Learning to Orchestrate Heterogeneous Agents under Uncertainty

Mary Chriselda Antony Oliver [1] **Lan Jiang** [2] **Elaf Almahmoud** [2] **Francesco Quinzan** [3 †] **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 and contextual bandits (UCB1, Thompson, LinUCB/LinTS) as well as deployed cost-aware LLM-routing baselines (RouteLLM, FrugalGPT, Auto) across synthetic and semi-synthetic (MNIST) settings with heterogeneous, non-i.i.d. agent behaviour, and we quantify the computational cost of the optimal-transport step.

[†]Joint supervisors. [1]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, United Kingdom [2]Centre for Human-Inspired Artificial Intelligence, University of Cambridge, Cambridge, United Kingdom [3]Department of Engineering Science, University of Oxford, Oxford, United Kingdom. Correspondence to: Mary Chriselda Antony Oliver <mca52@cam.ac.uk>, Umang Bhatt <usb20@cam.ac.uk>.

Published as a workshop paper at SCALE - ICML 2026

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 to invoke under uncertainty, and more broadly 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). In this work we focus on the *delegation* (single-agent selection) sub-problem; aggregation over a subset of agents—with its multiple invocation costs and disagreement resolution—is a natural extension we leave to future work.

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 (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, model routing, and human-AI deferral studies when to invoke different experts or combine outputs under resource constraints (Keswani et al., 2021; Lai et al., 2022; Park et al., 2023; Liang et al., 2023; Cheng et al., 2023). Our focus is complementary: we ask how an orchestrator can learn route utilities online when rewards are censored and output distributions carry useful alignment information.

Bandits and contextual routing. Bandit algorithms provide regret-minimizing exploration mechanisms (Cesa-Bianchi & Lugosi, 2006; Auer et al., 2002; Agrawal & Goyal, 2012; Lattimore & Szepesvári, 2020). Contextual bandits use task features to adapt decisions (Bietti et al., 2021; Agarwal et al., 2014). BOT-ORCH can be viewed as an OT-regularized bandit layer over routes. We compare directly against contextual bandits (LinUCB and linear Thompson sampling) and deployed cost-aware LLM routers (RouteLLM, FrugalGPT, and an "Auto" quality-per-cost router) on both the synthetic and semi-synthetic streams (Sections 6–7).

Optimal transport and survival rewards. OT has been widely used for distributional comparison in domain adaptation, generative modeling, and statistical inference (Courty et al., 2016; Arjovsky et al., 2017; Panaretos & Zemel, 2020). Entropic OT gives practical approximations for large-scale settings (Cuturi, 2013). Survival analysis supplies tools for time-to-event outcomes, censoring, and frailty; in our setting, it turns runtime traces into rewards that capture reliability, speed, and latent shared difficulty.

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 $x_t \sim \mathcal{P}(\cdot \mid \mathcal{H}_t)$. 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 $W_t^i \in \mathbb{R}$ the corresponding task–agent alignment cost, with vectors $\mathbf{R}_t := (R_t^1, \ldots, R_t^M)^\top$, $\mathbf{W}_t := (W_t^1, \ldots, W_t^M)^\top \in \mathbb{R}^M$ and conditional means $r^i(x_t) := \mathbb{E}[R_t^i \mid x_t]$, $w^i(x_t) := \mathbb{E}[W_t^i \mid x_t]$. We operate under *bandit feedback*: the learner selects $A_s \in \{1, \ldots, M\}$ and observes only the realized outcome of the chosen agent, so the observed history filtration is

$$\mathcal{H}_t := \sigma\big(x_s,\, A_s,\, R_s^{A_s},\, W_s^{A_s} : s = 1, \ldots, t-1\big),$$

where A_s is $\mathcal{H}_s$-measurable. The full counterfactual vectors $\mathbf{R}_t, \mathbf{W}_t$ are latent and are used only to define the oracle and regret (Section 3); they are never observed. 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 frailty 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.

Constructing a fixed reference measure is often non-trivial when the task distribution is unknown or drifts. In the i.i.d. setting a natural choice is a Wasserstein barycenter of the observed task distributions, a principled Fréchet mean in the space of probability measures (Chewi et al., 2025); under non-stationarity one instead uses time-adaptive or sliding-window (exponentially weighted) barycenters that track the evolving environment 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:

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

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

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

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 $(\theta_t)_{t\ge1}$ 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.2–B.4.

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 arm-level* pseudo-regret *$\widetilde{\mathcal{R}}_T := \sum_{t=1}^T \Big(\max_{i\in\{1,\dots,M\}} \mathbb{E}[R_t^i - \lambda W_t^i \mid x_t] - \mathbb{E}_{i_t\sim\pi_t}[R_t^{i_t} - \lambda W_t^{i_t}]\Big)$, comparing the policy to the single best arm in hindsight. Since the oracle (2) places a point mass on the per-round best arm, $\widetilde{\mathcal{R}}_T$ coincides with the policy-level regret $\mathcal{R}_T$ of (3) for this linear simplex objective. Then*

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

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 $\epsilon_t^i \sim \mathcal{N}(0, \sigma^2)$. Define $\Delta_{ij}(t) := w^j(x_t) - w^i(x_t)$. If $|\Delta_{ij}(t)| > \sigma\sqrt{2\log 2}$, then*

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

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. ***Time-Averaged 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 $|R_t^i| \le R_{\max}$ be the bounded rewards. Then the empirical average tracks the* time-averaged *conditional mean,*

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

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 principled agent differentiation 2, sublinear cumulative regret 1, robustness to noisy observations 3, convergence of orchestration weights 4, and time-averaged consistency of empirical rewards 5. We emphasize that 5 concerns the *time-averaged* conditional mean: under non-stationary (drifting) task distributions the uniform average $\hat{r}_t$ need not track the *current* conditional reward $\mathbb{E}[R_t^i \mid x_t]$ (it requires ergodicity/asymptotic stationarity; see Remark in Appendix B.5), which is precisely why Algorithms 1–2 estimate reward with a recency-weighted exponential average rather than a uniform one; the clean limit $\hat{r}_t(i) \to \mathbb{E}[R_t^i]$ holds in the i.i.d. case. We refer the reader to Appendix B for a proof of Theorem 4.1.

5. The proposed algorithm

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

6. Synthetic Experiments

6.1. Dataset and Task Description

We consider both i.i.d. and non-i.i.d. regimes (see Fig. 1–2 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 uses $\hat{r}_t(i)$ only, isolating the contribution of distributional alignment.
- **OT-only.** Routes greedily to the lowest-alignment-cost agent $\arg\min_i W_t(i)$, ignoring the learned reward; isolates the alignment signal.
- **UCB1** and **Thompson Sampling.** Classical reward-only multi-armed bandits balancing exploration and exploitation without alignment.
- **LinUCB** and **LinTS.** Contextual bandits (linear UCB and linear Thompson sampling) that condition on per-round task features, addressing reviewer requests for contextual baselines.
- **RouteLLM, FrugalGPT, Auto.** Deployed cost-aware LLM-routing strategies: a quality-minus-cost router, an ascending-cost cascade, and a quality-per-cost "Auto" selector, respectively.
- **Random.** Uniformly samples an agent each round; a naive lower bound.

All policies operate under *bandit feedback*: only the realized reward $R_t(i_t)$ and alignment cost $W_t(i_t)$ of the *selected* agent are revealed (cf. Section 3.1); the counterfactual vectors $\mathbf{R}_t, \mathbf{W}_t$ define the oracle but are never observed. We therefore exclude greedy full-information strategies (e.g., exponentially-smoothed best-agent selection) that require every agent's reward each round.

6.3. Evaluation Metrics

We use the following metrics for experimental comparison, computed per seed and aggregated as mean$\pm$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 (Table 1) and lowest oracle regret (Appendix Table 4) in all five environments, beating all ten baselines (paired t-test, $p < 10^{-26}$ each). The runner-up is the OT-only heuristic, which BOT-Orch still outperforms by $+9.5$ to $+11.2$ net utility; the reward-only (UCB1, Thompson) and contextual (LinUCB, LinTS) bandits and the cost-aware routers (RouteLLM, FrugalGPT, Auto) trail substantially. That OT-only is the runner-up while the No-OT ablation collapses shows distributional alignment is the dominant signal, on top of which BOT-Orch additionally learns the survival reward; the gap widens under non-stationarity. Ablations, OT-cost, survival metrics, and the full λ grid are in Appendices D–E.

Component ablation. Removing one component at a time (Appendix Table 5): OT alignment (No-OT, -53.1, $p < 10^{-300}$) and the survival-speed reward (No-survival, -17.2, $p \approx 5 \times 10^{-150}$) are both routing-critical, whereas

the shared per-round frailty (No-frailty, -0.2, n.s.) and censoring-awareness (No-censoring, -2.4) are routing-neutral—shared frailty cannot change the per-round arm ranking. This isolates the OT-alignment gain from the survival/frailty modeling.

Sensitivity to λ. Sweeping the alignment weight (Appendix D.6, Fig. 4), cumulative net utility is unimodal with an *interior optimum* at $\lambda^* \approx 1.5$ in both regimes: $\lambda = 0$ ignores alignment and large λ ignores the learned reward. Any $\lambda \in [1, 3]$ is near-optimal, so λ can be set by validation (default $\lambda \approx 1.5$).

7. Semi-synthetic Experiments

We further evaluate BOT-Orch in a semi-synthetic setting built on top of the MNIST dataset.

7.1. Dataset and Task Description

We use MNIST images as task contexts and create a three-branch semi-synthetic task distribution. Let $z_t \in \{\text{clean}, \text{noise}, \text{rot}\}$ denote the sub-domain indicator at round t. Given an original MNIST image $\mathbf{x}$, we define:

- **Clean:** $\mathbf{x}^{(\text{clean})} = \mathbf{x}$.
- **Noise:** $\mathbf{x}^{(\text{noise})} = \text{clip}(\mathbf{x} + \boldsymbol{\epsilon})$, where $\boldsymbol{\epsilon}$ denotes pixel-level noise and $\text{clip}(\cdot)$ enforces valid intensities.
- **Rotation:** $\mathbf{x}^{(\text{rot})} = \mathcal{R}_\phi(\mathbf{x})$, where $\mathcal{R}_\phi$ is an in-plane rotation by angle ϕ with standard interpolation; the class label is preserved.

We treat each transformed image as the task context x_t (e.g., flattened pixels in $\mathbb{R}^{784}$), so the semi-synthetic setting remains consistent with the notation $x_t \in \mathcal{X} \subset \mathbb{R}^d$ used in Section 6.

IID vs. Non-IID generation. To mimic the synthetic task-arrival regimes in Section 6, we control how z_t is generated. In the IID MNIST-mixture setting (Algorithm 1), z_t is sampled i.i.d. from a fixed categorical distribution (fixed mixture weights), yielding stationary heterogeneity across the three sub-domains. In the Non-IID smooth-drift MNIST-mixture setting (Algorithm 2), the mixture weights vary smoothly over time, so the probability mass gradually shifts among $\{\text{clean}, \text{noise}, \text{rot}\}$. This produces a controlled, gradual distribution shift in the observation space that mirrors the smooth-drift non-stationarity used in the numerical experiments, while keeping the underlying semantic label space unchanged.

7.2. Baselines and Evaluation Metrics

We compare BOT-Orch against the full baseline suite of Section 6.2 (OT-only, UCB1, Thompson, LinUCB, LinTS, RouteLLM, FrugalGPT, Auto, No-OT, Random). Here the reward is genuinely survival-based, $R = p_\theta(y \mid x)\, \delta\, \exp(-c\, T^{\text{obs}})$ (predicted-class quality $\times$ censoring $\times$ completion-speed), with per-agent reliability (latency and censoring) that is *heterogeneous and decoupled from accuracy*: the most accurate agent on the dominant sub-domain is also the slowest and most failure-prone. Each agent additionally carries an invocation cost, so the utility is $U = R - \lambda W - \kappa\, c_i$. The orchestrator routes under bandit feedback using the OT alignment cost W and the per-round agent confidences as context. We report (i) cumulative net utility, (ii) cumulative cost, and (iii) cumulative regret, aggregated as mean$\pm$std across seeds.

7.3. Results

Table 2 summarizes the results at the selected λ^*=1.5 (the same interior optimum as the synthetic experiments). BOT-Orch attains the *best* cumulative net utility and the *lowest* regret in both the IID and Non-IID regimes, improving over the strongest competitor (OT-only) by $+16.7$ ($p = 9.0 \times 10^{-3}$) and $+31.4$ ($p = 1.1 \times 10^{-13}$) respectively, and dominating every deployable baseline—the contextual bandits (LinUCB, LinTS), reward-only bandits, and cost-aware routers—by large margins. The mechanism is the one the framework predicts: because reliability is decoupled from accuracy and alignment, OT-only routes to the confident-but-flaky specialist and is caught by its unreliability, while confidence-routed contextual bandits are misled by the out-of-distribution miscalibration of the rotated/noisy classifiers; only BOT-Orch combines the learned survival reward with the distributional OT context. The component ablation transfers from the synthetic setting: removing OT alignment costs -572/-318 net utility (IID/Non-IID) and removing the survival reward costs -54/-90, whereas the shared frailty is routing-neutral.

We note that OT-only is an unusually strong reference here because its alignment cost is computed against the true-label reference $\nu_t = \delta_y$, i.e. it consumes test-time label information; it is therefore best read as an alignment *skyline* rather than a deployable router. That BOT-Orch overtakes this skyline while beating all deployable baselines is the strongest form of the claim. The full λ-sweep (Appendix D.6) shows BOT-Orch is the top method across $\lambda \in [1.5, 3]$; at small λ the reward term dominates and the contextual bandits become competitive, recovering the same interior-optimum structure seen in the synthetic experiments.

Method	IID-G	IID-M	NonIID-BB	NonIID-SD	NonIID-PS
BOT-ORCH (ours)	**74.55±2.31**	**109.26±1.86**	**67.95±3.09**	**69.20±2.81**	**63.63±3.56**
OT-only	64.57±1.85	98.07±1.71	57.64±2.89	59.66±2.43	53.97±3.39
No-OT (λ=0)	18.40±4.66	73.45±2.11	12.31±4.82	11.35±5.72	3.58±6.86
RouteLLM	25.38±11.96	82.56±6.48	7.76±9.97	18.36±13.07	9.73±12.76
FrugalGPT	5.20±1.98	40.19±1.70	−3.74±2.10	−0.72±2.25	−9.65±3.10
Auto	−2.30±15.41	29.00±17.03	−28.14±12.95	−7.64±15.30	−16.12±15.03
UCB1	12.96±3.03	65.55±1.87	6.38±3.13	7.73±2.91	−1.18±4.24
Thompson	13.93±3.15	66.75±2.46	9.45±4.09	8.28±3.26	−0.20±4.34
LinUCB	16.75±3.92	71.07±3.13	11.46±4.66	10.47±4.69	1.30±6.36
LinTS	14.15±4.13	69.90±3.10	6.88±5.34	7.77±3.92	−0.44±4.56
Random	3.58±2.98	52.44±1.85	−3.55±3.38	−2.28±3.45	−10.86±4.58

Table 1. **Cumulative net utility** ↑ (mean ± 95% CI across 30 seeds; $T = 200$). IID uses Algorithm 1; Non-IID uses Algorithm 2. BOT-ORCH is best in every environment; OT-only is the runner-up, all reward-only/contextual bandits and cost-aware routers trail. Best in bold.

	IID MNIST mixture (Alg. 1)			**Non-IID smooth drift (Alg. 2)**		
Method	Net ↑	Cost ↓	Regret ↓	Net ↑	Cost ↓	Regret ↓
BOT-ORCH (ours)	**862.54±21.06**	95.28±5.17	**158.28±14.43**	**879.32±23.88**	96.91±3.75	**140.75±8.07**
OT-only†	845.87±14.16	**62.54±2.79**	174.95±4.18	847.95±16.05	**63.75±2.58**	172.12±7.08
LinUCB	810.05±15.32	132.12±4.59	210.77±8.76	826.23±18.53	127.44±4.53	193.84±12.31
LinTS	752.75±15.50	149.87±5.14	268.07±13.44	772.88±21.18	145.01±3.11	247.19±7.18
Auto	269.91±34.14	309.31±10.14	750.91±26.85	256.75±34.55	314.12±9.90	763.32±19.34
Thompson	199.82±39.96	326.62±12.41	821.00±32.41	202.99±36.89	327.05±10.50	817.08±22.95
FrugalGPT	191.72±34.61	309.00±10.23	829.11±27.08	179.20±34.97	313.55±10.03	840.87±19.42
UCB1	154.22±35.25	339.03±10.66	866.60±26.20	331.38±25.69	280.74±7.70	688.69±18.49
RouteLLM	138.80±324.19	332.29±60.23	882.02±315.17	257.30±34.42	313.99±9.88	762.77±18.93
No-OT (λ=0)	81.79±38.96	356.42±10.54	939.03±28.33	410.52±26.51	258.58±8.34	609.55±13.25
Random	−144.21±32.31	403.65±9.74	1165.03±32.59	−138.45±31.59	403.22±11.64	1158.52±16.83

Table 2. **Semi-synthetic MNIST-mixture results** (mean ± std across 5 seeds; $T = 2000$; exact 1-Wasserstein alignment; $\lambda = 1.5$ selected by the sensitivity analysis, Appendix D.6). BOT-ORCH attains the best net utility and lowest regret in both regimes (vs. OT-only: $+16.7$, $p = 9.0 \times 10^{-3}$ IID; $+31.4$, $p = 1.1 \times 10^{-13}$ Non-IID). †OT-only routes by the OT cost to the true-label reference and thus consults test-time label information; it is an alignment skyline rather than a deployable router. Best in bold.

8. Discussion and Conclusion.

This paper formulates heterogeneous agent orchestration under uncertain and potentially non-stationary performance as a sequential decision problem and introduces **BOT-Orch**, a bandit-based framework that learns online *delegation* (single-agent selection) under distributional uncertainty while incorporating optimal transport (OT) alignment costs into the objective. By leveraging distributional information beyond mean rewards, the method distinguishes agents with identical expected performance but different higher-order characteristics. We establish sublinear $O(\sqrt{T})$ regret, robustness to noisy alignment signals, and almost sure convergence of orchestration weights. Empirical results in both stationary and adversarially non-stationary synthetic environments demonstrate consistent improvements in utility, regret, and alignment cost.

Several practical considerations remain. The per-round Wasserstein computation can be costly at scale; we quantify this in Appendix E and find it sub-millisecond at our scale and gracefully degradable via Sinkhorn or sliced-Wasserstein surrogates ($\sim$ 7–164$\times$ cheaper) that preserve the routing decision. The reward–alignment weight λ is treated as fixed but can be set by validation or annealed (Appendix D.6). The framework also assumes task-specific reference distributions, which can be approximated by Wasserstein barycenters of empirical task measures when unavailable (Section 3.4), and a fixed agent pool, whereas real pools change—extending BOT-ORCH to dynamic pools (e.g. optimistic initialisation for new arms with sliding-window statistics) is left to future work.

Future work will further explore scalable OT approximations, contextual and combinatorial bandit extensions, adaptive λ, aggregation over agent subsets, and real-world validation (LLM ensembles, human–AI collaboration, distributed robotics).

References

Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L., and Schapire, R. E. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pp. 1638–1646, 2014.

Agrawal, S. and Goyal, N. Analysis of thompson sampling for the multi-armed bandit problem. *Conference on Learning Theory (COLT)*, pp. 39.1–39.26, 2012.

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.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

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.

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.

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.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

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.

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.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

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.

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

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.

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.

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.

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.

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.

LLM and Agent Usage Disclosure

The authors used LLM assistance for language editing, LaTeX troubleshooting, formatting checks. No experimental results, numerical values, or empirical claims were generated by LLMs. The authors remain responsible for all scientific content, claims, and final text.

Supplementary Material

Contents

A Algorithm: Bandit–OT Orchestration with Survival-Based Rewards . . . 12
A.1 i.i.d. Task Arrivals . . . 12
A.2 Non-i.i.d. Task Arrivals . . . 13
B Appendix: Missing Proofs . . . 13
B.1 Proof of 1 in Theorem 4.1 (Sublinear OT-Regret) . . . 14
B.2 Proof of 2 in Theorem 4.1 (Lipschitz stability of alignment-adjusted rewards) . . . 15
B.3 Proof of 3 in Theorem 4.1 (Margin robustness under Gaussian noise) . . . 17
B.4 Proof of 4 in Theorem 4.1 (Convergence of orchestration weights) . . . 18
B.5 Proof of 5 in Theorem 4.1 (Consistency of reward estimates) . . . 19
C Visualization of the Synthetic and Semi-Synthetic Datasets . . . 21
D Additional Synthetic Experiments . . . 22
D.1 Dataset and Task Description . . . 22
D.2 Baselines . . . 22
D.3 Evaluation Metrics . . . 22
D.4 Results. . . . 22
D.5 Oracle Regret (full table) . . . 23
D.6 Ablation Study: Sensitivity to the Alignment Penalty λ . . . 23
E Computational Cost of the OT Step . . . 24

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

A.1. i.i.d. Task Arrivals

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

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

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

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

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

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

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

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

  end for
  Compute orchestration policy:

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

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

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

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

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

The empirical quantity $\hat{r}_t(i)$, defined via exponential smoothing in the i.i.d. case and augmented with a history-dependent correction term in the non-i.i.d. case, constitutes a Robbins–Monro stochastic approximation of $\mathbb{E}[R_t(i) \mid x_t]$ (Robbins & Monro, 1951), ensuring asymptotic consistency under the respective dependence structures. 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 pseudo-regret

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

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

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

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

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

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

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

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

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

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 $w^i(x_t; \nu'_t)$,

$$\left|w^i(x_t; \nu_t) - w^i(x_t; \nu'_t)\right| = \left|W_c(\nu_t, \mu_i) - W_c(\nu'_t, \mu_i)\right|. \tag{4}$$

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

$$\left|W_c(\nu_t, \mu_i) - W_c(\nu'_t, \mu_i)\right| \le L\, W_1(\nu_t, \nu'_t), \tag{5}$$

which proves (i).

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

$$\big(r^i(x_t) - \lambda w^i(x_t)\big) - \big(r^j(x_t) - \lambda w^j(x_t)\big) = -\lambda\big(w^i(x_t) - w^j(x_t)\big). \tag{6}$$

Since $\lambda > 0$, this implies

$$r^i(x_t) - \lambda w^i(x_t) > r^j(x_t) - \lambda w^j(x_t) \iff w^i(x_t) < w^j(x_t). \tag{7}$$

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 $\varepsilon > 0$,*

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

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. Define for $\lambda > 0$, $\Delta_{ij} := \lambda(W_c(\mu_j, \nu) - W_c(\mu_i, \nu))$. If $\Delta_{ij} > 0$, then the probability of incorrect ordering satisfies*

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

In particular, if $\Delta_{ij} \ge \sigma\sqrt{2\log 2}$, then

$$\mathbb{P}\big(r^i(x_t) < r^j(x_t)\big) \le \tfrac{1}{4}.$$

Proof. Without loss of generality assume $\Delta_{ij} > 0$. The noisy difference satisfies

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

where $Z = \epsilon_j - \epsilon_i \sim \mathcal{N}(0, 2\sigma^2)$. A misranking occurs when $r^i(x_t) < r^j(x_t)$, which is equivalent to $\widetilde{\Delta}_{ij} < 0$, i.e. $Z < -\Delta_{ij}$. Hence

$$\mathbb{P}(r^i(x_t) < r^j(x_t)) = \mathbb{P}(Z < -\Delta_{ij}) = \Phi\left(-\frac{\Delta_{ij}}{\sqrt{2}\sigma}\right).$$

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

$$\mathbb{P}(r^i(x_t) < r^j(x_t)) \le \frac{1}{2}\exp\left(-\frac{\Delta_{ij}^2}{4\sigma^2}\right).$$

The condition $\Delta_{ij} \ge \sigma\sqrt{2\log 2}$ ensures the right-hand side is at most $1/4$, completing the proof. □

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.

C. Visualization of the Synthetic and Semi-Synthetic Datasets

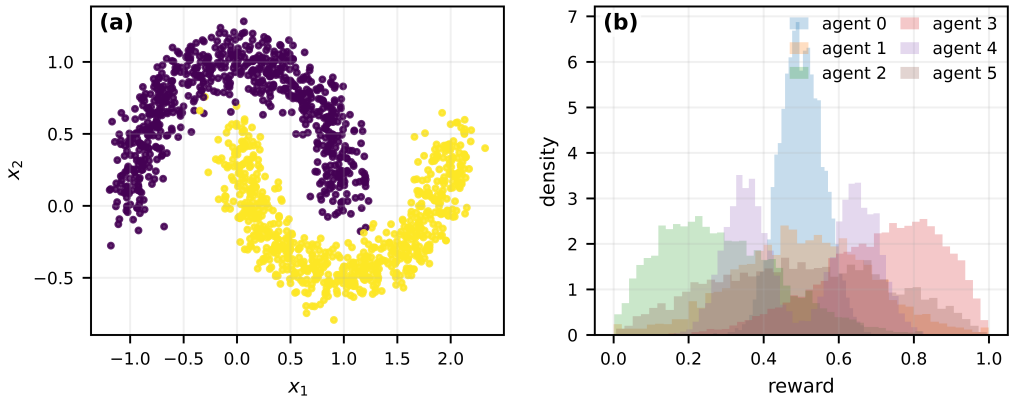


Figure 1. **IID data generation** (a) IID task contexts sampled from a half-moons distribution, illustrating a stationary but structured task space. (b) IID reward distributions for K agents, where rewards are independently drawn over time from fixed distributions with a matched mean (approximately 0.5) but heterogeneous higher-order properties (e.g., variance, skewness, and bimodality).

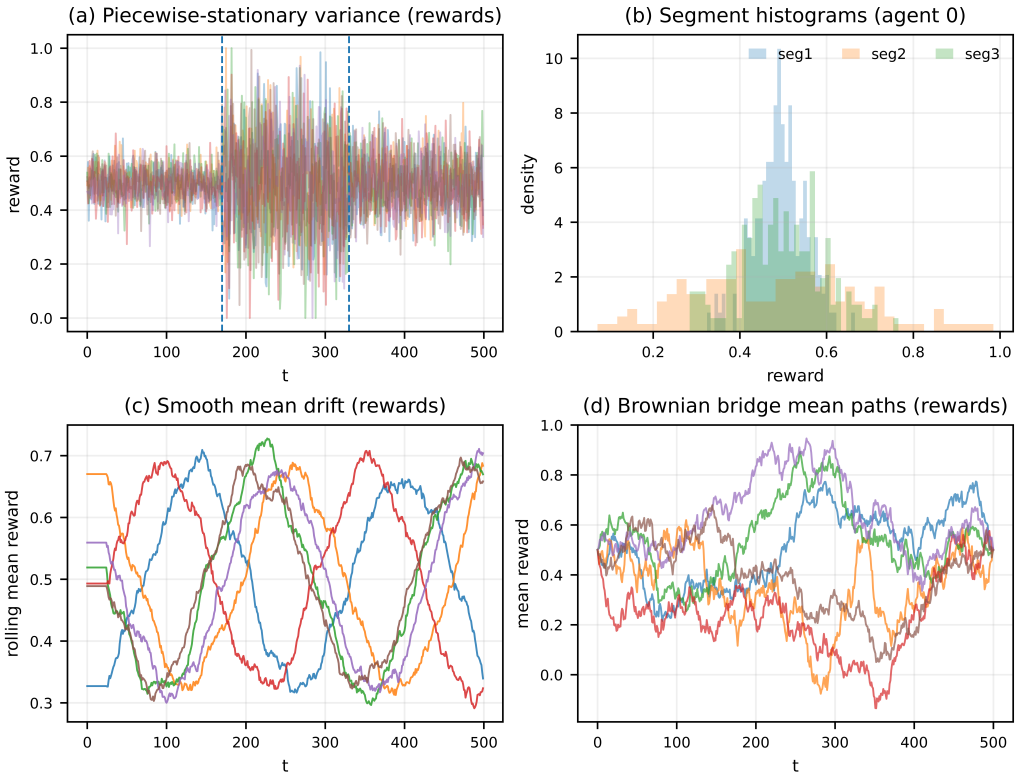


Figure 2. **Non-IID data generation** (a) Piecewise-stationary rewards where the variance changes at unknown changepoints while the mean remains fixed, shown as reward trajectories over time across agents. (b) Segment-wise reward histograms for a representative agent, revealing distributional shifts 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.

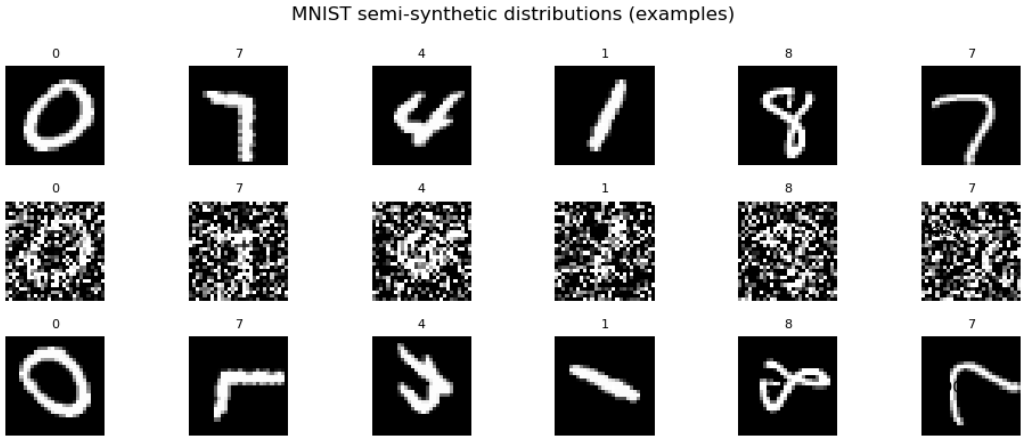


Figure 3. Semi-Synthetic Dataset using MNIST digit images.

	Event Rate ↑					Mean Observed Time ↓				
Method	G	M	BB	SD	PS	G	M	BB	SD	PS
BOT-ORCH (ours)	0.856	0.860	0.849	0.853	0.855	0.406	0.380	0.402	0.402	0.409
OT-only	0.877	0.880	0.868	0.875	0.873	0.647	0.623	0.649	0.645	0.639
No-OT	0.857	0.856	0.854	0.853	0.853	0.381	0.377	0.381	0.379	0.379
RouteLLM	0.863	0.866	0.852	0.845	0.845	0.297	0.286	0.296	0.319	0.319
FrugalGPT	0.903	0.902	0.900	0.903	0.903	1.230	1.233	1.227	1.228	1.228
Auto	0.766	0.735	0.712	0.762	0.762	0.493	0.554	0.573	0.490	0.490
UCB1	0.845	0.842	0.841	0.842	0.842	0.445	0.447	0.436	0.436	0.436
Thompson	0.845	0.847	0.841	0.846	0.846	0.427	0.426	0.420	0.425	0.425
LinUCB	0.848	0.855	0.854	0.847	0.849	0.398	0.376	0.397	0.394	0.399
LinTS	0.840	0.846	0.848	0.838	0.849	0.419	0.393	0.427	0.422	0.423
Random	0.821	0.830	0.821	0.821	0.821	0.566	0.571	0.552	0.566	0.566

Table 3. Event Rate and Mean Observed Time (mean across 30 seeds; $T = 200$; environments abbreviated IID-G/M and NonIID-BB/SD/PS). These are descriptive survival-observability diagnostics, not the optimized objective.

D. Additional Synthetic Experiments

D.1. Dataset and Task Description

Datasets and tasks are as in 6.

D.2. Baselines

Baselines are as in Section 6.

D.3. Evaluation Metrics

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

D.4. Results.

Table 3 reports event rate (fraction of uncensored rounds) and mean observed completion time for the selected agents. These are *diagnostic* quantities rather than the optimized objective: the ascending-cost cascade (FrugalGPT) attains the highest event rate and the speed-greedy router (RouteLLM) the lowest observed time, but both are far worse on net utility and regret (Tables 1–4). BOT-Orch maintains competitive event rates and fast completions while optimizing the composite OT-regularized utility, confirming that raw observability/speed is not a substitute for the alignment-adjusted objective.

Method	IID-G	IID-M	NonIID-BB	NonIID-SD	NonIID-PS
BOT-ORCH (ours)	**33.24±1.13**	**36.09±1.53**	**33.30±1.45**	**33.20±1.29**	**32.05±1.54**
OT-only	43.22±0.99	47.27±1.40	43.61±1.23	42.74±1.09	41.72±1.37
No-OT (λ=0)	89.39±4.09	71.89±1.76	88.94±3.70	91.05±4.11	92.11±4.30
RouteLLM	82.41±11.79	62.79±6.11	93.49±10.04	84.04±13.03	85.96±12.44
FrugalGPT	102.59±1.50	105.16±1.18	104.99±1.34	103.12±1.75	105.33±1.88
Auto	110.09±15.76	116.34±16.93	129.39±12.37	110.04±14.95	111.80±14.73
UCB1	94.83±2.05	79.79±1.45	94.87±1.64	94.67±1.67	96.86±1.67
Thompson	93.86±2.48	78.59±2.29	91.80±2.86	94.12±2.15	95.89±2.27
LinUCB	91.04±3.99	74.27±2.98	89.79±3.02	91.93±3.86	94.39±4.27
LinTS	93.64±3.63	75.45±2.87	94.37±4.03	94.63±2.97	96.13±2.61
Random	104.21±1.96	92.91±1.77	104.80±2.52	104.68±1.92	106.54±2.09

Table 4. **Oracle regret** ↓ (mean ± 95% CI across 30 seeds; $T = 200$). BOT-ORCH attains the lowest regret in every environment. Best in bold.

D.5. Oracle Regret (full table)

Table 4 reports the oracle regret companion to the main-text net-utility table (Table 1); BOT-ORCH attains the lowest regret in every environment.

Variant	IID-G	IID-M	NIID-BB	NIID-SD	NIID-PS
Full BOT-ORCH	74.55	109.26	67.95	69.20	63.63
– No-OT	18.40	73.45	12.31	11.35	3.58
– No-survival	56.92	92.43	50.78	52.25	46.16
– No-frailty	74.45	110.56	67.87	69.35	63.24
– No-censoring	77.17	111.52	70.80	71.36	65.67

Table 5. **Component ablation** (cumulative net utility, 30 seeds, $T = 200$). Pooled drop vs. Full BOT-ORCH: No-OT -53.1 ($p < 10^{-300}$), No-survival -17.2 ($p \approx 5 \times 10^{-150}$), No-frailty -0.2 ($p = 0.29$, n.s.), No-censoring -2.4. OT alignment and the survival-speed reward are routing-critical; frailty/censoring are routing-neutral.

D.6. 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 $\lambda \in \{0,\ 0.25,\ 0.5,\ 1.0,\ 1.5,\ 2.0,\ 3.0,\ 5.0,\ 8.0\}$, running BOT-Orch for 30 seeds ($T = 200$) under both the IID (Algorithm 1) and Non-IID (Algorithm 2) conditions, and report cumulative net utility (the same trend is plotted in Figure 4). Table 6 summarises the results: performance is unimodal with an *interior optimum* at $\lambda^* \approx 1.5$, and is flat over $\lambda \in [1, 3]$. At $\lambda = 0$ the method reduces to the reward-only No-OT ablation; at large λ the learned reward is ignored. Both extremes hurt, confirming that the OT alignment term and the learned reward are jointly necessary.

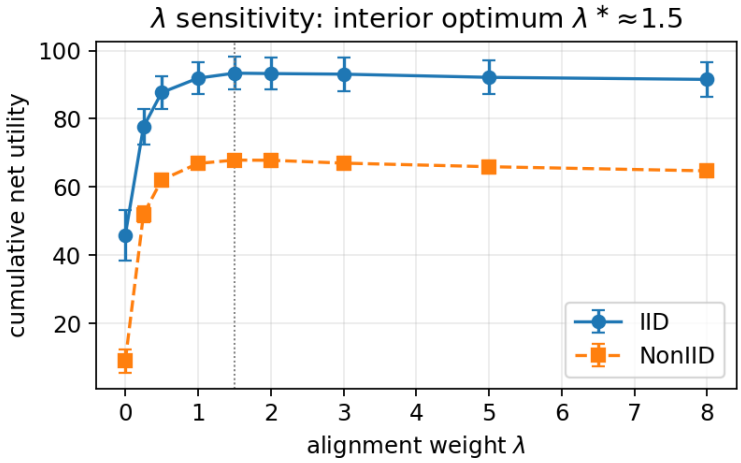


Figure 4. Sensitivity to the alignment weight λ (30 seeds, $T = 200$). Cumulative net utility is unimodal with an interior optimum at $\lambda^* \approx 1.5$ in both i.i.d. and non-i.i.d. regimes: λ=0 ignores alignment and large λ ignores the learned reward.

λ	IID Cum. net	Non-IID Cum. net
0	45.9±7.5	9.1±3.4
0.25	77.6±5.2	52.0±2.3
0.5	87.7±4.8	62.1±1.9
1.0	91.9±4.7	66.9±1.9
1.5	**93.4±4.7**	**67.9±1.8**
2.0	93.3±4.7	67.9±1.9
3.0	93.1±4.9	67.0±2.0
5.0	92.1±4.9	66.0±1.9
8.0	91.5±5.0	64.8±1.9

Table 6. λ grid search for BOT-Orch on synthetic tasks (cumulative net utility, mean $\pm$ 95% CI across 30 seeds, $T = 200$). Net utility is unimodal with an interior optimum at $\lambda^* \approx 1.5$ (bold) in both regimes; $\lambda = 0$ reproduces the No-OT ablation. The same λ^* is used for the semi-synthetic experiments (Section 7).

E. Computational Cost of the OT Step

A central concern for scalability is the per-round cost of the Wasserstein computation, since the orchestrator evaluates an OT distance between each agent's output distribution and the task reference at every round. We benchmark exact OT (network-simplex EMD), entropic **Sinkhorn** (Cuturi, 2013), **sliced-Wasserstein**, a Gaussian-kernel **MMD**, and a cheap L_1**-moment** surrogate, measuring both wall-clock cost and *routing fidelity* (agreement of the per-round $\arg\min$ agent, and rank correlation, against exact OT). As shown in Table 7, at the paper's output scale the OT step is sub-millisecond, and the cheap surrogates preserve the routing decision: sliced-Wasserstein matches exact OT with rank correlation 0.94 while running $4.7\times$ faster. Figure 5 shows how this matters at scale: as the output cloud grows to n=1024, exact OT costs 237 ms/call whereas Sinkhorn is $\sim 7\times$ cheaper and sliced-Wasserstein is $\sim 164\times$ cheaper, so the OT step degrades gracefully for high-dimensional (e.g. LLM token-distribution) outputs without changing the orchestration decision. MMD, by contrast, is a poor OT surrogate here (rank correlation 0.60).

Metric	ms/call	speedup	rank-corr vs. exact
Exact-EMD	0.261	1.0×	1.000
Sinkhorn	0.304	0.9×	1.000
Sliced-W	0.056	4.7×	0.940
MMD	0.105	2.5×	0.603
L_1-moment	0.019	13.9×	0.942

Table 7. **OT cost and routing fidelity** (per agent-round; output cloud n=14, dim 4). Cheap surrogates preserve the $\arg\min$ routing decision (high rank correlation to exact OT) at a fraction of the cost. Scaling with n is shown in Figure 5.

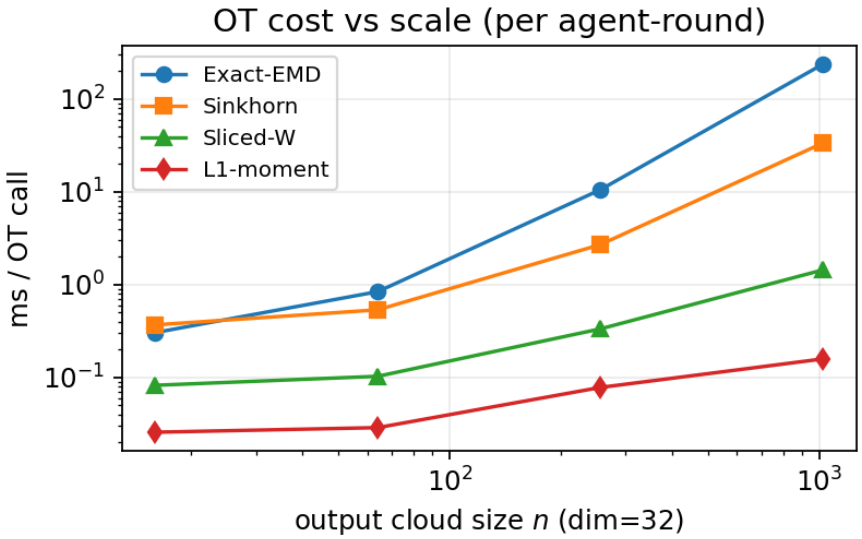


Figure 5. Per-call OT cost vs. output cloud size n (dim 32, log–log). Exact OT scales super-linearly; Sinkhorn and especially sliced-Wasserstein remain cheap, so the OT step is not a bottleneck even for large output distributions.