From Individual Excellence to Collective Sustainability: Seeking Strategic Equilibrium in Proactive Multi-Agent Teams

Anonymous ACL submission

Abstract

In heterogeneous team settings, proactive AI agents often suffer from Collaborative Myopia: a greedy optimization for immediate task accuracy that ignores long-term team longevity. This leads to the Individual-Centric Trap, where experts (e.g., PIs) are disproportionately overloaded while junior roles remain underutilized, incurring unobserved opportunity costs that erode collective sustainability. To resolve this efficiency-sustainability coupling problem, we propose GT-PMARL (Game-Theoretic Proactive Multi-Agent Reinforcement Learning). We reformulate team coordination as a strategic game that internalizes opportunity cost. Our framework employs: (1) a Positive-Unlabeled scorer to anchor intervention quality under sparse supervision, and (2) a Nash-Pareto competitive objective to seek an equilibrium between individual task excellence and collective load balancing. Empirical experiments on scientific workflows show that GT-PMARL effectively maintains high performance while preventing expert over-exploitation. Our work provides a game-theoretic foundation for building sustainable, balanced human-AI collaborative ecosystems.[1]

1 Introduction

Large language models (LLMs) have demonstrated significant progress in scientific task execution and workflow coordination (Hou et al., 2025; Schick et al., 2023; Qin et al., 2023). In modern digital laboratories, standardized tool interfaces like Model Context Protocol (MCP) have substantially expanded the external resources available to these agents, from code retrieval to biomedical data analysis (Liu et al., 2024; Wu et al., 2024a). However, this technological progress introduces a critical challenge in team coordination settings: how can LLM-based coordinators allocate tasks across heterogeneous team roles to simultaneously optimize for immediate task accuracy and long-term team sustainability? We refer to this as the efficiency-sustainability coupling problem.

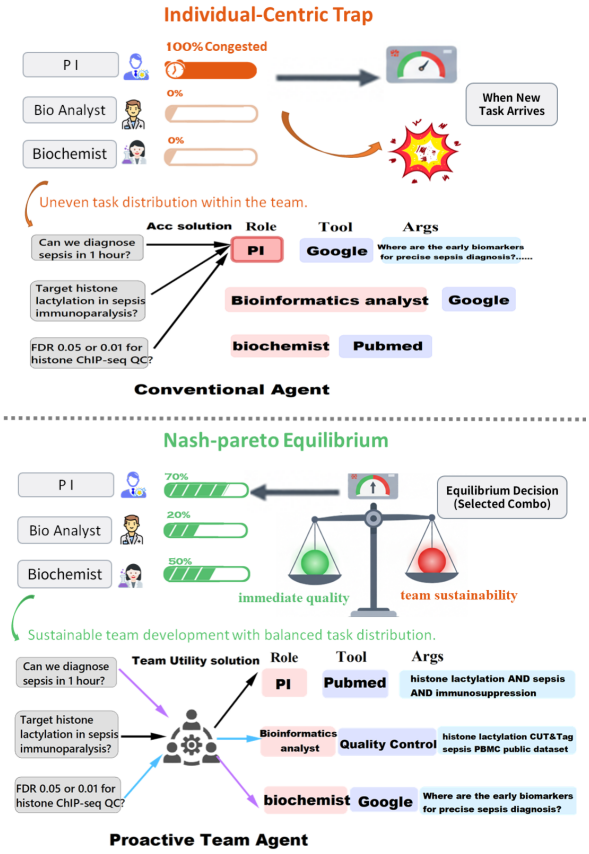

Figure 1: Top: Conventional agents fall into the Individual-Centric Trap, where tasks are greedily routed to the most capable member. Bottom: Our method seeks a Nash-Pareto Equilibrium, balancing immediate task quality with long-term team sustainability.

In practice, team coordination often defaults to accuracy-maximization heuristics, such as expert routing (Lin et al., 2024), multi-agent reinforcement learning (Gorsane et al., 2022), and reactive planning paradigms like ReAct (Yao et al., 2023). However, these frameworks overlook the systemic implications of cognitive overload and resource constraints in collaborative environments with fi-

[1] Code and data are provided in the supplementary material and will be made publicly available upon acceptance.

nite human bandwidth (Liang et al., 2022). Specifically, under reward signals calibrated to prioritize task completion, existing models exhibit the phenomenon of Collaborative Myopia: a tendency to concentrate the task load on the most capable individuals, thereby undermining long-term team performance (Qin et al., 2022). From an economic perspective, these systems neglect the Opportunity Cost of expending the limited time of domain experts on suboptimal task matching (Barron et al., 2024; Paz, 2025). As shown in Fig. 1, this leads to the Individual-Centric Trap, where high-stakes tasks are disproportionately assigned to Principal Investigators (PI), maximizing short-term metrics such as accuracy, while fragmenting the team's decision-making capacity and leaving junior roles underutilized.

Addressing the efficiency-sustainability coupling in scientific coordination requires navigating a delicate tension between immediate task fidelity and long-term team capacity building. This interaction constitutes a coupled constraint where resource distribution and solution quality function as inseparable facets of a unified optimization challenge. Beyond surface-level trade-offs, resolving this requires overcoming three systemic barriers.

Specifically, we identify the following. First, structural blindness to opportunity cost forces current agents into semantic matching rather than strategic allocation. By failing to internalize resource ceilings (e.g., expert bandwidth), models inadvertently succumb to the "Individual-Centric Trap," where misallocating high-value human capital to trivial tasks incurs an unobserved opportunity cost that erodes long-term team sustainability. Second, the combinatorial explosion of proactive planning necessitates navigating a tripartite *Role* $\times$ *Tool* $\times$ *Task* space to preemptively mitigate bottlenecks. This strategic requirement for high-quality "who-what-how" configurations exceeds the zero-shot capabilities of reactive LLMs, which struggle to simulate complex future constraints. Finally, the sparsity and equifinality of evaluative signals in collaborative data hinder policy refinement. Real-world trajectories typically offer only sparse successful interacting instances with multiple feasible paths, providing insufficient negative supervision to distinguish locally viable decisions from a globally optimal, sustainable equilibrium (Jaskie and Spanias, 2022).

To rectify these deficiencies, we propose GT-PMARL (Game-Theoretic Proactive Multi-Agent Reinforcement Learning), a dual-layer coordination framework that internalizes team sustainability as a first-class optimization constraint. We treat scientific collaboration as a multi-agent game where equitable and efficient allocations emerge from equilibrium dynamics rather than centralized decrees. Our work makes three primary contributions:

- **Dual-Layer Strategic Coordination.** We resolve the trade-off between exploration and exploitation via a hierarchical flow. The upper layer generates a diverse "Proactive Candidate Pool" under the quality constraints of a Positive-Unlabeled (NNPU) scorer. The lower layer refines these via MARL-based negotiation, where signals flow bidirectionally to contract the strategy space toward practical success.

- **Equilibrium as a Multi-Objective Optimizer.** We formulate task allocation as a competitive game where Nash Equilibrium and Pareto constraints prevent expert over-exploitation. By internalizing collective sustainability thresholds, our objective ensures that individual efficiency never compromises the team's long-term capacity.

- **Self-Amplifying Learning from Sparse supervision.** To overcome the scarcity of scientific failure labels, we leverage offline PU-learning to generalize from sparse successful interactions. This enables a self-amplifying supervision signal that discovers novel coordination patterns, providing a theoretical foundation for efficiency-equity-balanced human-AI ecosystems.

2 Related Work

The development of tool-augmented language models has marked a significant step in creating capable AI agents (Schick et al., 2023; Qin et al., 2023; Wu et al., 2024b). These frameworks, often leveraging reinforcement learning from human or model feedback to refine their policies (Yuan et al., 2023), have proven effective at decomposing tasks and interacting with external APIs. However, they predominantly operate under a single-agent, reactive paradigm, focusing on fulfilling a user's immediate request. This approach falls short in complex collaborative settings, which are inherently

multi-agent and require proactive resource allocation among team members with diverse expertise (O'Sullivan, 2003). While some work has explored multi-modal agents for specialized domains like medicine (Young et al., 2024), they still do not address the fundamental strategic dilemma of balancing team efficiency with equitable workload distribution. Our work bridges this gap by conceptualizing the problem not as a simple instruction-following task, but as a multi-agent coordination challenge.

Game-theoretic frameworks have been studied extensively to address the complexities of multi-agent interaction (Yang and Wang, 2020). A primary challenge in multi-agent reinforcement learning (MARL) is that decentralized agents often treat others as static parts of the environment, failing to account for their evolving strategies(Chen et al., 2025; Li et al., 2024; Ren et al., 2025). More advanced methods explicitly model the learning behavior of other agents to improve stability and foster cooperation (Hernandez-Leal et al., 2017). Our approach aligns with this latter line of inquiry but advances it by adopting a hierarchical game structure. We draw inspiration from Stackelberg games (Zheng et al., 2022; Ahamed et al., 2024; Chen et al., 2024), where a "leader" player's action constrains the subsequent "follower" players' best responses (Fiez et al., 2020). In our framework, the upper-layer planner acts as the leader, defining the strategy space (the candidate pool), while the lower-layer agents act as followers, competing to find an equilibrium within that space. This structure, combined with our use of Positive-Unlabeled (PU) learning to create a dense reward signal from sparse data, allows our system to implicitly optimize the trade-off between efficiency and fairness, a problem that remains a significant hurdle in both MARL and collaborative AI.

3 Preliminaries

In this section, we formalize the scientific team coordination task as a dynamic decision-making process and define the metrics for evaluating coordination quality and team sustainability.

3.1 Modeling Scientific Team Collaboration

Team State Representation. We model the operational state of a heterogeneous scientific team at each time step t as a composite triplet $\mathcal{T}_t = \langle \mathbf{w}_t, \mathbf{c}_t, \mathbf{s}_t \rangle$. Specifically, for a team consisting of N roles $\mathcal{R} = \{1, \ldots, N\}$, $\mathbf{w}_t = [w_1^t, \ldots, w_N^t]^\top \in \mathbb{R}_+^N$ represents the workload vector, where each scalar w_r^t quantifies the accumulated cognitive load and task pressure on role r. Complementary to this, $\mathbf{c}_t = [c_1^t, \ldots, c_N^t]^\top \in \mathbb{R}_+^N$ denotes the capacity vector, capturing the available bandwidth, resource quotas, and expertise-fit of each member. Finally, $\mathbf{s}_t \in \mathbb{R}^d$ encapsulates the task context, reflecting specific research requirements, urgency, and global constraints such as API limits or timeline milestones.

Allocation Decision and Team Dynamics. Given the state $\mathcal{T}_t$ and a task query q_t, the coordination agent must determine an allocation decision $\mathbf{a}_t = \langle f, \theta, r \rangle$, where $f \in \mathcal{F}$ identifies the external tool or resource, $\theta \in \Theta_f$ specifies the invocation parameters, and $r \in \mathcal{R}$ is the target role. The execution of an allocation imposes a load increment $\Delta w(\mathbf{a}_t, r)$ on the assigned role, driving the team state transition as follows:

$$w_r^{t+1} = w_r^t + \Delta w(\mathbf{a}_t, r) \tag{1}$$

For all non-assigned roles $r' \neq r$, the workload remains constant (i.e., $w_{r'}^{t+1} = w_{r'}^t$), reflecting the persistent nature of cognitive load in collaborative workflows.

Proactive Candidate Pool. To expand the decision space beyond reactive matching, our framework initiates the coordination process by generating a *Proactive Candidate Pool* $\mathcal{A}_t = \{\mathbf{a}_1, \ldots, \mathbf{a}_M\}$. This pool consists of M diverse, feasible allocation triplets that serve as the action space for the subsequent strategic selection.

3.2 Multi-Objective Evaluation Metrics

Effective coordination must balance immediate task performance with the long-term health of the team. We characterize this trade-off through two primary dimensions.

Task Execution Quality. The immediate utility of an allocation $\mathbf{a}$ is quantified by a quality scorer $S_\phi : \mathcal{A} \times \mathcal{T} \to [0, 1]$. This function estimates the conditional probability of task success:

$$Q(\mathbf{a} \mid \mathcal{T}_t) = P(\text{Success} \mid \mathbf{a}, \mathcal{T}_t) \approx S_\phi(\mathbf{a} \mid \mathcal{T}_t) \tag{2}$$

As detailed in Section 4.1, we leverage Positive-Unlabeled (PU) learning to train S_ϕ, addressing the inherent sparsity of expert-labeled successful coordination data.

Team Sustainability and Fairness. To avoid the *Individual-Centric Trap*—where senior experts

are over-utilized to maximize short-term accuracy at the cost of team exhaustion, we measure workload inequality using the Gini coefficient over the updated state $\mathbf{w}_{t+1}$:

$$\text{Gini}(\mathbf{a}, \mathcal{T}_t) = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} |w_i^{t+1} - w_j^{t+1}|}{2N^2 \bar{w}_{t+1}} \quad (3)$$

where $\bar{w}_{t+1}$ is the mean workload across all roles.

3.3 Allocation as Dynamic Resource Balancing

We reformulate the allocation challenge as a dynamic resource-balancing problem that internalizes the *Opportunity Cost* of human capital. By introducing a state-adaptive coefficient $\gamma_t(r) \in \mathbb{R}_+$, which represents the marginal cost of assigning work to role r, the optimal allocation $\mathbf{a}_t^*$ is selected by maximizing the following sustainability-aware objective:

$$\mathbf{a}_t^* = \arg\max_{\mathbf{a} \in \mathcal{A}_t} \{S_\phi(\mathbf{a} \mid \mathcal{T}_t) - \gamma_t(r) \cdot \Delta w(\mathbf{a}, r)\} \quad (4)$$

In this formulation, $\gamma_t(r)$ acts as a dynamic penalty that increases when role r approaches its capacity ceiling, effectively discouraging further burdening of overloaded roles. Solving this reformulated task requires the synergistic optimization of the PU-based scorer S_ϕ, the diversity-driven generation of $\mathcal{A}_t$, and the equilibrium-based learning of coordination coefficients, all of which are elaborated in the subsequent section.

4 Methodology

As show in Fig. 2, we present the architecture of **GT-PMARL** (Game-Theoretic Proactive Multi-Agent Reinforcement Learning), a dual-layer coordination framework designed to resolve the *efficiency-sustainability coupling* in scientific research teams. Unlike traditional methods that rely on reactive execution or greedy semantic matching, our framework adopts a "Generate-Negotiate" paradigm.

The architecture consists of two tightly coupled layers: (1) an **upper-layer** generative policy that proposes a high-entropy candidate pool under structural constraints, and (2) a **lower-layer** coordination engine that achieves a game-theoretic equilibrium among heterogeneous roles. Central to this framework is a Non-Negative Positive-Unlabeled (nnPU) scorer (Kiryo et al., 2017), which acts as a learned quality prior. It filters suboptimal configurations in the upper layer and serves as the foundational reward signal in the lower layer, ensuring that the team remains within a "scientific survival baseline" while pursuing equitable load distribution.

4.1 Strategy-Aware Quality Prior

A pivotal challenge in scientific coordination is the lack of explicit "failure" data—we know which allocations worked historically, but rarely what could have been optimized further. To resolve this, we propose a Non-Negative Positive-Unlabeled (nnPU) scoring framework. Rather than acting as a simple binary classifier, our scorer S_ϕ serves as the foundational prior and the structural linchpin that orchestrates the learning dynamics across both the generative and competitive layers of GT-PMARL.

4.1.1 Capturing Structural Harmony through Triplet Encoding

Existing models often treat task allocation as a flat semantic matching problem, failing to perceive the implicit logic required for scientific execution. We address this by designing a Structural Harmony Function $\mathcal{H}(\mathbf{a})$. Unlike vanilla encoders, $\mathcal{H}(\mathbf{a})$ is engineered to decipher the latent coordination grammar within the allocation triplet $\mathbf{a} = \langle r, f, \theta \rangle$:

$$\mathcal{H}(\mathbf{a}) = \text{MLP}\left(\mathbf{e}_q \oplus \mathbf{e}_{role} \oplus \mathbf{e}_{tool} \oplus \mathbf{e}_{args}\right) \quad (5)$$

where $\mathbf{e}$ represents deep contextual embeddings from the pre-trained `gte-Qwen2-7b` encoder. By utilizing dynamic feature fusion ($\oplus$), S_ϕ transcends keyword co-occurrence to internalize conditional dependencies, for instance, recognizing that the utility of a *Data Analyst* (r) is strictly conditioned on the choice of a *Statistical Tool* (f) and the rigor of its *Parameters* (θ). This structural awareness allows the model to penalize "semantic dissonance," providing a multi-dimensional reality check for all downstream strategy exploration.

4.1.2 Distilling Quality via Non-Negative Risk Minimization

To robustly distill this coordination logic from an unlabeled set $U = P \cup N$ (where successful trajectories P are sparse), we employ the Non-Negative Positive-Unlabeled (nnPU) objective. Given the class prior $\pi_p = \mathbb{P}(y = 1)$, we formulate the risk estimator to prevent the model from overfitting to

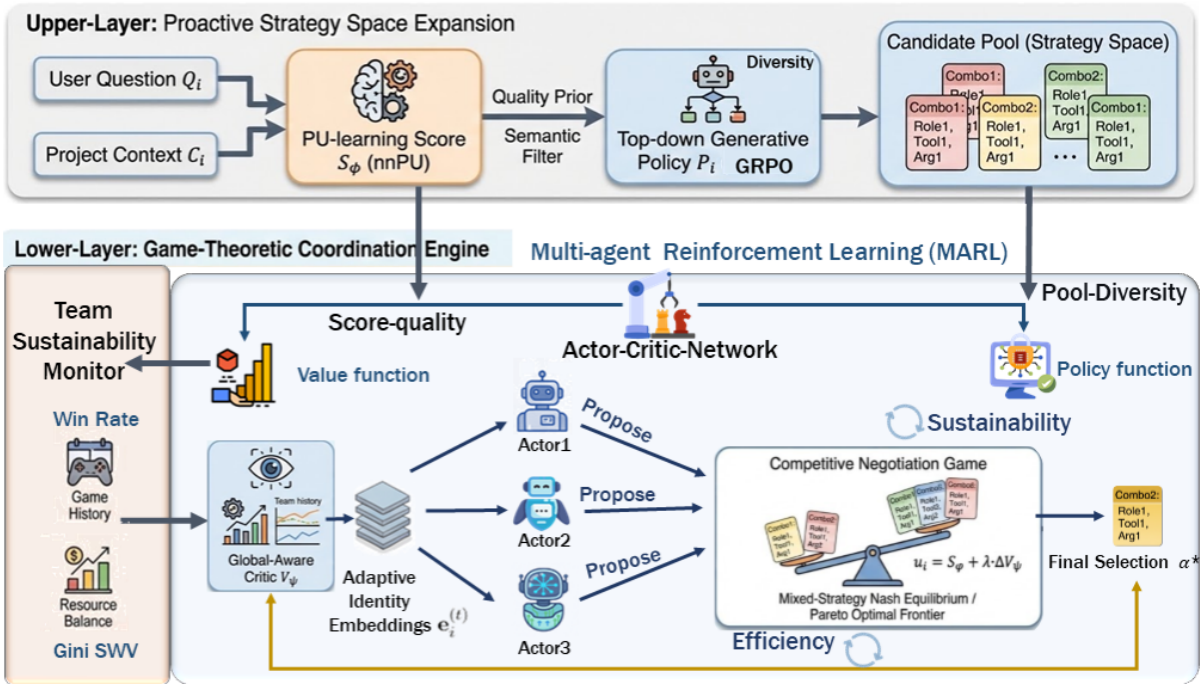

Figure 2: Overview of the GT-PMARL framework. The Upper-Layer (Strategy Expansion) generates a high-entropy candidate pool using Diversity-GRPO guided by the nnPU-based quality prior. The Lower-Layer (Coordination Engine) employs MARL to model the allocation as a competitive negotiation game. PU-scores serve as the "Rational Grounding" to mediate the trade-off between individual task efficiency and collective team sustainability, ultimately reaching a Nash-Pareto equilibrium.

latent negatives in U:

$$\mathcal{L}_{nnPU}(\phi) = \pi_p \mathcal{L}_P^+(S_\phi) + \max\{0, \mathcal{L}_U^-(S_\phi) - \pi_p \mathcal{L}_P^-(S_\phi)\} \quad (6)$$

where $\mathcal{L}^+$ and $\mathcal{L}^-$ denote the sigmoid loss for positive and negative classes, respectively. By enforcing a non-negative constraint on the estimated negative risk, we derive a robust scoring function $S_\phi(\mathbf{a} \mid \mathcal{T}_t) \in [0, 1]$ that serves as the Scientific Survival Baseline, a rigorous measure of whether a proposed allocation is operationally viable.

The uniqueness of S_ϕ lies in its dual-role as a connective tissue, synchronizing the two distinct training phases of our framework:

Generative Navigation. In the GRPO phase, S_ϕ functions as a *Quality Compass*. It projects a "success landscape" onto the vast combinatorial search space, providing a dense reward signal that guides the LLM to explore functionally diverse strategies without straying into the territory of scientific invalidity.

Equilibrium Grounding. In the MARL negotiation phase, S_ϕ provides the *Rational Grounding*. It acts as the "Base Payoff" that prevents agents from converging to "fair but scientifically invalid" allocations. By embedding S_ϕ into the game-theoretic reward, we ensure that team equity is pursued only upon a solid foundation of execution efficiency.

Through this clever structural design and cross-layer integration, the scorer ensures that the entire framework moves beyond simple semantic matching and toward **strategic resource optimization**.

4.2 Upper-Layer: Diversity-Driven Pool Generation

To resolve the *Collaborative Myopia* challenge, the upper layer must proactively identify a candidate set that is both high-quality and functionally heterogeneous. We optimize a generative policy π_θ using Group Relative Policy Optimization (GRPO), which refines strategies by comparing relative advantages within groups.

Structural Constraint and Strategy Menu Given a task context $\mathbf{s}_t$, π_θ generates a candidate pool $\mathcal{P}_t = \{\mathbf{a}_1, \ldots, \mathbf{a}_K\}$. To ensure $\mathcal{P}_t$ provides sufficient strategic maneuvering space for the lower layer, we define a composite reward R_{pool} to penalize functional redundancy:

$$R_{\text{pool}}(\mathcal{P}_t) = \omega_1 \bar{S}_\phi(\mathcal{P}_t) + \omega_2 \mathcal{D}(\mathcal{P}_t) + \omega_3 \mathcal{C}(\mathcal{P}_t) \quad (7)$$

where $\bar{S}_\phi$ is the mean coherence score from the $nnPU$ prior, ensuring scientific validity, and $\mathcal{C}$ enforces format compliance.

The diversity term $\mathcal{D}(\mathcal{P}_t)$ is quantified by the Taxonomic Coverage across three functional pillars:

$$\mathcal{D}(\mathcal{P}_t) = \frac{1}{3} \sum_{j \in \{U,E,C\}} \mathbb{I}(\mathcal{P}_t \cap \text{Pillar}_j \neq \emptyset) \quad (8)$$

where $\{U, E, C\}$ denote the Utility, Execution, and Coordination pillars. This incentivizes the policy to provide a "strategic menu" rather than collapsing into semantically identical PI-centric strategies.

4.3 Lower-Layer: Coordination via Implicit Equilibrium

Once the upper layer established the strategic possibilities, the coordination problem is transformed into a Strategic Selection Challenge: selecting a single strategy while simultaneously optimizing for immediate quality and long-term team sustainability. We model this as an Implicit Multi-Agent Game where team health is internalized as an equilibrium constraint.

4.3.1 Global-Aware Payoff Structure

Each agent i proposing a strategy $\mathbf{a}_i \in \mathcal{P}_t$ receives a payoff $\mathcal{U}_i$ that internalizes team-wide dynamics. This function acts as the objective for individual agents while respecting global constraints:

$$\begin{aligned} \mathcal{U}_i(\mathbf{a}_i \mid \mathcal{S}_t) = & \underbrace{S_\phi(\mathbf{a}_i)}_{\text{Quality}} + \beta \cdot \underbrace{\Delta V_\psi(\mathbf{a}_i, \mathcal{S}_t)}_{\text{Momentum}} \\ & - \lambda_t \cdot \underbrace{\Psi_i(\mathbf{a}_i, \mathcal{S}_t)}_{\text{Shadow Price}} \end{aligned} \quad (9)$$

Base Quality (S_ϕ): Grounded by the $nnPU$-prior, this ensures no strategy proposing an ineffective tool-role pairing can win the negotiation.

Sustainability Momentum (ΔV_ψ): A global-aware Critic V_ψ maintains a spatio-temporal estimate of the team state. The momentum is defined as:

$$\Delta V_\psi = \mathbb{E}[V_\psi(\mathcal{S}_{t+1}) \mid \mathbf{a}_i] - V_\psi(\mathcal{S}_t) \quad (10)$$

Negative values penalize strategies that deplete expert bandwidth or create future bottlenecks.

Shadow Price of Labor (Ψ_i): To prevent the *PI-Centric Trap*, we define a dynamic congestion penalty acting as a shadow price for specific human capital:

$$\Psi_i = \begin{cases} \tau \cdot (\frac{\text{Load}_i(t)}{\text{Cap}_i} - \theta) & \text{if overloaded} \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

where $\text{Load}_i(t)$ is the cumulative workload. As a role approaches saturation, its shadow price escalates, making it "economically expensive" within the equilibrium. The dynamic coefficients, particularly the shadow price λ_t, are derived from the team's historical state evolution (detailed in Appendix C)

4.3.2 Adaptive Identity and Specialization

A central Critic V_ψ maintains a spatio-temporal memory of team performance. It monitors role-specific metrics, such as cumulative win-rates and workload variance, to adjust the "cost of labor" in Eq.(9). This mechanism transforms the PI's capacity from a static resource into a dynamic constraint.

Simultaneously, agents are not static entities but evolving specialists. To encourage role-specific expertise, each agent is characterized by a learnable identity embedding $\mathbf{e}_i^{(t)}$. These embeddings are refined via gradient-based updates:

$$\mathbf{e}_i^{(t+1)} \leftarrow \mathbf{e}_i^{(t)} + \eta \nabla_{\mathbf{e}_i} \mathcal{J}_i \quad (12)$$

This allows agents to develop their Functional Niches, autonomously selecting strategies from $\mathcal{P}_t$ that align with their specialized capabilities and the team's evolving state.

4.3.3 Nash-Pareto Equilibrium Resolution

The final allocation $\mathbf{a}^*$ is found by computing a mixed-strategy Nash equilibrium σ^* restricted to the Pareto-optimal frontier $\mathcal{F}_P$ of all proposals:

$$\begin{aligned} \sigma_i^{(m+1)} & \leftarrow \text{BR}_i(\sigma_{-i}^{(m)}) \\ & = \arg\max_{\sigma_i'} \mathbb{E}_{\sigma_{-i}^{(m)}}[\mathcal{U}_i(\sigma_i', \sigma_{-i}^{(m)})] \end{aligned} \quad (13)$$

We solve this via Iterative Best Response (IBR) until convergence, representing the agents' attempt to reach a consensus under the Global-Aware Payoff. This typically converges in 2–3 iterations, yielding a selection that is both scientifically optimal and sustainable.

Upon executing $\mathbf{a}^*$, the real reward r_{real} triggers a synergistic feedback loop. The Critic refines V_ψ via TD-learning, while successful equilibrium selections are distilled into preference pairs to retrain the upper-layer policy.

5 Experiments

5.1 Dataset Construction

To evaluate the proposed framework, we constructed a novel benchmark by mining biomedical literature, specifically focusing on complex research scenarios that lack existing multi-agent modeling. We sourced 21,007 PubMed articles related to sepsis and filtered them into a high-quality subset of 11,792 multidisciplinary papers. Using this corpus, we employed a two-stage generation process: (1) extracting problem-solution pairs as positive instances (P); (2) annotating each solution with a ground-truth (Role, Tool, Argument) triplet. Remaining plausible but unverified action combinations formed our unlabeled set (U). This Positive-Unlabeled (PU) structure enables training a robust scorer to navigate a vast, sparsely labeled coordination space. The dataset is partitioned following an 8:1:1 ratio for training, validation, and testing.

Real-World Ecological Validation. To ensure practical validity, we curated a specialized dataset from four professional research teams (one technical logistics firm and three university labs) over 12 weeks. Using a custom-built assistant system, we conducted human-in-the-loop A/B testing to evaluate our framework on real-task queries and annotations. See Appendix A for demographics.

5.2 Experimental Setup

We benchmark our framework against three tiers of competitive baselines: (i) Base Models, including Llama-3B-base and the proprietary GPT-4o-base; (ii) Fine-tuning & RLHF Methods, which encompass standard SFT and GRPO variants on both Llama and Qwen backbones; and (iii) Advanced Agentic Paradigms, specifically the Qwen-plus series utilized in zero-shot and ReAct settings. All experiments were conducted on a computation node equipped with eight 80GB NVIDIA A800 GPUs. For the lower-layer MARL coordination engine, we instantiate three agents with a hidden dimension of 256, utilizing learning rates of 1×10^{-4} for actors and 1×10^{-8} for the critic over 2,000 training episodes. To ensure the robustness of our results, our proposed method (*ours*) is evaluated across three independent runs using random seeds 42, 215, and 3407, with the final results reported as the mean performance.

Model	Recall@1	Recall@2	Recall@5	Diversity
Qwen2.5-3B + GRPO	0.2636	0.3818	**0.5818**	0.5781
GT-PMARL (Ours)	**0.2727**	**0.4091**	0.5636	**0.9809**

Table 1: Performance of Candidate Generation (Upper-layer). The strategy-aware prior ensures high taxonomic diversity while preserving retrieval quality.

Method	Top-1 Acc	F1	MRR	Conf.
Balance Selection	0.3920	0.2633	0.5110	0.5680
Pareto Optimality	**0.4350**	**0.2921**	0.4692	0.7580
Nash Equilibrium	0.3780	0.2561	0.5098	0.7690
GT-PMARL (Ours)	0.4330	0.2846	**0.5225**	**0.7750**

Table 2: Selection Performance of the Lower-layer. Our game-theoretic engine provides the most stable ranking (MRR) and highest strategic confidence.

5.3 Results and Analysis

Validation of Proactive Team. As shown in Table 1, our diversity-driven generation policy (Ours-Upper) significantly outperforms GRPO in Taxonomic Diversity (0.9809), despite a slight trade-off in Recall@5. This functional richness in the candidate space enables the lower-layer to escape the "Individual-Centric Trap. "

Table 2 evaluates the lower-layer's coordination logic. GT-PMARL (Ours) achieves the best MRR (0.5225) and highest Avg. Confidence (0.7750), demonstrating that internalizing the "Scientific Survival Baseline" via PU-learning yields more robust strategic consensus than vanilla Nash solvers.

Primary Performance Superiority. As demonstrated in Table 3, GT-PMARL significantly outperforms all competitive baselines across every critical dimension. Notably, it achieves a Tool Selection Accuracy of 0.6273 and a Role Assignment Accuracy of 0.6000, surpassing the strongest proprietary models by 10.0% and 15.4% absolute margin, respectively. This performance gap indicates that our tiered coordination mechanism—which explicitly models the trade-off between task difficulty and role capacity—is fundamentally more effective for scientific delegation than simply scaling model parameters.

Fidelity and Strategic Argumentation. Beyond classification accuracy, we evaluate the execution logic via ROUGE-L. Our model achieves a score of 0.5563, a 63.8% relative improvement over Qwen-7b-SFT (0.3396). This suggests that our proactive candidate pool, filtered by the nnPU-prior, effectively prunes "semantically dissonant" strategies, providing high-precision arguments that match the

Models	Ans.Tool				Ans.Role				Ans.Args			Team Sustainable	
	Acc.	P.	R.	F1	Acc.	P.	R.	F1	R-1	R-2	R-L	Gini	SWV
Base Models													
llama-3b-base	0.3091	0.2715	0.3091	0.2687	0.4364	0.3904	0.4364	0.4005	0.2091	0.0429	0.1431	0.3697	0.1167
gpt-4o-base	0.4000	0.3811	0.4000	0.3665	0.3818	0.3570	0.3818	0.3575	0.1984	0.0750	0.1984	0.3273	0.1660
Supervised Fine-tuning (SFT) & RLHF Methods													
llama-3b-SFT	0.3200	0.3613	0.3200	0.2296	0.4200	0.5105	0.4200	0.4062	0.4264	0.1903	0.3088	**0.1467**	0.2862
Qwen-3b-SFT	0.2000	0.2185	0.2000	0.2079	0.2000	0.2042	0.2000	0.1962	0.3748	0.1308	0.2594	0.2400	0.2283
Qwen-3b-GRPO	0.3091	0.2652	0.1255	0.1468	0.2273	0.1182	0.0724	0.2596	0.1841	0.0319	0.1290	0.3212	0.1083
Qwen-3b-SFT+GRPO	0.2727	0.1691	0.1511	0.1574	0.2455	0.2032	0.1710	0.1505	0.1786	0.0328	0.1300	0.3515	0.1071
llama-3b-SFT+GRPO	0.3455	0.2150	0.0217	0.2068	0.2727	0.1933	0.2055	0.1519	0.0540	0.0096	0.0435	0.3394	0.0361
Qwen-7b-SFT	0.4200	0.3790	0.4200	0.3864	0.3400	0.3237	0.3400	0.3025	0.4369	0.2271	0.3396	0.2000	0.2875
Qwen-7b-GRPO	0.4091	0.3073	0.3244	0.2858	0.2921	0.2062	0.2116	0.1766	0.1806	0.0315	0.1370	0.3333	0.2337
Advanced Large Models & Agentic Methods													
Qwen-plus-base	0.5273	**0.6617**	0.5273	0.4934	0.4091	0.4135	0.4091	0.3931	0.2493	0.0573	0.1714	0.2909	0.1464
Qwen-plus-ReAct	0.4636	0.6458	0.4636	0.4008	0.4455	0.4541	0.4455	0.4026	0.2319	0.0499	0.1566	0.3576	0.1286
ours	**0.6273**	0.6538	**0.6154**	**0.6181**	**0.6000**	**0.6669**	**0.5504**	**0.5864**	**0.5924**	**0.4918**	**0.5563**	0.3697	**0.2881**

Table 3: Performance comparisons between our proposed method and various baseline models. The experiment evaluated the immediate accuracy and long-term group sustainability. The boldface represents the best performance in each category.

Model	Relevance	Usefulness	Personal.
Qwen-plus-base	4.2222	4.0000	4.4444
GT-PMARL (Ours)	**4.6667**	**4.5926**	**4.7778**

Table 4: Human evaluation ratings from the 12-week field study. Scores (1-5) reflect user satisfaction with live system suggestions.

Model Variant	Tool		Role		Ans	Avg
	Acc	F1	Acc	F1	R-L	F1
GT-PMARL (RoBERTa)	**0.4636**	**0.4127**	**0.4455**	**0.3268**	**0.4007**	**0.3697**
Training Strategy						
w/o GRPO	0.4273	0.3706	0.4545	0.2758	0.3584	0.3232
w/o PU-Score	0.4000	0.3865	0.4545	0.3458	0.3466	0.3662
w/o Multi-agent	0.3000	0.3014	0.1364	0.1382	0.1262	0.2198
MA Components						
w/o Game Coord.	0.4545	0.3907	0.4480	0.3022	0.3750	0.3464
w/o Nash Equil.	0.1909	0.1457	0.1000	0.0822	0.1837	0.1140
w/o Pareto Opt.	0.2364	0.1941	0.0909	0.0481	0.1625	0.1211
w/o Load Balance	0.1818	0.1571	0.1091	0.0907	0.1680	0.1239

Table 5: **Ablation Study.** Each component's contribution to Tool/Role selection and fidelity. Game-theoretic equilibrium is vital for structural stability.

rigorous requirements of scientific workflows.

Resolving the Efficiency-Sustainability Coupling. A critical finding is that GT-PMARL effectively resolves the "Individual-Centric Trap." While Llama-3B-SFT reports the lowest Gini coefficient (0.1467), its dismal accuracy reveals a "naive balancing" failure where tasks are distributed evenly but incorrectly. In contrast, GT-PMARL maintains a sustainable Gini of 0.3697 while maximizing the Social Welfare Value (SWV) to 0.2881. By internalizing opportunity costs via the Global-Aware Critic, our framework achieves high-fidelity execution without triggering the cognitive exhaustion characteristic of Collaborative Myopia. These findings are further validated in real-world deployment (Table 4), where GT-PMARL achieves superior Personalization (4.7778) ratings over Qwen-plus-base.

Ablation Study. The results of our ablation study (Table 5) quantify the contribution of each strategic component. While removing the PU-Score (Quality Prior) or Proactive Planning causes a steady decay in performance, ablating the game-theoretic mechanics triggers a catastrophic systemic collapse. Specifically, without the Nash Equilibrium or Pareto Optimality constraints, the Role F1 score plummets to near-random levels (e.g., 0.0822 and 0.0481). This confirms our core hypothesis: while the upper-layer generates a diverse "strategy menu," the lower-layer's equilibrium negotiation is the essential engine that resolves the efficiency-sustainability trade-off.

6 Conclusion and Future Work

This work introduces GT-PMARL to resolve the trade-off between task accuracy and team longevity. By embedding game-theoretic equilibrium into agent coordination, we break the "individual-centric trap". This represents a critical step toward sustainable human-AI collaboration, where immediate excellence and long-term collective potential are achieved in unison.

Limitations

While GT-PMARL demonstrates significant potential in achieving a strategic equilibrium between immediate efficiency and long-term team sustainability, we acknowledge the following limitations:

- Domain Specialization. Our evaluation is primarily grounded in biomedical research workflows (specifically sepsis-related literature). While the game-theoretic coordination framework and the nnPU-scoring mechanism are theoretically domain-agnostic, their effectiveness in other high-stakes collaborative environments, such as legal reasoning or industrial engineering, remains to be empirically verified. Incorporating more complex psychological or sociographic models into the MARL state space is a promising direction for future research.

- Computational Overhead of Proactive Planning. The dual-layer architecture involves an upper-layer candidate generation phase (GRPO) and a lower-layer negotiation phase. While this ensures high-quality and diverse strategy pools, it introduces additional computational latency during the "proactive thinking" stage compared to simple reactive agents. Future work will investigate lightweight distilling techniques to accelerate the coordination process without sacrificing the Nash-Pareto equilibrium.

- Limited Multi-modal Integration. Currently, our framework focuses on text-based scientific reasoning and tool invocation. Modern scientific collaboration often involves visual data, such as imaging artifacts and complex tables. Extending the framework to handle multi-modal inputs would provide a more comprehensive view of the team state.

Ethics Statement

We strictly adhere to the ACL Ethics Policy throughout the research process. The ethical considerations of this study are as follows:

Human Participants and Compensation: The real-world data collection and human-in-the-loop evaluation involved three biomedical research teams, each comprising one Principal Investigator (PI) and five PhD student researchers. These participants were recruited from established biomedical institutions. All participants were compensated at their standard institutional research rates for the time spent in experimental sessions. We ensured that the workload imposed during the study did not interfere with their primary academic or professional responsibilities.

Informed Consent and Privacy: Informed consent was obtained from all human participants prior to the study. Participants were briefed on the data collection process, and all interaction data were strictly anonymized to prevent the identification of individuals or specific institutional affiliations. The study does not involve sensitive, personal, or confidential medical information beyond publicly available PubMed literature metadata.

Data Source and Reproducibility: The dataset used for training the nnPU scorer and evaluating the GT-PMARL framework is derived from publicly available PubMed articles. We do not use any copyrighted material without proper attribution. To support the reproducibility of scientific AI research, we will release our curated benchmark, experimental protocols, and evaluation guidelines upon publication.

Potential Impact and Misuse: Our work aims to reduce the "Individual-Centric Trap" and prevent expert burnout in scientific teams. While this promotes a more equitable collaborative ecosystem, we emphasize that the AI coordinator is designed to assist rather than replace human leadership. Final strategic decisions should always remain under human oversight to mitigate potential risks of algorithmic bias in task allocation.

References

Sonya Ahamed, Gillian L Galford, Bindu Panikkar, Donna Rizzo, and Jennie C Stephens. 2024. Carbon collusion: Cooperation, competition, and climate obstruction in the global oil and gas extraction network. *Energy Policy*, 190:114103.

Kai Barron, Steffen Huck, and Philippe Jehiel. 2024. Everyday econometricians: Selection neglect and overoptimism when learning from others. *American Economic Journal: Microeconomics*, 16(3):162–198.

Geng Chen, Xiaoxian Kong, and Qingtian Zeng. 2024. Collaborative localization algorithm for joint node selection and power allocation based on cooperative games. In *Proceedings of the 2024 2nd International Conference on Computer, Internet of Things and Smart City*, pages 56–61.

Yiqun Chen, Jiaxin Mao, Yi Zhang, Dehong Ma, Long Xia, Jun Fan, Daiting Shi, Zhicong Cheng, Simiu

Gu, and Dawei Yin. 2025. Ma4div: Multi-agent reinforcement learning for search result diversification. In *Proceedings of the ACM on Web Conference 2025*, pages 1703–1715.

Tanner Fiez, Benjamin Chasnov, and Lillian Ratliff. 2020. Implicit learning dynamics in stackelberg games: Equilibria characterization, convergence analysis, and empirical study. In *International conference on machine learning*, pages 3133–3144. PMLR.

Rihab Gorsane, Omayma Mahjoub, Ruan John de Kock, Roland Dubb, Siddarth Singh, and Arnu Pretorius. 2022. Towards a standardised performance evaluation protocol for cooperative marl. *Advances in Neural Information Processing Systems*, 35:5510–5521.

Pablo Hernandez-Leal, Michael Kaisers, Tim Baarslag, and Enrique Munoz De Cote. 2017. A survey of learning in multiagent environments: Dealing with non-stationarity. *arXiv preprint arXiv:1707.09183*.

Xinyi Hou, Yanjie Zhao, Shenao Wang, and Haoyu Wang. 2025. Model context protocol (mcp): Landscape, security threats, and future research directions. *arXiv preprint arXiv:2503.23278*.

Kristen Jaskie and Andreas Spanias. 2022. *Positive unlabeled learning*. Morgan & Claypool Publishers.

Ryuichi Kiryo, Gang Niu, Marthinus C Du Plessis, and Masashi Sugiyama. 2017. Positive-unlabeled learning with non-negative risk estimator. *Advances in neural information processing systems*, 30.

Yanyan Li, Yijun Wang, and Yiwei Zhou. 2024. Multiagent deep reinforcement learning algorithms in starcraft ii: A review. *IEEE Access*.

Haitong Liang, Guangbo Hao, Oskar Z Olszewski, and Vikram Pakrashi. 2022. Ultra-low wide bandwidth vibrational energy harvesting using a statically balanced compliant mechanism. *International Journal of Mechanical Sciences*, 219:107130.

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Jinfa Huang, Junwu Zhang, Yatian Pang, Munan Ning, et al. 2024. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*.

Zuxin Liu, Thai Hoang, Jianguo Zhang, Ming Zhu, Tian Lan, Juntao Tan, Weiran Yao, Zhiwei Liu, Yihao Feng, Rithesh RN, et al. 2024. Apigen: Automated pipeline for generating verifiable and diverse function-calling datasets. *Advances in Neural Information Processing Systems*, 37:54463–54482.

Alan O'Sullivan. 2003. Dispersed collaboration in a multi-firm, multi-team product-development project. *Journal of Engineering and Technology Management*, 20(1-2):93–116.

Hugo Roger Paz. 2025. An agent-based simulation of regularity-driven student attrition: How institutional time-to-live constraints create a dropout trap in higher education. *arXiv preprint arXiv:2511.16243*.

Jiaqi Qin, Yi Zhang, Shixiong Fan, Xiaonan Hu, Yongqiang Huang, Zexin Lu, and Yan Liu. 2022. Multi-task short-term reactive and active load forecasting method based on attention-lstm model. *International Journal of Electrical Power & Energy Systems*, 135:107517.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*.

Tianyu Ren, Xuan Yao, Yang Li, and Xiao-Jun Zeng. 2025. Bottom-up reputation promotes cooperation with multi-agent reinforcement learning. *arXiv preprint arXiv:2502.01971*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551.

Fangzhou Wu, Shutong Wu, Yulong Cao, and Chaowei Xiao. 2024a. Wipi: A new web threat for llm-driven web agents. *arXiv preprint arXiv:2402.16965*.

Qinzhuo Wu, Wei Liu, Jian Luan, and Bin Wang. 2024b. Toolplanner: A tool augmented llm for multi granularity instructions with path planning and feedback. *arXiv preprint arXiv:2409.14826*.

Yaodong Yang and Jun Wang. 2020. An overview of multi-agent reinforcement learning from game theoretical perspective. *arXiv preprint arXiv:2011.00583*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. Rrhf: Rank responses to align language models with human feedback. *Advances in Neural Information Processing Systems*, 36:10935–10950.

Liyuan Zheng, Tanner Fiez, Zane Alumbaugh, Benjamin Chasnov, and Lillian J Ratliff. 2022. Stackelberg actor-critic: Game-theoretic reinforcement learning algorithms. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 9217–9224.

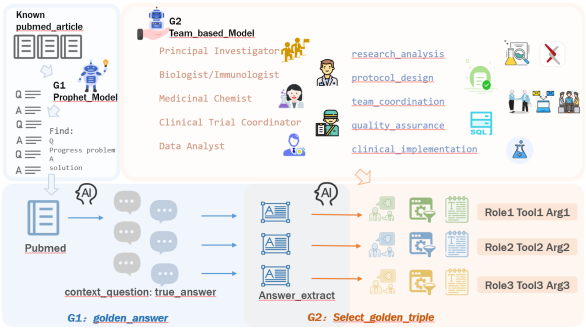

Figure 3: Data construction pipeline.

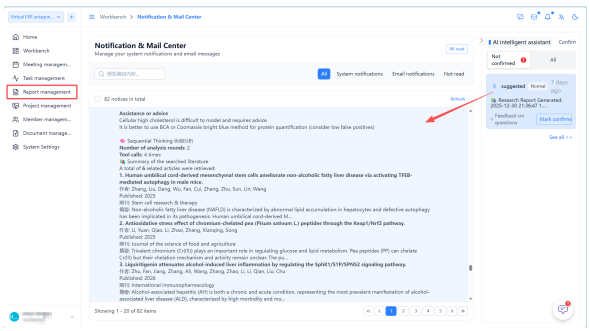

Figure 4: Proactive Coordination in the System Dashboard. This figure illustrates the agent's intervention logic. When a researcher submits a progress update via the "Report Management" portal, the assistant (right sidebar) proactively pushes a strategic recommendation (e.g., a Research Report generated via the Sequential Thinking module) instead of waiting for a manual query.

A Data Statistics

Simulation Data: To ensure annotation quality, we employed independent double-blind annotation by domain experts, achieving high inter-annotator agreement and data reliability, with the data construction pipeline shown in Figure 3.

To further illustrate the structure of our dataset and the logic behind strategy construction, we present two representative cases in Table **??**. These cases demonstrate how raw coordination issues are transformed into structured (Role, Tool, Argument) triplets for model training and evaluation.

A.1 Real-World Interaction Data

To validate the proactive coordination logic in authentic scenarios, we collected interaction data through a 12-week field study involving four heterogeneous teams: one from a technical logistics firm and three from university medical laboratories.

Collection Workflow: Users interacted with our system via the *Report Management* interface (see Figure 4). To evaluate the ecological validity of

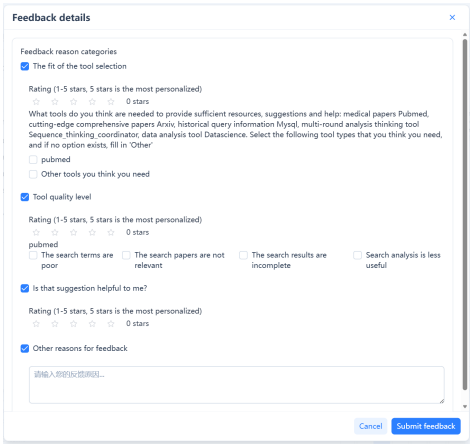

Figure 5: Multi-dimensional User Feedback Portal.

GT-PMARL in authentic scientific workflows, we conducted a 12-week field study involving four professional teams. This appendix details the participation profiles, the evaluation interface, and the feedback categories.

A.2 Experimental Setting and Participants

We deployed our system across four heterogeneous research environments:

- **Team 1 (Industrial):** A project team from a technical logistics company focused on supply chain optimization and biomedical logistics.
- **Teams 2–4 (Academic):** Three university-based medical research laboratories led by senior professors, specializing in sepsis pathology and bioinformatics.

A/B Testing Protocol: Teams updated their project milestones daily. We compared two proactive coordination engines: (1) Test A (Ours): The GT-PMARL engine, which internalizes team workload and opportunity costs. (2) Test B (Baseline): A standard proactive agent powered by GPT-5, which provides resource and task suggestions based purely on semantic relevance without strategic load-balancing.

A.3 User Feedback Interface

The evaluation data is harvested through a standardized feedback portal (Figure 5). Users provide feedback on three key dimensions: (1) Tool Selection Fit (relevance to the task), (2) Tool Quality Level (precision of retrieved content), and (3) Overall Helpfulness. This granular data is used to refine the PU-scorer and evaluate the social welfare value (SWV) of the coordination policy.

UI Feedback Item	Measurement Dimension	Linked Metric
Tool Relevance	Alignment of allocated resources	Qual. (S_ϕ)
Tool Quality	Precision of retrieved evidence	Strat. Arg.
Overall Helpfulness	Contextual utility for progress	SWV
Other (Free-text)	Subjective load perception	Gini Coeff.

Table 6: Mapping of UI Feedback Items to Research Metrics.

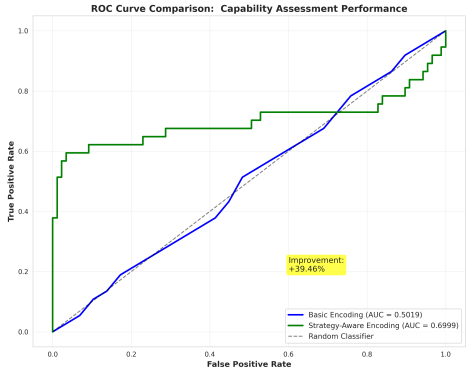

Figure 6: Performance gain in capability assessment. Comparison of ROC curves between the Basic Encoding baseline (AUC=0.5019) and our proposed Strategy-Aware Encoding (AUC=0.6999). The results demonstrate that the nnPU framework, combined with structural logic, achieves a significant improvement under sparse supervision.

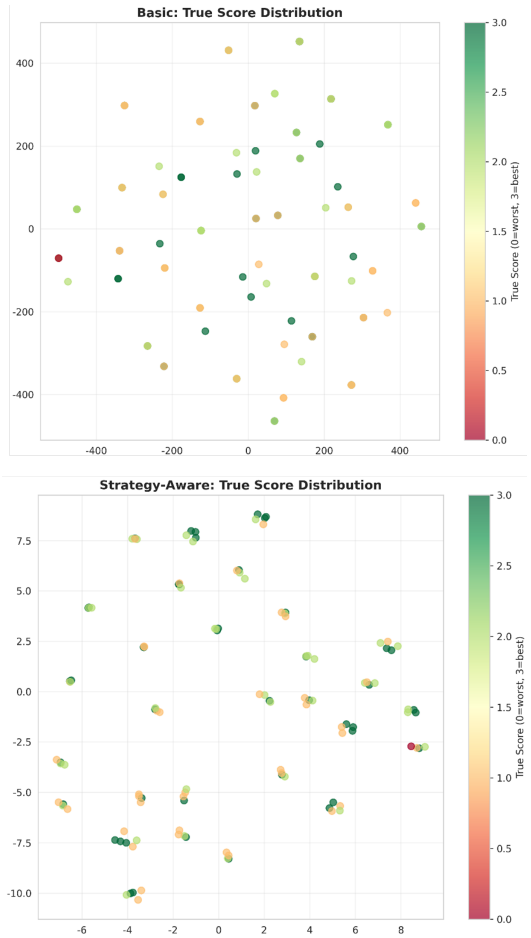

Figure 7: Visualization of the latent representation space via t-SNE. (Left) Basic encoding intermingles samples of different quality, failing to capture coordination logic. (Right) Strategy-Aware Triplet Encoding exhibits clear feature clustering, where high-quality strategies (green) are mapped into distinct functional regions, revealing the "coordination grammar" learned by the scorer.

B In-depth Analysis of the nnPU Quality Scorer

To ground the coordination logic of GT-PMARL, we utilize a Non-negative Positive-Unlabeled (nnPU) scorer. This appendix provides empirical evidence of the scorer's effectiveness in deciphering complex scientific strategies under sparse supervision.

B.1 Performance Gain over Sparse Supervision

In scientific coordination, we often only observe a few successful trajectories (Positive, P), while the vast majority of potential role-tool-task combinations remain Unlabeled (U). As shown in the ROC comparison (Fig. 6), a standard Positive-Negative (PN) approach results in an AUC of approximately 0.51, indicating a failure to distinguish quality due to the bias in U. By contrast, The Strategy-Aware Encoding achieves an AUC of 0.6999, a +39.46% improvement over Basic Encoding (AUC=0.5019). This demonstrates that the nnPU framework, combined with structural logic, achieves significant performance improvement under sparse supervision. This demonstrates that the non-negative risk estimator effectively corrects the gradient bias introduced by latent positives within the unlabeled set.

B.2 Latent Space and Strategy-Aware Encoding

We further analyze why the scorer performs well by visualizing the latent representation space using t-SNE. We compare the Basic Encoding (only semantic features) and our proposed Strategy-Aware Triplet Encoding (modeling the logical harmony between Role, Tool, and Arguments).

Table 7: Qualitative comparison of allocation strategies. Blue text highlights Ours' superior strategic depth in identifying leadership roles (Case 1) and proactive impact assessment (Case 2) compared to Baseline's reactive approach.

ID	Original Query (Q)	Raw Analysis (A)	Ground Truth (GT)	Baseline Model	Ours (GT-PMARL)
Case 1	Biomedical system integration is functional, but the project is **shelved** due to **field mapping** disagreements. Management suggested a site visit, but **stagnation** persists.	Align departmental needs. A project lead is required to bridge the gap, while technical staff verify field compatibility.	**Role:** Principal Investigator **Tool:** `team_coord` **Args:** *consensus meeting on mapping*	**tool:** quality_assurance **parameters:** Audit field mapping protocols, identify gaps, create validation checkpoints **role:** Quality Assurance Manager	*<think> Disagreements imply a **coordination deadlock** rather than just a protocol error. **Principal Investigator** is needed for high-level cross-dept. consensus. </think>* **tool:** `team_coordination` **parameters:** *organize consensus meeting for field definition* **role:** Principal Investigator
Case 2	Requirement **shifted** from "Library Push" to "Cost-based Inventory". This **frequent rework** caused team **dissatisfaction** and project **delays**.	Validate changes via analysis. Experts must assess the broader impact and set a fixed delivery cycle to avoid resource waste.	**Role:** Data Analyst **Tool:** `data_analysis` **Args:** *inventory logic impact assessment*	*<think> the problem is frequent rework. Tool "data_analysis" can help track changes and quantify impact. </think>* **tool:** data_analysis **parameters:** Track requirement change history, measure rework frequency, dashboards **role:** Data Analyst	*<think> Frequent shifts require **proactive validation** before technical rework. Focus on **pipeline optimization** to mitigate future delays. </think>* **tool:** `data_analysis` **parameters:** *evaluate inventory metrics for new pipeline logic* **role:** Data Analyst

B.3 Performance Gain over Sparse Supervision

As illustrated in Fig. 7, the latent space of the Basic model is fragmented, with P and U samples intermingled randomly. Conversely, the Strategy-Aware model exhibits clear feature clustering. High-quality strategies (indicated by True Scores) are mapped into distinct functional regions. This clustering suggests that the model has successfully learned the "coordination grammar", recognizing that the utility of a role is conditioned on specific tools and parameters—thereby facilitating a much sharper decision boundary for candidate filtering.

B.4 Sensitivity Analysis of Class Prior π_p

The class prior $\pi_p = \mathbb{P}(y = 1)$ is a critical hyperparameter representing the estimated proportion of high-quality strategies in the unlabeled pool. We conducted sensitivity experiments across $\pi_p \in \{0.2, 0.3, 0.4, 0.5\}$.

This work reveals that the model maintains robust performance across a reasonable range of priors. A prior of $\pi_p = 0.3$ (our default) provides the best balance between precision and recall. Overestimating the prior ($\pi_p > 0.5$) leads to a conservative bias, where the model becomes overly skeptical of unlabeled data, while under-estimating it results in a slight drop in the AUC of the positive class.

C Technical Derivations and Implementation Details

This appendix provides the specific mathematical formulations for the state evolution and the competitive refinement mechanisms used in our framework.

C.1 Dynamics of Team State Tracking

To enable the Global-Aware Critic to perceive long-term patterns, we implement an exponential moving average (EMA) to track the win-rate (WR) and workload persistence for each role r:

$$WR_r^{(t)} = (1 - \alpha)WR_r^{(t-1)} + \alpha \cdot \mathbb{I}(\text{role } r \in \mathbf{a}^*) \tag{14}$$

where $\alpha \in [0.01, 0.05]$ is the momentum coefficient. The cumulative load w_r^t is updated via a decay-and-increment process:

$$w_r^{t+1} = \eta \cdot w_r^t + \Delta w(\mathbf{a}^*, r) \tag{15}$$

where $\eta \in [0.95, 1.0]$ is the relaxation factor representing the temporal dissipation of cognitive pressure.

C.2 Non-linear Shadow Price Scaling (λ_t)

The opportunity-cost coefficient λ_t in Eq. (9) is adaptively scaled based on the current workload variance $\mathcal{V}(\mathbf{w}_t)$. We define a scarcity multiplier to protect expert bandwidth during high-imbalance periods:

$$\lambda_t = \lambda_{base} \cdot \exp\left(\gamma \cdot \frac{\mathcal{V}(\mathbf{w}_t) - \bar{\mathcal{V}}}{\sigma_{\mathcal{V}}}\right) \tag{16}$$

where γ is the sensitivity factor, $\bar{\mathcal{V}}$ is the historical mean variance, and $\sigma_{\mathcal{V}}$ is its standard deviation. This ensures λ_t escalates non-linearly as the team approaches the "PI-Centric Trap."

C.3 Regularized Iterative Best Response (IBR)

To ensure stable convergence of the Nash-Pareto game, we utilize a Boltzmann-regularized best response. At iteration m, the selection probability simplex is updated as:

$$P_i^{(m+1)}(\mathbf{a}_j) = \frac{\exp\left(\mathbb{E}_{\sigma_{-i}^{(m)}}[\mathcal{U}_i(\mathbf{a}_j, \sigma_{-i}^{(m)})]/\tau\right)}{\sum_{\mathbf{a}_k \in \mathcal{F}_P} \exp\left(\mathbb{E}_{\sigma_{-i}^{(m)}}[\mathcal{U}_i(\mathbf{a}_k, \sigma_{-i}^{(m)})]/\tau\right)} \tag{17}$$

where τ is the temperature parameter. The mixed-strategy Nash equilibrium σ^* is reached when $\|P^{(m+1)} - P^{(m)}\| < \delta$. In our implementation, convergence is typically achieved within $m \leq 3$ iterations.

C.4 Refinement of Specialized Identity Embeddings

The identity embedding $\mathbf{e}_i$ for agent i is refined through a functional drift loss $\mathcal{J}_i$, forcing the agent to specialize in its most competitive functional pillar:

$$\mathcal{J}_i = -\sum_{\mathbf{a}_k \in \mathcal{P}_t} \log P(\mathbf{a}_k \mid \mathcal{S}_t, \mathbf{e}_i) \cdot \text{Adv}_i(\mathbf{a}_k) + \mu \|\mathbf{e}_i - \bar{\mathbf{e}}\|^2 \tag{18}$$

where the first term is the specialized policy gradient and the second term is a diversity-maintaining constraint that prevents embeddings from collapsing into a single role-centroid.

place holder