Learning to Orchestrate Agents under Uncertainty

Mary Chriselda Antony Oliver [1] **Lan Jiang** [2] **Aaron Bundi Anampiu** [3] **Elaf Almahmoud** [2] **Francesco Quinzan** [4] **Umang Bhatt** [2]

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 an agent, accounting for reliability, cost, and uncertainty. We propose BOT-Orch, a lightweight framework that recasts 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.

[†]Francesco Quinzan and Umang Bhatt contributed equally as senior authors.

[1]Department of Applied Mathematics and Theoretical Physics, University of Cambridge [2]Centre for Human-Inspired Artificial Intelligence, University of Cambridge [3]African Institute for Mathematical Sciences, South Africa [4]Department of Engineering Science, University of Oxford. Correspondence to: Mary Chriselda Antony Oliver <mca52@cam.ac.uk>.

Accepted at the ICML 2026 Workshop on Decision-Making from Offline Datasets to Online Adaptation: Black-Box Optimization to Reinforcement Learning.

1. Introduction

Incorporating uncertainty in delegated decision-making process 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 frameworks (e.g., Dec-POMDPs, stochastic games, DCOPs) explicitly model uncertainty in 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 & Amato, 2016; Hansen et al., 2004; Heifetz et al., 2006; Shoham & 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 (including single-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 MARL methods implicitly assume similar reliability and reaction costs, and rarely address adaptive selection among heterogeneous agents of uncertain outputs.

Modern AI systems and single-agent orchestration frameworks mitigate 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 & Szepesvári, 2020; Bubeck & Cesa-Bianchi, 2012). In the special case, single-agent extensions of the MAB framework 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 & 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 & Zemel, 2020). However, OT remains underexplored in single-agent delegation 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 & Kochenderfer, 2024).

Our contribution.

- **Bandit-based orchestration with OT alignment:** We cast delegation over a single agent 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 (2-3 of Theorem 4.1), and convergence and consistency properties (4–5 of Theorem 4.1).
- **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 & 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 & Szepesvári, 2020; Chen, 2024; Tong, 2024; Letard et al., 2024). Recent surveys summarize advances in classical and contextual bandits and their applications (Chen, 2024; Letard et al., 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 with applications to health data science (Oliver et al., 2025a;c;b). This provides uncertainty-aware weights capturing the trade-off between confidence and disagreement. The framework is well-suited for distributional and uncertainty-aware reinforcement learning settings such as those in (Osband et al., 2013; Bellemare et al., 2017), offering a unified approach to adaptive multi-agent coordination with both theoretical guarantees and strong empirical performance.

3. Preliminaries and Notation

In this section, we introduce the relevant concepts that are foundational to our framework.

3.1. Agents, Tasks, and Task Space

Let $\mathcal{A} := \{a_1, \ldots, a_M\}$ denote a finite set of M agents. Time evolves over a finite horizon $t = 1, \ldots, T$. At each round t, a task $x_t \in \mathcal{X}$ is observed. We assume that tasks are generated by a history-dependent stochastic process $x_t \sim \mathcal{P}(\cdot \mid \mathcal{H}_t)$, where the history filtration is defined as

$$\mathcal{H}_t := \sigma\big(x_s, A_s, \mathbf{R}_s, \mathbf{W}_s : s = 1, \ldots, t-1\big),$$

and $A_s \in \{1, \ldots, M\}$ is $\mathcal{H}_s$-measurable and denotes the agent selected at time s. For each agent $a_i \in \mathcal{A}$ and task

x_t, let $R_t^i \in \mathbb{R}$ denote the (possibly counterfactual) reward that would be obtained if x_t were assigned to agent a_i, and let $W_t^i \in \mathbb{R}$ denote the corresponding task–agent alignment cost. Define vectors $\mathbf{R}_t := (R_t^1, \ldots, R_t^M)^\top \in \mathbb{R}^M$, $\mathbf{W}_t := (W_t^1, \ldots, W_t^M)^\top \in \mathbb{R}^M$. Analogously, we define the corresponding conditional mean quantities: $r^i(x_t) := \mathbb{E}[R_t^i \mid x_t]$, $w^i(x_t) := \mathbb{E}[W_t^i \mid x_t]$. Note that task arrivals may depend on past observations, environmental conditions, or previous assignments.

3.2. Correlated Multi-Agent Rewards

Let $\mathbf{R}_t \in \mathbb{R}^M$ denote the vector of (joint) rewards at round t. We model its conditional moments given the current task x_t and history $\mathcal{H}_t$. Assume there exists a function $\mathbf{r}$ and a history-dependent drift term $\mathbf{f} : \mathcal{H}_t \to \mathbb{R}^M$ such that

$$\mathbb{E}[\mathbf{R}_t \mid x_t, \mathcal{H}_t] = \mathbf{r}(x_t) + \mathbf{f}(\mathcal{H}_t),$$

and a positive semidefinite matrix Σ_t satisfying

$$\mathrm{Cov}[\mathbf{R}_t \mid x_t, \mathcal{H}_t] = \Sigma_t.$$

Here, Σ_t may vary over time, capturing non-stationarity and cross-agent correlations.

I.I.D. tasks as a special case. Suppose $x_t \overset{\text{i.i.d.}}{\sim} \mathcal{P}_X$. If, in addition, the reward process is conditionally independent of the past given the current task, i.e. $\mathbf{R}_t \perp \mathcal{H}_t \mid x_t$, then $\mathbf{f} \equiv 0$, and the model reduces to

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

for a constant covariance matrix Σ.

3.3. Survival-Based Rewards with Latent Frailty

In many systems, agent performance is naturally measured via time-to-event outcomes. For each task x_t and agent a_i, let $T_t(i) > 0$ denote the time-to-completion random variable. We allow right-censoring and introduce a censoring indicator $\delta_t(i) \in \{0, 1\}$, where $\delta_t(i) = 1$ indicates that the completion time is fully observed.

We define a baseline survival function

$$S_i(\tau \mid x_t) := \mathbb{P}(T_t(i) > \tau \mid x_t),$$

which captures heterogeneity across tasks and agents.

To model unobserved task-level difficulty, we introduce a latent frailty variable $\theta_t > 0$, assumed i.i.d. with $\mathbb{E}[\theta_t] = 1$. Conditional on (x_t, θ_t), we assume a proportional frailty model:

$$\mathbb{P}(T_t(i) > \tau \mid x_t, \theta_t) = S_i(\tau \mid x_t)^{\theta_t}.$$

Equivalently, θ_t acts as a multiplicative scaling of the baseline cumulative hazard. Under this construction, conditional on (x_t, θ_t), completion times are independent across agents: $T_t(i) \perp T_t(j) \mid (x_t, \theta_t)$, and marginalizing over θ_t induces dependence: $T_t(i) \not\perp T_t(j) \mid x_t$. Specific to our setting, one suitable choice for defining the fraility reward function is as follows,

$$R_t^i := \delta_t(i)\, S_i(T_t(i) \mid x_t)^{\theta_t}.$$

This survival-based formulation is suitable for modeling reliability, latency, and time-to-success in heterogeneous multi-agent systems.

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 $W_c(\nu_t, \mu_i)$ (Villani, 2003), 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.

We note that constructing a fixed reference measure is often non-trivial in practice, particularly when the task distribution is unknown or evolves over time. In the i.i.d. setting, a natural choice is to estimate a single representative measure of the environment using a Wasserstein barycenter of the observed task distributions, which provides a principled Fréchet mean in the space of probability measures (Chewi et al., 2025). This yields a stable global reference that summarises the stationary data-generating process. In contrast, in the non-i.i.d. setting, where the task distribution may drift or depend on historical interactions, a single global barycenter is generally insufficient to capture temporal heterogeneity. Instead, one may consider time-adaptive or history-dependent barycenters, computed over sliding windows or exponentially weighted empirical measures, thereby producing a sequence of local barycenters that track the evolving environment. This leads to a dynamic reference structure that better reflects non-stationarity while preserving the geometric advantages of the Wasserstein framework.

To incorporate randomness and modeling noise, we define the stochastic alignment cost

$$W_t^i := W_c(\nu_t, \mu_i) + \epsilon_t^i, \qquad \epsilon_t^i \sim \mathcal{N}(0, \sigma_i^2).$$

Its conditional mean is

$$w^i(x_t) := \mathbb{E}[W_t^i \mid x_t].$$

This formulation captures both systematic mismatch (via the Wasserstein term) and unpredictable variability (via the noise term ϵ_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 foundational 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 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 enable exploration and robustness to uncertainty in agent performance.

Per-round expected utility. The expected net reward under policy π_t is defined as:

$$R_t(\pi_t) := \mathbb{E}_{i_t \sim \pi_t}\left[r^{i_t}(x_t) - \lambda w^{i_t}(x_t)\right], \qquad \lambda > 0, \quad (1)$$

where $r^i(x_t)$ and $w^i(x_t)$ denote the conditional expected reward and alignment cost of assigning task x_t to agent a_i. Equivalently,

$$R_t(\pi_t) = \pi_t^\top \big(\mathbf{r}(x_t) - \lambda \mathbf{w}(x_t)\big).$$

Cumulative objective. The cumulative expected reward over a finite horizon T is

$$R(\pi_{1:T}) := \sum_{t=1}^{T} R_t(\pi_t).$$

3.6. Regret and Optimal policy.

The orchestrator selects policies without access to the true reward $\mathbf{r}(x_t)$ and alignment cost distributions $\mathbf{w}(x_t)$. We evaluate performance against an oracle that has full knowledge of these quantities. This motivates the notion of an optimal policy, which maximizes expected per-round utility under complete information. Since, the oracle is unavailable in practice, the learner deploys a sequence of policies $\{\pi_t\}_{t=1}^T$ based on partial information. The performance gap between the learned policy and the oracle is quantified via regret, which measures cumulative suboptimality over the horizon. Sublinear regret implies asymptotic convergence to oracle performance. The optimal (oracle) policy at each round is given by:

$$\pi_t^* := \arg\max_{\pi \in \Delta(\mathcal{A})} \pi^\top (\mathbf{r}(x_t) - \lambda \mathbf{w}(x_t)). \qquad (2)$$

Cumulative regret. We define the cumulative regret as the performance gap between the optimal policy and the learned policy:

$$\mathcal{R}_T := \sum_{t=1}^{T} \big(R_t(\pi_t^*) - R_t(\pi_t)\big). \qquad (3)$$

Remark 3.1. If tasks are i.i.d. and the environment is stationary, the problem reduces to a standard contextual bandit setting. Otherwise, the regret captures additional inefficiency arising from temporal dependence, non-stationary task arrivals, and stochastic alignment costs.

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 & Ben-David, 2014; Trillos & 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 & 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 model. The additional modeling 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.5.

Theorem 4.1 (Properties of OT-Regularized Orchestration)**.** *Assume (A1)–(A5) hold. Let $\mathcal{A} = \{a_1, \dots, a_M\}$ be a finite set of agents, and let $\pi_t \in \Delta(\mathcal{A})$ denote the orchestration policy at round t. Define the OT-regularized per-round utility $R_t(\pi_t)$ be as in (1) and let the cumulative regret $\mathcal{R}_T$ be as in (3) where the optimal policy π_t^* is as in (2). Let $\phi_t := \pi_t$ denote the policy weight vector. Assume initial conditions are well-defined: $\phi_0 \in \Delta(|\mathcal{A}| - 1)$. Then, the following statements hold for $i \in \{1, \cdots, M\}$:*

1. ***Sublinear OT-Regret.*** *Define the policy by $\pi_t(i) = \frac{w_t(i)}{\sum_{j=1}^M w_t(j)}$, and $w_{t+1}(i) = w_t(i)\exp\Big(\eta_t(R_t^i - \lambda W_t^i)\Big)$, where R_t^i and W_t^i are the realized reward and OT costs. Define the* static *pseudo-regret against the best fixed agent in hindsight as $\mathcal{R}_T := \max_{i\in\{1,\dots,M\}} \sum_{t=1}^T \mathbb{E}[R_t^i - \lambda W_t^i \mid x_t] - \sum_{t=1}^T \mathbb{E}_{i_t\sim\pi_t}[R_t^{i_t} - \lambda W_t^{i_t}]$. Then, the cumulative OT-regularized regret satisfies*

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

2. ***Structural OT-Optimality.*** *For any $i, j \in \mathcal{A}$, if $\mathbb{E}[R_t^i \mid x_t] = \mathbb{E}[R_t^j \mid x_t]$ and $w^i(x_t) < w^j(x_t)$, then the OT-regularized utilities satisfy*

$$\mathbb{E}[R_t^i \mid x_t] - \lambda w^i(x_t) > \mathbb{E}[R_t^j \mid x_t] - \lambda w^j(x_t).$$

3. ***Margin Robustness under Noisy Alignment.*** *Let observed alignment costs be $\widetilde{W}_t^i = W_t^i + \epsilon_t^i$, with $\epsilon_t^i \sim \mathcal{N}(0, \sigma^2)$, and suppose the base rewards are matched, $r^i(x_t) = r^j(x_t)$. Define $\Delta_{ij}(t) := w^j(x_t) - w^i(x_t)$. If $\Delta_{ij}(t) \geq 2\sigma\sqrt{\log 2}$, then*

$$\mathbb{P}\Big(r^i(x_t) - \lambda\widetilde{W}_t^i < r^j(x_t) - \lambda\widetilde{W}_t^j\Big) \leq \tfrac{1}{4}.$$

4. ***Convergence of Orchestration Weights.*** *Assume the policy update follows a stochastic approximation scheme: $\phi_{t+1} - \phi_t = \eta_t\Big(\mathrm{Softmax}(\mathbf{r}(x_t) - \lambda\mathbf{w}(x_t)) - \phi_t\Big)$, with step sizes η_t satisfying $\sum_t \eta_t = \infty$, $\sum_t \eta_t^2 < \infty$. Then ϕ_t converges almost surely to an invariant point ϕ_∞ of the mean-field ODE*

$$\dot{\phi} = \mathrm{Softmax}\big(\mathbb{E}[\mathbf{r}(x)] - \lambda\mathbb{E}[\mathbf{w}(x)]\big) - \phi.$$

5. ***Uniform Consistency of Empirical Rewards.*** *Let $\hat{r}_t(i) := \frac{1}{t}\sum_{s=1}^t R_s^i$ be the empirical average reward and let $|R_t^i| \leq R_{\max}$ be the bounded rewards. Then, we have*

$$\sup_{i\in\mathcal{A}}\left|\hat{r}_t(i) - \frac{1}{t}\sum_{s=1}^t \mathbb{E}[R_s^i \mid \mathcal{H}_{s-1}]\right| \xrightarrow[t\to\infty]{\text{a.s.}} 0.$$

In particular, in the i.i.d. case this implies

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

We note that Theorem 4.1 provides a foundational characterization of OT-regularized bandit learning. It integrates distributional alignment into the reward structure, yielding sublinear cumulative regret 1, principled agent differentiation 2, 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. We refer the reader to Appendix B for a proof of Theorem 4.1.

Remark 4.2 (Static vs. dynamic comparator). The regret in 1 is measured against the best *fixed* agent in hindsight (a static comparator), matching the exponential-weights analysis of Appendix B.1. A per-round (*dynamic*) oracle that re-selects the best agent every round is generally unattainable at rate $\mathcal{O}(\sqrt{T})$: under stochastic per-round fluctuations its regret grows linearly in T even in the i.i.d. regime, so controlling it would require an additional non-stationarity (variation) budget $V_T = \sum_t \|\bar{u}_t - \bar{u}_{t-1}\|_\infty$, which we do not assume here.

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 ques-

tions of convergence and adaptivity in non-stationary environments.

6. Synthetic Experiments

6.1. Dataset and Task Description

We consider both i.i.d. and non-i.i.d. regimes (see Fig. 5–6 in the Appendix) in order to evaluate BOT-Orch under both stationary and evolving environments. The i.i.d. settings isolate performance under stable task distributions, whereas the non-i.i.d. settings introduce different forms of temporal dependence and distributional shift.

In the stationary setting, we consider two environments. In **IID-G**, tasks x_t are sampled i.i.d. over $\mathcal{X}$ and agent rewards are drawn from fixed Gaussian distributions with matched means (≈ 0.5) but heterogeneous higher-order structure. This setting evaluates whether the orchestration policy can distinguish between agents beyond mean reward alone. In **IID-M**, tasks are sampled i.i.d. from a half-moons distribution, while rewards are drawn from fixed distributions with similar means but differing variance, skewness, and bimodality, thereby introducing additional distributional heterogeneity.

We then consider three forms of non-stationarity. In **NonIID-PS**, rewards are piecewise-stationary, with fixed means but variance shifts occurring at unknown change-points, inducing distributional shift without changing expected reward. In **NonIID-SD**, non-stationarity is introduced gradually through sinusoidal drift in the reward means over time. Finally, in **NonIID-BB**, latent reward means evolve according to a temporally correlated Brownian-bridge process with fixed endpoints, capturing smoothly varying latent dynamics.

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

We compare BOT-Orch against the following baselines. All baselines operate on the same task stream and are evaluated with identical metrics and random seeds.

- **BOT-Orch (ours).** Uses alignment-adjusted rewards $r_t(i) = \hat{r}_t(i) - \lambda W_t(i)$ and samples the selected agent i_t from a softmax policy over $r_t(i)$ (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 exclude baselines that require *full-information* feedback, i.e., access to the rewards of all agents at every round, since our setting assumes standard *bandit feedback*, where only the reward of the selected agent is observed. This includes simple greedy strategies based on exponentially smoothed reward estimates, which are not directly applicable under partial feedback without additional exploration or uncertainty-estimation mechanisms.

6.3. Evaluation Metrics

We use the following metrics for experimental comparison, computed per seed and aggregated as mean±95% confidence intervals across seeds.

- **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.
- **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 E, 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 E.5). As λ increases, BOT-Orch exhibits performance improvements, with large gains relative to $\lambda = 0$ and diminishing

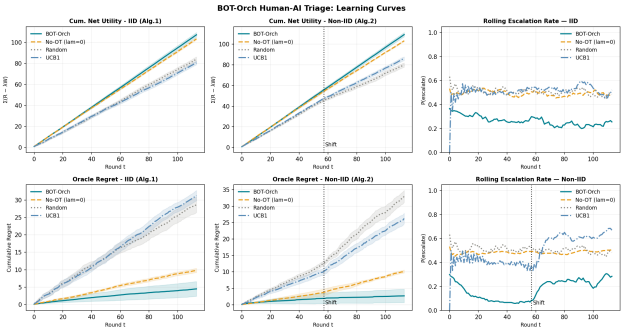

Figure 1. **Cumulative learning curves.** *Top row*: cumulative net utility. *Bottom row*: oracle regret. *Right column*: rolling escalation rate (window w=8 rounds). Left panels: IID condition (Algorithm 1); middle panels: Non-IID condition (Algorithm 2), with the dotted vertical line marking the shift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

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 F. 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 $\pm$ 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, namely **team accuracy**, defined as the fraction of patients correctly classified by the selected agent (AI or human), and **escalation rate**, defined as the fraction of patients routed to the human expert.

These additional metrics capture complementary aspects of orchestration quality beyond cumulative utility and regret. Together, these metrics provide insight into the trade-off between automation and human intervention, particularly under deployment shift and uncertainty.

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 G, 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. Conclusion

We introduced BOT-Orch, a framework for uncertainty-aware agent orchestration that formulates delegation as an OT-regularized bandit problem over heterogeneous agents. By combining sequential decision-making with distributional alignment, BOT-Orch jointly accounts for agent reliability, alignment quality, and uncertainty when selecting

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; $T = 200$). IID condition uses Algorithm 1; Non-IID uses Algorithm 2. Best in bold.

	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 ± 95% CI across 5 seeds; $T = 200$). IID condition uses Algorithm 1; Non-IID uses Algorithm 2. Best in bold.

agents under stochastic and potentially non-stationary environments. We established theoretical guarantees including sublinear OT-regularized regret, robustness to noisy alignment, and convergence of orchestration weights. Empirically, BOT-Orch consistently outperformed standard bandit and heuristic baselines across synthetic and semi-synthetic settings, particularly under distribution shift and heterogeneous agent behaviour.

More broadly, our results suggest that incorporating distributional structure directly into orchestration policies can substantially improve robustness and adaptability in modern multi-agent AI systems. This is especially relevant for emerging ecosystems composed of specialised models, reasoning systems, and tool-using agents, where orchestration increasingly plays a central role in overall system performance.

Several limitations remain. First, computing Wasserstein distances can become computationally expensive in high-dimensional settings, potentially limiting scalability in real-time applications. Second, the alignment parameter λ introduces a reward–alignment trade-off that may require calibration across environments. Third, the framework assumes access to suitable task-specific reference distributions, which may be difficult to obtain or estimate reliably in open-world settings.

Future work will investigate more scalable alignment mechanisms, adaptive calibration strategies, and evaluation in larger-scale real-world orchestration environments. We also believe that extending uncertainty-aware orchestration beyond single-agent delegation toward richer coordination and interaction settings is a promising direction for further study.

References

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

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein gan. In *ICML*, 2017.

Arjun, K., Parlevliet, D., Wang, H., and Yazdani, A. 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.

Baheri, A. and Kochenderfer, M. J. 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.

Bellemare, M. G., Dabney, W., and Munos, R. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, pp. 449–458, 2017.

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

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

	Cumulative Net Utility				Oracle Regret			
Environment	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			
Environment	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.

Bousquet, O., Gelly, S., Tolstikhin, I., Simon-Gabriel, C.-J., and Schoelkopf, B. From optimal transport to generative modeling: the vegan cookbook. *arXiv preprint arXiv:1705.07642*, 2017.

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

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

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

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

Chewi, S., Niles-Weed, J., and Rigollet, P. *Wasserstein Barycenters*, pp. 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.

Collins, K. M., Chen, V., Sucholutsky, I., Kirk, H. R., Sadek, M., Sargeant, H., Talwalkar, A., Weller, A., and Bhatt, U. Modulating language model experiences through frictions. *arXiv preprint arXiv:2407.12804*, 2024.

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9): 1853–1865, 2016.

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

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

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

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

Heifetz, A., Meier, M., and Schipper, B. Interactive unawareness. *Journal of Economic Theory*, 130(1):78–94, 2006.

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

Keswani, V., Lease, M., and Kenthapadi, K. Towards unbiased and accurate deferral to multiple experts. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 154–165, 2021.

Lai, V., Carton, S., Bhatnagar, R., Liao, Q. V., Zhang, Y., and Tan, C. 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*, pp. 1–18, 2022.

Landgren, P., Srivastava, V., and Leonard, N. E. Distributed cooperative decision-making in multiarmed bandits: Frequentist and bayesian algorithms. *CDC*, 2016.

Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press, 2020.

Letard, A., Gutowski, N., Camp, O., and Amghar, T. Bandit algorithms: A comprehensive review and their dynamic selection from a portfolio for multicriteria top-k recommendation. *Expert Systems with Applications*, 246: 123151, 2024.

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

Lu, C., Holt, S., Fanconi, C., Chan, A., Foerster, J., van der Schaar, M., and Lange, R. Discovering preference optimization algorithms with and for large language models. *Advances in Neural Information Processing Systems*, 37: 86528–86573, 2024.

Olfati-Saber, R., Fax, J. A., and Murray, R. M. 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.

Oliehoek, F. A. and Amato, C. *A Concise Introduction to Decentralized POMDPs*. Springer, 2016.

Oliver, M. C. A., Graham, M., Manolopoulou, I., Medley, G. F., Pellis, L., Pouwels, K. B., Thorpe, M., and Hollingsworth, T. D. Uncertainty quantification in cost-effectiveness analysis for stochastic-based infectious disease models: Insights from surveillance on lymphatic filariasis. *Journal of Theoretical Biology*, pp. 112197, 2025a.

Oliver, M. C. A., Hartman, E., and Needham, T. Conic formulations of transport metrics for unbalanced measure networks and hypernetworks. *arXiv preprint arXiv:2508.10888*, 2025b.

Oliver, M. C. A., Roberts, M., Schönlieb, C.-B., and Thorpe, M. Laplace learning in wasserstein space. *arXiv preprint arXiv:2511.13229*, 2025c.

Osband, I., Russo, D., and Van Roy, B. (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pp. 3003–3011, 2013.

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

Park, J. S., O'Brien, J., Cai, C. J., et al. Generative agents: Interactive simulacra of human behavior. *UIST*, 2023.

Peyré, G. and Cuturi, M. *Computational Optimal Transport*. Foundations and Trends in Machine Learning, 2019.

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

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

Rizk, Y., Awad, M., and Tunstel, E. W. 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.

Robbins, H. and Monro, S. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, 1951. doi: 10.1214/aoms/1177729586.

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

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

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

Sunehag, P., Lever, G., Gruslys, A., et al. Value-decomposition networks for cooperative multi-agent learning. In *AAMAS*, 2018.

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

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

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

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

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

Wurman, P. R., D'Andrea, R., and Mountz, M. 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.

Xie, M., Chen, S., and Lv, C. 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 . . . 12

B Missing Proofs . . . 13

C Empirical Validation of the Theoretical Claims . . . 21

D Visualization of the Synthetic Datasets . . . 22

E Additional Synthetic Experiments . . . 23

F Semi-Synthetic Experiment Settings . . . 24

G Additional Semi-Synthetic Experiments and Figures . . . 25

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 output distribution μ_i and OT cost $\mathcal{W}_t(i) = W_c(\nu_t, \mu_i) + \epsilon_t(i)$ {all arms; no agent invoked yet}
  end for
  Compute orchestration policy from the *previous* reward estimates:
    $$\pi_t(i) = \frac{\exp(\eta_t[\hat{r}_{t-1}(i) - \lambda \mathcal{W}_t(i)])}{\sum_{j=1}^{M} \exp(\eta_t[\hat{r}_{t-1}(j) - \lambda \mathcal{W}_t(j)])}$$
  Sample and execute agent $i_t \sim \pi_t$
  Observe survival time $T_t(i_t)$ and censoring $\delta_t(i_t)$ {bandit feedback: chosen arm only}
  Compute reward $R_t(i_t) = \delta_t(i_t)\, S_{i_t}(T_t(i_t) \mid x_t)^{\theta_t}$
  Update only the chosen arm: $\hat{r}_t(i_t) = \alpha \hat{r}_{t-1}(i_t) + (1-\alpha) R_t(i_t)$; $\hat{r}_t(i) = \hat{r}_{t-1}(i)$ for $i \neq i_t$
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 output distribution μ_i and OT cost $\mathcal{W}_t(i) = W_c(\nu_t, \mu_i) + \epsilon_t(i)$ {all arms; no agent invoked yet}
  end for
  Compute orchestration policy from the *previous* history-corrected estimates:
    $$\pi_t(i) = \frac{\exp(\eta_t[\hat{r}_{t-1}(i) - \lambda \mathcal{W}_t(i)])}{\sum_{j=1}^{M} \exp(\eta_t[\hat{r}_{t-1}(j) - \lambda \mathcal{W}_t(j)])}$$
  Sample and execute agent $i_t \sim \pi_t$
  Observe survival time $T_t(i_t)$, censoring $\delta_t(i_t)$, reward $R_t(i_t) = \delta_t(i_t)\, S_{i_t}(T_t(i_t) \mid x_t)^{\theta_t}$ {bandit feedback}
  Update only the chosen arm (with history correction): {temporal dependence from past observed rewards}
    $\hat{r}_t(i_t) = \alpha \hat{r}_{t-1}(i_t) + (1-\alpha) R_t(i_t) + f_{i_t}(\mathbf{R}_{1:t-1})$; $\hat{r}_t(i) = \hat{r}_{t-1}(i)$ for $i \neq i_t$
  Update history $\mathcal{H}_{t+1} = \mathcal{H}_t \cup \{(x_t, i_t, R_t(i_t), \mathcal{W}_t)\}$
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

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

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 $\hat{r}_t(i)$ is a Robbins–Monro stochastic approximation of $\mathbb{E}[R_t(i) \mid x_t]$ (Robbins & Monro, 1951). With a decaying step size $\alpha_t = 1 - \frac{1}{t}$ it reduces to the running average $\frac{1}{t}\sum_{s \le t} R_s(i)$, for which strong consistency holds (Theorem B.7). The deployed fixed-α exponential smoothing is instead an asymptotically biased *tracking* estimator—retaining a stationary variance of order $\frac{1-\alpha}{1+\alpha}\mathrm{Var}(R)$—which we adopt to follow non-stationary (drifting) rewards; the history term $f_i(\cdot)$ plays the analogous role in the non-i.i.d. case. Substituting this estimator into the theoretical objective yields a computable approximation of $r_t(i)$, 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 1 in Theorem 4.1 (Sublinear OT-Regret)

Theorem B.1 (Sublinear OT-Regret)**.** *Assume* (A1)–(A5)*. Let the BOT–Orch policy be defined by*

$$\pi_t(i) = \frac{w_t(i)}{\sum_{j=1}^{M} w_t(j)}, \qquad w_{t+1}(i) = w_t(i)\exp\Big(\eta_t(R_t^i - \lambda W_t^i)\Big),$$

where R_t^i and W_t^i are the realized reward and OT costs. Define the static *pseudo-regret against the best fixed agent in hindsight*

$$\mathcal{R}_T := \max_{i\in\{1,\dots,M\}} \sum_{t=1}^{T} \mathbb{E}[R_t^i - \lambda W_t^i \mid x_t] - \sum_{t=1}^{T} \mathbb{E}_{i_t\sim\pi_t}[R_t^{i_t} - \lambda W_t^{i_t}].$$

If the inverse-temperature schedule satisfies $\eta_t \asymp t^{-1/2}$ by Assumption (A5), then

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

Proof. The expected per-round utility of a policy $\pi_t \in \Delta(\mathcal{A})$ is

$$R_t(\pi_t) = \sum_{i=1}^{M} \pi_t(i)\big(r^i(x_t) - \lambda w^i(x_t)\big).$$

For each agent i, define the realized instantaneous utility

$$u_t^i := R_t^i - \lambda W_t^i,$$

and its conditional expectation

$$\bar{u}_t^i := r^i(x_t) - \lambda w^i(x_t).$$

We compare against a fixed action $i^* \in \{1,\dots,M\}$. The oracle policy selects

$$R_t(\pi_t^*) = \max_{1\le i\le M}\big(r^i(x_t) - \lambda w^i(x_t)\big),$$

i.e. a point mass on the best action in each round.

Define the log-partition potential

$$\Phi_t := \log \sum_{i=1}^{M} w_t(i).$$

By assumption (A1), there exist constants $a < b$ such that for all t, i,

$$u_t^i \in [a, b].$$

Step 1: Constant inverse temperature. Assume $\eta_t \equiv \eta > 0$. Then

$$\Phi_{t+1} - \Phi_t = \log \sum_{i=1}^{M} \pi_t(i)\exp\big(\eta u_t^i\big).$$

Applying Hoeffding's lemma,

$$\Phi_{t+1} - \Phi_t \le \eta \sum_{i=1}^{M} \pi_t(i) u_t^i + \frac{\eta^2}{8}(b-a)^2.$$

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

$$\Phi_{T+1} - \Phi_1 \le \eta \sum_{t=1}^{T}\sum_{i=1}^{M} \pi_t(i) u_t^i + \frac{\eta^2 T}{8}(b-a)^2.$$

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

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

With uniform initialization $w_1(i) = 1$, we obtain

$$\sum_{t=1}^{T} u_t^{i^*} - \sum_{t=1}^{T}\sum_{i=1}^{M} \pi_t(i) u_t^i \leq \frac{\log M}{\eta} + \frac{\eta T}{8}(b-a)^2.$$

Choosing

$$\eta = \sqrt{\frac{8 \log M}{T(b-a)^2}}$$

yields

$$\mathcal{R}_T \leq (b-a)\sqrt{\frac{T \log M}{2}}.$$

Taking conditional expectations with respect to $\mathcal{H}_{t-1}$,

$$\mathbb{E}[u_t^i \mid x_t, \mathcal{H}_{t-1}] = r^i(x_t) - \lambda w^i(x_t),$$

gives the same regret bound in expectation.

Step 2: Time-varying inverse temperature. Let $\eta_t \asymp t^{-1/2}$. Then

$$\Phi_{t+1} - \Phi_t \leq \eta_t \sum_{i=1}^{M} \pi_t(i) u_t^i + \frac{\eta_t^2}{8}(b-a)^2.$$

Summing over t yields

$$\mathcal{R}_T \leq \frac{\log M}{\eta_T} + \frac{(b-a)^2}{8} \sum_{t=1}^{T} \eta_t^2.$$

Since $\eta_t = ct^{-1/2}$,

$$\frac{1}{\eta_T} = O(\sqrt{T}), \qquad \sum_{t=1}^{T} \eta_t^2 = O(\log T),$$

so

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

If u_t^i admits a martingale decomposition with bounded increments, Azuma–Hoeffding implies that deviations between realized and conditional cumulative utilities are $O(\sqrt{T})$ (up to log factors), which does not change the overall regret rate. □

B.2. Proof of 2 in Theorem 4.1 (Lipschitz stability of alignment-adjusted rewards)

Proposition B.2 (OT geometry and preference stability)**.** *Assume (A1). Let $\mathcal{A} = \{a_1, \ldots, a_M\}$ be the set of agents, $\mathcal{X} \subset \mathbb{R}^d$ be the measurable space and $\mathcal{Y}$ denote the outcome space. For $i \in \{1, \cdots, M\}$ let each agent be denoted as a_i and let $\mu_i \in \mathcal{P}(\mathcal{Y})$ be its source outcome distribution. For each task $x_t \in \mathcal{X}$, let $\nu(x_t) \in \mathcal{P}(\mathcal{Y})$ denote the task-induced target outcome distribution. Define the cost as $W_t^i := W_c(\nu_t, \mu_i) + \epsilon_t^i$, such that $\mathbb{E}[\epsilon_t^i \mid x_t] = 0$. Let the conditional cost $w^i(x_t) := \mathbb{E}[W_t^i \mid x_t] = W_c(\nu_t, \mu_i)$ and conditional reward as $r^i(x_t) := \mathbb{E}[R_t^i \mid x_t]$. Then, the following statements hold:*

(i) *For any $\nu_t, \nu_t' \in \mathcal{P}(\mathcal{Y})$, we have*

$$\left|W_c(\nu_t, \mu_i) - W_c(\nu_t', \mu_i)\right| \leq L\, W_1(\nu_t, \nu_t').$$

(ii) *If $r^i(x_t) = r^j(x_t)$, then for any $\lambda > 0$, we have*

$$W_c(\nu_t, \mu_i) < W_c(\nu_t, \mu_j) \iff r^i(x_t) - \lambda w^i(x_t) > r^j(x_t) - \lambda w^j(x_t).$$

Proof. We outline the proof in a sequence.

(i) By definition of $w^i(x_t; \nu_t)$ and $w^i(x_t; \nu'_t)$,

$$\left|w^i(x_t; \nu_t) - w^i(x_t; \nu'_t)\right| = \left|W_c(\nu_t, \mu_i) - W_c(\nu'_t, \mu_i)\right|. \tag{1}$$

Since c is L-Lipschitz by (A1), standard Wasserstein stability implies

$$\left|W_c(\nu_t, \mu_i) - W_c(\nu'_t, \mu_i)\right| \le L\, W_1(\nu_t, \nu'_t), \tag{2}$$

which proves (i).

(ii) Using $r^i(x_t) = r^j(x_t)$,

$$\big(r^i(x_t) - \lambda w^i(x_t)\big) - \big(r^j(x_t) - \lambda w^j(x_t)\big) = -\lambda\big(w^i(x_t) - w^j(x_t)\big). \tag{3}$$

Since $\lambda > 0$, this implies

$$r^i(x_t) - \lambda w^i(x_t) > r^j(x_t) - \lambda w^j(x_t) \iff w^i(x_t) < w^j(x_t). \tag{4}$$

Substituting $w^i(x_t) = W_c(\nu_t, \mu_i)$ completes the result. □

Lemma B.3 (Concentration of frailty-adjusted rewards)**.** *Assume* (A3)–(A4)*. Fix $x_t \in \mathcal{X}$ and $a_i \in \mathcal{A}$, and define the frailty reward as $R_t^i := \delta_t(i)\, S_i(T_t(i) \mid x_t)^{\theta_t}$, where $0 \le S_i(\tau \mid x_t) \le 1$, $\theta_t > 0$, and $\delta_t(i) \in \{0, 1\}$. Then, the centered reward $R_t^i - \mathbb{E}[R_t^i \mid x_t]$ is conditionally sub-Gaussian. More precisely, for every $\varepsilon > 0$,*

$$\mathbb{P}\Big(\big|R_t^i - \mathbb{E}[R_t^i \mid x_t]\big| > \varepsilon \,\Big|\, x_t\Big) \le 2e^{-2\varepsilon^2}.$$

Proof. Since $0 \le S_i(T_t(i) \mid x_t) \le 1$ and $\delta_t(i) \in \{0, 1\}$, we have

$$0 \le R_t^i \le 1 \qquad \text{almost surely.}$$

Fix x_t and define

$$m := \mathbb{E}[R_t^i \mid x_t], \qquad X := R_t^i - m.$$

Then $m \in [0, 1]$ and $\mathbb{E}[X \mid x_t] = 0$.

We first bound the conditional moment generating function. For any $\lambda \in \mathbb{R}$, convexity of $y \mapsto e^{\lambda y}$ on $[0, 1]$ implies that for every $y \in [0, 1]$,

$$e^{\lambda y} \le (1 - y)e^0 + ye^\lambda = 1 - y + ye^\lambda.$$

Applying this pointwise inequality to R_t^i and taking conditional expectation yields

$$\mathbb{E}[e^{\lambda R_t^i} \mid x_t] \le 1 + m(e^\lambda - 1).$$

Therefore,

$$\mathbb{E}[e^{\lambda X} \mid x_t] = e^{-\lambda m}\mathbb{E}[e^{\lambda R_t^i} \mid x_t] \le e^{-\lambda m}\big(1 + m(e^\lambda - 1)\big) =: \phi(\lambda, m).$$

Define $\psi(\lambda) := \log \phi(\lambda, m)$. A direct computation gives

$$\psi(0) = 0, \qquad \psi'(0) = 0, \qquad \psi''(\lambda) = \frac{m(1 - m)e^\lambda}{(1 - m + me^\lambda)^2}.$$

A direct simplification shows that

$$\psi''(\lambda) = \frac{1}{4\cosh^2(\lambda/2)} \le \frac{1}{4} \qquad \text{for all } \lambda \in \mathbb{R},\ m \in [0, 1].$$

Now apply Taylor's theorem with integral remainder:

$$\psi(\lambda) = \psi(0) + \psi'(0)\lambda + \int_0^\lambda (\lambda - s)\psi''(s)\, ds.$$

Using $\psi(0) = \psi'(0) = 0$ and $\psi''(s) \le \frac{1}{4}$, we obtain

$$\psi(\lambda) \le \int_0^\lambda (\lambda - s)\frac{1}{4}\, ds = \frac{\lambda^2}{8}.$$

Hence,

$$\mathbb{E}[e^{\lambda X} \mid x_t] \le \exp\left(\frac{\lambda^2}{8}\right),$$

so X is conditionally sub-Gaussian with variance proxy $1/4$.

Finally, for any $\lambda > 0$, Markov's inequality gives

$$\mathbb{P}(X > \varepsilon \mid x_t) \le \exp\left(-\lambda\varepsilon + \frac{\lambda^2}{8}\right).$$

Optimizing in λ yields $\lambda = 4\varepsilon$, hence

$$\mathbb{P}(X > \varepsilon \mid x_t) \le \exp(-2\varepsilon^2).$$

Applying the same argument to $-X$ yields

$$\mathbb{P}(|X| > \varepsilon \mid x_t) \le 2e^{-2\varepsilon^2}.$$

□

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

Lemma B.4 (Margin robustness under Gaussian noise)**.** *Let $\widetilde{W}_c(\mu_i, \nu) = W_c(\mu_i, \nu) + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ are independent, and suppose the two agents have matched conditional rewards $r^i(x_t) = r^j(x_t)$. Define the alignment margin $\Delta_{ij} := W_c(\mu_j, \nu) - W_c(\mu_i, \nu)$. Writing the alignment-adjusted utility $U^k := r^k(x_t) - \lambda\widetilde{W}_c(\mu_k, \nu)$ ($\lambda > 0$), if $\Delta_{ij} > 0$ then the probability of mis-ordering satisfies*

$$\mathbb{P}\big(U^i < U^j\big) = \Phi\left(-\frac{\Delta_{ij}}{\sqrt{2}\sigma}\right) \le \frac{1}{2}\exp\left(-\frac{\Delta_{ij}^2}{4\sigma^2}\right).$$

In particular, if $\Delta_{ij} \ge 2\sigma\sqrt{\log 2}$, then the mis-ordering probability is at most $\frac{1}{4}$.

Proof. Without loss of generality assume $\Delta_{ij} > 0$. Because the base rewards are matched and $\lambda > 0$, a mis-ordering $r^i(x_t) - \lambda\widetilde{W}_c(\mu_i, \nu) < r^j(x_t) - \lambda\widetilde{W}_c(\mu_j, \nu)$ is equivalent to $\widetilde{W}_c(\mu_i, \nu) > \widetilde{W}_c(\mu_j, \nu)$, i.e. $\epsilon_i - \epsilon_j > \Delta_{ij}$; the factor λ cancels. With $Z := \epsilon_i - \epsilon_j \sim \mathcal{N}(0, 2\sigma^2)$,

$$\mathbb{P}(\text{mis-order}) = \mathbb{P}(Z > \Delta_{ij}) = \Phi\left(-\frac{\Delta_{ij}}{\sqrt{2}\sigma}\right).$$

Using the Gaussian tail bound $\Phi(-x) \le \frac{1}{2}e^{-x^2/2}$ for $x \ge 0$ with $x = \Delta_{ij}/(\sqrt{2}\sigma)$, we obtain

$$\mathbb{P}(\text{mis-order}) \le \frac{1}{2}\exp\left(-\frac{\Delta_{ij}^2}{4\sigma^2}\right).$$

The condition $\Delta_{ij} \ge 2\sigma\sqrt{\log 2}$ gives $\Delta_{ij}^2/(4\sigma^2) \ge \log 2$, so the right-hand side is at most $\frac{1}{2}e^{-\log 2} = 1/4$, completing the proof. (The previously stated threshold $\sigma\sqrt{2\log 2}$ yields only $\frac{1}{2}e^{-\log 2/2} \approx 0.354$.) □

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

Theorem B.5 (Almost-sure convergence of orchestration weights)**.** *Assume (A1)–(A5). Let $(\gamma_t)_{t\geq 0}$ satisfy*

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

Assume the task process (x_t) is stationary and independent of (ϕ_t), and that there exists a measurable function $\bar{r} : \mathcal{X} \to \mathbb{R}^M$ such that

$$\mathbb{E}[R_t^i - \lambda W_t^i \mid x_t] = \bar{r}^i(x_t).$$

Then (ϕ_t) converges almost surely to the internally chain transitive set of the ODE

$$\dot{\phi} = \mathrm{Softmax}(\bar{r}(x)) - \phi,$$

where $\bar{r}(x)$ is evaluated under the stationary distribution of x_t.

Proof. Let $\mathcal{F}_t = \sigma(\phi_s, x_s, R_s, W_s : s \leq t)$.

The update can be written as

$$\phi_{t+1} = \phi_t + \gamma_t\big(\mathrm{Softmax}(u_t) - \phi_t\big),$$

where $u_t := (R_t^i - \lambda W_t^i)_{i=1}^M$.

Define the drift:

$$H(\phi_t) := \mathbb{E}[\mathrm{Softmax}(u_t) \mid \mathcal{F}_{t-1}] - \phi_t,$$

and the noise term:

$$\xi_{t+1} := \mathrm{Softmax}(u_t) - \mathbb{E}[\mathrm{Softmax}(u_t) \mid \mathcal{F}_{t-1}].$$

Then

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

By boundedness of rewards, $\mathrm{Softmax}(u_t)$ takes values in the simplex, hence ξ_{t+1} is a martingale difference sequence with bounded second moments.

Now use the stationarity assumption: since x_t is stationary and independent of ϕ_t,

$$\mathbb{E}[\mathrm{Softmax}(u_t) \mid \mathcal{F}_{t-1}] = \mathbb{E}_{x\sim\pi_x}\big[\mathrm{Softmax}(\bar{r}(x))\big],$$

where π_x is the invariant distribution of (x_t).

Define the averaged drift

$$\bar{H}(\phi) := \mathbb{E}_{x\sim\pi_x}\big[\mathrm{Softmax}(\bar{r}(x))\big] - \phi.$$

Then the recursion becomes a standard Robbins–Monro scheme:

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

The function $\bar{H}$ is Lipschitz on the simplex, and the noise term has bounded second moment. The step-size conditions ensure

$$\sum_t \gamma_t = \infty, \qquad \sum_t \gamma_t^2 < \infty,$$

so the stochastic approximation framework applies.

By the stochastic approximation theorem of Benaïm (1999), the piecewise linear interpolation of (ϕ_t) is an asymptotic pseudo-trajectory of the ODE

$$\dot{\phi} = \bar{H}(\phi).$$

Therefore, the limit set of (ϕ_t) is contained in the internally chain transitive invariant sets of this ODE.

Since the ODE is a linear contraction toward the fixed point

$$\phi^* = \mathbb{E}_{x\sim\pi_x}[\mathrm{Softmax}(\bar{r}(x))],$$

all internally chain transitive sets reduce to this equilibrium, which is globally asymptotically stable.

Hence,

$$\phi_t \xrightarrow{a.s.} \phi^*.$$

□

Remark B.6. The proof follows a stochastic approximation argument in which the orchestration weights evolve as a Robbins–Monro recursion driven by a softmax response to OT-adjusted rewards. The key step is the decomposition of the update into a deterministic drift term and a martingale difference noise, allowing the dynamics to be linked to a limiting ordinary differential equation via the asymptotic pseudo-trajectory framework. Intuitively, the softmax operator induces a smooth, entropy-regularised selection mechanism that continuously reweights agents according to their relative OT-adjusted performance. Under step sizes satisfying the usual summability conditions, stochastic fluctuations vanish asymptotically, and the evolution tracks the averaged response of this softmax mechanism under the stationary task distribution.

The main structural novelty lies in the fact that the drift is not driven by exogenous rewards but by optimal-transport-adjusted utilities, which couple agent performance through a geometric cost. This embeds a stochastic approximation scheme into a geometry-aware, mean-field decision system, where the limiting ODE describes an entropy-regularised flow toward stationary softmax equilibria.

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

Theorem B.7 (Empirical reward decomposition and consistency)**.** *Assume (A2)–(A4) and bounded rewards $|R_t^i| \le R_{\max}$. Define*

$$r^i(x_t) := \mathbb{E}[R_t^i \mid x_t], \qquad \hat{r}_t(i) := \frac{1}{t}\sum_{s=1}^{t} R_s^i.$$

Then the following hold:

(i) i.i.d. case. *If $(x_t, \mathbf{R}_t)_{t\ge1}$ are i.i.d., then for every $i \in \mathcal{A}$,*

$$\hat{r}_t(i) \xrightarrow[t\to\infty]{\text{a.s.}} \mathbb{E}[R_t^i].$$

(ii) general (possibly non-i.i.d.) case. *Let $\mathcal{H}_{t-1}$ be the filtration generated by $(x_s, \mathbf{R}_s)_{s\le t-1}$, and define the martingale difference sequence $\xi_t(i) := R_t^i - \mathbb{E}[R_t^i \mid \mathcal{H}_{t-1}]$. Assume $\mathbb{E}[\xi_t(i) \mid \mathcal{H}_{t-1}] = 0$ and $|\xi_t(i)| \le R_{\max}$ a.s. Then, for every $i \in \mathcal{A}$,*

$$\hat{r}_t(i) - \frac{1}{t}\sum_{s=1}^{t} \mathbb{E}[R_s^i \mid \mathcal{H}_{s-1}] \xrightarrow[t\to\infty]{\text{a.s.}} 0.$$

Moreover, since $\mathcal{A}$ is finite,

$$\sup_{i\in\mathcal{A}} \left| \hat{r}_t(i) - \frac{1}{t}\sum_{s=1}^{t} \mathbb{E}[R_s^i \mid \mathcal{H}_{s-1}] \right| \xrightarrow[t\to\infty]{\text{a.s.}} 0.$$

Proof. We treat the two cases separately.

(i) i.i.d. case. If $(x_t, \mathbf{R}_t)$ are i.i.d., then $(R_t^i)_{t\ge1}$ is an i.i.d. integrable sequence. Hence, by the strong law of large numbers,

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

(ii) general non-i.i.d. case. Let $\mathcal{H}_t := \sigma(x_s, \mathbf{R}_s : s \le t)$ and write

$$R_t^i = \mathbb{E}[R_t^i \mid \mathcal{H}_{t-1}] + \xi_t(i), \qquad \xi_t(i) := R_t^i - \mathbb{E}[R_t^i \mid \mathcal{H}_{t-1}].$$

Then $(\xi_t(i), \mathcal{H}_t)$ is a martingale difference sequence with

$$\mathbb{E}[\xi_t(i) \mid \mathcal{H}_{t-1}] = 0, \qquad \mathbb{E}[\xi_t(i)^2 \mid \mathcal{H}_{t-1}] \le R_{\max}^2.$$

Summing over t gives

$$\hat{r}_t(i) = \frac{1}{t}\sum_{s=1}^{t} \mathbb{E}[R_s^i \mid \mathcal{H}_{s-1}] + \frac{1}{t}\sum_{s=1}^{t} \xi_s(i).$$

Hence,

$$\hat{r}_t(i) - \frac{1}{t}\sum_{s=1}^{t} \mathbb{E}[R_s^i \mid \mathcal{H}_{s-1}] = \frac{1}{t}\sum_{s=1}^{t} \xi_s(i).$$

Since

$$\sum_{t=1}^{\infty} \frac{\mathbb{E}[\xi_t(i)^2 \mid \mathcal{H}_{t-1}]}{t^2} \le R_{\max}^2 \sum_{t=1}^{\infty} \frac{1}{t^2} < \infty,$$

the martingale strong law implies

$$\frac{1}{t}\sum_{s=1}^{t} \xi_s(i) \xrightarrow[t\to\infty]{\text{a.s.}} 0.$$

Therefore,

$$\hat{r}_t(i) - \frac{1}{t}\sum_{s=1}^{t} \mathbb{E}[R_s^i \mid \mathcal{H}_{s-1}] \xrightarrow[t\to\infty]{\text{a.s.}} 0.$$

Finally, since $\mathcal{A}$ is finite, taking a maximum preserves almost sure convergence:

$$\max_{i\in\mathcal{A}} \left| \frac{1}{t}\sum_{s=1}^{t} \xi_s(i) \right| \xrightarrow[t\to\infty]{\text{a.s.}} 0.$$

□

Remark B.8. In the i.i.d. case, the empirical average $\hat{r}_t(i)$ is formed from independent samples drawn from a fixed distribution, so the strong law of large numbers implies convergence to the constant mean $\mathbb{E}[R_t^i]$.

In the non-i.i.d. case, the rewards are adapted to the history and admit the decomposition into a predictable component $\mathbb{E}[R_t^i \mid \mathcal{H}_{t-1}]$ and a martingale difference noise term. Consequently, the empirical average splits into the time-average of the predictable process and a martingale average, and boundedness ensures that the latter vanishes almost surely. This yields that $\hat{r}_t(i)$ tracks the time-averaged conditional expectation

$$\frac{1}{t}\sum_{s=1}^{t} \mathbb{E}[R_s^i \mid \mathcal{H}_{s-1}],$$

up to an asymptotically negligible martingale error. However, full convergence of $\hat{r}_t(i)$ is not guaranteed in general, since the predictable process $\mathbb{E}[R_t^i \mid \mathcal{H}_{t-1}]$ may itself fail to converge or even fail to admit a Cesàro limit (for instance, under oscillatory or adversarial dynamics). Additional structure, such as ergodicity or asymptotic stationarity of the conditional mean process, is required to ensure convergence of the empirical averages.

We emphasise that Theorem B.7 concerns the running average $\hat{r}_t(i) = \frac{1}{t}\sum_{s\le t} R_s^i$ (Robbins–Monro with step $1/t$), which is the consistent estimator. The algorithms in Appendix B instead deploy a fixed-α exponential moving average: this is *not* consistent (it retains a stationary variance of order $\frac{1-\alpha}{1+\alpha}\mathrm{Var}(R)$) but is preferred for tracking non-stationary rewards. The i.i.d. consistency claim therefore applies verbatim to the running-average instantiation ($\alpha_t = 1 - 1/t$); the empirical convergence of both estimators is shown in Appendix C.

C. Empirical Validation of the Theoretical Claims

Under strict bandit feedback (only the invoked agent's reward observed; cf. Algorithms 1–2), we empirically corroborate the corrected statements of Theorem 4.1. All panels use the synthetic suite at T=200, 30 seeds.

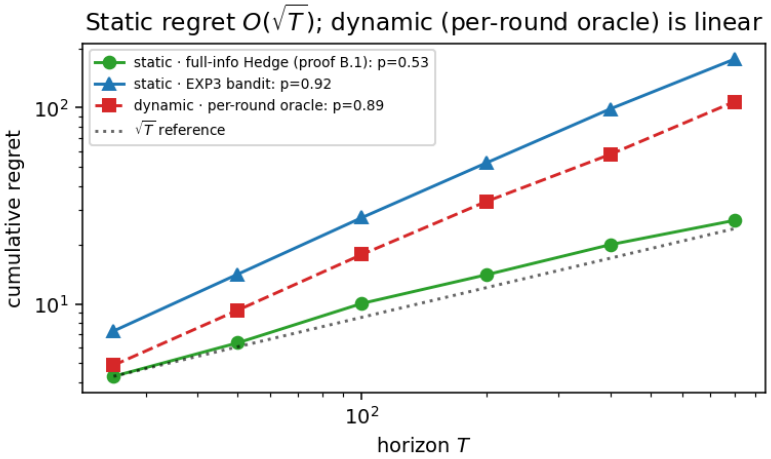

Figure 2. **Static regret is $\mathcal{O}(\sqrt{T})$ (1); the per-round oracle is linear.** The full-information exponential-weights (Hedge) procedure analysed in Appendix B.1 attains the $\sqrt{T}$ rate (log–log slope 0.53); an importance-weighted EXP3 bandit variant is sublinear ($0.92 < 1$). In contrast, regret against the per-round (dynamic) oracle grows linearly (slope 0.89) even in i.i.d. tasks, confirming Remark 4.2 and motivating the fixed-comparator formulation.

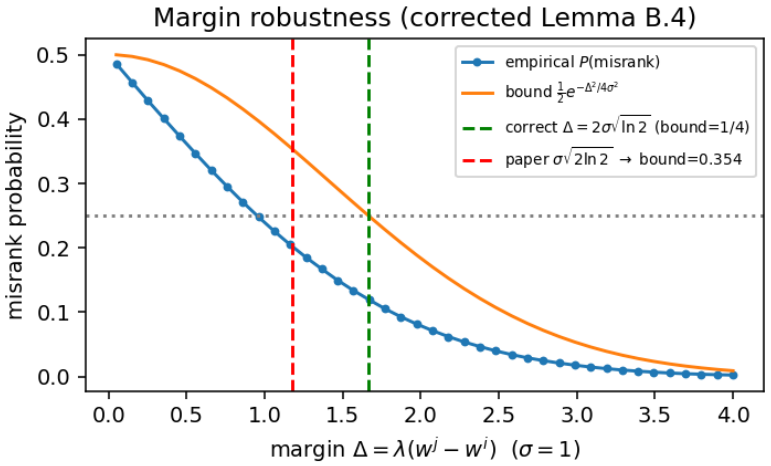

Figure 3. **Margin robustness (3, Lemma B.4).** Empirical mis-ordering probability vs. margin Δ (σ=1). It never exceeds the bound $\frac{1}{2}e^{-\Delta^2/4\sigma^2}$, which equals exactly $1/4$ at the corrected threshold $\Delta = 2\sigma\sqrt{\ln 2}$ (green) and 0.354 at the previously stated $\sigma\sqrt{2\ln 2}$ (red).

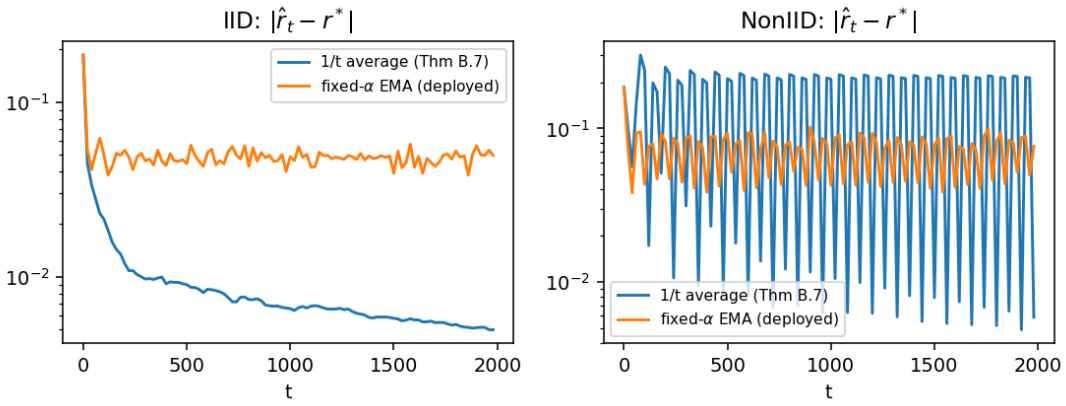

Figure 4. **Consistency (5, Theorem B.7).** The $1/t$ running average converges in the i.i.d. case ($|\hat{r}_t - r^*| \to 0$ at the $1/\sqrt{t}$ rate) while the deployed fixed-α EMA plateaus; under drift (right) the EMA instead tracks the moving mean, justifying its use for non-stationarity.

D. Visualization of the Synthetic Datasets

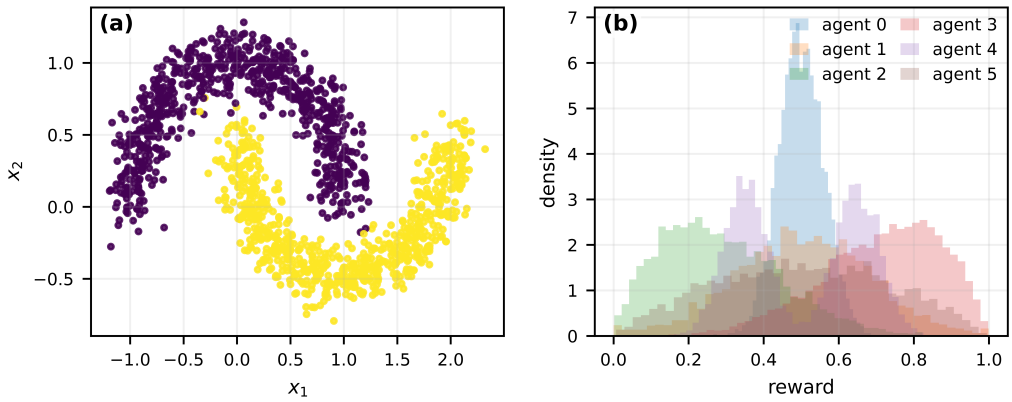

Figure 5. **IID data generation** (a) IID task contexts sampled from a half-moons distribution, illustrating a stationary but structured task space. (b) IID reward distributions for K agents, where rewards are independently drawn over time from fixed distributions with a matched mean (approximately 0.5) but heterogeneous higher-order properties (e.g., variance, skewness, and bimodality).

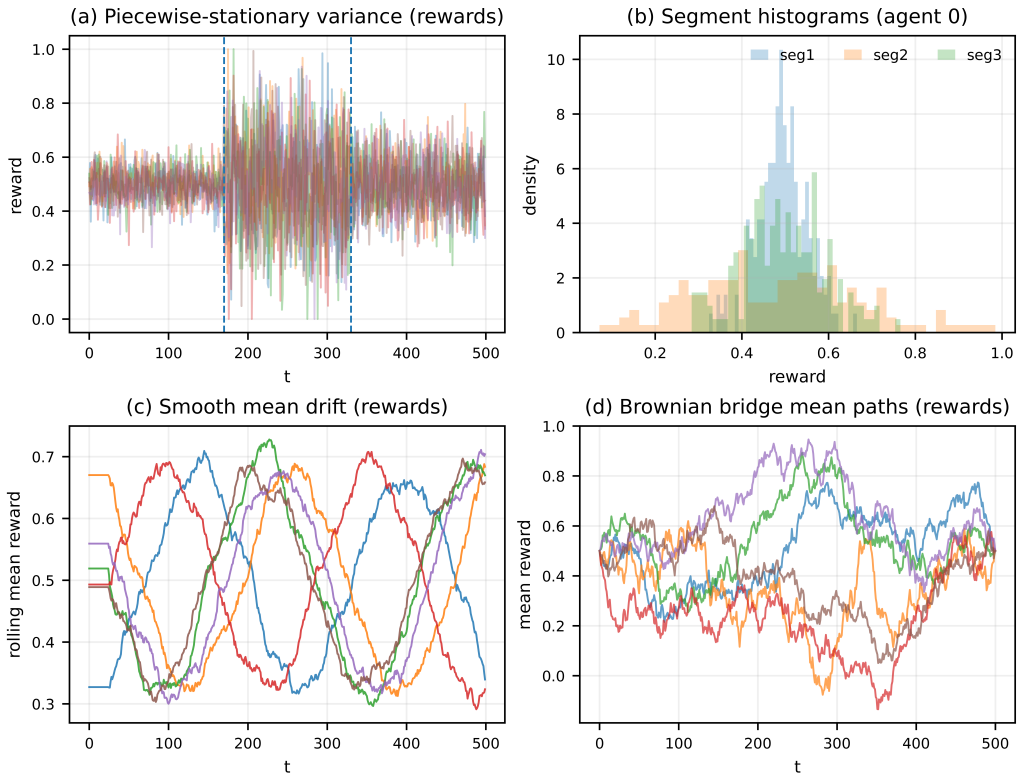

Figure 6. **Non-IID data generation** (a) Piecewise-stationary rewards where the variance changes at unknown changepoints while the mean remains fixed, shown as reward trajectories over time across agents. (b) Segment-wise reward histograms for a representative agent, revealing distributional shifts 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.

E. Additional Synthetic Experiments

E.1. Dataset and Task Description

Datasets and tasks are as in 6.

E.2. Baselines

Baselines are as in Section 6.

E.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.

E.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.

	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; $T = 200$).

	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 ± 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.

E.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.

F. 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

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

(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:

$$W_t(\text{AI}) = 1 - R_t(\text{AI}) \quad \text{(probability AI is wrong on patient } t\text{)}, \tag{5}$$

$$W_t(\text{human}) = 1 - p_h(x_t) \quad \text{(probability human is wrong on patient } t\text{)}, \tag{6}$$

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.

G. Additional Semi-Synthetic Experiments and Figures

G.1. Diagnostic Analysis

Figure 7 and 8 provide diagnostic analyses.

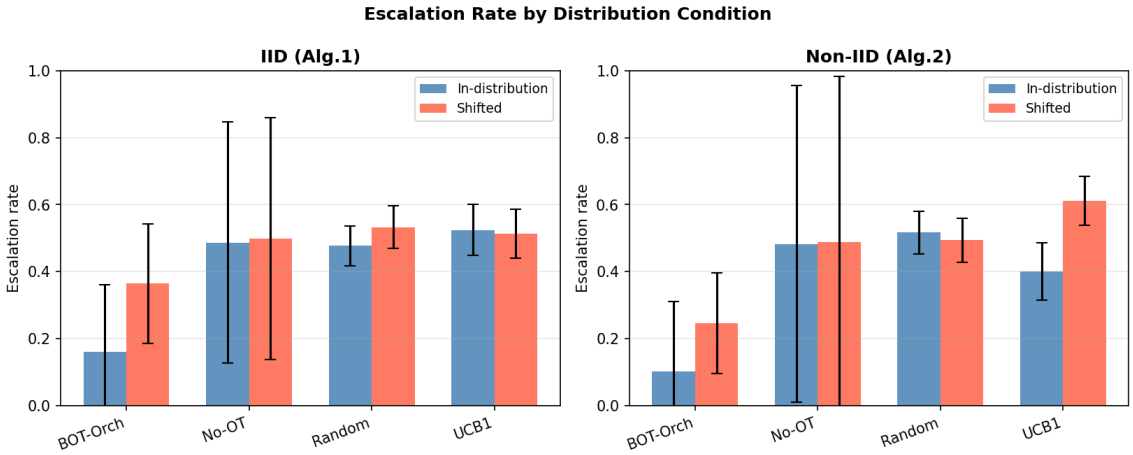

Figure 7. **Escalation rate by distribution condition.** Mean escalation rate for in-distribution patients (blue) and shifted patients (red) per method. Error bars show standard deviation across seeds. *Left*: IID condition; *right*: Non-IID condition. BOT-Orch achieves a higher escalation rate on shifted 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.

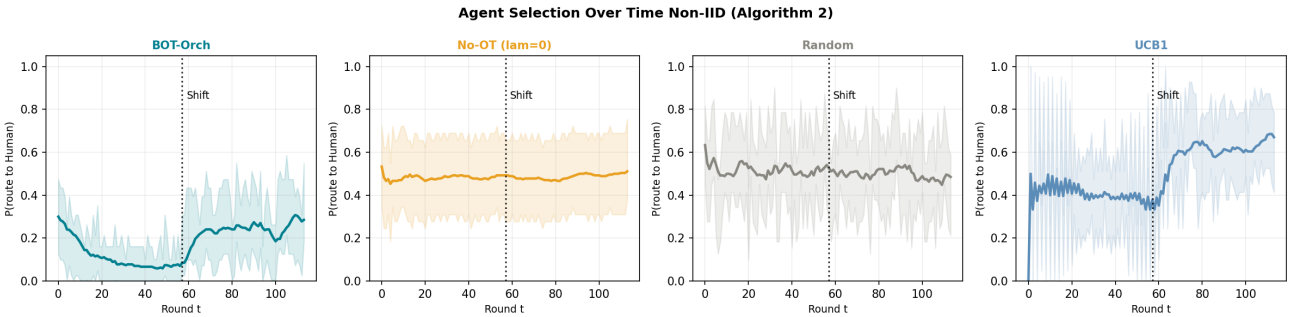

Figure 8. **Agent selection trajectories, Non-IID condition (Algorithm 2).** Rolling probability of routing to the human expert (window w=8 rounds) for each method over the deployment episode. The dotted vertical line marks the shift onset at round 57. BOT-Orch (top-left): routing probability rises after 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.

G.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).

G.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 9 summarise the results.

Sensitivity and Optimal λ. In the IID condition (Figure 9, 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 9, 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 > $W_t(\text{AI})$≈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 λ≈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 ±5.53 at λ=2.0 to ±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.

	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 $\pm$ 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.

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 $3\times$ 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.

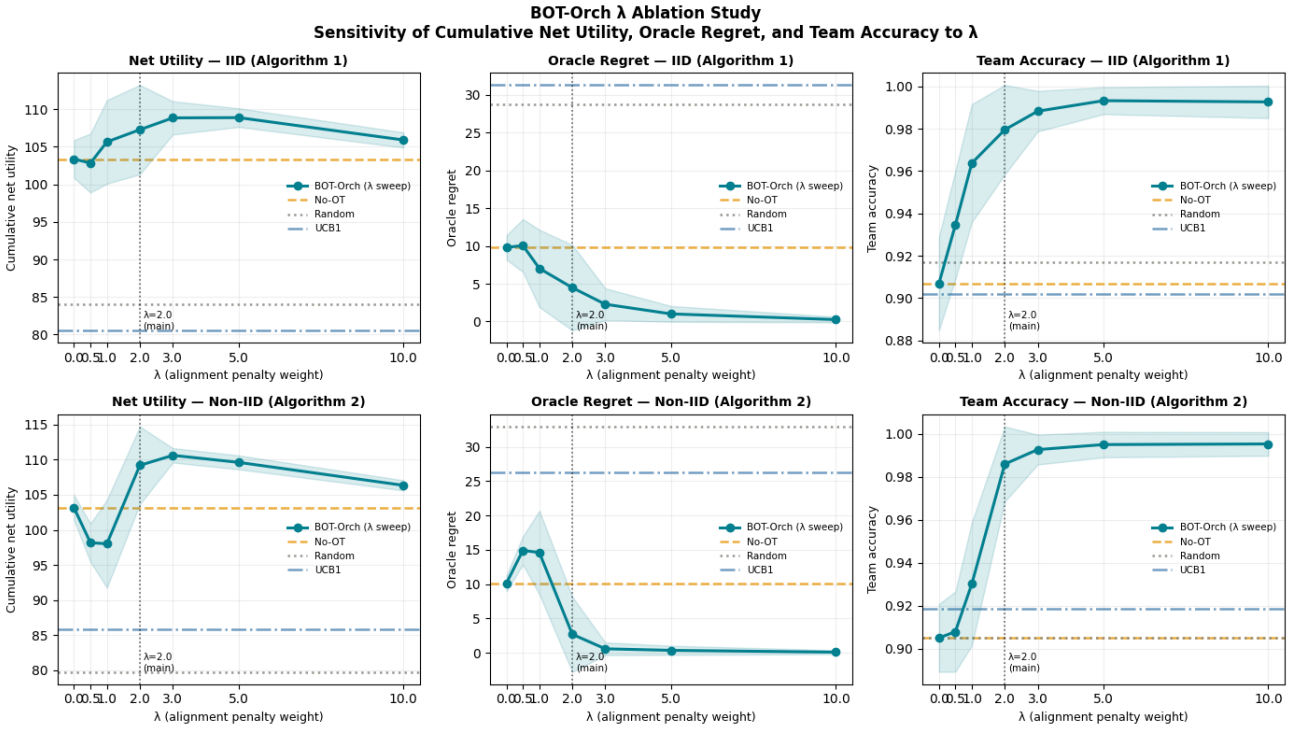

Figure 9. λ **sensitivity curves.** Cumulative net utility (*left*), oracle regret (*centre*), and team accuracy (*right*) as functions of λ for BOT-Orch (teal line with $\pm$1 std shading). Horizontal dashed lines show the No-OT, Random, and UCB1 baselines. The dotted vertical line marks λ=2.0 (initial 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.

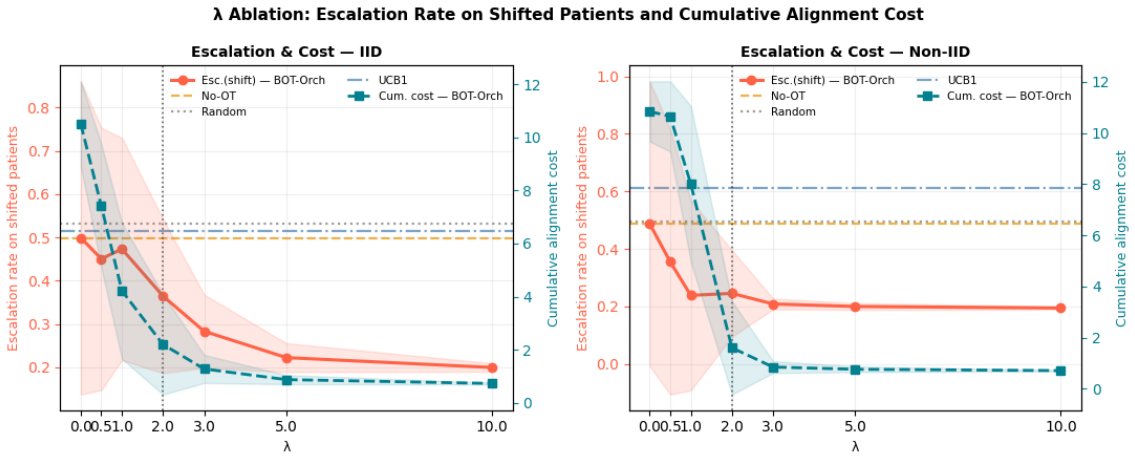

Figure 10. **Escalation rate on shifted patients (left axis, red) and cumulative alignment cost (right axis, teal) as functions of λ.** Horizontal dashed lines show No-OT, Random, and UCB1 escalation rates. Alignment cost decreases monotonically with λ, confirming Theorem 4.2 Part 1. *Left*: IID condition; *right*: Non-IID condition.