Learning to Orchestrate Heterogeneous Agents under Uncertainty

Anonymous Author(s)
Affiliation
Address
email

Abstract

Adaptive orchestration of heterogeneous agents requires making sequential delegation decisions under uncertain and evolving agent behaviour, e.g., coordinating specialised AI models with varying reliability, cost, and response quality. While prior work on agent orchestration focuses on performance or cost, uncertainty in agent reliability and output distributions is typically not modelled explicitly at the orchestration level. In this work, we study the problem of adaptive orchestration of heterogeneous agents under uncertainty, where a meta-controller must decide when to delegate to a single agent, accounting for reliability, cost, and uncertainty. We propose BOT-Orch, a lightweight framework that casts orchestration as a bandit problem over agents, regularized by OT distances between agent output distributions and task-specific reference distributions. We show that the regularised orchestration enjoys $\mathcal{O}(\sqrt{T})$ regret under standard assumptions, and provably induces preference ordering among agents with identical mean rewards but differing distributional alignment. Empirically, we demonstrate that BOT-Orch outperforms standard bandit and heuristic baselines in synthetic but adversarial task allocation settings with heterogeneous, non-i.i.d. agent behaviour.

1 Introduction

Incorporating uncertainty when coordinating multiple decision-making entities is a fundamental challenge across machine learning and autonomous systems [Wurman et al., 2007, Olfati-Saber et al., 2007, Shalev-Shwartz, 2012]. In real-world environments, where stochasticity is dominant (e.g., partial and noisy observations), variability in agent capabilities makes centralized control impractical and can break classical solutions. Classical multi-agent frameworks (e.g., Dec-POMDPs, stochastic games, DCOPs) explicitly model uncertainty in multi-agent decision-making but often assume known system models, have limited scalability, or simplify agent interactions, restricting applicability in complex coordination settings [Bernstein et al., 2002, Oliehoek and Amato, 2016, Hansen et al., 2004, Heifetz et al., 2006, Shoham and Leyton-Brown, 2008, Fioretto et al., 2018]. In practice, heterogeneity in agent reliability, operational cost, and resource constraints further complicates coordination, especially at scale where performance, risk, and cost must be jointly managed [Rizk et al., 2019, Arjun et al., 2025]. Surveys in distributed decision-making and heterogeneous multi-agent coordination consistently identify these issues as key barriers to scalable deployment [Olfati-Saber et al., 2007, Rizk et al., 2019].

Multi-agent reinforcement learning (MARL) provides tools for learning cooperative behaviors under partial observability and decentralised execution [Buşoniu et al., 2008, Hernandez-Leal et al., 2019]. Value decomposition and centralized training with decentralized policies can maximize joint rewards even with distinct roles and observations [Sunehag et al., 2018, Rashid et al., 2018]. However, many

Submitted to the ICML 2026 Workshop on "Decision-Making from Offline Datasets to Online Adaptation: Black-Box Optimization to Reinforcement Learning" (DEMO). Do not distribute.

MARL methods implicitly assume similar reliability and reaction costs, and rarely address adaptive selection among heterogeneous agents of uncertain outputs.

Modern AI systems and multi-agent orchestration frameworks sharpen these challenges. In planning, reasoning, and AI agent teams, systems increasingly rely on specialized agents with different expertise, reliability, and computational cost. Effective orchestration requires deciding which agent(s) to invoke and when to combine multiple predictions for robustness. Recent work shows that dynamic selection and composition conditioned on context can outperform static pipelines [Park et al., 2023, Liang et al., 2023, Cheng et al., 2023].

A natural abstraction for sequential decision-making under uncertainty is the multi-armed bandit (MAB) framework, capturing the exploration–exploitation trade-off [Lattimore and Szepesvári, 2020, Bubeck and Cesa-Bianchi, 2012]. Multi-agent extensions study cooperation, regret minimization, and communication among learners [Landgren et al., 2016, Gupta et al., 2021]. Yet most bandit models treat agents as interchangeable up to mean reward and ignore query costs, limiting their suitability for orchestration with heterogeneous reliability, non-stationarity, and explicit invocation costs.

Optimal Transport (OT) offers a complementary way to compare distributions and quantify discrepancies between uncertain outcomes [Villani, 2003, Peyré and Cuturi, 2019]. OT is widely used for uncertainty-aware distributional comparison in domain adaptation [Courty et al., 2016], generative modeling [Arjovsky et al., 2017, Bousquet et al., 2017], and statistical inference [Panaretos and Zemel, 2020]. However, OT remains underused in multi-agent coordination as a mechanism to compare agent output distributions and guide adaptive orchestration. Recent work suggests OT can be fruitfully combined with multi-agent reinforcement learning for scalability and alignment in complex environments [Baheri and Kochenderfer, 2024].

Our contribution.

- **Bandit-based orchestration with OT alignment:** We cast delegation over heterogeneous agents as a stochastic bandit regularized by OT distances between agent output distributions and task-specific references, enabling uncertainty-aware alignment and adaptive decision-making.
- **Theoretical guarantees:** We establish sublinear OT-regularized regret (1 of Theorem 4.1), structural optimality and robustness to noisy alignment (3), and convergence and consistency properties (5–4).
- **Empirical validation under heterogeneity and shift:** We evaluate BOT-Orch across synthetic and semi-synthetic settings, including a human–AI triage scenario under deployment shift (Section 6-7), demonstrating consistent improvements over standard bandit and heuristic baselines in both i.i.d. and non-i.i.d. environments.

2 Related Work

Agent orchestration. Prior work on agent orchestration largely focuses on selecting agents based on expected utility or offline accuracy [Keswani et al., 2021, Lai et al., 2022, Rasal and Hauer, 2024]. However, these approaches often overlook practical constraints such as availability, cost, and capability. Recent work on human–AI orchestration highlights how inter-agent interactions shape system-level decision networks [Collins et al., 2024], yet uncertainty-aware adaptive orchestration under realistic constraints remains underexplored. We address this gap via uncertainty-aware OT-based orchestration. In a related direction, DiscoPOP [Lu et al., 2024] learns loss functions without human input, optimizing over objectives. This suggests that automatic objective discovery could extend to orchestration.

Bandit approaches. Multi-armed bandits provide a natural framework for sequential decision making under uncertainty, balancing exploration and exploitation [Lattimore and Szepesvári, 2020, Chen, 2024, Tong, 2024, Anonymous, 2024]. Recent surveys summarize advances in classical and contextual bandits and their applications [Chen, 2024, Anonymous, 2024], while empirical studies highlight design choices affecting performance [Bietti et al., 2021]. Emerging work connects bandits with large language models in complex environments [Xie et al., 2026]. Here, we use bandits to learn orchestrations that maximize expected utility across heterogeneous agents with uncertain performance.

OT for uncertainty-aware orchestration. Optimal transport provides a principled framework for comparing probability distributions via geometrically meaningful discrepancies [Villani et al., 2008]. We use OT to quantify uncertainty in heterogeneous agent outputs through distributional disagreement and variability, building on recent work in OT-based uncertainty quantification [Oliver et al., 2025a,c,b]. This yields uncertainty-aware weights capturing both confidence and disagreement. The framework connects to distributional and uncertainty-aware reinforcement learning [Osband et al., 2013, Bellemare et al., 2017], providing a unified approach to adaptive multi-agent coordination with theoretical guarantees and strong empirical performance.

3 Preliminaries and Notation

3.1 Agents, Tasks, and Task Space

Let $\mathcal{A} := \{a_1, \ldots, a_M\}$ denote a finite set of M agents, and let $\mathcal{X} \subset \mathbb{R}^d$ be a measurable task space endowed with its Borel σ-algebra. Time evolves over a finite horizon of T rounds indexed by $t = 1, \ldots, T$. At each round t, a task $x_t \in \mathcal{X}$ arrives, forming a stochastic process $(x_t)_{t=1}^T$. For each round s and each agent $a_i \in \mathcal{A}$, let $R_s^i \in \mathbb{R}$ denote the reward that agent a_i would obtain from handling task x_s, and let $W_s^i \in \mathbb{R}$ denote the task–agent alignment cost (or mismatch cost) associated with assigning task x_s to agent a_i. We collect these quantities into vectors $\mathbf{R}_s := (R_s^1, \ldots, R_s^M)^\top \in \mathbb{R}^M$ and $\mathbf{W}_s := (W_s^1, \ldots, W_s^M)^\top \in \mathbb{R}^M$. Task arrivals may depend on past observations, environmental conditions, or previous assignments. Define the history up to (but not including) time t as $\mathcal{H}_t := \{(x_s, \mathbf{R}_s, \mathbf{W}_s)\}_{s=1}^{t-1}$. Then each task is drawn from a history-dependent conditional distribution $x_t \sim \mathcal{P}(\cdot \mid \mathcal{H}_t)$, $t = 1, \ldots, T$, where $\mathcal{P}(\cdot \mid \mathcal{H}_t)$ is a probability measure over $\mathcal{X}$ conditioned on the past history.

3.2 Correlated Multi-Agent Rewards

Recall that $\mathbf{R}_t \in \mathbb{R}^M$ denotes the vector of rewards for all agents at round t (defined in Section 2.1). In general, rewards may depend on both the current task and past system outcomes. We model this via the conditional first and second moments:

$$\mathbb{E}[\mathbf{R}_t \mid x_t, \mathcal{H}_t] = \mathbf{r}(x_t) + \mathbf{f}(\mathbf{R}_{1:t-1}),\ \mathrm{Cov}[\mathbf{R}_t \mid x_t, \mathcal{H}_t] = \Sigma_t.$$

Here $\mathbf{r} : \mathcal{X} \to \mathbb{R}^M$ is the baseline task-dependent mean reward vector, $\mathbf{f} : \mathbb{R}^{M\times(t-1)} \to \mathbb{R}^M$ captures temporal dependence on past realized rewards (e.g., learning effects, fatigue, or system congestion), and $\Sigma_t \in \mathbb{R}^{M\times M}$ is a time-varying covariance matrix that models uncertainty and cross-agent correlations at time t.

I.I.D. tasks as a special case. A common simplifying assumption is that tasks are independent and identically distributed:

$$x_t \overset{\text{i.i.d.}}{\sim} \mathcal{P}_X, \quad t = 1, \ldots, T,$$

for some fixed distribution $\mathcal{P}_X$ over $\mathcal{X}$. If, in addition, rewards do not depend on past outcomes, then the history $\mathcal{H}_t$ provides no additional information once x_t is known. In this case, the temporal dependence term vanishes and the covariance becomes time-invariant:

$$\mathbb{E}[\mathbf{R}_t \mid x_t] = \mathbf{r}(x_t), \qquad \mathrm{Cov}[\mathbf{R}_t \mid x_t] = \Sigma,$$

where Σ is a constant matrix. This corresponds exactly to the special case of the correlated reward model above with $\mathbf{f} \equiv 0$ and $\Sigma_t \equiv \Sigma$ for all t. Thus, the i.i.d. setting removes both temporal dependence in the mean rewards and time variation in the covariance structure.

3.3 Survival-Based Rewards with Latent Frailty

In many real-world systems, agent performance is naturally measured by the *time to complete a task* or *time until failure*, as in service response times, human or robotic workflows, or component lifetimes. Such outcomes are commonly modelled using *survival analysis*, where the primary variable is a *time-to-event*. Formally, let $T_t(i) > 0$ denote the time-to-event for agent a_i on task x_t. Since event times may be only partially observed, we introduce a *censoring indicator* $\delta_t(i) \in \{0, 1\}$, where $\delta_t(i) = 1$ indicates a fully observed event and $\delta_t(i) = 0$ indicates right-censoring.

Let $S_i(\tau \mid x_t) = \mathbb{P}\big(T_t(i) > \tau \mid x_t\big)$ denote the *survival function* of agent a_i, capturing the probability that the task is not completed by time τ, and allowing for heterogeneous performance across agents and tasks. To model unobserved task-level factors affecting all agents, such as difficulty or system load, we introduce a *latent frailty variable* $\theta_t > 0$, a shared random effect that multiplicatively scales task difficulty.

The survival model is given by $\mathbb{P}\big(T_t(i) > \tau \mid x_t, \theta_t\big) = S_i(\tau \mid x_t)^{\theta_t}$, $R_t(i) = \delta_t(i)\, S_i\big(T_t(i) \mid x_t\big)^{\theta_t}$. Under this formulation, agents on the same task share θ_t, inducing positive correlation in outcomes: more difficult tasks lead to longer completion times for all agents. This survival-based reward model is well-suited to settings involving reliability, speed, or risk, where hidden task-level factors jointly influence performance.

3.4 OT-Based Alignment Costs

Beyond rewards, we also model how well an agent's capabilities align with the requirements of a task. In many applications, both tasks and agents are naturally described by *distributions* rather than single feature vectors. For example, an agent may produce a distribution of outcomes (quality levels, response times, error types), while a task may specify a desired target distribution over outcomes.

Let $\mu_i \in \mathcal{P}(\mathcal{Y})$ denote the outcome distribution induced by agent a_i over a measurable space $\mathcal{Y}$, and let $\nu_t \in \mathcal{P}(\mathcal{Y})$ represent the reference or desired outcome distribution associated with task x_t. We quantify the mismatch between an agent and a task using the Wasserstein distance $W_c(\nu_t, \mu_i)$, which measures the minimal cost of transporting mass from one distribution to the other under ground cost c. This provides a geometrically meaningful notion of alignment that accounts for the full distribution of outcomes rather than just summary statistics.

To incorporate randomness and modeling noise, we define the stochastic alignment cost $\mathcal{W}_t(i) := W_c(\nu_t, \mu_i) + \epsilon_t(i)$, with $\epsilon_t(i) \sim \mathcal{N}(0, \sigma_i^2)$. This formulation captures both systematic mismatch (via the Wasserstein term) and unpredictable variability (via $\epsilon_t(i)$). Introducing alignment costs at this stage allows us to jointly model (i) how well an agent is suited to a task and (ii) the stochastic rewards that result from performing it, providing a unified framework for assignment decisions under uncertainty.

3.5 Orchestration Policy and Objective

We now formalize the decision-making problem faced by the orchestrator. At each round t, after observing the task x_t and past history $\mathcal{H}_t$, the orchestrator selects a *randomized assignment policy* $\pi_t \in \Delta(\mathcal{A})$, where $\Delta(\mathcal{A})$ denotes the probability simplex over agents. An agent $i_t \sim \pi_t$ is then sampled and assigned to handle task x_t. Randomized policies allow exploration and robustness to uncertainty in agent performance.

Given a policy π_t, the expected net reward at round t is defined as

$$R_t(\pi_t) := \pi_t^\top \Big(\mathbf{r}(x_t) - \lambda\, \mathbb{E}[\mathcal{W}_t]\Big), \qquad \lambda > 0,$$

where $\mathbf{r}(x_t) \in \mathbb{R}^M$ is the vector of expected task-dependent rewards and $\mathcal{W}_t \in \mathbb{R}^M$ is the vector of alignment costs. The scalar parameter $\lambda > 0$ controls the trade-off between maximizing performance and minimizing mismatch costs. Thus, $R_t(\pi_t)$ represents the expected utility of assigning the task according to π_t, balancing reward quality against alignment penalties.

The overall objective is to maximize the cumulative expected net reward over the time horizon:

$$J(\pi_{1:T}) := \sum_{t=1}^{T} \mathbb{E}[R_t(\pi_t)],$$

where the expectation is taken over task arrivals, reward randomness, alignment noise, and the orchestrator's own randomization. This objective formalizes the goal of learning an assignment strategy that performs well on average while accounting for uncertainty and task-agent compatibility.

We may also consider a finite set of candidate orchestration strategies $\mathcal{S} = \{s_1, \ldots, s_K\}$, where each strategy s_k specifies a rule for selecting policies $\pi_t^{(k)}$ based on the observed history.

3.6 Regret and Optimal Policy

The optimal policy for known rewards and alignment costs is

$$\pi_t^* := \arg\max_{\pi \in \Delta(\mathcal{A})} \pi^\top \Big(\mathbf{r}(x_t) - \lambda\, \mathbb{E}[\mathcal{W}_t] \Big), \tag{1}$$

with cumulative regret

$$\mathcal{R}_T := \sum_{t=1}^{T} (\pi_t^*)^\top \big(\mathbf{r}(x_t) - \lambda\, \mathbb{E}[\mathcal{W}_t] \big) - \sum_{t=1}^{T} \mathbb{E}[R_t(\pi_t)]. \tag{2}$$

For i.i.d. tasks, the regret simplifies due to stationarity. For non-i.i.d. tasks, $\mathcal{R}_T$ captures efficiency loss arising from temporal correlations, history-dependent distributions, and uncertainty in strategy selection.

4 Theoretical Properties

We begin by stating the standing assumptions that remain in force throughout.

(A1) The ground cost $c : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is L-Lipschitz and bounded.

(A2) Rewards are uniformly bounded: $0 \le R_t(i) \le R_{\max}$ for all t and all actions $i \in \mathcal{A}$.

(A3) The frailty variables $(\theta_t)_{t \ge 1}$ admit finite exponential moments.

(A4) Conditional on (x_t, θ_t), censoring is independent of survival time.

(A5) The learning rate satisfies $\eta_t = O(t^{-1/2})$.

Assumptions (A1)–(A5) ensure: (i) stability of Wasserstein distances under empirical perturbations; (ii) sub-exponential behaviour of frailty-adjusted rewards; (iii) well-posedness of the induced softmax stochastic approximation dynamics.

We remark that these conditions are standard in the literature on stochastic approximation, Wasserstein-based learning, and frailty-adjusted reward models [Ambrosio et al., 2008, Shalev-Shwartz and Ben-David, 2014, Trillos and Slepčev, 2016]. In practice, they are not restrictive: Lipschitz and bounded costs are typical in OT and multi-agent learning applications, bounded rewards naturally arise in reinforcement learning, and learning rates of order $t^{-1/2}$ are widely used to guarantee convergence. Moreover, our framework generalizes several prior works by allowing frailty variables with arbitrary distributions admitting finite exponential moments, rather than restricting to specific parametric forms [Del Barrio and Loubes, 2019, Wang et al., 2020].

4.1 Main Results

Following the assumptions in Section 4, we now present the main theoretical guarantees for the OT-regularised bandit model. The regret analysis is conducted for a general exponential-weights (softmax) procedure applied to bounded, OT-regularised reward signals. Importantly, the resulting guarantees depend only on the boundedness of these rewards and are independent of the specific generative model.

The additional modelling components introduced in the setup, namely, correlated rewards, survival-based frailty, and non-i.i.d. task generation—serve as a motivating probabilistic framework in which such bounded reward processes naturally arise. They are not directly used in the regret derivation, but instead provide one possible instantiation of the abstract reward model. Proofs are provided in Appendices B.1–B.3.

Theorem 4.1 (Performance and Convergence of OT-Regularized Bandits)**.** *Let $\mathcal{A}$ be a finite set of agents, and let $(\tilde{r}_t(i))_{i \in \mathcal{A}}$ denote the BOT–Orch alignment-adjusted rewards defined by*

$$\tilde{r}_t(i) = \hat{r}_t(i) - \lambda W_c(\mu_i, \nu_t),$$

where $\hat{r}_t(i)$ is the empirical reward, μ_i is the agent-specific outcome distribution, ν_t is the target distribution at time t, and W_c is the ground-cost Wasserstein distance. Assume (A1)–(A5) hold.

Environment	BOT-Orch	No-OT (λ=0)	Random	UCB1 (MAB)
IID				
IID-G	**537.43±16.57**	656.31±22.64	673.99±19.08	664.89±29.85
IID-M	**459.84±9.27**	545.85±17.74	553.07±19.16	549.48±7.40
Non-IID				
NonIID-BB	**410.04±71.06**	503.14±91.24	520.28±101.61	496.60±79.57
NonIID-PS	**571.43±18.71**	713.81±25.79	734.65±30.08	729.51±28.34
NonIID-SD	**564.13±18.11**	717.24±22.51	707.69±18.33	708.60±16.06

Table 1: Cumulative Alignment Cost (mean ± 95% CI across 5 seeds; $T = 200$). Best in bold.

Let $i_t^* = \arg\max_{i \in \mathcal{A}} \tilde{r}_t(i)$ denote the optimal agent at round t, and define the OT-regularized cumulative regret by

$$\mathcal{R}_T := \sum_{t=1}^{T} \big(\tilde{r}_t(i_t^*) - \tilde{r}_t(i_t)\big),$$

where i_t is the agent chosen by the BOT–Orch policy at round t. Let $\phi_t = \pi_t(\cdot)$ denote the strategy-mixing vector. Suppose that the initial conditions are well-defined, i.e., $\hat{r}_0(i)$ is finite for all $i \in \mathcal{A}$, and $\phi_0 \in \Delta^{|\mathcal{A}|-1}$. Then, the following statements hold:

1. ***Sublinear OT-Regret.*** *The cumulative OT-regularized regret satisfies $\mathcal{R}_T = O(\sqrt{T})$.*
2. ***Structural OT-Optimality.*** *For any two agents $i, j \in \mathcal{A}$ with $\hat{r}_t(i) = \hat{r}_t(j)$, if $W_c(\mu_i, \nu_t) < W_c(\mu_j, \nu_t)$, then $\tilde{r}_t(i) > \tilde{r}_t(j)$.*
3. ***Margin Robustness under Noisy Alignment.*** *Suppose the observed Wasserstein distances are perturbed by independent Gaussian noise $\epsilon_t(i) \sim \mathcal{N}(0, \sigma^2)$, defining $\widetilde{W}_c(\mu_i, \nu_t) = W_c(\mu_i, \nu_t) + \epsilon_t(i)$, and let $\Delta_{ij} = W_c(\mu_j, \nu_t) - W_c(\mu_i, \nu_t)$. If it holds $|\Delta_{ij}| > \sigma\sqrt{2\log 2}$, then $\mathbb{P}\big(\tilde{r}_t(i) < \tilde{r}_t(j)\big) < 1/4$.*
4. ***Convergence of Orchestration Weights.*** *The sequence of policy vectors $(\phi_t)_{t \geq 1}$ converges almost surely to a stable equilibrium ϕ_∞ of the limiting ODE $\dot{\phi} = \mathrm{Softmax}\big(\mathbb{E}[\tilde{r}]\big) - \phi$.*
5. ***Uniform Consistency of Empirical Rewards.*** *Let $\hat{r}_t(i)$ denote the empirical average of frailty-adjusted rewards for agent i. Then,*

$$\sup_{i \in \mathcal{A}} \big|\hat{r}_t(i) - \mathbb{E}[R_t(i) \mid x_t]\big| \xrightarrow[t \to \infty]{\text{a.s.}} 0.$$

We note that Theorem 4.1 provides a foundational characterization of OT-regularized bandit learning. It integrates distributional alignment into the reward structure, yielding principled agent differentiation 2, sublinear cumulative regret 1, robustness to noisy observations 3, convergence of orchestration weights 4, and uniform consistency of empirical rewards 5. The framework enables heterogeneous agents to learn coordinated policies under distributionally-aware uncertainty, supporting robust, adaptive orchestration in stochastic, partially observable environments. When task-specific reference distributions are unavailable, they can be constructed using Wasserstein barycenters [Chewi et al., 2025] of observed empirical task distributions. We refer the reader to Appendix B for a proof of Theorem 4.1.

5 The proposed algorithm

BOT-Orch combines bandit-based selection with OT alignment and survival-based rewards to handle heterogeneous, non-stationary agents. In the *i.i.d. task setting* (Algorithm 1 in the Appendix), a Boltzmann policy selects agents using exponentially smoothed rewards penalized by OT misalignment, balancing exploitation and alignment, while survival rewards capture latent difficulty, censoring, and reliability. The *non-i.i.d. extension* (Algorithm 2 in the Appendix) allows history-dependent task distributions and reward updates, handling temporal correlations, non-stationarity, and regime shifts via a correction term that encodes memory effects. Overall, BOT-Orch is a *risk-aware, alignment-regularized bandit algorithm* in distributional space, where OT enforces task-agent compatibility and survival rewards provide robustness to censoring and heterogeneity. Its modular design accommodates alternative OT solvers, survival models, and exploration schemes, and motivates regret analysis under composite reward–cost objectives as well as questions of convergence and adaptivity in non-stationary environments.

	Cumulative Net Utility				Oracle Regret			
Environment	BOT-Orch	No-OT (λ=0)	Random	UCB1 (MAB)	BOT-Orch	No-OT (λ=0)	Random	UCB1 (MAB)
IID								
IID-G	**-467.528±17.60**	-588.34±23.19	-605.71±18.91	-603.31±29.13	**122.65±5.68**	243.47±14.26	260.83±11.59	258.43±18.11
IID-M	**-386.10±10.04**	-478.54±20.49	-482.81±18.65	-484.57±9.35	**74.70±5.45**	166.25±11.99	170.52±10.99	172.28±8.40
Non-IID								
NonIID-BB	**-335.25±70.42**	-434.01±86.69	-449.34±100.73	-426.09±79.48	**76.85±7.42**	175.61±16.44	190.94±31.44	167.69±19.20
NonIID-PS	**-498.28±22.49**	-642.77±25.89	-666.10±29.31	-663.22±27.77	**126.77±11.68**	271.26±11.84	294.58±20.25	291.71±14.45
NonIID-SD	**-492.92±18.74**	-644.68±21.62	-637.82±18.54	-643.69±17.20	**128.77±13.30**	280.53±10.81	273.67±10.74	279.54±11.52

Table 2: Cumulative Net Utility and Oracle Regret (mean $\pm$ 95% CI across 5 seeds; $T = 200$). Best in bold.

6 Synthetic Experiments

6.1 Dataset and Task Description

We consider both i.i.d and non-i.i.d. regimes (see Fig. 2-3 in the Appendix):

- **IID-G:** Tasks x_t sampled i.i.d. over $\mathcal{X}$. Agent rewards drawn i.i.d. from fixed Gaussian distributions with matched mean (≈ 0.5) but heterogeneous higher-order shape.
- **IID-M:** Tasks x_t sampled i.i.d. from a half-moons distribution. Agent rewards drawn i.i.d. from fixed distributions with matched mean (≈ 0.5) but differing variance/skewness/bimodality.
- **NonIID-PS:** Piecewise-stationary rewards: mean fixed, variance shifts at unknown changepoints, inducing distributional shifts without mean change.
- **NonIID-SD:** Smooth non-stationarity via sinusoidal drift in reward means.
- **NonIID-BB:** Latent reward means follow a temporally correlated Brownian-bridge path with fixed endpoints.

All experiments were conducted on a standard x86_64 CPU platform using 2 CPU cores, 13.6 GB RAM, and 107 GB disk storage.

6.2 Baselines

- **BOT-Orch (ours).** Uses alignment-adjusted rewards $\tilde{r}_t(i) = \hat{r}_t(i) - \lambda W_t(i)$ and samples the selected agent i_t from a softmax policy over $\tilde{r}_t(i)$ (Algorithm 1–2). This couples OT alignment with sequential exploration/exploitation.
- **No-OT ($\lambda = 0$).** Ablation removing the OT alignment term. The policy is computed from $\hat{r}_t(i)$ only. This isolates the contribution of distributional alignment.
- **Random.** Uniformly samples an agent each round. This serves as a naive lower bound that does not learn from observations.
- **UCB1 (MAB).** A classical multi-armed bandit baseline that selects the agent with the highest upper confidence bound based on empirical reward estimates, balancing exploration and exploitation without OT alignment.

We omit Greedy-EMA, which picks at each round the agent with the highest exponentially-smoothed reward estimate, because typical implementations assume *full-information* rewards for all agents each round, unlike our *bandit feedback* setting (only the chosen agent is observed). A bandit-compatible "greedy" would require exploration/confidence bounds and is essentially covered by standard bandit baselines (e.g., UCB1).

6.3 Evaluation Metrics

- **Cumulative net utility.** We evaluate the OT-regularized net utility $U_t(i_t) = R_t(i_t) - \lambda W_t(i_t)$, and report the cumulative net utility $\sum_{t=1}^{T} U_t(i_t)$. This captures the overall performance when explicitly trading off reward quality and alignment cost via λ.
- **Cumulative alignment cost.** We report $\sum_{t=1}^{T} W_t(i_t)$, the cumulative OT alignment cost incurred by the selected agent assignments. Lower values indicate more distributionally aligned agent-task matching.

	Cumulative Net Utility				Oracle Regret			
	BOT-Orch	No-OT (λ=0)	Random	UCB1	BOT-Orch	No-OT (λ=0)	Random	UCB1
IID	**108.84±2.22**	103.37±2.51	83.98±5.92	80.55±4.88	**2.29±2.11**	9.80±1.61	28.72±6.03	31.31±3.98
Non-IID	**110.61±1.03**	103.17±1.80	79.78±5.52	85.82±4.91	**0.59±0.93**	10.14±1.17	32.97±4.87	26.26±4.28

	Team Accuracy				Cumulative Alignment Cost			
	BOT-Orch	No-OT (λ=0)	Random	UCB1	BOT-Orch	No-OT (λ=0)	Random	UCB1
IID	**0.980±0.010**	0.907±0.022	0.917±0.020	0.902±0.026	**1.28±0.53**	10.51±1.61	10.28±2.01	11.14±1.33
Non-IID	**0.993±0.007**	0.905±0.016	0.905±0.024	0.919±0.025	**0.85±0.23**	10.84±1.18	11.69±1.62	9.46±1.43

Table 3: Deployment metrics across all four methods and both experimental conditions (Mean ± 95% CI across 30 seeds, T=114, λ=3.0). IID condition uses Algorithm 1; Non-IID uses Algorithm 2 (ID patients rounds 1–57, shifted rounds 58–114). Best in bold.

- **Oracle regret.** We report OT-regularized cumulative regret relative to the best alignment-adjusted agent at each round: $\sum_{t=1}^{T}(\max_i U_t(i) - U_t(i_t))$ measuring efficiency loss due to suboptimal selection under uncertainty and non-stationarity.

6.4 Results

BOT-Orch achieves the highest cumulative net utility and lowest oracle regret across all environments (Table 1-2), consistently outperforming No-OT, Random, and UCB1. The improvement is particularly pronounced in non-i.i.d. settings, where BOT-Orch maintains low regret under distributional shifts such as piecewise variance changes and smooth drift. These results indicate that incorporating OT-based alignment enables more effective adaptation to heterogeneous and non-stationary agent behaviour.

In Appendix D, we evaluate BOT-Orch on additional synthetic benchmarks using survival-based metrics, showing that BOT-Orch achieves best performance. We further conduct an ablation study on the parameter λ (Appendix D.5). As λ increases, BOT-Orch exhibits performance improvements, with large gains relative to $\lambda = 0$ and diminishing returns at higher values.

7 Semi-Synthetic Experiments: Human-AI Triage Under Deployment Shift

7.1 Dataset and Task Description

We use the Breast Cancer Wisconsin (Diagnostic) dataset, consisting of 569 patient cases with 30 numerical features and a binary target (malignant vs. benign). The data is split into 60% train, 20% calibration, 10% Test-ID, and 10% Test-Shift. To simulate deployment shift, we perturb a subset of features in Test-Shift with additive Gaussian noise $\mathcal{N}(0, 0.64)$ and a $+0.5$ standard-unit bias, modelling a change in patient population at deployment. At each round t, a patient x_t arrives and the selected agent receives a reward of 1 if it classifies the patient correctly and 0 otherwise, under standard bandit feedback.

We consider two agents: an AI classifier and a human proxy with complementary accuracy, where the human performs better on shifted cases while the AI performs better in-distribution. Full model details and accuracy statistics are provided in Appendix E. All experiments were conducted on a standard x86_64 CPU platform using 2 CPU cores, 13.6 GB RAM, and 107 GB disk storage.

7.2 Baselines

We compare the same four methods used in the synthetic experiments, now applied to the clinical triage task. We use $\alpha = 0.90$, $\eta = 5.0$, $\lambda = 3.0$, and, for the non-i.i.d. variant, the same history correction form as Algorithm 2 with $\beta = 0.05$. Each episode has $T = 114$ rounds, and we average over $n = 30$ random seeds (patient orderings). We report mean ± 95% CI across seeds.

7.3 Evaluation Metrics

From Section 6.3, we use cumulative net utility, cumulative alignment cost, and oracle regret, defined as above. We additionally report metrics specific to the human–AI deferral setting:

- **Team accuracy**: fraction of patients correctly classified by the selected agent (AI or human).

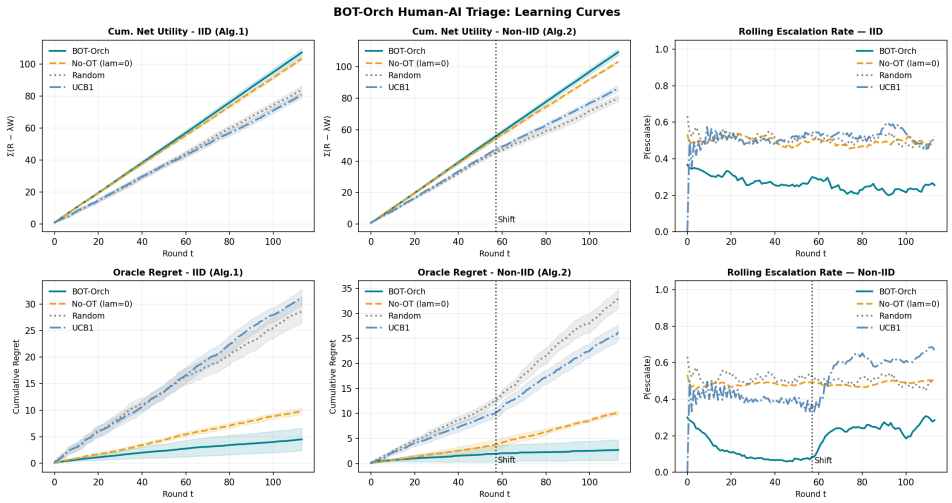

Figure 1: **Cumulative learning curves.** *Top row*: cumulative net utility. *Bottom row*: oracle regret. *Right column*: rolling escalation rate (window w=8 rounds). Left panels: IID condition (Algorithm 1); middle panels: Non-IID condition (Algorithm 2), with the dotted vertical line marking the shift onset at round 57; right panels: escalation rate evolution.

- **Escalation rate**: fraction of patients routed to the human expert.

7.4 Results

Table 3 summarizes the results. BOT-Orch achieves the highest cumulative net utility and lowest oracle regret in both regimes, outperforming all baselines. In the IID condition, it improves net utility by $+5.5$ over No-OT while reducing alignment cost by an order of magnitude. In the Non-IID setting, the gain is larger, reaching 110.61 ± 1.03 net utility and 0.59 ± 0.93 regret compared to 103.17 ± 1.80 and 10.14 ± 1.18 for No-OT.

The No-OT ablation confirms that the OT term drives the improvement: removing it increases regret and alignment cost. BOT-Orch also exhibits targeted escalation under shift, routing fewer patients overall while increasing escalation on shifted cases, indicating effective adaptation to distributional mismatch. Figure 1 further shows that BOT-Orch separates from all baselines early and maintains the performance gap over time. In the Non-IID setting, the escalation rate increases after the shift, demonstrating online adaptation to the changing population. Furthermore, the rolling escalation rate in Figure 1 shows how the policy adapts over time, increasing reliance on the human expert when distributional shift or uncertainty rises.

Additional results are presented in Appendix F, including more results on the escalation rate, diagnostic analysis, and an ablation study on the parameter λ. We observe that, as λ increases, performance improves markedly peaking around $\lambda \approx 3.0$, after which gains saturate or slightly diminish at very high values.

8 Discussion and Conclusion.

This paper formulates heterogeneous agent orchestration under uncertain as a sequential decision problem and introduces BOT-Orch, a bandit-based framework that jointly learns delegation while incorporating optimal transport alignment costs. While BOT-Orch provides a principled approach with strong theoretical and empirical performance, it has various practical limitations. First, computing Wasserstein distances can be expensive, especially in high-dimensional settings, limiting real-time use. Second, the reward–alignment trade-off is controlled by a scalar parameter that may require tuning across environments. Third, the framework assumes access to task-specific reference distributions, which may be unavailable in open or non-stationary settings. Addressing these challenges remains an important direction for future work.

References

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Basel, 2nd edition, 2008.

Anonymous. Bandit algorithms: A comprehensive review and their dynamic selection from a portfolio for multicriteria top-k recommendation. *Expert Systems with Applications*, 246:123151, 2024. doi: 10.1016/j.eswa.2024.123151.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. In *ICML*, 2017.

Krishna Arjun, David Parlevliet, Hai Wang, and Amirmehdi Yazdani. Optimizing coalition formation strategies for scalable multi-robot task allocation: A comprehensive survey. *Robotics*, 14(7): 93, 2025. doi: 10.3390/robotics14070093. URL `https://www.mdpi.com/2218-6581/14/7/93`.

Ali Baheri and Mykel J. Kochenderfer. The synergy between optimal transport theory and multi-agent reinforcement learning. *arXiv preprint arXiv:2401.10949*, 2024. Explores integration of optimal transport theory with MARL for policy alignment, resource distribution, and non-stationarity mitigation.

Marc G Bellemare, Will Dabney, and Remi Munos. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, pages 449–458, 2017.

Michel Benaïm. Dynamics of stochastic approximation algorithms. *Annals of Probability*, 27(1): 361–414, 1999. doi: 10.1214/aop/1019160366.

Daniel S. Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. The complexity of decentralized control of markov decision processes. *Mathematics of Operations Research*, 27(4): 819–840, 2002.

Alberto Bietti, Alekh Agarwal, and John Langford. A contextual bandit bake-off. In *Journal of Machine Learning Research*, volume 22, pages 1–49, 2021.

Olivier Bousquet, Sylvain Gelly, Ilya Tolstikhin, Carl-Johann Simon-Gabriel, and Bernhard Schoelkopf. From optimal transport to generative modeling: the vegan cookbook. *arXiv preprint arXiv:1705.07642*, 2017.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

Lucian Buşoniu, Robert Babuška, and Bart De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 38(2): 156–172, 2008.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Qiufan Chen. A survey on contextual multi-armed bandits. *Applied and Computational Engineering*, 53:287–295, 2024. doi: 10.54254/2755-2721/53/20241593.

Yao Cheng et al. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023.

Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. *Wasserstein Barycenters*, pages 211–227. Springer Nature Switzerland, Cham, 2025. ISBN 978-3-031-85160-5. doi: 10.1007/978-3-031-85160-5_8. URL `https://doi.org/10.1007/978-3-031-85160-5_8`.

Katherine M Collins, Valerie Chen, Ilia Sucholutsky, Hannah Rose Kirk, Malak Sadek, Holli Sargeant, Ameet Talwalkar, Adrian Weller, and Umang Bhatt. Modulating language model experiences through frictions. *arXiv preprint arXiv:2407.12804*, 2024.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

Eustasio Del Barrio and Jean-Michel Loubes. Frailty models and empirical process theory. *Annals of Statistics*, 47(5):2519–2543, 2019.

Ferdinando Fioretto, Enrico Pontelli, and William Yeoh. Distributed constraint optimization problems and applications: A survey. *Journal of Artificial Intelligence Research*, 61:623–698, 2018.

Samarth Gupta, Shreyas Chaudhari, Gauri Joshi, and Osman Yağan. Multi-armed bandits with correlated arms. *IEEE Transactions on Information Theory*, 67(10):6711–6732, 2021.

Eric A. Hansen, Daniel S. Bernstein, and Shlomo Zilberstein. Dynamic programming for partially observable stochastic games. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, 2004.

Aviad Heifetz, Martin Meier, and Burkhard Schipper. Interactive unawareness. *Journal of Economic Theory*, 130(1):78–94, 2006.

Pablo Hernandez-Leal, Bilal Kartal, and Matthew E. Taylor. A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 33:750–797, 2019.

Vijay Keswani, Matthew Lease, and Krishnaram Kenthapadi. Towards unbiased and accurate deferral to multiple experts. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 154–165, 2021.

Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q Vera Liao, Yunfeng Zhang, and Chenhao Tan. Human-ai collaboration via conditional delegation: A case study of content moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2022.

Peter Landgren, Vaibhav Srivastava, and Naomi E. Leonard. Distributed cooperative decision-making in multiarmed bandits: Frequentist and bayesian algorithms. *CDC*, 2016.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Jacky Liang, Wenjia Wang, et al. Taskmatrix.ai: Completing tasks by connecting foundation models with millions of apis. *arXiv preprint arXiv:2303.16434*, 2023.

Chris Lu, Samuel Holt, Claudio Fanconi, Alex Chan, Jakob Foerster, Mihaela van der Schaar, and Robert Lange. Discovering preference optimization algorithms with and for large language models. *Advances in Neural Information Processing Systems*, 37:86528–86573, 2024.

Reza Olfati-Saber, J. Alex Fax, and Richard M. Murray. Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE*, 95(1):215–233, 2007. URL `https://doi.org/10.1109/JPROC.2006.887293`.

Frans A. Oliehoek and Christopher Amato. *A Concise Introduction to Decentralized POMDPs*. Springer, 2016.

Mary Chriselda Antony Oliver, Matthew Graham, Ioanna Manolopoulou, Graham F Medley, Lorenzo Pellis, Koen B Pouwels, Matthew Thorpe, and T Deirdre Hollingsworth. Uncertainty quantification in cost-effectiveness analysis for stochastic-based infectious disease models: Insights from surveillance on lymphatic filariasis. *Journal of Theoretical Biology*, page 112197, 2025a.

Mary Chriselda Antony Oliver, Emmanuel Hartman, and Tom Needham. Conic formulations of transport metrics for unbalanced measure networks and hypernetworks. *arXiv preprint arXiv:2508.10888*, 2025b.

Mary Chriselda Antony Oliver, Michael Roberts, Carola-Bibiane Schönlieb, and Matthew Thorpe. Laplace learning in wasserstein space. *arXiv preprint arXiv:2511.13229*, 2025c.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pages 3003–3011, 2013.

Victor M. Panaretos and Yoav Zemel. *Statistical aspects of Wasserstein distances*. Annual Review of Statistics and Its Application, 2020.

Joon Sung Park, Joseph O'Brien, Carrie J. Cai, et al. Generative agents: Interactive simulacra of human behavior. *UIST*, 2023.

Gabriel Peyré and Marco Cuturi. *Computational Optimal Transport*. Foundations and Trends in Machine Learning, 2019.

Sumedh Rasal and EJ Hauer. Navigating complexity: Orchestrated problem solving with multi-agent llms. *arXiv preprint arXiv:2402.16713*, 2024.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, et al. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *ICML*, 2018.

Yara Rizk, Mariette Awad, and Edward W. Tunstel. Cooperative heterogeneous multi-robot systems: A survey. *ACM Computing Surveys*, 52(2):1–31, 2019. doi: 10.1145/3303848. URL `https://doi.org/10.1145/3303848`.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4:107–194, 2012.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Yoav Shoham and Kevin Leyton-Brown. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press, 2008. ISBN 9780521899437.

Peter Sunehag, Guy Lever, Audrunas Gruslys, et al. Value-decomposition networks for cooperative multi-agent learning. In *AAMAS*, 2018.

Ruoyi Tong. A survey of the application and technical improvement of the multi-armed bandit. *Applied and Computational Engineering*, 77:25–31, 2024.

Nicolas Garcia Trillos and Dejan Slepčev. A variational approach to the consistency of spectral clustering. *Applied and Computational Harmonic Analysis*, 40(2):274–319, 2016.

Cédric Villani. *Topics in Optimal Transportation*. American Mathematical Society, 2003.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2008.

Peter Walters. *An Introduction to Ergodic Theory*, volume 79 of *Graduate Texts in Mathematics*. Springer, 1982. ISBN 978-0-387-90626-6.

Qiang Wang et al. Stochastic approximation for survival models with frailty terms. *Statistics in Medicine*, 39(10):1405–1420, 2020.

Peter R. Wurman, Raffaello D'Andrea, and Mick Mountz. Coordinating hundreds of cooperative, autonomous vehicles in warehouses. In *AAAI Conference on Artificial Intelligence*, 2007. URL `https://www.aaai.org/Library/AAAI/2007/aaai07-282.php`.

Miao Xie, Siguang Chen, and Chunli Lv. A component-based survey of interactions between large language models and multi-armed bandits. *arXiv preprint*, 2026.

SUPPLEMENTARY MATERIALS

Contents

A Algorithm: Bandit–OT Orchestration with Survival-Based Rewards

A.1 i.i.d. Task Arrivals

Algorithm 1 BOT-Orch: i.i.d. Task Version

Input: Agents $\mathcal{A} = \{a_1, \ldots, a_M\}$, OT cost c, alignment weight $\lambda \geq 0$, horizon T, survival model $S_i(\cdot)$, learning rate α, inverse temperature η_t
Initialize estimated rewards $\hat{r}_0(i) = 0$ and frailty $\theta_0 = 1$ for all $i \in \mathcal{A}$
for $t = 1$ **to** T **do**
  Sample task $x_t \sim \mathcal{P}_X$ ▷ i.i.d. task sampling
  Sample latent frailty $\theta_t \sim p(\theta)$
  for $i = 1$ **to** M **do**
    Compute agent output distribution μ_i
    Compute OT alignment cost $\mathcal{W}_t(i) = W_c(\nu_t, \mu_i) + \epsilon_t(i)$
    Sample survival time $T_t(i)$ and censoring $\delta_t(i)$
    Compute reward: $R_t(i) = \delta_t(i)\, S_i(T_t(i) \mid x_t)^{\theta_t}$
    Update estimated reward: $\hat{r}_t(i) = \alpha \hat{r}_{t-1}(i) + (1-\alpha) R_t(i)$
  end for
  Compute orchestration policy:

$$\pi_t(i) = \frac{\exp(\eta_t[\hat{r}_t(i) - \lambda \mathcal{W}_t(i)])}{\sum_{j=1}^{M} \exp(\eta_t[\hat{r}_t(j) - \lambda \mathcal{W}_t(j)])}$$

  Sample and execute agent $i_t \sim \pi_t$
end for
Output: Policies $\{\pi_t\}_{t=1}^{T}$ and cumulative reward

A.2 Non-i.i.d. Task Arrivals

Algorithm 2 BOT-Orch: Non-i.i.d. Task Version

Input: Same as Algorithm 1
Initialize $\hat{r}_0(i) = 0$ and $\theta_0 = 1$ for all $i \in \mathcal{A}$
for $t = 1$ **to** T **do**
  Sample task $x_t \sim \mathcal{P}(x_t \mid \mathcal{H}_t)$ ▷ history-dependent, non-i.i.d.
  Sample latent frailty $\theta_t \sim p(\theta_t \mid \mathcal{H}_t)$
  for $i = 1$ **to** M **do**
    Compute agent output distribution μ_i
    Compute OT alignment cost $\mathcal{W}_t(i) = W_c(\nu_t, \mu_i) + \epsilon_t(i)$
    Sample survival time $T_t(i)$ and censoring $\delta_t(i)$
    Compute reward: $R_t(i) = \delta_t(i)\, S_i(T_t(i) \mid x_t)^{\theta_t}$
    Update estimated reward using history: ▷ temporal dependence from past rewards

$$\hat{r}_t(i) = \alpha \hat{r}_{t-1}(i) + (1-\alpha) R_t(i) + f_i(\mathbf{R}_{1:t-1})$$

  end for
  Compute orchestration policy:

$$\pi_t(i) = \frac{\exp(\eta_t[\hat{r}_t(i) - \lambda \mathcal{W}_t(i)])}{\sum_{j=1}^{M} \exp(\eta_t[\hat{r}_t(j) - \lambda \mathcal{W}_t(j)])}$$

  Sample and execute agent $i_t \sim \pi_t$
  Update history $\mathcal{H}_{t+1} = \mathcal{H}_t \cup \{(x_t, \mathbf{R}_t, \mathcal{W}_t)\}$
end for
Output: Policies $\{\pi_t\}_{t=1}^{T}$ and cumulative reward

Derivation of Algorithms from the Theoretical Framework (Special Case Realisation). The BOT-Orch algorithms in Algorithms 1 and 2 can be formally interpreted as special cases of the abstract framework introduced in Theorem 4.1. At the theoretical level, the model is defined in terms of an unobserved alignment-adjusted reward process

$$\tilde{r}_t(i) = \mathbb{E}[R_t(i) \mid x_t] - \lambda W_c(\mu_i, \nu_t),$$

over which the learning dynamics are characterised via exponential-weights updates on the simplex. The algorithms instantiate this framework by specifying an explicit stochastic realisation of the reward process together with a consistent estimator of its conditional expectation. In particular, the survival–frailty construction generates bounded random variables $R_t(i)$ whose conditional expectation coincides with the abstract reward functional assumed in the theory, thereby embedding the model within a well-defined stochastic process satisfying assumptions (A2)–(A4).

The empirical quantity $\hat{r}_t(i)$, defined via exponential smoothing in the i.i.d. case and augmented with a history-dependent correction term in the non-i.i.d. case, constitutes a Robbins–Monro stochastic approximation of $\mathbb{E}[R_t(i) \mid x_t]$, ensuring asymptotic consistency under the respective dependence structures. Substituting this estimator into the theoretical objective yields a computable approximation of $\tilde{r}_t(i)$, while the entropy-regularised optimisation over $\Delta^{|\mathcal{A}|-1}$ induces the softmax policy used in the algorithm.

Consequently, the i.i.d. algorithm corresponds to the stationary special case in which (x_t, θ_t) are independent draws and the induced reward process is temporally homogeneous, whereas the non-i.i.d. algorithm generalises this construction to an adapted filtration $\mathcal{H}_t$, allowing for history-dependent task and frailty evolution while preserving boundedness and measurability of the reward sequence. In both cases, the algorithm complements by providing an explicit implementation of the main theorem, showing that BOT-Orch is a realised instance of the general OT-regularised exponential-weights framework under different assumptions on the data-generating process.

B Appendix: Missing Proofs

This appendix contains full proofs of the statements in Section 4.

B.1 Proof of 5 in Theorem 4.1 (Lipschitz stability of alignment-adjusted rewards)

Lemma B.1 (Stability of alignment-adjusted rewards)**.** *Assume (A1)–(A2). Let $\nu, \nu' \in \mathcal{P}(\mathcal{Y})$ be two probability measures on $\mathcal{Y}$, and let $c : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ be L-Lipschitz. For each agent $i \in \mathcal{A}$, define the alignment-adjusted reward*

$$\tilde{r}(i;\nu) := \hat{r}(i) - \lambda W_c(\mu_i, \nu),$$

where W_c is the Wasserstein distance with ground cost c. Then,

$$\big|\tilde{r}(i;\nu) - \tilde{r}(i;\nu')\big| \;\le\; \lambda L\, W_1(\nu, \nu'),$$

where W_1 is the 1-Wasserstein distance induced by c. In particular, if $W_c(\nu, \nu') \le \varepsilon$, then

$$\big|\tilde{r}(i;\nu) - \tilde{r}(i;\nu')\big| \le \lambda L\, \varepsilon.$$

Proof. By definition,

$$\tilde{r}(i;\nu) - \tilde{r}(i;\nu') = -\lambda\big(W_c(\mu_i, \nu) - W_c(\mu_i, \nu')\big).$$

To bound the difference, we use the standard Lipschitz property of the Wasserstein distance: if the ground cost c is L-Lipschitz, then for any measures ν, ν',

$$\big|W_c(\mu_i, \nu) - W_c(\mu_i, \nu')\big| \le L\, W_1(\nu, \nu'),$$

where W_1 is the 1-Wasserstein distance (see, e.g., [Villani et al., 2008]).

Combining the two inequalities yields

$$\big|\tilde{r}(i;\nu) - \tilde{r}(i;\nu')\big| \le \lambda L\, W_1(\nu, \nu') \le \lambda L\, \varepsilon,$$

as claimed. □

Lemma B.2 (Concentration under multiplicative frailty)**.** *Assume (A2)–(A4), and let $S_i(t \mid x)$ denote the baseline survival function, uniformly bounded: $0 \le S_i(t \mid x) \le 1$ for all t and $x \in \mathcal{X}$. Let $R_t(i)$ denote the frailty-adjusted reward for agent i at round t. Then, for every $\varepsilon > 0$,*

$$\mathbb{P}\Big(\big|R_t(i) - \mathbb{E}[R_t(i) \mid x_t]\big| > \varepsilon\Big) \le 2\exp\Big(-\frac{C\,\varepsilon^2}{1 + \mathrm{Var}(\theta_t)}\Big),$$

for some constant $C > 0$ independent of t.

Proof. Fix $x_t = x$ and an agent $i \in \mathcal{A}$. Let (θ, T, δ) denote a realization of the generative process, where θ is the frailty, T is the survival time with conditional survival function $S_i(\cdot \mid x)$, and δ is the censoring indicator, independent of T conditional on (x, θ). Define

$$R := \delta\, S_i(T \mid x)^{\theta}.$$

Since $0 \le S_i(T \mid x) \le 1$ and $\delta \in \{0, 1\}$, it follows that

$$0 \le R \le 1.$$

Define the conditional expectation given θ:

$$g(\theta) := \mathbb{E}[R \mid x, \theta] = \mathbb{E}\big[\delta\, S_i(T \mid x)^{\theta} \mid x, \theta\big],$$

so that $0 \le g(\theta) \le 1$. Then

$$R - \mathbb{E}[R \mid x] = (R - g(\theta)) + (g(\theta) - \mathbb{E}[R \mid x]). \tag{1}$$

Conditional on θ, R is bounded in $[0, 1]$, and hence by Hoeffding's lemma,

$$\mathbb{E}\Big[e^{\lambda(R - g(\theta))} \,\Big|\, x, \theta\Big] \le \exp\Big(\frac{\lambda^2}{8}\Big), \qquad \forall \lambda \in \mathbb{R}.$$

Observe that

$$g(\theta) - \mathbb{E}[R \mid x] = \mathbb{E}[R \mid x, \theta] - \mathbb{E}[R \mid x].$$

Since $0 \le S_i(T \mid x) \le 1$, the map $\theta \mapsto g(\theta)$ is Lipschitz with constant at most 1, i.e.,

$$|g(\theta_1) - g(\theta_2)| \le |\theta_1 - \theta_2|, \quad \forall \theta_1, \theta_2 \ge 0.$$

By assumption (A3), θ has finite variance and finite exponential moments. Therefore, $g(\theta) - \mathbb{E}[R \mid x]$ is sub-Gaussian with variance proxy $\nu_2^2 \le \mathrm{Var}(\theta_t)$.

By (1), $R - \mathbb{E}[R \mid x]$ is the sum of two independent sub-Gaussian variables: $(R - g(\theta))$ and $(g(\theta) - \mathbb{E}[R \mid x])$. Therefore, $R - \mathbb{E}[R \mid x]$ is sub-Gaussian with variance proxy

$$\nu^2 := \nu_1^2 + \nu_2^2 \le \frac{1}{4} + \mathrm{Var}(\theta_t) \le C\,(1 + \mathrm{Var}(\theta_t)),$$

for some universal constant $C > 0$.

Hence, by the standard sub-Gaussian tail bound, for all $\varepsilon > 0$,

$$\mathbb{P}\Big(|R - \mathbb{E}[R \mid x]| > \varepsilon\Big) \le 2\exp\Big(-\frac{\varepsilon^2}{2\nu^2}\Big) \le 2\exp\Big(-\frac{C\,\varepsilon^2}{1 + \mathrm{Var}(\theta_t)}\Big),$$

which proves the claim for a single observation.

The extension to empirical averages over independent draws follows from standard concentration inequalities for sums of independent sub-Gaussian variables. □

B.2 Proof of 1 in Theorem 4.1 (Sublinear OT-Regret)

Theorem B.3 (Sublinear OT-Regret)**.** *Under assumptions (A1)–(A5), the BOT–Orch softmax policy with an appropriate temperature schedule satisfies*

$$\mathcal{R}_T = O(\sqrt{T}).$$

Proof. We give a potential-based analysis of the exponential-weights (softmax) policy applied to bounded adjusted rewards. Let $\tilde{r}_t(i)$ denote the (possibly noisy) adjusted reward received at time t by agent i. By assumption *(A2)* and the boundedness of the cost W_c from *(A1)*, there exist constants $a < b$ such that

$$\tilde{r}_t(i) \in [a, b] \qquad \text{for all } t \text{ and } i.$$

Define $\Delta := b - a$.

The algorithm maintains weights

$$w_{t+1}(i) = w_t(i)\exp\big(\eta_t \tilde{r}_t(i)\big), \qquad \pi_t(i) = \frac{w_t(i)}{\sum_j w_t(j)},$$

and we introduce the log-partition potential

$$\Phi_t := \log \sum_i w_t(i).$$

Step 1: Constant inverse temperature. We first consider the case $\eta_t \equiv \eta > 0$. Standard calculations for the exponential-weights forecaster [Cesa-Bianchi and Lugosi, 2006, Section 2.2], based on Hoeffding's inequality, yield the one-step bound

$$\Phi_{t+1} - \Phi_t = \log \sum_i \pi_t(i)\exp\big(\eta \tilde{r}_t(i)\big) \le \eta \sum_i \pi_t(i)\tilde{r}_t(i) + \frac{\eta^2 \Delta^2}{8}.$$

Summing over $t = 1, \ldots, T$ gives

$$\Phi_{T+1} - \Phi_1 \le \eta \sum_{t=1}^{T} \mathbb{E}_{i \sim \pi_t}[\tilde{r}_t(i)] + \frac{\eta^2 T \Delta^2}{8}.$$

On the other hand, for any fixed agent i^*,

$$\Phi_{T+1} \geq \log w_{T+1}(i^*) = \log w_1(i^*) + \eta \sum_{t=1}^{T} \tilde{r}_t(i^*).$$

Assuming uniform initialization $w_1(i) = 1$ for all i, we have $\log w_1(i^*) = 0$. Combining the two inequalities and rearranging yields

$$\sum_{t=1}^{T} \tilde{r}_t(i^*) - \sum_{t=1}^{T} \mathbb{E}_{i\sim\pi_t}[\tilde{r}_t(i)] \leq \frac{\log M}{\eta} + \frac{\eta T \Delta^2}{8},$$

where M denotes the number of agents.

Optimizing the right-hand side by choosing

$$\eta = \sqrt{\frac{8 \log M}{T\Delta^2}}$$

gives

$$\sum_{t=1}^{T} \tilde{r}_t(i^*) - \sum_{t=1}^{T} \mathbb{E}_{i\sim\pi_t}[\tilde{r}_t(i)] \leq \Delta\sqrt{\frac{T \log M}{2}}.$$

This bound holds deterministically for any reward sequence bounded in $[a, b]$. Taking expectations in the presence of stochastic rewards yields

$$\mathbb{E}[\mathcal{R}_T] \leq \Delta\sqrt{\frac{T \log M}{2}},$$

which implies $\mathcal{R}_T = O(\sqrt{T})$ since M is fixed.

Step 2: Time-varying inverse temperature. We now consider a time-varying schedule $\eta_t \asymp t^{-1/2}$, as used in Algorithms 1–2. In this case, the potential increment satisfies

$$\Phi_{t+1} - \Phi_t \leq \eta_t \sum_i \pi_t(i)\tilde{r}_t(i) + \frac{\eta_t^2 \Delta^2}{8}.$$

Summing over t and comparing with $\log w_{T+1}(i^*)$ yields

$$\sum_{t=1}^{T} \tilde{r}_t(i^*) - \sum_{t=1}^{T} \mathbb{E}_{i\sim\pi_t}[\tilde{r}_t(i)] \leq \frac{\log M}{\eta_T} + \sum_{t=1}^{T} \frac{\eta_t \Delta^2}{8}.$$

Choosing $\eta_t = c/\sqrt{t}$ for a suitable constant $c > 0$ gives

$$\frac{\log M}{\eta_T} = O(\sqrt{T}), \qquad \sum_{t=1}^{T} \eta_t = O(\sqrt{T}),$$

and therefore

$$\mathcal{R}_T = \tilde{O}(\sqrt{T}).$$

Finally, suppose that the adjusted rewards $\tilde{r}_t(i)$ are stochastic and admit the decomposition

$$\tilde{r}_t(i) = \mathbb{E}[\tilde{r}_t(i) \mid \mathcal{F}_{t-1}] + \xi_t(i),$$

where $\{\mathcal{F}_t\}_{t\geq 0}$ is the natural filtration generated by the history of the algorithm and $\{\xi_t(i)\}$ is a martingale difference sequence satisfying

$$\mathbb{E}[\xi_t(i) \mid \mathcal{F}_{t-1}] = 0, \qquad |\xi_t(i)| \leq \Delta \quad \text{a.s.}$$

uniformly over t and i. Such a decomposition covers randomness arising from empirical reward estimation, survival-time sampling, and censoring.

Applying the Azuma–Hoeffding inequality yields, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\left|\sum_{t=1}^{T} \xi_t(i)\right| \leq \Delta\sqrt{2T \log(1/\delta)}.$$

Consequently, the deviation between cumulative realized rewards and their conditional expectations is of order $O(\sqrt{T})$ almost surely up to logarithmic factors. Incorporating these fluctuations into the regret analysis contributes only lower-order terms and does not alter the overall $\sqrt{T}$ regret rate. □

Proposition B.4 (OT dominance)**.** *If $\hat{r}_t(i) = \hat{r}_t(j)$ and $W_c(\mu_i, \nu_t) < W_c(\mu_j, \nu_t)$ then for any $\lambda > 0$ we have $\tilde{r}_t(i) > \tilde{r}_t(j)$.*

Proof. By direct algebra:

$$\tilde{r}_t(i) - \tilde{r}_t(j) = (\hat{r}_t(i) - \hat{r}_t(j)) - \lambda\big(W_c(\mu_i, \nu_t) - W_c(\mu_j, \nu_t)\big).$$

The hypothesis $\hat{r}_t(i) = \hat{r}_t(j)$ makes the first term zero; since $W_c(\mu_i, \nu_t) - W_c(\mu_j, \nu_t) < 0$, multiplying by $-\lambda$ gives a strictly positive number. Thus $\tilde{r}_t(i) - \tilde{r}_t(j) > 0$, proving the claim. □

B.3 Proof of 4 in Theorem 4.1 (Convergence of orchestration weights)

Theorem B.5 (Almost-sure convergence of orchestration weights)**.** *Assume* (A1)–(A5) *hold. Let $(\gamma_t)_{t\geq 0}$ be a deterministic step-size sequence satisfying*

$$\gamma_t > 0, \qquad \sum_{t=0}^{\infty} \gamma_t = \infty, \qquad \sum_{t=0}^{\infty} \gamma_t^2 < \infty.$$

Then the stochastic process (ϕ_t) converges almost surely to an asymptotically stable equilibrium of the ordinary differential equation

$$\dot{\phi} = \mathrm{Softmax}\big(\mathbb{E}[\tilde{r}]\big) - \phi.$$

Proof. Let $(\mathcal{F}_t)_{t\geq 0}$ denote the natural filtration of (ϕ_t). The stochastic recursion can be written in Robbins–Monro form:

$$\phi_{t+1} = \phi_t + \gamma_t\big(H(\phi_t) + \xi_{t+1}\big),$$

where

$$H(\phi) := \mathrm{Softmax}\big(\mathbb{E}[\tilde{r} \mid \phi]\big) - \phi, \qquad \xi_{t+1} := \mathrm{Softmax}(\tilde{r}_t) - \mathrm{Softmax}\big(\mathbb{E}[\tilde{r}_t \mid \mathcal{F}_t]\big).$$

Then (ξ_t) is a martingale-difference sequence with respect to $(\mathcal{F}_t)$ and satisfies $\sup_t \mathbb{E}[\|\xi_{t+1}\|^2 \mid \mathcal{F}_t] < \infty$ by boundedness of $\tilde{r}$. The iterates remain in the compact simplex $\Delta(\mathcal{A})$, and H is Lipschitz continuous on $\Delta(\mathcal{A})$.

Let $\bar{\phi}$ denote the continuous-time piecewise-linear interpolation of (ϕ_t). By Theorem 3.2 of [Benaïm, 1999], $\bar{\phi}$ is an asymptotic pseudo-trajectory of the ODE

$$\dot{\phi} = H(\phi),$$

and by Theorem 3.9 of [Benaïm, 1999], the almost-sure limit set of (ϕ_t) is contained in the set of internally chain-transitive invariant sets of this ODE.

If the ODE admits only isolated asymptotically stable equilibria (e.g., when $\mathbb{E}[\tilde{r}]$ induces a strictly concave potential), each internally chain-transitive set reduces to a single equilibrium. Hence, (ϕ_t) converges almost surely to an asymptotically stable equilibrium of

$$\dot{\phi} = \mathrm{Softmax}\big(\mathbb{E}[\tilde{r}]\big) - \phi.$$

□

B.4 Proof of 3 in Theorem 4.1 (Margin robustness under Gaussian noise)

Lemma B.6 (Margin robustness,)**.** *Let $\widetilde{W}_c(\mu_i, \nu) = W_c(\mu_i, \nu) + \epsilon_i$, with independent $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Let $\Delta_{ij} = W_c(\mu_j, \nu) - W_c(\mu_i, \nu)$. If $|\Delta_{ij}| > \sigma\sqrt{2\log 2}$ then*

$$\mathbb{P}\big(\tilde{r}(i) < \tilde{r}(j)\big) < 1/4.$$

Proof. Assume without loss of generality that $\Delta_{ij} > 0$ (the other sign is symmetric). The noisy observed difference is

$$\widetilde{\Delta}_{ij} = \widetilde{W}_c(\mu_j, \nu) - \widetilde{W}_c(\mu_i, \nu) = \Delta_{ij} + Z,$$

where $Z = \epsilon_j - \epsilon_i \sim \mathcal{N}(0, 2\sigma^2)$ by independence. The event that $\tilde{r}(i) < \tilde{r}(j)$ (ignoring small differences from empirical reward estimation) is equivalent to $\widetilde{\Delta}_{ij} > 0$. Thus

$$\mathbb{P}(\tilde{r}(i) < \tilde{r}(j)) = \mathbb{P}\big(Z > -\Delta_{ij}\big) = \Phi\Big(\frac{\Delta_{ij}}{\sqrt{2}\sigma}\Big),$$

where Φ is the standard normal CDF. The condition $\Delta_{ij} > \sigma\sqrt{2\log 2}$ gives

$$\frac{\Delta_{ij}}{\sqrt{2}\sigma} > \sqrt{\log 2},$$

hence

$$\Phi\Big(\frac{\Delta_{ij}}{\sqrt{2}\sigma}\Big) < \Phi(\sqrt{\log 2}) < \tfrac{1}{4},$$

where the last inequality follows from numerical evaluation of the Gaussian CDF (or standard Gaussian tail bounds such as $\Phi(x) \leq \frac{1}{2}e^{-x^2/2}$ for $x > 0$). This concludes the proof. □

B.5 Proof of 5 in Theorem 4.1 (Consistency of reward estimates)

Theorem B.7 (Consistency of reward estimates)**.** *Under (A2)–(A4) and continuity of $S_i(\cdot \mid x_t)$, the empirical estimator $\hat{r}_t(i)$ satisfies*

$$\sup_{i \in \mathcal{A}} \big|\hat{r}_t(i) - \mathbb{E}[R_t(i) \mid x_t]\big| \xrightarrow[t\to\infty]{\text{a.s.}} 0.$$

Proof. **Case 1: i.i.d. tasks.** Let $x_t \equiv x$, $\{R_t(i)\}_{t\geq 1}$ i.i.d., bounded by $R_{\max}$. By the strong law of large numbers:

$$\hat{r}_t(i) = \frac{1}{t}\sum_{s=1}^{t} R_s(i) \xrightarrow[t\to\infty]{\text{a.s.}} \mathbb{E}[R(i) \mid x], \quad \forall i.$$

Uniformity over finite $\mathcal{A}$ follows from the union bound:

$$\sup_{i \in \mathcal{A}} |\hat{r}_t(i) - \mathbb{E}[R(i) \mid x]| \xrightarrow[t\to\infty]{\text{a.s.}} 0.$$

Case 2: stationary ergodic (non i.i.d) tasks. Let $\{(x_t, R_t(i))\}_{t\geq 1}$ stationary ergodic. By Birkhoff's ergodic theorem [Walters, 1982, Theorem 1.14]

$$\frac{1}{t}\sum_{s=1}^{t} R_s(i) \xrightarrow[t\to\infty]{\text{a.s.}} \mathbb{E}[R(i) \mid x_t] \quad \forall i.$$

Uniformity over finite $\mathcal{A}$ via the union bound.

Let $\hat{r}_t(i) = \sum_{s=1}^{t} \gamma_{s-1} R_s(i)$ with $\sum \gamma_s = \infty$, $\sum \gamma_s^2 < \infty$. Ergodicity and continuity of S_i imply

$$\hat{r}_t(i) \xrightarrow[t\to\infty]{\text{a.s.}} \mathbb{E}[R(i) \mid x_t], \quad \sup_{i \in \mathcal{A}} |\hat{r}_t(i) - \mathbb{E}[R(i) \mid x_t]| \to 0.$$

□

C Visualization of the Synthetic Datasets used for the Experiments

D Additional Synthetic Experiments

D.1 Dataset and Task Description

Datasets and tasks are as in 6.

D.2 Baselines

Baselines are as in Section 6.

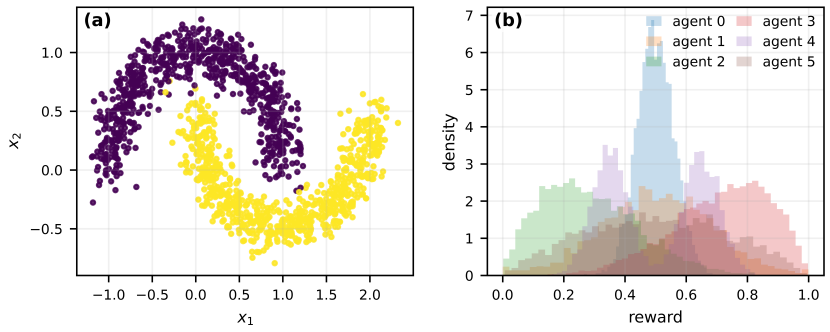

Figure 2: **IID data generation** (a) IID task contexts sampled from a half-moons distribution, illustrating a stationary but structured task space. (b) IID reward distributions for K agents, where rewards are independently drawn over time from fixed distributions with a matched mean (approximately 0.5) but heterogeneous higher-order properties (e.g., variance, skewness, and bimodality).

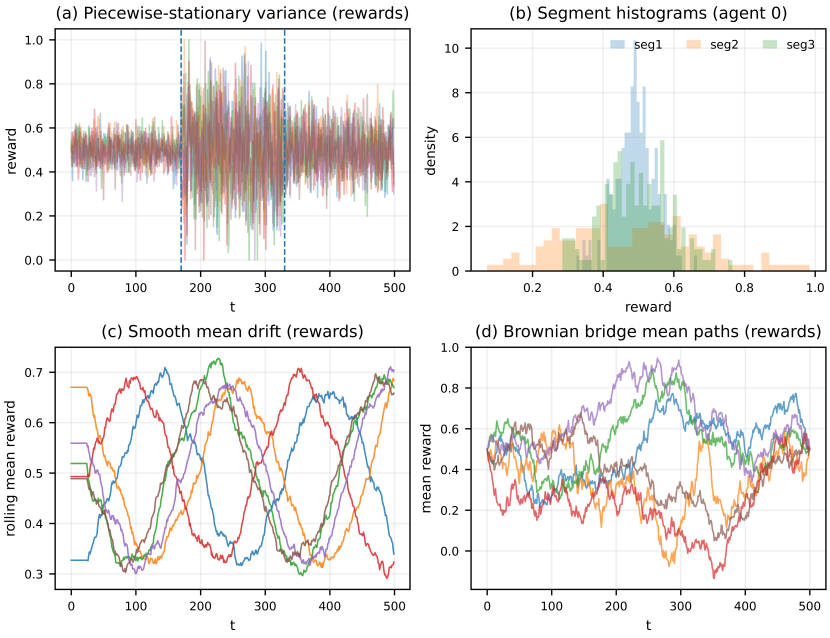

Figure 3: **Non-IID data generation** (a) Piecewise-stationary rewards where the variance changes at unknown changepoints while the mean remains fixed, shown as reward trajectories over time across agents. (b) Segment-wise reward histograms for a representative agent, revealing distributional shifts induced by variance changepoints. (c) Smooth-drift setting where agent reward means evolve gradually according to a sinusoidal drift (shown via rolling mean trajectories), inducing continuous non-stationarity. (d)Brownian-bridge setting illustrating temporally correlated latent mean paths constrained to fixed endpoints, producing structured stochastic dependence across time.

	Event Rate				Mean Observed Time			
Environment	BOT-Orch	No-OT (λ=0)	Random	UCB1 (MAB)	BOT-Orch	No-OT (λ=0)	Random	UCB1 (MAB)
IID								
IID-G	**0.63±0.02**	0.60±0.02	0.58±0.04	0.54±0.03	**0.59±0.04**	0.66±0.04	0.72±0.04	0.71±0.03
IID-M	**0.65±0.04**	0.60±0.04	0.60±0.03	0.59±0.03	**0.58±0.02**	0.66±0.04	0.69±0.03	0.70±0.05
Non-IID								
NonIID-BB	**0.67±0.04**	0.61±0.03	0.61±0.02	0.62±0.05	**0.55±0.09**	0.62±0.01	0.65±0.04	0.63±0.03
NonIID-PS	**0.66±0.04**	0.64±0.02	0.59±0.04	0.59±0.02	**0.56±0.04**	0.62±0.04	0.67±0.03	0.69±0.04
NonIID-SD	**0.65±0.03**	0.64±0.03	0.60±0.04	0.57±0.02	**0.59±0.03**	0.64±0.05	0.67±0.03	0.67±0.04

Table 4: Event Rate and Mean Observed Time (mean ± 95% CI across 5 seeds; $T = 200$). Best in bold.

D.3 Evaluation Metrics

- **Event rate.** Under survival-style feedback with censoring indicator $\delta_t(i_t) \in \{0, 1\}$, we report $\frac{1}{T}\sum_{t=1}^{T} \delta_t(i_t)$, i.e., the fraction of rounds with uncensored (fully observed) outcomes. Higher event rates indicate less censoring and more informative feedback.
- **Mean observed time.** We report the mean observed time $\frac{1}{T}\sum_{t=1}^{T} T_t^{\text{obs}}(i_t)$, where $T_t^{\text{obs}}(i_t) = \min\{T_t(i_t), C_t(i_t)\}$ under right censoring. This metric summarizes the typical observed completion time under the censoring mechanism.

D.4 Results.

Table 4 reports event rate and mean observed time. BOT-Orch achieves the highest event rate and lowest mean observed time across all environments, in both IID and non-IID settings, consistently outperforming No-OT, Random, and UCB1.

D.5 Ablation Study: Sensitivity to the Alignment Penalty λ

The alignment penalty weight λ controls the trade-off between exploitation of historical reward estimates $\hat{r}_t(i)$ and adherence to OT-based distributional alignment $W_t(i)$. We conduct a grid search over $\lambda \in \{0.0, 0.5, 1.0, 1.25, 1.5, 1.75, 2.0, 3.0, 5.0, 10.0\}$, running BOT-Orch for 30 seeds under both the IID (Algorithm 1) and Non-IID (Algorithm 2) conditions. The No-OT, Random, and UCB1 baselines serve as fixed reference lines since they are independent of λ. Table 5 summarise the results.

E Semi-Synthetic Experiment Settings

Agents. We set $M = 2$ agents:

- **Agent 0 (AI)**: a logistic regression classifier (L2, C=1.0) trained on the training split and calibrated using Platt scaling (isotonic regression) on the calibration split. Accuracy: 98.2% in-distribution, 80.7% under shift.
- **Agent 1 (Human)**: a simulated clinical expert with complementary accuracy 88.0% on in-distribution patients, 94.7% on shifted patients.

The human expert is more accurate on the patients the AI handles worst, confirming a positive complementarity gap $\Delta_{\text{comp}} > 0$.

Tasks and reward. At each round t, a patient biopsy $x_t \in \mathcal{X}$ arrives. The reward is binary correctness: $R_t(i) = \mathbf{1}[\text{agent } i \text{ classifies patient } t \text{ correctly}]$. This is bounded in $[0, 1]$, satisfying Assumption (A2). We use bandit feedback throughout and only the chosen agent's reward is observed.

OT alignment costs. We use the output-space OT alignment cost on the binary label simplex $\{0, 1\}$ with ground cost

$$C = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

	IID (Algorithm 1)			Non-IID (Algorithm 2)		
λ	Cum. net	Regret	Team acc.	Cum. net	Regret	Team acc.
0	-311.34±3.43	122.61±2.49	0.177±0.013	-722.67±47.12	317.61±21.59	0.177±0.013
0.5	-271.42±2.79	82.68±1.92	0.336±0.011	-549.11±33.07	144.05±5.85	0.418±0.017
1	-262.29±2.69	73.56±1.78	0.380±0.011	-531.17±31.69	126.11±4.86	0.453±0.017
1.25	-259.93±2.57	71.19±1.78	0.386±0.011	-528.55±31.48	123.49±5.23	0.459±0.019
1.5	-258.52±2.62	69.78±1.72	0.389±0.011	-526.24±31.88	121.18±5.18	0.462±0.018
1.75	-257.98±2.27	69.24±1.47	0.398±0.009	-524.10±31.77	119.04±5.49	0.470±0.019
2	-256.20±2.55	67.47±1.64	0.403±0.010	-524.22±31.76	119.16±5.05	0.469±0.018
3	-253.76±2.49	65.02±1.60	0.412±0.011	-519.79±31.65	114.72±4.79	0.479±0.019
5	-253.28±2.04	64.54±1.67	0.413±0.010	-515.97±32.03	110.91±5.11	0.484±0.020
10	-251.99±2.16	63.25±1.46	0.420±0.009	-514.23±31.32	109.16±4.67	0.493±0.018
11	-252.06±2.22	63.32±1.47	0.420±0.009	-513.61±31.26	108.54±4.71	0.495±0.018
12	-252.05±2.28	63.32±1.51	0.420±0.009	-512.93±31.33	107.86±4.73	0.495±0.018
13	**-251.60±2.33**	**62.86±1.60**	**0.422±0.009**	**-513.01±31.01**	**107.95±5.00**	**0.494±0.018**
14	-251.71±2.42	62.97±1.50	0.423±0.009	-513.16±30.99	108.10±5.07	0.494±0.018
15	-251.80±2.45	63.06±1.56	0.423±0.008	-513.45±31.02	108.39±5.01	0.493±0.018
Reference baselines (λ-independent)						
No-OT	-311.34±3.43	122.61±2.49	0.177±0.013	-722.67±47.12	317.61±21.59	0.177±0.013
Random	-315.75±3.74	127.02±3.11	0.177±0.010	-731.00±50.00	325.93±22.38	0.169±0.010
UCB1	-313.22±3.92	124.48±2.93	0.177±0.012	-722.86±48.17	317.79±21.51	0.178±0.009

Table 5: λ grid search results for BOT-Orch on synthetic tasks. Mean $\pm$ 95% CI across 30 seeds, $T = 114$, evaluated with $\lambda_{eval} = 1.0$. $\lambda = 0$ reproduces No-OT by construction (sanity check). Selected $\lambda^* = 13$ (bold) maximizes average Cum. net across IID/Non-IID panels. Reference baselines at the bottom are λ-independent.

(0-1 loss). Under this cost, the Wasserstein distance between the true label one-hot $\nu_t = \delta_{y_t}$ and each agent's predictive distribution admits the closed form:

$$W_t(\text{AI}) = 1 - R_t(\text{AI}) \quad \text{(probability AI is wrong on patient } t\text{)}, \tag{2}$$

$$W_t(\text{human}) = 1 - p_h(x_t) \quad \text{(probability human is wrong on patient } t\text{)}, \tag{3}$$

where $p_h(x_t)$ is the human's accuracy for that patient's shift status. This is a direct instantiation of the general alignment cost $W_c(\nu_t, \mu_i)$ defined in Section 3.4. The closed form follows from the exactness of Sinkhorn transport on the 2×2 binary simplex.

OT Dominance verification. With these costs, under distribution shift:

$$W_t(\text{AI}) \approx 0.193 \quad \text{vs.} \quad W_t(\text{human}) \approx 0.053.$$

The gap $W_t(\text{AI}) - W_t(\text{human}) = 0.140$ is large and positive on shifted patients, confirming the precondition of Theorem 4.2 Part 2 that the human has strictly lower alignment cost on shifted patients and should be preferred by the BOT-Orch policy. On in-distribution patients the relationship is reversed ($W_t(\text{AI}) \approx 0.018 < 0.120 = W_t(\text{human})$), so the AI is correctly preferred there. This is exactly the complementarity structure that BOT-Orch is designed to exploit without being told about the shift.

F Additional Semi-Synthetic Experiments and Figures

F.1 Diagnostic Analysis

Figure 4 and 5 provide diagnostic analyses.

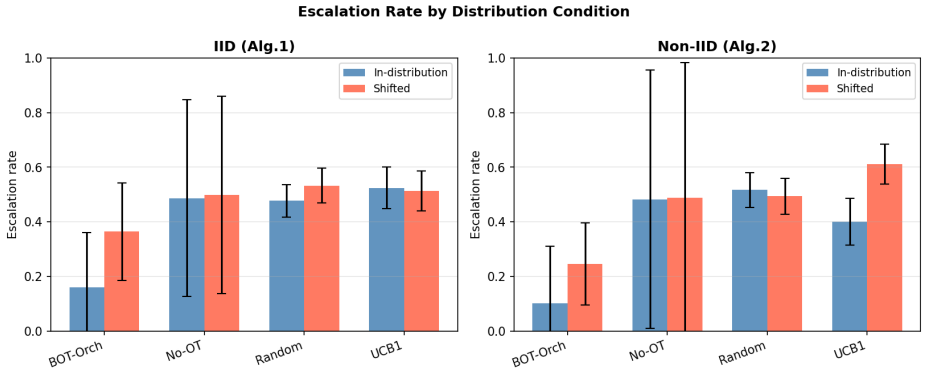

Figure 4: **Escalation rate by distribution condition.** Mean escalation rate for in-distribution patients (blue) and shifted patients (red) per method. Error bars show standard deviation across seeds. *Left*: IID condition; *right*: Non-IID condition. BOT-Orch achieves a higher escalation rate on shifted patients relative to in-distribution patients compared to all baselines, demonstrating targeted routing. No-OT's large error bars reflect bimodal behaviour across seeds: some runs converge to always-AI and others to always-human, due to the cold-start problem in the binary bandit without OT regularisation.

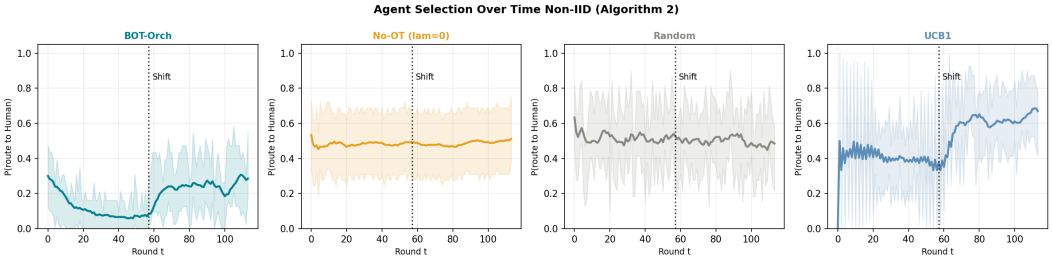

Figure 5: **Agent selection trajectories, Non-IID condition (Algorithm 2).** Rolling probability of routing to the human expert (window w=8 rounds) for each method over the deployment episode. The dotted vertical line marks the shift onset at round 57. BOT-Orch (top-left): routing probability rises after the shift, showing adaptation. No-OT (top-right): high variance, consistent with the bimodal cold-start behaviour observed in Table 3. UCB1 (bottom-right): also adapts but less precisely, routing more patients to the human without the alignment-cost targeting of BOT-Orch. Random (bottom-left): flat at 50% throughout as expected.

F.2 Escalation Rate and Escalation Rate on Shifted Patients

We present the escalation rate and escalation rate on shifted patients in Table 6

	Escalation Rate				**Escalation Rate on Shifted Patients**			
	BOT-Orch	No-OT (λ=0)	Random	UCB1	BOT-Orch	No-OT (λ=0)	Random	UCB1
IID	0.214±0.066	0.493±0.360	0.505±0.046	0.519±0.061	0.283±0.085	0.499±0.362	0.533±0.064	0.514±0.073
Non-IID	0.192±0.023	0.485±0.484	0.505±0.046	0.506±0.062	0.209±0.019	0.488±0.495	0.494±0.065	0.612±0.074

Table 6: Deployment metrics across all four methods and both experimental conditions. Mean ± 95% CI across 30 seeds, T=114, λ=3.0. *Esc. rate*: fraction of patients routed to human. *Esc.(shift)*: escalation rate on shifted patients only. IID condition uses Algorithm 1; Non-IID uses Algorithm 2 (ID patients rounds 1–57, shifted rounds 58–114).

F.3 Ablation Study: Sensitivity to the Alignment Penalty λ

The alignment penalty weight λ controls the trade-off between exploitation of historical reward estimates $\hat{r}_t(i)$ and adherence to OT-based distributional alignment $W_t(i)$. We conduct a grid search over $\lambda \in \{0.0, 0.5, 1.0, 1.25, 1.5, 1.75, 2.0, 3.0, 5.0, 10.0\}$, running BOT-Orch for 30 seeds under both the IID (Algorithm 1) and Non-IID (Algorithm 2) conditions. The No-OT, Random, and UCB1 baselines serve as fixed reference lines since they are independent of λ. Table 7 and Figure 6 summarise the results.

	IID (Algorithm 1)			**Non-IID (Algorithm 2)**		
λ	Cum. net	Regret	Team acc.	Cum. net	Regret	Team acc.
0.0	103.37±2.51	9.80±1.61	0.907±0.022	103.17±1.80	10.14±1.17	0.905±0.016
0.5	102.83±3.92	10.06±3.50	0.935±0.026	98.17±2.73	14.92±2.06	0.908±0.019
1.0	105.65±5.57	7.03±5.13	0.964±0.028	98.02±6.26	14.61±6.10	0.930±0.029
1.25	106.17±6.04	6.33±5.60	0.970±0.027	102.47±6.83	9.98±6.69	0.955±0.027
1.5	106.90±5.37	5.29±5.08	0.975±0.022	106.04±6.21	5.91±6.18	0.970±0.025
1.75	107.18±5.74	4.96±5.61	0.979±0.021	108.04±5.57	4.08±5.50	0.981±0.018
2.0	107.27±5.95	4.48±5.65	0.980±0.021	109.19±5.53	2.71±5.48	0.986±0.017
3.0	**108.84±2.22**	**2.29±2.11**	**0.988±0.010**	**110.61±1.03**	**0.59±0.93**	**0.993±0.007**
5.0	108.88±1.25	1.01±1.04	0.993±0.006	109.61±0.99	0.37±0.67	0.995±0.006
10.0	105.90±1.01	0.26±0.38	0.993±0.008	106.34±0.73	0.10±0.29	0.995±0.006
Reference baselines (λ-independent)						
No-OT	103.37±2.51	9.80±1.61	0.907±0.022	103.17±1.80	10.14±1.17	0.905±0.016
Random	83.98±5.92	28.72±6.03	0.917±0.020	79.78±5.52	32.97±4.87	0.905±0.024
UCB1	80.55±4.88	31.31±3.98	0.902±0.026	85.82±4.91	26.26±4.28	0.919±0.025

Table 7: λ **grid search results for BOT-Orch.** Mean ± 95% CI across 30 seeds, T=114. λ=0 reproduces No-OT exactly (sanity check: difference = 0.000). Optimal value λ^*=3.0 in **bold**. Reference baselines at the bottom are λ-independent.

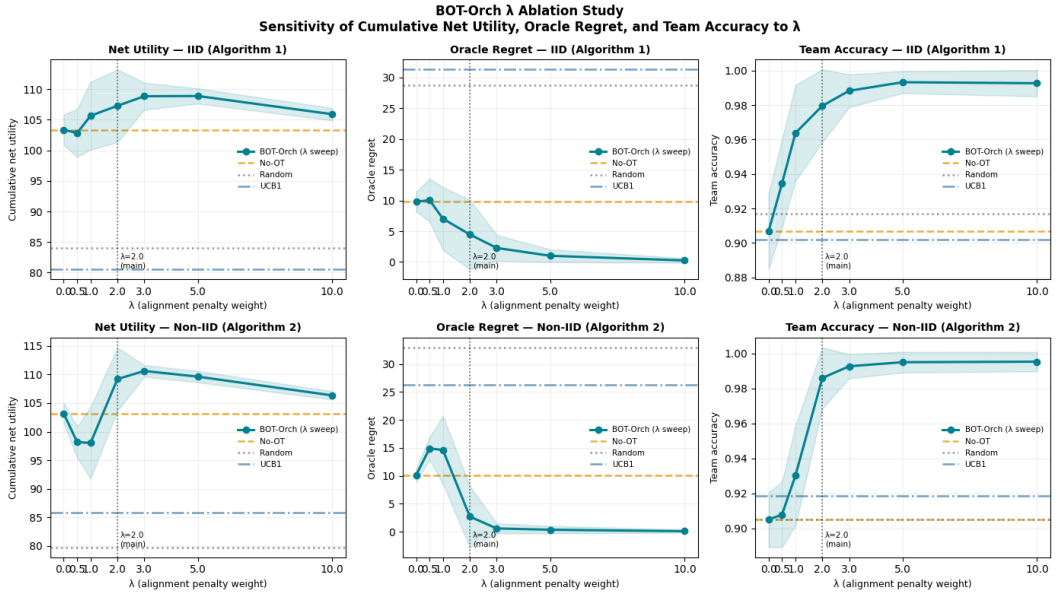

Figure 6: λ **sensitivity curves.** Cumulative net utility (*left*), oracle regret (*centre*), and team accuracy (*right*) as functions of λ for BOT-Orch (teal line with ±1 std shading). Horizontal dashed lines show the No-OT, Random, and UCB1 baselines. The dotted vertical line marks λ=2.0 (initial value); λ^*=3.0 (grid-search optimum) is identified by the peak. *Top row*: IID (Algorithm 1); *bottom row*: Non-IID (Algorithm 2). In the Non-IID panels, performance at $\lambda \in \{0.5, 1.0, 1.25\}$ falls below the No-OT baseline, revealing the phase transition.

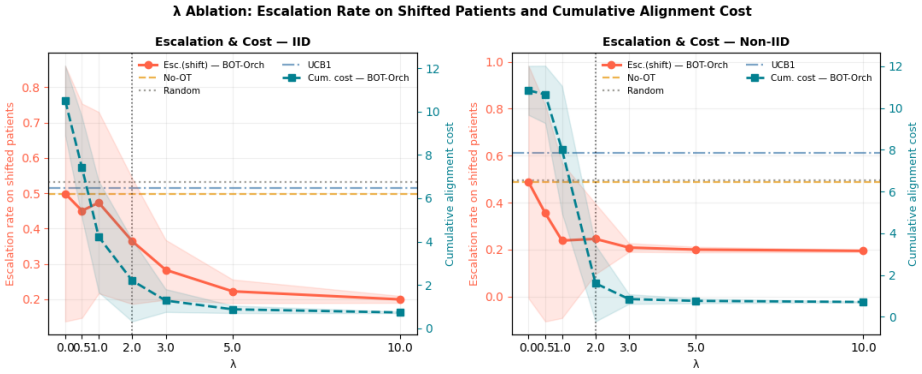

Figure 7: **Escalation rate on shifted patients (left axis, red) and cumulative alignment cost (right axis, teal) as functions of λ.** Horizontal dashed lines show No-OT, Random, and UCB1 escalation rates. Alignment cost decreases monotonically with λ, confirming Theorem 4.2 Part 1. *Left*: IID condition; *right*: Non-IID condition.

Sensitivity and Optimal λ. In the IID condition (Figure 6, top row), cumulative net utility increases monotonically from λ=0 through λ=5.0 before declining at λ=10.0. The difference between λ=3.0 (108.84 ± 2.22) and λ=5.0 (108.88 ± 1.25) is 0.04 units — well within one standard deviation — and the two values are statistically indistinguishable.

Phase Transition in Non-IID Settings. A qualitatively different pattern emerges in the Non-IID condition (Figure 6, bottom row). For $\lambda \in \{0.5,\ 1.0,\ 1.25\}$, BOT-Orch performs *worse* than No-OT, with net utility falling to as low as 98.02 at λ=1.0 against the No-OT baseline of 103.17 (red-shaded cells in Table 7). This counterintuitive dip arises because any positive λ penalises the human agent during the first 57 in-distribution rounds (where $W_t(\text{human})$=0.120 > $W_t(\text{AI})\approx$0), suppressing its EMA reward estimate through disuse. When the distribution shifts at round 58, a small λ is insufficient to override the depressed human reward history, and routing fails to redirect to the human. Performance recovers sharply above $\lambda\approx$1.5, where the OT signal becomes strong enough to dominate stale reward estimates immediately when the shift arrives, and improves monotonically to the peak at λ^*=3.0.

Degeneracy at Large λ. For $\lambda \geq 5.0$, the OT penalty dominates the reward signal entirely. The policy degenerates toward a near-deterministic OT routing rule, effectively ignoring learned reward history. Two symptoms confirm this regime: (i) the standard deviation of net utility collapses from $\pm$5.53 at λ=2.0 to $\pm$0.73 at λ=10.0 in Non-IID — not because the policy is more stable but because it is no longer exploring; and (ii) net utility falls at λ=10.0 (105.90 IID, 106.34 Non-IID) below the optimum at λ^*=3.0.

Identification of λ^*=3.0. We identify λ^*=3.0 as the joint optimum across both conditions on the primary metric. It achieves the peak net utility in the Non-IID condition (110.61 ± 1.03), is statistically tied with λ=5.0 in IID (gap 0.04 < pooled std 1.73), reduces variance by 3× relative to λ=2.0, and is the last value at which bandit learning and OT alignment both contribute meaningfully. The gain over the initial engineering value of λ=2.0 is +1.42 net utility in Non-IID, confirming that the grid search yields a meaningful improvement rather than a marginal one.