies are

$$\begin{aligned} \pi(Z, U \mid \mathcal{S}) &= \pi(Z \mid \mathcal{S}) \prod_{i=1}^{N} \pi^{\text{sci}}_{\theta_i}(u_i \mid \mathcal{S}, z^i), \\ \pi^{\text{par}}(Z, U \mid \mathcal{S}) &= \prod_{i=1}^{N} \big[q_i(z^i \mid \mathcal{S})\, \pi^{\text{sci}}_{\theta_i}(u_i \mid \mathcal{S}, z^i) \big], \end{aligned} \tag{34}$$

so that the parallel comparison retains every Scientist model and matches each unit's assignment distribution. Let $\kappa_i(dy \mid \mathcal{S}, z^i)$ be the conditional law of unit i's utility after its Scientist responds, the Executor implements the response, and the evaluator scores it; a failed call returns utility 0. The conditional independence and per-unit dependence assumptions of Section 5.2 give

$$\mathbb{P}_\pi(dy_{1:N} \mid \mathcal{S}) = \int \prod_{i=1}^{N} \kappa_i(dy_i \mid \mathcal{S}, z^i)\, \pi(dZ \mid \mathcal{S}), \tag{35}$$

and likewise for π^{par}: the same kernels occur in both laws, and only the coupling of the assignments changes. All probabilities and expectations below condition on this fixed snapshot and these fixed models.

Proof of Theorem 5.3. Step 1: marginal matching and joint failure. For any measurable $A \subseteq [0, 1]$, integrating Eq. (35) over the other utilities yields $\Pr_\pi(Y_i \in A \mid \mathcal{S}) = \int \kappa_i(A \mid \mathcal{S}, z^i)\, q_i(dz^i \mid \mathcal{S}) = \Pr_{\pi^{\text{par}}}(Y_i \in A \mid \mathcal{S})$, so the unit marginals, including failure probabilities, agree; setting $A = [0, y]$ defines their common CDF $H_i(y)$. For the parallel law the product assignment measure gives

$$H_{\pi^{\text{par}}}(y) = \int \prod_{i=1}^{N} \kappa_i([0, y] \mid \mathcal{S}, z^i) \prod_{i=1}^{N} q_i(dz^i \mid \mathcal{S}) = \prod_{i=1}^{N} H_i(y), \tag{36}$$

whereas for π the same integral against $\pi(dZ \mid \mathcal{S})$ equals $H_\pi(y)$.

Step 2: from threshold coverage to expected best utility. Put $M_Y = \max_i Y_i$. For each realized vector, $\max\{b, M_Y\} = b + \int_b^1 \mathbf{1}\{M_Y > y\}\, dy$, since the integral is the length of $[b, M_Y)$ when $M_Y > b$ and zero otherwise; the integrand is nonnegative and bounded, so expectation and integration interchange and

$$\mathbb{E}_{\pi'}[B_{\text{new}} \mid \mathcal{S}] = b + \int_b^1 \big(1 - H_{\pi'}(y)\big)\, dy, \qquad \pi' \in \{\pi, \pi^{\text{par}}\}. \tag{37}$$

With Eq. (36) this gives the closed form of $\mathbb{E}_{\pi^{\text{par}}}[B_{\text{new}}]$ and, by subtraction, the identity in part (i). Atoms in the utility distributions cause no difficulty because the pointwise identity uses the strict event $M_Y > y$.

Step 3: the ceiling. For any $c \in [b, 1]$ and any realization, $\max\{b, Y_1, \dots, Y_N\} \le c + \sum_i (Y_i - c)^+$, because $b \le c$ and $Y_i \le c + (Y_i - c)^+$ for each i. Taking $\mathbb{E}_\pi$ and using $\mathbb{E}(Y_i - c)^+ = \int_c^1 (1 - H_i(y))\, dy$, which depends on the marginals only, gives $\mathbb{E}_\pi[B_{\text{new}}] \le c + \sum_i \int_c^1 (1 - H_i(y))\, dy$; minimizing over c yields $\mathbb{E}_\pi[B_{\text{new}}] \le \mathrm{MN}(H)$ and hence the inequality in part (i). This is the Meilijson–Nádas bound for the $N + 1$ variables $b, Y_1, \dots, Y_N$, whose minimization over all real c is attained in $[b, 1]$ (for $c < b$ the term $(b - c)^+$ raises the expression to at least its value at $c = b$, and for $c > 1$ it equals c); it is attained by some coupling of the marginals (Meilijson & Nádas, 1979), as Lai & Robbins (1976) showed for identically distributed marginals, hence sharp over all matched laws, but it need not be attained by a law of the form (35), whose dependence is created through the assignments alone.

Step 4: the potential. By Lemma A.1, the validated candidates X of the round satisfy $\mathcal{A}(\mathcal{S} \oplus X) = \mathcal{A}(\mathcal{S}) + \sum_{c \in X} (e^{s(c)/\tau} - 1) = \mathcal{A}(\mathcal{S}) + \Delta\mathcal{A}$, since a failed unit has $Y_i = 0$ and contributes $e^0 - 1 = 0$; hence $r^\Phi = \tau \log \frac{\mathcal{A}(\mathcal{S}) + \Delta\mathcal{A}}{\mathcal{A}(\mathcal{S})} = \psi(\Delta\mathcal{A})$. The function ψ is strictly concave and increasing on $[0, \infty)$, and $\mathbb{E}_\pi[\Delta\mathcal{A}] = \sum_i \int_0^1 (e^{y/\tau} - 1)\, dH_i(y)$ is determined by the marginals, so it is the same under every matched law. Jensen's inequality gives $\mathbb{E}_\pi[\psi(\Delta\mathcal{A})] \le \psi(\mathbb{E}[\Delta\mathcal{A}])$, with equality if and only if $\Delta\mathcal{A}$ is π-a.s. constant by strict concavity. Under π^{par} the summands of $\Delta\mathcal{A}$ are independent, so $\Delta\mathcal{A}$ is degenerate only if every summand is, that is, only if every Y_i is degenerate, and otherwise the inequality is strict.

Step 5: negative association. Suppose $(Y_1, \dots, Y_N)$ is negatively associated under π. Negative association is preserved by coordinatewise increasing maps (Joag-Dev & Proschan, 1983, Property P6), so $(e^{Y_i/\tau} - 1)_i$ is negatively associated, and for negatively associated summands the sum is dominated in the convex order by the sum of independent copies with the same marginals (Shao, 2000, Theorem 1): $\mathbb{E}_\pi[\varphi(\Delta\mathcal{A})] \le \mathbb{E}_{\pi^{\text{par}}}[\varphi(\Delta\mathcal{A})]$ for every convex φ, since under π^{par} the summands are exactly such independent copies. Taking $\varphi = -\psi$ gives part (iii). Permutation distributions are negatively associated (Joag-Dev & Proschan, 1983, Theorem 2.11); the utility vector is one whenever N distinct directions are assigned to the N units by a uniformly random permutation and each unit's utility is determined by its direction. The property extends to utilities of the form $Y_i = f_i(X_i, \epsilon_i)$ with X_i the direction-determined value, $\epsilon_1, \dots, \epsilon_N$ independent of each other and of the permutation, and each f_i coordinatewise increasing, because the union of independent negatively associated families is negatively associated and increasing functions of disjoint subsets preserve the property (Joag-Dev & Proschan, 1983, Properties P7 and P6). Without such structure the condition can fail, for instance when the noise scale depends on the direction. □

Remark A.8 (Why part (ii) is stated separately)**.** *The sign of the best-of-round gain in part (i) does not control the sign of the potential gain. Take* $N = 2$*,* $\tau = 0.05$*,* $K = 48$*, a root of utility* $b = 0.40$ *and no other node, three directions whose utilities are deterministically* 0.40*,* 0.45*, and* 0.50*, unit marginals* $(0.32, 0.48, 0.20)$ *over them, and the assignment coupling that places mass* 0.04 *on* $(0.40, 0.40)$*,* 0.28 *on each of* $(0.40, 0.45)$ *and* $(0.45, 0.40)$*, and* 0.20 *on each of* $(0.45, 0.50)$ *and* $(0.50, 0.45)$*. Then* $\prod_i H_i(y) - H_\pi(y) \ge 0$ *for every* y *and the best-of-round gain is* $+5.1 \times 10^{-3}$*, but the potential gain is* -8×10^{-5}*: the coupling reduces joint failure on the diagonal, yet the law of* $\Delta\mathcal{A}$ *it induces is not dominated in the convex order by that of* π^{par} *(its variance is even slightly smaller), so the concavity of* ψ *does not favor it. The potential is a function of the whole utility vector rather than of its maximum, and is governed by the convex order of* $\Delta\mathcal{A}$*, which parts (ii)–(iii) address directly.*

Two-unit interpretation. When $N = 2$, let $I_i(y) = \mathbf{1}\{Y_i \le y\}$. Marginal matching yields $\prod_i H_i(y) - H_\pi(y) = -\operatorname{Cov}_\pi\big(I_1(y), I_2(y) \mid \mathcal{S}\big)$: negatively correlated threshold failures translate directly into improved best-of-round utility while preserving both marginal utility distributions.

A.5.2 A Complementary-Direction Construction

Let $N \ge 2$ and suppose the snapshot offers N feasible research directions. For every unit, the same direction deterministically yields a validated utility $s_{\text{hi}} > b = B(\mathcal{S})$ and each of the other $N - 1$ directions yields utility $s_{\text{lo}} \in [0, b]$. Under the parallel law, every unit independently chooses a uniformly random direction. Under the coordinated law, draw a uniformly random permutation of the N directions and assign one to each unit. Every unit has the same uniform assignment marginal under both laws, and both make exactly N evaluator calls.

For $y \in [b, s_{\mathrm{hi}})$ each marginal CDF is $(N-1)/N$. The parallel law misses the successful direction with probability $(1-1/N)^N$, whereas the permutation always assigns it to one unit; for $y \geq s_{\mathrm{hi}}$ both joint CDFs equal 1. Hence

$$\begin{aligned} \prod_i H_i(y) - H_{\mathrm{perm}}(y) &= (1-1/N)^N \, \mathbf{1}\{b \leq y < s_{\mathrm{hi}}\}, \\ \mathbb{E}_{\mathrm{perm}}[B_{\mathrm{new}}] - \mathbb{E}_{\mathrm{par}}[B_{\mathrm{new}}] &= (s_{\mathrm{hi}} - b)(1-1/N)^N. \end{aligned} \tag{38}$$

In particular, for $N = 2$ the probability of finding utility s_{hi} is $3/4$ under independent assignments and 1 under coordination, although each unit finds it with probability $1/2$ in both systems. The permutation law attains the ceiling of part (i): each unit's utility is s_{hi} with probability $1/N$ and s_{lo} otherwise, so for $c \in [b, s_{\mathrm{hi}}]$ the expression $c + \sum_i \mathbb{E}(Y_i - c)^+$ equals $c + (s_{\mathrm{hi}} - c) = s_{\mathrm{hi}}$, whence $\mathrm{MN}(H) = s_{\mathrm{hi}} = \mathbb{E}_{\mathrm{perm}}[B_{\mathrm{new}}]$.

It also attains the ceiling of part (ii) for every $\tau > 0$. Let R_{hi} count the units assigned to the successful direction; then $\Delta\mathcal{A} = N(e^{s_{\mathrm{lo}}/\tau} - 1) + R_{\mathrm{hi}}\,(e^{s_{\mathrm{hi}}/\tau} - e^{s_{\mathrm{lo}}/\tau})$. Under the permutation law $R_{\mathrm{hi}} = 1$, so $\Delta\mathcal{A}$ is constant and $\mathbb{E}_{\mathrm{perm}}[\psi(\Delta\mathcal{A})] = \psi(\mathbb{E}[\Delta\mathcal{A}])$; under independent assignments $R_{\mathrm{hi}} \sim \mathrm{Binomial}(N, 1/N)$ has mean 1 and positive variance, so strict concavity gives $\mathbb{E}_{\mathrm{par}}[\psi(\Delta\mathcal{A})] < \psi(\mathbb{E}[\Delta\mathcal{A}])$, and the potential gap $\psi(\mathbb{E}[\Delta\mathcal{A}]) - \mathbb{E}_{\mathrm{par}}[\psi(\Delta\mathcal{A})]$ is strictly positive.

A full-support version is obtained from the assignment law $\pi^\lambda = (1-\lambda)\pi^{\mathrm{par}} + \lambda\pi^{\mathrm{perm}}$ with $0 < \lambda < 1$. This mixture preserves every uniform marginal and assigns positive probability to every one of the N^N portfolios, and both the best-of-round and the potential gains over π^{par} equal λ times their values under π^{perm}, since expectations are affine in the law. In particular, the expected round reward of the Coordinator increases linearly in the weight λ it places on the complementary law.

A.6 Private Adaptation and Proof of Proposition 5.4

A.6.1 Setting

Fix a snapshot $\mathcal{S}$ and a full portfolio Z. Let $\bar{r}^\Phi(W_1, \ldots, W_N) = \bar{r}^\Phi(W_{1:N}; \mathcal{S}, Z)$ be the expected round reward $\mathbb{E}[r^\Phi \mid \mathcal{S}, Z]$ when Scientist i responds from effective weights W_i, with the expectation over the Scientists' responses, execution, and evaluation; conditioning on the full Z retains every unit's contribution to the shared reward, although Scientist i observes only $(\mathcal{S}, z^i)$. Every Scientist starts from the common model with weights W_0. Each private adapter and the shared comparator have effective weights $W(U,V) = W_0 + \gamma UV$ with $U \in \mathbb{R}^{d_{\mathrm{out}} \times r}$, $V \in \mathbb{R}^{r \times d_{\mathrm{in}}}$, scale $\gamma > 0$, and the common zero-output initialization $(U, V) = (0, V_0)$, with the implemented settings in Appendix B.3. Write $P_V = V_0^\top V_0$ and

$$\begin{aligned} G_i &= \nabla_{W_i} \bar{r}^\Phi(W_0, \ldots, W_0), \qquad \Gamma_i = G_i V_0^\top, \qquad \bar{G} = \tfrac{1}{N}\textstyle\sum_i G_i, \qquad \bar{\Gamma} = \tfrac{1}{N}\textstyle\sum_i \Gamma_i, \\ \mathcal{D}(Z;\mathcal{S}) &= \tfrac{1}{N^2}\textstyle\sum_{i<j} \|\Gamma_i - \Gamma_j\|_F^2, \qquad \mathcal{D}_\phi(\mathcal{S}) = \mathbb{E}_{Z \sim \pi_\phi^{\mathrm{coord}}(\cdot|\mathcal{S})}[\mathcal{D}(Z;\mathcal{S})]. \end{aligned} \tag{39}$$

For multiple adapted matrices, regard each W_i as their concatenated collection; inner products and squared norms are summed over matrices, and $\|V_0\|_2$ denotes the largest factor spectral norm. By Proposition 5.2(iv) applied to Scientist i's block at the fixed portfolio, G_i is a population sequence-score gradient; the distinction from the token-normalized surrogate is given in Appendix A.3.1.

The two schemes are compared on the joint objective $\bar{r}^\Phi$. Under *private* adaptation, each Scientist takes one population factor-gradient ascent step of size η on its own adapter, giving W_i^+. Under a *shared* adapter, one adapter trained on the pooled responses with the same mean-reduced surrogate takes a step of the same size; its population gradient is the average $\bar{G}$ of the per-Scientist gradients, so its update is the average of the private updates, and every Scientist then responds from the same weights W_{sh}^+. The improvements per Scientist are

$$\begin{aligned} \mathcal{I}_{\mathrm{priv}}(Z) &= \tfrac{1}{N}\big[\bar{r}^\Phi(W_1^+, \ldots, W_N^+) - \bar{r}^\Phi(W_0, \ldots, W_0)\big], \\ \mathcal{I}_{\mathrm{sh}}(Z) &= \tfrac{1}{N}\big[\bar{r}^\Phi(W_{\mathrm{sh}}^+, \ldots, W_{\mathrm{sh}}^+) - \bar{r}^\Phi(W_0, \ldots, W_0)\big]. \end{aligned} \tag{40}$$

A sum-reduced pooled loss would instead move the shared adapter by N times as much; the comparison below is at matched per-sample step size, which is the mean-reduced form of Eq. (6).

Lemma A.9 (Pairwise variance identity). *For matrices $A_1, \ldots, A_N$ with mean $\bar{A} = \frac{1}{N}\sum_i A_i$, $\frac{1}{N}\sum_i \|A_i\|_F^2 - \|\bar{A}\|_F^2 = \frac{1}{N^2}\sum_{i<j} \|A_i - A_j\|_F^2$.*

Proof. Expanding the squares, $\sum_{i<j}\|A_i - A_j\|_F^2 = \frac{1}{2}\sum_{i,j}(\|A_i\|_F^2 + \|A_j\|_F^2 - 2\langle A_i, A_j\rangle_F) = N\sum_i \|A_i\|_F^2 - \|\sum_i A_i\|_F^2$. Divide by N^2 and use $\sum_i A_i = N\bar{A}$. □

Lemma A.10 (Private-adapter specialization)**.** *In the setting above, one population-gradient step gives $\Delta W_i^+ = \eta\gamma^2 G_i P_V$ for private adapters and $\Delta W_{\mathrm{sh}}^+ = \eta\gamma^2 \bar{G} P_V = \frac{1}{N}\sum_i \Delta W_i^+$ for the shared adapter, so $\Delta W_i^+ - \Delta W_j^+ = \eta\gamma^2(\Gamma_i - \Gamma_j)V_0$. If $\bar{r}^\Phi(\cdot; \mathcal{S}, Z)$ has an L_r-Lipschitz gradient on the update segments, then*

$$\mathcal{I}_{\mathrm{priv}}(Z) - \mathcal{I}_{\mathrm{sh}}(Z) = \eta\gamma^2 \mathcal{D}(Z; \mathcal{S}) + R_\eta(Z), \qquad |R_\eta(Z)| \le L_r \eta^2 \gamma^4 \|V_0\|_2^2 \, \tfrac{1}{N}\textstyle\sum_i \|\Gamma_i\|_F^2. \quad (41)$$

If moreover $\mathbb{E}_Z[\frac{1}{N}\sum_i \|\Gamma_i\|_F^2 \mid \mathcal{S}] < \infty$, then $\mathbb{E}_Z[\mathcal{I}_{\mathrm{priv}} - \mathcal{I}_{\mathrm{sh}} \mid \mathcal{S}] = \eta\gamma^2 \mathcal{D}_\phi(\mathcal{S}) + O(\eta^2)$, which is strictly positive for all sufficiently small $\eta > 0$ whenever $\mathcal{D}_\phi(\mathcal{S}) > 0$; if $\mathcal{D}_\phi(\mathcal{S}) = 0$, the two schemes produce the same weights almost surely.

Proof. Step 1: effective updates. For any differentiable $\bar{r}^\Phi$, the differential $dW = \gamma\, dU\, V + \gamma\, U\, dV$ gives $\nabla_U \bar{r}^\Phi(W(U,V)) = \gamma \nabla_W \bar{r}^\Phi(W(U,V))V^\top$ and $\nabla_V \bar{r}^\Phi(W(U,V)) = \gamma U^\top \nabla_W \bar{r}^\Phi(W(U,V))$, by cyclicity of the trace. At $(0, V_0)$, Scientist i therefore has factor gradients $(\gamma\Gamma_i, 0)$, so $U_i^+ = \eta\gamma\Gamma_i$, $V_i^+ = V_0$, and

$$\Delta W_i^+ = \gamma U_i^+ V_i^+ = \eta\gamma^2 G_i P_V, \qquad \Delta W_{\mathrm{sh}}^+ = \eta\gamma^2 \bar{G} P_V = \tfrac{1}{N}\textstyle\sum_i \Delta W_i^+, \quad (42)$$

and $\Delta W_i^+ - \Delta W_j^+ = \eta\gamma^2(\Gamma_i - \Gamma_j)V_0$.

Step 2: first-order gains and remainders. For any matrix A of the shape of G_i, cyclicity of the trace and the spectral-norm inequality give

$$\langle A, AP_V\rangle_F = \|AV_0^\top\|_F^2, \qquad \|AP_V\|_F = \|(AV_0^\top)V_0\|_F \le \|AV_0^\top\|_F \|V_0\|_2. \quad (43)$$

For a joint displacement $\Delta = (\Delta_1, \dots, \Delta_N)$, the fundamental theorem of calculus yields

$$\bar{r}^\Phi(W_0+\Delta_1, \dots, W_0+\Delta_N) - \bar{r}^\Phi(W_0, \dots, W_0) = \textstyle\sum_i \langle G_i, \Delta_i\rangle_F + \varrho(\Delta), \;\; |\varrho(\Delta)| \le \frac{L_r}{2}\sum_i \|\Delta_i\|_F^2, \quad (44)$$

by integrating the Lipschitz bound along the segment and applying Cauchy–Schwarz. For the private step, $\sum_i \langle G_i, \Delta W_i^+\rangle_F = \eta\gamma^2 \sum_i \|\Gamma_i\|_F^2$ by Eq. (43); for the shared step, $\sum_i \langle G_i, \Delta W_{\mathrm{sh}}^+\rangle_F = \eta\gamma^2 N \langle \bar{G}, \bar{G} P_V\rangle_F = \eta\gamma^2 N \|\bar{\Gamma}\|_F^2$. Dividing by N,

$$\mathcal{I}_{\mathrm{priv}}(Z) = \eta\gamma^2 \tfrac{1}{N}\textstyle\sum_i \|\Gamma_i\|_F^2 + \frac{1}{N}\varrho(\Delta W_{1:N}^+), \qquad \mathcal{I}_{\mathrm{sh}}(Z) = \eta\gamma^2 \|\bar{\Gamma}\|_F^2 + \frac{1}{N}\varrho(\Delta W_{\mathrm{sh}}^+, \dots, \Delta W_{\mathrm{sh}}^+), \quad (45)$$

and Lemma A.9 turns the difference of the first-order terms into $\eta\gamma^2 \mathcal{D}(Z; \mathcal{S})$, which is Eq. (41) with $R_\eta(Z)$ the difference of the two remainders. By Eqs. (43)–(44) and $\|\bar{\Gamma}\|_F^2 \le \frac{1}{N}\sum_i \|\Gamma_i\|_F^2$,

$$|R_\eta(Z)| \le \frac{L\eta^2\gamma^4}{2}\Big[\tfrac{1}{N}\textstyle\sum_i \|G_i P_V\|_F^2 + \|\bar{G} P_V\|_F^2\Big] \le L_r \eta^2 \gamma^4 \|V_0\|_2^2 \, \tfrac{1}{N}\textstyle\sum_i \|\Gamma_i\|_F^2. \quad (46)$$

Step 3: averaging over portfolios. Taking conditional expectations in Eq. (41) and using Eq. (46) with the integrability assumption gives $\mathbb{E}_Z[\mathcal{I}_{\mathrm{priv}} - \mathcal{I}_{\mathrm{sh}} \mid \mathcal{S}] \ge \eta\gamma^2 \mathcal{D}_\phi(\mathcal{S}) - C_\phi(\mathcal{S})\eta^2$ with $C_\phi(\mathcal{S}) = L_r \gamma^4 \|V_0\|_2^2 \, \mathbb{E}_Z[\frac{1}{N}\sum_i \|\Gamma_i\|_F^2 \mid \mathcal{S}] < \infty$, which is positive for $0 < \eta < \gamma^2 \mathcal{D}_\phi(\mathcal{S})/C_\phi(\mathcal{S})$. If $\mathcal{D}_\phi(\mathcal{S}) = 0$, then $\Gamma_i = \bar{\Gamma}$ for every i almost surely, and Eq. (42) gives $\Delta W_i^+ = \Delta W_{\mathrm{sh}}^+$. □

The zero-output initialization can also be imposed at a later common model by absorbing its effective weights into W_0 and initializing fresh factors $(0, V_0)$ there.

A.6.2 Proof of Proposition 5.4

Proof of Proposition 5.4. The first-order terms of Eq. (45), rewritten with Lemma A.9, are $\eta\gamma^2(\|\bar{\Gamma}\|_F^2 + \mathcal{D}(Z; \mathcal{S}))$ for private adapters and $\eta\gamma^2\|\bar{\Gamma}\|_F^2$ for the shared adapter; for differentiable $\bar{r}^\Phi$ they are η times the derivatives at $\eta = 0$ of $\eta \mapsto \bar{r}^\Phi(W_0 + \Delta W_{1:N}^+(\eta))$ by the chain rule, and under the Lipschitz condition of Lemma A.10 the remainders are $O(\eta^2)$ by Eq. (46).

For the second claim, suppose $z^1 = \cdots = z^N = z$ and that the assignment is the only Scientist-specific input. Define $\Psi(u_1, \dots, u_N) = \mathbb{E}[\Phi_{\tau,K}(\mathcal{S} \oplus X) - \Phi_{\tau,K}(\mathcal{S}) \mid u_1, \dots, u_N]$, where execution and evaluation act identically and independently on each $(\mathcal{S}, z, u_i)$. The potential depends only on the

multiset of validated utilities (Lemma A.1), so Ψ is permutation-invariant, and $\bar{r}^{\Phi}(W_1, \ldots, W_N) = \mathbb{E}[\Psi(u_1, \ldots, u_N)]$ with $u_i \sim \pi^{\mathrm{sci}}_{W_i}(\cdot \mid \mathcal{S}, z)$ drawn independently. Hence $\bar{r}^{\Phi}$ is a symmetric function of $(W_1, \ldots, W_N)$, and its partial gradients at the symmetric point $(W_0, \ldots, W_0)$ coincide: $G_i = G_j$ for all i, j. Consequently $\Gamma_i = \Gamma_j$, $\mathcal{D}(Z; \mathcal{S}) = 0$, and by Eq. (42) the private updates are identical, preserving adapter symmetry. □

A.7 Proof of Theorem 5.6

Throughout, $F(\Theta) = J_{\tau,K}(\Theta)$ on Ω, $F^\star = \sup_\Omega F$, $J_T^{\mathcal{H}} = \sup_{\Pi \in \mathcal{H}} J_T(\Pi; \mathcal{P})$, $\varepsilon_{\mathrm{cov}} = s^\star - J_T^{\mathcal{H}}$, and $\varepsilon_{\mathrm{opt}}(m) = F^\star - \mathbb{E}[F(\Theta^{(m)})]$. The proof decomposes the regret into $\varepsilon_{\mathrm{cov}}$, the smoothing cost, and $\varepsilon_{\mathrm{opt}}(M)$ (Lemma A.11), and bounds $\varepsilon_{\mathrm{opt}}(M)$ by a rate for two-block ascent with biased estimates (Proposition A.13). Lemma A.12 bounds $\varepsilon_{\mathrm{cov}}$ under a coverage condition, and Appendix A.7.3 extends everything to an arbitrary starting state.

A.7.1 Decomposition and Coverage

Lemma A.11 (Regret decomposition)**.** *For every $\Theta \in \Omega$, $s^\star - J_T(\Pi_\Theta; \mathcal{P}) \le \varepsilon_{\mathrm{cov}} + \tau \log(K+1) + F^\star - F(\Theta)$.*

Proof. By Proposition 5.2(i), for every $\Theta' \in \Omega$,

$$F(\Theta') \le J_T(\Pi_{\Theta'}; \mathcal{P}) \le F(\Theta') + \tau \log(K+1). \tag{47}$$

Taking the supremum over $\Theta' \in \Omega$ in the right inequality gives $J_T^{\mathcal{H}} \le F^\star + \tau \log(K+1)$, and the left inequality at $\Theta' = \Theta$ gives $J_T(\Pi_\Theta; \mathcal{P}) \ge F(\Theta)$. Hence $s^\star - J_T(\Pi_\Theta; \mathcal{P}) = \big(s^\star - J_T^{\mathcal{H}}\big) + \big(J_T^{\mathcal{H}} - J_T(\Pi_\Theta; \mathcal{P})\big) \le \varepsilon_{\mathrm{cov}} + \tau \log(K+1) + F^\star - F(\Theta)$. □

Lemma A.12 (Coverage)**.** *Suppose some $\Pi^\dagger \in \mathcal{H}$ produces at every evaluator call, conditioned on any history, a valid candidate of utility at least $s^\star - \epsilon$ with probability at least $p_\epsilon > 0$. Then $\varepsilon_{\mathrm{cov}} \le \epsilon + (s^\star - s_0) e^{-K p_\epsilon}$.*

Proof. Index the evaluator calls of a run of $\Pi^\dagger$ by $j = 1, \ldots, K$, ordering the calls of one round arbitrarily, let $\mathcal{K}_{j-1}$ be the history before call j, including the outcomes of earlier calls of the same round, let S_j be the event that call j returns a validated candidate with utility at least $s^\star - \epsilon$, and let $\mathsf{E}_j = \bigcap_{j' \le j} \mathsf{S}_{j'}^c$, with E_0 the sure event. The hypothesis states $\mathbb{P}(\mathsf{S}_j \mid \mathcal{K}_{j-1}) \ge p_\epsilon$. Since $\mathsf{E}_{j-1} \in \mathcal{K}_{j-1}$ and $\mathsf{E}_j = \mathsf{E}_{j-1} \cap \mathsf{S}_j^c$,

$$\mathbb{P}(\mathsf{E}_j) = \mathbb{E}\big[\mathbf{1}\{\mathsf{E}_{j-1}\}\, \mathbb{P}(\mathsf{S}_j^c \mid \mathcal{K}_{j-1})\big] \le (1 - p_\epsilon)\, \mathbb{P}(\mathsf{E}_{j-1}), \qquad \text{so} \qquad \mathbb{P}(\mathsf{E}_K) \le (1-p_\epsilon)^K \le e^{-K p_\epsilon}. \tag{48}$$

On E_K^c some validated candidate has utility at least $s^\star - \epsilon$, so $s^\star - B(\mathcal{S}_{\mathrm{end}}) \le \epsilon$; on E_K the root remains, so $B(\mathcal{S}_{\mathrm{end}}) \ge s_0$. Therefore $s^\star - J_T(\Pi^\dagger; \mathcal{P}) = \mathbb{E}[s^\star - B(\mathcal{S}_{\mathrm{end}})] \le \epsilon\, \mathbb{P}(\mathsf{E}_K^c) + (s^\star - s_0)\, \mathbb{P}(\mathsf{E}_K) \le \epsilon + (s^\star - s_0) e^{-K p_\epsilon}$, and $\varepsilon_{\mathrm{cov}} \le s^\star - J_T(\Pi^\dagger; \mathcal{P})$ because $\Pi^\dagger \in \mathcal{H}$. □

A.7.2 Two-Block Ascent with Biased Estimates

Write $\vartheta = (\theta_1, \ldots, \theta_N)$ for the stacked Scientist block. As in Section 5.3, $\widehat{\nabla}_m$ is supported on the block b_m updated at step m and is zero outside it, and $\nabla_{\mathsf{b}} F$ denotes the gradient of F restricted to block b and padded with zeros likewise, so that $\|\nabla F\|^2 = \|\nabla_\phi F\|^2 + \|\nabla_\vartheta F\|^2$. The updates of Algorithm 1 are

$$\Theta^{(m+1)} = \Theta^{(m)} + \eta\, \widehat{\nabla}_m, \qquad m = 0, \ldots, 2C-1, \qquad \mathsf{b}_{2j} = \phi, \quad \mathsf{b}_{2j+1} = \vartheta, \tag{49}$$

where $\Theta^{(m)}$ is $\mathcal{F}_m$-measurable and the filtrations are nested. Since Ω is convex and almost surely contains every iterate, all update segments lie in Ω, and the condition $\mu \le L$ in (A1) gives $q = 1 - \eta\mu/4 \in [15/16, 1)$ for $\eta \le 1/(4L)$; this condition is automatic whenever Ω contains $\Theta + \nabla F(\Theta)/L$ for every $\Theta \in \Omega$ (Step 3 below).

Assumption (A2) is where the algorithm enters. The estimate $\widehat{\nabla}_m$ is the negative gradient of Eq. (6) on the group collected at $\Theta^{(m)}$, evaluated at the behavior model, where the importance ratios equal

one and clipping is inactive. By Proposition 5.2(iv), the leave-one-out group estimate is unbiased for the current-round part of the block gradient at the snapshot at which the group is drawn; the bias of $\widehat{\nabla}_m$ with respect to $\nabla_{\mathsf{b}_m} F(\Theta^{(m)})$ therefore consists of the omitted downstream terms $R^{\Phi}_{t+1:\mathrm{end}}$ of Eq. (7), the dependence of F on rounds other than the one at which the group is drawn, and the deviations of the implemented surrogate from the leave-one-out estimate (Appendix A.3.1); δ bounds their sum. The variance bound σ^2 is that of a group estimate, which averages G forks.

Proposition A.13 (Two-block ascent with biased estimates)**.** *Under (A1)–(A2), for* $0 < \eta \le 1/(4L)$ *and* $q = 1 - \eta\mu/4$*,*

$$\varepsilon_{\mathrm{opt}}(2C) \le q^C \varepsilon_{\mathrm{opt}}(0) + \frac{5L\eta\sigma^2}{\mu} + \frac{8\delta^2}{\mu}. \tag{50}$$

For an odd number of updates, $\varepsilon_{\mathrm{opt}}(2C+1) \le \varepsilon_{\mathrm{opt}}(2C) + \frac{3}{4}\eta\delta^2 + \frac{1}{2}L\eta^2\sigma^2$.

Proof. Step 1: one block update. Condition on $\mathcal{F}_m$, write $v_m = \nabla_{\mathsf{b}_m} F(\Theta^{(m)})$, and decompose $\widehat{\nabla}_m = v_m + e_m + \xi_m$ with $e_m = \bar{\nabla}_m - v_m$, $\bar{\nabla}_m = \mathbb{E}[\widehat{\nabla}_m \mid \mathcal{F}_m]$, so that $\|e_m\| \le \delta$, $\mathbb{E}[\xi_m \mid \mathcal{F}_m] = 0$, and $\mathbb{E}[\|\xi_m\|^2 \mid \mathcal{F}_m] \le \sigma^2$ by (A2). By (A1) and since $\widehat{\nabla}_m$ is supported on b_m, $F(\Theta^{(m+1)}) \ge F(\Theta^{(m)}) + \eta\langle v_m, \widehat{\nabla}_m\rangle - \frac{L\eta^2}{2}\|\widehat{\nabla}_m\|^2$. Now $\mathbb{E}[\langle v_m, \widehat{\nabla}_m\rangle \mid \mathcal{F}_m] = \|v_m\|^2 + \langle v_m, e_m\rangle \ge \frac{1}{2}\|v_m\|^2 - \frac{1}{2}\delta^2$ and $\mathbb{E}[\|\widehat{\nabla}_m\|^2 \mid \mathcal{F}_m] = \|v_m + e_m\|^2 + \mathbb{E}[\|\xi_m\|^2 \mid \mathcal{F}_m] \le 2\|v_m\|^2 + 2\delta^2 + \sigma^2$, hence

$$\begin{aligned}\mathbb{E}[F(\Theta^{(m+1)}) \mid \mathcal{F}_m] &\ge F(\Theta^{(m)}) + \eta\Big(\frac{1}{2} - L\eta\Big)\|v_m\|^2 - \eta\delta^2\Big(\frac{1}{2} + L\eta\Big) - \frac{L\eta^2\sigma^2}{2} \\ &\ge F(\Theta^{(m)}) + \frac{\eta}{4}\|v_m\|^2 - \frac{3}{4}\eta\delta^2 - \frac{L\eta^2\sigma^2}{2},\end{aligned} \tag{51}$$

where the last step uses $\eta \le 1/(4L)$.

Step 2: a complete cycle. Write $v = \nabla_\phi F(\Theta^{(2j)})$, $w_0 = \nabla_\vartheta F(\Theta^{(2j)})$, and $w = \nabla_\vartheta F(\Theta^{(2j+1)})$. By (A1), $\|w - w_0\| \le L\eta\|\widehat{\nabla}_{2j}\|$, and $\|a\|^2 \ge \frac{1}{2}\|a'\|^2 - \|a - a'\|^2$ gives

$$\|w\|^2 \ge \tfrac{1}{2}\|w_0\|^2 - L^2\eta^2\|\widehat{\nabla}_{2j}\|^2, \qquad \mathbb{E}[\|\widehat{\nabla}_{2j}\|^2 \mid \mathcal{F}_{2j}] \le 2\|v\|^2 + 2\delta^2 + \sigma^2. \tag{52}$$

Applying Eq. (51) at $m = 2j+1$, taking $\mathbb{E}[\,\cdot \mid \mathcal{F}_{2j}]$, inserting Eq. (52), and then applying Eq. (51) at $m = 2j$,

$$\begin{aligned}\mathbb{E}[F(\Theta^{(2j+2)}) \mid \mathcal{F}_{2j}] &\ge F(\Theta^{(2j)}) + \Big(\frac{\eta}{4} - \frac{L^2\eta^3}{2}\Big)\|v\|^2 + \frac{\eta}{8}\|w_0\|^2 \\ &\quad - \eta\delta^2\Big(\frac{3}{2} + \frac{L^2\eta^2}{2}\Big) - L\eta^2\sigma^2\Big(1 + \frac{L\eta}{4}\Big) \\ &\ge F(\Theta^{(2j)}) + \frac{\eta}{8}\|\nabla F(\Theta^{(2j)})\|^2 - 2\eta\delta^2 - \frac{5}{4}L\eta^2\sigma^2 \\ &\ge F(\Theta^{(2j)}) + \frac{\eta\mu}{4}\big(F^\star - F(\Theta^{(2j)})\big) - 2\eta\delta^2 - \frac{5}{4}L\eta^2\sigma^2,\end{aligned} \tag{53}$$

where the third line uses $\eta \le 1/(4L)$, so that $\frac{\eta}{4} - \frac{L^2\eta^3}{2} \ge \frac{7\eta}{32} \ge \frac{\eta}{8}$, $\frac{3}{2} + \frac{L^2\eta^2}{2} \le 2$, and $1 + \frac{L\eta}{4} \le \frac{5}{4}$, and the last step uses the PL inequality in (A1). Subtracting from $F^\star$ and taking expectations, $\varepsilon_{\mathrm{opt}}(2j+2) \le q\,\varepsilon_{\mathrm{opt}}(2j) + 2\eta\delta^2 + \frac{5}{4}L\eta^2\sigma^2$, and unrolling over C cycles with $\sum_{j<C} q^j \le \frac{1}{1-q} = \frac{4}{\eta\mu}$,

$$\varepsilon_{\mathrm{opt}}(2C) \le q^C\varepsilon_{\mathrm{opt}}(0) + \frac{4}{\eta\mu}\Big(2\eta\delta^2 + \frac{5}{4}L\eta^2\sigma^2\Big) = q^C\varepsilon_{\mathrm{opt}}(0) + \frac{8\delta^2}{\mu} + \frac{5L\eta\sigma^2}{\mu}. \tag{54}$$

Step 3: an unmatched final block. Dropping the nonnegative gradient term in Eq. (51) at $m = 2C$ and taking expectations gives the odd-M statement. Finally, if $\Theta + \nabla F(\Theta)/L \in \Omega$, then by smoothness $F^\star - F(\Theta) \ge F(\Theta + \nabla F(\Theta)/L) - F(\Theta) \ge \frac{1}{2L}\|\nabla F(\Theta)\|^2$, which with the PL inequality gives $\mu \le L$ wherever $\nabla F \ne 0$, so the condition $\mu \le L$ in (A1) is then implied. □

Proof of Theorem 5.6. Apply Lemma A.11 at $\Theta = \Theta^{(M)}$ and take expectations over the $M = 2C$ updates: $s^\star - \mathbb{E}[J_T(\Pi_{\Theta^{(M)}}; \mathcal{P})] \le \varepsilon_{\mathrm{cov}} + \tau\log(K+1) + \varepsilon_{\mathrm{opt}}(2C)$. Bound the last term by Proposition A.13, noting $\varepsilon_{\mathrm{opt}}(0) = F^\star - F(\Theta^{(0)})$ since $\Theta^{(0)}$ is deterministic. □

A.7.3 Arbitrary Starting States and the Interleaved Run

For a state $\mathcal{S}$ with $k(\mathcal{S})$ evaluator calls remaining, let $J_T(\Pi;\mathcal{S})$ and $J_{\tau,K}(\Theta;\mathcal{S})$ be the expected best utility and the expected potential at budget exhaustion when Π, respectively Π_Θ, is run from $\mathcal{S}$ without further updates, and write $F_{\mathcal{S}}(\Theta) = J_{\tau,K}(\Theta;\mathcal{S})$, $F^\star_{\mathcal{S}} = \sup_\Omega F_{\mathcal{S}}$, and $\varepsilon_{\mathrm{cov}}(\mathcal{S}) = s^\star - \sup_{\Pi\in\mathcal{H}} J_T(\Pi;\mathcal{S})$. Since $B(\mathcal{S}) \geq s_0 \geq 0$, Proposition 5.2(i) applies to every state reached from $\mathcal{S}$, so the proof of Lemma A.11 gives, for every $\Theta \in \Omega$,

$$s^\star - J_T(\Pi_\Theta;\mathcal{S}) \leq \varepsilon_{\mathrm{cov}}(\mathcal{S}) + \tau\log(K+1) + F^\star_{\mathcal{S}} - F_{\mathcal{S}}(\Theta), \tag{55}$$

and Lemma A.12 holds with the coverage condition imposed on rollouts from $\mathcal{S}$, K replaced by $k(\mathcal{S})$, and s_0 by $B(\mathcal{S})$, giving $\varepsilon_{\mathrm{cov}}(\mathcal{S}) \leq \epsilon + (s^\star - B(\mathcal{S}))e^{-k(\mathcal{S})p_\epsilon}$. For a fixed starting state, Theorem 5.6 therefore holds verbatim with $\mathcal{P}$ replaced by $\mathcal{S}$ and (A1)–(A2) imposed on $F_{\mathcal{S}}$.

In MELD the adaptation groups and the final inference phase share one budget: candidates produced during adaptation are committed to the tree, and once the remaining budget no longer admits a full group, the remainder is a rollout of $\Pi_{\Theta^{(M)}}$ without further updates (Appendix B.3). Let $\mathcal{S}_{\mathrm{ad}}$ be the state at which adaptation stops; $(\Theta^{(M)}, \mathcal{S}_{\mathrm{ad}})$ is random through the adaptation run, and conditionally on it the final phase is a rollout of $\Pi_{\Theta^{(M)}}$ from $\mathcal{S}_{\mathrm{ad}}$, so by the tower property $\mathfrak{R}_T(\textsc{Meld}) = s^\star - \mathbb{E}[B(\mathcal{S}_{\mathrm{end}})] = \mathbb{E}\big[s^\star - J_T(\Pi_{\Theta^{(M)}};\mathcal{S}_{\mathrm{ad}})\big]$. Applying Eq. (55) pathwise at $(\Theta^{(M)}, \mathcal{S}_{\mathrm{ad}})$,

$$\mathfrak{R}_T(\textsc{Meld}) \leq \mathbb{E}[\varepsilon_{\mathrm{cov}}(\mathcal{S}_{\mathrm{ad}})] + \tau\log(K+1) + \mathbb{E}\big[F^\star_{\mathcal{S}_{\mathrm{ad}}} - F_{\mathcal{S}_{\mathrm{ad}}}(\Theta^{(M)})\big]. \tag{56}$$

The last expectation is not covered by Proposition A.13 directly, because $F_{\mathcal{S}_{\mathrm{ad}}}$ depends on the whole adaptation run and is not $\mathcal{F}_m$-measurable. We therefore assume a snapshot-mismatch bound: (A3) there is $\varepsilon_{\mathrm{snap}} \geq 0$ such that $\big|\big(F^\star_{\mathcal{S}_{\mathrm{ad}}} - F_{\mathcal{S}_{\mathrm{ad}}}(\Theta)\big) - \big(F^\star - F(\Theta)\big)\big| \leq \varepsilon_{\mathrm{snap}}$ for all $\Theta \in \Omega$ almost surely, where F is the objective of Theorem 5.6 from the initial state. Then $\mathbb{E}[F^\star_{\mathcal{S}_{\mathrm{ad}}} - F_{\mathcal{S}_{\mathrm{ad}}}(\Theta^{(M)})] \leq \varepsilon_{\mathrm{opt}}(M) + \varepsilon_{\mathrm{snap}}$, and Proposition A.13 yields

$$\mathfrak{R}_T(\textsc{Meld}) \leq \mathbb{E}[\varepsilon_{\mathrm{cov}}(\mathcal{S}_{\mathrm{ad}})] + \tau\log(K+1) + q^C\big(F^\star - F(\Theta^{(0)})\big) + \frac{5L\eta\sigma^2 + 8\delta^2}{\mu} + \varepsilon_{\mathrm{snap}}. \tag{57}$$

The coverage term involves the length $k(\mathcal{S}_{\mathrm{ad}})$ of the inference phase and the incumbent $B(\mathcal{S}_{\mathrm{ad}})$ that adaptation produced, and $\varepsilon_{\mathrm{snap}}$ measures how much the surrogate optimality gap changes when it is measured from $\mathcal{S}_{\mathrm{ad}}$ rather than from the initial state, which is controlled by the budget and the candidates committed during adaptation.

Remark A.14 (Rate, noise, bias, and parallelism)**.** *With $\delta = \sigma = 0$, Eq. (50) gives geometric decay to zero; the two floors are the price of gradient noise and bias under a Polyak–Łojasiewicz condition (see Karimi et al., 2016 for the stochastic case). The cross-block term in Eq. (52) is what the Gauss–Seidel order pays for evaluating the Scientist gradient after the Coordinator has moved; it costs a constant factor, not a dependence on N. Parallel Scientist updates avoid an explicit $1/(N+1)$ block-selection factor because the exponent counts Coordinator–Scientist cycles rather than individual blocks; L, μ, σ^2, and δ may still depend on N.*

Remark A.15 (Private-gradient energy decomposition)**.** *At any Scientist update, write $g_i = \nabla_{\theta_i} F$, $\bar{g} = \frac{1}{N}\sum_i g_i$, and $\mathcal{D}^\vartheta = \frac{1}{N}\sum_{i<j}\|g_i - g_j\|^2$. Lemma A.9 gives $\|\nabla_\vartheta F\|^2 = \sum_i \|g_i\|^2 = N\|\bar{g}\|^2 + \mathcal{D}^\vartheta$, so dispersion contributes additively to the gradient energy that drives Eq. (51). At a common model $\theta_1 = \cdots = \theta_N = \theta$, a shared adapter with the mean-reduced pooled loss moves every Scientist by $\eta\bar{g}$ while private steps move Scientist i by ηg_i; their first-order ascents are $\eta N\|\bar{g}\|^2$ and $\eta(N\|\bar{g}\|^2 + \mathcal{D}^\vartheta)$, the parameter-space counterpart of Lemma A.10. A uniform PL constant, however, lower-bounds the ratio of gradient energy to objective gap along all iterates; a larger dispersion at one snapshot does not by itself yield a larger μ.*

B Experimental Details

B.1 Tasks, Metrics, and Reporting Anchors

The 12 MLGym tasks span tabular prediction, image classification, natural-language understanding and modeling, game-theoretic strategy design, and sequential control. They comprise 11 of the

13 MLGym-Bench tasks (all except image captioning and 3-SAT) and the Titanic task distributed with the MLGym framework. Table 2 lists the 12 tasks with their metric, training timeout, and the two anchors used for normalization: the shipped baseline score and the human-best reference. Across task families, agents edit ML pipelines, train classifiers or language models, implement game strategies, and train control policies under task-specific evaluators. The 10 MLE-bench tasks of Table 1 are Kaggle competitions on tabular, text, image, scientific, and medical data. Nine are from the 75-competition MLE-bench test set (Plant Pathology is the 2020 FGVC7 competition), and Spaceship Titanic is from its development split.

Table 2: Tasks, metrics, timeouts, and reporting anchors. Human-best references are independently measured as described below; for Titanic and MNLI, re-measured baseline values are shown in parentheses.

Task	Metric	Timeout	Baseline	Human best
Titanic	accuracy	1800 s	0.766 (0.943)	0.830
House Prices	R^2	1800 s	0.880	0.990
Fashion-MNIST	accuracy	1800 s	0.848	0.968
CIFAR-10	accuracy	1800 s	0.497	0.994
MNLI	accuracy (%)	2400 s	52.51 (50.81)	92.5
Language Modeling	validation loss	2400 s	4.238	3.280
Battle of the Sexes	mean payoff	1800 s	1.023	1.667
Iterated Prisoner's Dilemma	mean payoff	1800 s	2.372	3.000
Colonel Blotto	mean score	1800 s	-0.248	0.500
MountainCar	mean reward	1800 s	53.26	100.0
Breakout	mean reward	1800 s	41.93	100.0
MetaMaze	mean reward	1800 s	20.12	100.0

Human-best references. We independently measure the Human Best anchors by evaluating strong human-designed reference solutions under each task's standard benchmark protocol. They are used only for cross-task reporting normalization.

B.2 Private Evaluators

We score all candidates with private evaluators that run outside the agent workspace. Test labels, evaluator state, and protected scoring files are inaccessible to the research agents and are restored before each validation. For Fashion-MNIST, CIFAR-10, and MNLI, we hold out a test split from the labeled data and keep its labels in the private evaluator. For House Prices, we use MLGym's split of the 1460 labeled rows (1022 train, 146 validation, 292 test) with test labels moved to the private directory. Image and language candidates run in a sandbox with network access disabled and task data exposed through a permission-enforcing local filesystem.

Game-theory evaluators. Each scoring plays 10 rounds $\times$ 10,000 Monte-Carlo simulations and reports the mean payoff per round. The evaluator isolates opponent randomness from the candidate, protects the opponent and scorer files, and rebuilds a clean evaluation directory for each scoring.

Control-task evaluators. MountainCar, Breakout, and MetaMaze share one evaluator, byte-identical to MLGym's up to the environment id and episode cap. A scoring loads the submission's five checkpoints and rolls each out on five fixed seeds with 256 vectorized environments (6,400 rollouts), reporting the mean reward. Rollouts are stochastic but fully determined by the fixed seeds within one software build; because frozen seeds produce different networks under different jax/flax versions, the baseline anchors in Table 2 are re-measured on our stack.

Language Modeling evaluator. The score is cross-entropy on a fixed 10,485,760-token FineWeb validation shard (320 deterministic batches of 32×1024 tokens, GPT-2 BPE). The training script never reads the validation shard; inside the sandbox the shard path is blanked, and the private scorer, which runs outside the sandbox, reads the real one.

B.3 MODELS, HARNESS, AND CANDIDATE EXECUTION

Executor and judge. All methods share the same Executor, Claude Haiku 4.5 operating a tool loop with read, write, run, and validate actions inside the candidate workspace. The Executor is frozen throughout; it implements the plan it is given and returns the candidate together with execution and validation feedback. A per-batch judge, also Haiku 4.5, discards malformed hypotheses (empty, off-schema, or inexecutable plans) before execution in every arm; a discarded hypothesis is regenerated and does not consume a scored candidate.

Policies and adapters. MELD uses Qwen3-4B-Instruct-2507 as the frozen backbone for both the Coordinator and the Scientists, served locally in bfloat16, with four independent LoRA adapters: one coordination adapter ϕ and one private Scientist adapter θ_i per unit (N=3). Each adapter has rank r=32, scaling α=64, no dropout, and is applied to the seven linear projections of every transformer block (q, k, v, o, gate, up, down); each adapter has 66M parameters. The Coordinator emits the complete portfolio in one schema-constrained JSON completion of at most 1200 tokens containing, for each unit, a parent identifier chosen from a retrieved shortlist of at most 16 tree nodes (at most two nodes from the same family) and a free-text directive with three fields (research question, observed bottleneck, constraints). Each Scientist emits a hypothesis and an execution plan of at most 3400 tokens for its assignment. All adapters are reset to their initial checkpoint at the start of every task.

Adaptation. Adaptation follows Section 4.3. A group consists of G=8 forks of one frozen snapshot, each executing all N=3 units, hence 24 scored candidates per group; Coordinator groups and Scientist groups alternate strictly, and each group is followed by one optimizer step on its active block, after which the updated serving adapters are used for the next group. When the remaining scored-candidate budget is insufficient for another full matched-snapshot group, the remainder is used for inference without an additional update. The leave-one-out advantage of Eq. (6) is standardized by the within-group standard deviation, broadcast to all response tokens, and combined with the asymmetric clipping range and the KL penalty to the frozen task-start policy. Table 3 lists the values.

Table 3: Test-time adaptation hyperparameters for MELD.

Backbone	Qwen3-4B-Instruct-2507 (frozen, bfloat16)
Adapters	1 Coordinator + 3 Scientist LoRA, r=32, α=64, dropout 0
Research units N / forks per group G	3 / 8
Candidates per group	$G \times N = 24$
Optimizer steps per group	1
Clipping range $[1-\epsilon_{\text{lo}}, 1+\epsilon_{\text{hi}}]$	$[0.8,\ 1.28]$
KL coefficient β_{KL} (to task-start policy)	0.01
Learning rate / optimizer / gradient clip	10^{-5} / AdamW / 1.0
Sampling temperature / top-p	1.0 / 1.0
Max new tokens, Coordinator / Scientist	1200 / 3400
Potential temperature τ	0.05
Parent shortlist size / family cap	16 / 2

B.4 BASELINE IMPLEMENTATIONS

The three baselines are re-implemented in the same harness as MELD while preserving the core search mechanisms of their cited methods. AIRA (Toledo et al., 2025) formulates research as graph-based search over candidate solutions: it selects previously evaluated nodes, applies LLM-driven Draft, Debug, and Improve operators, and adds the evaluated artifact back to the search graph. MLEvolve (Du et al., 2026) uses Progressive MCGS, augmenting tree search with graph-based reference edges for cross-branch information flow, a progressive exploration-to-exploitation schedule, and retrospective memory for task-specific experience reuse. CORAL (Qu et al., 2026) runs multiple autonomous research agents asynchronously; agents explore and reflect in parallel, communicate discoveries through shared persistent memory, and use periodic heartbeat-style interventions to sustain and redirect the search. All three use Haiku 4.5 as a frozen Scientist together with the shared Executor, evaluator, tree bookkeeping, and judge. We additionally run ARTS* (Juneja et al., 2026), the test-time-training variant of ARTS, in which a single reasoning Scientist inspects prior execution logs, diagnoses whether failures stem from faulty implementations or weak hypotheses, selects which

tree node to expand, and is periodically fine-tuned with LoRA to distill the search tree into its weights. ARTS* uses the Qwen3-4B-Instruct-2507 backbone of MELD together with the shared Executor, evaluator, tree bookkeeping, and judge.

B.5 ABLATION PROTOCOL

Variants and budget. We compare full MELD, Coordinator-only, Scientist-only, and Frozen on Iterated Prisoner's Dilemma (PD) and Colonel Blotto (CB), using three independent seeds per variant. The variants share the task-start policy initialization, backbone, prompts, sampling settings, Executor, judge, and tree bookkeeping. They differ only in which optimizer steps are enabled: full MELD updates both policy levels, Coordinator-only updates the Coordinator, Scientist-only updates the private Scientist adapters, and Frozen disables both updates while retaining the same tree-based search. Each run uses 72 candidate evaluations, comprising two 24-candidate adaptation groups and a final 24-candidate evaluation batch reserved before the run. The reserved batch is excluded from training. Timeouts and validation failures consume budget in every variant. Model and optimization settings follow Appendix B.3.

Adaptation schedule. The first group, denoted b_{pre}, is collected from the task-start snapshot before any parameter update. Following the Coordinator-group protocol of Section 4.3, it consists of $G = 8$ independent forks, each sampling a fresh joint portfolio for $N = 3$ research units and executing their Scientists' responses. The snapshot and behavior policies remain fixed during collection. Full MELD and Coordinator-only then take one optimizer step on the Coordinator adapter; Scientist-only and Frozen skip this step. Validated candidates are committed to the discovery tree in every variant.

The second group, denoted b_{mid}, follows the Scientist-group protocol. From the resulting snapshot, the current Coordinator samples one portfolio, which is held fixed across eight forks. Each Scientist independently samples its response in each fork, producing another 24 candidate evaluations. Full MELD and Scientist-only then take one optimizer step on each private Scientist adapter, with these updates performed in parallel; Coordinator-only and Frozen skip this step. Validated candidates are again committed to the tree. After each enabled update, the updated serving adapters are used for subsequent generation. Full MELD therefore completes one Coordinator–Scientist adaptation cycle before the final batch is collected.

Post-cycle evaluation. After the second adaptation group has been processed and all enabled updates applied, we generate a fresh 24-candidate batch b_{post} using the resulting policies. This batch consists of eight independent forks of the post-adaptation snapshot. Each fork samples a new joint portfolio and fresh responses from all three Scientists. Thus, b_{pre} and b_{post} use the same portfolio-resampling scheme, whereas b_{mid} holds one portfolio fixed for Scientist training.

All policy parameters, the discovery state, and the parent shortlist remain fixed throughout collection of b_{post}. Candidates and feedback from one evaluation fork are not exposed to the others, and no optimizer step is taken on this batch. Each variant is evaluated from the tree produced by its own adaptation groups, so the comparison includes how the enabled updates shape the subsequent discovery process.

Reported statistics. A valid candidate counts as a threshold hit when its raw private-evaluator score (Appendix B.2) is at least 2.62 on PD or 0.245 on CB, and the same thresholds apply to all variants and seeds. For each seed and each reported batch, we compute the median raw score over valid candidates and the number of threshold hits among the 24 attempted evaluations. Failed evaluations count as non-hits.

Table 4 reports $b_{\mathrm{pre}} \rightarrow b_{\mathrm{post}}$, averaging each statistic over the three seeds. Batch medians are computed separately within each seed before averaging, rather than by pooling candidates across seeds. Both statistics use only candidates generated in the corresponding batch, not candidates accumulated in the tree.

B.6 AGGREGATE METRICS

Figure 2 is computed from the three-seed mean per-task best scores of Table 1 and the anchors of Table 2. Scores are direction-corrected and normalized against the shipped baseline and human-best

Table 4: Adaptation ablation on Prisoner's Dilemma (PD) and Colonel Blotto (CB), averaged over three independent seeds. Each entry reports $b_{\text{pre}} \to b_{\text{post}}$: the initial 24-candidate batch and a fresh 24-candidate batch generated after all enabled updates in one scheduled Coordinator–Scientist cycle. "Median" is the mean of the per-seed medians over valid candidates; "hits" is the mean number of threshold hits (score at least 2.62 on PD or 0.245 on CB) per 24 attempted evaluations.

	PD		CB	
Variant	Median	Hits / 24	Median	Hits / 24
Full MELD	$2.35 \to 2.57$	$2 \to 10$	$0.000 \to 0.010$	$2 \to 6$
Coordinator-only	$2.35 \to 2.44$	$2 \to 5$	$0.000 \to 0.004$	$2 \to 3$
Scientist-only	$2.35 \to 2.48$	$2 \to 6$	$0.000 \to 0.006$	$2 \to 4$
Frozen	$2.35 \to 2.37$	$2 \to 3$	$0.000 \to 0.002$	$2 \to 2$

reference, $z = \text{clip}_{[0,1]}\left((\tilde{s} - \tilde{s}_0)/(\tilde{s}_{\text{human}} - \tilde{s}_0)\right)$, where $\tilde{s}$ increases with quality; Table 5 lists the resulting values. Median, Mean, and IQM (the mean of the middle two quartiles) are taken over the 12 tasks; Optimality Gap is $\mathbb{E}[1 - z]$, which under clipping equals $1 - \text{Mean}$. Confidence intervals are 95% percentile intervals from a bootstrap that resamples the 12 task-level three-seed means with replacement (5,000 replicates, fixed bootstrap seed). Titanic clips to $z{=}1$ for every method, since all exceed the 0.830 reference.

Table 5: Normalized, clipped scores z behind Figure 2.

Task	MLEvolve	AIRA	CORAL	ARTS*	MELD
Titanic	1.00	1.00	1.00	1.00	1.00
House Prices	0.45	0.45	0.40	0.46	0.57
Fashion-MNIST	0.74	0.82	0.79	0.81	0.91
CIFAR-10	0.92	0.94	0.94	0.93	0.93
MNLI	0.79	0.79	0.79	0.79	0.79
Language Modeling	0.49	0.58	0.60	0.59	0.60
Battle of the Sexes	0.66	0.66	0.66	0.66	0.66
Prisoner's Dilemma	0.43	0.42	0.42	0.43	0.44
Colonel Blotto	0.67	0.67	0.67	0.67	0.67
MountainCar	0.31	0.38	0.61	0.67	0.86
Breakout	0.14	0.55	0.51	0.63	0.74
MetaMaze	0.48	0.65	0.69	0.69	0.76

B.7 PROTOCOL AND COMPUTE

Budget accounting. Within each task, all methods receive the same nominal scored-candidate budget, so comparisons are matched by scored candidates rather than wall-clock time. Every candidate is executed under the task training timeout and scored by the private evaluator. A candidate that exceeds the timeout or fails validation consumes one evaluator call, receives failure utility $s_{\min}{=}0$ for the training potential, and is not added to the derivation tree.

Scores and seeds. Main benchmark results and adaptation ablations are averaged over three independent seeds. Table 1 reports the mean of each seed's best held-out score within budget, and Figure 1 aggregates the corresponding anytime best-so-far trajectories. Table 4 analogously averages the per-seed ablation statistics. Normalized scores use the anchors of Table 2; this reporting scale is distinct from the bounded training utility $s(c) = g_m(q_m(c)) \in [0, 1]$ used by the potential, which serves only task-time adaptation.

Compute. Experiments run on four NVIDIA A100 GPUs. Each method run takes approximately 9 hours, with test-time adaptation updating only the LoRA blocks described in Table 3.

Broader impact. More capable automated discovery systems can reduce the cost of exploring large hypothesis spaces and accelerate iterative scientific and machine-learning experimentation. The same

autonomy can also propagate erroneous hypotheses or be applied in sensitive domains; practical deployments should therefore retain task-appropriate evaluation and human oversight.