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

Tong Zhang[♣] **Yang Wu**[♡] **Yufei Shi**[♢]
Rujing Yao[♣] **Zhuoren Jiang**[♠] **Xiaozhong Liu**[♡*]
[♣]Nankai University [♡]Worcester Polytechnic Institute
[♢]The Hong Kong Polytechnic University [♠]Zhejiang University
{tongzhangnk, rjyao}@mail.nankai.edu.cn {ywu19, xliu14}@wpi.edu
yufei1999.shi@connect.polyu.hk jiangzhuoren@zju.edu.cn

Abstract

In heterogeneous scientific teams, proactive team agents can serve as effective assistants regarding the research progress of the project. However, proactive agents always suffer from collaborative myopia: a greedy optimization for immediate task accuracy which ignore the long-term goal of team sustainability. This leads to the Individual-centric Trap, where capable experts (e.g., PIs) are disproportionately overloaded while Junior roles remain underutilized. Therefore, neglecting opportunity costs in task allocation can implicitly erodes the enduring performance of the team. To solve this imbalance between efficiency and sustainability, we propose GT-PMARL (Game-Theoretic Proactive Multi-Agent Reinforcement Learning). By internalizing the opportunity cost as a key consideration in individual decision-making, the collaboration logic of agents has been reshaped. Our framework employs: (1) a Positive-Unlabeled scorer to anchor intervention quality under sparse supervision; (2) a Nash-Pareto competitive objective to seek an equilibrium between individual task excellence and collective load balancing. Empirical experiments in scientific workflows show that GT-PMARL effectively maintains high performance while preventing experts from overdeveloping. Our work provides a scalable paradigm for building a sustainable and balanced human-AI collaborative ecosystem.

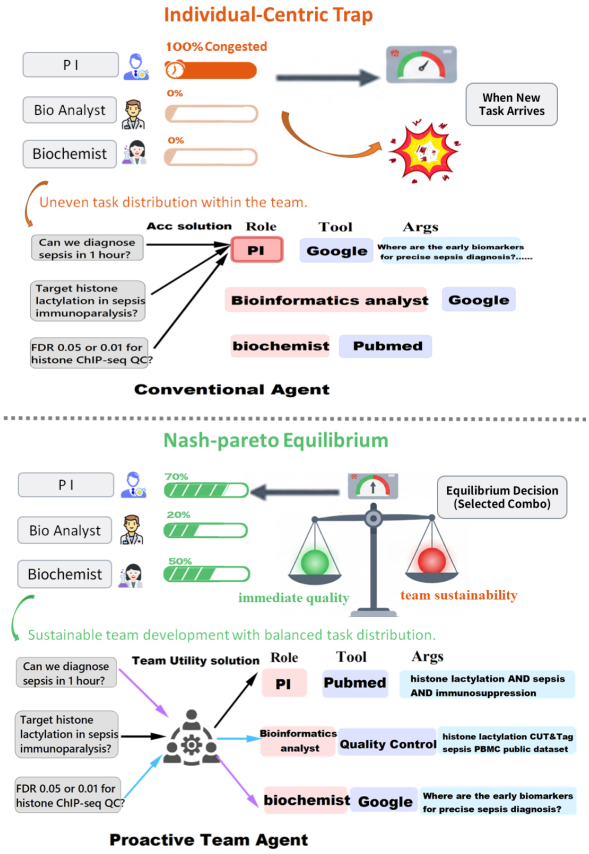

Figure 1: Top: Conventional agents fall into the Individual-Centric Trap, where tasks are greedily routed to the most capable member. Bottom: Our method seeks a Nash-Pareto Equilibrium, balancing immediate task quality with long-term team sustainability.

1 Introduction

Large language models (LLMs) have made significant progress in scientific tasks through tool execution workflows (Hou et al., 2025; Schick et al., 2023; Qin et al., 2023). In modern laboratories, where interdisciplinary cooperation becomes increasingly close, the ability to orchestrate complex workflows and explore scientific knowledge is particularly critical. In this context, LLMs are evolving into a bridge for enhancing research collaboration, with their core lying in the ability to achieve intelligent decision-making planning and supplementation of expertise through tool invocation. For example, standardized tool interfaces like Model Context Protocol (MCP) expanded the external resources available to these agents, from code retrieval to biomedical data analysis (Liu et al., 2024; Wu et al., 2024a). However, team collaboration face a critical challenge: how can LLM-based coordinators allocate tasks across heterogeneous

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 38552–38565

team roles to simultaneously optimize for immediate task accuracy and long-term team sustainability? We refer to this as the efficiency-sustainability coupling problem.

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, most of these frameworks stay at the level of solving immediate problems passively, and may ignore the long-term constraints of team members' cognitive overload and resource constraints in a limited collaborative environment. Specifically, when the reward signal is calibrated to give priority to the immediate task, these models can't avoid the probability of cooperative myopia: it tends to concentrate the task load on the most capable individual, thus undermining the long-term performance of the team (Qin et al., 2022). From an economic point of view, these systems ignore the opportunity cost of spending the limited time of domain experts on sub-optimal task matching (Barron et al., 2024; Paz, 2025). As shown in figure 1, this leads to the trap of taking individuals as the center, and high-risk tasks are disproportionately assigned to the Principal Investigator (PI).

The super team agent, which can balance team goals and individual goals, can act as the intelligent brain to promote the project in the development of team projects, thus forming a coupling constraint, in which resource allocation and solution quality play an inseparable role in the unified optimization challenge. In addition to the apparent trade-off, there are three systematic obstacles to overcome to solve this problem.

Specifically, we have determined the following. First of all, agents usually make decisions through semantic matching rather than team strategy adaptation. Due to the failure to embed the constraints of resource upper limit (such as expert workload, cost), the agent can't quantify the long-term cost of decision-making, and the model will inadvertently succumb to the "short-termism trap centered on individuals". Secondly, the combination of active planning explodes, and agents need to explore in the huge *Role* $\times$ *Tool* $\times$ *Task* space to pre-empt the bottleneck. This strategic demand for high-quality "who-what-how" configuration exceeds the attention space of reactive model. Finally, the sparsity and equivalence of supervision signals in collaborative data hinder the refinement of strategies. The trajectory of the real world usually only provides sparse examples of successful interactions with multiple feasible paths, which provides insufficient negative supervision to distinguish local feasible decisions from global sustainable equilibrium strategies (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

Tool-augmented models have mainly relied on passive reasoning to empower agents (Schick et al., 2023; Qin et al., 2023; Wu et al., 2024b), these agentic frameworks have proven effective at decomposing tasks and interacting with external

APIs. However, they always run under a reactive paradigm (Yuan et al., 2023), focusing on satisfying the user's immediate request, is static and struggles to adapt to the diverse situations in real-world scenarios. From reactive to proactive, agents anticipate user intent and assist in task implementation through forward-looking planning. In the context of scientific teamwork, such proactive team agents can become reliable decision-making assistants (O'Sullivan, 2003). Prior research has demonstrated that proactive assistance provides a significantly better user experience than passive response(Lu et al., 2024). Furthermore, in collaboration settings, active intention inference can bring the best collaboration performance (Zhang et al., 2024). However, scientific collaboration is essentially multi-agent and requires the active allocation of resources among team members with different professional knowledge (O'Sullivan, 2003). Although some studies have explored multi-agent systems in specialized fields, they often overlook the fundamental strategic dilemma of balancing immediate efficiency with long-term team sustainability. In the absence of a coordination mechanism at the team level, even proactive agents are vulnerable to the influence of "cooperative short-sightedness". Our work Bridges this gap by redefining the problem as a multi-agent coordination challenge rather than a simple instruction following task.

Game theoretic frameworks have been widely applied to solve the complexity issues of multi-agents interaction. Multi-agent Reinforcement Learning (MARL) encounters a challenge where agents treat other agents as a static part of the environment. The combination of game theory with MARL provides a way for solving the problems of strategy optimization and coordination in non-stationary environments (Hernandez-Leal et al., 2017). For this reason, research focus has shifted to the game theory method of explicitly modeling the interaction between agents (Chen et al., 2025; Li et al., 2024; Ren et al., 2025).We draw inspiration from games (Zheng et al., 2022; Ahamed et al., 2024; Chen et al., 2024), where a player's action constrains the other players' best responses (Fiez et al., 2020). To prevent the system from not converging or converging to an inefficient equilibrium state, this paper requires that the stable state of the system must simultaneously be on the Pareto optimal frontier through the nash-pareto architecture. In our framework, the upper-level planner defining the strategy space, while the lower-level agents act as team members, competing within this space to find the trade of accuracy and team balance (Zheng et al., 2022). 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.

3 Preliminaries

In this section, we formalize the scientific team coordination task into a dynamic decision-making process, and define the indicators to evaluate the coordination quality and team sustainability.

3.1 Modeling Scientific Team Collaboration

Team State Representation. We model the operational state of the heterogeneous science team at each time step t as a composite triplet $\mathcal{T}_t = \langle \mathbf{w}_t, \mathbf{c}_t \rangle$. Specifically, A team composed of N roles $\mathcal{R} = \{1, \ldots, N\}$, $\mathbf{w}_t = [w_1^t, \ldots, w_N^t]^\top \in \mathbb{R}^N$ represent the workload distribution vector, $\mathbf{c}_t \in \mathbb{R}^d$ is the task context embedding, encoding the current research situation.

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 $\mathbf{a}_t = \langle r, f, \theta \rangle$, where $r \in \mathcal{R}$ is the target role, $f \in \mathcal{F}$ identifies the external tool or resource and $\theta \in \Theta_f$ specifies the invocation parameters. 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:

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

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* $\mathcal{P}_t = \{\mathbf{a}_1, \ldots, \mathbf{a}_M\}$. 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 $S_\phi : \mathcal{P} \times \mathcal{T} \to [0, 1]$. 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 $\mathbf{w}_{t+1}$:

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

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:

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

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 $\mathcal{P}_t$, 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 closely coupled layers: (1) the **upper-layer** generation strategy of high-entropy candidate pool is proposed under structural constraints, and (2) the **lower-layer** coordination engine that realizes game theory balance among heterogeneous roles. The core of this framework is a nnPU scorer (Kiryo et al., 2017), which is used as a prior benchmark to filter the poor configuration in the upper layer and as a basic accuracy reward signal in the lower layer.

4.1 Prior of Strategic Perception Quality

The scorer S_ϕ does not act as a simple binary classifier, but as a feature test of basic prior and structural logic relationship, coordinating the learning dynamic optimal scheme of GT-PMARL generation layer and competition layer.

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 $\mathbf{a} = \langle r, f, \theta \rangle$:

$$S_\phi(\mathbf{a} \mid \mathbf{q}) = \sigma\left(\text{MLP}\phi\left(\text{Encoder}\left(\mathcal{T}(\mathbf{q}, r, f, \theta)\right)\right)\right) \tag{5}$$

where $\mathcal{T}$ denotes a structured prompt that synthesizes the task context $\mathbf{q}$, the agent's role profile r, the tool's functional tags f, and the implementation arguments θ. In this architecture, the pre-trained `gte-Qwen2-7b` encoder functions as a deep feature extractor, while the trainable *Ordinal MLP* S_ϕ serves as the decision head.

By internalizing conditional dependence, this structural awareness allows the model to punish semantic inconsistency and provide multidimensional reality check for all downstream strategy exploration in GRPO and MARL layers.

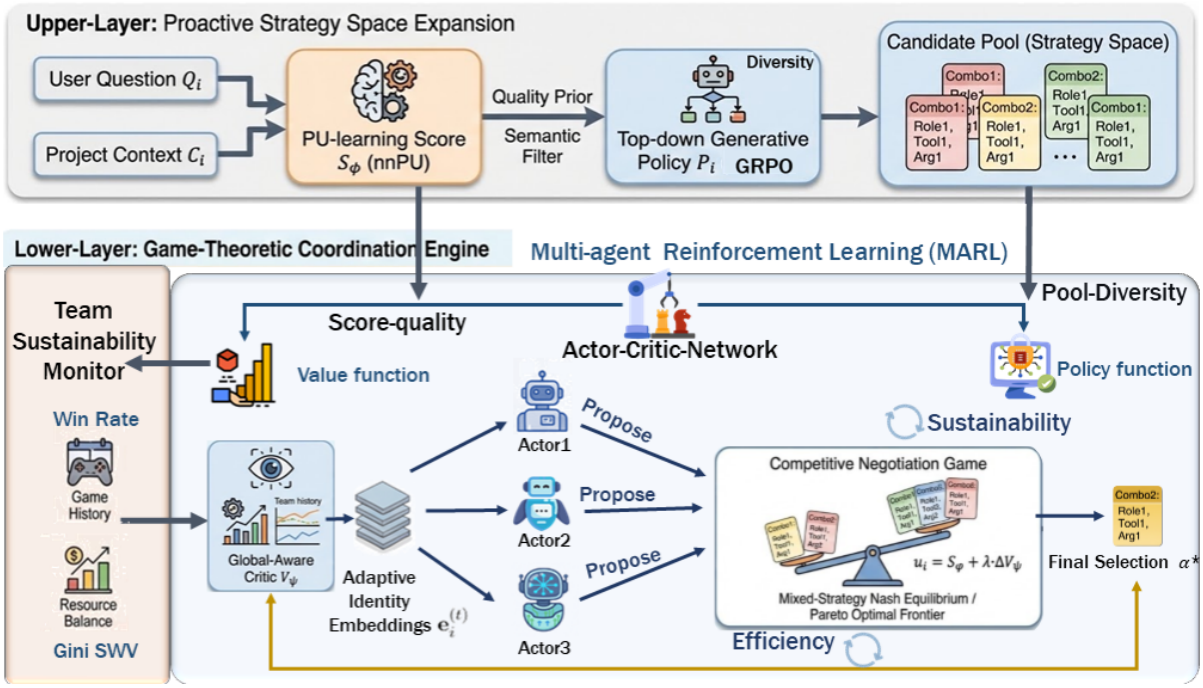

Figure 2: Overview of the GT-PMARL framework. The Upper-Layer (Strategy Expansion) generates a high-entropy candidate pool using Diversity-GRPO guided by the nnPU-based quality prior. The Lower-Layer (Coordination Engine) employs MARL to model the allocation as a competitive negotiation game. PU-scores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

4.1.2 Distilling Quality via Non-Negative Risk Minimization

One of the key challenges in the coordinated implementation among team members is the lack of clear negative labels; Although the successful trajectory is recorded, the sub-optimal task-role pairing is rarely recorded. In order to solve this problem, we adopt the non-negative Positive-Unlabeled (nnPU) scoring framework, unlabeled set $U = P \cup N$ (where successful trajectories P are sparse). Given the class prior $\pi_p = \mathbb{P}(y = 1)$, we formulate the risk estimator to prevent the model from overfitting to 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 $S_\phi(\mathbf{a} \mid \mathcal{T}_t) \in [0, 1]$ as the baseline of scientific effect, which indicates the effectiveness of the current strategic plan.

The uniqueness of S_ϕ lies in its important role as connective tissue, which restricts the different training stages of the upper and lower frames synchronously:

Generative Navigation. In GRPO stage, S_ϕ plays a role of high-quality strategy guidance. It projects the successful problem-solving path into a vast combinatorial search space, provides intensive reward signals, and guides LLM to explore effective strategies with diverse functions

Equilibrium Grounding. In the process of MARL collaboration, S_ϕ plays the role of "basic income" to prevent agents from converging to fair but scientifically invalid distribution. By embedding S_ϕ into the reward of game theory, we ensure that team fairness can only be pursued on the solid basis 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

In order to solve the challenge of text collaborative myopia, the upper layer needs to provide a sufficiently differentiated and rich set of optional strategies. We use Group Relative Policy Optimiza-

tion (GRPO) to optimize a generation strategy π_θ, and keep the whole pool at a high level by comparing the relative advantages within the group. Given the task context $\mathbf{s}_t$, π_θ generates a candidate pool $\mathcal{P}_t = \{\mathbf{a}_1, \ldots, \mathbf{a}_K\}$. To ensure that $\mathcal{P}_t$ provides enough strategic operating space for the bottom, we define a compound reward of R_{pool}:

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

where $\bar{S}_\phi$ is the mean coherence score from the $nnPU$ prior, ensuring accuracy, $\mathcal{D}$ means semantic diversity and $\mathcal{C}$ means taxonomic coverage.

The diversity term $\mathcal{C}(\mathcal{P}_t)$ is quantified by the taxonomic coverage across three functional pillars:

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

Where $\{U, E, C\}$ stands for utility, execution and coordination pillar. This incentive policy provides a strategic menu rather than a simple semantic match.

4.3 Lower-Layer: Coordination via Implicit Equilibrium

When the upper level provides a candidate pool of strategic possibilities, the coordination problem becomes a strategic choice challenge: choosing a single strategy while optimizing the current quality and long-term team sustainability. We model it as an implicit multi-agent game, and team health is internalized as a balanced 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 Internalized the dynamics of the entire team. Unlike conventional MARL which only considers task success, our framework has redefined the objective function to incorporate sustainability constraints:

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

Base Quality (S_ϕ): Grounded by the $NNPU$ Score, $S_\phi \in [0, 1]$ represents the scientific validity of the proposed (Role, Tool, Args) triplet. This ensures that the accuracy of baseline execution will not be sacrificed due to balance.

Sustainability Momentum (ΔV_ψ): the Critic V_ψ is conditioned on the agent identity to perceive how a specific allocation affects the future state $\mathcal{S}_{t+1}$. This term represents the potential gain in long-term team health:

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

Shadow Price (Ψ_i): To avoid the Individual-Centric Trap, Ψ_i means a nonlinear congestion penalty. the shadow price follows:

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

where κ means the penalty strength. When the agent exceeds the threshold, the penalty constraint of shadow price is used to avoid its work overload, which makes the task turn to other agents.

4.3.2 Adaptive Identity and Specialization

In order to cultivate the heterogeneity of different subjects in the team, a special actor network is used to realize it. Each agent i maintains a learnable identity representation $\mathbf{e}_i$ and an inherent decision deviation $\mathbf{b}_i$. The policy π_i integrates the basic task state $\mathcal{S}_{base}$ and the dynamic team state $\mathcal{S}_{dyn}$ (workload, win rate):

$$\text{Logits}_i = \text{MLP}_{\theta_i}(\mathcal{S}_{base} \oplus \mathcal{S}_{dyn} \oplus \mathbf{e}_i) + \mathbf{b}_i \tag{12}$$

Both $\mathbf{e}_i$ and $\mathbf{b}_i$ are updated:

$$\begin{aligned} &\left[\mathbf{e}_i \mathbf{b}_i\right]^{(t+1)} = \left[\mathbf{e}_i \mathbf{b}_i\right]^{(t)} + \eta \nabla_{\{\mathbf{e}_i, \mathbf{b}_i\}} \mathcal{J}(\boldsymbol{\phi}_i) \\ &\text{where} \quad \mathcal{J}(\boldsymbol{\phi}_i) = \mathbb{E}_{\mathbf{a}_t \sim \pi_i}\left[\log \pi_i(\mathbf{a}_t \mid \mathcal{S}_t) \cdot \hat{A}_t\right] \end{aligned} \tag{13}$$

where $\hat{A}_t$ represents the dominant value estimated by critics with global awareness.

This mechanism helps agents to further subdivide their functions. For example, primary agents pay attention to low-entropy tasks $\mathbf{b}_i$, and form their own preference style.

4.3.3 Nash-Pareto Decision

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:

$$\begin{aligned} \sigma_i^{(m+1)} &\leftarrow \text{BR}_i(\sigma_{-i}^{(m)}) \\ &= \arg\max_{\sigma_i'} \mathbb{E}_{\sigma_{-i}^{(m)}}[\mathcal{U}_i(\sigma_i', \sigma_{-i}^{(m)})] \end{aligned} \tag{14}$$

We solve this problem by 2-3 rounds of iterative best game (IBR) until it converges, which shows that agents reach a consensus under the global perceived benefits. After executing $\mathbf{a}^*$, and critics update V_ψ through TD-learning.

5 Experiments

5.1 Dataset Construction

To evaluate the current algorithm framework, we obtained biomedical literature from pubmed database to build a benchmark data set. We collected and screened out 11,792 high-quality papers related to sepsis. Based on corpus, we adopt a two-stage generation process: (1) re-extracting problem-solution pairs from literature facts as positive samples (P); (2) Annotate the corresponding solution (P) with the basic facts (roles, tools and arguments), and generate the corresponding unverified random action combination to form an unlabeled set (U). By constructing the set of positive samples and unlabeled samples, we can train a robust scorer, so as to navigate effectively in the sparsely labeled and huge collaborative action space. The data set is partitioned according to the ratio of 8:1:1 for training, verification 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 compare our method with three levels of baseline models (i) Base Models, including Llama-3B-base (ii) Fine-tuning & RLHF, including sft and grpo variants on Llama and Qwen models, and (iii) Advanced Agentic Paradigms, including Qwen-plus large-scale models. The experiments are all carried out on the computing nodes of NVIDIA GPUs with A800-80GB. In the lower marl framework, this paper instantiates three actor agents and trains 2000 epochs, in which the learning rate of three actors is 1×10^{-4} and the one critic learning rate is 1×10^{-3}. In order to ensure the robustness of the results, we used random seeds 42, 215 and 3407 to evaluate in three independent runs, and the final results were reported as average 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 (Upper-layer) is obviously superior to GRPO in classification diversity (0.9809), although there is a slight compromise in Recall@5. This functional richness in the candidate space enables the lower class to escape the trap of individual-centered.

Table 2 evaluates lower-level coordination logic. GT-PMARL (ours) achieved the best MRR (0.5225) and the highest average score. The confidence level (0.7750) shows that this method produces a stronger strategic consensus than the common Nash solver.

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	
	Accuracy	Macro-P	Macro-R	Macro-F1	Accuracy	Macro-P	Macro-R	Macro-F1	R-1	R-2	R-L	Gini	SWV
Base Models													
llama-3b-base	0.2091	0.1065	0.0571	0.0706	0.0818	0.0681	0.0203	0.0301	0.1477	0.0307	0.1030	**0.1455**	0.0955
Qwen-3b-base	0.2455	0.1579	0.1234	0.1274	0.2727	0.1395	0.1359	0.1223	0.1984	0.0750	0.1984	0.1273	0.1119
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	0.0596	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	0.3636	0.2295	0.2295	0.2194	0.2727	0.1907	0.2007	0.1466	0.0547	0.0112	0.0449	0.3212	0.0377
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 ours and various baseline models. The experiment evaluated the immediate accuracy and long-term group sustainability. we report Macro-average metrics. Note that *Macro-F1* is calculated as the arithmetic mean of per-class F1-scores ($F1_{macro} = \frac{1}{N}\sum F1_i$), rather than the harmonic mean of macro-averaged Precision and Recall, which provides a more robust evaluation. 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	0.3616	0.4545	0.2729	0.3584	0.3172
w/o PU-Score	0.4000	0.3865	0.4545	0.2933	0.3466	0.3399
w/o Multi-agent	0.3000	0.3014	0.3364	0.1382	0.1262	0.2198
MA Components						
w/o Game Coord.	0.4545	0.4467	0.4400	0.3022	0.3750	0.3744

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 reports the lowest Gini coefficient (0.1455), 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. We found that removing PU value or active planning will lead to performance degradation, and we further clarified that the basic principle of deleting each component will bring about a slight effect reduction (0.3697 vs 0.3172, 0.3399, 0.2198). 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

We 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". The proposed framework introduces a decentralized dual-layer architecture that internalizes opportunity costs, supported by an innovative nnPU-based reward model that captures the "coordination grammar" of scientific logic under sparse supervision. 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, as well as a logistics company team. 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 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.

Potential Impact and Misuse: Our work aims to reduce the Individual-Centric Trap. 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.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (72574198).

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*, 12:167452–167470.

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.

Yaxi Lu, Shenzhi Yang, Cheng Qian, Guirong Chen, Qinyu Luo, Yesai Wu, Huadong Wang, Xin Cong, Zhong Zhang, Yankai Lin, et al. 2024. Proactive agent: Shifting llm agents from reactive responses to active assistance. *arXiv preprint arXiv:2410.12361*.

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. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18315–18339.

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)*.

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.

Ceyao Zhang, Kaijie Yang, Siyi Hu, Zihao Wang, Guanghe Li, Yihang Sun, Cheng Zhang, Zhaowei Zhang, Anji Liu, Song-Chun Zhu, et al. 2024. Proagent: building proactive cooperative agents with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17591–17599.

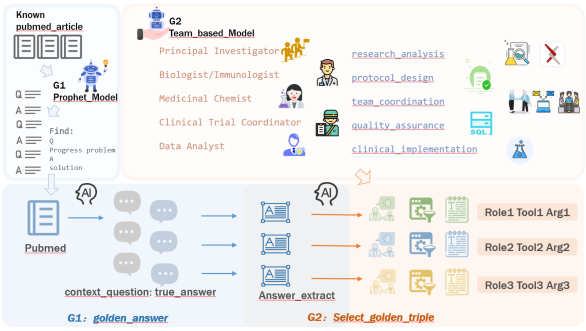

Figure 3: Data construction pipeline.

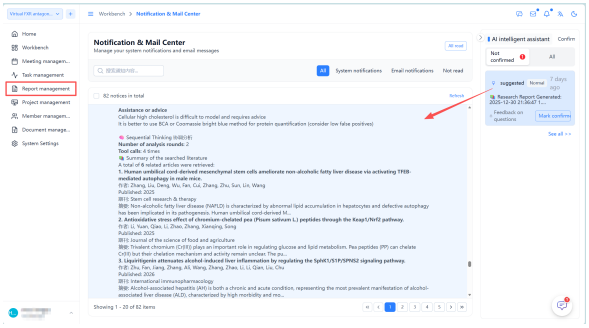

Figure 4: Proactive Coordination in the System Dashboard. This figure illustrates the agent's intervention logic. When a researcher submits a progress update via the "Report Management" portal, the assistant (right sidebar) proactively pushes a strategic recommendation (e.g., a Research Report generated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

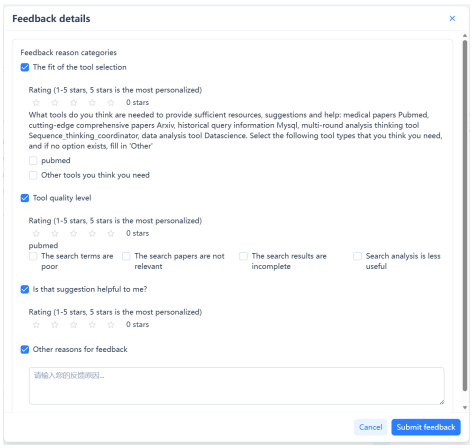

Figure 5: Multi-dimensional User Feedback Portal.

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.

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 7. 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 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 base model, 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

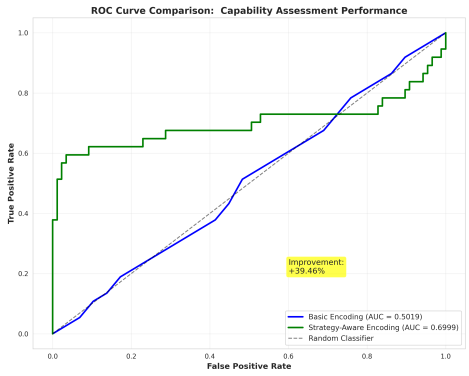

Figure 6: Performance gain in capability assessment. Comparison of ROC curves between the Basic Encoding baseline (AUC=0.5019) and our proposed Strategy-Aware Encoding (AUC=0.6999). The results demonstrate that the nnPU framework, combined with structural logic, achieves a significant improvement under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

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.

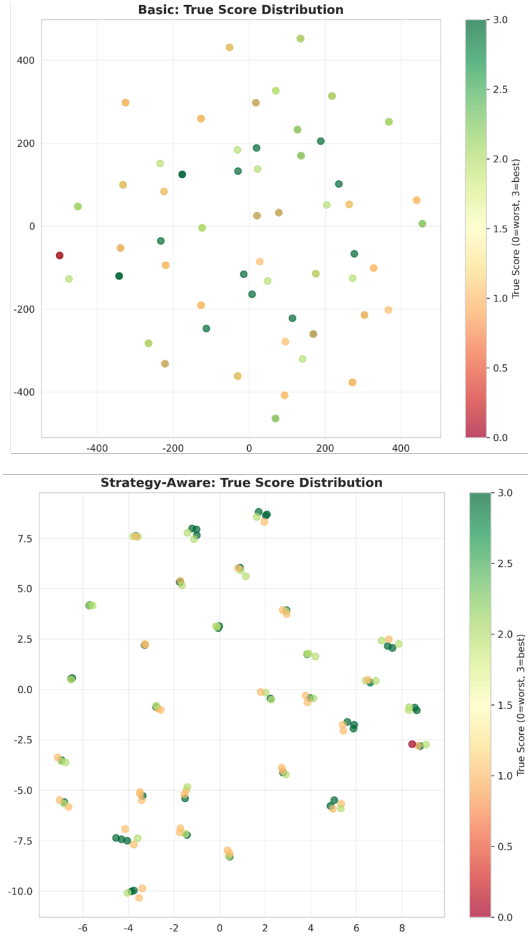

Figure 7: Visualization of the latent representation space via t-SNE. (Left) Basic encoding intermingles samples of different quality, failing to capture coordination logic. (Right) Strategy-Aware Triplet Encoding exhibits clear feature clustering, where high-quality strategies (green) are mapped into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

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.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).

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. Over-estimating 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:

$$WR_r^{(t)} = (1 - \alpha)WR_r^{(t-1)} + \alpha \cdot \mathbb{I}(\text{role } r \in \mathbf{a}^*) \tag{15}$$

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:

$$w_r^{t+1} = \eta \cdot w_r^t + \Delta w(\mathbf{a}^*, r) \tag{16}$$

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:

$$\lambda_t = \lambda_{base} \cdot \exp\left(\gamma \cdot \frac{\mathcal{V}(\mathbf{w}_t) - \bar{\mathcal{V}}}{\sigma_{\mathcal{V}}}\right) \tag{17}$$

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."