POST-TRAINING AT THE EDGE OF DETECTABILITY: A GAME-THEORETIC APPROACH TO FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning (RL) fine-tuning is widely used in language model training to improve performance on a target task while limiting drift from a reference policy. A standard way to balance this trade-off is via a KL-regularized RL objective, although this formulation does not by itself provide a principled way to set the regularization coefficient. In practice, the coefficient is typically chosen heuristically or via hyperparameter search, which can lead to unnecessary overhead in training cost or undesirable reward–retention trade-offs. We instead propose a game-theoretic framework that gives this trade-off an explicit statistical interpretation. Specifically, we study a sequential game in which an agent chooses a policy to maximize cumulative reward while a monitor observes policy outputs over time and tests for deviations from the reference policy. Although not originating from the same perspective, we show that the resulting equilibrium policy can nonetheless be expressed as the solution to a KL-regularized RL problem for an optimal regularization parameter that can be viewed as maximizing reward per unit of statistical distinguishability. Drawing on classical results from concave-convex fractional programming, we provide a principled method for learning this equilibrium coefficient via reduction to the KL-regularized RL objective, thus allowing for flexible integration into standard fine-tuning pipelines. In experiments with `Qwen3-8B` and `Llama-3.2-1B`, we show that our methods result in competitive reward-retention trade-offs in a continual learning setting, and illustrate how our framework may be used to audit API providers serving open-source models.

1 INTRODUCTION

A common goal in large language model (LLM) fine-tuning is to improve performance on a target objective while preserving useful behaviors inherited from a reference policy. These objectives may be at odds with one another, and there are many possible ways to formalize the resulting trade-off. One of the most widely used approaches is Kullback–Leibler (KL)-regularized reinforcement learning (RL), which augments the reward objective with a penalty for deviating from the reference policy (Jaques et al., 2017; 2019; Neu et al., 2017; Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022). In this formulation, the regularization coefficient $\beta \geq 0$ determines how aggressively the fine-tuned policy is penalized for deviating from the reference policy (e.g., a pre-trained LLM). When β is too small, fine-tuning may increase reward at the cost of significantly changing the model's behavior. When β is too large, the model remains close to the reference policy, but may fail to adequately learn the target task. In practice, this coefficient is often selected by manual tuning or via grid search (see, e.g., Ouyang et al., 2022; Zhang et al., 2023b; Lin et al., 2024; Tang et al., 2024). However this approach can waste compute and lead to undesirable reward–retention trade-offs. Some implementations instead adapt β online to match a prescribed KL target, but this still requires the learner to specify the target divergence and update heuristic in advance (Schulman et al., 2017; Ziegler et al., 2019).

We consider an alternative notion of deviation from the reference policy. Rather than measuring deviations directly through a distance or similarity metric, we ask how difficult it is to distinguish the fine-tuned policy from the reference policy based on its outputs. This naturally leads to a game between an agent who seeks to maximize reward, and a monitor attempting to detect deviations from the reference policy. In this game, an agent deploys a policy to maximize their utility, while a monitor observes policy outputs as they are generated and aims to detect whether those outputs have

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.

been generated by the (intended) reference policy or by some other (unintended) one. If the monitor detects a deviation from the reference policy, deployment is terminated. The agent therefore faces a trade-off between increasing reward and remaining statistically indistinguishable from the reference policy. We call this the *sequential detection game*. Our key contributions are as follows:

Game-theoretic formulation for RL fine-tuning (Section 3). We provide a new perspective on RL fine-tuning through the lens of sequential detection. Starting from a game-theoretic formulation, we prove that the agent's Nash equilibrium strategy in the sequential detection game is to fine-tune the reference policy with the standard KL-regularized RL objective. Thus, rather than introducing a new fine-tuning objective, our game-theoretic formulation recovers one that is already used in practice, which allows existing RL fine-tuning algorithms to be applied directly. At the same time, it gives the KL penalty an operational interpretation as the statistical cost of becoming distinguishable from the reference policy under sequential monitoring. Moreover, the equilibrium identifies the regularization coefficient $\beta^\star$ that optimally trades off between reward maximization and statistical distinguishability. Unlike conventional RL fine-tuning, where β is chosen heuristically, the equilibrium coefficient is determined solely by the reward function, prompt distribution, and reference policy.

Learning the optimal regularization strength (Section 4). Building on our equilibrium analysis, we present a stochastic bisection algorithm (Algorithm 1) that estimates $\beta^\star$ to ϵ-precision by solving $O(\log(1/\epsilon))$ RL sub-problems. Algorithm 1 may be interpreted as a game-theoretically principled approach to RL fine-tuning: rather than relying on a separate hyperparameter search, it adaptively computes the equilibrium regularization strength. The analysis of Algorithm 1 combines classical techniques from the literature on fractional programming (Dinkelbach, 1967; Schaible, 1976) and sequential hypothesis testing (Robbins & Siegmund, 1974).

Continual learning and model auditing experiments (Section 5). We evaluate our framework in both continual learning and model auditing settings using `Qwen3-8B` (Yang et al., 2025) and `Llama-3.2-1B` (Meta AI, 2024). In continual learning, we find preliminary evidence that the equilibrium regularization coefficient yields better reward–retention trade-offs than compute-matched manual tuning. In model auditing, we show that the monitor's equilibrium-prescribed test detects hidden model modifications by open-source API providers more quickly than natural baselines while controlling the type-I error rate.

2 PRELIMINARIES

We now review the key ingredients from RL fine-tuning, sequential hypothesis testing, and game theory that will later be combined in our analysis.

RL fine-tuning. Consider a prompt space $\mathcal{X}$ and response space $\mathcal{Y}$. A policy π is a mapping from prompts to distributions over responses, i.e., $\pi : \mathcal{X} \to \Delta(\mathcal{Y})$. We use $\pi(y|x)$ to denote the probability of receiving response y given prompt x, and $y \sim \pi(\cdot|x)$ to denote a sample from the conditional distribution over responses given x.[1]

Given a reference policy π_{ref}, a reward function $r : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, the KL-regularized RL fine-tuning objective for a regularization parameter $\beta \geq 0$ is defined by

$$\max_{\pi} \left\{ \mathbb{E}_{y\sim\pi(\cdot|x)}[r(x,y)] - \beta \cdot \mathbb{E}\left[\mathrm{KL}(\pi(\cdot|x), \pi_{\mathrm{ref}}(\cdot|x))\right] \right\}, \tag{1}$$

where $\mathrm{KL}(\pi(\cdot|x), \pi_{\mathrm{ref}}(\cdot|x)) := \sum_{y\in\mathcal{Y}} \pi(y|x) \log \frac{\pi(y|x)}{\pi_{\mathrm{ref}}(y|x)}$ is the KL divergence between $\pi(\cdot|x)$ and $\pi_{\mathrm{ref}}(\cdot|x)$, and expectations are (also) taken over the prompt distribution $\mathcal{D}$. This optimization admits a closed-form optimal solution (Donsker & Varadhan, 1975; Rafailov et al., 2023) of the form

$$\pi_\beta(y|x) := \pi_{\mathrm{ref}}(y|x) \frac{\exp(r(x,y)/\beta)}{Z_\beta(x)}, \tag{2}$$

where $Z_\beta(x) := \mathbb{E}_{y\sim\pi_{\mathrm{ref}}(\cdot|x)}[\exp(r(x,y)/\beta)]$ is the partition function. We call π_β the tilt of π_{ref}. Since $Z_\beta(x)$ is generally intractable to compute due to the size of $\mathcal{Y}$, RL methods like Proximal Policy Optimization (Schulman et al., 2017) and Group Relative Policy Optimization (Shao et al., 2024) are typically used to approximately solve Equation equation 1.

[1] LLMs generate tokens auto-regressively, but for our purposes it suffices to consider only the induced distribution over complete responses.

Sequential hypothesis testing is a paradigm of statistical inference for which type-I errors (false positive rates) are controlled not only at pre-determined sample sizes, but also at stopping times (Wald, 1945; 1947). Informally, sequential hypothesis tests allow analysts to "peek" at their data routinely to adaptively stop experiments and make conclusions for data-dependent reasons.

In the case of simple nulls and simple alternatives, one sees a sequence of observations $z_1, z_2, \ldots$ from a distribution P_z and the goal is to determine whether $P_z = P$ (the null) or $P_z = Q$ (the alternative). A sequential hypothesis test, $\phi \equiv (\phi_t)_{t\in\mathbb{N}}$, is a sequence of functions $\phi_t \equiv \phi(z_1, \ldots, z_t)$ outputting either "reject" or "do not reject" at each time step $t \in \mathbb{N}$. For a fixed $\alpha \in (0, 1)$, a test is said to be α-*correct* if it controls the type-I error at all sample sizes simultaneously; i.e., $P(\exists\, t \in \mathbb{N} : \phi_t \text{ rejects}) \leq \alpha$. Letting $\tau_\alpha \equiv \tau_\alpha(\phi) = \inf\{t \in \mathbb{N} : \phi_t \text{ rejects}\}$ be the *stopping time*, notice that α-correctness is equivalent to the condition $P(\tau_\alpha < \infty) \leq \alpha$. A test is *power-one* if $Q(\tau < \infty) = 1$.

For any α-correct, power-one sequential test, $\mathbb{E}_Q[\tau_\alpha] \geq \frac{\log(1/\alpha)}{\mathrm{KL}(Q,P)}$ (Wald, 1945). This lower bound is known to be tight in the sense that there exist valid α-correct, power-one tests for which

$$\lim_{\alpha\downarrow 0} \frac{\mathbb{E}_Q[\tau_\alpha]}{\log(1/\alpha)} = \frac{1}{\mathrm{KL}(Q, P)}. \tag{3}$$

A canonical test with this leading-order behavior is the sequential probability ratio test (SPRT) (Wald, 1945) given by $\phi_t = \mathbb{1}\{\prod_{i=1}^t \frac{dQ(z_i)}{dP(z_i)} \geq \frac{1}{\alpha}\}$.

Nash equilibria. A two-player game consists of strategy spaces $\mathcal{S}_1, \mathcal{S}_2$ and utility functions $u_1, u_2 : \mathcal{S}_1 \times \mathcal{S}_2 \to \mathbb{R}$. Given a strategy $s_2 \in \mathcal{S}_2$ for player 2, a strategy $s_1 \in \mathcal{S}_1$ is a *best response* for player 1 if $s_1 \in \arg\max_{s_1' \in \mathcal{S}_1} u_1(s_1', s_2)$. Best responses for player 2 are defined analogously. A strategy profile (s_1^*, s_2^*) is a *Nash equilibrium* if both strategies are simultaneously best responses, i.e. $u_1(s_1^*, s_2^*) \geq u_1(s_1, s_2^*)$ for all $s_1 \in \mathcal{S}_1$ and $u_2(s_1^*, s_2^*) \geq u_2(s_1^*, s_2)$ for all $s_2 \in \mathcal{S}_2$. Thus, at a Nash equilibrium, neither player can improve their utility by unilaterally changing their strategy; i.e., each player's behavior is optimal given the behavior of the other.

3 The Sequential Detection Game

We explore differences between reference policies and target policies from the perspective of a monitor aiming to quickly distinguish one from the other in a two-player (general sum) game against a strategic agent.

Sequential detection game:
Consider a setup where a monitor intends to deploy a policy π_{ref} while an agent potentially deploys a policy $\tilde{\pi} \neq \pi_{\mathrm{ref}}$ in place of π_{ref}. The monitor intends to detect such a change from prompt-response outputs while controlling the type-I error rate at a desired level $\alpha \in (0, 1)$. Concretely, for each time step $t \in \mathbb{N}$:

1. A prompt $x_t \sim \mathcal{D}$ is generated from the prompt distribution $\mathcal{D}$ and a response $y_t \sim \tilde{\pi}(\cdot \mid x_t)$ is generated from the agent's deployed policy.
2. The monitor updates a sequential test $\phi_t \equiv \phi((x_1, y_1), \ldots, (x_t, y_t))$ subject to the type-I error constraint: $\mathbb{P}_{y\sim\pi_{\mathrm{ref}}}(\exists\, t \in \mathbb{N} : \phi_t \text{ rejects}) \leq \alpha$.
3. If ϕ_t rejects, then the monitor stops deployment, and the agent receives utility $\sum_{i=1}^t r(x_i, y_i)$ according to reward function $r : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, while the monitor receives utility $-t$.

In this sequential detection game, the monitor stops deployment if and only if ϕ_t rejects. In other words, their stopping rule is characterized by the stopping time $\tau_\alpha = \inf\{t \in \mathbb{N} : \phi_t \text{ rejects}\}$, with the convention that $\tau_\alpha = \infty$ if ϕ_t never rejects. From this vantage point, the agent wishes to deploy a policy $\tilde{\pi}$ to maximize their expected cumulative reward before detection, $\mathbb{E}_{y\sim\tilde{\pi}(\cdot|x)}[\sum_{i=1}^{\tau_\alpha} r(x_i, y_i)]$, while the monitor wishes to minimize $\mathbb{E}_{y\sim\tilde{\pi}(\cdot|x)}[\tau_\alpha]$.

The agent's optimal choice of $\tilde{\pi}$ turns out to be precisely a KL-regularized tilt of the reference policy, while the monitor's best response turns out to be a likelihood ratio test between π_{ref} and $\tilde{\pi}$. To state these facts formally, we first require two assumptions on the agent's reward function r.

Assumption 3.1 (The agent can and must strategize to benefit)**.** *It is possible for the agent to achieve positive expected utility. That is, there exists a policy π such that $\mathbb{E}_{y\sim\pi(\cdot|x)}[r(x,y)] > 0$. Moreover, they must strategize in order to do so, i.e., $\mu_{\mathrm{ref}} := \mathbb{E}_{y\sim\pi_{\mathrm{ref}}(\cdot|x)}[r(x,y)] < 0$.*

Assumption 3.2 (The monitor can always devise a detection strategy)**.** *For every $\pi' \neq \pi_{\mathrm{ref}}$, there exists a test $(\phi'_t)_{t\in\mathbb{N}}$ with a finite expected stopping time, i.e., $\mathbb{E}_{\pi'}[\tau_\alpha(\phi')] < \infty$.*

Assumption 3.1 provides sufficient conditions for the agent to want to participate in the game and to learn a non-trivial strategy.[2] For the necessity of strategization in Assumption 3.1, it is mathematically equivalent to assume that $\mu_{\mathrm{ref}} - c \leq 0$, where $c \geq 0$ is the minimum improvement in expected reward required for the agent to want to update the reference policy. The constant c may also be interpreted as capturing the cost associated with model training (e.g., monetary costs, time spent, effort exerted, etc.). Assumption 3.2 is an analogous assumption on the monitor's incentive to play the game. It is a weak assumption that is satisfied for likelihood ratios in all but pathological cases which we eschew for the purposes of this paper.

In what follows, we characterize the equilibrium of the sequential detection game. These statements should be interpreted as holding in the so-called "high-confidence" regime, i.e., where $\alpha \downarrow 0$. This technicality arises because of the discrete-time nature of the sequential tests we consider and does not affect the main interpretation of our result. Concretely, Robbins & Siegmund (1974) prove an information-theoretic lower bound $\log(1/\alpha)/\mathrm{KL}(Q,P)$ for the expected stopping time of any α-correct sequential test. This bound is tight at the leading $\log(1/\alpha)$ order, but the corresponding upper bounds can include lower-order corrections due to boundary overshoot, e.g., because the log-likelihood ratio is updated at discrete times, it can cross the $\log(1/\alpha)$ boundary and not hit it exactly. Therefore, the following equilibrium characterization should be interpreted as stating that unilateral deviations cannot improve either player's utility *at the leading* $\log(1/\alpha)$ *order*.

Equilibrium characterization. We begin with the monitor's equilibrium strategy. Suppose the agent is playing policy $\tilde{\pi}$. Targeting the true agent policy $\tilde{\pi}$ maximizes the expected evidence accumulated per sample and therefore minimizes the expected stopping time. Moreover, it is known that SPRTs exhibit the optimal leading order behavior of Equation 3 (Siegmund, 1985). Therefore the monitor's best-response to agent policy $\tilde{\pi}$ is to test for it using an SPRT.

The agent's goal is to pick their policy to maximize $\mathbb{E}\left[\sum_{t=1}^{\tau_\alpha} r(x_t,y_t)\right]$, which can be written as $\mathbb{E}[\tau_\alpha]\cdot\mathbb{E}_{y\sim\pi(\cdot|x)}[r(x,y)]$ by Wald's equation, since the expected stopping time will be finite under Assumptions 3.1 and 3.2. By the Robbins & Siegmund lower bound, the agent can guarantee themselves utility at least $\log(1/\alpha)\mathbb{E}_{y\sim\pi(\cdot|x)}[r(x,y)]/\mathbb{E}[\mathrm{KL}(\pi(\cdot|x),\pi_{\mathrm{ref}}(\cdot|x))]$, which is also an (asymptotic) upper bound on their utility under the monitor's best-response SPRT. Since neither $\mathbb{E}_{y\sim\pi(\cdot|x)}[r(x,y)]$ nor $\mathbb{E}[\mathrm{KL}(\pi(\cdot|x),\pi_{\mathrm{ref}}(\cdot|x))]$ depend on α, the agent's optimization takes the form

$$\max_{\pi}\ \mathbb{E}_{y\sim\pi(\cdot|x)}[r(x,y)]/\mathbb{E}[\mathrm{KL}(\pi(\cdot|x),\pi_{\mathrm{ref}}(\cdot|x))], \tag{4}$$

where we use the convention that $0/0 = 0$. Equation 4 is a fractional program, with a concave numerator and a convex, strictly positive denominator (excluding the singular point $\pi = \pi_{\mathrm{ref}}$, which cannot be optimal under Assumption 3.1). As such, we can draw on the rich literature on concave-convex fractional programming (see, e.g., Dinkelbach, 1967; Schaible, 1976), which shows that while such fractional optimization problems are not concave, they are quasi-concave and can be solved iteratively through a sequence of concave optimization problems.

Theorem 3.3. *Suppose that for each iteration k, the quantities $\pi^{(k)}$ and $\beta^{(k)}$ are chosen by $\pi^{(k)} \in \mathrm{argmax}_\pi \left\{\mathbb{E}_{y\sim\pi(\cdot|x)}[r(x,y)] - \beta^{(k)}\mathbb{E}[\mathrm{KL}(\pi(\cdot|x),\pi_{\mathrm{ref}}(\cdot|x))]\right\}$ and $\beta^{(k+1)} \leftarrow \frac{\mathbb{E}_{y\sim\pi^{(k)}(\cdot|x)}[r(x,y)]}{\mathbb{E}[\mathrm{KL}(\pi^{(k)}(\cdot|x),\pi_{\mathrm{ref}}(\cdot|x))]}$. Then it holds that $\beta^{(k)} \to \beta^\star$, where $\pi_{\beta^\star} \in \mathrm{argmax}_\pi\{\mathbb{E}_{y\sim\pi(\cdot|x)}[r(x,y)] - \beta^\star\mathbb{E}[\mathrm{KL}(\pi(\cdot \mid x),\pi_{\mathrm{ref}}(\cdot \mid x)]\}$ solves (4).*

Notice that $\pi_{\beta^\star}$ is precisely the optimizer of the RL fine-tuning objective in Equation 1 with regularization coefficient $\beta^\star$. In other words, the optimal policy for the agent is the reference policy π_{ref} tilted by $\beta^\star$ as in 2.

Discussion. Taken together, the results in this section imply that the agent playing policy $\pi_{\beta^\star}$ and the monitor testing for policy $\pi_{\beta^\star}$ is a Nash equilibrium in the sequential detection game. The

[2] If the latter half of Assumption 3.1 does not hold, the agent can play π_{ref} and collect infinite reward.

ALGORITHM 1: Stochastic bisection method

Input: $\epsilon > 0, \delta \in (0,1), \beta_{\text{high}} \in \mathbb{R}_+, n_0 \in \mathbb{N}, \beta_{\text{low}} := 0$
while $\beta_{\text{high}} - \beta_{\text{low}} > \epsilon$ **do**
  $\widehat{\beta} = \frac{1}{2}(\beta_{\text{high}} + \beta_{\text{low}}), n = n_0, \pi_{\widehat{\beta}} \leftarrow \texttt{RL-Oracle}(\pi_{\text{ref}}, \widehat{\beta})$
  while True **do**
    if $\widehat{M}_n(\widehat{\beta}) + \text{rad}(\widehat{\beta}, n, \delta) < 0$ **then**
      $\beta_{\text{high}} \leftarrow \widehat{\beta}$; break
    else if $\widehat{M}_n(\widehat{\beta}) - \text{rad}(\widehat{\beta}, n, \delta) > 0$ **then**
      $\beta_{\text{low}} \leftarrow \widehat{\beta}$; break
    else
      $n \leftarrow 2n$
return $\texttt{RL-Oracle}(\pi_{\text{ref}}, \beta_{\text{low}})$

equilibrium characterization yields a game-theoretic derivation for KL-regularized RL, and provides an operational interpretation of the KL penalty as the statistical cost of becoming distinguishable from the reference policy. In particular, the optimal regularization coefficient $\beta^\star$ is the maximum achievable reward per unit of statistical distinguishability, and can be interpreted as the shadow price of distinguishability.

This equilibrium characterization is also robust to tie-breaking. In Appendix A.1, we show that while $\pi_{\beta^\star}$ is not a unique agent best-response, the monitor's SPRT with $\mathcal{H}_1 : \pi = \pi_{\beta^\star}$ is still an α-correct, power-one test for *any* agent best response. Therefore, the monitor's test remains valid even if the agent breaks ties in an unknown or arbitrary way.

4 Reduction to KL-Regularized RL

We now turn our attention from characterizing the equilibrium policy $\pi_{\beta^\star}$ to learning it. Traditional fractional programming approaches cannot be applied off-the-shelf, as they rely on the ability to evaluate quantities like $\mathbb{E}_{y\sim\pi(\cdot|x)}[r(x,y)]$ and $\mathbb{E}[\text{KL}(\pi(\cdot|x), \pi_{\text{ref}}(\cdot|x))]$ exactly for an arbitrary policy π, which is generally not feasible when π is an LLM. Instead, we show that $\beta^\star$ can be characterized as the unique root of a one-dimensional function and estimate it from data using a stochastic bisection procedure, in the spirit of stochastic root-finding methods (Robbins & Monro, 1951).

Consider the function $M(\beta) := \mathbb{E}_{y\sim\pi_\beta(\cdot|x)}[r(x,y)] - \beta \cdot \mathbb{E}[\text{KL}(\pi_\beta(\cdot|x), \pi_{\text{ref}}(\cdot|x))]$. Intuitively, $M(\beta)$ measures whether the reward that is obtained by deviating from the reference policy exceeds the statistical cost imposed by the KL penalty for the given β. When β is too small, the reward dominates and $M(\beta) > 0$. When β is too large, the KL penalty dominates and $M(\beta) < 0$. The equilibrium coefficient $\beta^\star$ is precisely the point where these forces balance. Therefore, computing $\beta^\star$ reduces to finding the unique root of $M(\beta)$. This justification follows from an analysis of Dinkelbach (1967) (see Lemma A.1).

Since $M(\beta)$ is not available in closed form, we use $\widehat{\mathbb{E}}$ to denote the sample average and instead consider the empirical estimate of $M(\beta)$ using $n \geq 1$ samples, i.e. $\widehat{M}_n(\beta) := \widehat{\mathbb{E}}_{y\sim\pi_\beta(\cdot|x)}[r(x,y)] - \beta \cdot \widehat{\mathbb{E}}[\text{KL}(\pi_\beta(\cdot|x), \pi_{\text{ref}}(\cdot|x))]$.

Our algorithm proceeds as follows: Set $\beta_{\text{low}} = 0$. Given an upper bound $\beta_{\text{high}} \geq \beta^\star$, set $\widehat{\beta} := 0.5(\beta_{\text{high}} + \beta_{\text{low}})$ and collect enough samples n such that $\widehat{M}_n(\widehat{\beta}) \pm \text{rad}(\widehat{\beta}, n, \delta)$ is bounded away from zero, where $\text{rad}(\widehat{\beta}, n, \delta)$ is a confidence radius that holds with probability $1 - \delta$. Depending on the sign of $\widehat{M}_n(\widehat{\beta})$, set either $\beta_{\text{high}} \leftarrow \widehat{\beta}$ or $\beta_{\text{low}} \leftarrow \widehat{\beta}$, and repeat this process until $|\beta_{\text{high}} - \beta_{\text{low}}|$ is within the desired precision ϵ. The full procedure is outlined in Algorithm 1.

Theorem 4.1. *If $\beta_{\text{high}} \geq \beta^\star$ and for every β queried by Algorithm 1 $\mathbb{P}(\exists\ n \geq 1 : \left|\widehat{M}_n(\beta) - M(\beta)\right| > \text{rad}(\beta, n, \delta)) \leq \delta$, then with probability at least $1 - \delta \log_2(\beta_{\text{high}}/\epsilon)$, Algorithm 1 returns a policy $\pi_{\beta_{\text{low}}}$ satisfying $0 \leq \beta^\star - \beta_{\text{low}} \leq \epsilon$ in $\log_2(\beta_{\text{high}}/\epsilon)$ bisection iterations.*

We instantiate Algorithm 1 with specific choices of $(\beta_{\text{high}}, \text{rad}(\cdot,\cdot,\cdot))$ in the Appendix (Corollaries B.5 and B.7). In either instantiation, Algorithm 1 needs $n = O(M(\widehat{\beta})^{-2})$ samples in each iteration to determine the sign of $M(\widehat{\beta})$ with high probability. Therefore, as $\widehat{\beta} \to \beta^\star$ and $|M(\widehat{\beta})| \to 0$,

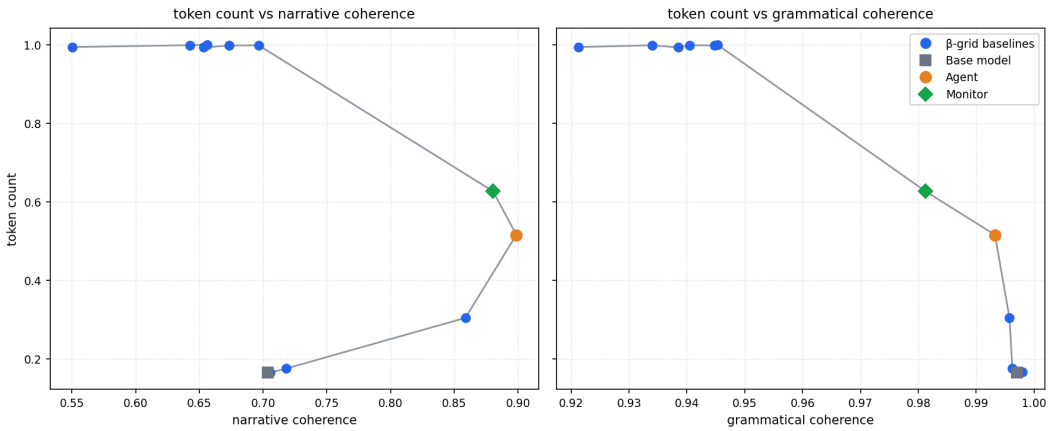


Figure 1: Reward–retention trade-offs for `Qwen3-8B` with shift $\rho = 0.1$. The vertical axis reports response length in tokens, and the horizontal axes report narrative and grammatical coherence, as judged by `GPT-5-mini` (Singh et al., 2025). Blue circles show policies trained on the compute-matched exponential grid over β, the gray square denotes the reference policy; and the orange circle and green diamond denote two independent stochastic-bisection runs. Each point is the average of 400 independent responses. Higher values are better on both axes.

we should expect the number of samples we need in each iteration of stochastic bisection to increase. As a practical relaxation, we forego the confidence sequence in our experiments and bisect based on the sign of $\widehat{M}_n(\widehat{\beta})$ after observing a sufficiently large number of samples (e.g., $n = 4096$).

Algorithm 1 posits access to an "RL oracle" which, given a reference policy π_{ref} and regularization parameter β, returns the KL-regularized tilt π_β. While exact, polynomial-time oracles generally do not exist due to the difficulty of computing the partition function, there is a rich literature of approximate oracles that have been designed to solve the KL-regularized RL problem (e.g., PPO, GRPO, and variants thereof). Consequently, our theoretical guarantees for Algorithm 1 should be interpreted as characterizing the complexity of identifying the equilibrium coefficient, assuming that each RL subproblem can be solved sufficiently accurately.

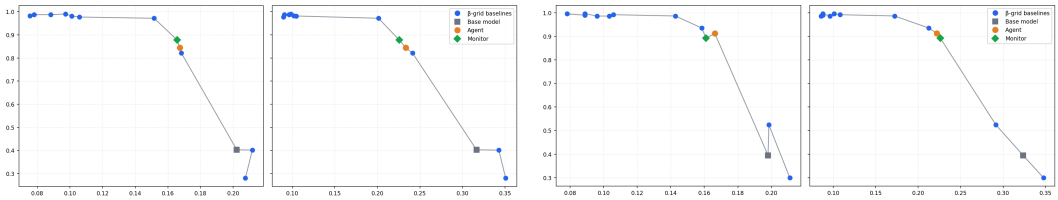
Warm-start version. It is possible to "warm start" Algorithm 1 given an arbitrary setting of β_{high} as follows: Compute $\widehat{M}_n(\beta_{\text{high}}) + \text{rad}(\beta_{\text{high}}, n, \delta)$. If it is bounded below zero, run Algorithm 1 as normal. Otherwise set $\beta_{\text{low}} \leftarrow \beta_{\text{high}}$, $\beta_{\text{high}} \leftarrow 2\beta_{\text{high}}$, and repeat. This procedure will produce an ε-approximate solution with high probability. Furthermore, it makes $\log_2(\beta_{\text{high}}/\epsilon)$ calls to the RL oracle if β_{high} is a valid upper bound on $\beta^\star$. If it is not, the number of oracle calls is still $O\left(\log_2\left(\sigma^2/(|\mu_{\text{ref}}|\epsilon)\right)\right)$ if rewards are σ-sub-Gaussian by Lemma B.3.

Finally, it is worth noting that in our game formulation, even if the agent and the monitor run separate copies of Algorithm 1, the agent and monitor utilities will still be close to their equilibrium values. See Appendix B.1 for more details.

5 Experiments

We evaluate our methods empirically in two settings: a continual learning task and a model auditing task. In the continual learning task, the goal is to improve performance on a new target reward while preserving existing behavioral properties of the reference policy. In the model auditing task, an auditor tries to detect whether a model has been strategically fine-tuned from its outputs. Taken together, these experiments test the two main operational interpretations of our framework: stochastic bisection as a replacement for manual KL coefficient search, and the equilibrium likelihood-ratio test as a practical auditing rule.

Experiment setup. The two sets of experiments use the same setup. In each round, the language model is given the same prompt ("Once upon a time") and generates a completion, with the agent receiving reward equal to the number of characters in the response.



(a) `Llama-3.2-1B` with shift $\rho = 0.1$. (b) `Llama-3.2-1B` with shift $\rho = 0.2$.

Figure 2: Reward–retention trade-offs for `Llama-3.2-1B` with shift $\rho = 0.1$ (left) and $\rho = 0.2$ (right). Plotting conventions, axis labels, and plot titles are the same as in Figure 1. In both settings, Algorithm 1 selects policies near the elbow of the empirical trade-off curve traced out by the compute-matched β grid.

Before bisection, the reward is calibrated by sampling completions from the reference policy and subtracting off the raw mean reward, plus a fixed margin ρ. Under the reference policy, the calibrated reward therefore has mean approximately equal to $-\rho$. The calibrated reward is used throughout KL-constrained fine-tuning and bisection; because adding a constant to the reward does not change the optimal policy at any fixed β, this shift does not alter the fine-tuning objective. Evaluation metrics on Pareto plots are reported in the original, uncalibrated units.

5.1 Continual Learning

In our continual learning setting, the response-length reward function plays the role of the skill to be improved, while the reference policy's ability to produce coherent responses is the behavior to be retained. Although this is a single model update rather than a multi-task setting, it isolates the central continual learning trade-off: aggressively optimizing a new objective can alter or degrade useful behaviors inherited from the reference policy.[3] For each trained policy, we measure the target behavior using token count and retention using both narrative and grammatical coherence.[4]

We compare Algorithm 1 against the standard practice of manually sweeping over the regularization coefficient. This baseline trains one policy for each value in an exponentially spaced grid over β, using the same number of fine-tuning runs as Algorithm 1. After inspecting the results, a practitioner could select whichever grid point best matches their desired balance between length and coherence.

Our results are summarized in Figures 1 and 2. Higher values are desirable on both axes, so policies toward the upper-right corner of each panel achieve more favorable reward–retention trade-offs. The points labeled `Agent` and `Monitor` are produced by two independent runs of Algorithm 1. We retain these labels for consistency with the auditing experiment in Section 5.2; in this experiment, the two points illustrate the variation induced by GRPO RL oracle calls. Blue circles correspond to points on an exponential grid between 0 and β_{high}; the reference policy is denoted by a gray square.

Figure 1 shows the results for `Qwen3-8B` with shift $\rho = 0.1$. Policies that generate the longest responses exhibit lower coherence, while policies that preserve the highest narrative and grammatical coherence produce shorter responses. Interestingly, narrative coherence actually increases with token count up to a point, before decreasing once the token count gets too large. Each run of Algorithm 1 selects an intermediate policy near the bends of the empirical trade-off curves, in contrast to the compute-matched grid, which results in models closer to either extreme.

Figure 2 shows the corresponding results for `Llama-3.2-1B` with shifts $\rho = 0.1$ and $\rho = 0.2$. The same qualitative pattern holds in both settings: the bisection policies lie near the "elbow" of the curve traced out by the grid, avoiding both the high-retention regime in which little progress is made on the length objective, and the high-length regime in which coherence falls sharply.[5]

[3] This is commonly observed in the literature on RLHF and is often referred to as the "alignment tax" (e.g., Ouyang et al., 2022; Lin et al., 2024; Zhang et al., 2024; Jang et al., 2024).

[4] We used `GPT-5-mini` and `GPT-5-nano` as judges to evaluate narrative and grammatical coherence (Zheng et al., 2023). All main body plots use `GPT-5-mini`. Results using `GPT-5-nano` are very similar and are in Appendix C.

[5] `Llama-3.2-1B` scores considerably lower than `Qwen3-8B` in terms of narrative and grammatical coherence for all values of β, and does not exhibit the same improvement in narrative coherence as token count increases. This is likely due to it being a much smaller model.

Table 1: Model-auditing results at level $\alpha = 0.05$. π_A SPRT uses the true deployed policy and is thus an oracle comparator; π_M SPRT uses the monitor's independently trained equilibrium policy. Lower stopping times and false positive rates are better. Averages are taken over 100 samples.

Model and margin	Metric	π_A SPRT	π_M SPRT	Grid mixture
`Qwen3-8B`, $\rho = 0.1$	Avg. stopping time	1.00	1.08	1.02
	False positive rate	0.00	0.00	0.09
`Llama-3.2-1B`, $\rho = 0.01$	Avg. stopping time	2.49	6.61	20.36
	False positive rate	0.04	0.06	0.20
`Llama-3.2-1B`, $\rho = 0.1$	Avg. stopping time	1.89	5.31	9.47
	False positive rate	0.03	0.01	0.15
`Llama-3.2-1B`, $\rho = 0.2$	Avg. stopping time	1.75	4.40	6.23
	False positive rate	0.04	0.03	0.09

Across all settings, we find that Algorithm 1 selects a policy near the elbow of the empirical reward–retention frontier. In some instances (e.g., Fig. 1, left), Algorithm 1 finds desirable regions of the frontier that are not reached by grid search. However even when Algorithm 1 and grid search find similar parts of the frontier, grid search requires an additional post-hoc selection step, whereas Algorithm 1 resolves this trade-off adaptively by concentrating its fine-tuning runs around $\beta^\star$.

5.2 Model Auditing

Consider a model-hosting platform that sets inference costs such that they are profitable on a per-token basis. Such a platform has an economic incentive to fine-tune its models to produce unnecessarily long responses, while remaining difficult to distinguish from the advertised model (Velasco et al., 2026; Cao et al., 2026).

A third-party auditor (e.g., a watchdog organization) can easily observe the model's outputs, but verifying the parameters of the deployed model is significantly more difficult. Nevertheless, the auditor may want to detect this strategic behavior when it exists, while controlling the probability of falsely flagging the model when it does not. Our theory precisely characterizes the equilibrium of this game, and gives the auditor a way to empirically test for this type of bad behavior.

Starting from the same reference policy and using the token length reward described above, the agent and monitor independently run Algorithm 1 with the same RL oracle budget. Let π_A denote the policy obtained by the agent and let π_M denote the policy obtained (independently) by the monitor/auditor. The agent deploys π_{A}, whereas the monitor uses π_{M} as the alternative hypothesis in its sequential test. In other words, the auditor knows the reference policy and the incentive being optimized, but they do not know the agent's realized policy or share its fine-tuning randomness.

Each auditing trial uses a source policy $\pi_{\mathrm{src}} \in \{\pi_A, \pi_{\mathrm{ref}}\}$. Sampling from π_A represents the presence of strategic fine-tuning, whereas π_{ref} represents honest deployment. At time t, a completion $y_t \sim \pi_{\mathrm{src}}(\cdot|x)$ is generated and scored under both π_M and π_{ref}. The monitor uses the SPRT which rejects if $L_t > \log(1/\alpha)$ for any $t \geq 1$, where $L_t := \sum_{s=1}^{t} \log(\pi_M(y_s|x)/\pi_{\mathrm{ref}}(y_s|x))$ and the log likelihood is computed by summing the token-level log probabilities along the realized completion.

We compare the equilibrium test against a uniform mixture likelihood ratio test (described in Appendix C). We also compare against the "skyline" SPRT which uses the agent's actual deployed policy as its alternative. While the agent's policy is not available to the auditor, it gives a lower bound on the detection speed attainable by any SPRT. Finally, all tests are evaluated on the same sequence of completions from π_{src}; the candidate policies are used only to score these observations.

Table 1 summarizes the results. For `Qwen3-8B`, the strategic modification is sufficiently pronounced that every test detects it after approximately one observation. The agent SPRT stops after exactly one completion on average, while the independently trained monitor stops after 1.08 completions. The uniform mixture test obtains a slightly better stopping time of 1.02, albeit at the cost of a false positive rate that is higher than the 0.05 threshold. The differences between testing strategies are clearer for `Llama-3.2-1B`. As expected, the oracle agent SPRT detects the deployed policy fastest. Nevertheless, the monitor test is significantly faster than the uniform mixture test at every

calibration margin. The composite grid tests also reject more frequently under the reference policy, with empirical false-positive rates between 0.08 and 0.2, compared to the monitor's false positive rate which is between 0.01 and 0.06.

6 Related Work

Reward–KL trade-offs in RL fine-tuning. KL regularization is a central component of modern LLM fine-tuning, particularly in reinforcement learning from human feedback, where it is used to improve a target reward while limiting drift from a reference model. Several works study the trade-off between reward and KL regularization empirically (see, e.g., Bai et al., 2022; Gao et al., 2023). Korbak et al. (2022) provide a complementary theoretical interpretation of KL-regularized RL through the lens of variational inference, where the reference model is analogous to a prior and the reward function provides evidence.

Game-theoretic applications of hypothesis testing. A small but growing line of work uses statistical hypothesis testing procedures as strategic instruments in games. Gauthier et al. (2026) use sequential tests to monitor whether agents in a multi-agent system are behaving consistently with equilibrium play. Capitaine et al. (2026) use sequential testing in infinitely repeated games, where they show a "test-then-punish strategy" can sustain cooperation by detecting deviations and triggering punishment. Hu et al. (2024) are closer in spirit, as they also study game-theoretic hypothesis testing against strategic evasion. They model an attacker who manipulates observations to evade a Neyman–Pearson test and characterize equilibrium distortion strategies and detector responses in Stackelberg and signaling games. Our settings differ in both the object being manipulated and the resulting equilibrium structure.

Continual learning. Sun et al. (2019) uses LLMs to generate pseudo-samples of previous tasks for training alongside data for the new task. Razdaibiedina et al. (2023) avoid modifying the reference policy by learning task-specific "soft prompts" and concatenating newly learned prompts with earlier ones. Zhang et al. (2024) study continual learning in RLHF and use sample-wise "balance weights" to regulate the trade-off between policy learning and knowledge retention. Like us, Zhang et al. (2023a) explicitly consider policies that are tilted from a previous task's policy, but they treat the KL coefficient as a hyperparameter.

AI auditing. Several recent works formalize LLM auditing as a statistical testing problem over model outputs. Gao et al. (2025) introduce model equality testing, which asks whether a black-box API is serving the same model as a claimed reference model. Richter et al. (2025) frame auditing for behavioral shifts in LLMs as a sequential testing problem. Our auditing setup is aligned with this statistical view, but differs in that the audited model is itself trained strategically to avoid detection.

Strategic learning. More broadly, our work belongs to the literature on strategic machine learning, where the deployment of a learning algorithm influences the behavior of other agents. Strategic classification studies settings in which individuals manipulate their features to receive favorable decisions (Hardt et al., 2016), while performative prediction (Perdomo et al., 2020) analyzes learning problems where the underlying data distribution changes in response to the deployed model. Subsequent work extends these ideas to multi-step settings, where strategic responses can accumulate over time (Harris et al., 2021; Brown et al., 2022), like in our setting.

7 Conclusions

We introduced the sequential detection game, a framework for RL fine-tuning in which an agent seeks to maximize reward while remaining difficult to distinguish from a reference policy. We showed that the Nash equilibrium recovers the KL-regularized RL objective while determining the regularization coefficient endogenously, giving it an explicit operational interpretation rather than treating it as a tunable hyperparameter. We then developed a stochastic bisection algorithm for estimating the equilibrium coefficient and presented empirical evidence that it yields competitive reward–retention trade-offs in a continual learning setting, while also enabling principled auditing of LLM APIs. Finally, while further work is needed to close the gap between our equilibrium analysis and large-scale RL fine-tuning, we hope that this work will encourage practitioners to use methods whose hyperparameters are derived from principled objectives, rather than chosen heuristically.

AI USE STATEMENT

In this work, we used generative AI tools for finding related work, polishing the writing, some code generation for experiments, and aiding in proving some mathematical claims. We have not used generative AI tools for drafting sections of the paper, developing theoretical models or conceptual frameworks, interpreting results, developing a synthetic dataset, or for anything else. We have reviewed all AI-assisted work. We manually checked that all LLM-suggested references were relevant and we did our own manual literature survey as well, LLM-generated code was verified and tested for correctness, and LLM-generated proofs were manually rewritten and checked. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Gavin Brown, Shlomi Hod, and Iden Kalemaj. Performative Prediction in a Stateful World. In *International Conference on Artificial Intelligence and Statistics*, pp. 6045–6061. PMLR, 2022.

Yuhan Cao, Yu Wang, Sitong Liu, Miao Li, Yixin Tao, and Tianxing He. Pay for the Second-best Service: A Game-theoretic Approach Against Dishonest LLM Providers. In *Proceedings of the ACM Web Conference 2026*, pp. 249–260, 2026.

Aymeric Capitaine, Antoine Scheid, Etienne Boursier, Alain Durmus, and Michael I Jordan. Test-then-Punish: A Statistical Approach to Repeated Games. *arXiv preprint arXiv:2603.05619*, 2026.

Werner Dinkelbach. On Nonlinear Fractional Programming. *Management Science*, 13(7):492–498, 1967.

Monroe D Donsker and SR Srinivasa Varadhan. On a Variational Formula for the Principal Eigenvalue for Operators with Maximum Principle. *Proceedings of the National Academy of Sciences*, 72(3):780–783, 1975.

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.

Leo Gao, John Schulman, and Jacob Hilton. Scaling Laws for Reward Model Overoptimization. In *International Conference on Machine Learning*, pp. 10835–10866. PMLR, 2023.

Etienne Gauthier, Francis Bach, and Michael I Jordan. Betting on Equilibrium: Monitoring Strategic Behavior in Multi-Agent Systems. *arXiv preprint arXiv:2601.05427*, 2026.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic Classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pp. 111–122, 2016.

Keegan Harris, Hoda Heidari, and Steven Z Wu. Stateful Strategic Regression. *Advances in Neural Information Processing Systems*, 34:28728–28741, 2021.

Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, Nonparametric, Nonasymptotic Confidence Sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank Adaptation of Large Language Models. *International Conference on Learning Representations*, 1(2):3, 2022.

Yinan Hu, Juntao Chen, and Quanyan Zhu. Game-theoretic Neyman-Pearson Detection to Combat Strategic Evasion. *IEEE Transactions on Information Forensics and Security*, 20:516–530, 2024.

Youngsoo Jang, Geon-Hyeong Kim, Byoungjip Kim, Yu Jin Kim, Honglak Lee, and Moontae Lee. Degeneration-free Policy Optimization: RL Fine-Tuning for Language Models without Degeneration. In *International Conference on Machine Learning*, 2024.

Natasha Jaques, Shixiang Gu, Dzmitry Bahdanau, José Miguel Hernández-Lobato, Richard E Turner, and Douglas Eck. Sequence Tutor: Conservative Fine-Tuning of Sequence Generation Models with KL-Control. In *International Conference on Machine Learning*, pp. 1645–1654. PMLR, 2017.

Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. Way Off-policy Batch Deep Reinforcement Learning of Implicit Human Preferences in Dialog. *arXiv preprint arXiv:1907.00456*, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Tomasz Korbak, Ethan Perez, and Christopher Buckley. RL with KL Penalties is Better Viewed as Bayesian Inference. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1083–1091, 2022.

Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, et al. Mitigating the Alignment Tax of RLHF. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 580–606, 2024.

Meta AI. Llama-3.2-1B. `https://huggingface.co/meta-llama/Llama-3.2-1B`, 2024. Model Card. Accessed: 2026-06-10.

Gergely Neu, Anders Jonsson, and Vicenç Gómez. A Unified View of Entropy-regularized Markov Decision Processes. *arXiv preprint arXiv:1705.07798*, 2017.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training Language Models to Follow Instructions with Human Feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative Prediction. In *International Conference on Machine Learning*, pp. 7599–7609. PMLR, 2020.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

Anastasia Razdaibiedina, Yuning Mao, Rui Hou, Madian Khabsa, Mike Lewis, and Amjad Almahairi. Progressive Prompts: Continual Learning for Language Models. *arXiv preprint arXiv:2301.12314*, 2023.

Leo Richter, Xuanli He, Pasquale Minervini, and Matt Kusner. An Auditing Test to Detect Behavioral Shift in Language Models. In *International Conference on Learning Representations*, volume 2025, pp. 49313–49339, 2025.

Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, pp. 400–407, 1951.

Herbert Robbins and David Siegmund. The Expected Sample Size of Some Tests of Power One. *The Annals of Statistics*, pp. 415–436, 1974.

Siegfried Schaible. Fractional Programming. II, on Dinkelbach's Algorithm. *Management Science*, 22(8):868–873, 1976.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300*, 2024.

David Siegmund. *Sequential Analysis: Tests and Confidence Intervals*. Springer Series in Statistics. Springer, New York, NY, 1985. ISBN 978-0-387-96134-7. doi: 10.1007/978-1-4757-1862-1.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. OpenAI GPT-5 System Card. *arXiv preprint arXiv:2601.03267*, 2025.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to Summarize with Human Feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

Fan-Keng Sun, Cheng-Hao Ho, and Hung-Yi Lee. Lamol: Language Modeling for Lifelong Language Learning. *arXiv preprint arXiv:1909.03329*, 2019.

Yunhao Tang, Daniel Zhaohan Guo, Zeyu Zheng, Daniele Calandriello, Yuan Cao, Eugene Tarassov, Rémi Munos, Bernardo Ávila Pires, Michal Valko, Yong Cheng, et al. Understanding the Performance Gap Between Online and Offline Alignment Algorithms. *arXiv preprint arXiv:2405.08448*, 2024.

Thinking Machines Lab. Tinker: A Flexible API for Efficiently Fine-Tuning Open-Source Models with LoRA. `https://thinkingmachines.ai/tinker/`, 2025. Accessed: 2026-06-18.

Ander Artola Velasco, Dimitrios Rontogiannis, Stratis Tsirtsis, and Manuel Gomez-Rodriguez. Test-Time Compute Games. *arXiv preprint arXiv:2601.21839*, 2026.

Abraham Wald. Sequential Tests of Statistical Hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945.

Abraham Wald. *Sequential Analysis*. Wiley, New York, 1947.

Abraham Wald and Jacob Wolfowitz. Optimum Character of the Sequential Probability Ratio Test. *The Annals of Mathematical Statistics*, pp. 326–339, 1948.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025.

Han Zhang, Lin Gui, Yuanzhao Zhai, Hui Wang, Yu Lei, and Ruifeng Xu. COPF: Continual Learning Human Preference Through Optimal Policy Fitting. *arXiv preprint arXiv:2310.15694*, 2023a.

Han Zhang, Yu Lei, Lin Gui, Min Yang, Yulan He, Hui Wang, and Ruifeng Xu. CPPO: Continual Learning for Reinforcement Learning with Human Feedback. In *The Twelfth International Conference on Learning Representations*, 2024.

Tianjun Zhang, Fangchen Liu, Justin Wong, Pieter Abbeel, and Joseph E Gonzalez. The Wisdom of Hindsight Makes Language Models Better Instruction Followers. In *International Conference on Machine Learning*, pp. 41414–41428. PMLR, 2023b.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning Language Models from Human Preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A Appendix for Section 3: The Sequential Detection Game

Lemma A.1 (Dinkelbach (1967))**.** *The following properties are true for $M(\beta)$:*

1. $M(\beta^\star) = 0$
2. $M(\beta) < 0$ *for all* $\beta > \beta^\star$
3. $M(\beta) > 0$ *for all* $\beta < \beta^\star$
4. $M(\beta)$ *is convex.*

Lemma A.2. *Let* $\tau_\alpha := \inf\{\tau \ : \ \sum_{t=1}^{\tau} \log \frac{dQ}{dP}(X_t) \geq \log(1/\alpha)\}$ *and assume that* $Q << P$, $\mathbb{E}_R[\tau_\alpha] < \infty$, *and* $\mathbb{E}_R[\log \frac{dQ}{dP}(X)] < \infty$. *Then*

$$\mathbb{E}_R[\tau_\alpha] \geq \frac{\log(1/\alpha)}{\mathbb{E}_R[\log \frac{dQ}{dP}(X)]}.$$

Proof. By Wald's identity,

$$\mathbb{E}_R\left[\sum_{t=1}^{\tau_\alpha} \log \frac{dQ}{dP}(X_t)\right] = \mathbb{E}_R[\tau_\alpha] \cdot \mathbb{E}_R\left[\log \frac{dQ}{dP}(X)\right].$$

Our definition of τ_α implies that

$$\mathbb{E}_R[\tau_\alpha] \cdot \mathbb{E}_R\left[\log \frac{dQ}{dP}(X)\right] \geq \log(1/\alpha).$$

Dividing both sides by $\mathbb{E}_R[\log \frac{dQ}{dP}(X)]$ obtains the result. □

A.1 Robustness to tie-breaking

While $\pi_{\beta^\star}$ is an optimal policy for the agent to play against the monitor's equilibrium SPRT, it is not the only one.

Theorem A.3. *Let* $d(\pi) := \mathbb{E}[\mathrm{KL}(\pi(\cdot|x), \pi_{\mathrm{ref}}(\cdot|x)) - \mathrm{KL}(\pi(\cdot|x), \pi_{\beta^\star}(\cdot|x))]$. *Suppose that* $\tau_\alpha := \inf\left\{\tau \ : \ \sum_{t=1}^{\tau} \log \frac{\pi_{\beta^\star}(x_t, y_t)}{\pi_{\mathrm{ref}}(x_t, y_t)} \geq \log(1/\alpha)\right\}$ *and the agent plays policy* π *such that* $d(\pi) > 0$. *We have that*

$$\max_\pi \mathbb{E}_{x_t \sim \mathcal{D}, y_t \sim \pi(\cdot|x_t)}\left[\sum_{t=1}^{\tau_\alpha} r(x_t, y_t)\right] \geq \beta^\star \log(1/\alpha).$$

Proof. We have that

$$\max_\pi \mathbb{E}_{x_t \sim \mathcal{D}, y_t \sim \pi(\cdot|x_t)}\left[\sum_{t=1}^{\tau_\alpha} r(x_t, y_t)\right] = \max_\pi \mathbb{E}_{x_t \sim \mathcal{D}, y_t \sim \pi(\cdot|x_t)}[\tau_\alpha] \cdot \mathbb{E}_{y \sim \pi(\cdot|x)}[r(x, y)],$$

where we know that $d(\pi) > 0$ (and hence $\mathbb{E}_{y \sim \pi(\cdot|x)}[\tau_\alpha] < \infty$) by Assumption 3.1. Applying Lemma A.2, we have that

$$\max_\pi \mathbb{E}_{x_t \sim \mathcal{D}, y_t \sim \pi(\cdot|x_t)}\left[\sum_{t=1}^{\tau_\alpha} r(x_t, y_t)\right] \geq \max_\pi \frac{\mathbb{E}_{y \sim \pi(\cdot|x)}[r(x, y)] \log(1/\alpha)}{\mathbb{E}_{y \sim \pi(\cdot|x)}\left[\log \frac{\pi_{\beta^\star}(y|x)}{\pi_{\mathrm{ref}}(y|x)}\right]}$$

The joint distribution induced by $\pi_{\beta^\star}$ is absolutely continuous with respect to π_{ref} since $\pi_{\beta^\star}$ is a tilt of π_{ref}. Focusing on the denominator, we have that

$$\mathbb{E}_{y \sim \pi(\cdot|x)}\left[\log \frac{\pi_{\beta^\star}(y|x)}{\pi_{\mathrm{ref}}(y|x)}\right] = \frac{1}{\beta^\star}\mathbb{E}_{y \sim \pi(\cdot|x)}[r(x, y)] - \mathbb{E}[\log Z_{\beta^\star}(x)]$$

Setting this aside for a moment, recall that

$$\begin{aligned} M(\beta^\star) &= \mathbb{E}_{y\sim\pi_{\beta^\star}(\cdot|x)}[r(x,y)] - \beta^\star \mathbb{E}[\mathrm{KL}(\pi_{\beta^\star}(\cdot|x), \pi_{\mathrm{ref}}(\cdot|x))] \\ &= \beta^\star \mathbb{E}[\log Z_{\beta^\star}(x)] = 0 \end{aligned}$$

where the last equality follows from Lemma A.1. Since we know that $\beta^\star > 0$ by Assumption 3.1, we can conclude that $\mathbb{E}[\log Z_{\beta^\star}(x)] = 0$. Substituting this all back in, we can conclude that

$$\frac{\mathbb{E}_{y\sim\pi(\cdot|x)}[r(x,y)]\log(1/\alpha)}{\mathbb{E}_{y\sim\pi(\cdot|x)}\left[\log\frac{\pi_{\beta^\star}(y|x)}{\pi_{\mathrm{ref}}(y|x)}\right]} = \beta^\star \log(1/\alpha),$$

which is independent of the agent's policy π as long as $d(\pi) > 0$. □

Therefore, we can conclude that if the monitor is deploying a sequential likelihood ratio test between $\pi_{\beta^\star}$ and π_{ref}, the agent can guarantee themselves utility at least $\beta^\star \log(1/\alpha)$ by playing any policy π such that $d(\pi) > 0$ (including $\pi = \pi_{\beta^\star}$). Moreover as $\alpha \downarrow 0$, this is the optimal utility attainable for the agent under this test.

The test with stopping time $\tau_\alpha := \inf\{\tau \;:\; \sum_{t=1}^{\tau} \log\frac{\pi_{\beta^\star}(x_t,y_t)}{\pi_{\mathrm{ref}}(x_t,y_t)} \geq \log(1/\alpha)\}$ is also a valid sequential test for testing the null hypothesis $\mathcal{H}_0 \;:\; \pi = \pi_{\mathrm{ref}}$ versus the *composite* alternative $\mathcal{H}_1 \;:\; \pi$ is s.t. $d(\pi) > 0$.

To see this, let $L_t = \log\frac{\pi_{\beta^\star}(y_t|x_t)}{\pi_{\mathrm{ref}}(y_t|x_t)}$ and note that under $\mathcal{H}_0$, $\exp\left(\sum_{s=1}^{t} L_s\right)$ is a non-negative martingale with mean 1. By Ville's inequality, we have that

$$\mathbb{P}_{\mathcal{H}_0}\left(\sup_{t>0}\exp\left(\sum_{s=1}^{t} L_s\right) \geq \frac{1}{\alpha}\right) \leq \alpha,$$

which implies that $\mathbb{P}_{\mathcal{H}_0}(\tau_\alpha < \infty) \leq \alpha$.

Under $\mathcal{H}_1$, $\mathbb{E}_{y\sim\pi(\cdot|x)}[L_t] = d(\pi) > 0$, so by the law of large numbers we have that $\sum_{s=1}^{t} L_s \to \infty$ almost surely, and so $\tau_\alpha < \infty$ almost surely.

Therefore, even if the agent breaks ties among best-response policies in an arbitrary or unknown way, this remains a valid power-1 sequential test for the composite alternative $\mathcal{H}_1 \;:\; \pi$ is s.t. $d(\pi) > 0$.

B Appendix for Section 4: Reduction to KL-Regularized RL

Assumption B.1 (Sub-Gaussian rewards)**.** *Rewards are sub-Gaussian with parameter σ, i.e., for any π and all $\lambda \in \mathbb{R}$,*

$$\log \mathbb{E}_{y\sim\pi(\cdot|x)}\left[\exp(\lambda(r(x,y) - \mu))\right] \leq \frac{\sigma^2\lambda^2}{2},$$

where $\mu := \mathbb{E}_{y\sim\pi(\cdot|x)}[r(x,y)]$.

Lemma B.2 (Howard et al. (2021))**.** *Given a sequence of i.i.d. observations $(X)_{n=1}^{\infty}$ from a σ-sub-Gaussian distribution, we have that*

$$\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X]\right| \leq 1.7\sigma\sqrt{\frac{0.72\log(10.4/\delta) + \log\log(2n)}{n}}$$

with probability at least $1-\delta$, simultaneously for every $n \geq 1$.

Lemma B.3. *Under Assumption B.1, $\beta^\star \leq \frac{\sigma^2}{2|\mu_{\mathrm{ref}}|}$.*

Proof. Consider the Donsker-Varadhan variational formula (Donsker & Varadhan, 1975), which says that for two distributions P, Q such that $P \ll Q$ and function f which is measurable,

$$\mathbb{E}_P[f] \leq \log \mathbb{E}_Q[\exp(f)] + \mathrm{KL}(P, Q).$$

Let P be the joint distribution over (x, y) such that $x \sim \mathcal{D}$, $y \sim \pi(\cdot|x)$ for some arbitrary policy π, Q be the joint distribution such that $x \sim \mathcal{D}$, $y \sim \pi_{\text{ref}}(\cdot|x)$, and $f = \lambda(r(x, y) - \mu_{\text{ref}})$ for some arbitrary $\lambda \in \mathbb{R}$. Plugging this in, we get that

$$\mathbb{E}_{y\sim\pi(\cdot|x)}[r(x,y)] - \mu_{\text{ref}} \leq \frac{1}{\lambda}\left(\log \mathbb{E}_{y\sim\pi_{\text{ref}}(\cdot|x)}[\exp(\lambda(r(x,y) - \mu_{\text{ref}}))] + \mathbb{E}[\text{KL}(\pi(\cdot|x), \pi_{\text{ref}}(\cdot|x))]\right).$$

Observe that $P \ll Q$ is implicit in our setting as if this fails to hold, then $\text{KL}(\pi\|\pi_{\text{ref}}) = \infty$ so such a policy π could never be optimal. Applying Assumption B.1, we have that

$$\begin{aligned}\mathbb{E}_{y\sim\pi(\cdot|x)}[r(x,y)] - \mu_{\text{ref}} &\leq \frac{1}{\lambda}\left(\frac{\sigma^2\lambda^2}{2} + \mathbb{E}[\text{KL}(\pi(\cdot|x), \pi_{\text{ref}}(\cdot|x))]\right) \\ &= \frac{\sigma^2\lambda}{2} + \frac{1}{\lambda}\mathbb{E}[\text{KL}(\pi(\cdot|x), \pi_{\text{ref}}(\cdot|x))].\end{aligned}$$

Since λ is arbitrary, we can pick it to minimize the right hand side. Applying first-order conditions, we get that the optimal choice of λ is

$$\lambda = \sqrt{\frac{2\mathbb{E}[\text{KL}(\pi(\cdot|x), \pi_{\text{ref}}(\cdot|x))]}{\sigma^2}}.$$

Plugging this into our bound, we get that

$$\mathbb{E}_{y\sim\pi(\cdot|x)}[r(x,y)] - \mu_{\text{ref}} \leq \sqrt{2\sigma^2\mathbb{E}[\text{KL}(\pi(\cdot|x), \pi_{\text{ref}}(\cdot|x))]}.$$

Rearranging terms and subtracting $\beta\mathbb{E}[\text{KL}(\pi(\cdot|x), \pi_{\text{ref}}(\cdot|x))]$ from each side of the inequality, we get

$$\begin{aligned}\mathbb{E}_{y\sim\pi(\cdot|x)}[r(x,y)] - \beta\mathbb{E}[\text{KL}(\pi(\cdot|x), \pi_{\text{ref}}(\cdot|x))] \leq \mu_{\text{ref}} &+ \sqrt{2}\sigma\sqrt{\mathbb{E}[\text{KL}(\pi(\cdot|x), \pi_{\text{ref}}(\cdot|x))]} \\ &- \beta\mathbb{E}[\text{KL}(\pi(\cdot|x), \pi_{\text{ref}}(\cdot|x))].\end{aligned}$$

Our choice of π was arbitrary, so we take the max over π on both sides to get

$$\begin{aligned}M(\beta) &\leq \mu_{\text{ref}} + \max_{\pi}\left(\sqrt{2}\sigma\sqrt{\mathbb{E}[\text{KL}(\pi(\cdot|x), \pi_{\text{ref}}(\cdot|x))]} - \beta\mathbb{E}[\text{KL}(\pi(\cdot|x), \pi_{\text{ref}}(\cdot|x))]\right) \\ &= \mu_{\text{ref}} + \max_{k\geq 0}(\sqrt{2}\sigma\sqrt{k} - \beta k).\end{aligned}$$

Since $\sqrt{2}\sigma\sqrt{k} - \beta k$ is concave in k, we can apply the first order condition to compute the maximum, which is obtained at $k = \frac{\sigma^2}{2\beta^2}$. Substituting this back in, we get that

$$M(\beta) \leq \mu_{\text{ref}} + \frac{\sigma^2}{2\beta}.$$

Using this inequality, we can see that $M(\beta) < 0$ if $\mu_{\text{ref}} + \frac{\sigma^2}{2\beta} < 0$, or equivalently, if $\beta > \frac{\sigma^2}{-2\mu_{\text{ref}}}$. Finally, taking the contrapositive, we get that if $M(\beta) \geq 0$, then $\beta \leq \frac{\sigma^2}{-2\mu_{\text{ref}}}$. Since $M(\beta^\star) = 0$, we can conclude that $\beta^\star \leq \frac{\sigma^2}{-2\mu_{\text{ref}}} = \frac{\sigma^2}{2|\mu_{\text{ref}}|}$. □

Theorem 4.1. *If $\beta_{\text{high}} \geq \beta^\star$ and for every β queried by Algorithm 1 $\mathbb{P}(\exists\ n \geq 1 : \left|\widehat{M}_n(\beta) - M(\beta)\right| > \text{rad}(\beta, n, \delta)) \leq \delta$, then with probability at least $1 - \delta\log_2(\beta_{\text{high}}/\epsilon)$, Algorithm 1 returns a policy $\pi_{\beta_{\text{low}}}$ satisfying $0 \leq \beta^\star - \beta_{\text{low}} \leq \epsilon$ in $\log_2(\beta_{\text{high}}/\epsilon)$ bisection iterations.*

Proof. The proof consists of two separate parts:

1. Showing the correctness of Algorithm 1, conditioned on the clean event
2. Bounding the number of loop iterations in Algorithm 1 (and therefore the number of high probability events)

Correctness: $\beta_{\text{low}} = 0$ is a valid lower bound on $\beta^\star$, since Assumption 3.1 implies that it is possible for the agent to get non-negative cumulative expected reward. $\beta_{\text{high}} = \frac{\sigma^2}{2|\mu_{\text{ref}}|}$ is a valid upper bound on $\beta^\star$ by Lemma B.3.

Under the clean event, $M(\widehat{\beta}) < 0$ if $\widehat{M}_n(\widehat{\beta}) + \text{rad}(\widehat{\beta}, n, \delta) < 0$, so we can conclude that $\beta_{\text{high}} > \widehat{\beta} > \beta^\star$ by parts 1 and 2 of Lemma A.1. On the other hand, if $\widehat{M}_n(\widehat{\beta}) - \text{rad}(\widehat{\beta}, n, \delta) > 0$ then $M(\widehat{\beta}) > 0$ under the clean event, so we can conclude that $\beta_{\text{low}} < \widehat{\beta} < \beta^\star$ by parts 1 and 3 of Lemma A.1. Therefore since $M(\beta)$ is a continuous function on the interval $[\beta_{\text{low}}, \beta_{\text{high}}]$, under the clean event, $\beta_{\text{high}} - \beta_{\text{low}} \leq \epsilon$ and $\beta^\star \in [\beta_{\text{low}}, \beta_{\text{high}}]$ once the while loop terminates.

Number of iterations: Our bisection method takes $\log_2(\frac{\beta_{\text{high}} - \beta_{\text{low}}}{\epsilon}) = \log_2(\frac{\sigma^2}{2|\mu_{\text{ref}}|\epsilon})$ iterations to ensure that the final parameter returned is within an additive error of ϵ. Assuming a confidence sequence on $\widehat{M}_n(\beta)$ that is valid for all $n \geq 1$ and any fixed β, our high probability statement requires taking the union bound over each bisection method iteration. □

Lemma B.4. *Under Assumption B.1, if π_{ref} is such that $\pi_{\text{ref}}(\iota|h) \geq \gamma$ for any token $\iota \in \mathcal{Y}$ and token prefix h, then for any fixed β,*

$$|\widehat{M}_n(\beta) - M(\beta)| \leq 1.7\,(\sigma + \beta m \log(1/\gamma)) \sqrt{\frac{0.72 \log(20.8/\delta) + \log\log(2n)}{n}},$$

with probability at least $1 - \delta$ simultaneously for all $n \geq 1$, where m is the maximum token output sequence length. Therefore, it suffices to set

$$\text{rad}(\beta, n, \delta) = 1.7\,(\sigma + \beta m \log(1/\gamma)) \sqrt{\frac{0.72 \log(20.8/\delta) + \log\log(2n)}{n}}$$

in Algorithm 1.

Proof. For an arbitrary policy π with $x_1, \ldots, x_n \sim \mathcal{D}$ and $y_i \sim \pi(\cdot|x_i)$, we have that

$$\left| \frac{1}{n} \sum_{i=1}^{n} r(x_i, y_i) - \mathbb{E}_{y \sim \pi(\cdot|x)}[r(x, y)] \right| \leq 1.7\sigma \sqrt{\frac{0.72 \log(10.4/\delta) + \log\log(2n)}{n}},$$

with probability at least $1 - \delta$, simultaneously for every $n \geq 1$ by Lemma B.2. Turning our attention to the KL term, we have that

$$\begin{aligned}
\mathbb{E}[\text{KL}(\pi(\cdot|x), \pi_{\text{ref}}(\cdot|x))] &= \mathbb{E}_{y \sim \pi(\cdot|x)} \left[\sum_{j=1}^{m} \log \left(\frac{\pi(y_j | y_{1:j-1}, x)}{\pi_{\text{ref}}(y_j | y_{1:j-1}, x)} \right) \right] \\
&= \mathbb{E} \left[\sum_{j=1}^{m} \mathbb{E}_{y_{1:j-1} \sim \pi(\cdot|x)} \left[\sum_{\iota \in \mathcal{Y}} \pi(\iota | y_{1:j-1}, x) \log \left(\frac{\pi(\iota | y_{1:j-1}, x)}{\pi_{\text{ref}}(\iota | y_{1:j-1}, x)} \right) \right] \right].
\end{aligned}$$

Let $G_j(x) = \mathbb{E}_{y_{1:j-1} \sim \pi(\cdot|x)} \left[\sum_{\iota \in \mathcal{Y}} \pi(\iota | y_{1:j-1}, x) \log \left(\frac{\pi(\iota | y_{1:j-1}, x)}{\pi_{\text{ref}}(\iota | y_{1:j-1}, x)} \right) \right]$, and observe that $0 \leq G_j(x) \leq \log(1/\gamma)$.

Consider the one-sample estimator $\widehat{G}_j(x) := \sum_{\iota \in \mathcal{Y}} \pi(\iota | y_{1:j-1}, x) \log \left(\frac{\pi(\iota | y_{1:j-1}, x)}{\pi_{\text{ref}}(\iota | y_{1:j-1}, x)} \right)$ and note that it is unbiased and computable given access to π, π_{ref}, x, and $y \sim \pi(\cdot|x)$. We can estimate $\mathbb{E}[\text{KL}(\pi(\cdot|x), \pi_{\text{ref}}(\cdot|x))] = \mathbb{E}[\sum_{j=1}^{m} G_j(x)]$ as $\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \widehat{G}_j(x_i)$. Observing that $\sum_{j=1}^{m} G_j(x)$ is sub-Gaussian with parameter at most $m \log(1/\gamma)$, we can apply Lemma B.2 to get

$$\left| \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \widehat{G}_j(x_i) - \mathbb{E}[\text{KL}(\pi(\cdot|x), \pi_{\text{ref}}(\cdot|x))] \right| \leq 1.7 m \log(1/\gamma) \sqrt{\frac{0.72 \log(10.4/\delta) + \log\log(2n)}{n}},$$

with probability at least $1 - \delta$, simultaneously for all $n \geq 1$. Combining this with our high-probability guarantee for $\mathbb{E}_{y \sim \pi(\cdot|x)}[r(x, y)]$ and taking a union bound gets us the desired result. □

Corollary B.5. *Under Assumption B.1, if* $\beta_{\text{high}} = \frac{\sigma^2}{2|\mu_{\text{ref}}|}$, $\pi_{\text{ref}}(\iota|h) \geq \gamma > 0$ *for any token* $\iota \in \mathcal{Y}$ *and token prefix* h*, and*

$$\text{rad}(\beta, n, \delta) = 1.7\,(\sigma + \beta m \log(1/\gamma))\,\sqrt{(0.72\log(20.8/\delta) + \log\log(2n))/n},$$

then with probability at least $1 - \delta \log_2\left(\frac{\sigma^2}{2|\mu_{\text{ref}}|\epsilon}\right)$*, Algorithm 1 returns a policy* $\pi_{\beta_{\text{low}}}$ *satisfying* $0 \leq \beta^\star - \beta_{\text{low}} \leq \epsilon$ *in* $\log_2\left(\frac{\sigma^2}{2|\mu_{\text{ref}}|\epsilon}\right)$ *bisection iterations.*

Proof. Corollary B.5 follows immediately from Theorem 4.1 and Lemma B.4. □

Assumption B.6 (Conditional sub-Gaussianity)**.** *Rewards satisfy Assumption B.1. Moreover, under the reference policy, rewards are uniformly conditionally sub-Gaussian, i.e., for every prompt* $x \in \mathcal{X}$ *and all* $\lambda \in \mathbb{R}$*,*

$$\log \mathbb{E}_{y\sim\pi_{\text{ref}}(\cdot|x)}\left[\exp\left(\lambda\big(r(x,y) - \mu_{\text{ref}}(x)\big)\right)\right] \leq \frac{\sigma^2\lambda^2}{2},$$

where $\mu_{\text{ref}}(x) := \mathbb{E}_{y\sim\pi_{\text{ref}}(\cdot|x)}[r(x,y)]$.

Corollary B.7. *Under Assumption B.6, if*

$$\text{rad}(\beta, n, \delta) = 1.7\sigma\sqrt{\left(2 + \frac{\sigma^2}{8\beta^2}\right)\frac{0.72\log(10.4/\delta) + \log\log(2n)}{n}},$$

and $\beta_{\text{high}} = \frac{\sigma^2}{2|\mu_{\text{ref}}|}$*, then Algorithm 1 returns a policy* $\pi_{\beta_{\text{low}}}$ *satisfying* $0 \leq \beta^\star - \beta_{\text{low}} \leq \epsilon$ *with probability at least* $1 - \delta\log_2\left(\frac{\sigma^2}{2|\mu_{\text{ref}}|\epsilon}\right)$ *in* $\log_2\left(\frac{\sigma^2}{2|\mu_{\text{ref}}|\epsilon}\right)$ *bisection iterations.*

Proof. We have that

$$\log\frac{\pi_\beta(y|x)}{\pi_{\text{ref}}(y|x)} = \frac{r(x,y)}{\beta} - \log Z_\beta(x).$$

and so for each fixed x,

$$\text{KL}(\pi_\beta(\cdot|x), \pi_{\text{ref}}(\cdot|x)) = \mathbb{E}_{y\sim\pi_\beta(\cdot|x)}\left[\frac{r(x,y)}{\beta} - \log Z_\beta(x)\right],$$

which implies that $M(\beta) = \mathbb{E}[\beta \log Z_\beta(x)]$. By Jensen's inequality and Assumption B.6, we have that

$$0 \leq \beta\log Z_\beta(x) - \mathbb{E}_{y\sim\pi_{\text{ref}}(\cdot|x)}[r(x,y)] \leq \frac{\sigma^2}{2\beta}$$

and $\mathbb{E}_{y\sim\pi_{\text{ref}}(\cdot|x)}[r(x,y)] - \mu_{\text{ref}}$ is σ-sub-Gaussian over $x \sim \mathcal{D}$. Hence by Cauchy–Schwarz and Hoeffding's lemma,

$$\mathbb{E}\exp\left(\lambda\left(\beta\log Z_\beta(x) - M(\beta)\right)\right) \leq \exp\left(\frac{\lambda^2}{2}\left(2\sigma^2 + \frac{\sigma^4}{8\beta^2}\right)\right).$$

Applying Lemma B.2 with sub-Gaussian parameter $\sqrt{2\sigma^2 + \frac{\sigma^4}{8\beta^2}}$ gives the desired confidence sequence. The result now follows from Theorem 4.1. □

B.1 Approximate Equilibria

Theorem B.8. *The following guarantees hold when the monitor and the agent use the tilted policy* $\pi_{\beta_{\text{low}}}$ *returned by Algorithm 1:*

1. *If the monitor is deploying an SPRT between* $\pi_{\beta_{\text{low}}}$ *and* π_{ref}*, the agent can guarantee themselves utility at least* $\beta_{\text{low}}\log(1/\alpha)$ *by playing any policy* π *such that* $\mathbb{E}_{y\sim\pi(\cdot|x)}\left[\log\frac{\pi_{\beta_{\text{low}}}(y|x)}{\pi_{\text{ref}}(y|x)}\right] > 0$ *(including* $\pi = \pi_{\beta_{\text{low}}}$*).*[6]

[6] Recall that $0 \leq \beta^\star - \beta_{\text{low}} \leq \epsilon$ with high probability, by Theorem 4.1.

2. *If the agent is playing policy $\pi_{\beta_{\text{low}}}$ and the monitor is playing a power-one simple-vs-simple SPRT, then their best response is to test $\mathcal{H}_0 : \pi = \pi_{\text{ref}}$ versus $\mathcal{H}_1 : \pi = \pi_{\beta_{\text{low}}}$. Moreover, as $\alpha \downarrow 0$, this is the optimal such test out of all power-one sequential tests for the monitor to play (Wald & Wolfowitz, 1948).*

Proof.

$$\max_{\pi} \mathbb{E}_{y_t \sim \pi(\cdot|x_t)} \left[\sum_{t=1}^{\tau_\alpha} r(x_t, y_t) \right] = \max_{\pi} \mathbb{E}_{y_t \sim \pi(\cdot|x)}[\tau_\alpha] \cdot \mathbb{E}_{y \sim \pi(\cdot|x)}[r(x,y)]$$
$$\geq \max_{\pi} \frac{E_{y \sim \pi(\cdot|x)}[r(x,y)] \log(1/\alpha)}{\mathbb{E}_{y \sim \pi(\cdot|x)} \left[\log \frac{\pi_{\beta_{\text{low}}}(y|x)}{\pi_{\text{ref}}(y|x)} \right]}$$

where the inequality follows from Lemma A.2 and the fact that $\beta_{\text{low}} > 0$ with high probability. We can rewrite

$$\mathbb{E}_{y \sim \pi(\cdot|x)}[\log \frac{\pi_{\beta_{\text{low}}}(y|x)}{\pi_{\text{ref}}(y|x)}] = \frac{1}{\beta_{\text{low}}} \mathbb{E}_{y \sim \pi(\cdot|x)}[r(x,y)] - \frac{1}{\beta_{\text{low}}} M(\beta_{\text{low}}).$$

Since $\beta_{\text{low}} \leq \beta^\star$, we know that $M(\beta_{\text{low}}) > 0$, and so

$$\max_{\pi} \frac{E_{y \sim \pi(\cdot|x)}[r(x,y)] \log(1/\alpha)}{\mathbb{E}_{y \sim \pi(\cdot|x)}[\log \frac{\pi_{\beta_{\text{low}}}(y|x)}{\pi_{\text{ref}}(y|x)}]} \geq \beta_{\text{low}} \log(1/\alpha)$$

□

Observe that even if each agent runs Algorithm 1 separately, if they stop after the same number of iterations they will arrive at the same value for β_{low} with high probability. With that being said, it is still possible to say something about the quality of the equilibrium whenever the monitor and the agent run Algorithm 1 for a different number of iterations.

Theorem B.9. *Let β_A (resp. β_M) be the agent's (resp. monitor's) computation of β_{low} and suppose that $\beta_A \leq \beta_M$, i.e., the clean event holds and the monitor runs Algorithm 1 for at least as long as the agent. Then:*

1. *If the monitor is deploying a SPRT between π_{β_M} and π_{ref}, the agent can guarantee themselves utility at least $(\beta^\star - \epsilon)\log(1/\alpha)$ by playing policy π_{β_A}.*

2. *If the agent is playing policy π_{β_A} and the monitor plays a power-one simple SPRT between $\mathcal{H}_0 : \pi = \pi_{\text{ref}}$ and $\mathcal{H}_1 : \pi = \pi_{\beta_M}$, then as $\alpha \downarrow 0$*

$$\tau^* \leq \mathbb{E}_{y \sim \pi_{\beta_A}}[\tau_\alpha] \leq \frac{\beta_M}{\beta_A} \cdot \tau^*,$$

where τ^* is the best possible expected stopping time.

Proof. Part 1: The key step is to show that $\mathbb{E}_{y \sim \pi_A(\cdot|x)}[\log \frac{\pi_M(y|x)}{\pi_{\text{ref}}(y|x)}] > 0$. To see this, observe that

$$\mathbb{E}_{y \sim \pi_{\beta_A}(\cdot|x)} \left[\log \frac{\pi_{\beta_M}(y|x)}{\pi_{\text{ref}}(y|x)} \right] = \frac{1}{\beta_M} (\mathbb{E}_{y \sim \pi_{\beta_A}(\cdot|x)}[r(x,y)] - M(\beta_M))$$

and

$$\mathbb{E}_{y \sim \pi_{\beta_A}(\cdot|x)}[r(x,y)] = M(\beta_A) + \beta_A \mathbb{E}[\text{KL}(\pi_{\beta_A}(\cdot|x), \pi_{\text{ref}}(\cdot|x))] > M(\beta_A) \geq M(\beta_M)$$

where the last inequality follows from the fact that $\beta_A \leq \beta_M$. Given this, we can bound

$$\mathbb{E}_{y_t \sim \pi_{\beta_A}(\cdot|x_t)} \left[\sum_{t=1}^{\tau_\alpha} r(x_t, y_t) \right] = \mathbb{E}_{y_t \sim \pi_{\beta_A}(\cdot|x)}[\tau_\alpha] \cdot \mathbb{E}_{y \sim \pi_{\beta_A}(\cdot|x)}[r(x,y)]$$
$$\geq \frac{E_{y \sim \pi_{\beta_A}(\cdot|x)}[r(x,y)] \log(1/\alpha)}{\mathbb{E}_{y \sim \pi_{\beta_A}(\cdot|x)} \left[\log \frac{\pi_{\beta_M}(y|x)}{\pi_{\text{ref}}(y|x)} \right]}$$
$$= \frac{E_{y \sim \pi_{\beta_A}(\cdot|x)}[r(x,y)] \log(1/\alpha)}{\frac{1}{\beta_M} \mathbb{E}_{y \sim \pi_{\beta_A}(\cdot|x)}[r(x,y)] - \frac{1}{\beta_M} M(\beta_M)}$$
$$\geq \beta_M \log(1/\alpha) \geq (\beta^\star - \epsilon) \log(1/\alpha)$$

Part 2: Let $\tau_\alpha(\beta_M)$ be the stopping time when the monitor is using a simple versus simple SPRT with alternative hypothesis $\pi = \pi_{\beta_M}$, and $\tau_\alpha(\beta_A)$ be the counterfactual stopping time when using alternative hypothesis $\pi = \pi_{\beta_A}$. Consider the limit where $\alpha \downarrow 0$. We have that

$$\begin{aligned}\mathbb{E}_{y\sim\pi_{\beta_A}(\cdot|x)}\left[\log\frac{\pi_{\beta_M}(\cdot|x)}{\pi_{\text{ref}}(\cdot|x)}\right] &= \frac{1}{\beta_M}\mathbb{E}_{y\sim\pi_{\beta_A}(\cdot|x)}[r(x,y)] - \frac{1}{\beta_M}M(\beta_M)\\ &\geq \frac{1}{\beta_M}\mathbb{E}_{y\sim\pi_{\beta_A}(\cdot|x)}[r(x,y)] - \frac{1}{\beta_M}M(\beta_A)\\ &= \frac{\beta_A}{\beta_M}\left(\frac{1}{\beta_A}\mathbb{E}_{y\sim\pi_{\beta_A}(\cdot|x)}[r(x,y)] - \frac{1}{\beta_A}M(\beta_A)\right)\end{aligned}$$

Therefore,

$$\begin{aligned}\frac{\mathbb{E}_{y\sim\pi_{\beta_A}}[\tau_\alpha]}{\tau^*} &= \frac{\mathbb{E}_{y\sim\pi_{\beta_A}}[\tau_\alpha(\beta_M)]}{\mathbb{E}_{y\sim\pi_{\beta_A}}[\tau_\alpha(\beta_A)]}\\ &= \frac{\mathbb{E}_{y\sim\pi_{\beta_A}(\cdot|x)}\left[\log\frac{\pi_{\beta_A}(y|x)}{\pi_{\text{ref}}(y|x)}\right]}{\mathbb{E}_{y\sim\pi_{\beta_A}(\cdot|x)}\left[\log\frac{\pi_{\beta_M}(y|x)}{\pi_{\text{ref}}(y|x)}\right]} \leq \frac{\beta_M}{\beta_A}.\end{aligned}$$

□

C Appendix for Section 5: Experiments

Policies are fine-tuned with rank 16 LoRA adapters (Hu et al., 2022) using GRPO (Shao et al., 2024) as the RL oracle and Adam (Kingma & Ba, 2014) as the optimizer with learning rate 1e-4, temperature 1.0, top-p 1.0, and a completion cap of 2048 tokens. Each call to the RL oracle consists of 15 optimization steps with 400 rollouts each. All models were trained using the Tinker API (Thinking Machines Lab, 2025).

C.1 Continual Learning

Judge prompts. For each trained policy, we evaluated generated completions using two separate LLM-judge prompts: one for narrative coherence and one for grammatical coherence. The same prompts were used for both `GPT-5-mini` and `GPT-5-nano`.

Narrative coherence system prompt:

> You are an expert literary critic evaluating short stories written by a small language model. Your task is to rate the NARRATIVE COHERENCE of a story on an integer scale from 0 to 10.
>
> Narrative coherence means:
>
> The story has a recognizable beginning, middle, and end. Events follow each other in a sensible causal or temporal order. Characters and settings remain consistent throughout. The story arrives at some kind of resolution or conclusion. Ignore grammar and spelling mistakes – those are evaluated separately. Focus only on the structure and coherence of the narrative.
>
> Respond with ONLY a single integer between 0 and 10, with no other text. 0 = no discernible narrative; 10 = a complete, well-structured short story.

Grammatical coherence system prompt:

> You are an expert linguist evaluating short stories written by a small language model. Your task is to rate the GRAMMATICAL COHERENCE of a story on an integer scale from 0 to 10.
>
> Grammatical coherence means:
>
> Sentences are syntactically well-formed. Subject-verb agreement, tense, pronouns, and articles are used correctly. Punctuation and capitalization roughly follow standard English conventions. Words are spelled correctly and used in

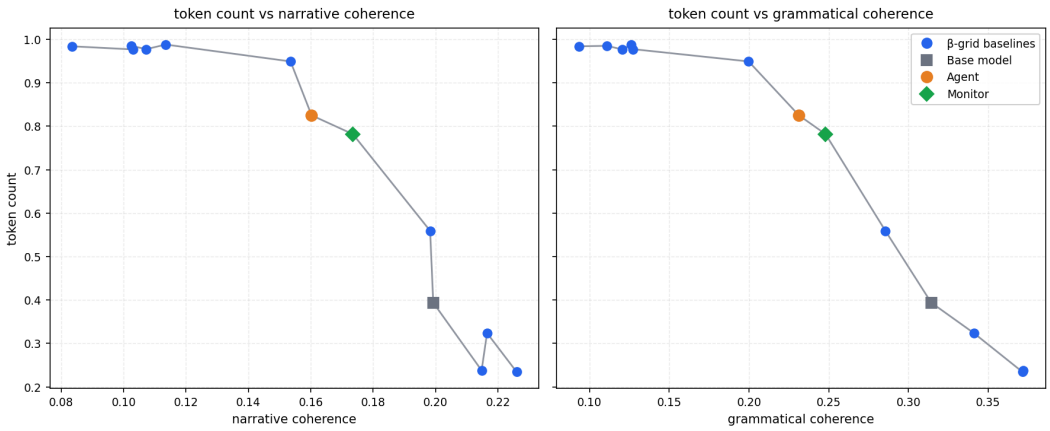


Figure 3: Reward–retention trade-offs for `Llama-3.2-1B` with shift $\rho = 0.01$. Plotting conventions are the same as in Figure 1. In both settings, Algorithm 1 selects policies near the elbow of the empirical trade-off curve traced out by the compute-matched β grid.

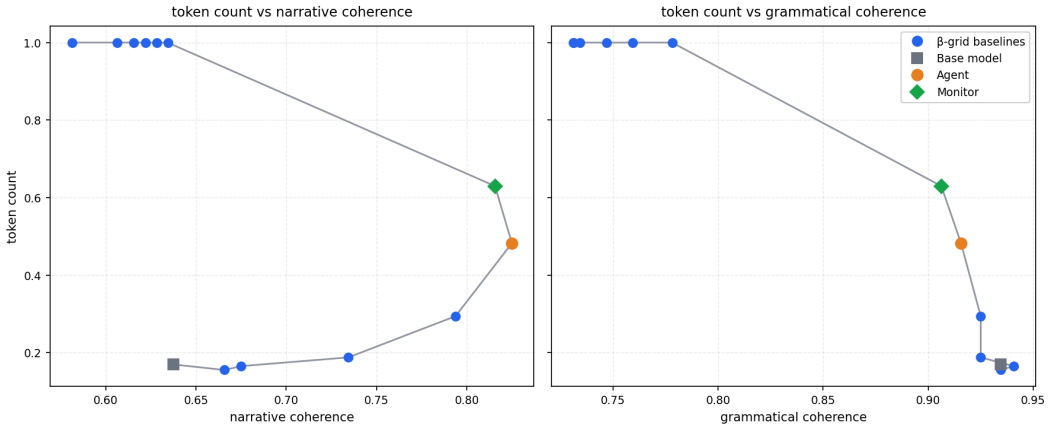


Figure 4: Reward–retention trade-offs for `Qwen3-8B` with shift $\rho = 0.1$ and `GPT-5-nano`-as-a-judge. All other plotting conventions are the same as in Figure 1.

> plausible contexts. Ignore plot quality and narrative structure – those are evaluated separately. Focus only on syntactic and morphological correctness.
>
> Respond with ONLY a single integer between 0 and 10, with no other text. 0 = essentially ungrammatical; 10 = fully grammatical, idiomatic English.

Figure 3 reports the additional `Llama-3.2-1B` experiment with calibration margin $\rho = 0.01$, using `GPT-5-mini` as the judge.

Figures 4–7 repeat the continual-learning experiments using `GPT-5-nano` as the judge. The absolute coherence scores differ slightly from the `GPT-5-mini` scores, but the qualitative pattern is unchanged: Across `Qwen3-8B` and `Llama-3.2-1B`, and across the tested calibration margins, the stochastic-bisection policies fall in the transition region of the Pareto frontier rather than at either extreme of the compute-matched β grid.

Plot details. This subsection contains the β values that are used in all figures, and the order in which they are plotted. For `Qwen3-8B`, all β values are listed in clockwise order. For `Llama-3.2-1B`, all β values are listed from left to right.

`Qwen3-8B` with shift $\rho = 0.1$ details:

- Figure 1, left: 0.00058, $2.693e-05$, 0.000125, 0.00269, 0.0, $5.802e-06$, 0.008545 (monitor), 0.008545 (agent), 0.0125, 0.0580, 0.2693, 1.25, base model.

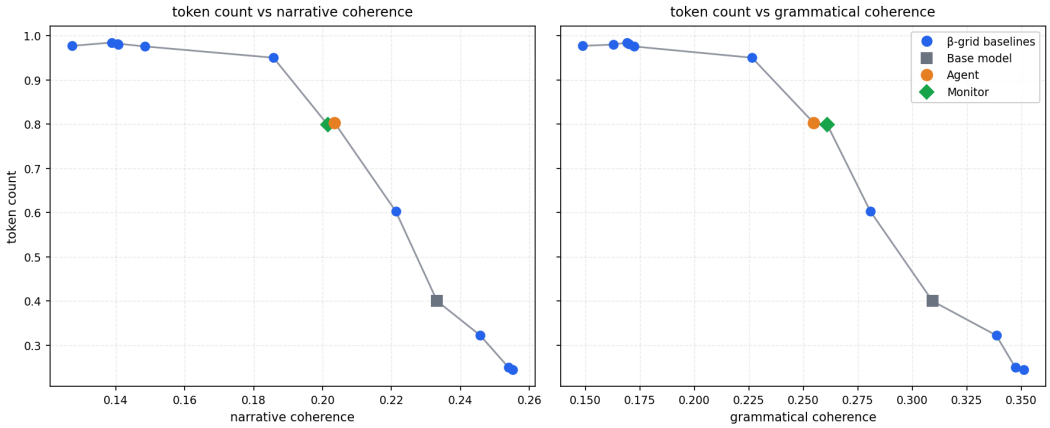


Figure 5: Reward–retention trade-offs for `Llama-3.2-1B` with shift $\rho = 0.01$ and `GPT-5-nano`-as-a-judge.

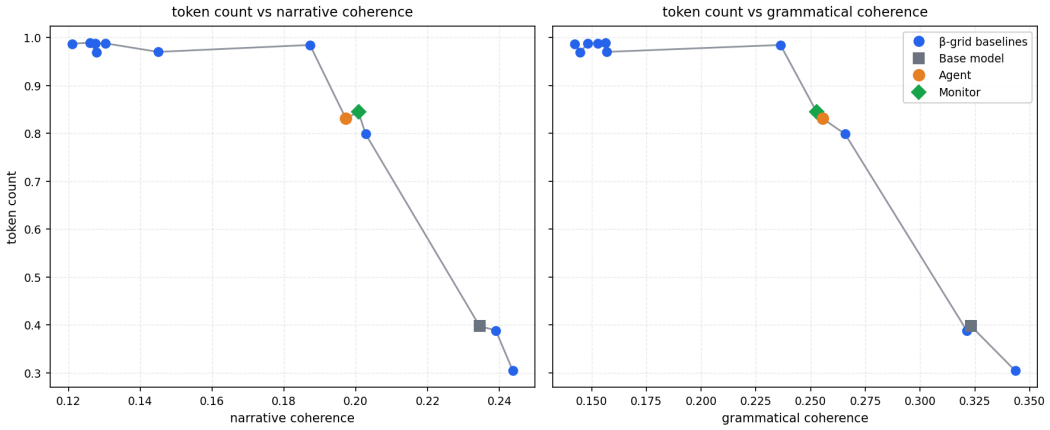


Figure 6: Reward–retention trade-offs for `Llama-3.2-1B` with shift $\rho = 0.1$ and `GPT-5-nano`-as-a-judge.

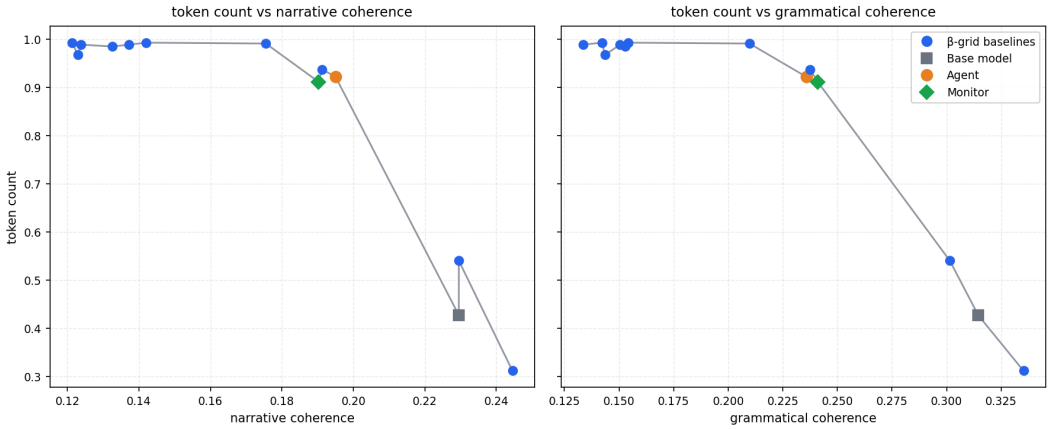


Figure 7: Reward–retention trade-offs for `Llama-3.2-1B` with shift $\rho = 0.2$ and `GPT-5-nano`-as-a-judge.

- Figure 1, right: 0.00058, $5.802e-06$, 0.00269, $2.693e-05$, 0.0, 0.000125, 0.008545 (monitor), 0.008545 (agent), 0.0125, 0.0580, 0.2693, 1.25, base model.
- Figure 4, left: 0.00058, 0.0, $2.693e-05$, 0.00269, $5.802e-06$, 0.000125, 0.008545 (monitor), 0.008545 (agent), 0.0125, 0.0580, 1.25, 0.2693, base model.
- Figure 4, right: 0.0, $2.693e-05$, 0.00058, 0.000125, 0.00269, $5.802e-06$, 0.008545 (monitor), 0.008545 (agent), 0.0125, 0.0580, 1.25, 0.2693, base model.

`Llama-3.2-1B` with shift $\rho = 0.01$ details:

- Figure 3, left: 0.0004, 0.00884, 0.0, 0.0019, $8.8437e-05$, 0.041, 0.093 (agent), 0.093 (monitor), 0.19, base model, 4.10, 0.884, 19.05
- Figure 3, right: 0.0004, 0.0, 0.00884, $8.8437e-05$, 0.0019, 0.041, 0.093 (agent), 0.093 (monitor), 0.19, base model, 0.884, 19.05, 4.10
- Figure 5, left: 0.0004, 0.0019, 0.0, 0.00884, $8.8437e-05$, 0.041, 0.093 (monitor), 0.093 (agent), 0.19, base model, 0.884, 19.05, 4.10
- Figure 5, right: 0.0004, 0.0, 0.0019, 0.00884, $8.8437e-05$, 0.041, 0.093 (agent), 0.093 (monitor), 0.19, base model, 0.884, 19.05, 4.10

`Llama-3.2-1B` with shift $\rho = 0.1$ details:

- Figure 2a, left: 0.0, 0.0042, $4.202e-05$, 0.000195, 0.000905, $9.0536e-06$, 0.0195, 0.08381 (monitor), 0.08381 (agent), base model, 0.0905, 0.4202, 1.95
- Figure 2a, right: $9.0536e-06$, 0.0042, $4.202e-05$, 0.000195, 0.0, 0.000905, 0.0195, 0.08381 (monitor), 0.08381 (agent), 0.0905, base model, 0.4202, 1.95
- Figure 6, left: 0.0042, 0.0, $4.202e{-}05$, 0.000195, $9.0536e{-}06$, 0.000905, 0.0195, 0.08381 (agent), 0.08381 (monitor), 0.0905, base model, 0.4202, 1.95
- Figure 6, right: 0.0042, $4.202e-05$, $9.0536e-06$, 0.000195, 0.0, 0.000905, 0.0195, 0.08381 (monitor), 0.08381 (agent), 0.0905, 0.4202, base model, 1.95

`Llama-3.2-1B` with shift $\rho = 0.2$ details:

- Figure 2b, left: $9.352e-05$, 0.002, 0.000434, $4.34e-06$, 0.0, $2.0148e-05$, 0.00935, 0.0434, 0.0603 (monitor), 0.0603 (agent), base model, 0.201, 0.935
- Figure 2b, right: $4.34e-06$, 0.002, $9.352e-05$, 0.0, 0.000434, $2.0148e-05$, 0.00935, 0.0434, 0.0603 (agent), 0.0603 (monitor), 0.201, base model, 0.935
- Figure 7, left: $9.352e-05$, 0.002, $4.34e-06$, 0.000434, 0.0, $2.0148e-05$, 0.00935, 0.0603 (monitor), 0.0434, 0.0603 (agent), base model, 0.201, 0.935
- Figure 7, right: $4.34e-06$, $9.352e-05$, 0.002, 0.0, 0.000434, $2.0148e-05$, 0.00935, 0.0603 (agent), 0.0434, 0.0603 (monitor), 0.201, base model, 0.935

C.2 Model Auditing

As a baseline, we fine-tune surrogate policies on an exponential β-grid (the same one as in Appendix C.1) and, at each timestep, compute per-arm log-likelihood ratios against the reference on the same observed pairs. The mixture likelihood-ratio test aggregates evidence through a uniform mixture over grid points, stopping when the mixture statistic $\Lambda_t = \sum_k w_k \exp(L_{k,t})$ exceeds $1/\alpha$, where $L_{k,t}$ is the cumulative log-LR for arm k and $w_k = 1/K$. Under strategic sampling we compare the stopping times of both composites to those of the monitor and the bisection-optimal agent; under honest sampling we report the fraction of trials on which each composite crosses its threshold.