TEACH ME NO REGRET: STATISTICAL ANALYSIS OF ACCESSIBILITY AND REALIZABILITY IN REGRET-BASED POST-TRAINING.

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models are increasingly used in interactive settings where they must repeatedly act, observe feedback, and adapt their behavior over time, with a central challenge in such tasks being the learning of effective exploration–exploitation strategies. Regret-minimization fine-tuning (RMFT) has been proposed to address this issue, but depends both on generating useful trajectories and on reproducing their behavior with the information available at deployment. In this paper, we study those two requirements. For regret distributions with arbitrary atoms and ties, we derive a finite-iteration bound combining initial lower-tail accessibility with a selection-to-deployment discrepancy. A finite-sample result separates different error sources, while margin conditions translate regret into rollout-averaged action error. We also solve repeated population imitation in a binary model with noisy observations: the policy converges geometrically to the best admissible decision rule, although an information-dependent gap to the evaluator's oracle persists. Experiments with a lightweight language model in stochastic, contextual, and language-based recommendation bandits show contrasting responses to direct RMFT and teacher-distilled initialization. The resulting framework distinguishes the quality of selected trajectories from the performance achieved by the deployed policy and provides a basis for interpreting initialization-dependent post-training gains.

1 INTRODUCTION

Large Language Models (LLMs) have been used for a variety of tasks, ranging from text summarization (Goyal et al. (2022)) and code generation (Chen et al. (2021b)) to medical reasoning (Singhal et al. (2023)), scientific discovery (Boiko et al. (2023)), data interpretation (Shen et al. (2025)), and social and political modeling (Helm et al. (2026)). As many real-world decision-making problems are naturally specified and experienced through language, LLMs are natural candidates for acting as decision agents in such environments. However, when used off-the-shelf, LLMs often underperform, even for simple online decision tasks, exhibiting poor exploration and exploitation behavior. Recent literature proposes distillation (Nie et al. (2025); Laskin et al. (2022)) and reinforcement learning-based (Schmied et al. (2025)) post-training methods to enhance the performance of LLMs in decision making, although these methods cannot adequately generalize to purely language-grounded tasks.

In Park et al. (2025a), the authors propose a framework for post-training LLMs for language-grounded decision tasks using low-regret model-generated trajectories, managing to link regret-based learning with the sampling-based generative process of an LLM. However, its effectiveness raises a question that is especially relevant in sequential environments: *When does selecting better trajectories produce a better deployed policy ?* Two requirements must be distinguished. First, the current proposal must generate useful trajectories frequently enough for the available sampling budget. We call this *low-regret accessibility*. Second, the decisions represented in the selected data must be reproducible by a policy with the information available at deployment. Selection can favor trajectories with lucky early actions or favorable external observations, while limited policy expressivity, fitting error, and generalization across scenarios can further impede this transfer.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We analyze the statistical and information constraints of the scenario-wise RMFT procedure introduced by Park et al. (2025a). Our contribution has three parts. First, we give an atom-compatible finite-iteration bound that separates concentration of the initial regret distribution's lower tail from errors in transferring selected behavior to deployed policies. Second, we bound the effects of finite trajectory samples and noisy scores and translate regret bounds into contextual action error under margin conditions. Third, we derive the exact dynamics of repeated population imitation with noisy observations, establishing geometric convergence to the best admissible policy together with an irreducible gap to the informed oracle. Accessibility guarantees for exploratory mixtures and distillation connect the analysis to proposal design. The bandit experiments document how direct RMFT and distilled initialization perform across three environments.

1.1 Related Work

1.1.1 Language Models in Decision-Making Environments

A growing literature studies LLMs as autonomous agents that reason, plan, and act in interactive environments. Early agentic frameworks such as ReAct (Yao et al. (2022)) and Reflexion (Shinn et al. (2023)) combine language-based reasoning with sequential interaction and feedback, while related work has explored LLM-driven planning and embodied decision-making in open-world environments (Zihao et al. (2024); Wang et al. (2023)). More application-oriented systems use LLM agents for software engineering (Yang et al. (2024a); Wang et al. (2025)), enterprise and knowledge-work tasks (Drouin et al. (2024)), medical decision-making (Kim et al. (2024)), and cybersecurity (Zhang et al. (2024)). Despite this progress in complex application domains, the fundamental decision-making behavior of LLMs remains poorly understood, motivating recent work in controlled online learning environments. Regarding the native decision-making performance of LLMs, the authors in Krishnamurthy et al. (2024) study LLMs directly in stochastic bandits and find that robust exploration generally does not emerge without substantial prompting, requiring both chain-of-thought and externally summarized interaction statistics. The generalization from stochastic exploration to regret in online learning and repeated games is studied in Park et al. (2025b). It has been shown that LLMs can learn online decision algorithms from reward feedback in classical (Lin et al. (2024)) and in contextual (Monea et al. (2024)) settings under implicit reasoning limitations, while the generalization of implicit learning to two-player zero-sum games was presented in Shi et al. (2024) and to matrix and Markov zero-sum games in Dong et al.. These results suggest that although LLMs showcase non-trivial sequential adaptation, pre-training alone is not enough to lead to effective exploration-exploitation behavior.

A complementary line of work studies how decision-making structure can be introduced through prompting or post-training. Algorithm Distillation (Laskin et al. (2022)) trains sequence models on learning histories generated by reinforcement-learning algorithms. Building on this idea, the authors in Nie et al. (2025) improve the exploration performance of LLMs through algorithm-guided distillation using trajectories generated by established exploration strategies. Other approaches avoid relying on a fixed teacher algorithm, such as PAPRIKA (Tajwar et al. (2025)) that fine-tunes LLMs on synthetic interaction data. Alternatively, reinforcement learning or selective supervision can be used to improve sequential behavior. In Schmied et al. (2025), the authors identify sources of poor decision-making behavior, and use reinforcement-learning fine-tuning on self-generated reasoning trajectories to mitigate them. The authors in Chen et al. (2026) study both SFT- and RL-based bandit training and show that improvements in average regret can induce an excessive exploitation bias. Related agent-tuning methods use privileged teacher feedback (Choudhury & Sodhi (2025)) or selectively fine-tune on critical decision steps from expert trajectories (Chen et al. (2025)). Most closely related to our work, the RMFT process is introduced in Park et al. (2025a), which iteratively selects the model's own low-regret trajectories for supervised fine-tuning, allowing post-training to operate directly on model-generated reasoning.

1.1.2 Statistical Theory for Transformer Language Models

Recent literature treats LLMs as stochastic black-box models that can be studied using statistical inference, rather than simply large neural networks. This includes Duderstadt et al. (2023), where the authors propose comparing foundation models through the geometry of their embeddings, while the first consistency conditions for such model-level representations were presented in Acharyya et al. (2024). This was extended to real model-level inference tasks in Helm et al. (2024a). The crucial

point of this line of work is that the statistical target is the model itself as an observable random object, and not its interior. In the same direction, Helm et al. (2023) provide a formal statistical pattern recognition framework to examine whether human and machine-generated content are distinguishable under context. This idea was later transferred from single-model testing to collections of interacting LLMs in Helm et al. (2024b). For black-box APIs, this perspective takes a more classical frequentist form. Gao et al. (2025) formulate model-distortion detection as a two-sample testing problem between API outputs and a reference distribution, while Zhu et al. (2025) introduce query-efficient rank-based tests. Related approaches use response orderings for contamination testing (Oren et al. (2024)) and detect changes under arbitrary model interventions through the DBPA framework (Rauba et al. (2025)).

The theory of in-context learning also provides the conceptual background for treating an LLM not only as a text generator, but as a statistical estimation machine. Akyürek et al. (2022) show that transformers can implement gradient descent and closed-form ridge regression in linear problems, while trained in-context learners can approximate least-squares, ridge, and even Bayesian estimators. Following this, it was proven in Von Oswald et al. (2023) that a linear self-attention layer is equivalent to a gradient-descent step, while multiple layers learn deep representations for non-linear regression. In Bai et al. (2023), the authors show that a transformer can choose in-context a different algorithm for each sequence, reaching nearly Bayes-optimal behavior in mixed-noise linear models, while this concept was extended to Bayesian inference for generalized linear models and latent factor models in Reuter et al. (2025), with posterior samples comparable to Markov chain Monte Carlo and variational inference. Related work then extends this algorithmic view of transformers to sequential decision-making problems, with Chen et al. (2021a) showing that reinforcement learning can be cast as conditional sequence modeling, which was later refined in Laskin et al. (2022) and Lin et al. (2024), and was extended in bandit problems (Mukherjee et al. (2024)) and games (Shi et al. (2024); Dong et al.), as mentioned previously.

Recent work has also begun to study post-training procedures directly at the level of the distribution induced by model outputs. In Sessa et al. (2025), Best-of-N Distillation (BOND) is formulated as a distribution-matching procedure that moves the model toward the distribution induced by selecting high-reward generations, while Yang et al. (2024b) provide a theoretical analysis of iterative Best-of-N distillation and its sample efficiency. This is closely related to classical elite-sample methods such as the cross-entropy method (Rubinstein (1999)). In sequential decision-making environments, the relevant statistical target is not simple concentration on high-reward outputs, but recovery of an optimal context-dependent action rule. Classical contextual-bandit theory studies this connection through margin conditions that control how reward or value estimation error translates into decision error and regret (Rigollet & Zeevi (2010); Bastani et al. (2021); Hu et al. (2020)).

Finally, the distinction between selected and deployed behavior is closely related to return-conditioned supervised learning (RCSL). Paster et al. (2022) show that conditioning on fortunate outcomes can prescribe behavior whose performance is not reproduced under the true environment dynamics, while Brandfonbrener et al. (2022) identify assumptions for optimality of RCSL and examples demonstrating their necessity. Building on this distinction, we study repeated scenario-wise regret selection through the discrepancy between selected and deployed regret laws. Evaluating expected reward gaps removes realized reward noise from the score, while selection can still bias exogenous observations within a hidden scenario. Section 4.1 isolates this effect and derives the resulting population imitation dynamics.

2 Decision policies and regret-based selection

2.1 Natural Language-Based Contextual Bandit Environments

A contextual multi-armed bandit (MAB) (Barto et al. (1981); Abe & Long (1999); Langford & Zhang (2007)) environment is characterized by two spaces, a feature space X that contains information for both the user and an action, called the context space, and a space of actions $\mathcal{A}$. A detailed background of algorithms used for numerical contextual MAB environments is provided in Tanković & Šajina (2025). These methods include upper confidence bound (UCB) algorithms (Li et al. (2010); Chu et al. (2011); Filippi et al. (2010); Zhou et al. (2020)), which construct upper confidence bounds for each arm's reward, ϵ-greedy methods (Sutton et al. (1998); Langford & Zhang (2007)) that probabilisti-

cally balance exploration and exploitation, posterior sampling (or Thompson sampling) techniques (Chapelle & Li (2011); Agrawal & Goyal (2013); Dumitrascu et al. (2018)) that parameterize the reward model and maintain the posterior over these parameters, and generalized linear bandits (Zhang & Sugiyama (2023); Foster & Krishnamurthy (2021)) that can model non-linear relationships between context and expected rewards through the appropriate link function.

In a language-grounded contextual bandit environment, at round $t \leq T$, the agent observes a signal of context X_t via a natural language description and history H_t, encoded by a prompt $q_t = \Psi(X_t, H_t)$. A language-model policy induces $\pi_\theta(a \mid q_t)$ on a finite action set $\mathcal{A}$. Any latent preferences unavailable from the observed textual representation of X_t remain part of the environment. Indicative examples of such contextual environments can be found in Monea et al. (2024); Baheri & Alm (2023); Harris & Slivkins (2025). Conditional on an environment instance $S \sim \rho$, rewards have means $\mu_s(x, a) \in [0, 1]$. Then, an optimal action is characterized as $a_s^*(x) \in \arg\max_a \mu_s(x, a)$, and we define:

$$\Delta_s(x, a) = \mu_s(x, a_s^*(x)) - \mu_s(x, a), \qquad \overline{R}_s(\tau) = \frac{1}{T}\sum_{t=1}^{T} \Delta_s(X_t, A_t) \in [0, 1]. \tag{1}$$

The corresponding cumulative pseudo-regret is $T\,\mathbb{E}[\overline{R}_S]$. This benchmark uses the instance-specific oracle, which can have more information than an admissible policy. The selection and transfer results also apply to any measurable score in $[0, 1]$, including regret defined using evaluator-only latent variables recorded in the full trajectory. The action-law conversion in Section 3.3 specifically requires gaps of the form (1) on the stated context space.

Algorithm 1: RMFT with selection within each scenario (Park et al., 2025a)

Input: Initial policy θ_0, iterations J, scenarios per iteration M, rollouts L, retained trajectories k, and a scoring rule.

1. For $j = 0, \ldots, J-1$, initialize a training dataset D_j.
2. For each sampled or prescribed scenario s_i, generate L independent rollouts from $P_{\theta_j}^{s_i}$.
3. Within that scenario, retain exactly k rollouts with smallest scores, breaking ties uniformly; add their model outputs and prompts to D_j.
4. After processing all M scenarios, fit the shared policy θ_{j+1} to D_j by supervised fine-tuning, then continue to the next iteration.

2.2 An elite operator that allows ties

For fixed $\alpha \in (0, 1)$ and law ν on $[0, 1]$, we denote by F_ν its CDF and $q_u(\nu) = \inf\{r : F_\nu(r) \geq u\}$ for $u \in (0, 1]$. For a trajectory law P and $c = q_\alpha(\nu_P)$, we define

$$w_{\alpha,P}(\tau) = \mathbf{1}\{\overline{R}(\tau) < c\} + \lambda_{\alpha,P}\mathbf{1}\{\overline{R}(\tau) = c\}, \quad \lambda_{\alpha,P} = \frac{\alpha - P(\overline{R} < c)}{P(\overline{R} = c)}. \tag{2}$$

When the denominator is zero, we set $\lambda_{\alpha,P} = 0$, implying $P(\overline{R} < c) = \alpha$. The elite law is: $\mathsf{T}_\alpha P(\,\mathrm{d}\tau) = \alpha^{-1} w_{\alpha,P}(\tau) P(\,\mathrm{d}\tau)$. Thus, exactly an α fraction of population mass is retained, including a suitable fraction of any boundary atom and ties are randomized independently of trajectory content. The operator T_α can be considered as the population analogue of the scenario-wise selection step in Algorithm 1.

The induced regret law, denoted $\mathsf{B}_\alpha\nu$, with B_1 defined as the identity, satisfies:

$$F_{\mathsf{B}_\alpha\nu}(r) = \min\{F_\nu(r)/\alpha, 1\}, \qquad q_u(\mathsf{B}_\alpha\nu) = q_{\alpha u}(\nu) \quad \text{for almost every } u. \tag{3}$$

Applying this operation separately to P_θ^s preserves the scenario weights ρ, unlike conditioning the mixture $\int P_\theta^s \rho(\,\mathrm{d}s)$ on one global regret threshold. At the population level, supervised imitation minimizes the expected negative model log-likelihood under $\int \mathsf{T}_\alpha P_\theta^s \rho(\,\mathrm{d}s)$, within the chosen policy class, i.e., it is a projection onto policies that preserve the environment dynamics and observation constraints. Pointwise consistency of the empirical likelihood objective with random tie-breaking is established in Appendix B.

3 FROM SELECTED TRAJECTORIES TO DEPLOYED DECISIONS

In this section we quantify how closely the regret distribution of the next deployed policy after fine-tuning tracks the ideal elite-selected regret distribution, and propagate that discrepancy across RMFT iterations to bound final regret and decision error. Proofs of the theoretical results are provided in Appendices A–C.

Throughout the section, for laws on $[0, 1]$, we use the one-dimensional Wasserstein distance, given by:

$$W_1(\nu, \xi) = \int_0^1 |F_\nu(r) - F_\xi(r)| \, \mathrm{d}r = \int_0^1 |q_u(\nu) - q_u(\xi)| \, \mathrm{d}u. \tag{4}$$

The following result establishes how elite selection transforms regret distributions.

Lemma 3.1 (Stability and repeated selection)**.** *For arbitrary regret laws, including laws with atoms,*

$$W_1(\mathsf{B}_\alpha \nu, \mathsf{B}_\alpha \xi) \le \alpha^{-1} W_1(\nu, \xi), \qquad \mathsf{B}_\alpha^j \nu = \mathsf{B}_{\alpha^j} \nu.$$

Moreover, the mean of $\mathsf{B}_u \nu$ *is* $u^{-1} \int_0^u q_v(\nu) \, \mathrm{d}v$.

3.1 A CONDITIONAL FINITE-ITERATION GUARANTEE

For $\widehat{\nu}_j^s = \nu_{\theta_j}^s$ the actual deployed regret law, we define the ideal *regret-law reference*: $\nu_j^{\circ,s} = \mathsf{B}_{\alpha^j} \nu_{\theta_0}^s$, which does not need to be realizable by an admissible policy. We then set:

$$d_j = \mathbb{E}_{s\sim\rho}\left[W_1\left(\widehat{\nu}_{j+1}^s, \mathsf{B}_\alpha \widehat{\nu}_j^s\right)\right], \qquad A_0(u) = \mathbb{E}_{s\sim\rho}[q_u(\nu_{\theta_0}^s)]. \tag{5}$$

The residual d_j measures selection-to-deployment transfer and includes restrictions imposed by information and the policy class, as well as different error sources, e.g., estimation or imperfect optimization. There is also a connection to supervised fitting. By letting $\mathcal{A}_j$ be the smallest population trajectory KL divergence from the elite targets attainable by the shared policy class, then if empirical imitation loss approximates its population objective uniformly within g_j and is minimized within o_j, we have:

$$d_j \le \sqrt{(\mathcal{A}_j + 2g_j + o_j)/2}. \tag{6}$$

Appendix A.5 states the likelihood assumptions and proves this bound.

The following theorem bounds final deployed regret by initial low-regret accessibility plus accumulated selection-to-deployment errors.

Theorem 3.1 (Regret transfer at a fixed horizon)**.** *For any sequence of admissible policies starting at* θ_0 *and any integer* $J \ge 0$ *(with an empty error sum when* $J = 0$*),*

$$\mathbb{E}_{s,\tau\sim P_{\theta_J}^s}[\overline{R}_s(\tau)] \le \frac{1}{\alpha^J} \int_0^{\alpha^J} A_0(u) \, \mathrm{d}u + \sum_{j=0}^{J-1} \alpha^{-(J-1-j)} d_j. \tag{7}$$

The first term measures the quality available in the initial lower tail, while the second quantifies the failure to reproduce successive selected targets. For a quantitative accessibility envelope, suppose

$$A_0(u) \le r_{\mathrm{acc}} + K u^{1/\beta} \quad (0 < u \le u_0), \qquad K, \beta > 0, \quad r_{\mathrm{acc}} \ge 0, \quad u_0 \in (0, 1]. \tag{8}$$

Then for $\alpha^J \le u_0$, the first term in (7) is at most $r_{\mathrm{acc}} + \frac{\beta}{\beta+1} K \alpha^{J/\beta}$. The offset permits an initially inaccessible neighborhood of zero. If instead each initial law has a zero-regret atom of mass at least $p_* > 0$, the ideal term is exactly zero once $\alpha^J \le p_*$.

If a uniform residual bound is available, the decaying accessibility term and accumulated error define a computable upper-bound criterion for choosing an iteration count, derived in Appendix A.3.

3.2 FINITE TRAJECTORY SAMPLES AND NOISY SCORES

We now give an explicit specialization for a fixed collection of M scenarios, with ρ uniform on them. For $\alpha = k/L$, at iteration j and scenario s, we sample L independent rollouts with exact regrets

$R_{j,s,\ell}$, rank clipped scores $Z_{j,s,\ell} \in [0,1]$, and let $\widetilde{\nu}_{j,s}$ be the empirical law of the *true* regrets of the k selected trajectories. We then define:

$$e_j^{\mathrm{fit}} = \frac{1}{M}\sum_s W_1(\widehat{\nu}_{j+1}^s, \widetilde{\nu}_{j,s}), \qquad e_j^{\mathrm{score}} = \frac{1}{ML}\sum_{s,\ell} |Z_{j,s,\ell} - R_{j,s,\ell}|.$$

Theorem 3.2 (Finite-sample residual bound)**.** *Fix $J_{\max} \geq 1$ and $\delta \in (0,1)$. Conditional on each iteration's past, assume the rollouts within every scenario are i.i.d. from the current policy. With probability at least $1-\delta$, simultaneously for $j < J_{\max}$,*

$$d_j \leq e_j^{\mathrm{fit}} + \frac{1}{\alpha}\sqrt{\frac{\log(2MJ_{\max}/\delta)}{2L}} + \frac{2}{\alpha}e_j^{\mathrm{score}}. \tag{9}$$

Substituting these residual bounds into Theorem 3.1 gives a simultaneous guarantee over the stated iteration range.

For an additional probabilistic score bound, suppose the clipped errors are independent across trajectories conditional on the sampled rollouts and past, with conditional mean absolute error at most m_n. Hoeffding's inequality and a union bound yield, with probability at least $1-\delta_{\mathrm{score}}$,

$$e_j^{\mathrm{score}} \leq m_n + \sqrt{\frac{\log(MJ_{\max}/\delta_{\mathrm{score}})}{2L}} \quad \text{for all } j < J_{\max}, \tag{10}$$

with a conditional mean-square score error of at most σ^2/n implying: $m_n \leq \sigma/\sqrt{n}$.

3.3 What contextual decision recovery means

For an independent $I \sim \mathrm{Unif}\{1,\dots,T\}$, we define $p_\theta(a \mid s,x) = \mathbb{P}_\theta(A_I = a \mid S = s, X_I = x)$. Assuming that contexts are exogenous with rollout-average law v_s common across policies and evaluate on $\rho(\,\mathrm{d}s)v_s(\,\mathrm{d}x)$, then under a unique optimal action and a hard gap $\Delta_s(x,a) \geq \gamma > 0$ for $a \neq a_s^*(x)$, we obtain:

$$\mathbb{E}_{s,x}\,\|p_{\theta_J}(\cdot \mid s,x) - \delta_{a_s^*(x)}\|_1 \leq \min\{2,\, 2r_J/\gamma\}, \quad r_J = \mathbb{E}[\overline{R}_S(\tau)]. \tag{11}$$

Thus, any informative regret bound above yields an action-law bound. A softer alternative lets $g_s(x)$ be the smallest suboptimal gap and assumes $\mathbb{P}_{\rho v}(g_S(X) \leq z) \leq C_{\mathrm{m}} z^\zeta$ for $z > 0$, with a unique optimum almost surely. The same decision error is at most $2\inf_{z>0}\{C_{\mathrm{m}} z^\zeta + r_J/z\}$, capped at two. For a different evaluation context law with density ratio at most C_{cov} relative to the actual rollout law, the hard-margin bound gains this factor.

4 Accessibility, information, and diagnostics

In this section we show that low-regret accessibility alone is not enough, as selected good trajectories must also be realizable by the deployed policy. The following proposition establishes that accessibility does not need to be realizable through a counter-example.

Proposition 4.1. *Consider $T = 1$, two equally likely hidden scenarios $S \in \{0,1\}$, two actions, and reward $\mathbf{1}\{A = S\}$. The initial observation is assumed to be identical in both scenarios. Under the uniform initial policy, each conditional regret law is $\frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$ and for $\alpha \leq 1/2$, selection within each scenario gives $\mathsf{B}_\alpha \nu_0^s = \delta_0$. Nevertheless, every admissible policy has expected regret $1/2$, and its average Wasserstein discrepancy from these selected targets is $1/2$.*

4.1 Exact imitation with noisy information

The following result shows that repeated RMFT imitation in a simple noisy-information model can converge geometrically to the best policy the agent can actually implement, while still leaving a positive regret floor because the agent does not observe enough information to match the evaluator's oracle.

Proposition 4.2 (Repeated population imitation)**.** *Let $T = 1$, let $S \in \{0,1\}$ be uniform, and let the observed context be $X = S \oplus N$, where $N \sim \mathrm{Bernoulli}(\eta)$ is independent of S and $0 < \eta < 1/2$*

and the reward is $\mathbf{1}\{A = S\}$. *Starting from a uniform policy, retaining a fraction* $\alpha \in (0, 1/2]$ *separately in each scenario, randomizing ties independently, and fitting the selected data by exact population maximum likelihood over all shared policies* $\pi(a \mid x)$, *then for* $q_j = \pi_j(A = X \mid X = x)$ *independently of* x, *the iterates satisfy:*

$$q_{j+1} = \frac{(1-\eta)q_j}{(1-\eta)q_j + \eta(1-q_j)}, \qquad q_j = \frac{(1-\eta)^j}{(1-\eta)^j + \eta^j}. \tag{12}$$

Their expected regret and transfer residual are

$$r_j = \eta + (1-2\eta)\frac{\eta^j}{(1-\eta)^j + \eta^j}, \qquad d_j = r_{j+1}. \tag{13}$$

Thus $r_j \downarrow \eta$, *the minimum over all admissible policies, with* $r_j - \eta \le (1-2\eta)[\eta/(1-\eta)]^j$. *At* $\eta = 1/2$, *the uniform policy remains unchanged and* $r_j = 1/2$; *at* $\eta = 0$, *one update achieves zero regret.*

Each initial scenario has zero-regret mass $1/2$, irrespective of η, and every elite regret law is δ_0. Nevertheless, replaying the fitted policy restores the original observation noise. For $\eta > 0$, the oracle's zero regret is unattainable even though imitation converges to the best policy given X. Tightening α below $1/2$ cannot remove this information floor. The exact update controls the deployed performance in this example; it is not a guarantee for approximate neural fitting or general multi-round tasks. The proof is in Appendix A.8.

Accessibility at a finite budget. For a useful threshold ε, we denote $p_s(\varepsilon) = P^s_\theta(\overline{R}_s \le \varepsilon)$. Then, a batch of L trajectories contains at least one useful trajectory with probability $1-(1-p_s(\varepsilon))^L$ and the total number of useful trajectories available is $\mathrm{Binomial}(L, p_s(\varepsilon))$. This highlights the accessibility bottleneck in applied settings, since a positive but very small p_s can be irrelevant for practical batch sizes.

Exploration and distillation. An episode-level mixture $\widetilde{P}^s = (1-w)P^s + wP^s_{\mathrm{aux}}$ guarantees $p^s_{\widetilde{P}}(\varepsilon) \ge w p^s_{\mathrm{aux}}(\varepsilon)$, assuming that we select the auxiliary policy once per episode. If $q_u(\nu^s_{\mathrm{aux}}) \le Ku^{1/\beta}$ for $u \le u_0$, then $q_u(\widetilde{\nu}^s) \le K(u/w)^{1/\beta}$ for $u \le wu_0$.

For policy distillation with the same environment and observation interface, a per-round average teacher-occupancy KL error at most ϵ_D gives: $d_{\mathrm{TV}}(P^s_T, P^s_D) \le \delta_D = \sqrt{T\epsilon_D/2}$. Consequently, $p^s_D(\varepsilon) \ge [p^s_T(\varepsilon) - \delta_D]_+$ and $q_u(\nu^s_D) \le q_{u+\delta_D}(\nu^s_T)$ for $u + \delta_D < 1$. Additional details are presented in Appendix C.

Measuring transfer. Define the population selection gain $G_j = \mathbb{E}_s[\mathbb{E}_{\widehat{\nu}^s_j} R - \mathbb{E}_{\mathsf{B}_\alpha \widehat{\nu}^s_j} R] \ge 0$. Then:

$$r_{j+1} \le r_j - G_j + d_j. \tag{14}$$

Hence $d_j < G_j$ is a sufficient condition for one-step improvement. Appendix D.3 defines prospective diagnostics for these quantities, including changes in the context distribution induced by selection.

5 Numerical Experiments

In this section, we evaluate the RMFT process in both numerical and language-grounded contextual environments, with experiments designed to isolate the gains obtained from regret-based post-training relative to an untuned LLM decision maker. We use the lightweight Qwen2.5-0.5B-Instruct model deliberately: by limiting the capabilities of the base model, we can attribute improvements more directly to the RMFT procedure itself rather than to strong pre-existing decision-making capabilities or model scale. Actions are sampled from normalized next-token scores for action labels, using textual context and history summaries. Following Park et al. (2025a), we use instantaneous regret as the supervision signal for elite selection during the training phase.

Table 1 summarizes cumulative pseudo-regret for the four training conditions. These comparisons characterize performance after task-specific offline training under the evaluation prior. Appendix D records the experimental settings and includes additional experiments using policy mixtures for the Qwen2.5-3B-Instruct model, and Appendix E provides indicative prompt-response pairs for each experiment.

Table 1: Cumulative pseudo-regret (reported mean ± standard error) for LLM agents (KD denotes policy distillation). Offline information and training budgets differ across methods.

Policy	Stationary (150 rounds)	Contextual (150 rounds)	Recommendation (100 Rounds)
Vanilla LLM	32.01 ± 5.62	33.06 ± 4.92	30.77 ± 0.61
Vanilla + RMFT	31.86 ± 5.88	33.56 ± 4.69	9.10 ± 0.31
KD	12.47 ± 2.09	26.67 ± 3.57	—
KD + RMFT	1.90 ± 0.48	17.99 ± 2.41	—

5.1 Stationary and contextual bandits

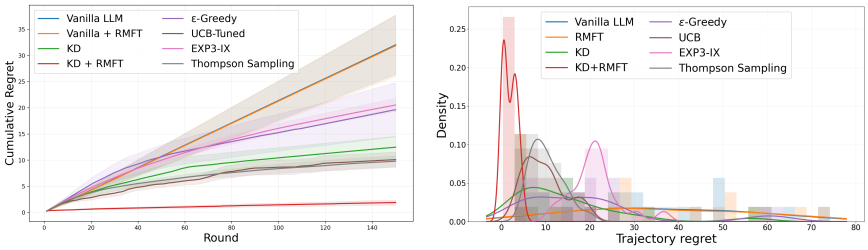

Figure 1: Comparison of algorithms for stationary stochastic bandit environments. Left: Cumulative regret over time for LLMs and classical baselines. Right: Distribution of cumulative regret across evaluation seeds.

In the stationary task, each scenario assigns fixed Bernoulli success probabilities to five arms. The agent observes arm-level pull counts, successes, empirical means, and recent outcomes. We compare direct RMFT, distillation of an ϵ-greedy teacher, and RMFT initialized from the distilled model. Baselines include ϵ-greedy, UCB-family methods (Auer et al., 2002a), EXP3-IX (Neu, 2015), and Thompson sampling (Thompson et al., 1933); Appendix D.2 reports their results.

Direct RMFT yields similar mean regret to the base model (32.01 versus 31.86), whereas RMFT initialized from the distilled policy reduces mean regret from 12.47 to 1.90. The contrast shows a dependence on initialization in this comparison and is consistent with changes in both useful-tail accessibility and selection-to-deployment transfer. Distillation can affect both quantities, so this comparison does not identify their separate contributions. Figure 1 shows the regret trajectories and final-regret distributions. Comparisons with classical online baselines should account for the LLM's additional offline training under the task prior.

In the contextual task, $X_t \in [-1, 1]$ and arm a has mean $\mu_s(x, a) = \text{sigmoid}(c_{s,a} + d_{s,a}x)$. Parameters are fixed within a scenario and resampled across scenarios. The construction makes multiple arms optimal on non-negligible context regions and excludes neighborhoods of optimal-action boundaries. This creates a favorable separated setting; a uniform margin across the full instance prior additionally depends on the parameter and exclusion rules. The contextual teacher is ridge ϵ-greedy. Baselines include LinUCB (Li et al., 2010), an EXP3-family method (Auer et al., 2002b; Neu, 2015), and linear Thompson sampling (Agrawal & Goyal, 2013); results appear in Appendix D.2.

The contextual task exhibits a similar dependence on initialization. Direct RMFT changes mean regret from 33.06 to 33.56, while applying RMFT after distillation reduces it from

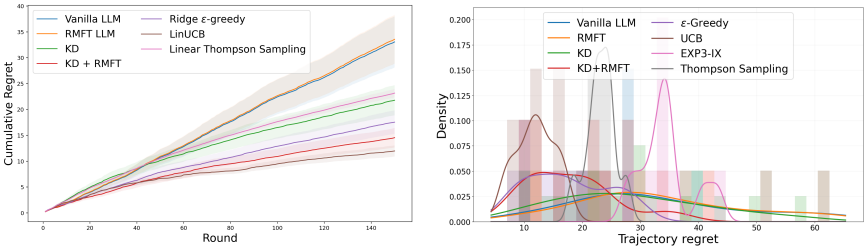

Figure 2: Comparison of algorithms for contextual stochastic bandit environments. Left: Cumulative Regret over time for LLMs and classical baselines. Right: Distribution of cumulative regret across evaluation seeds.

26.67 to 17.99.
The latter is close to the ridge ϵ-greedy teacher's 17.54, although LinUCB achieves a lower mean of 11.98. Figure 2 shows the corresponding comparisons. These results document improved performance after distilled initialization without establishing a general advantage over contextual-bandit methods.

5.2 Language-Grounded Contextual Product Recommendation

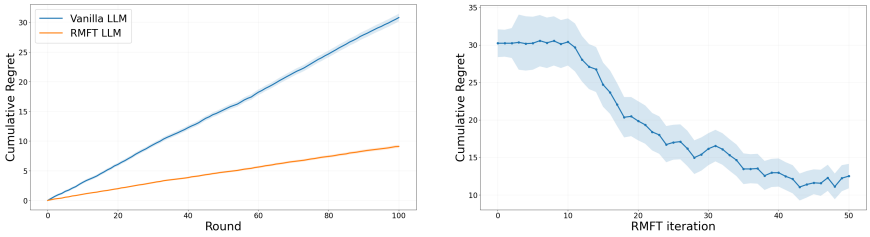

Figure 3: Comparison of algorithms for language-grounded contextual product recommendation. Left: Cumulative regret over time for LLMs and classical baselines. Right: Evolution of regret across RMFT iterations.

We consider a language-grounded contextual bandit environment based on real product information from the Amazon Reviews 2023 dataset. Unlike the previous synthetic experiments, both the context and the available actions are presented to the decision maker in language form. Each environment instance contains five semantically related products represented through titles, descriptions, and selected reviews constructed using BLaIR (Hou et al. (2024)), a language model family pre-trained on Amazon Reviews 2023 dataset, and at every round, each user is associated with a latent vector of attraction probabilities $\mu_t \in [0, 1]^5$. Based on this attraction vector, a natural language description reflecting the user's preferences is generated and once the LLM observes this linguistic user request, the textual descriptions of the candidate products, and history of past observations, it outputs a distribution over the available products and an action is sampled. The environment then samples a reward from the Bernoulli distribution parameterized by the latent attraction probabilities. We compare the vanilla LLM and RMFT applied to the initial model, and we do not use policy distillation, as contextual teachers did not yield any noteworthy improvement even when using frozen language encoder models to transform textual context to numerical form. We present the cumulative regret and the evolution of the regret per RMFT iteration in Figure 3.

6 Conclusion

In this work, we studied when regret-based self-training can turn language models into effective contextual decision makers. By treating regret-based selection and policy improvement as distinct operations, our analysis separates the quality available in a proposal's lower tail from the ability of a fitted policy to reproduce selected behavior under deployment constraints. The finite-iteration and finite-sample bounds make this distinction explicit for scenario-wise RMFT with regret atoms and ties. In the tractable noisy-observation model, repeated population imitation converges geometrically to the best admissible policy even though the informed oracle remains unattainable. The bandit experiments document contrasting responses to direct RMFT and distilled initialization. Together, these results motivate evaluating the proposal, the selected trajectories, and the deployed policy as separate components of regret-based post-training.

AI Use Statement

Generative AI tools assisted with revising the theoretical framework, formulating and proving claims, interpreting reported results, proposing diagnostic experiments, checking literature, editing text, and writing mathematical verification code. The authors are responsible for reviewing this assistance and for the final results and claims.

REFERENCES

Naoki Abe and Philip M Long. Associative reinforcement learning using linear probabilistic concepts. In *ICML*, pp. 3–11, 1999.

Aranyak Acharyya, Michael W Trosset, Carey E Priebe, and Hayden S Helm. Consistent estimation of generative model representations in the data kernel perspective space. *arXiv preprint arXiv:2409.17308*, 2024.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pp. 127–135. PMLR, 2013.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. *arXiv preprint arXiv:2211.15661*, 2022.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002a.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002b.

Ali Baheri and Cecilia O Alm. Llms-augmented contextual bandit. *arXiv preprint arXiv:2311.02268*, 2023.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *Advances in neural information processing systems*, 36:57125–57211, 2023.

Andrew G Barto, Richard S Sutton, and Peter S Brouwer. Associative search network: A reinforcement learning associative memory. *Biological cybernetics*, 40(3):201–211, 1981.

Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits. *Management Science*, 67(3):1329–1349, 2021.

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

David Brandfonbrener, Alberto Bietti, Jacob Buckman, Romain Laroche, and Joan Bruna. When does return-conditioned supervised learning work for offline reinforcement learning? *Advances in Neural Information Processing Systems*, 35:1542–1553, 2022.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24, 2011.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021a.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021b.

Sanxing Chen, Xiaoyin Chen, Yukun Huang, Roy Xie, and Bhuwan Dhingra. When greedy wins: Emergent exploitation bias in meta-bandit llm training. In *International Conference on Learning Representations*, volume 2026, pp. 57389–57421, 2026.

Zhixun Chen, Ming Li, Yuxuan Huang, Yali Du, Meng Fang, and Tianyi Zhou. Atlas: Agent tuning via learning critical steps. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25334–25349, 2025.

Sanjiban Choudhury and Paloma Sodhi. Better than your teacher: Llm agents that learn from privileged ai feedback. In *International Conference on Learning Representations*, volume 2025, pp. 10457–10493, 2025.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 208–214. JMLR Workshop and Conference Proceedings, 2011.

Jing Dong, Nanxiang Zhou, and Baoxiang Wang. Provable strategic in-context learning of transformers.

Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H Laradji, Manuel Del Verme, Tom Marty, Léo Boisvert, Megh Thakkar, Quentin Cappart, David Vazquez, et al. Workarena: How capable are web agents at solving common knowledge work tasks? *arXiv preprint arXiv:2403.07718*, 2024.

Brandon Duderstadt, Hayden S Helm, and Carey E Priebe. Comparing foundation models using data kernels. *arXiv preprint arXiv:2305.05126*, 2023.

Bianca Dumitrascu, Karen Feng, and Barbara Engelhardt. Pg-ts: Improved thompson sampling for logistic contextual bandits. *Advances in neural information processing systems*, 31, 2018.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. *Advances in neural information processing systems*, 23, 2010.

Dylan J Foster and Akshay Krishnamurthy. Efficient first-order contextual bandits: Prediction, allocation, and triangular discrimination. *Advances in Neural Information Processing Systems*, 34: 18907–18919, 2021.

Irena Gao, Percy Liang, and Carlos Guestrin. Model equality testing: Which model is this api serving? In *International Conference on Learning Representations*, volume 2025, pp. 86369–86382, 2025.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*, 2022.

Keegan Harris and Aleksandrs Slivkins. Should you use your large language model to explore or exploit? *arXiv preprint arXiv:2502.00225*, 2025.

Hayden Helm, Carey E Priebe, and Weiwei Yang. A statistical turing test for generative models. *arXiv preprint arXiv:2309.08913*, 2023.

Hayden Helm, Aranyak Acharyya, Youngser Park, Brandon Duderstadt, and Carey Priebe. Statistical inference on black-box generative models in the data kernel perspective space. *arXiv preprint arXiv:2410.01106*, 2024a.

Hayden Helm, Brandon Duderstadt, Youngser Park, and Carey Priebe. Tracking the perspectives of interacting language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1508–1519, 2024b.

Hayden Helm, Tianyi Chen, Harvey McGuinness, Paige Lee, Brandon Duderstadt, and Carey Priebe. Toward a digital twin of us congress. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 10148–10160, 2026.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*, 2024.

Yichun Hu, Nathan Kallus, and Xiaojie Mao. Smooth contextual bandits: Bridging the parametric and non-differentiable regret regimes. In *Conference on Learning Theory*, pp. 2007–2010. PMLR, 2020.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37: 79410–79452, 2024.

Akshay Krishnamurthy, Keegan Harris, Dylan J Foster, Cyril Zhang, and Aleksandrs Slivkins. Can large language models explore in-context? *Advances in Neural Information Processing Systems*, 37:120124–120158, 2024.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. *Advances in neural information processing systems*, 20, 2007.

Michael Laskin, Luyu Wang, Junhyuk Oh, Emilio Parisotto, Stephen Spencer, Richie Steigerwald, DJ Strouse, Steven Hansen, Angelos Filos, Ethan Brooks, et al. In-context reinforcement learning with algorithm distillation. *arXiv preprint arXiv:2210.14215*, 2022.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670, 2010.

Licong Lin, Yu Bai, and Song Mei. Transformers as decision makers: Provable in-context reinforcement learning via supervised pretraining. In *International Conference on Learning Representations*, volume 2024, pp. 31176–31233, 2024.

Pascal Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, pp. 1269–1283, 1990.

Giovanni Monea, Antoine Bosselut, Kianté Brantley, and Yoav Artzi. Llms are in-context bandit reinforcement learners. *arXiv preprint arXiv:2410.05362*, 2024.

Subhojyoti Mukherjee, Josiah P Hanna, Qiaomin Xie, and Robert Nowak. Pretraining decision transformers with reward prediction for in-context multi-task structured bandit learning. *arXiv preprint arXiv:2406.05064*, 2024.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. *Advances in Neural Information Processing Systems*, 28, 2015.

Allen Nie, Yi Su, Bo Chang, Jonathan Lee, Ed H Chi, Quoc V Le, and Minmin Chen. Evolve: Evaluating and optimizing llms forin-context exploration. In *The Exploration in AI Today Workshop at ICML 2025*, 2025.

Yonatan Oren, Nicole Meister, Niladri Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. Proving test set contamination in black-box language models. In *International Conference on Learning Representations*, volume 2024, pp. 16354–16372, 2024.

Chanwoo Park, Ziyang Chen, Asuman Ozdaglar, and Kaiqing Zhang. Post-training llms as better decision-making agents: A regret-minimization approach. *arXiv preprint arXiv:2511.04393*, 2025a.

Chanwoo Park, Xiangyu Liu, Asuman Ozdaglar, and Kaiqing Zhang. Do llm agents have regret? a case study in online learning and games. In *International Conference on Learning Representations*, volume 2025, pp. 41214–41234, 2025b.

Keiran Paster, Sheila McIlraith, and Jimmy Ba. You can't count on luck: Why decision transformers and rvs fail in stochastic environments. *Advances in neural information processing systems*, 35: 38966–38979, 2022.

Paulius Rauba, Qiyao Wei, and Mihaela Van Der Schaar. Statistical hypothesis testing for auditing robustness in language models. *arXiv preprint arXiv:2506.07947*, 2025.

Arik Reuter, Tim GJ Rudner, Vincent Fortuin, and David Rügamer. Can transformers learn full bayesian inference in context? *arXiv preprint arXiv:2501.16825*, 2025.

Philippe Rigollet and Assaf Zeevi. Nonparametric bandits with covariates. *arXiv preprint arXiv:1003.1630*, 2010.

Reuven Rubinstein. The cross-entropy method for combinatorial and continuous optimization. *Methodology and computing in applied probability*, 1(2):127–190, 1999.

Thomas Schmied, Jörg Bornschein, Jordi Grau-Moya, Markus Wulfmeier, and Razvan Pascanu. Llms are greedy agents: Effects of rl fine-tuning on decision-making abilities. *arXiv preprint arXiv:2504.16078*, 2025.

Pier Giuseppe Sessa, Robert Dadashi, Léonard Hussenot-Desenonges, Johan Ferret, Nino Vieillard, Alexandre Ramé, Bobak Shahriari, Sarah Perrin, Abram Friesen, Geoffrey Cideron, et al. Bond: Aligning llms with best-of-n distillation. In *International Conference on Learning Representations*, volume 2025, pp. 59040–59060, 2025.

Cencheng Shen, Darren Edge, Jonathan Larson, and Carey E Priebe. Explaining categorical feature interactions using graph covariance and llms. *Applied Network Science*, 11(1):12, 2025.

Chengshuai Shi, Kun Yang, Jing Yang, and Cong Shen. Transformers as game players: Provable in-context game-playing capabilities of pre-trained models. *Advances in neural information processing systems*, 37:132001–132049, 2024.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Fahim Tajwar, Yiding Jiang, Abitha Thankaraj, Sumaita Sadia Rahman, J Zico Kolter, Jeff Schneider, and Ruslan Salakhutdinov. Training a generally curious agent. *arXiv preprint arXiv:2502.17543*, 2025.

Nikola Tanković and Robert Šajina. Scalable and interpretable contextual bandits: A literature review and retail offer prototype. In *International Conference on Data Analytics & Management*, pp. 527–538. Springer, 2025.

R William Thompson et al. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pp. 35151–35174. PMLR, 2023.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

Xingyao Wang, Boxuan Li, Yufan Song, Frank F Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, et al. Openhands: An open platform for ai software developers as generalist agents. In *International Conference on Learning Representations*, volume 2025, pp. 65882–65919, 2025.

John Yang, Carlos Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. Swe-agent: Agent-computer interfaces enable automated software engineering. *Advances in Neural Information Processing Systems*, 37:50528–50652, 2024a.

Tong Yang, Jincheng Mei, Hanjun Dai, Zixin Wen, Shicong Cen, Dale Schuurmans, Yuejie Chi, and Bo Dai. Faster wind: Accelerating iterative best-of-n distillation for llm alignment. *arXiv preprint arXiv:2410.20727*, 2024b.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Andy K Zhang, Neil Perry, Riya Dulepet, Joey Ji, Celeste Menders, Justin W Lin, Eliot Jones, Gashon Hussein, Samantha Liu, Donovan Jasper, et al. Cybench: A framework for evaluating cybersecurity capabilities and risks of language models. *arXiv preprint arXiv:2408.08926*, 40, 2024.

Yu-Jie Zhang and Masashi Sugiyama. Online (multinomial) logistic bandit: Improved regret and constant computation cost. *Advances in Neural Information Processing Systems*, 36:29741–29782, 2023.

Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural contextual bandits with ucb-based exploration. In *International conference on machine learning*, pp. 11492–11502. PMLR, 2020.

Xiaoyuan Zhu, Yaowen Ye, Tianyi Qiu, Hanlin Zhu, Sijun Tan, Ajraf Mannan, Jonathan Michala, Raluca Ada Popa, and Willie Neiswanger. Auditing black-box llm apis with a rank-based uniformity test. *arXiv preprint arXiv:2506.06975*, 2025.

Wang Zihao, Cai Shaofei, Chen Guanzhou, et al. Describe, explain, plan and select: Interactive planning with large language models enables open-world multitask agents. In *Proceedings of the 37th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc*, pp. 34153–34189, 2024.

SUPPLEMENTARY MATERIAL

The Supplementary Material is organized as follows. Appendices A–C provide the proofs and assumptions. Appendix D describes the experimental protocol, and Appendix E gives representative interaction excerpts. The source package includes exact finite-support checks of the mathematical identities and a direct enumeration check of Proposition 4.2. Finally, Appendix F contains additional experiments focusing directly on probability measures over actions and studying how different error sources affect the RMFT process.

CONTENTS

1 Introduction **1**

1.1 Related Work 2

1.1.1 Language Models in Decision-Making Environments 2

1.1.2 Statistical Theory for Transformer Language Models 2

2 Decision policies and regret-based selection **3**

2.1 Natural Language-Based Contextual Bandit Environments 3

2.2 An elite operator that allows ties 4

3 From selected trajectories to deployed decisions **5**

3.1 A conditional finite-iteration guarantee 5

3.2 Finite trajectory samples and noisy scores 5

3.3 What contextual decision recovery means 6

4 Accessibility, information, and diagnostics **6**

4.1 Exact imitation with noisy information 6

5 Numerical Experiments **7**

5.1 Stationary and contextual bandits 8

5.2 Language-Grounded Contextual Product Recommendation 9

6 Conclusion **9**

A Proofs and mathematical scope **17**

A.1 Elite selection with atoms 17

A.2 Proof of the transfer theorem 17

A.3 Minimizing a uniform upper bound 18

A.4 Finite samples and arbitrary score perturbations 18

A.5 A population likelihood route to the transfer residual 19

A.6 Contextual action-law recovery 20

A.7 Information constraints and one-step improvement 20

A.8 Proof of repeated imitation with noisy information 21

B Consistency of the empirical imitation objective **21**

C Accessibility through mixtures and distillation **22**

D Experimental protocol and reported results **23**

D.1 Main-model protocol 23

D.2 Classical baseline results 23

D.3 Exploratory proposals and prospective diagnostics 24

E Representative interaction excerpts **24**

E.1 Stochastic Setting 24

E.2 Contextual Setting 26

E.3 Amazon Product Recommendation Setting 28

F Additional descriptive plots **29**

A PROOFS AND MATHEMATICAL SCOPE

All trajectory and scenario spaces are taken to be standard Borel spaces, so the regular conditional laws used in the paper exist. Reward gaps and policies are measurable, and the interaction horizon T is fixed. Expectations of normalized regret are finite because regret lies in $[0, 1]$. Statements about action-law recovery concern the actual rollout occupancy measure, not a pointwise limit over every history.

A.1 ELITE SELECTION WITH ATOMS

Let $c = q_\alpha(\nu)$ and $m = \nu(\{c\})$. By the quantile definition, $F_\nu(c^-) \le \alpha \le F_\nu(c)$. If $m > 0$, the weight $\lambda = (\alpha - F_\nu(c^-))/m$ lies in $[0, 1]$, and $\int w_{\alpha,P}\,\mathrm{d}P = \alpha$. If $m = 0$, both CDF values at c equal α, so the same normalization holds with $\lambda = 0$. Thus (2) defines a probability law in either case.

For $r < c$, its regret CDF is $F_\nu(r)/\alpha$. For $r \ge c$, all retained mass has regret at most r, so this CDF is one. This proves the first identity in (3); the corresponding generalized inverse gives its quantile identity, up to null sets in the quantile index.

Proof of Lemma 3.1. Using the quantile representation of W_1,

$$\begin{aligned} W_1(\mathsf{B}_\alpha\nu, \mathsf{B}_\alpha\xi) &= \int_0^1 |q_{\alpha u}(\nu) - q_{\alpha u}(\xi)|\,\mathrm{d}u \\ &= \frac{1}{\alpha}\int_0^\alpha |q_v(\nu) - q_v(\xi)|\,\mathrm{d}v \le \frac{1}{\alpha}W_1(\nu, \xi). \end{aligned}$$

Alternatively, the same inequality follows by integrating the CDF identity and noting that $z \mapsto \min\{z/\alpha, 1\}$ is $1/\alpha$-Lipschitz. Composing the CDF map gives $F_{\mathsf{B}^j_\alpha\nu}(r) = \min\{F_\nu(r)/\alpha^j, 1\}$, which is precisely $\mathsf{B}_{\alpha^j}\nu$. Finally, the mean of a bounded law is the integral of its quantile function. Substitution of $v = uu'$ gives $\mathbb{E}_{\mathsf{B}_u\nu} R = u^{-1}\int_0^u q_v(\nu)\,\mathrm{d}v$. □

Scope of the composition identity. The identity concerns the regret marginal, which is all that is needed in the transfer theorem. Policy fitting is an intervening operation in RMFT, so the identity must not be used to equate an actual sequence of neural policies with iterated elite conditioning.

Hard gaps and zero-regret atoms. If all positive per-round gaps exceed $\gamma > 0$, then any positive trajectory regret is at least γ/T. Therefore, an initial quantile envelope tending to zero entails positive mass at zero regret. This is consistent with our operator and all the Wasserstein bounds. In particular, neither the consistency result below nor the noisy-selection argument requires a continuous regret CDF.

A.2 PROOF OF THE TRANSFER THEOREM

Proof of Theorem 3.1. Let $e_j = \mathbb{E}_s\, W_1(\widehat{\nu}^s_j, \nu^{\circ,s}_j)$. The actual and ideal processes start from the same regret laws, so $e_0 = 0$. The triangle inequality and Lemma 3.1 imply

$$\begin{aligned} e_{j+1} &\le \mathbb{E}_s\, W_1(\widehat{\nu}^s_{j+1}, \mathsf{B}_\alpha\widehat{\nu}^s_j) + \mathbb{E}_s\, W_1(\mathsf{B}_\alpha\widehat{\nu}^s_j, \mathsf{B}_\alpha\nu^{\circ,s}_j) \\ &\le d_j + \alpha^{-1}e_j. \end{aligned}$$

Unrolling gives $e_J \le \sum_{j=0}^{J-1}\alpha^{-(J-1-j)}d_j$. Because the identity function on $[0, 1]$ is 1-Lipschitz, differences in mean regret are bounded by W_1. Consequently,

$$\begin{aligned} r_J &\le \mathbb{E}_s\,\mathbb{E}_{\nu^{\circ,s}_J} R + e_J \\ &= \mathbb{E}_s\left[\frac{1}{\alpha^J}\int_0^{\alpha^J} q_u(\nu^s_{\theta_0})\,\mathrm{d}u\right] + e_J \\ &= \frac{1}{\alpha^J}\int_0^{\alpha^J} A_0(u)\,\mathrm{d}u + e_J. \end{aligned}$$

The interchange uses nonnegativity and Tonelli's theorem. This proves (7). The cap at one uses the support of normalized regret. □

Under (8) and $\alpha^J \le u_0$,

$$\frac{1}{\alpha^J}\int_0^{\alpha^J}(r_{\mathrm{acc}} + Ku^{1/\beta})\,\mathrm{d}u = r_{\mathrm{acc}} + \frac{\beta}{\beta+1}K\alpha^{J/\beta}.$$

If $\nu^s_{\theta_0}(\{0\}) \ge p_*$ almost surely in s, then $q_u(\nu^s_{\theta_0}) = 0$ for $u \le p_*$, proving the zero-atom statement. For $J = 0$, the reference mean is simply the initial mean; the same decomposition holds with an empty error sum.

A.3 Minimizing a uniform upper bound

Selecting an iteration count from an error bound. Fix a maximum number $J_{\max}$ in advance and suppose $d_j \le b$ holds throughout that range. Using the looser ideal bound $r_{\mathrm{acc}} + K\alpha^{J/\beta}$ gives the computable criterion

$$U(J) = r_{\mathrm{acc}} + K\alpha^{J/\beta} + B(\alpha^{-J} - 1), \qquad B = \frac{\alpha b}{1-\alpha}, \tag{15}$$

for $J_0 \le J \le J_{\max}$, where $J_0 = \lceil \log u_0 / \log \alpha \rceil$. For $b > 0$, its continuous minimizer before clipping to this range is

$$J_{\mathrm{cont}} = \frac{\log(K/(\beta B))}{(1+\beta^{-1})\log(1/\alpha)}.$$

The minimizing integer is found by evaluating the admissible neighboring integers after clipping; for $b = 0$, take the largest admissible J. This optimizes a worst-case upper bound. It does not prove an empirical stopping optimum, divergence after that index, or an unavoidable positive error floor.

Assume $0 < u_0 \le 1$, $J_{\max} \ge J_0$, and $d_j \le b$ for all $j < J_{\max}$. Summing the geometric series yields

$$\sum_{j=0}^{J-1} \alpha^{-(J-1-j)} b = \frac{\alpha b}{1-\alpha}(\alpha^{-J} - 1).$$

Put $c_1 = \beta^{-1}\log(1/\alpha)$ and $c_2 = \log(1/\alpha)$. For $K, b > 0$, the part of $U(J)$ that depends on J is $Ke^{-c_1 J} + Be^{c_2 J}$, which is strictly convex. Its derivative vanishes when $c_1 Ke^{-c_1 J} = c_2 Be^{c_2 J}$. Since $c_1/c_2 = 1/\beta$, the solution is the displayed J_{cont}. Clipping it to $[J_0, J_{\max}]$ and checking its floor and ceiling gives the exact minimizer over admissible integers. If $b = 0$, the criterion decreases in J; if the admissible range is empty, the accessibility envelope cannot be used on the prescribed horizon.

If b is constructed from a high-probability error bound, that bound must use the predeclared $J_{\max}$ throughout. Replacing it during optimization by a different, iteration-dependent budget is not justified by this argument. Moreover, a nonzero upper bound on d_j is not a lower bound on actual policy error. An error floor requires separate evidence, such as the information constraint in Proposition 4.1.

A.4 Finite samples and arbitrary score perturbations

We first establish a deterministic lemma used in Theorem 3.2.

Lemma A.1 (Perturbed top-k selection)**.** *Let $r_1, \dots, r_L, z_1, \dots, z_L \in [0,1]$ and $\alpha = k/L$. Let ν_L be the empirical law of the r_ℓ, let I_z be any set of exactly k indices with smallest z_ℓ, and let $\widetilde{\nu}_z = k^{-1}\sum_{\ell \in I_z}\delta_{r_\ell}$. Then*

$$W_1(\widetilde{\nu}_z, \mathsf{B}_\alpha \nu_L) \le \frac{2}{k}\sum_{\ell=1}^{L}|z_\ell - r_\ell|.$$

Proof. Let $\xi_L = L^{-1}\sum_\ell \delta_{z_\ell}$. The law of the selected scores is $\mathsf{B}_\alpha \xi_L$, including in the presence of ties because all boundary scores have the same scalar value. Pair each selected true regret with its own score:

$$W_1(\widetilde{\nu}_z, \mathsf{B}_\alpha \xi_L) \le \frac{1}{k}\sum_{\ell \in I_z}|r_\ell - z_\ell|.$$

The pairing of all L observations gives $W_1(\xi_L, \nu_L) \le L^{-1} \sum_\ell |r_\ell - z_\ell|$. Apply Lemma 3.1 and the triangle inequality. Each of the two resulting terms is at most $k^{-1} \sum_\ell |r_\ell - z_\ell|$. □

The lemma controls regret values of the selected data, not the total-variation distance between selected trajectory sets. Large changes in trajectory identity at nearly equal scores can have small regret-law distance.

Proof of Theorem 3.2. Let $\nu_{L,j,s} = L^{-1} \sum_\ell \delta_{R_{j,s,\ell}}$. Conditional on the past, the Dvoretzky–Kiefer–Wolfowitz inequality with Massart's constant (Massart, 1990) gives

$$\mathbb{P}\left(\sup_r |F_{\nu_{L,j,s}}(r) - F_{\widehat{\nu}_j^s}(r)| > a_L \;\middle|\; \text{past} \right) \le \frac{\delta}{M J_{\max}}, \quad a_L = \sqrt{\frac{\log(2MJ_{\max}/\delta)}{2L}}.$$

It holds for any CDF, including a discrete one. Integrating on $[0,1]$ bounds $W_1(\nu_{L,j,s}, \widehat{\nu}_j^s)$ by a_L. A union bound, with iterated conditioning for adaptively chosen policies, makes these bounds simultaneous over all scenario–iteration pairs.

On that event, the triangle inequality, Lemma 3.1, and Lemma A.1 imply

$$\begin{aligned} W_1(\widehat{\nu}_{j+1}^s, \mathsf{B}_\alpha \widehat{\nu}_j^s) &\le W_1(\widehat{\nu}_{j+1}^s, \widetilde{\nu}_{j,s}) + W_1(\widetilde{\nu}_{j,s}, \mathsf{B}_\alpha \nu_{L,j,s}) \\ &\quad + W_1(\mathsf{B}_\alpha \nu_{L,j,s}, \mathsf{B}_\alpha \widehat{\nu}_j^s) \\ &\le W_1(\widehat{\nu}_{j+1}^s, \widetilde{\nu}_{j,s}) + \frac{2}{k} \sum_\ell |Z_{j,s,\ell} - R_{j,s,\ell}| + \frac{a_L}{\alpha}. \end{aligned}$$

Averaging over the fixed M scenarios proves the theorem. No independence between the fitted policy and its training data is required because e_j^{fit} remains an explicit residual. □

For (10), let $D_{j,s,\ell} = |Z_{j,s,\ell} - R_{j,s,\ell}| \in [0,1]$. Conditional independence and $\mathbb{E}[D_{j,s,\ell} \mid \text{rollouts,past}] \le m_n$ permit Hoeffding's inequality for each scenario and a union bound over $MJ_{\max}$ pairs. Averaging the resulting bounds over s proves (10). If an unclipped estimator has conditional mean-square error at most σ^2/n, clipping to $[0,1]$ cannot increase its error from $R \in [0,1]$; Cauchy–Schwarz therefore gives $m_n \le \sigma/\sqrt{n}$. Conditional unbiasedness and variance σ^2/n are sufficient, but not necessary, for this condition.

What the finite-sample result does not establish. The theorem is conditional on small fitting residuals. These include the inability of a shared policy to realize selected data and are not pure optimization errors. The DKW specialization is conservative; for example, at $\alpha = 0.1$, $L = 200$, $M = 50$, $J_{\max} = 50$, and $\delta = 0.05$, its sampling term alone is approximately 1.70. It therefore gives no nontrivial numerical certificate for the reported 50-iteration LLM runs. The decomposition identifies quantities for a sharper problem-specific analysis or independent audit; it is not used to certify those experimental gains.

Scenario sampling. For an arbitrary evaluation law ρ, Theorem 3.1 remains valid with the residual defined by (5) under that same law. Theorem 3.2 establishes a sampling bound only when ρ is the fixed finite scenario design. If training scenarios are resampled, or deployment uses new scenarios, a further generalization argument is needed to bound (5). Selection of k trajectories separately in each training scenario does not establish that argument.

A.5 A POPULATION LIKELIHOOD ROUTE TO THE TRANSFER RESIDUAL

Fix iteration j and write $Q_j^s = \mathsf{T}_\alpha P_{\theta_j}^s$. Let $\ell_\theta(\tau)$ contain all model-dependent log-likelihood terms in the trajectory, and define

$$L_j(\theta) = -\,\mathbb{E}_s\, \mathbb{E}_{Q_j^s}\, \ell_\theta(\tau), \qquad K_j(\theta) = \mathbb{E}_s\, D_{\mathrm{KL}}(Q_j^s \| P_\theta^s), \qquad \mathcal{A}_j = \inf_\theta K_j(\theta).$$

Assume these objectives are finite on the candidate class and $K_j(\theta) - L_j(\theta)$ is independent of θ. Common environment kernels and a common initial law give this identity whenever the relevant entropy and likelihood terms are finite. The candidate policies share the same observation constraints. Selection can change the conditional distribution of external observations in Q_j^s, so the best KL approximation need not be zero.

Proposition A.1 (Uniform imitation error and realizability)**.** *Suppose an empirical objective* $\widehat{L}_j$ *satisfies*

$$\sup_\theta |\widehat{L}_j(\theta) - L_j(\theta)| \le g_j, \qquad \widehat{L}_j(\theta_{j+1}) \le \inf_\theta \widehat{L}_j(\theta) + o_j.$$

Then (6) *holds.*

Proof. For every candidate θ, comparison through the empirical objective gives $L_j(\theta_{j+1}) \le L_j(\theta) + 2g_j + o_j$. The objective identity and an infimum over candidates imply $K_j(\theta_{j+1}) \le \mathcal{A}_j + 2g_j + o_j$. On $[0,1]$, W_1 is bounded by total variation. Data processing under the regret map and Pinsker's inequality therefore give, for each scenario,

$$W_1(\widehat{\nu}^s_{j+1}, \mathsf{B}_\alpha \widehat{\nu}^s_j) \le d_{\mathrm{TV}}(P^s_{\theta_{j+1}}, Q^s_j) \le \sqrt{D_{\mathrm{KL}}(Q^s_j \| P^s_{\theta_{j+1}})/2}.$$

Averaging and applying Jensen's inequality proves the claim. □

This compares population trajectory laws, rather than a continuous rollout law with a finite empirical trajectory measure. It is an alternative route to d_j, not an additional term to add to Theorem 3.2. A uniform bound g_j must be established for the actual selection, scoring, and scenario design; the pointwise result in Appendix B does not establish it. In Proposition 4.1, the smallest population KL is $\log 2$ despite exact fitting, illustrating the realizability term. We do not claim a uniform bound for the experimental LoRA training.

A.6 Contextual action-law recovery

Choose an independent uniform round I, and use the joint law of (S, X_I, A_I) induced by the actual rollout policy. Its conditional action distribution is $p_\theta(a \mid s, x)$, so

$$\mathbb{E}[\overline{R}_S(\tau)] = \mathbb{E}_{s,x} \sum_a p_\theta(a \mid s, x) \Delta_s(x, a).$$

If the optimum is unique and every suboptimal gap is at least γ, then

$$\sum_a p_\theta(a \mid s, x) \Delta_s(x, a) \ge \gamma[1 - p_\theta(a^*_s(x) \mid s, x)].$$

The distance to the deterministic optimum is exactly $\|p_\theta(\cdot \mid s, x) - \delta_{a^*_s(x)}\|_1 = 2[1 - p_\theta(a^*_s(x) \mid s, x)]$. Integration proves (11). For a different evaluation law absolutely continuous with density ratio at most C_{cov}, integrate this nonnegative error against the density ratio to obtain the stated factor. This controls the same conditional action law under a changed weighting of contexts; it is not a theorem about rerunning an agent under changed history dynamics.

For the soft-margin version, on the event $g_s(x) > z$, any suboptimal action incurs gap larger than z. Hence

$$\mathbb{P}(A_I \ne a^*_S(X_I)) \le \mathbb{P}(g_S(X_I) \le z) + r_J/z \le C_{\mathrm{m}} z^\zeta + r_J/z.$$

Multiply by two and minimize over $z > 0$. When $r_J > 0$, choosing $z = (r_J/(\zeta C_{\mathrm{m}}))^{1/(\zeta+1)}$ gives

$$2(\zeta+1) C_{\mathrm{m}}^{1/(\zeta+1)} \zeta^{-\zeta/(\zeta+1)} r_J^{\zeta/(\zeta+1)},$$

capped at two. When $r_J = 0$, taking $z \downarrow 0$ gives zero error. If there are multiple optimal actions on positive-measure contexts, one can instead bound the probability of choosing outside the optimal-action set; distance to one arbitrarily chosen deterministic optimum is then not the appropriate target.

A.7 Information constraints and one-step improvement

Proof of Proposition 4.1. Any admissible policy chooses action 1 with some probability q independent of the hidden scenario. Its regrets are q in expectation under $s = 0$ and $1 - q$ under $s = 1$, giving unconditional expected regret $1/2$. For the uniform initial policy, each scenario has zero-regret mass $1/2$, so selecting any $\alpha \le 1/2$ produces the regret law δ_0 in both scenarios. For any law on $[0,1]$, $W_1(\nu, \delta_0) = \mathbb{E}_\nu R$. The average discrepancy between any deployed conditional laws and these elite targets is therefore $[q + (1-q)]/2 = 1/2$. □

With population supervised imitation and $\alpha \le 1/2$, the balanced selected dataset prescribes action 0 in scenario 0 and action 1 in scenario 1, despite identical prompts. Its action log-likelihood is $\frac{1}{2}\log(1-q) + \frac{1}{2}\log q$, maximized at $q = 1/2$. Thus even exact population fitting returns the same initial policy. If S is observed, a policy can choose $A = S$ and achieve zero regret. This example separates information from sample support and fitting capacity.

Equation (14) follows directly from

$$r_{j+1} \le \mathbb{E}_s \, \mathbb{E}_{\mathsf{B}_\alpha \widehat{\nu}_j^s} \, R + d_j = r_j - G_j + d_j.$$

The gain G_j is nonnegative because retaining the lower tail cannot increase the mean. This is a sufficient condition for improvement; $d_j \ge G_j$ alone does not prove deterioration.

A.8 Proof of repeated imitation with noisy information

Proof of Proposition 4.2. Suppose $\pi_j(A = x \mid X = x) = q_j$ for both contexts, initially $q_0 = 1/2$. For either scenario, the probability of zero regret is

$$u_j = (1-\eta)q_j + \eta(1-q_j) = \eta + (1-2\eta)q_j.$$

For $q_j \ge 1/2$, we have $u_j \ge 1/2 \ge \alpha$. All selected trajectories therefore have $A = S$. Independent randomization among these ties makes their conditional law independent of α. Since u_j is the same in both scenarios, the joint selected law is

$$Q_j(s,x,a) = \frac{\frac{1}{2}\mathbb{P}(X = x \mid S = s)\pi_j(a \mid x)\mathbf{1}\{a = s\}}{u_j}.$$

Summing over s, a gives $Q_j(X = x) = 1/2$. Maximizing $\sum_{x,a} Q_j(x,a)\log\pi(a \mid x)$ over all conditional action probabilities returns $\pi_{j+1}(a \mid x) = Q_j(A = a \mid X = x)$. In particular,

$$q_{j+1} = Q_j(A = x \mid X = x) = \frac{(1-\eta)q_j}{(1-\eta)q_j + \eta(1-q_j)}.$$

This is symmetric in x and at least q_j for $q_j \in [1/2, 1]$, proving the induction. For $0 < \eta < 1/2$, its odds satisfy

$$\frac{q_{j+1}}{1-q_{j+1}} = \frac{1-\eta}{\eta}\frac{q_j}{1-q_j}.$$

The initial odds are one, yielding (12). Fresh rollouts retain the original observation channel, so their probability of an incorrect action is $(1-\eta)(1-q_j) + \eta q_j = \eta + (1-2\eta)(1-q_j)$, which proves (13) and the geometric bound. Given $X = x$, the posterior probability of $S = x$ is $1-\eta$. The best admissible action is consequently $A = x$, with error η; no policy observing only X can do better. Since every selected regret law is δ_0, $W_1(\nu_{j+1}^s, \delta_0) = r_{j+1}$ in each scenario, establishing the residual formula. Substitution in the update gives the two endpoint cases directly. □

In this example the oracle action-law error is exactly $2r_j$, hence converges to 2η. This is compatible with convergence to the best admissible policy: the scenario-specific oracle has additional information. The result requires the stated symmetry, exact population fitting, and the full class of binary policies; finite batches and restricted fitting require a separate analysis.

B Consistency of the empirical imitation objective

We state consistency at a fixed scenario and checkpoint, permitting arbitrary atoms. Let τ_ℓ be i.i.d. from P, let U_ℓ be independent $\mathrm{Unif}(0,1)$ variables, and select the k_L lexicographically smallest pairs $(\overline{R}(\tau_\ell), U_\ell)$, where $k_L/L \to \alpha \in (0,1)$. Denote this index set by I_L. Let $\ell_\theta(\tau)$ be the model-dependent log-likelihood of the retained model outputs for a fixed candidate θ.

Proposition B.1 (Pointwise likelihood consistency with ties)**.** *If* $\mathbb{E}_P\,|\ell_\theta(\tau)| < \infty$*, then*

$$\frac{1}{k_L}\sum_{\ell \in I_L} \ell_\theta(\tau_\ell) \xrightarrow{\mathbb{P}} \mathbb{E}_{\mathsf{T}_\alpha P}\,\ell_\theta(\tau).$$

Proof. Let F be the CDF of $\overline{R}$ under P and set

$$V = F(\overline{R}^-) + U[F(\overline{R}) - F(\overline{R}^-)].$$

The randomized probability integral transform is uniform on $[0, 1]$. Ordering $(\overline{R}, U)$ agrees almost surely with ordering V, and the event $\{V \le \alpha\}$ selects exactly the weights in (2). For atomless values, this statement follows by the usual probability integral transform; on an atom of mass m, it distributes the atom uniformly across its CDF interval of length m.

The order statistic $V_{(k_L)}$ converges in probability to α. Put $G = \ell_\theta(\tau)$. On $\{|V_{(k_L)} - \alpha| \le h\}$, the difference between the averages with indicators $\mathbf{1}\{V_\ell \le V_{(k_L)}\}$ and $\mathbf{1}\{V_\ell \le \alpha\}$, each normalized by L, is in absolute value at most $L^{-1} \sum_\ell |G_\ell| \mathbf{1}\{|V_\ell - \alpha| \le h\}$. For fixed h, the law of large numbers gives limit $\mathbb{E}[|G|\mathbf{1}\{|V - \alpha| \le h\}]$. Integrability and $\mathbb{P}(V = \alpha) = 0$ make this tend to zero as $h \downarrow 0$. Another law of large numbers gives convergence of the fixed-threshold average to $\mathbb{E}[G\mathbf{1}\{V \le \alpha\}]$. Multiplying by $L/k_L \to 1/\alpha$ proves the proposition. □

For a fixed finite scenario design, averaging this statement across scenarios gives the corresponding population objective. Uniform convergence over θ, existence or uniqueness of optimizers, and convergence of the supervised training algorithm require additional hypotheses and are not consequences of the proposition.

C ACCESSIBILITY THROUGH MIXTURES AND DISTILLATION

Proposition C.1 (Episode-level exploration)**.** *Fix a scenario s and let $\widetilde{P}^s = (1 - w)P^s + wP^s_{\mathrm{aux}}$, where $0 < w \le 1$. For every regret threshold ε, $p^s_{\widetilde{P}}(\varepsilon) \ge w p^s_{\mathrm{aux}}(\varepsilon)$. If $q_v(\nu^s_{\mathrm{aux}}) \le Kv^{1/\beta}$ for $0 < v \le u_0$, then $q_u(\widetilde{\nu}^s) \le K(u/w)^{1/\beta}$ for $0 < u \le wu_0$.*

Proof. The mixture dominates wP^s_{aux} on every measurable event, proving the first statement. At $r = q_{u/w}(\nu^s_{\mathrm{aux}})$, the auxiliary law assigns at least u/w mass to $\{\overline{R}_s \le r\}$. The mixture therefore assigns at least u mass to that event, implying the quantile bound. □

Implementing this trajectory mixture means choosing the base or auxiliary policy once at the beginning of an episode and retaining that choice. If instead every action kernel is $(1 - w)\pi(\cdot \mid h, x) + w\pi_{\mathrm{aux}}(\cdot \mid h, x)$, its likelihood along an auxiliary trajectory is at least w^T times the auxiliary action likelihood. With the same external dynamics, this gives trajectory-measure domination by $w^T P^s_{\mathrm{aux}}$, not generally by wP^s_{aux}. If the auxiliary policy has privileged observations, it may be usable as a data generator during training but is not thereby an admissible deployed exploration policy.

Proposition C.2 (Additive transfer by distillation)**.** *Fix a scenario s. Suppose teacher and distilled policies share the environment, initial law, and observation interface, and*

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{P^s_T}[D_{\mathrm{KL}}(\pi_T(\cdot \mid H_t, X_t) \| \pi_D(\cdot \mid H_t, X_t))] \le \epsilon_D.$$

With $\delta_D = \sqrt{T\epsilon_D/2}$, for all ε,

$$p^s_D(\varepsilon) \ge [p^s_T(\varepsilon) - \delta_D]_+.$$

For $0 < u$ with $u + \delta_D < 1$, $q_u(\nu^s_D) \le q_{u+\delta_D}(\nu^s_T)$.

Proof. The KL chain rule cancels the identical environment kernels, giving

$$D_{\mathrm{KL}}(P^s_T \| P^s_D) = \sum_{t=1}^{T} \mathbb{E}_{P^s_T} D_{\mathrm{KL}}(\pi_T(\cdot \mid H_t, X_t) \| \pi_D(\cdot \mid H_t, X_t)) \le T\epsilon_D.$$

Pinsker's inequality implies $d_{\mathrm{TV}}(P^s_T, P^s_D) \le \delta_D$. Apply this to the event $\{\overline{R}_s \le \varepsilon\}$ for the tail-mass bound. Taking $\varepsilon = q_{u+\delta_D}(\nu^s_T)$ gives student mass at least u, proving the quantile statement. □

The condition must hold for each scenario to obtain each conditional quantile guarantee. An average KL bound over scenarios instead gives a joint TV bound and controls an average tail probability; it does not automatically give a uniform conditional tail bound. For fixed $\delta_D > 0$, the result cannot certify arbitrarily deep-tail access. In particular, an additive approximation bound is not a multiplicative preservation guarantee.

D EXPERIMENTAL PROTOCOL AND REPORTED RESULTS

D.1 MAIN-MODEL PROTOCOL

Table 2 records the main-model settings. Appendix D.3 discusses the separate Qwen2.5-3B auxiliary study.

Table 2: Main-model training configuration.

Item	Setting
Base model	Qwen2.5-0.5B-Instruct
Actions per scenario	5
Synthetic evaluation horizon	150
RMFT iterations	50
Scenarios per iteration / rollouts per scenario	50 / 200
Retained trajectories / elite fraction	20 / 0.1
Distillation scenarios / teacher rollouts each	4,000 / 1
Fine-tuning method / epochs per stage	LoRA / 1
Learning rate	$2 \cdot 10^{-4}$
Training / evaluation seed offsets	1,000 / 100,000

Evaluation and uncertainty. We compare mean cumulative pseudo-regret under the task's instance prior. The error quantities in Table 1 and the baseline table are reported as standard errors. The comparisons are descriptive and do not establish statistical significance or equivalence.

Regret scale. The theoretical score is normalized pseudo-regret in Eq. (1); the reported evaluation metric is cumulative pseudo-regret. At a common horizon, these differ by a positive constant and induce the same trajectory ranking. For a general implemented score Z, Theorem 3.2 quantifies its deviation from the normalized target through e_j^{score}.

Synthetic instance generation. Stationary scenarios use five Bernoulli arms. Contextual means have the form $\mathrm{sigmoid}(c_{s,a} + d_{s,a}x)$, with coefficients fixed within an episode and resampled across instances. Removing contexts near optimal-action boundaries creates separation within an instance; a uniform hard margin also depends on the prior and exclusion rule.

Model interface and fitting. The deployed policy uses action-label scores under textual prompts. Appendix E illustrates the observation summaries and resulting action probabilities.

Recommendation construction. Products use Amazon Reviews 2023 text with BLaIR-based processing; preferences, requests, and satisfaction rewards are simulated. A semantic match between the request and product description can lower regret without feedback-driven adaptation.

D.2 CLASSICAL BASELINE RESULTS

The contextual reward specification is logistic, whereas LinUCB and linear Thompson sampling use linear reward models. The results therefore characterize their performance under this task specification.

Table 3: Reported classical baseline cumulative pseudo-regret. These methods are evaluated online; the LLM variants additionally use offline data and post-training under the instance prior.

Environment	Policy	Cumulative pseudo-regret
Stationary	ϵ-greedy	19.61 ± 5.15
	UCB	10.09 ± 1.46
	EXP3-IX	20.52 ± 1.25
	Thompson sampling	9.78 ± 1.03
Contextual	Ridge ϵ-greedy	17.54 ± 2.26
	LinUCB	11.98 ± 1.06
	EXP3	34.40 ± 1.48
	Linear Thompson sampling	23.16 ± 0.79

D.3 Exploratory proposals and prospective diagnostics

We also present an auxiliary experiment comparing direct RMFT with uniform and ϵ-greedy exploration mixtures for Qwen2.5-3B-Instruct in the stationary bandit task in Figure 4. Proposition C.1 describes the accessibility guarantee for a mixture chosen once per episode; independent action-level mixing induces a different trajectory law.

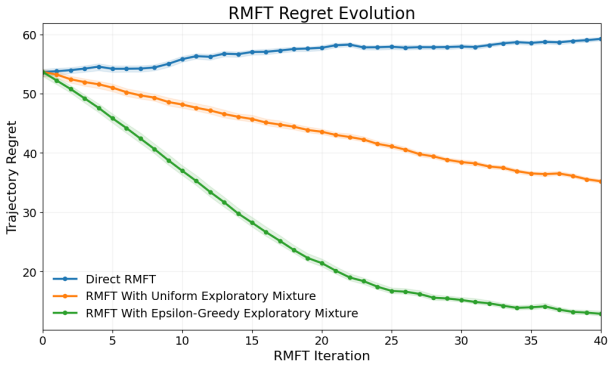

Figure 4: Evolution of the average regret per RMFT iteration for the Qwen2.5-3B-Instruct model with different exploration mixtures.

The transfer analysis suggests three diagnostic quantities at a fixed normalized-regret threshold ε:

$$p_j^s(\varepsilon) = \mathbb{P}^s_{\theta_j}(\overline{R}_s \le \varepsilon),$$
$$m_j^{\mathrm{elite},s} = \mathbb{E}_{\mathsf{B}_\alpha \nu_j^s} R,$$
$$m_{j+1}^{\mathrm{fresh},s} = \mathbb{E}_{\nu_{j+1}^s} R.$$

These distinguish useful-tail access, selected-trajectory quality, and deployed performance. An empirical Wasserstein distance between independently estimated elite and fresh-rollout regret laws additionally estimates the transfer residual, subject to sampling uncertainty. The difference of their means is a complementary summary and generally differs from their Wasserstein distance. Context frequencies before and after selection can reveal changes in the selected observation distribution. These quantities describe a prospective diagnostic design; the reported LLM experiments do not separately estimate all of them.

E Representative interaction excerpts

In this section, we present some prompts given to the best-performing LLMs from Section 5 and their responses for each setting.

E.1 Stochastic Setting

We present an interaction between the LLM and the environment that showcases the exploration and exploitation balancing behavior of the model. After initially leaving one arm unvisited, the policy explores this arm and rapidly shifts toward exploiting it after observing a positive reward, while continuing to retain a small amount of exploratory probability mass in later rounds.

(. . . Previous interactions . . .)

SYSTEM + USER PROMPT — ROUND 33

System prompt.
You are a stochastic bandit policy. Use the observed history to assign a probability to every available arm.

User prompt.
`round=33`
`available_arms=[0, 1, 2, 3, 4]`

Within this episode, each arm has a fixed but unknown probability of reward 1. Arm probabilities are newly sampled in other episodes, so arm labels have no persistent meaning. You observe only the reward of the sampled arm.

`observed_arm_statistics:`

arm	pulls	reward_1s	empirical_mean	last_reward
0	4	2	0.500	0
1	14	8	0.571	0
2	7	4	0.571	1
3	8	4	0.500	0
4	**0**	**0**	**NA**	**NA**

Arm 4 is the only arm that has not yet been explored.

The policy is represented by the model's normalized next-token scores for the arm labels in `available_arms`. Assign higher score to arms that should be selected more often while preserving any desired exploration.
`POLICY_ARM_LABEL:`

AGENT RESPONSE — EXPLORING THE UNVISITED ARM

Policy:

$$\pi_{33} = [0.089,\ 0.488,\ 0.218,\ 0.098,\ \mathbf{0.107}].$$

Although the current observations favor arm 1, the policy retains **10.7% probability on the only unvisited arm**.

Selected action: Arm 4

Observed reward: 1.

USER PROMPT — ROUND 34

`round=34`
`available_arms=[0, 1, 2, 3, 4]`

Within this episode, each arm has a fixed but unknown probability of reward 1. Arm probabilities are newly sampled in other episodes, so arm labels have no persistent meaning. You observe only the reward of the sampled arm.

`observed_arm_statistics:`

arm	pulls	reward_1s	empirical_mean	last_reward
0	4	2	0.500	0
1	14	8	0.571	0
2	7	4	0.571	1
3	8	4	0.500	0
4	**1**	**1**	**1.000**	**1**

The policy is represented by the model's normalized next-token scores for the arm labels in `available_arms`. Assign higher score to arms that should be selected more often while preserving any desired exploration.
`POLICY_ARM_LABEL:`

AGENT RESPONSE — RAPID SHIFT TO EXPLOITATION

After observing a reward of 1 from the previously unvisited arm, the policy changes sharply:

$$\pi_{34} = [0.015,\ 0.017,\ 0.015,\ 0.015,\ \mathbf{0.937}].$$

The probability assigned to arm 4 increases from $\mathbf{10.7}\%$ to $\mathbf{93.7}\%$ after the first observation.

Selected action: Arm 4

Observed reward: 1.

(. . . Next 5 interactions, predominantly exploiting Arm 4 . . .)

USER PROMPT — ROUND 40

`round=40`
`available_arms=[0, 1, 2, 3, 4]`

Within this episode, each arm has a fixed but unknown probability of reward 1. Arm probabilities are newly sampled in other episodes, so arm labels have no persistent meaning. You observe only the reward of the sampled arm.

`observed_arm_statistics:`

arm	pulls	reward_1s	empirical_mean	last_reward
0	4	2	0.500	0
1	14	8	0.571	0
2	7	4	0.571	1
3	9	4	0.444	0
4	**6**	**6**	**1.000**	**1**

The policy is represented by the model's normalized next-token scores for the arm labels in `available_arms`. Assign higher score to arms that should be selected more often while preserving any desired exploration.
`POLICY_ARM_LABEL:`

AGENT RESPONSE — CONTINUED EXPLORATION

The policy continues to place the overwhelming majority of its probability mass on the empirically best arm:

$$\pi_{40} = [0.0168,\ 0.0168,\ \mathbf{0.0154},\ 0.0154,\ \mathbf{0.9355}].$$

Nevertheless, the stochastic policy still retains probability mass on the remaining alternatives. In this round, an exploratory action is sampled:

Selected action: Arm 2

despite the policy assigning $\mathbf{93.55}\%$ probability to Arm 4.
Observed reward: 1.

(. . . Subsequent interactions . . .)

E.2 Contextual Setting

In this section, we provide an indicative interaction generated by the LLM in contextual environments. Unlike the stochastic setting, the key behavior now is context-dependent adaptation, where despite recently receiving several positive rewards from the same arm, the model needs to change its policy when the observed context changes and maintain the new action preference for similar contexts.

SYSTEM + USER PROMPT — ROUND 4

System prompt.
You are a contextual stochastic-bandit policy. For the current context, assign a probability to every available arm. The reward relationship is stationary within this episode but newly sampled in other episodes, so arm labels do not have persistent meaning.

User prompt.
`round=4`
`current_context=[0.9092]`
`available_arms=[0, 1, 2, 3, 4]`

Observed contextual history (most recent window):

round	x0	arm	reward
0	-0.0895	3	0
1	-0.3700	3	1
2	0.4726	2	1
3	0.3996	2	1

Rewards are Bernoulli. Use both context and history to infer which arms are promising at the current context while retaining useful exploration. The policy is represented by normalized next-token scores for the arm labels.
`POLICY_ARM_LABEL:`

AGENT RESPONSE — POSITIVE CONTEXT

For the strongly positive current context $\mathbf{x_4 = 0.9092}$, the model assigns most of its probability mass to **Arm 2**:

$$\pi_4 = [0.0130,\ 0.0167,\ \mathbf{0.9119},\ 0.0427,\ 0.0157].$$

Selected action: Arm 2

Observed reward: 1.

USER PROMPT — ROUND 5

`round=5`
`current_context=[-0.2305]`
`available_arms=[0, 1, 2, 3, 4]`

Observed contextual history (most recent window):

round	x0	arm	reward
0	-0.0895	3	0
1	-0.3700	3	1
2	0.4726	2	1
3	0.3996	2	1
4	**0.9092**	**2**	**1**

Rewards are Bernoulli. Use both context and history to infer which arms are promising at the current context while retaining useful exploration. The policy is represented by normalized next-token scores for the arm labels.
`POLICY_ARM_LABEL:`

AGENT RESPONSE — CONTEXT-DEPENDENT POLICY SHIFT

Although Arm 2 has now produced **three consecutive observed rewards of 1 at positive contexts**, the current context changes substantially from 0.9092 to

$$\mathbf{x_5 = -0.2305}.$$

The model responds by sharply reversing its action preference:

$$\pi_5 = [0.0224,\ 0.0186,\ 0.0394,\ \mathbf{0.8958},\ 0.0239].$$

In particular,

$$\pi_4(2) = \mathbf{91.2}\% \quad \longrightarrow \quad \pi_5(2) = \mathbf{3.9}\%,$$

while

$$\pi_4(3) = \mathbf{4.3}\% \quad \longrightarrow \quad \pi_5(3) = \mathbf{89.6}\%.$$

Selected action: Arm 3

Observed reward: 1.

USER PROMPT — ROUND 6

```
round=6
current_context=[-0.2914]
available_arms=[0, 1, 2, 3, 4]
```

Observed contextual history (most recent window):

round	x0	arm	reward
0	-0.0895	3	0
1	-0.3700	3	1
2	0.4726	2	1
3	0.3996	2	1
4	0.9092	2	1
5	**-0.2305**	**3**	**1**

Rewards are Bernoulli. Use both context and history to infer which arms are promising at the current context while retaining useful exploration. The policy is represented by normalized next-token scores for the arm labels.
`POLICY_ARM_LABEL:`

AGENT RESPONSE — CONSISTENT BEHAVIOR AT A SIMILAR CONTEXT

The new context remains in the same negative region,

$$\mathbf{x_6} = -\mathbf{0.2914},$$

and the model becomes even more confident in Arm 3:

$$\pi_6 = [0.0090,\ 0.0109,\ 0.0110,\ \mathbf{0.9524},\ 0.0167].$$

Selected action: Arm 3

Thus, after shifting away from Arm 2 when the context changes sign, the model **maintains the context-dependent preference** when presented with a nearby negative context.

E.3 Amazon Product Recommendation Setting

We finally present an indicative interaction from the language-grounded Amazon product recommendation environment. In this setting, both the user context and the candidate actions are presented in natural language. To keep the example concise, we only provide information in the prompt that is directly relevant, with extensive product descriptions and user reviews being abbreviated.

SYSTEM + USER PROMPT — AMAZON PRODUCT RECOMMENDATION

System prompt.
You are a stochastic product-recommendation policy in a contextual bandit. Five candidate products remain fixed during the episode, but different users have different needs. Use the current

request, product information, and observed recommendation outcomes to assign probability mass over all available products.

User prompt.
`round=3`

AVAILABLE PRODUCTS

Product 0: Hair Curlers Magic Styling Kit
Product description, features, and selected user reviews abbreviated.

Product 1: Rose Quartz Contour Face Roller and Gua Sha Sculptor
Features include a **face roller and gua sha made with real 100% Brazilian rose quartz**, designed for facial massage, anti-aging and anti-wrinkle use, and convenient **at-home facial treatment**.
Additional description and selected reviews abbreviated.

Product 2: Cordless Hair Curler
Product description, features, and selected user reviews abbreviated.

Product 3: Oil-Absorbing Volcanic Face Roller / Jade Roller
Product description, features, and selected user reviews abbreviated.

Product 4: Eyelash Curler
Features emphasize **strong iron plating with a rubber handle**.
Additional description and selected reviews abbreviated.

CURRENT USER REQUEST
I am looking for something that emphasizes **a face roller and gua sha made with real 100% Brazilian rose quartz**, known for anti-aging and anti-wrinkle use, and suitable as an **at-home treatment**. I care less for products centered around **strong iron plating with a rubber handle**.

PREVIOUS INTERACTIONS
Previous recommendation outcomes abbreviated.

The reward is binary user satisfaction. Infer which product is most suitable for the current request while retaining useful exploration. The policy is represented by normalized next-token scores for the product labels.
`POLICY_ARM_LABEL:`

AGENT RESPONSE — SUCCESSFUL LANGUAGE-GROUNDED RECOMMENDATION

The model places the overwhelming majority of its probability mass on **Product 1**:

$$\pi_3 = [0.047,\ \mathbf{0.935},\ 0.005,\ 0.006,\ 0.007].$$

Selected action: Product 1

The selected product directly matches the user's stated preference for a **Brazilian rose-quartz face roller and gua sha** intended for an **at-home facial treatment**.
Observed user-satisfaction reward: 1.

F ADDITIONAL DESCRIPTIVE PLOTS

The plots comprise model studies and idealized probability-measure experiments.

Model family and size. Figures 5 and 6 show base-model regret distributions for Qwen2.5 and Qwen3.5, describing finite-budget tail accessibility. These base-model distributions characterize the proposals and do not establish post-training robustness across model families or scales.

Probability-measure illustrations. These experiments use action distributions at contexts $\{-1, -0.5, 0, 0.5, 1\}$ and vary the elite fraction, accessibility, and margin. One perturbation adds a

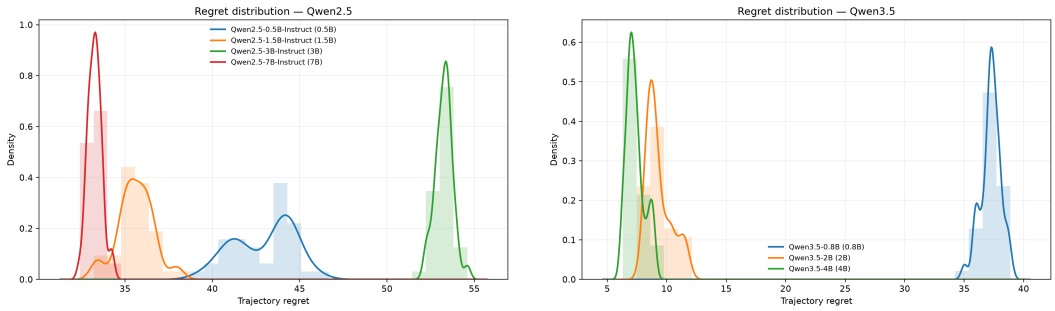

Figure 5: Regret distributions in the stationary stochastic bandit environment for Qwen2.5 (left) and Qwen3.5 (right) checkpoints.

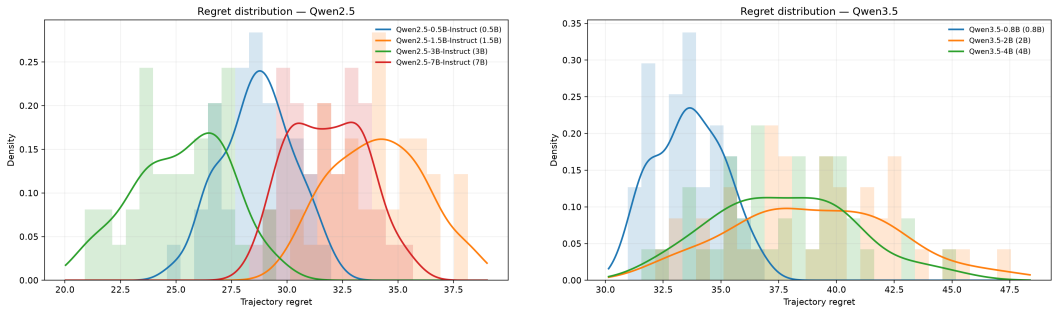

Figure 6: Regret distributions in the contextual bandit environment for Qwen2.5 (left) and Qwen3.5 (right) checkpoints.

fresh random distribution to the imitation objective with weight 0.3; another adds Gaussian score noise with standard deviation 0.3. These perturbations illustrate sensitivity to changes in the imitation objective and score; they do not directly measure the deployed residual in Theorem 3.1. Changes in the margin can also change the scale of regret. The plotted ordinate is decision-law error, rather than mean regret. Figure 7 varies the elite fraction; Figure 8 compares the two perturbations; and Figure 9 varies accessibility and margin. These illustrations are distinct from the exact binary-observation model in Proposition 4.2.

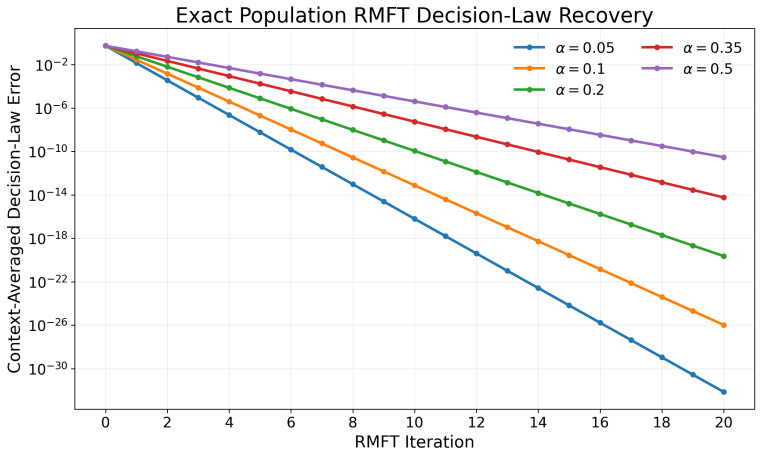

Figure 7: Decision-law error by iteration in the idealized probability-measure experiment for different elite fractions α. The experiment operates on action distributions.

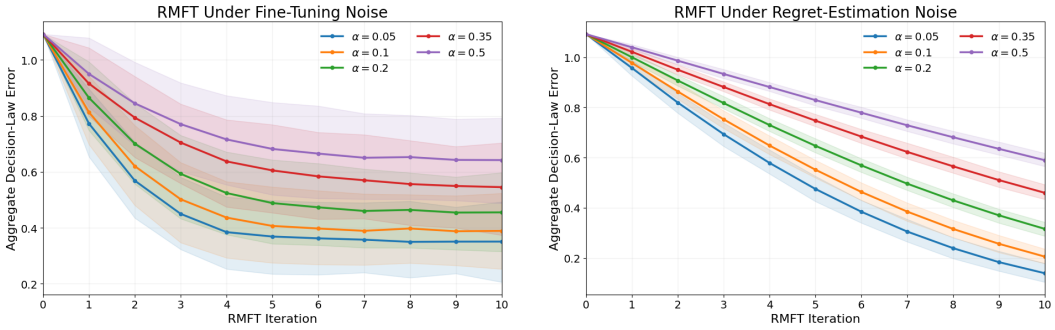

Figure 8: Decision-law error by iteration for different elite fractions α and error sources. Left: a perturbation of the imitation objective. Right: regret-score noise.

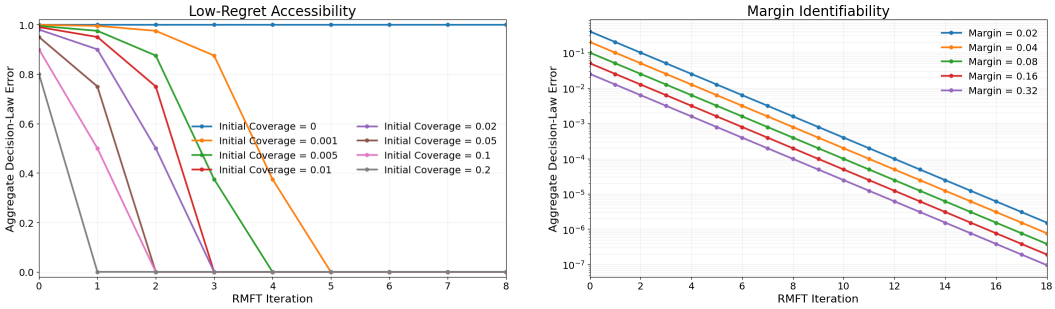

Figure 9: Decision-law error by iteration in the idealized probability-measure experiments. Left: varying initial low-regret accessibility. Right: varying the margin.