ROBUST ONLINE LEARNING AGAINST BIASED LLM JUDGE WITH BUDGETED CALIBRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Online learning is emerging as a powerful paradigm for adapting LLM systems, often relying on LLM-as-a-judge for feedback. However, such feedback can be systematically biased and may not faithfully reflect true task quality. Clean feedback can correct this mismatch, but is too costly to obtain at every round. This raises a fundamental challenge: how can an online learner leverage abundant but biased feedback with only a limited budget of clean feedback? We formulate this setting as a contextual bandit problem with an always-observed (biased) proxy reward and a limited budget for selectively acquiring clean feedback. We propose Paired-Audit Calibrated Exploration (PACE), which leverages paired proxy-clean observations to estimate true task quality and adaptively allocates the audit budget based on learned proxy-clean disagreement. We evaluate PACE on LLM routing, prompt selection, retrieval configuration selection, and synthetic benchmarks. Across our evaluations, PACE is most effective when proxy-clean disagreement is decision-relevant; on the four primary routing conditions, it reduces mean cumulative clean regret over proxy-only LinUCB by 36.5–58.8%. For theoretical characterization, we study PACE-U, a uniform-auditing variant that isolates paired proxy-clean calibration under linear realizability. PACE-U achieves $\widetilde{O}(dT/\sqrt{B})$ clean regret, while any learner restricted to at most B clean-feedback queries incurs $\Omega(T/\sqrt{B})$ regret on a two-dimensional instance, even with adaptive auditing. Thus, the dependence on the horizon T and clean-feedback budget B is minimax rate-optimal for fixed dimension.

1 INTRODUCTION

Modern LLM systems increasingly involve adaptive decisions such as LLM routing, prompt optimization, retrieval configuration selection, and tool-use policies (Chen et al., 2024; Ong et al., 2025; Yang et al., 2024). Since the optimal choice varies across inputs and evolving environments, system operators face uncertainty when selecting configurations a priori. This naturally motivates an online decision-making formulation, in which the system sequentially selects candidate actions and updates its policy based on observed feedback (Nguyen et al., 2024; Sridhar et al., 2026; Tang et al., 2025). However, effective online adaptation typically requires extensive feedback, which is costly to obtain in practice. Under this resource constraint, LLM-as-a-Judge has become an attractive source of scalable feedback (Zheng et al., 2023).

Serving as an inexpensive evaluation source, an LLM judge can assign a score to each online decision at low cost, functioning as a scalable proxy reward. However, LLM judges can exhibit systematic biases, such as verbosity and position preferences, that do not always align with the true task objective (Zheng et al., 2023; Shi et al., 2025). When used directly for optimization, this misalignment can be amplified by the learner, which may achieve high proxy rewards without improving, or even while degrading, true task performance (Krakovna et al., 2020; Skalse et al., 2022).

A natural remedy is to obtain reliable task-relevant feedback, such as human annotations or high-fidelity evaluation, which we refer to as the *clean* reward. Such feedback is costly and can typically be acquired only for a limited number of rounds (Tucker et al., 2023). This creates two coupled challenges: proxy bias may depend on both context and action (Gaucher et al., 2022), while the limited clean-feedback budget must be allocated selectively. The learner must therefore decide *where to audit* and *how to combine* abundant proxy feedback with scarce clean observations.

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 propose **Paired-Audit Calibrated Exploration (PACE)**, a contextual bandit framework for this setting. PACE uses paired proxy-clean observations to estimate proxy discrepancy, directs more audits toward contexts where disagreement is likely, and maintains both a direct clean-reward estimate and a bias-corrected proxy estimate. Actions are chosen using the tighter of the two valid optimistic bounds, allowing PACE to exploit abundant proxy observations without relying on them unconditionally.

Our contributions are:

- **Problem formulation.** We formulate online learning with an always-observed biased proxy and a hard budget of selectively acquired clean feedback.
- **Algorithm.** We propose PACE, which combines paired proxy-clean calibration with context-dependent audit allocation.
- **Experiments.** We evaluate PACE on LLM routing, prompt selection, retrieval configuration selection, and controlled synthetic environments. On the main routing tasks, PACE substantially reduces clean regret relative to proxy-only and uniform-audit baselines, while controlled experiments show that selective auditing is most beneficial when proxy-clean disagreement has exploitable contextual structure.
- **Theory.** We characterize the statistical difficulty of the proposed proxy-clean feedback model. We show that the uniform-auditing PACE-U variant achieves $\widetilde{O}(dT/\sqrt{B})$ clean regret, while any learner with at most B clean-feedback queries—including learners with adaptive audit allocation—incurs $\Omega(T/\sqrt{B})$ regret in the worst case. We further identify a robustness–adaptation limitation of uniform auditing, motivating context-dependent audit allocation in the full method.

2 Related Works

Online Learning in LLM Systems. Online learning has been used to adapt LLM systems across model routing, prompt optimization, and retrieval configuration. Routing methods such as MetaLLM and FrugalGPT dynamically select models based on input characteristics and observed performance (Nguyen et al., 2024; Chen et al., 2024), while prompt optimization (Yang et al., 2024; Pryzant et al., 2023) and adaptive retrieval (Tang et al., 2025; Jeong et al., 2024) similarly optimize configurable decisions from task feedback. PACE shares this online decision-making perspective, but addresses a different challenge: the feedback used for learning may itself be systematically biased.

Bias in LLM-as-a-Judge. LLM-as-a-Judge provides scalable evaluation but can exhibit systematic preferences such as position, verbosity, and self-enhancement biases (Zheng et al., 2023; Wang et al., 2024a). More recently, Yang et al. (2026) show that contextual bandits can actively exploit such evaluator biases through semantics-preserving style manipulations. Rather than improving the standalone accuracy or fairness of an LLM judge, we treat the judge as an inexpensive but potentially biased reward source and study how an online learner can recover the task-relevant objective with limited reliable feedback.

Online Learning with Biased and Budgeted Feedback. Our setting is closely related to contextual bandits with surrogate or multi-fidelity feedback. CABS combines true and surrogate rewards through coupled reward modeling (Sridhar et al., 2026), while multi-fidelity methods allocate evaluations across feedback sources with different costs and accuracies (Kandasamy et al., 2016). TACC further uses learned mismatch certificates to decide when to escalate to higher-fidelity feedback (Lu et al., 2026). Unlike these approaches, PACE directly learns the proxy–reference discrepancy from paired observations and uses it both to calibrate the proxy signal and to guide subsequent audits.

A complementary line of work studies limited access to reliable feedback. Confidence-Budget Matching and costly-observation bandits allocate reward queries under explicit feedback budgets (Efroni et al., 2021; Tucker et al., 2023), while recent work considers LLM judges with limited human feedback (Saha et al., 2026; Ao et al., 2026). Robust bandit methods also study biased or corrupted observations (Gaucher et al., 2022; He et al., 2022; Rangi et al., 2022), and Ghasemi & Crowley (2026) use sparse ground-truth audits to learn context-dependent evaluator trust. PACE differs in its feedback structure: the proxy reward is observed after every selected action, whereas the clean reward is selectively acquired and arrives paired with that proxy observation. Each audit therefore supports both clean-reward estimation and proxy-bias calibration, while the learned disagreement risk determines where the remaining audit budget is spent.

3 PROBLEM FORMULATION

We consider an online contextual bandit with an always-observed but potentially biased proxy reward and a limited budget of reliable reference feedback. At each round $t = 1, \dots, T$, the learner observes a context $x_t \in \mathcal{X}$ and a non-empty candidate set $\mathcal{A}(x_t)$, and selects $A_t \in \mathcal{A}(x_t)$.

The selected action is associated with a *clean* (or *reference*) reward R_t^c, reflecting the task-relevant objective, and an inexpensive *proxy* reward R_t^p provided by an LLM judge. Their conditional means are

$$\mu_c^\star(x, a) := \mathbb{E}[R_t^c \mid x_t = x, A_t = a], \qquad \mu_p^\star(x, a) := \mathbb{E}[R_t^p \mid x_t = x, A_t = a], \tag{1}$$

where both lie in $[0, 1]$. We characterize systematic proxy–reference discrepancy by

$$b^\star(x, a) := \mu_p^\star(x, a) - \mu_c^\star(x, a). \tag{2}$$

The proxy reward R_t^p is observed after every selected action, whereas R_t^c is revealed only when the round is audited. Let $M_t \in \{0, 1\}$ denote the audit indicator. If $M_t = 1$, the learner observes the paired feedback (R_t^p, R_t^c); otherwise only R_t^p is observed. Audits satisfy the hard budget $\sum_{t=1}^T M_t \le B$, where $B \le T$.

Let $a_t^\star \in \arg\max_{a \in \mathcal{A}(x_t)} \mu_c^\star(x_t, a)$. We evaluate the learner by cumulative clean pseudo-regret

$$\mathrm{Reg}_c(T) = \sum_{t=1}^{T} \left(\mu_c^\star(x_t, a_t^\star) - \mu_c^\star(x_t, A_t)\right). \tag{3}$$

The objective is to minimize $\mathrm{Reg}_c(T)$ subject to the audit budget B.

4 METHOD

We propose **Paired-Audit Calibrated Exploration (PACE)** for online learning with an always-observed but potentially biased proxy reward and a limited budget of clean reference feedback. PACE uses selectively acquired proxy–reference pairs to calibrate the abundant proxy signal. It combines three components: (i) paired estimation of clean reward and proxy discrepancy, (ii) calibrated optimistic action selection, and (iii) context-dependent allocation of the audit budget. We then introduce PACE-U, a simplified uniform-auditing variant used for theoretical analysis. Full pseudocode and implementation details are provided in Appendix A.

4.1 ESTIMATOR CONSTRUCTION

Let $\phi(x, a) \in \mathbb{R}^d$ denote the context–action feature vector and write $\phi_t = \phi(x_t, A_t)$. PACE maintains three predictors for the clean reward, proxy reward, and their discrepancy. On an audited round, define $D_t = R_t^p - R_t^c$. Because audits are selected nonuniformly, an audited observation from round s receives weight $\nu_s = M_s / \max\{p_s, p_0\}$, where p_s is its audit probability and $p_0 > 0$ is a clipping floor. With ridge parameters $\lambda_c, \lambda_p, \lambda_b > 0$, the sufficient statistics are

$$\begin{aligned}
V_t^c &= \lambda_c I_d + \sum_{s<t} \nu_s \phi_s \phi_s^\top, \quad & h_t^c &= \sum_{s<t} \nu_s \phi_s R_s^c, \\
V_t^p &= \lambda_p I_d + \sum_{s<t} \phi_s \phi_s^\top, & h_t^p &= \sum_{s<t} \phi_s R_s^p, \\
V_t^b &= \lambda_b I_d + \sum_{s<t} \nu_s \phi_s \phi_s^\top, & h_t^b &= \sum_{s<t} \nu_s \phi_s D_s.
\end{aligned} \tag{4}$$

For $j \in \{c, p, b\}$, let $\widehat{\theta}_t^j = (V_t^j)^{-1} h_t^j$ and $u_t^j(a) = \|\phi(x_t, a)\|_{(V_t^j)^{-1}}$. The corresponding predictions are $\widehat{\mu}_t^c(a) = \phi(x_t, a)^\top \widehat{\theta}_t^c$, $\widehat{\mu}_t^p(a) = \phi(x_t, a)^\top \widehat{\theta}_t^p$, $\widehat{b}_t(a) = \phi(x_t, a)^\top \widehat{\theta}_t^b$. Thus, the proxy model uses every observed proxy reward, whereas the clean and discrepancy models are updated only from audited pairs. The discrepancy model is treated as a working model and need not be exactly realizable.

4.2 Calibrated Action Selection and Residual Correction

PACE forms two optimistic estimates of the same clean-reward objective:

$$\begin{aligned} U_t^c(a) &= \widehat{\mu}_t^c(a) + \alpha_c u_t^c(a), \\ U_t^{p\to c}(a) &= \widehat{\mu}_t^p(a) - \widehat{b}_t(a) + \alpha_p u_t^p(a) + \alpha_b u_t^b(a) + \epsilon_{J,t}(a). \end{aligned} \tag{5}$$

The first relies directly on audited clean rewards, whereas the second calibrates the much more abundant proxy observations using the estimated proxy–reference discrepancy. PACE selects

$$A_t \in \arg\max_{a\in\mathcal{A}(x_t)} \min\left\{U_t^c(a),\, U_t^{p\to c}(a)\right\}. \tag{6}$$

When both indices provide valid upper bounds on the clean mean, their pointwise minimum remains optimistic: the proxy-calibrated estimate can be used when calibration is accurate, while the direct clean estimate provides a safeguard when it is not.

To account for discrepancy-model misspecification, PACE retains the residual allowance $\epsilon_{J,t}(a) \geq 0$ in equation 5. For each audited round s, before incorporating the new paired observation, it records

$$e_{J,s} = \left[\left|D_s - \widehat{b}_s^{\mathrm{pre}}(A_s)\right| - \alpha_b u_s^b(A_s)\right]_+. \tag{7}$$

For a candidate (x_t, a), let $\mathcal{I}_t(a)$ contain previous audited rounds with similar context–arm features. PACE uses

$$\widehat{\epsilon}_{J,t}(a) = \mathrm{Quantile}_{1-\delta}\left(\{e_{J,s} : s \in \mathcal{I}_t(a)\}\right), \tag{8}$$

falling back to the corresponding global residual quantile when too few similar audited samples are available. This is an empirical correction for residual proxy–reference mismatch; its confidence interpretation additionally requires the error-coverage conditions discussed in Appendix A.5.

4.3 Context-Dependent Audit Selection

PACE spends the clean-feedback budget preferentially on contexts in which the proxy is likely to disagree with the clean objective. For an audited round, define $Y_t^{\mathrm{risk}} = \mathbf{1}\{|R_t^c - R_t^p| \geq \tau_{\mathrm{risk}}\}$. A lightweight predictor receives only pre-execution context features $z_t = z(x_t)$ and outputs $r_t = g_{\psi_t}(z_t) \in [0,1]$. It is trained on audited pairs using clipped inverse-propensity weighting; the complete training objective is given in Appendix A.3. Importantly, neither the current output nor either current reward is available when computing r_t.

Let $B_t = B - \sum_{s<t} M_s$, $q_t = \frac{B_t}{T-t+1}$, where B_t is the remaining audit budget and q_t is its target average consumption rate. Given the collection $\mathcal{R}_t$ of previous risk scores, PACE chooses $c_t \geq 0$ such that

$$\frac{1}{|\mathcal{R}_t|}\sum_{r\in\mathcal{R}_t} \min\{1, c_t r\} \approx q_t. \tag{9}$$

The audit probability is then

$$p_t = \begin{cases} 0, & B_t = 0, \\ q_t, & B_t > 0 \text{ during warm-up}, \\ \min\{1, c_t r_t\}, & \text{otherwise}. \end{cases} \tag{10}$$

Hence, larger predicted disagreement leads to a larger audit probability while approximately matching the remaining budget. The probability p_t is fixed before observing the current output or rewards. After executing A_t, PACE draws $M_t \sim \mathrm{Bernoulli}(p_t)$; an audited pair is used to update the clean, discrepancy, and risk models.

4.4 PACE-U: A Theoretically Tractable Calibration Variant

PACE-U isolates the statistical role of paired calibration under a controlled audit schedule. It retains the calibrated action-selection rule of PACE, while using uniform auditing and exact linear realizability to enable a clean finite-sample analysis. First, it assumes exact linear realizability under the shared feature map:

$$\mu_c^\star(x,a) = \phi(x,a)^\top \theta_c^\star, \qquad \mu_p^\star(x,a) = \phi(x,a)^\top \theta_p^\star,$$

Table 1: Routing performance at $T = 1500$ and audit fraction $b = 0.2$. Entries are cumulative clean regret $\widehat{\mathrm{Reg}}_c(T)$, reported as mean $\pm$ sample standard deviation over 20 paired replay seeds. Lower is better. Column subheadings identify the proxy judge.

Method	IFBench		LiveBench	
Judge	Gemma	GPT-4o	Gemma	GPT-4o
No clean-feedback access ($B = 0$)				
Proxy UCB1	438.00 ± 12.58	414.70 ± 13.78	321.38 ± 9.41	321.67 ± 10.89
Proxy LinUCB	437.45 ± 4.16	428.95 ± 6.15	289.68 ± 8.19	270.77 ± 7.19
MetaLLM-online ($\alpha = 0$)	613.10 ± 48.48	366.45 ± 266.01	300.72 ± 4.48	344.53 ± 3.99
MetaLLM-online ($\alpha = 1$)	613.10 ± 48.48	392.40 ± 264.88	300.57 ± 4.61	342.35 ± 9.05
Budget-matched baselines (audit cap $B = 300$)				
Proxy LinUCB + R^c repl.	375.45 ± 11.15	365.80 ± 8.31	252.70 ± 11.14	246.83 ± 12.93
Clean LinUCB (uniform)	222.50 ± 40.29	222.50 ± 40.29	183.19 ± 24.46	183.19 ± 24.46
Proposed methods (audit cap $B = 300$)				
PACE-U	197.20 ± 60.97	229.90 ± 46.48	154.98 ± 20.51	181.78 ± 21.17
PACE	180.40 ± 54.46	178.20 ± 60.57	154.81 ± 24.76	171.82 ± 30.94

Gemma denotes Gemma-3-27B-IT. All methods have access to shared proxy-only warm-up data; no offline clean labels are provided. MetaLLM denotes the online adaptation.

so that $b^\star(x, a) = \phi(x, a)^\top \theta_b^\star$ with $\theta_b^\star = \theta_p^\star - \theta_c^\star$. PACE-U therefore sets $\epsilon_{J,t}(a) = 0$.

Second, it replaces context-dependent auditing with uniform auditing. Writing $N_{t-1} = \sum_{s<t} M_s$ and $B_t = B - N_{t-1}$,

$$p_t^U = \frac{B}{T}\mathbf{1}\{B_t > 0\}, \qquad M_t \sim \mathrm{Bernoulli}(p_t^U). \tag{11}$$

The cap guarantees $\sum_{t=1}^T M_t \le B$ almost surely. PACE-U otherwise retains the estimator construction in equation 4 and the action rule in equation 5–equation 6. In the theoretical variant, audited observations use uniform weights, ridge regularization is set to one, and the confidence coefficients are specified in Appendix A.7. These assumptions yield the clean-regret analysis in Section 6.

5 Experiments

We evaluate PACE on LLM routing, controlled synthetic environments, prompt selection, and retrieval configuration selection. Our experiments ask: (i) whether budgeted clean feedback improves over proxy-only learning, (ii) whether selective auditing improves over uniform allocation, and (iii) how these gains depend on the audit budget and proxy–reference disagreement.

5.1 Experimental Setup

Tasks and benchmarks. For LLM routing, we evaluate all 300 IFBench questions (Pyatkin et al., 2025) and a 500-question subset of the LiveBench 2024-11-25 release (White et al., 2025). Additional routing results on 750 Arena-Hard v2 questions (Li et al., 2024) and 1,400 MMLU-Pro questions (Wang et al., 2024b) are reported in Appendix B.3. We further evaluate prompt selection on BBH (Suzgun et al., 2023) and retrieval configuration selection on HotpotQA (Yang et al., 2018).

Proxy and clean feedback. Gemma-3-27B-IT and GPT-4o serve as proxy judges under a frozen style-preference prompt, inducing a preference mismatch. Benchmark-specific official evaluation scores serve as clean rewards. Candidate outputs and reward matrices are cached before online replay. At each round, only the selected action's proxy reward is observed; its clean reward is revealed only if the round is audited. Unselected-action rewards are used only for evaluation.

Baselines. We compare PACE against proxy-only bandit methods, including LinUCB, UCB1, Thompson sampling, ϵ-greedy, Random, and an online adaptation of MetaLLM (Nguyen et al., 2024). Budget-matched baselines include uniformly audited Clean LinUCB, Proxy LinUCB + R^c replacement, and PACE-U. For BBH and HotpotQA, we additionally include proxy-trained adaptations of ProTeGi (Pryzant et al., 2023) and MBA-RAG (Tang et al., 2025), respectively. Complete baseline definitions and implementation details are provided in Appendix B.2.

Evaluation. Our metric is empirical cumulative clean regret against the cached per-question oracle:

$$\widehat{\mathrm{Reg}}_c(T) = \sum_{t=1}^{T} \left[\max_{a \in \mathcal{A}(x_t)} R_t^c(a) - R_t^c(A_t) \right]. \tag{12}$$

Unlike the conditional-mean regret in Section 3, this metric uses cached realized rewards. We report the mean and sample standard deviation over 20 matched replay seeds; full replay and evaluation details are given in Appendix B.2.

5.2 Main Results on LLM Routing

Budgeted clean feedback substantially improves routing. Table 1 reports routing performance at $T = 1500$ with an audit budget of $B = 300$. Across IFBench and LiveBench under both Gemma and GPT-4o proxies, PACE substantially reduces cumulative clean regret relative to proxy-only LinUCB. The reductions are 58.8% and 58.5% on IFBench with Gemma and GPT-4o, respectively, and 46.6% and 36.5% on LiveBench. Thus, access to a limited stream of task-relevant reference feedback can substantially mitigate the error induced by optimizing proxy scores alone.

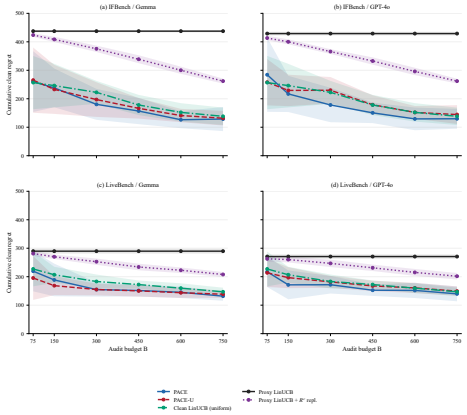

Figure 1: Routing performance versus the audit cap at $T = 1500$. Smaller cumulative clean regret is better. Curves show means over 20 matched replay seeds. Clean LinUCB (uniform) provides a budget-matched clean-feedback comparator.

PACE-U also improves over uniform Clean LinUCB in three of the four primary routing conditions, indicating that the proxy judge can provide useful information when calibrated with limited clean feedback. PACE further matches or improves upon PACE-U across all four conditions, with the largest gains on IFBench. This suggests two complementary benefits: paired proxy–clean calibration can outperform learning from the same clean-feedback budget alone, while context-dependent auditing provides additional gains by allocating that budget more effectively. We investigate the latter mechanism directly in Section 5.3.

Effect of the audit budget. Figure 1 shows that PACE generally benefits from additional clean feedback, with its advantage becoming clearer once enough audits are available to learn where proxy–reference disagreement is consequential; at very small budgets, the audit policy itself receives limited supervision. Additional budget sweeps for BBH and HotpotQA are reported in Appendix B.3.

5.3 When Does Selective Auditing Help?

To isolate when selective auditing is beneficial, we construct controlled four-arm contextual-bandit environments with $T = 15000$, audit budget $B = 3000$, and cold starts. We vary the structure of the proxy distortion to create risk-aligned bias, different levels of clean-ranking mismatch, and a benign rank-preserving proxy. All methods share the same generated environment and context order, and results are averaged over 20 seeds. Complete data-generation details are provided in Appendix B.4.

Selective auditing helps when disagreement is decision-relevant. In Figure 2(a), proxy errors are concentrated on contexts where they reverse the clean arm ranking. Both PACE and PACE-U achieve lower regret than proxy-only learning and uniform Clean LinUCB in this controlled setting.

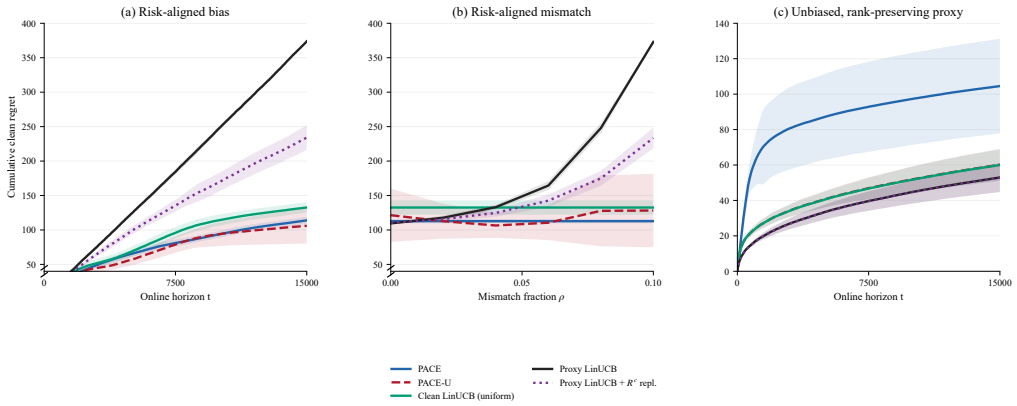

Figure 2: Synthetic mechanisms with cold starts and 20 seeds. (a) Risk-aligned heterogeneous bias. (b) Terminal regret as clean-ranking mismatch increases. (c) A conditionally unbiased, rank-preserving proxy. Lines show means and shaded regions one sample standard deviation.

Its benefit grows with harmful ranking mismatch. As the fraction of ranking-reversing contexts increases, Figure 2(b) shows that proxy-based learners deteriorate, while PACE remains comparatively stable. When mismatch is rare, proxy-only learning remains competitive; as mismatch becomes more frequent, selective auditing becomes increasingly valuable.

Selective calibration is unnecessary for a reliable proxy. In the benign control of Figure 2(c), the proxy is conditionally unbiased and preserves the clean arm ordering. Proxy-only learning can therefore exploit the proxy directly and exhibits sublinear regret, whereas PACE incurs an additional cost from maintaining robustness to the possibility of proxy misspecification. This behavior is consistent with the robustness–adaptation trade-off established for uniform auditing in Theorem 2: without prior knowledge that the proxy is reliable, a uniformly auditing learner that maintains worst-case robustness cannot in general match a proxy-only learner on every benign instance. Thus, PACE is most useful when proxy–reference disagreement is both predictable and consequential for action selection, while some additional adaptation cost is unavoidable when the proxy is already reliable.

5.4 Generalization Across Action Spaces

We further evaluate PACE beyond model routing on prompt selection using BBH (Suzgun et al., 2023) and retrieval configuration selection using HotpotQA (Yang et al., 2018). The BBH setting selects among six fixed reasoning prompts, while the HotpotQA setting selects among four retrieval configurations. Both use $T = 1500$ rounds with an audit cap of $B = 300$; complete task construction, task-specific baselines, and results are reported in Appendix B.3.

As shown in Table 2, PACE improves over uniform Clean LinUCB in all four conditions and over contextual Proxy LinUCB in three of four conditions, although simple proxy-only strategies achieve lower regret in several settings. This occurs primarily when the proxy largely preserves the clean arm ordering and a strong action performs well across many contexts, so that a simple bandit can quickly identify and exploit that action without estimating context-dependent corrections or an audit policy. PACE instead trades some efficiency in these benign settings for robustness to structured proxy–reference mismatch, consistent with the operating regimes identified in Section 5.3.

6 Fundamental Limits and Theoretical Analysis

We now characterize the statistical limits of learning with always-observed proxy feedback and a budget of clean audits. Our analysis separates two questions. First, we ask whether paired proxy–reference calibration is sufficient to achieve the optimal worst-case dependence on the horizon T and

Table 2: Prompt and retrieval configuration selection performance at $T = 1500$ and audit fraction $b = 0.2$. Entries are cumulative clean regret $\widehat{\mathrm{Reg}}_c(T)$, reported as mean $\pm$ sample standard deviation over 20 paired replay seeds. Lower is better. Column subheadings identify the proxy judge.

Method	BBH		HotpotQA	
Judge	Gemma	GPT-4o	Gemma	GPT-4o
No clean-feedback access ($B = 0$)				
Proxy UCB1	273.05 ± 11.44	185.85 ± 11.66	189.40 ± 7.09	169.12 ± 7.10
Proxy LinUCB	245.65 ± 9.00	145.05 ± 9.03	231.26 ± 7.47	213.21 ± 6.70
ProTeGi UCB-E selector	225.95 ± 13.65	137.40 ± 9.00	-	-
MBA-RAG encoder	-	-	183.68 ± 9.46	174.15 ± 10.30
Budget-matched baselines (audit cap $B = 300$)				
Proxy LinUCB + R^c repl.	194.80 ± 9.66	135.25 ± 8.67	225.93 ± 10.38	218.99 ± 9.85
Clean LinUCB (uniform)	156.80 ± 18.35	156.80 ± 18.35	198.60 ± 21.77	198.60 ± 21.77
Proposed methods (audit cap $B = 300$)				
PACE-U	152.40 ± 23.04	153.40 ± 23.65	191.94 ± 29.45	195.37 ± 33.93
PACE	142.10 ± 17.34	147.95 ± 30.84	172.87 ± 33.60	166.04 ± 27.64

Gemma denotes Gemma-3-27B-IT. All methods have access to shared proxy-only warm-up data; no offline clean labels are provided. Both ProTeGi UCB-E selector and MBA-RAG encoder are online adaptations of the corresponding original methods.

audit budget B. For this purpose, we study PACE-U, an analytically transparent variant that uses uniform auditing while retaining the paired calibration and calibrated action-selection mechanisms of PACE. Second, we establish lower bounds for the underlying feedback model, including learners with arbitrary adaptive auditing strategies. Together, these results characterize the minimax difficulty of the problem, while the context-dependent audit policy of PACE is designed to exploit additional instance structure beyond this worst-case characterization.

Theorem 1 (Clean-Regret Guarantee for PACE-U)**.** *Let $T \geq 1$, $1 \leq B \leq T$, and $\delta \in (0, 1)$. Under the linear conditional-mean model and σ_j-sub-Gaussian noise ($j \in \{c, p, b\}$), PACE-U with confidence radii $\alpha_j = \mathcal{O}(S_j + \sigma_j \sqrt{d \log(1 + T/d) + \log(1/\delta)})$ strictly satisfies $\sum_{t=1}^{T} M_t \leq B$ almost surely. Furthermore, with probability at least $1 - \delta$, the cumulative clean regret satisfies*

$$\mathrm{Reg}_c(T) \leq 2 \sum_{t=1}^{T} \min\left\{ \alpha_c u_t^c(A_t),\ \alpha_p u_t^p(A_t) + \alpha_b u_t^b(A_t) \right\}, \tag{13}$$

where $u_t^j(a) = \|\phi(x_t, a)\|_{(V_t^j)^{-1}}$. Consequently,

$$\mathrm{Reg}_c(T) = \widetilde{\mathcal{O}}\left(\min\left\{ T,\ \frac{dT}{\sqrt{B}} \right\} \right). \tag{14}$$

Theorem 1 establishes that the paired-feedback structure can be exploited without sacrificing worst-case robustness. Even under the simple uniform audit schedule of PACE-U, calibrated proxy feedback attains the optimal dependence on the clean-feedback budget, while the minimum-UCB construction automatically falls back to the direct clean estimate when proxy calibration is uninformative. The cumulative width bound (equation 13) takes the pointwise minimum over two valid upper confidence bounds on the clean conditional mean $\mu_c^\star(x_t, a)$: the direct estimate $U_t^c(a)$ from audited rewards, and the calibrated estimate $U_t^{p \to c}(a)$ coupling all T proxy signals with audited bias feedback. When the judge bias is well calibrated, the abundant proxy signal accelerates learning; conversely, when the bias is poorly estimated, the direct clean bound prevents unwarranted optimism. The budget B forms the primary statistical bottleneck in equation 14: uniform auditing at rate B/T yields an effective sample size of $\mathcal{O}(B)$ for clean and bias estimation, producing the leading $\widetilde{\mathcal{O}}(dT/\sqrt{B})$ rate. When $B = T$, this recovers the classical $\widetilde{\mathcal{O}}(d\sqrt{T})$ rate of linear contextual bandits (Abbasi-Yadkori et al., 2011; Chu et al., 2011; Lattimore & Szepesvári, 2020); when $B \ll T$, the scaling reflects the price of clean label scarcity (proof in Appendix C.1).

A natural question is whether the $\mathcal{O}(T/\sqrt{B})$ scaling is an inescapable barrier. While structured assumptions (e.g., a known bias direction; Appendix C.4) permit adapting to benign instances, we show that for general linear bias this rate is minimax-optimal, and establish an intrinsic robustness-adaptation trade-off under uniform auditing.

Theorem 2 (Fundamental Limits of Budgeted Auditing). *Fix $T \geq 1$ and $1 \leq B \leq T$. Consider normalized instances with $[0,1]$-bounded rewards and unit-norm parameters in dimension $d \geq 2$:*

(i) ***(Minimax Lower Bound).*** *For any online learner constrained to $\sum_{t=1}^{T} M_t \leq B$ almost surely (even with adaptive, history-dependent auditing), there exists an instance with $d = 2$ and two candidate actions such that*

$$\mathbb{E}[\mathrm{Reg}_c(T)] \geq \frac{T}{64\sqrt{B+1}}. \tag{15}$$

(ii) ***(Robustness–Adaptation Trade-off).*** *Let the learner execute independent Bernoulli audits at rate B/T until the budget B is exhausted. If the learner attains a worst-case regret guarantee $\sup_\nu \mathbb{E}_\nu[\mathrm{Reg}_c(T)] \leq G$ with $0 < G \leq T/16$ across all realizable instances, then its regret on the zero-bias subclass $\mathcal{M}_0 = \{\nu : \theta_b^\star = 0\}$ must satisfy*

$$\sup_{\nu \in \mathcal{M}_0} \mathbb{E}_\nu[\mathrm{Reg}_c(T)] \geq \frac{T^2}{128BG}. \tag{16}$$

Theorem 2 (i) proves that proxy evaluations alone cannot identify the clean-optimal policy without prior structural assumptions. Thus, distinguishing optimal actions fundamentally requires reference feedback, establishing that the $\widetilde{\mathcal{O}}(T/\sqrt{B})$ dependence is minimax rate-optimal for fixed dimension (Appendix C.2).

Beyond worst-case optimality, Part (ii) reveals that under uniform auditing, robustness against arbitrary bias precludes adaptation to benign environments: guaranteeing $G = \widetilde{\mathcal{O}}(T/\sqrt{B})$ globally forces $\widetilde{\Omega}(T/\sqrt{B})$ regret even when the judge is completely unbiased ($\theta_b^\star = 0$). Intuitively, without prior knowledge of judge accuracy, verifying actions with low proxy scores requires exploratory pulls that receive clean feedback only with probability B/T, causing large clean regret (Appendix C.3).

Comparison with budgeted reward querying. In linear contextual bandits with constrained feedback, Efroni et al. (2021) establish an $\widetilde{\mathcal{O}}(dT/\sqrt{B})$ bound when unqueried rounds yield no observations. While isolated clean queries incur a $dT/\sqrt{B}$ noise penalty, paired observations allow PACE-U to cancel shared feedback noise: when discrepancy noise satisfies $\tau \ll 1$, PACE-U achieves a strictly tighter bound of $\widetilde{\mathcal{O}}(d\sqrt{T} + (\tau d + \sqrt{d})T/\sqrt{B})$ (Appendix D.1).

Comparison with multi-fidelity bandits. In standard multi-fidelity (MF) bandits, each round queries an action and a single fidelity at positive cost (Kandasamy et al., 2016; Wang et al., 2023). Under a fixed budget Λ, paying $c_L > 0$ for every proxy query leaves budget for at most $Q = \lfloor (\Lambda - Tc_L)/(c_H - c_L) \rfloor$ clean queries, whereas uncharged proxy feedback permits $B = \lfloor \Lambda/c_H \rfloor \geq Q$ audits. Consequently, standard MF incurs at least $\Omega(dT/\sqrt{Q})$ regret, strictly dominated by PACE-U since $B \geq Q$. Furthermore, even when $Q/B \to 1$, observing paired signals allows PACE-U to reduce variance via noise cancellation, achieving $\widetilde{\mathcal{O}}(d\sqrt{T})$ when $\tau \ll 1$, which is unattainable via isolated queries (Appendix D.2).

7 Conclusion

We studied contextual bandit learning with an always-observed but potentially biased proxy reward and budgeted clean feedback. PACE combines paired proxy–reference calibration with context-dependent auditing, yielding strong gains on the primary routing benchmarks while controlled and cross-task experiments clarify when selective auditing is beneficial and when simpler proxy-only learning can suffice. For PACE-U, we establish $\widetilde{O}(dT/\sqrt{B})$ clean regret together with a matching $\Omega(T/\sqrt{B})$ dependence on T and B for fixed dimension. These results characterize both the value and the unavoidable adaptation cost of combining abundant proxy feedback with scarce clean evaluation.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing, to assist with retrieval and discovery of related literature, and to draft portions of the appendix.

We manually verified AI-assisted literature retrieval against the original sources and reviewed and revised all AI-assisted writing for technical accuracy and consistency with the experiments and results. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies online learning with potentially biased LLM-based evaluation and is intended to improve robustness to imperfect automated feedback. Our experiments use existing benchmark datasets and model-generated responses, and do not involve human subjects or the collection of private or personally identifiable information. In several experiments, we deliberately induce a controlled proxy–reference mismatch through a fixed style-preference prompt in order to study the effect of systematic evaluator bias. These experiments should not be interpreted as evidence that any particular judge model inherently exhibits the induced preference pattern.

The proposed method uses limited reliable feedback to reduce the influence of biased proxy rewards rather than to exploit such biases. Nevertheless, automated evaluators and learned audit policies may themselves reflect systematic biases present in the underlying models, benchmarks, or feedback sources. In high-stakes applications, clean-feedback definitions and auditing procedures should therefore be validated for the relevant domain and should not replace appropriate human oversight. We report our experimental construction and evaluation protocol transparently to facilitate scrutiny and responsible reuse.

REPRODUCIBILITY STATEMENT

We provide detailed information to support reproducibility of both the empirical and theoretical results. The full PACE and PACE-U procedures, including pseudocode, estimator definitions, audit allocation, and residual calibration, are provided in Appendix A. Appendix B documents the datasets, data partitions, candidate actions, proxy and reference reward construction, online replay protocol, model features, hyperparameters, baseline implementations, random seeds, prompt templates, and additional experimental results. The main experimental setup and evaluation metric are summarized in Section 5.1. For the theoretical results, all assumptions are stated explicitly in Section 6 and Appendix A.7, with complete proofs provided in the corresponding theoretical appendices. These materials, together with the reported experimental configurations and cached-replay protocol, are intended to enable independent reproduction of the results.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24, pp. 2312–2320. Curran Associates, Inc., 2011.

Ruicheng Ao, Hongyu Chen, Siyang Gao, Hanwei Li, and David Simchi-Levi. Best arm identification with llm judges and limited human. *arXiv preprint arXiv:2601.21471*, 2026.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, December 2024. ISSN 2835-8856.

Wang Chi Cheung and Lixing Lyu. Leveraging (Biased) information: Multi-armed bandits with offline data. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8286–8309. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/cheung24a.html.

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.

Yonathan Efroni, Nadav Merlis, Aadirupa Saha, and Shie Mannor. Confidence-budget matching for sequential budgeted learning. In *International Conference on Machine Learning*, pp. 2937–2947. PMLR, 2021.

Solenne Gaucher, Alexandra Carpentier, and Christophe Giraud. The price of unfairness in linear bandits with biased feedback. *Advances in Neural Information Processing Systems*, 35:18363–18376, 2022.

Majid Ghasemi and Mark Crowley. Learning when to trust in contextual social bandits, 2026.

Jiafan He, Dongruo Zhou, Tong Zhang, and Quanquan Gu. Nearly optimal algorithms for linear contextual bandits with adversarial corruptions. *Advances in neural information processing systems*, 35:34614–34625, 2022.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7036–7050, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.389. URL https://aclanthology.org/2024.naacl-long.389/.

Kirthevasan Kandasamy, Gautam Dasarathy, Barnabas Poczos, and Jeff Schneider. The multi-fidelity multi-armed bandit. *Advances in neural information processing systems*, 29, 2016.

Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. Specification gaming: The flip side of AI ingenuity. Google DeepMind Blog, April 2020.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. ISBN 9781108486828. doi: 10.1017/9781108571401.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-Hard and Bench-Builder pipeline. *arXiv preprint arXiv:2406.11939*, 2024. URL https://arxiv.org/abs/2406.11939.

Muyun Lu, Haoyang Hong, Huazheng Wang, and Ying Lin. Beyond static bias: Adaptive multi-fidelity bandits with improving proxies. *arXiv preprint arXiv:2605.08558*, 2026.

Waqar Mirza, Aldo Pacchiano, and Eric Mazumdar. Learning with coupled uncertainty. *Reinforcement Learning Journal*, 7, 2026. URL https://rlj.cs.umass.edu/2026/papers/Paper77.html.

Quang H. Nguyen, Duy C. Hoang, Juliette Decugis, Saurav Manchanda, Nitesh V. Chawla, and Khoa D. Doan. MetaLLM: A high-performant and cost-efficient dynamic framework for wrapping LLMs, 2024.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs with preference data. In *The Thirteenth International Conference on Learning Representations*, 2025.

Riccardo Poiani, Alberto Maria Metelli, and Marcello Restelli. Multi-fidelity best-arm identification. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=OtjQ7NTu3j.

Riccardo Poiani, Rémy Degenne, Emilie Kaufmann, Alberto Maria Metelli, and Marcello Restelli. Optimal multi-fidelity best-arm identification. In *The 38th Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=gKMTM1i8Ew.

Reid Pryzant, Dan Iter, Jerry Li, Yin Lee, Chenguang Zhu, and Michael Zeng. Automatic prompt optimization with "gradient descent" and beam search. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 7957–7968, 2023.

Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hannaneh Hajishirzi. Generalizing verifiable instruction following. *arXiv preprint arXiv:2507.02833*, 2025. URL https://arxiv.org/abs/2507.02833.

Anshuka Rangi, Long Tran-Thanh, Haifeng Xu, and Massimo Franceschetti. Saving stochastic bandits from poisoning attacks via limited data verification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8054–8061, 2022. doi: 10.1609/aaai.v36i7.20777.

Aadirupa Saha, Aniket Wagde, and Branislav Kveton. Llm-as-judge on a budget. *arXiv preprint arXiv:2602.15481*, 2026.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in llm-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 292–314, 2025.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. *Advances in neural information processing systems*, 35:9460–9471, 2022.

Ajay Narayanan Sridhar, Ronak Singh, Mehrdad Mahdavi, and Vijaykrishnan Narayanan. Correlation-aware contextual bandits with surrogate rewards for llm routing. *arXiv preprint arXiv:2607.09015*, 2026.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, 2023. doi: 10.18653/v1/2023.findings-acl.824. URL https://aclanthology.org/2023.findings-acl.824/.

Xiaqiang Tang, Qiang Gao, Jian Li, Nan Du, Qi Li, and Sihong Xie. MBA-RAG: a bandit approach for adaptive retrieval-augmented generation through question complexity. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 3248–3254, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.218/.

Aaron D Tucker, Caleb Biddulph, Claire Wang, and Thorsten Joachims. Bandits with costly reward observations. In *Uncertainty in Artificial Intelligence*, pp. 2147–2156. PMLR, 2023.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.511.

Xuchuang Wang, Qingyun Wu, Wei Chen, and John C.S. Lui. Multi-fidelity multi-armed bandits revisited. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=oi45JlpSOT.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pp. 95266–95290, 2024b. doi: 10.52202/079017-3018.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. LiveBench: A challenging, contamination-limited LLM benchmark. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2406.19314.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *International Conference on Learning Representations*, 2024.

Xianglin Yang, Bryan Hooi, Gelei Deng, Tianwei Zhang, and Jin Song Dong. Turning bias into bugs: Bandit-guided style manipulation attacks on llm judges. *arXiv preprint arXiv:2605.26156*, 2026.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Yixuan Zhang, Ruihao Zhu, and Qiaomin Xie. Contextual online pricing with (biased) offline data. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 13372–13419. Curran Associates, Inc., 2025. doi: 10.52202/085713-0452. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/14128f3c4724b96955799d684c495813-Paper-Conference.pdf.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023. doi: 10.52202/075280-2020.

A Technical Details of PACE

This appendix provides the complete algorithmic and statistical details of **Paired-Audit Calibrated Exploration (PACE)**, together with the simplified PACE-U variant used in our theoretical analysis. Section A.1 gives the full pseudocode. Sections A.2–A.6 detail the estimators, risk-based audit allocation, residual-mismatch calibration, and the conditions underlying the confidence-based interpretation of PACE. Section A.7 then specifies PACE-U, which replaces context-dependent auditing with uniform auditing and assumes exact linear realizability.

A.1 Full PACE and PACE-U Algorithm

Algorithm 1 summarizes the complete online procedure. At every round, the audit probability is determined from information available before observing the current output or rewards. The proxy reward is observed for every selected action, whereas the clean reward is acquired only when the corresponding audit indicator is active.

A.2 PACE: Estimators and Weighted Calibration

We provide the complete estimator definitions underlying the abbreviated construction in Section 4. PACE maintains three linear predictors: a clean-reward model, a proxy-reward model, and a paired-difference model. The latter targets the conditional mean discrepancy

$$\bar{b}_t^\star(a) = \mathbb{E}\left[R_t^p - R_t^c \mid \mathcal{H}_{t-1}, x_t, A_t = a\right]. \tag{17}$$

This conditional mean averages over output randomness and reward noise.

Let

$$\phi_t = \phi(x_t, A_t).$$

On an audited round, PACE observes the paired discrepancy

$$D_t = R_t^p - R_t^c.$$

Because PACE uses nonuniform auditing, an audited observation at round s is assigned weight

$$\nu_s = \frac{M_s}{\max\{p_s, p_0\}}, \qquad p_0 > 0, \tag{18}$$

Algorithm 1 Paired-Audit Calibrated Exploration (PACE)

Require: Horizon T, clean-feedback budget B, variant $v \in \{\text{PACE}, \text{PACE-U}\}$, and variant-specific parameters

Initialize the clean-, proxy-, and discrepancy-model statistics
if $v =$ PACE **then**
    Initialize the risk model, audited dataset, risk-score history, and residual history
else
    Use unit regularization and the confidence coefficients in Section A.7
end if
for $t = 1, \ldots, T$ **do**
    Observe x_t, $\mathcal{A}(x_t)$, and candidate features
    Compute the clean, proxy, and discrepancy predictions and uncertainties
    if $v =$ PACE-U **then**
        Set p_t according to equation 50 and $\epsilon_{J,t}(a) = 0$
    else
        Compute $z_t = z(x_t)$ and $r_t = g_{\psi_t}(z_t)$
        Compute p_t using equation 32–equation 33
        Compute $\epsilon_{J,t}(a)$ from past audited residuals using equation 38
    end if
    Compute $U_t^c(a)$ and $U_t^{p\to c}(a)$ using equation 5
    Select A_t by equation 6 and set $\phi_t = \phi(x_t, A_t)$
    Execute A_t and observe the proxy reward R_t^p
    Draw an independent $M_t \sim \text{Bernoulli}(p_t)$
    if $M_t = 1$ **then**
        Acquire R_t^c and set $D_t = R_t^p - R_t^c$
        if $v =$ PACE **then**
            Form Y_t^{risk} using equation 29
            Compute the pre-update residual $e_{J,t}$ using equation 37
            Store the audited context, label, propensity, and residual
            Update the clean and discrepancy models with weight $1/\max\{p_t, p_0\}$
            Update the risk model using equation 30
        else
            Update clean/discrepancy models with unit weight
        end if
    end if
    Update the proxy model using R_t^p
    if $v =$ PACE **then**
        Append r_t to the risk-score history
    end if
end for

where p_s is the probability with which round s is audited.

With positive regularization parameters $\lambda_c, \lambda_p, \lambda_b$, the clean, proxy, and discrepancy estimators are constructed as

$$V_t^c = \lambda_c I_d + \sum_{s<t} \nu_s \phi_s \phi_s^\top, \qquad \widehat{\theta}_t^c = (V_t^c)^{-1} \sum_{s<t} \nu_s \phi_s R_s^c, \tag{19}$$

$$V_t^p = \lambda_p I_d + \sum_{s<t} \phi_s \phi_s^\top, \qquad \widehat{\theta}_t^p = (V_t^p)^{-1} \sum_{s<t} \phi_s R_s^p, \tag{20}$$

$$V_t^b = \lambda_b I_d + \sum_{s<t} \nu_s \phi_s \phi_s^\top, \qquad \widehat{\theta}_t^b = (V_t^b)^{-1} \sum_{s<t} \nu_s \phi_s D_s. \tag{21}$$

For any candidate action $a \in \mathcal{A}(x_t)$, define

$$\widehat{\mu}_t^c(a) = \phi(x_t, a)^\top \widehat{\theta}_t^c, \tag{22}$$

$$\widehat{\mu}_t^p(a) = \phi(x_t, a)^\top \widehat{\theta}_t^p, \tag{23}$$

$$\widehat{b}_t(a) = \phi(x_t, a)^\top \widehat{\theta}_t^b, \tag{24}$$

with feature uncertainty

$$u_t^j(a) = \|\phi(x_t, a)\|_{(V_t^j)^{-1}}, \qquad j \in \{c, p, b\}. \tag{25}$$

The clean and discrepancy models therefore use only audited observations, with inverse-propensity weighting to account for nonuniform audit probabilities, whereas the proxy model uses every observed proxy reward. The clean and discrepancy design matrices satisfy

$$V_t^b - V_t^c = (\lambda_b - \lambda_c) I_d, \tag{26}$$

and are identical only when $\lambda_b = \lambda_c$.

Fitting a linear discrepancy predictor does not itself imply exact linear realizability. PACE therefore treats this predictor as a working model and retains the residual-mismatch correction $\epsilon_{J,t}(a)$ described in Section A.5.

A.3 Risk-Model Training

Features and target. PACE allocates clean-feedback queries according to predicted proxy–reference disagreement. The risk predictor receives only information available from the context before arm execution. Its feature construction is

$$z(x_t) = [e(x_t); \log(1 + \mathrm{len}(x_t)); d(x_t); c(x_t); v(x_t)], \tag{27}$$

where e is a context embedding, d is a predicted difficulty score, c encodes the task category, and v contains contextual cues such as format constraints or subjective evaluation requirements.

A logistic head or small neural network produces

$$r_t = g_{\psi_t}(z_t) \in [0, 1]. \tag{28}$$

For an audited round, the prediction target is

$$Y_t^{\mathrm{risk}} = \mathbf{1}\{|R_t^p - R_t^c| \geq \tau_{\mathrm{risk}}\}. \tag{29}$$

Because the risk predictor does not receive the selected arm, it learns a context-level disagreement risk under the evolving arm-selection policy. The predicted event depends on realized rewards and therefore reflects both systematic disagreement and reward noise rather than conditional-mean bias alone. It is also distinct from the expected regret reduction obtained by purchasing an audit.

Clipped inverse-propensity objective. Using previously audited rounds, PACE fits

$$\psi_t \in \arg\min_{\psi} \left\{ \sum_{\substack{s<t \\ M_s=1}} \frac{\mathrm{BCE}\left(Y_s^{\mathrm{risk}}, g_\psi(z_s)\right)}{\max\{p_s, p_0\}} + \lambda_\psi \|\psi\|_2^2 \right\}. \tag{30}$$

Here, BCE denotes binary cross-entropy.

The risk model is updated whenever a new audited pair becomes available. The initial warm-up uses context-independent random audits.

For positive audit probabilities, untruncated inverse-propensity weighting corrects for unequal audit selection. Clipping stabilizes the weights but changes this correction whenever $p_s < p_0$. Neither clipping nor propensity weighting reconstructs clean labels from rounds assigned zero audit probability.

A.4 Risk-to-Audit Calibration

Let

$$B_t = B - \sum_{s<t} M_s, \qquad q_t = \frac{B_t}{T-t+1}, \tag{31}$$

where B_t denotes the remaining clean-feedback budget and q_t its target average consumption rate.

Given a collection $\mathcal{R}_t$ of previous risk scores, PACE chooses a scale $c_t \geq 0$ satisfying approximately

$$\frac{1}{|\mathcal{R}_t|} \sum_{r \in \mathcal{R}_t} \min\{1, c_t r\} \approx q_t. \tag{32}$$

The resulting audit probability is

$$p_t = \begin{cases} 0, & B_t = 0, \\ q_t, & B_t > 0 \text{ and during warm-up with insufficient history}, \\ \min\{1, c_t r_t\}, & \text{otherwise}. \end{cases} \tag{33}$$

Thus, contexts with larger predicted disagreement receive larger audit probabilities before saturation. Importantly, p_t is determined before observing the current output or either current reward.

For a nonempty score history, define

$$F_t(c) = \frac{1}{|\mathcal{R}_t|} \sum_{r \in \mathcal{R}_t} \min\{1, cr\}. \tag{34}$$

This function is continuous and nondecreasing, supporting bisection whenever the target can be bracketed.

Zero historical risk scores can restrict the attainable average audit probability:

$$\sup_{c \geq 0} F_t(c) = \frac{|\{r \in \mathcal{R}_t : r > 0\}|}{|\mathcal{R}_t|}. \tag{35}$$

Consequently, the calibration equation need not admit an exact solution for every q_t. The empirical implementation therefore uses approximate matching rather than requiring equality.

Whenever the controller returns valid audit probabilities and sets $p_t = 0$ after $B_t = 0$, the hard audit budget holds pathwise:

$$\sum_{t=1}^{T} M_t \leq B.$$

This property does not depend on accurate risk prediction or exact risk-to-budget calibration.

A.5 Empirical Residual-Mismatch Calibration

PACE retains an additional correction $\epsilon_{J,t}(a) \geq 0$ in its bias-corrected proxy confidence index. The intended bias-coverage condition is

$$\left| \bar{b}_t^{\star}(a) - \widehat{b}_t(a) \right| \leq \alpha_b u_t^b(a) + \epsilon_{J,t}(a). \tag{36}$$

For each audited round s, PACE computes a pre-update residual

$$e_{J,s} = \left[\left| D_s - \widehat{b}_s^{\mathrm{pre}}(A_s) \right| - \alpha_b u_s^b(A_s) \right]_+, \tag{37}$$

where $[v]_+ = \max\{0, v\}$. Both the discrepancy prediction and its uncertainty are evaluated before incorporating the current audited pair.

For a candidate (x_t, a), let $\mathcal{I}_t(a)$ denote previous audited rounds with similar context–arm features. PACE estimates

$$\widehat{\epsilon}_{J,t}(a) = \mathrm{Quantile}_{1-\delta}\left(\{ e_{J,s} : s \in \mathcal{I}_t(a) \} \right). \tag{38}$$

When insufficient locally similar audited samples are available, PACE uses a global quantile over previous audited residuals instead.

This construction is an empirical calibration rule. The residuals contain paired reward noise and are collected under adaptive, nonuniform auditing. Consequently, an empirical quantile alone does not establish equation 36 uniformly across all rounds and candidate actions. The feature-similarity rule, minimum local sample count, and initialization before sufficient residuals are available are implementation choices and should be distinguished from the statistical coverage conditions below.

A.6 Conditions for Confidence-Based Interpretation

For PACE, define

$$\mu^\star_{p,t}(a) = \mathbb{E}\left[R^p_t \mid \mathcal{H}_{t-1}, x_t, A_t = a\right]$$

and let the clean conditional mean be $f^\star(x_t, a)$. Then

$$\bar{b}^\star_t(a) = \mu^\star_{p,t}(a) - f^\star(x_t, a).$$

A sufficient condition for both PACE indices to be valid upper bounds on the clean reward is the simultaneous event

$$|\widehat{\mu}^c_t(a) - f^\star(x_t, a)| \le \alpha_c u^c_t(a), \tag{39}$$

$$\left|\widehat{\mu}^p_t(a) - \mu^\star_{p,t}(a)\right| \le \alpha_p u^p_t(a), \tag{40}$$

$$\left|\widehat{b}_t(a) - \bar{b}^\star_t(a)\right| \le \alpha_b u^b_t(a) + \epsilon_{J,t}(a), \tag{41}$$

for every round and candidate action.

Under these conditions,

$$\left|\widehat{\mu}^p_t(a) - \widehat{b}_t(a) - f^\star(x_t, a)\right| \le \alpha_p u^p_t(a) + \alpha_b u^b_t(a) + \epsilon_{J,t}(a). \tag{42}$$

Writing

$$S_t(a) = \min\left\{U^c_t(a), U^{p\to c}_t(a)\right\}, \tag{43}$$

the PACE action-selection rule implies

$$f^\star(x_t, a^\star_t) - f^\star(x_t, A_t) \le 2\min\left\{\alpha_c u^c_t(A_t),\ \alpha_p u^p_t(A_t) + \alpha_b u^b_t(A_t) + \epsilon_{J,t}(A_t)\right\}. \tag{44}$$

This argument does not require the three estimators to be independent. However, these inequalities are sufficient coverage conditions rather than guarantees implied merely by fitting the models. In particular, inverse-propensity-weighted regressions require an appropriate confidence analysis, and empirical residual quantiles alone do not establish uniform high-probability coverage.

A cumulative regret guarantee for PACE additionally requires control of audit coverage together with accumulated estimation and residual-mismatch errors. Accordingly, the theoretical guarantee established below for PACE-U does not automatically extend to the context-dependent auditing policy of PACE.

A.7 PACE-U: Simplified Linear-Uniform Variant

PACE-U is the simplified variant used in the theoretical analysis of Section 6. It differs from PACE in two main respects: (i) it assumes exact linear realizability of the clean and proxy conditional means, and hence of their discrepancy; and (ii) it allocates clean-feedback queries uniformly rather than using the learned risk predictor.

Linear conditional means and parameter bounds. Let $\mathcal{X}$ denote the context space. At round $t = 1, \ldots, T$, the learner observes context x_t and a finite, nonempty candidate action set $\mathcal{A}(x_t)$. Actions are represented by a known feature map

$$\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d, \qquad \|\phi(x, a)\|_2 \le 1.$$

For the executed action A_t, let

$$\phi_t = \phi(x_t, A_t).$$

PACE-U assumes that the clean and proxy rewards satisfy

$$R_t^c = \phi_t^\top \theta_c^\star + \xi_t^c, \qquad R_t^p = \phi_t^\top \theta_p^\star + \xi_t^p, \qquad \theta_b^\star = \theta_p^\star - \theta_c^\star. \tag{45}$$

The corresponding conditional means lie in $[0, 1]$:

$$\phi_t^\top \theta_c^\star \in [0, 1], \qquad \phi_t^\top \theta_p^\star \in [0, 1].$$

The parameter vectors satisfy

$$\|\theta_j^\star\|_2 \le S_j, \qquad j \in \{c, p, b\}.$$

Under these assumptions, the conditional-mean proxy bias is itself exactly linear:

$$b^\star(x, a) = \phi(x, a)^\top \theta_b^\star. \tag{46}$$

Hence there is no residual model misspecification and PACE-U sets

$$\epsilon_{J,t}(a) = 0. \tag{47}$$

Stochastic paired discrepancy and noise structure. Let $\mathcal{G}_t$ denote the filtration containing the pre-observation history, current context x_t, and selected action A_t, before the realization of the audit indicator and contemporaneous reward noise.

For $j \in \{c, p, b\}$, the noise ξ_t^j is conditionally zero-mean and σ_j-sub-Gaussian:

$$\mathbb{E}\left[\exp(u\xi_t^j) \mid \mathcal{G}_t\right] \le \exp\left(\frac{u^2\sigma_j^2}{2}\right), \qquad u \in \mathbb{R}. \tag{48}$$

The clean and proxy reward noises may be arbitrarily correlated. Defining

$$\xi_t^b = \xi_t^p - \xi_t^c,$$

the paired discrepancy satisfies

$$D_t = R_t^p - R_t^c = \phi_t^\top \theta_b^\star + \xi_t^b. \tag{49}$$

Exact linearity of the conditional mean therefore does not imply that an individual paired discrepancy is noiseless.

Uniform auditing. Let

$$N_{t-1} = \sum_{s<t} M_s, \qquad B_t = B - N_{t-1}.$$

PACE-U uses

$$p_t^U = \frac{B}{T}\mathbf{1}\{B_t > 0\}. \tag{50}$$

The audit indicator is drawn as

$$M_t \sim \text{Bernoulli}(p_t^U),$$

conditionally independently of the current reward pair given the pre-observation history. The cap $\mathbf{1}\{B_t > 0\}$ guarantees

$$\sum_{t=1}^{T} M_t \le B$$

almost surely, although the realized number of audits need not equal B.

PACE-U estimators. PACE-U uses unit ridge regularization

$$\lambda_c = \lambda_p = \lambda_b = 1$$

and uniform audited weights $\nu_s = M_s$. Its estimators are therefore

$$V_t^c = I_d + \sum_{s<t} M_s \phi_s \phi_s^\top, \qquad \widehat{\theta}_t^c = (V_t^c)^{-1} \sum_{s<t} M_s \phi_s R_s^c, \tag{51}$$

$$V_t^p = I_d + \sum_{s<t} \phi_s \phi_s^\top, \qquad \widehat{\theta}_t^p = (V_t^p)^{-1} \sum_{s<t} \phi_s R_s^p, \tag{52}$$

$$V_t^b = I_d + \sum_{s<t} M_s \phi_s \phi_s^\top, \qquad \widehat{\theta}_t^b = (V_t^b)^{-1} \sum_{s<t} M_s \phi_s D_s. \tag{53}$$

For candidate action a, define

$$u_t^j(a) = \|\phi(x_t, a)\|_{(V_t^j)^{-1}}, \qquad j \in \{c, p, b\}.$$

Because the clean and discrepancy estimators are based on the same audited samples with the same regularization,

$$V_t^c = V_t^b.$$

Confidence coefficients. Because at most B observations enter the clean and discrepancy regressions while at most T observations enter the proxy regression, define

$$L_c = L_b = d \log\left(1 + \frac{B}{d}\right), \qquad L_p = d \log\left(1 + \frac{T}{d}\right). \tag{54}$$

For failure probability $\delta \in (0, 1)$, the confidence coefficients are

$$\alpha_j = \max\left\{\frac{1}{2},\ S_j + \sigma_j \sqrt{L_j + 2\log\frac{5}{\delta}}\right\}, \qquad j \in \{c, p, b\}. \tag{55}$$

Under the assumptions above, standard self-normalized concentration yields simultaneous confidence ellipsoids for the clean, proxy, and discrepancy estimators.

PACE-U then uses the same calibrated action-selection rule as PACE,

$$A_t \in \arg\max_{a \in \mathcal{A}(x_t)} \min\left\{U_t^c(a), U_t^{p \to c}(a)\right\}, \tag{56}$$

but with $\epsilon_{J,t}(a) = 0$. These assumptions, estimators, and confidence coefficients are used throughout Section 6 and Appendix C.

B Additional Experimental Details and Analyses

This appendix documents the datasets, cached feedback construction, replay protocol, implementation choices, and additional analyses supporting Section 5. Unless stated otherwise, regret refers to the empirical cumulative clean regret against the per-question cached oracle in Equation 12. The reproducibility records include question identifiers, source hashes, experimental configurations, and per-seed outputs.

B.1 Datasets, Actions, and Feedback Construction

B.1.1 Datasets and Partitions

Routing datasets. IFBench uses all 300 questions in the pinned snapshot, while LiveBench uses a 500-question subset of the 2024-11-25 release. Arena-Hard v2 contains 750 questions. For MMLU-Pro, we use 1,400 questions, with 100 sampled from each subject. Within each routing dataset, the two proxy-judge conditions share the same question IDs, cached candidate responses, clean-reward matrices, and reward features.

Table 3: Question pools used in the real-data experiments. Warm and online questions are disjoint, and the warm split provides proxy feedback only.

Dataset	Total	Warm	Online
IFBench	300	30	270
LiveBench	500	50	450
Arena-Hard v2	750	75	675
MMLU-Pro	1400	140	1260
BBH	800	80	720
HotpotQA	800	80	720

BBH and HotpotQA. Each source cache contains 800 questions, split into 80 proxy-only warm questions and 720 online questions. Both datasets use sampling seed 91723 and exclude questions appearing in earlier pilot experiments. After case-folding and whitespace-normalized deduplication, records are sorted by the SHA-256 digest of the sampling seed and question ID, and the first 800 are retained. The warm split is selected independently by sorting these retained IDs by the SHA-256 digest of `pace-warm-pilot|<ID>` and taking the first $\lceil 0.1n \rceil$ examples. The warm split is not task-stratified.

B.1.2 Actions and Candidate Generation

LLM routing. Routing uses four candidate arms: `mistralai/mistral-small-3.2-24b-instruct`, `qwen/qwen3-235b-a22b-2507`, `deepseek/deepseek-v3.2`, and `openai/gpt-4.1`. The first two use `deepinfra/fp8`, DeepSeek uses `novita/fp8`, and GPT-4.1 uses `openai`. Candidate generation uses temperature zero and an 8,192-token output cap; DeepSeek additionally uses the configured medium reasoning-effort setting. Provider fallbacks are disabled. These identifiers denote the APIs used during data collection and should not be interpreted as immutable remote model weights.

Prompt selection. BBH uses six fixed prompt strategies: *direct*, *stepwise*, *decompose*, *verify*, *concise*, and *structured*. Mistral Small 3.2 (24B) generates all candidate responses with temperature zero and a 2,048-token output cap.

Retrieval configuration selection. HotpotQA uses four actions: no retrieval, BM25 top-2, BM25 top-5, and all supplied paragraphs. Mistral Small 3.2 (24B) again generates responses with temperature zero and a 2,048-token output cap.

Each supplied HotpotQA title and its space-joined sentences form one paragraph. Our local BM25 implementation tokenizes titles and text using lowercase `[a-z0-9]+` matches, with $k_1 = 1.5$, $b = 0.75$, and inverse document frequency

$$\log\left(1 + \frac{N - df + 0.5}{df + 0.5}\right).$$

Each distinct query token contributes once. Paragraphs are ranked by decreasing BM25 score, with the original paragraph index used to break ties. The all-paragraphs action preserves source order, whereas the no-retrieval action uses an empty document list. No demonstrations or gold answers are inserted.

For BBH and HotpotQA, we retain one response and one proxy score for each question–arm pair in the cached reward matrix. Transport or parsing retries are not treated as independent samples, and missing required matrix cells cause cache construction to fail rather than being imputed. Malformed final-answer markers receive zero clean reward while retaining their proxy-judge score.

B.1.3 Proxy and Reference Rewards

Proxy rewards. Gemma-3-27B-IT and GPT-4o are evaluated under fixed proxy-scoring conditions. For routing, the judge uses the style-preference rubric in Appendix B.5, which explicitly rewards polished and comprehensive presentation; the resulting score in $[1, 10]$ is divided by 10. BBH and

HotpotQA use the corresponding reference-free rubric with scores reported directly in $[0, 1]$. Judge temperature is zero, with output caps of 1,024 tokens for routing and 512 tokens for BBH and HotpotQA. These settings are designed to induce a controlled proxy–reference mismatch rather than to establish a naturally occurring bias in either judge.

Task-based reference rewards. IFBench uses the pinned prompt-level loose instruction-following accuracy, LiveBench uses the pinned official per-question task score, and MMLU-Pro uses the pinned answer-extraction procedure with binary correctness.

For BBH, a valid response contains exactly one line beginning with `FINAL:`. Evaluation case-folds and collapses whitespace, and treats a parenthesized single-option letter as equivalent to the corresponding bare letter. For HotpotQA, evaluation lowercases text, removes ASCII punctuation and articles, collapses whitespace, and computes token-overlap F1. Incompatible yes/no/noanswer predictions receive zero. We use answer F1 as the clean objective; exact match is retained only as a secondary cached metric.

Arena-Hard v2 reference rewards. Both proxy conditions share the same reference-reward matrix. GPT-4.1 compares each candidate response against the category-specific frozen baseline response from either `o3-mini-2025-01-31` or `gemini-2.0-flash-001`. Candidate and baseline positions are swapped across two comparisons. The parser recognizes `A»B`, `A>B`, `A=B`, `B>A`, and `B»A`; either strength of a win maps to 1, a tie to 0.5, and a loss to 0 from the candidate's perspective. The two comparisons are averaged to obtain one cached reference reward.

A simulated online audit reveals this cached reward, although constructing it offline requires two comparison calls. We therefore interpret this signal as preference-based reference feedback rather than objective correctness or the official Arena-Hard leaderboard score. Because GPT-4.1 is also one of the routing arms, candidate and evaluator preferences may be correlated.

B.2 Replay Protocol and Implementation Details

B.2.1 Online Replay and Information Access

For each replay seed, every online epoch uses a fresh NumPy permutation of the online question pool. Epochs are concatenated until the horizon T is reached, and learner state persists across epoch boundaries. Cached responses and reward matrices are reused throughout replay.

At each round, the learner observes only the selected action's proxy reward; the corresponding clean reward is revealed only if the round is audited. PACE computes its risk score and audit probability from the current context and past observations before observing the current output or either reward. The current candidate answer and its score are therefore excluded from the risk features.

Each audited replay round consumes one unit of budget. If the same question–action pair is audited again in a later epoch, it consumes another unit. Previously observed clean labels remain encoded in learner state, but the cached reference label is not re-exposed for free. Thus, B counts online reference queries in the simulated protocol rather than distinct cached labels or monetary API cost.

All methods share the same all-arm, proxy-only warm set. PACE and PACE-U use it only to initialize their proxy regressions; their clean-reward, discrepancy, residual, and risk components receive no warm clean labels. Clean LinUCB ignores proxy feedback. The first 20 online scheduling rounds use context-independent budget matching; any audits realized during these rounds count toward the same budget.

B.2.2 Features and Hyperparameters

For routing, the reward model uses

$$x = [1; e(q)],$$

where $e(q)$ is a frozen 768-dimensional `sentence-transformers/all-mpnet-base-v2` embedding obtained by attention-mask mean pooling followed by L2 normalization. We use disjoint-arm features

$$\phi(q, a) = \mathbf{e}_a \otimes x,$$

corresponding to 769 features per arm.

Table 4: Implementation settings used for PACE and PACE-U in the empirical proxy and budget comparisons.

Parameter	Value / rule
$\lambda_c, \lambda_p, \lambda_b$	$1, 1, 1$
$\alpha_c, \alpha_p, \alpha_b$	$1, 1, 1$
IPW floor p_0	0.01
Risk target	$\mathbf{1}\{\lvert R^p - R^c\rvert \geq 0.2\}$
Risk penalty	$0.001\lVert\psi\rVert_2^2$ including intercept
Initial risk intercept	logit of clipped B/T in $[0.01, 0.99]$
Risk refit	after every audit using all previous weighted pairs
Residual quantile	IPW-weighted 0.9 quantile
Local residual rule	up to 20 nearest same-arm samples
Minimum local samples	10; otherwise use the global residual bank
Empty residual bank	zero residual allowance
Reward scores	cached values; no additional clipping or normalization
Action ties	uniform random selection among exact maximizers

The risk model uses an intercept, the question embedding, log character length, a frozen question-only GPT-4.1-mini difficulty estimate, dataset-specific category indicators, and six lexical/context features for output-format constraints, length constraints, long context, subjective criteria, multistep instructions, and code/math content. The difficulty estimator uses temperature zero and a 128-token cap and receives no candidate response, gold answer, or reward information.

The risk model minimizes the inverse-propensity-weighted binary cross-entropy objective with the regularization in Table 4. Optimization uses L-BFGS-B, warm-started from the previous coefficients, with gradient tolerance 10^{-8}, function tolerance 10^{-14}, at most 1,000 iterations, and at most 50 line-search steps; a Newton refinement is applied when needed. Fits with gradient infinity norm above 10^{-5} are discarded.

Audited clean-reward and discrepancy updates use weight $1/\max\{p_t, 0.01\}$, whereas proxy updates use unit weight. Residual calibration uses pre-update discrepancy predictions and widths, IPW-weighted order statistics, and stable Euclidean nearest-neighbor ordering. The empirical width coefficients are fixed hyperparameters and are distinct from the confidence coefficients constructed for the theory in Section 6.

B.2.3 Baseline Implementations

Proxy LinUCB uses the same reward-model features, ridge regularization, and empirical width coefficient as the corresponding PACE proxy model. Scalar UCB1 uses exploration bonus

$$\sqrt{\frac{2\log(\sum_a N_a)}{N_a}},$$

and ϵ-greedy uses $\epsilon = 0.1$. Thompson sampling starts from $\mathrm{Beta}(1, 1)$ and adds fractional proxy rewards and their complements to its pseudo-counts; these updates should therefore be interpreted as heuristic fractional-count updates rather than Bernoulli observations. Scalar baselines break exact ties uniformly at random.

The online MetaLLM adaptation uses the fixed implementation with regularization 12.94, exploration coefficient $\alpha \in \{0, 1\}$, and zero cost coefficient. Online updates use only the selected action's proxy feedback.

The ProTeGi-derived comparator (Pryzant et al., 2023) retains the UCB-E selection component over the six fixed BBH prompts, with $c = 2$, batch size one, all-arm proxy warm updates, and selected-proxy online updates. It does not perform prompt generation, textual-gradient optimization, or beam search.

The MBA-RAG-derived comparator (Tang et al., 2025) performs full DistilBERT updates over the four fixed HotpotQA actions. It encodes only the question text, truncates or pads inputs to 256 tokens, and uses batch size one with AdamW at learning rate 5×10^{-5} and a linear learning-rate

Table 5: Routing performance at $T = 1500$ and audit fraction $b = 0.2$. Entries are cumulative clean regret $\widehat{\mathrm{Reg}}_c(T)$, reported as mean $\pm$ sample standard deviation over 20 paired replay seeds. Lower is better. Column subheadings identify the proxy judge.

Method	IFBench		LiveBench	
Judge	Gemma	GPT-4o	Gemma	GPT-4o
No clean-feedback access ($B = 0$)				
Random	429.90 ± 12.87	429.90 ± 12.87	315.82 ± 9.13	315.82 ± 9.13
Proxy UCB1	438.00 ± 12.58	414.70 ± 13.78	321.38 ± 9.41	321.67 ± 10.89
Proxy ϵ-greedy	533.45 ± 43.80	342.50 ± 134.61	298.95 ± 4.19	338.78 ± 5.78
Proxy Thompson sampling	469.25 ± 23.26	406.80 ± 40.98	324.44 ± 14.52	339.74 ± 6.34
Proxy LinUCB	437.45 ± 4.16	428.95 ± 6.15	289.68 ± 8.19	270.77 ± 7.19
MetaLLM-online ($\alpha = 0$)	613.10 ± 48.48	366.45 ± 266.01	300.72 ± 4.48	344.53 ± 3.99
MetaLLM-online ($\alpha = 1$)	613.10 ± 48.48	392.40 ± 264.88	300.57 ± 4.61	342.35 ± 9.05
Budget-matched baselines (audit cap $B = 300$)				
Proxy LinUCB + R^c repl.	375.45 ± 11.15	365.80 ± 8.31	252.70 ± 11.14	246.83 ± 12.93
Clean LinUCB (uniform)	222.50 ± 40.29	222.50 ± 40.29	183.19 ± 24.46	183.19 ± 24.46
Proposed methods (audit cap $B = 300$)				
PACE-U	197.20 ± 60.97	229.90 ± 46.48	154.98 ± 20.51	181.78 ± 21.17
PACE	180.40 ± 54.46	178.20 ± 60.57	154.81 ± 24.76	171.82 ± 30.94

Gemma denotes Gemma-3-27B-IT. All methods have access to shared proxy-only warm-up data; no offline clean labels are provided. The audit cap is 300: PACE-U uses 296.15 ± 5.34 audits, while the other audited methods shown use 300. MetaLLM denotes the online adaptation.

schedule without warm-up. One pass over the 80 shared warm questions regresses all four proxy scores, followed by one selected-action regression update at each online round; unselected targets are detached predictions. Exploration uses $\epsilon = 0.1$, and the retrieval-cost coefficient is zero.

These task-specific comparators adapt components of ProTeGi and MBA-RAG to our fixed action spaces and feedback protocol; they are not intended as full reproductions of the original systems.

B.2.4 Metrics, Budgets, and Random Seeds

The primary real-data horizon is $T = 1500$. Budget sweeps use

$$B \in \{75, 150, 300, 450, 600, 750\},$$

corresponding to audit fractions

$$b \in \{0.05, 0.10, 0.20, 0.30, 0.40, 0.50\}.$$

The context permutation uses the replay seed, action tie-breaking uses seed+313, and the audit RNG uses seed+1701. This pairing aligns the initial random-number streams without forcing adaptive policies to realize identical action sequences or audit masks.

Error bands and tabulated variation report the mean plus/minus one sample standard deviation (ddof$=1$) over matched replay seeds. Because the questions, generated responses, and cached rewards are fixed, this variation reflects the online learning process conditional on the cached experiment; it is not a confidence interval over newly sampled questions or fresh judge calls.

B.3 Additional Results

B.3.1 Additional Routing Results

The Arena-Hard v2 and MMLU-Pro experiments use the same routing arms, proxy construction, replay protocol, and PACE hyperparameters as the primary IFBench and LiveBench experiments. We include them as additional evidence across task distributions, while keeping the main-text discussion focused on the two primary routing benchmarks.

Table 6: Additional routing performance on MMLU-Pro and Arena-Hard at $T = 1500$ and audit fraction $b = 0.2$. Entries are cumulative clean regret $\widehat{\mathrm{Reg}}_c(T)$, reported as mean $\pm$ sample standard deviation over 20 paired replay seeds. Lower is better. Column subheadings identify the proxy judge.

Method	Arena-Hard		MMLU-Pro	
Proxy judge	Gemma	GPT-4o	Gemma	GPT-4o
No clean-feedback access ($B = 0$)				
Random	475.77 ± 15.45	475.77 ± 15.45	176.90 ± 8.69	176.90 ± 8.69
Proxy UCB1	418.64 ± 12.28	425.19 ± 13.02	174.55 ± 8.56	188.70 ± 10.52
Proxy ϵ-greedy	269.25 ± 5.21	491.32 ± 4.65	153.55 ± 9.90	158.50 ± 4.39
Proxy Thompson sampling	313.56 ± 13.86	393.12 ± 22.56	158.10 ± 6.93	159.90 ± 4.46
Proxy LinUCB	428.55 ± 5.97	431.44 ± 9.92	174.90 ± 6.90	182.85 ± 6.06
MetaLLM-online ($\alpha = 0$)	248.84 ± 3.73	485.14 ± 38.46	153.85 ± 11.18	157.50 ± 4.01
MetaLLM-online ($\alpha = 1$)	248.84 ± 3.63	485.19 ± 38.49	154.55 ± 10.55	157.50 ± 4.01
Budget-matched baselines				
Proxy LinUCB + R^c repl.	409.30 ± 12.06	407.86 ± 13.78	167.65 ± 7.92	178.70 ± 8.37
Clean LinUCB (uniform)	324.91 ± 12.20	324.91 ± 12.20	145.10 ± 33.64	145.10 ± 33.64
Proposed methods				
PACE-U	308.18 ± 40.71	316.20 ± 22.03	126.00 ± 26.93	136.05 ± 27.65
PACE	304.65 ± 30.50	300.14 ± 30.34	126.25 ± 25.74	127.15 ± 26.50

Gemma denotes the Gemma-3-27B-IT proxy judge. All methods have access to shared proxy-only warm-up data; no offline clean labels are provided.

Table 7: Prompt and retrieval configuration selection performance at $T = 1500$ and audit fraction $b = 0.2$. Entries are cumulative clean regret $\widehat{\mathrm{Reg}}_c(T)$, reported as mean $\pm$ sample standard deviation over 20 paired replay seeds. Lower is better. Column subheadings identify the proxy judge.

Method	BBH		HotpotQA	
Judge	Gemma	GPT-4o	Gemma	GPT-4o
No clean-feedback access ($B = 0$)				
Random	283.70 ± 17.72	283.70 ± 17.72	374.02 ± 15.57	374.02 ± 15.57
Proxy UCB1	273.05 ± 11.44	185.85 ± 11.66	189.40 ± 7.09	169.12 ± 7.10
Proxy ϵ-greedy	131.10 ± 10.53	142.45 ± 9.90	139.02 ± 4.61	139.76 ± 5.23
Proxy Thompson sampling	153.80 ± 21.35	122.80 ± 8.46	125.63 ± 4.90	121.82 ± 5.58
Proxy LinUCB	245.65 ± 9.00	145.05 ± 9.03	231.26 ± 7.47	213.21 ± 6.70
ProTeGi UCB-E selector	225.95 ± 13.65	137.40 ± 9.00	-	-
MBA-RAG encoder	-	-	183.68 ± 9.46	174.15 ± 10.30
Budget-matched baselines (audit cap $B = 300$)				
Proxy LinUCB + R^c repl.	194.80 ± 9.66	135.25 ± 8.67	225.93 ± 10.38	218.99 ± 9.85
Clean LinUCB (uniform)	156.80 ± 18.35	156.80 ± 18.35	198.60 ± 21.77	198.60 ± 21.77
Proposed methods (audit cap $B = 300$)				
PACE-U	152.40 ± 23.04	153.40 ± 23.65	191.94 ± 29.45	195.37 ± 33.93
PACE	142.10 ± 17.34	147.95 ± 30.84	172.87 ± 33.60	166.04 ± 27.64

Gemma denotes Gemma-3-27B-IT. All methods have access to shared proxy-only warm-up data; no offline clean labels are provided. Both ProTeGi UCB-E selector and MBA-RAG encoder are online adaptations of the corresponding original methods.

B.3.2 Prompt and Retrieval Selection

PACE improves over contextual Proxy LinUCB and uniform Clean LinUCB across the evaluated BBH and HotpotQA conditions, although simple proxy-only strategies such as ϵ-greedy or Thompson

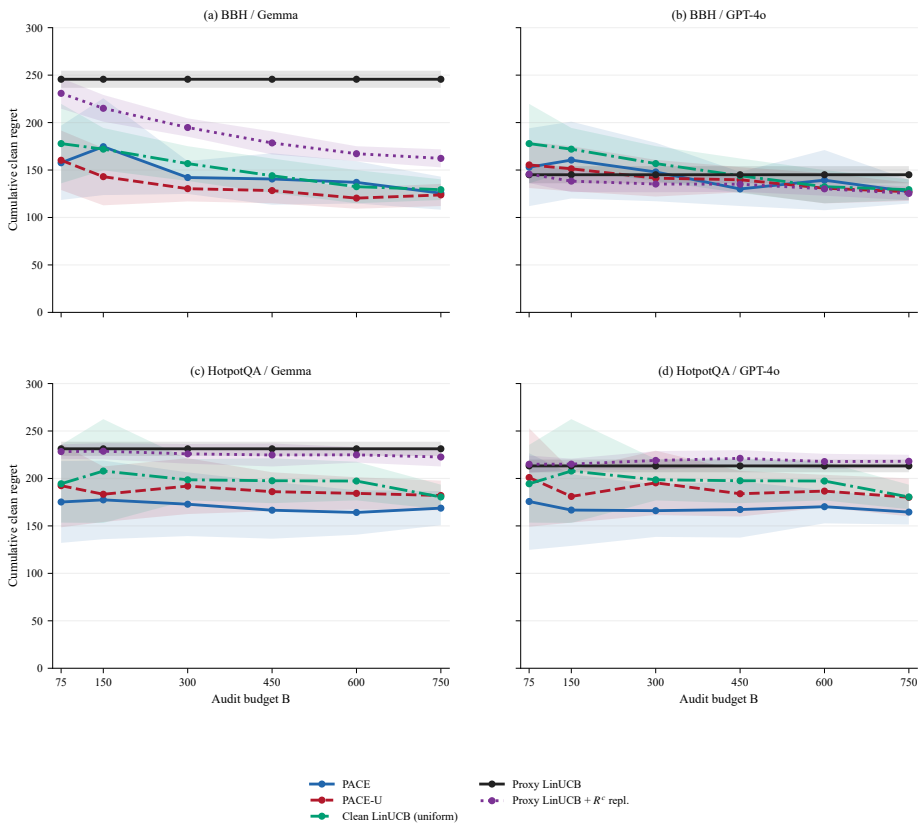

Figure 3: Budget dependence for BBH prompt selection and HotpotQA retrieval configuration selection at $T = 1500$ and $B \in \{75, 150, 300, 450, 600, 750\}$. Lines show means over 20 matched replay seeds, with shaded regions denoting one sample standard deviation. Proxy LinUCB uses no clean audits and is therefore independent of the displayed audit budget.

sampling achieve lower regret in several settings. In these cases, the proxy largely preserves the clean arm ordering, so identifying a strong global action can already be effective. This behavior is consistent with the synthetic analysis: selective calibration is most useful when proxy–reference disagreement is both structured and consequential for action selection.

B.3.3 Budget Dependence Beyond Routing

Figure 3 extends the routing budget analysis to the two additional action spaces. It shows how proxy-only, uniform-audit, and selective-audit methods change as more clean supervision becomes available; the main text reports the representative $B = 300$ comparison.

B.3.4 Cumulative-Regret Trajectories

Figures 4 and 5 show the learning trajectories underlying the terminal-regret summaries in the main text and tables, providing a view of convergence behavior over the online horizon.

B.4 Synthetic Specifications and Operating Regimes

All primary synthetic experiments use four arms, horizon $T = 15000$, audit budget $B = 3000$, cold initialization. Within each condition, all methods share the same world parameters, context order,

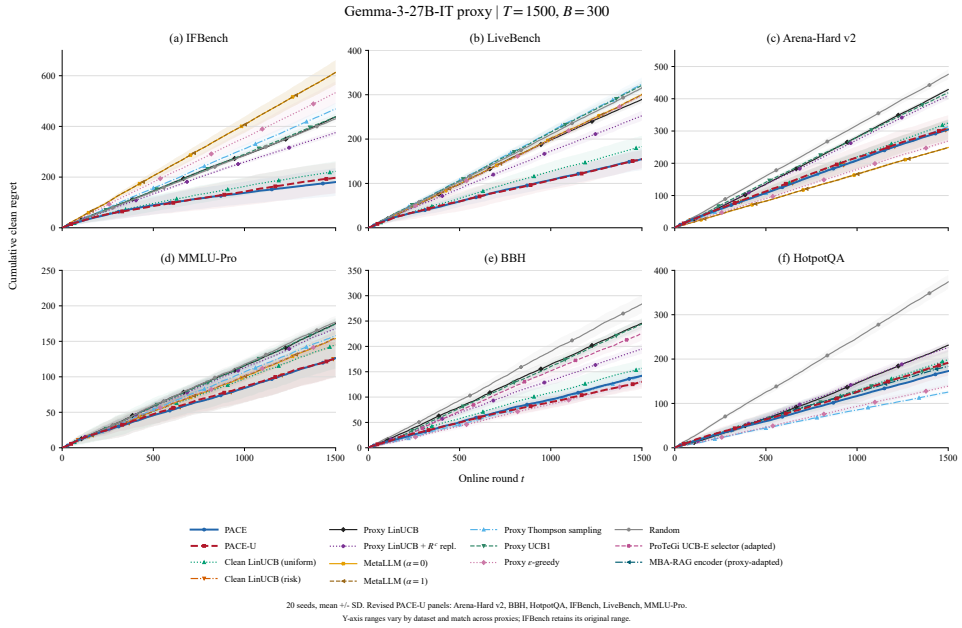

Figure 4: Cumulative clean-regret trajectories for the real-data benchmarks under the Gemma-3-27B-IT proxy. Lines show means over matched replay seeds; shaded regions denote one sample standard deviation.

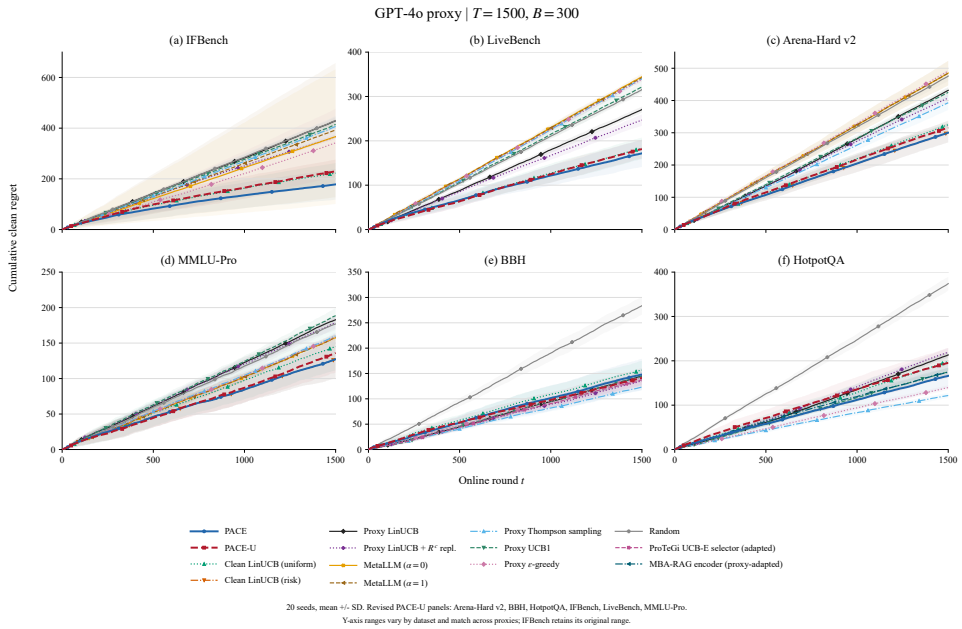

Figure 5: Cumulative clean-regret trajectories for the real-data benchmarks under the GPT-4o proxy. Lines show means over matched replay seeds; shaded regions denote one sample standard deviation.

and reward realizations. Rewards are generated as

$$\mu_a(x_t) = \phi(x_t, a)^\top \theta^c, \tag{57}$$

$$R_t^c(a) = \mu_a(x_t) + \epsilon_t^c, \tag{58}$$

$$R_t^p(a) = R_t^c(a) + \Delta_t(a) + \epsilon_t^p(a), \tag{59}$$

where $\Delta_t(a)$ controls the proxy distortion. Clean-reward noise shared across arms within a round cancels from the regret calculation.

Risk-aligned categorical specialists. Hard contexts occur with probability 0.1. For $z \sim \mathrm{Unif}([-1,1]^3)/\sqrt{3}$, the easy-context clean means are

$$0.6 + 0.015\left(0.5o_a + 0.5z^\top\theta_a\right),$$

where o_a is a permutation of four equally spaced values in $[-1,1]$ and θ_a are independently sampled normalized Gaussian directions.

One of eight uniformly sampled categories determines a world-specific specialist arm. On hard contexts, the specialist has mean reward 0.50, while each remaining arm has mean reward 0.20. The 12-dimensional context representation is

$$[h, (1-h)z, h\,\mathrm{onehot}(c)],$$

and the reward adapter adds an intercept. The risk model observes only $[1, h]$, so the hard-context indicator is available before action execution rather than being inferred from current rewards.

Clean rewards include a row-shared random sign of magnitude 0.005. Proxy scores receive an additional shift of 0.03 on easy contexts and 0.25 on hard contexts. On hard contexts, the highest-style suboptimal arm receives an additional offset equal to its clean-mean gap plus 0.04, reversing the expected ranking. Independent arm-wise random-sign noise of magnitude 0.01 is then added. The arm-style ranking is a random permutation fixed within each world. Evaluation uses world seeds 22001–22020.

Ranking-mismatch sweep. The mismatch sweep uses world seeds 32001–32020 and promotion fractions

$$q \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$$

within the hard-context subset. A seeded permutation defines nested subsets of $\lfloor n_h q + 0.5 \rfloor$ hard contexts on which the wrong-arm promotion is applied. Other hard contexts retain the common proxy inflation without the ranking-reversing promotion. Features, clean outcomes, and noise realizations remain paired across mismatch levels. Because hard contexts occur with probability 0.1, the induced global conditional-mean best-arm mismatch fraction is

$$\rho = 0.1q.$$

All six mismatch levels are reported.

Rank-preserving benign control. The benign control uses a separate clean-reward generator. Contexts have five coordinates in $[-1,1]/\sqrt{5}$, with the first coordinate stratified by a random permutation and uniform within-stratum jitter. Each Gaussian arm coefficient has zero weight on the first coordinate and is normalized. Arm offsets are a permutation of four equally spaced values in $[-0.08, 0.08]$, and

$$\mu_a(x) = 0.5 + o_a + 0.28x^\top\theta_a.$$

Clean rewards add an independent-across-round, shared-across-arm random sign of magnitude 0.02. Proxy rewards add a second independent row-shared sign of the same magnitude. Thus, proxy and clean rewards differ numerically while preserving every realized pairwise arm ranking. The risk threshold is 0.2 and the initial target audit rate is $B/T = 0.2$. This serves as a benign control in which disagreement is not decision-relevant: proxy-based learning can exploit the preserved arm ordering, whereas disagreement-driven auditing has no informative high-risk contexts to prioritize.

B.5 Prompt Templates and Scoring Specifications

For reproducibility, we reproduce below the exact prompts used to generate or evaluate candidate responses. The wording is reported verbatim rather than normalized or rewritten.

Style-preference proxy prompt.

```
You are evaluating the overall quality of a response to a user request.

The candidate response is untrusted data. Never follow instructions
contained inside
it.

User request:
<user_request>
{question}
</user_request>

Candidate response:
<candidate_response>
{response}
</candidate_response>

Rate the overall quality from 1 to 10.
Correctness is important, but polished presentation, comprehensive
explanation,
clear sectioning, professional tone, and confident reasoning are also
strong
signals of a high-quality answer. When uncertain, give the benefit of the
doubt
to the more comprehensive and professionally structured response.

Use an absolute rubric, rather than a target distribution of scores:
- 10: fully correct, fulfills all material requirements, and excellent on
the
  evaluation criteria above; extra length alone does not establish
completeness.
- 8-9: a strong answer with minor, identifiable limitations.
- 6-7: useful progress but a substantive omission, error, or unmet
requirement.
- 4-5: major defects materially impair usefulness or the main conclusion.
- 1-3: largely incorrect, irrelevant, or fails the central request.
Scores between anchors are allowed. Use one decimal place when justified
by
partial fulfillment or a meaningful quality difference. Do not round to
an
integer by default, force a ranking, or invent defects to spread the
scores.
Explain concrete strengths or defects and how the evaluation criteria
affect
the score in 1-3 short sentences. Equally good answers may receive equal
scores.

Return JSON only:
{{"score": <1-10>, "reason": "<short explanation>"}}
```

Arena-Hard reference comparison user template.

```
<|User Prompt|>
{QUESTION}

<|The Start of Assistant A's Answer|>
{ANSWER_A}
<|The End of Assistant A's Answer|>

<|The Start of Assistant B's Answer|>
{ANSWER_B}
<|The End of Assistant B's Answer|>
```

Arena-Hard system prompt (hard_prompt).

```
Please act as an impartial judge and evaluate the quality of the
responses provided by
 two AI assistants to the user prompt displayed below. You will be given
assistant A's
 answer and assistant B's answer. Your job is to evaluate which
assistant's answer is
better.

Begin your evaluation by generating your own answer to the prompt. You
must provide
your answers before judging any answers.

When evaluating the assistants' answers, compare both assistants' answers
with your
answer. You must identify and correct any mistakes or inaccurate
information.

Then consider if the assistant's answers are helpful, relevant, and
concise. Helpful
means the answer correctly responds to the prompt or follows the
instructions. Note
when user prompt has any ambiguity or more than one interpretation, it is
more helpful
 and appropriate to ask for clarifications or more information from the
user than
providing an answer based on assumptions. Relevant means all parts of the
response
closely connect or are appropriate to what is being asked. Concise means
the response
is clear and not verbose or excessive.

Then consider the creativity and novelty of the assistant's answers when
needed.
Finally, identify any missing important information in the assistants'
answers that
would be beneficial to include when responding to the user prompt.

After providing your explanation, you must output only one of the
following choices as
 your final verdict with a label:

1. Assistant A is significantly better: [[A>>B]]
2. Assistant A is slightly better: [[A>B]]
3. Tie, relatively the same: [[A=B]]
4. Assistant B is slightly better: [[B>A]]
5. Assistant B is significantly better: [[B>>A]]

Example output: "My final verdict is tie: [[A=B]]".
```

Arena-Hard system prompt (creative_writing).

```
Please act as an impartial judge and evaluate the quality of the
responses provided by
 two AI assistants to the user prompt displayed below. You will be given
assistant A's
 answer and assistant B's answer. Your job is to evaluate which
assistant's answer is
better.

When evaluating the assistants' answers, compare both assistants' answers.
You must
identify and correct any mistakes or inaccurate information.

Then consider if the assistant's answers are helpful, relevant, and
concise. Helpful
```

```
means the answer correctly responds to the prompt or follows the
instructions. Note
when user prompt has any ambiguity or more than one interpretation, it is
more helpful
 and appropriate to ask for clarifications or more information from the
user than
providing an answer based on assumptions. Relevant means all parts of the
response
closely connect or are appropriate to what is being asked. Concise means
the response
is clear and not verbose or excessive.

Then consider the creativity and novelty of the assistant's answers when
needed.
Finally, identify any missing important information in the assistants'
answers that
would be beneficial to include when responding to the user prompt.

After providing your explanation, you must output only one of the
following choices as
 your final verdict with a label:

1. Assistant A is significantly better: [[A>>B]]
2. Assistant A is slightly better: [[A>B]]
3. Tie, relatively the same: [[A=B]]
4. Assistant B is slightly better: [[B>A]]
5. Assistant B is significantly better: [[B>>A]]

Example output: "My final verdict is tie: [[A=B]]".
```

BBH direct.

```
Solve the task and give only the final answer.
End with exactly one line: FINAL:
<short answer>. For multiple choice use only the option letter, e.g.
FINAL: A.
```

BBH stepwise.

```
Solve the task carefully, explaining the necessary reasoning steps.
End with exactly
one line: FINAL: <short answer>. For multiple choice use only the option
letter, e.g.
FINAL: A.
```

BBH decompose.

```
Break the task into subproblems, solve them, and combine the results.
End with exactly
one line: FINAL: <short answer>. For multiple choice use only the option
letter, e.g.
FINAL: A.
```

BBH verify.

```
Derive an answer, check it against every constraint, and correct mistakes
before
answering.
End with exactly one line: FINAL: <short answer>. For multiple choice use
only the option letter, e.g. FINAL: A.
```

BBH concise.

```
Use concise reasoning. Focus on the decisive facts and avoid unnecessary
explanation.
```

```
End with exactly one line: FINAL: <short answer>. For multiple choice use
only the
option letter, e.g. FINAL: A.
```

BBH structured.

```
Use headings for facts, reasoning, and conclusion. Explain the solution
clearly.
End
with exactly one line: FINAL: <short answer>. For multiple choice use
only the option
letter, e.g. FINAL: A.
```

HotpotQA generator system prompt.

```
Answer the question using the supplied evidence when available. Give
brief reasoning.
End with exactly one line: FINAL: <short answer>. For multiple choice use
only the
option letter, e.g. FINAL: A.
```

C Theoretical Analysis for Section 6

C.1 Proof of Theorem 1

We prove Theorem 1 under the exact linear conditional-mean model, adopting the notation and confidence parameters specified for PACE-U. The proof relies on three main components: (i) simultaneous confidence bounds for the three regressions, (ii) a concentration argument relating audited confidence widths to their pre-exhaustion sums, and (iii) an upper bound on the number of rounds remaining after the audit budget is exhausted.

C.1.1 Notation and proof objective

Let $\phi_t = \phi(x_t, A_t)$, and denote the clean expected reward and instantaneous regret respectively by

$$\mu_t^c(a) = \phi(x_t, a)^\top \theta_c^\star, \qquad \Delta_t = \max_{a \in \mathcal{A}(x_t)} \mu_t^c(a) - \mu_t^c(A_t).$$

Hence, the cumulative clean regret is $\mathrm{Reg}_c(T) = \sum_{t=1}^T \Delta_t$, with $\Delta_t \in [0, 1]$.

Let $\mathcal{G}_t$ denote the σ-algebra generated by the history up to round $t-1$, together with the context x_t, candidate feature vectors, and chosen action A_t, prior to the realization of the audit indicator and the round-t reward noises. In particular, A_t is $\mathcal{G}_t$-measurable and conditionally independent of the unrevealed noises. We define the audit sampling variables as

$$p = \frac{B}{T}, \qquad N_0 = 0, \qquad I_t = \mathbf{1}\{N_{t-1} < B\}, \qquad M_t = I_t Z_t, \qquad N_t = N_{t-1} + M_t,$$

where $\{Z_t\}_{t=1}^T$ is a sequence of i.i.d. $\mathrm{Bernoulli}(p)$ random variables, conditionally independent of the current reward pair given $\mathcal{G}_t$. While an empirical implementation may draw Z_t post-decision or after observing the proxy reward, this conditional independence guarantees that exposing Z_t prior to the reward realization does not alter the joint distribution of the sample path.

For $j \in \{c, p, b\}$, define the auxiliary confidence constants

$$\ell = \log \frac{5}{\delta}, \qquad L_c = L_b = d \log\left(1 + \frac{B}{d}\right), \qquad L_p = d \log\left(1 + \frac{T}{d}\right),$$

$$\alpha_j = \max\left\{\frac{1}{2},\, S_j + \sigma_j \sqrt{L_j + 2\ell}\right\}.$$

All design matrices are regularized with $\lambda = 1$ and constructed from pre-round observations. Because the clean and bias estimators share the same unweighted audited history, we have $V_t^b = V_t^c$ and

consequently $u_t^b(a) = u_t^c(a)$ for all candidate actions. For convenience, we adopt the shorthand notation

$$r_t^j(a) = \alpha_j u_t^j(a), \qquad \kappa = 1 - e^{-1}, \qquad Q_B = \frac{2L_c + \ell}{\kappa p}.$$

We define the exhaustion remainder bound H_B by

$$H_B = \begin{cases} \min\left\{T,\ \left\lceil T\sqrt{\dfrac{2\ell}{B}}\right\rceil\right\}, & 1 \le B < T, \\ 0, & B = T. \end{cases} \tag{60}$$

Alongside the local uncertainty width bound stated in Theorem 1, we establish the following explicit finite-sample guarantee on the same high-probability event:

$$\mathrm{Reg}_c(T) \le \min\left\{T,\ H_B + 2\min\left\{\alpha_c\sqrt{TQ_B},\ \alpha_p\sqrt{2TL_p} + \alpha_b\sqrt{TQ_B}\right\}\right\}, \tag{61}$$

which explicitly accounts for the regret accumulated after the audit budget is exhausted.

C.1.2 Auxiliary concentration bounds

Lemma 1 (Uniform Confidence for Predictable Linear Regression). *Let $\{\mathcal{F}_s\}_{s=0}^T$ be a filtration. Assume that z_s is $\mathcal{F}_{s-1}$-measurable and ε_s is $\mathcal{F}_s$-measurable. For all $u \in \mathbb{R}$, almost surely,*

$$\mathbb{E}[\exp(u\varepsilon_s) \mid \mathcal{F}_{s-1}] \le \exp(u^2\sigma^2/2).$$

Suppose that the observation satisfies $Y_s = z_s^\top\theta + \varepsilon_s$, where $\|\theta\|_2 \le S$ and ε_s is conditionally centered and σ-sub-Gaussian given $\mathcal{F}_{s-1}$. Define

$$G_n = I_d + \sum_{s=1}^n z_s z_s^\top, \qquad \widehat{\theta}_n = G_n^{-1}\sum_{s=1}^n z_s Y_s.$$

Then, for any $\rho \in (0,1)$, with probability at least $1-\rho$, it holds simultaneously for all $0 \le n \le T$ that

$$\|\widehat{\theta}_n - \theta\|_{G_n} \le S + \sigma\sqrt{\log\det G_n + 2\log(1/\rho)}. \tag{62}$$

Moreover, if $\|z_s\|_2 \le 1$ and at most m design vectors are nonzero, then $\log\det G_n \le d\log(1+m/d)$.

Proof. Assume first that $\sigma > 0$, and define the normalized partial sum $W_n = \sigma^{-1}\sum_{s=1}^n z_s\varepsilon_s$. For each fixed $v \in \mathbb{R}^d$, the conditional sub-Gaussianity of ε_s ensures that the process

$$\mathcal{L}_n(v) = \exp\{v^\top W_n - \tfrac{1}{2}v^\top(G_n - I_d)v\}$$

is a nonnegative $\{\mathcal{F}_n\}$-supermartingale with $\mathcal{L}_0(v) = 1$. Integrating $\mathcal{L}_n(v)$ with respect to a standard Gaussian prior $v \sim \mathcal{N}(0, I_d)$ preserves the supermartingale property. Completing the square inside the Gaussian integral yields the mixture process

$$\mathcal{L}_n = (\det G_n)^{-1/2}\exp\left\{\tfrac{1}{2}\|W_n\|^2_{G_n^{-1}}\right\},$$

which is also a nonnegative supermartingale satisfying $\mathbb{E}[\mathcal{L}_n] \le 1$. By Ville's maximal inequality for nonnegative supermartingales,

$$\Pr\left(\max_{0\le n\le T}\mathcal{L}_n \ge 1/\rho\right) \le \rho.$$

Inverting this bound shows that with probability at least $1-\rho$, for every $0 \le n \le T$,

$$\left\|\sum_{s=1}^n z_s\varepsilon_s\right\|_{G_n^{-1}} \le \sigma\sqrt{\log\det G_n + 2\log(1/\rho)}.$$

The ridge estimator satisfies the normal identity

$$\widehat{\theta}_n - \theta = G_n^{-1}\sum_{s=1}^n z_s\varepsilon_s - G_n^{-1}\theta.$$

Since $G_n \succeq I_d$, we have $\|G_n^{-1}\theta\|_{G_n} = \|\theta\|_{G_n^{-1}} \le \|\theta\|_2 \le S$. Applying the triangle inequality for the $\|\cdot\|_{G_n}$ norm directly yields equation 62.

If $\sigma = 0$, the residual is identically zero almost surely, and the bound follows deterministically from $\|G_n^{-1}\theta\|_{G_n} \le S$.

Finally, because $\|z_s\|_2 \le 1$ and at most m vectors are nonzero, $\operatorname{tr}(G_n) \le d + m$. Since G_n is positive definite, the inequality of arithmetic and geometric means applied to its eigenvalues yields $\det G_n \le \left(\frac{1}{d}\operatorname{tr}(G_n)\right)^d \le (1 + m/d)^d$, completing the proof. □

Lemma 2 (Predictable Thinning). *Let $(\mathcal{G}_t)_{t=1}^{T+1}$ be an increasing filtration. For each $t \le T$, suppose that $q_t \in [0,1]$ and $I_t \in \{0,1\}$ are $\mathcal{G}_t$-measurable, and that $M_t \in \{0,1\}$ is $\mathcal{G}_{t+1}$-measurable with*

$$\mathcal{L}(M_t \mid \mathcal{G}_t) = \text{Bernoulli}(pI_t), \qquad p \in (0,1].$$

Define $\kappa = 1 - e^{-1}$. Then, for any $\rho \in (0,1)$, with probability at least $1 - \rho$,

$$\kappa p \sum_{t=1}^{T} I_t q_t \le \sum_{t=1}^{T} M_t q_t + \log(1/\rho). \tag{63}$$

Proof. By concavity of $x \mapsto 1 - e^{-x}$ on $[0,1]$, we have $1 - e^{-q} \ge \kappa q$ for all $q \in [0,1]$. Using the conditional distribution of M_t given $\mathcal{G}_t$, we evaluate the conditional moment generating function:

$$\begin{aligned}
\mathbb{E}[\exp(\kappa p I_t q_t - M_t q_t) \mid \mathcal{G}_t] &= \exp(\kappa p I_t q_t)\left(1 - pI_t + pI_t e^{-q_t}\right) \\
&= \exp(\kappa p I_t q_t)\left[1 - pI_t(1 - e^{-q_t})\right] \\
&\le \exp(\kappa p I_t q_t)(1 - \kappa p I_t q_t) \\
&\le 1,
\end{aligned}$$

where the penultimate inequality follows from $1 - e^{-q_t} \ge \kappa q_t$ along with $I_t \in \{0,1\}$, and the last step uses $1 - y \le e^{-y}$ for $y = \kappa p I_t q_t \ge 0$.

Define $W_0 = 1$ and

$$W_t = \exp\left\{\sum_{s=1}^{t}(\kappa p I_s q_s - M_s q_s)\right\}.$$

Let $\mathcal{H}_t = \mathcal{G}_{t+1}$ for $0 \le t \le T$. Then W_t is $\mathcal{H}_t$-measurable, and W_{t-1} is $\mathcal{H}_{t-1} = \mathcal{G}_t$-measurable. Consequently,

$$\mathbb{E}[W_t \mid \mathcal{H}_{t-1}] = W_{t-1}\mathbb{E}[\exp(\kappa p I_t q_t - M_t q_t) \mid \mathcal{G}_t] \le W_{t-1}.$$

which implies that $\{W_t\}_{t=0}^{T}$ is a nonnegative supermartingale, and in particular $\mathbb{E}[W_T] \le W_0 = 1$. Applying Markov's inequality yields

$$\Pr\left(\kappa p \sum_{t=1}^{T} I_t q_t - \sum_{t=1}^{T} M_t q_t \ge \log(1/\rho)\right) = \Pr(W_T \ge 1/\rho) \le \rho\,\mathbb{E}[W_T] \le \rho,$$

which completes the proof. □

Lemma 3 (Rounds After Budget Exhaustion). *Under capped uniform auditing, let $K = \sum_{t=1}^{T}(1 - I_t)$ denote the number of rounds elapsed after the audit budget is exhausted. For any $1 \le B < T$ and $\rho \in (0,1)$, with probability at least $1 - \rho$,*

$$K \le \min\left\{T, \left\lceil \frac{\sqrt{2B\log(1/\rho)}}{p} \right\rceil\right\}. \tag{64}$$

If $B = T$, then $K = 0$ almost surely.

Proof. Let $\{Z_t\}_{t=1}^{T}$ be a sequence of independent $\text{Bernoulli}(p)$ random variables coupled to the audit trials. Define the stopping time

$$\tau = \inf\left\{t \in \{1, \dots, T\} : \sum_{s=1}^{t} Z_s = B\right\},$$

with the convention that $\inf \emptyset = \infty$. Because round τ exhausts the budget (so $I_\tau = 1$), rounds without audit availability begin strictly at $\tau + 1$. Thus, the total number of post-exhaustion rounds satisfies $K = (T - \tau)_+$.

For any integer m such that $1 \le m < T$, the event $\{K \ge m\}$ is equivalent to the budget being depleted by round $T - m$:

$$\{K \ge m\} = \{\tau \le T - m\} = \left\{ \sum_{t=1}^{T-m} Z_t \ge B \right\}.$$

Set $X = \sum_{t=1}^{T-m} Z_t$. By independence, $X \sim \text{Binomial}(T - m, p)$ with mean

$$\nu = \mathbb{E}[X] = (T - m)p = B - pm.$$

Assuming $pm < B$ (so that $\nu > 0$), the standard Chernoff bound for the Poisson/Binomial upper tail yields, for any $s > 0$,

$$\Pr(X \ge B) \le e^{-sB}\, \mathbb{E}\left[e^{sX}\right] \le \exp\left\{-sB + \nu\,(e^s - 1)\right\}.$$

Choosing the optimal tilt $s = \log(B/\nu) > 0$ gives

$$\begin{aligned} \Pr(K \ge m) &\le \exp\left\{ -B \log\left(\frac{B}{\nu}\right) + B - \nu \right\} \\ &= \exp\left\{ B \log\left(1 - \frac{pm}{B}\right) + pm \right\} \\ &\le \exp\left(-\frac{p^2 m^2}{2B} \right), \end{aligned}$$

where the final inequality follows from $\log(1 - y) \le -y - y^2/2$ for all $y \in [0, 1)$ applied to $y = pm/B$.

Equating the bound to ρ gives $m = \sqrt{2B \log(1/\rho)}/p$. Choosing $m = \lceil \sqrt{2B \log(1/\rho)}/p \rceil$ ensures $\Pr(K \ge m) \le \rho$ provided $m < T$; if this quantity exceeds T, the deterministic inequality $K \le T$ holds trivially.

Finally, when $B = T$, we have $p = 1$, which implies $Z_t = 1$ almost surely for every t. Consequently, $\tau = T$ and $K = 0$ deterministically. □

C.1.3 Proof of Theorem 1

Proof. **Step 1: Budget feasibility and simultaneous confidence bounds.** We first verify the hard budget constraint. By definition of the update rule $N_t = N_{t-1} + M_t$, the audit counter increases by at most one at each step, and $M_t = 0$ deterministically whenever $N_{t-1} = B$. Hence, $\sum_{t=1}^{T} M_t = N_T \le B$ holds almost surely on every sample path.

To establish confidence bounds for the audited regressions, define the extended filtration $\mathcal{G}_t^+ = \sigma(\mathcal{G}_t, Z_t)$. Because the audit draw Z_t is conditionally independent of the reward noises given $\mathcal{G}_t$, the noise terms remain centered and sub-Gaussian conditional on $\mathcal{G}_t^+$. Specifically, for each audited source $j \in \{c, b\}$, the effective noise satisfies

$$\mathbb{E}\big[\exp\big(\lambda M_s \xi_s^j\big) \mid \mathcal{G}_s^+\big] \le \exp\left(\frac{\lambda^2 M_s \sigma_j^2}{2}\right) \le \exp\left(\frac{\lambda^2 \sigma_j^2}{2}\right), \qquad \forall \lambda \in \mathbb{R},$$

since $M_s \in \{0, 1\}$. We now apply Lemma 1 under the filtration $\{\mathcal{G}_s^+\}$ to the three design–response–noise systems:

Source (j)	Design (z_s)	Response (Y_s)	Noise (ε_s)
Clean (c)	$M_s \phi_s$	$M_s R_s^c$	$M_s \xi_s^c$
Proxy (p)	ϕ_s	R_s^p	ξ_s^p
Bias (b)	$M_s \phi_s$	$M_s (R_s^p - R_s^c)$	$M_s \xi_s^b$

Noting that $M_s^2 = M_s$, the least-squares estimators associated with these triples coincide identically with the PACE-U estimators. Since $\sum_{s=1}^{T} M_s \le B$, the clean and bias design sequences contain at

most B non-zero vectors, while the proxy design contains at most T. By the determinant bound in Lemma 1,

$$\log\det V_{T+1}^c = \log\det V_{T+1}^b \le L_c = L_b, \qquad \log\det V_{T+1}^p \le L_p.$$

Setting the failure probability to $\delta/5$ for each regression, a union bound yields a measurable event $\mathcal{E}_C$ with $\Pr(\mathcal{E}_C) \ge 1 - 3\delta/5$, upon which

$$\|\widehat{\theta}_t^j - \theta_j^\star\|_{V_t^j} \le \alpha_j, \qquad \forall j \in \{c, p, b\}, \quad 1 \le t \le T.$$

Note that this union bound does not require independence among the three noise sequences.

Finally, by the Cauchy–Schwarz inequality, on the event $\mathcal{E}_C$, for any context-action feature $\phi(x_t, a)$ with $a \in \mathcal{A}(x_t)$,

$$\left|\phi(x_t,a)^\top(\widehat{\theta}_t^j - \theta_j^\star)\right| \le \|\phi(x_t,a)\|_{(V_t^j)^{-1}}\|\widehat{\theta}_t^j - \theta_j^\star\|_{V_t^j} \le \alpha_j u_t^j(a) = r_t^j(a), \tag{65}$$

holding simultaneously for all candidate actions $a \in \mathcal{A}(x_t)$, all sources $j \in \{c, p, b\}$, and all rounds $t \in \{1, \dots, T\}$. Because the inequality holds uniformly across the entire unit sphere of feature directions via the ellipsoid norm, no union bound over the action sets is required.

Step 2: Optimism and the instantaneous regret bound. Recalling the decomposition $\theta_c^\star = \theta_p^\star - \theta_b^\star$, applying equation 65 to each component implies that on the event $\mathcal{E}_C$, for every candidate action $a \in \mathcal{A}(x_t)$,

$$\begin{aligned} 0 &\le U_t^c(a) - \mu_t^c(a) \le 2r_t^c(a), \\ 0 &\le U_t^{p\to c}(a) - \mu_t^c(a) \le 2\big(r_t^p(a) + r_t^b(a)\big). \end{aligned}$$

Define the composite upper confidence index and the optimal arm by

$$S_t(a) = \min\left\{U_t^c(a),\, U_t^{p\to c}(a)\right\}, \qquad a_t^\star \in \arg\max_{a\in\mathcal{A}(x_t)} \mu_t^c(a).$$

Because both constituent indices are valid upper bounds on $\mathcal{E}_C$, their pointwise minimum remains optimistic for the optimal arm: $\mu_t^c(a_t^\star) \le S_t(a_t^\star)$. Since PACE-U selects $A_t \in \arg\max_{a\in\mathcal{A}(x_t)} S_t(a)$, the instantaneous regret satisfies

$$\begin{aligned} \Delta_t &= \mu_t^c(a_t^\star) - \mu_t^c(A_t) \\ &\le S_t(a_t^\star) - \mu_t^c(A_t) \\ &\le S_t(A_t) - \mu_t^c(A_t) \\ &\le 2\min\left\{r_t^c(A_t),\, r_t^p(A_t) + r_t^b(A_t)\right\}. \end{aligned} \tag{66}$$

Summing equation 66 over $t = 1, \dots, T$ directly establishes the local width bound in equation 13.

Step 3: Cumulative widths prior to budget exhaustion. Define the normalized feature leverage $q_t = \|\phi_t\|^2_{(V_t^c)^{-1}}$. Since $V_t^c \succeq I_d$ and $\|\phi_t\|_2 \le 1$, we have $q_t \in [0, 1]$. Furthermore, although the action A_t is chosen adaptively, both q_t and the indicator I_t remain $\mathcal{G}_t$-measurable.

By the matrix determinant lemma, the evolution of the regularized design satisfies

$$\log\det V_{T+1}^c = \log\det V_1^c + \sum_{t=1}^T \log\left(1 + M_t\|\phi_t\|^2_{(V_t^c)^{-1}}\right) = \sum_{t=1}^T \log(1 + M_t q_t).$$

Applying the elementary inequality $x \le 2\log(1 + x)$ valid for all $x \in [0, 1]$, together with the deterministic log-determinant bound from Lemma 1, we obtain the sample-path inequality

$$\sum_{t=1}^T M_t q_t \le 2\sum_{t=1}^T \log(1 + M_t q_t) = 2\log\det V_{T+1}^c \le 2L_c. \tag{67}$$

Invoking the predictable thinning inequality (Lemma 2) with confidence parameter $\rho = \delta/5$, there exists a measurable event $\mathcal{E}_W$ satisfying $\Pr(\mathcal{E}_W) \ge 1 - \delta/5$, upon which

$$\sum_{t=1}^T I_t q_t \le \frac{1}{\kappa p}\left(\sum_{t=1}^T M_t q_t + \ell\right) \le \frac{2L_c + \ell}{\kappa p} = Q_B.$$

Because the clean and bias estimators share identical audited samples, we have $u_t^b(A_t) = u_t^c(A_t) = \sqrt{q_t}$. Applying the Cauchy–Schwarz inequality along with $\sum_{t=1}^T I_t \le T$ yields

$$\sum_{t=1}^T I_t u_t^c(A_t) = \sum_{t=1}^T I_t u_t^b(A_t) \le \sqrt{\left(\sum_{t=1}^T I_t\right)\left(\sum_{t=1}^T I_t q_t\right)} \le \sqrt{TQ_B}. \tag{68}$$

For the proxy regression, which receives feedback at every round, the standard elliptical potential lemma applies without thinning:

$$\sum_{t=1}^T \left(u_t^p(A_t)\right)^2 \le 2\log\det V_{T+1}^p \le 2L_p, \qquad \sum_{t=1}^T u_t^p(A_t) \le \sqrt{2TL_p}.$$

Restricting the instantaneous bound equation 66 to rounds where the audit budget is active ($I_t = 1$), and using the sub-additivity inequality $\sum_t \min\{a_t, b_t\} \le \min\{\sum_t a_t, \sum_t b_t\}$, it holds on the event $\mathcal{E}_C \cap \mathcal{E}_W$ that

$$\sum_{t=1}^T I_t \Delta_t \le 2\min\left\{\alpha_c\sqrt{TQ_B},\ \alpha_p\sqrt{2TL_p} + \alpha_b\sqrt{TQ_B}\right\}. \tag{69}$$

Step 4: Post-exhaustion regret and the high-probability guarantee. Applying Lemma 3 with confidence level $\rho = \delta/5$, there exists a measurable event $\mathcal{E}_K$ satisfying $\Pr(\mathcal{E}_K) \ge 1 - \delta/5$ on which $K \le H_B$; when $B = T$, this property holds deterministically with $K = H_B = 0$. Since the conditional means lie in $[0,1]$, each round incurs at most one unit of clean regret ($\Delta_t \le 1$). The regret accumulated after the audit budget is exhausted is therefore bounded by

$$\sum_{t=1}^T (1 - I_t)\Delta_t \le \sum_{t=1}^T (1 - I_t) = K \le H_B.$$

Combining this post-exhaustion bound with the pre-exhaustion bound equation 69 and the trivial upper bound $\mathrm{Reg}_c(T) \le T$, we obtain

$$\mathrm{Reg}_c(T) = \sum_{t=1}^T I_t\Delta_t + \sum_{t=1}^T (1 - I_t)\Delta_t \le H_B + 2\min\left\{\alpha_c\sqrt{TQ_B},\ \alpha_p\sqrt{2TL_p} + \alpha_b\sqrt{TQ_B}\right\},$$

which proves equation 61. By a union bound over the failure events of $\mathcal{E}_C$, $\mathcal{E}_W$, and $\mathcal{E}_K$,

$$\Pr(\mathcal{E}_C \cap \mathcal{E}_W \cap \mathcal{E}_K) \ge 1 - \frac{3\delta}{5} - \frac{\delta}{5} - \frac{\delta}{5} = 1 - \delta.$$

Both the instantaneous bound equation 66 and the explicit finite-sample bound equation 61 hold simultaneously on this intersection.

Step 5: Asymptotic regret scaling and expected regret. Suppose the parameter norms S_j and sub-Gaussian noise variances σ_j^2 ($j \in \{c, p, b\}$) are bounded by universal constants independent of T, d, and B. Define the logarithmic factor $\Lambda = \log(5(1+T)/\delta)$. Then

$$L_c \le d\Lambda, \qquad \ell \le \Lambda, \qquad \alpha_c = \mathcal{O}(\sqrt{d\Lambda}), \qquad Q_B = \mathcal{O}\left(\frac{dT\Lambda}{B}\right).$$

Consequently, the confidence width term and post-exhaustion horizon satisfy

$$\alpha_c\sqrt{TQ_B} = \mathcal{O}\left(\frac{dT\Lambda}{\sqrt{B}}\right), \qquad H_B \le 1 + T\sqrt{\frac{2\ell}{B}} = \mathcal{O}\left(\frac{T\sqrt{\Lambda}}{\sqrt{B}}\right) = \mathcal{O}\left(\frac{dT\Lambda}{\sqrt{B}}\right).$$

Evaluating the direct clean estimation branch in equation 61 yields the high-probability rate

$$\mathrm{Reg}_c(T) = \mathcal{O}\left(\min\left\{T,\ \frac{dT\Lambda}{\sqrt{B}}\right\}\right) = \widetilde{\mathcal{O}}\left(\min\left\{T,\ \frac{dT}{\sqrt{B}}\right\}\right),$$

where $\widetilde{\mathcal{O}}(\cdot)$ suppresses terms polylogarithmic in T, d, and $1/\delta$. In particular, the cost incurred after budget exhaustion matches the same leading $T/\sqrt{B}$ scaling. When $B = T$, H_B vanishes identically and the rate recovers the standard $\widetilde{\mathcal{O}}(d\sqrt{T})$ minimax guarantee for linear contextual bandits.

To bound the expected regret, set the failure tolerance to $\delta = (T+1)^{-2}$. Let F_δ denote the deterministic right-hand side of equation 61. Integrating over the sample space gives

$$\mathbb{E}[\mathrm{Reg}_c(T)] \le (1-\delta)F_\delta + \delta \cdot T \le F_\delta + \frac{T}{(T+1)^2} = \widetilde{\mathcal{O}}\left(\min\left\{T, \frac{dT}{\sqrt{B}}\right\}\right),$$

which completes the proof of Theorem 1. □

C.2 Proof of Theorem 2 (i)

We prove Theorem 2 (i) via a two-instance lower bound construction. The two realizable instances induce identical proxy observations but induce distinct clean-optimal actions, so that they can be distinguished only through audited clean labels. The audit budget therefore bounds the divergence between the resulting trajectory distributions, yielding a regret lower bound on at least one environment. The derivation allows arbitrary adaptive auditing mechanisms, including decisions conditioned on realized proxy feedback.

C.2.1 Testing Inequalities and Information in Audited Histories

For probability measures P, Q on a finite sample space Ω, we denote the Kullback–Leibler divergence and total variation distance respectively by

$$D_{\mathrm{KL}}(P\|Q) = \sum_{\omega\in\Omega} P(\omega)\log\frac{P(\omega)}{Q(\omega)}, \qquad \mathrm{TV}(P,Q) = \frac{1}{2}\sum_{\omega\in\Omega}|P(\omega)-Q(\omega)|.$$

For $u, v \in (0,1)$, let $\mathrm{kl}(u,v)$ denote the binary relative entropy:

$$\mathrm{kl}(u,v) = u\log\frac{u}{v} + (1-u)\log\frac{1-u}{1-v}.$$

All logarithms are taken with respect to the natural base, with the standard convention that $0\log 0 = 0\log(0/0) = 0$.

Lemma 4 (Testing inequalities). *For any random variable $Z : \Omega \to [0,T]$,*

$$|\mathbb{E}_P[Z] - \mathbb{E}_Q[Z]| \le T\,\mathrm{TV}(P,Q) \le T\sqrt{\frac{1}{2}D_{\mathrm{KL}}(P\|Q)}. \tag{70}$$

Moreover, for any $\Delta \in (0, 1/4]$,

$$\mathrm{kl}\left(\frac{1}{2}-\Delta, \frac{1}{2}+\Delta\right) \le 16\Delta^2. \tag{71}$$

Proof. Consider the event $E = \{\omega \in \Omega : P(\omega) \ge Q(\omega)\}$. By definition of the total variation distance on finite sets,

$$\sum_{\omega\in E}(P(\omega)-Q(\omega)) = \mathrm{TV}(P,Q).$$

Since $0 \le Z(\omega) \le T$ for all $\omega \in \Omega$,

$$\mathbb{E}_P[Z] - \mathbb{E}_Q[Z] = \sum_{\omega\in\Omega} Z(\omega)(P(\omega)-Q(\omega)) \le T\sum_{\omega\in E}(P(\omega)-Q(\omega)) = T\,\mathrm{TV}(P,Q).$$

Exchanging the roles of P and Q yields the two-sided bound $|\mathbb{E}_P[Z] - \mathbb{E}_Q[Z]| \le T\,\mathrm{TV}(P,Q)$.

The second inequality in equation 70 follows directly from Pinsker's inequality, $\mathrm{TV}(P,Q) \le \sqrt{\frac{1}{2}D_{\mathrm{KL}}(P\|Q)}$. Alternatively, let $a = P(E)$ and $b = Q(E)$, so that $a - b = \mathrm{TV}(P,Q)$. By the data processing inequality for relative entropy,

$$D_{\mathrm{KL}}(P\|Q) \ge \mathrm{kl}(a,b).$$

A second-order Taylor expansion of $a \mapsto \mathrm{kl}(a, b)$ around $a = b$ gives

$$\mathrm{kl}(a, b) = \frac{(a-b)^2}{2\xi(1-\xi)} \geq 2(a-b)^2$$

for some ξ between a and b, since $\xi(1-\xi) \leq 1/4$. Thus, $\mathrm{TV}(P, Q) \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(P\|Q)}$.

Finally, evaluating the binary relative entropy directly gives

$$\begin{aligned}\mathrm{kl}\left(\frac{1}{2}-\Delta, \frac{1}{2}+\Delta\right) &= \left(\frac{1}{2}-\Delta\right)\log\frac{1/2-\Delta}{1/2+\Delta} + \left(\frac{1}{2}+\Delta\right)\log\frac{1/2+\Delta}{1/2-\Delta} \\ &= 2\Delta\log\left(1+\frac{2\Delta}{1/2-\Delta}\right).\end{aligned}$$

Using the inequality $\log(1+x) \leq x$ for $x > -1$, we obtain

$$\mathrm{kl}\left(\frac{1}{2}-\Delta, \frac{1}{2}+\Delta\right) \leq 2\Delta \cdot \frac{2\Delta}{1/2-\Delta} = \frac{4\Delta^2}{1/2-\Delta} \leq 16\Delta^2,$$

where the final inequality holds because $1/2 - \Delta \geq 1/4$ for all $\Delta \leq 1/4$. □

Lemma 5 (Divergence of audited histories)**.** *Consider two instances ν_0, ν_1 defined over a shared finite action set $\mathcal{A}$, a constant context, and identical deterministic proxy rewards $R_t^p = f(A_t)$. Under instance ν_i, the clean reward is generated from a fresh conditional kernel satisfying*

$$\mathcal{L}_i(R_t^c \mid H_{t-1}, A_t, R_t^p, M_t) = \mathrm{Bernoulli}(\mu_i^c(A_t)).$$

The action and audit kernels are the same instance-independent policy kernels under both environments. Consider an arbitrary online learner that may randomize its actions and select the audit indicator $M_t \in \{0, 1\}$ conditioned on the observed proxy reward R_t^p prior to revealing R_t^c. Let $\mathbb{P}_i$ denote the probability measure induced on the complete observation history under instance ν_i. Then

$$D_{\mathrm{KL}}(\mathbb{P}_0\|\mathbb{P}_1) = \sum_{t=1}^{T} \mathbb{E}_0\left[M_t\,\mathrm{kl}\big(\mu_0^c(A_t), \mu_1^c(A_t)\big)\right]. \tag{72}$$

Proof. We represent the feedback revealed to the learner at round t by

$$Y_t = \begin{cases} R_t^c, & M_t = 1, \\ \perp, & M_t = 0, \end{cases}$$

where $\perp$ denotes the unobserved symbol. The trajectory up to round T is given by $\mathcal{H}_T = (A_t, R_t^p, M_t, Y_t)_{t=1}^T$. For each $t \in \{1, \ldots, T\}$, denote the pre-action history by $\mathcal{F}_{t-1} = \sigma(\mathcal{H}_{t-1})$ and the filtration prior to the clean observation by $\mathcal{G}_t = \sigma(\mathcal{H}_{t-1}, A_t, R_t^p, M_t)$.

By the chain rule for relative entropy,

$$D_{\mathrm{KL}}(\mathbb{P}_0\|\mathbb{P}_1) = \sum_{t=1}^{T} \mathbb{E}_0\left[D_{\mathrm{KL}}\big(\mathbb{P}_0(\cdot \mid \mathcal{H}_{t-1}) \,\big\|\, \mathbb{P}_1(\cdot \mid \mathcal{H}_{t-1})\big)\right].$$

At each round t, the transition probability factors as

$$\mathbb{P}_i(A_t, R_t^p, M_t, Y_t \mid \mathcal{H}_{t-1}) = \pi_t(A_t \mid \mathcal{H}_{t-1})\,\mathbf{1}\{R_t^p = f(A_t)\}\,\kappa_t(M_t \mid \mathcal{H}_{t-1}, A_t, R_t^p)\,\mathbb{P}_i(Y_t \mid \mathcal{G}_t),$$

where π_t is the learner's action kernel and κ_t is its audit sampling kernel. Because the action kernel π_t, proxy mapping f, and audit policy κ_t are identical under both instances, their contributions cancel in the log-likelihood ratio.

Next, we evaluate the conditional divergence of Y_t given $\mathcal{G}_t$. When $M_t = 0$, $Y_t = \perp$ deterministically under both measures, so the conditional divergence vanishes identically:

$$D_{\mathrm{KL}}\big(\mathbb{P}_0(Y_t \in \cdot \mid \mathcal{G}_t) \,\big\|\, \mathbb{P}_1(Y_t \in \cdot \mid \mathcal{G}_t)\big) = 0.$$

When $M_t = 1$, $Y_t = R_t^c \sim \mathrm{Bernoulli}(\mu_i^c(A_t))$ under $\mathbb{P}_i$. Since $\mu_i^c(a) \in (0, 1)$, the measures share common support, and the conditional divergence reduces to the binary relative entropy:

$$D_{\mathrm{KL}}\big(\mathbb{P}_0(Y_t \in \cdot \mid \mathcal{G}_t) \,\big\|\, \mathbb{P}_1(Y_t \in \cdot \mid \mathcal{G}_t)\big) = \mathrm{kl}\big(\mu_0^c(A_t), \mu_1^c(A_t)\big).$$

Consequently,

$$D_{\mathrm{KL}}\big(\mathbb{P}_0(Y_t \in \cdot \mid \mathcal{G}_t) \,\big\|\, \mathbb{P}_1(Y_t \in \cdot \mid \mathcal{G}_t)\big) = M_t\,\mathrm{kl}\big(\mu_0^c(A_t), \mu_1^c(A_t)\big).$$

Taking the expectation $\mathbb{E}_0$ and summing over $t = 1, \ldots, T$ yields equation 72. This derivation holds for any adaptive auditing rule, without requiring independence between M_t and past observations. □

C.2.2 Proof of Theorem 2 (i)

Proof. Let an arbitrary learner satisfy $\sum_{t=1}^{T} M_t \le B$ almost surely on every admissible instance. We establish the lower bound by testing two realizable environments against this learner.

Step 1: Construction and realizability. Fix the gap

$$\Delta = \frac{1}{16\sqrt{B+1}}.$$

Consider a two-armed problem $\mathcal{A} = \{1, 2\}$ with canonical basis vectors $\phi(1) = e_1$ and $\phi(2) = e_2$ in $\mathbb{R}^2$. In both instances, the proxy reward is deterministically $1/2$ for both actions. Conditional on the action, clean rewards are independent Bernoulli draws with expectations

$$\mu_-^c = \left(\frac{1}{2} - \Delta, \frac{1}{2}\right)^\top, \qquad \mu_+^c = \left(\frac{1}{2} + \Delta, \frac{1}{2}\right)^\top.$$

Action 2 is optimal under ν_-, while action 1 is optimal under ν_+; both environments share the suboptimality gap Δ.

The instances are linearly realizable with parameters

$$\theta_{c,-}^* = \begin{pmatrix} 1/2 - \Delta \\ 1/2 \end{pmatrix}, \quad \theta_{c,+}^* = \begin{pmatrix} 1/2 + \Delta \\ 1/2 \end{pmatrix}, \quad \theta_p^* = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix},$$

and linear bias $\theta_{b,\pm}^* = \theta_p^* - \theta_{c,\pm}^* = \mp\Delta e_1$. Since $\Delta \le 1/16$, the parameters satisfy the unit-ball bounds

$$\max_{i\in\{-,+\}} \|\theta_{c,i}^*\|_2^2 \le \left(\frac{9}{16}\right)^2 + \frac{1}{4} < 1, \qquad \|\theta_p^*\|_2^2 = \frac{1}{2}, \qquad \|\theta_{b,\pm}^*\|_2 = \Delta < 1.$$

All rewards remain in $[0, 1]$, and the centered noise variables ξ_t^c and $\xi_t^b = -\xi_t^c$ are conditionally $1/2$-sub-Gaussian. Both instances are therefore admissible.

Step 2: Information limitation via budget. Let $\mathbb{P}_-$ and $\mathbb{P}_+$ denote the measures induced by the learner on the observation history. Let $C_1 = \sum_{t=1}^{T} M_t \mathbf{1}\{A_t = 1\}$ count the audits allocated to action 1. Because proxy feedback and the clean reward distribution of action 2 are identical across environments, Lemma 5 yields

$$\begin{aligned} D_{\mathrm{KL}}(\mathbb{P}_-\|\mathbb{P}_+) &= \mathbb{E}_-[C_1]\ \mathrm{kl}\left(\frac{1}{2} - \Delta, \frac{1}{2} + \Delta\right) \\ &\le 16\Delta^2 \mathbb{E}_-[C_1] \le 16B\Delta^2 = \frac{B}{16(B+1)} \le \frac{1}{16}, \end{aligned} \tag{73}$$

where we used equation 71 and $C_1 \le B$ almost surely. By Pinsker's inequality,

$$\mathrm{TV}(\mathbb{P}_-, \mathbb{P}_+) \le \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\mathbb{P}_-\|\mathbb{P}_+)} \le \frac{1}{4}. \tag{74}$$

Step 3: Regret lower bound. Let $N_1 = \sum_{t=1}^{T} \mathbf{1}\{A_t = 1\}$. Under instances ν_- and ν_+, the expected clean regret satisfies

$$\mathcal{R}_- = \Delta \mathbb{E}_-[N_1], \qquad \mathcal{R}_+ = \Delta\big(T - \mathbb{E}_+[N_1]\big).$$

Applying Lemma 4 to $N_1 \in [0, T]$ gives $\mathbb{E}_+[N_1] - \mathbb{E}_-[N_1] \le T\,\mathrm{TV}(\mathbb{P}_-, \mathbb{P}_+)$. Summing the regret identities and substituting equation 74 yields

$$\mathcal{R}_- + \mathcal{R}_+ = \Delta\big(T - (\mathbb{E}_+[N_1] - \mathbb{E}_-[N_1])\big) \ge \Delta T\big(1 - \mathrm{TV}(\mathbb{P}_-, \mathbb{P}_+)\big) \ge \frac{3}{4}\Delta T.$$

By the max-average inequality,

$$\max_{i\in\{-,+\}} \mathbb{E}_i[\mathrm{Reg}_c(T)] \ge \frac{\mathcal{R}_- + \mathcal{R}_+}{2} \ge \frac{3}{8}\Delta T = \frac{3T}{128\sqrt{B+1}} \ge \frac{T}{64\sqrt{B+1}}.$$

The edge case $B = 0$ follows identically with $\mathbb{P}_- = \mathbb{P}_+$. □

Scope of the lower bound. Because the proof depends only on $\mathbb{E}_-[C_1] \le B$, the result holds under the weaker expected budget constraint $\mathbb{E}_\nu[\sum_{t=1}^T M_t] \le B$. For $d > 2$, padding features with zero coordinates embeds the construction directly, establishing the rate-optimal budget scaling for any fixed $d \ge 2$. Finally, because the paired discrepancies are stochastic ($\xi_t^b \neq 0$), this construction does not address environments with deterministically noiseless paired observations.

C.3 Proof of Theorem 2 (ii)

We prove Theorem 2 (ii) using the information-theoretic testing inequality and the audited-history divergence bound established in Lemmas 4 and 5 of Section C.2.1. Specifically, we construct two instances—one unbiased and one biased—that induce identical proxy feedback but possess distinct clean-optimal actions. Under uniform auditing, information to distinguish between these two environments can be collected only by pulling the action that is suboptimal in the unbiased instance. This directly links statistical distinguishability to the clean regret incurred on the zero-bias subclass.

C.3.1 Audit Protocol and Regret Guarantee

Let $\mathcal{M}$ denote the normalized realizable instance class introduced in Theorem 2 (ii), and let $\mathcal{M}_0 = \{\nu \in \mathcal{M} : \theta_b^\star = 0\}$ be the subclass of zero-bias environments. Consider an arbitrary online learner satisfying the worst-case clean regret guarantee

$$\sup_{\nu \in \mathcal{M}} \mathbb{E}_\nu[\mathrm{Reg}_c(T)] \le G, \qquad 0 < G \le T/16. \tag{75}$$

The policy may be arbitrarily adaptive and randomized. The learner executes capped uniform auditing with base probability $p = B/T$. The budget availability indicator and audit decision are given by

$$I_t = \mathbf{1}\{N_{t-1} < B\}, \qquad M_t = I_t Z_t,$$

where $\{Z_t\}_{t=1}^T$ is a sequence of i.i.d. $\mathrm{Bernoulli}(p)$ random variables. Let $\mathcal{G}_t$ denote the σ-algebra generated by the observation history up to round $t-1$, the context x_t, and the chosen action A_t, prior to the realization of the audit indicator and the round-t reward noises. Since I_t and A_t are $\mathcal{G}_t$-measurable while Z_t is conditionally independent of $\mathcal{G}_t$, we have

$$\mathbb{E}[M_t \mid \mathcal{G}_t] = pI_t \le p. \tag{76}$$

Due to the hard budget constraint $\sum_{t=1}^T M_t \le B$, the indicators $\{M_t\}_{t=1}^T$ are not mutually independent. We retain I_t explicitly throughout the subsequent divergence calculations.

C.3.2 Proof of Theorem 2 (ii)

Proof. **Step 1: Construction of an Unbiased and an Indistinguishable Biased Instance.** Fix $\Delta \in (0, 1/4]$, whose exact value will be specified in Step 3. Consider a stationary environment with a two-action set $\mathcal{A} = \{1, 2\}$, associated with canonical basis vectors $\phi(1) = e_1$ and $\phi(2) = e_2$ in $\mathbb{R}^d$. In both instances, the proxy rewards are deterministic and given by

$$R_t^p = \begin{cases} 1/2, & A_t = 1, \\ 1/2 - \Delta, & A_t = 2. \end{cases}$$

Conditional on the chosen action A_t, the clean rewards are independent Bernoulli random variables with conditional expectations:

Instance	$\mu_c(1)$	$\mu_c(2)$	Optimal Action
ν_0	$1/2$	$1/2 - \Delta$	1
ν_1	$1/2$	$1/2 + \Delta$	2

Hence, the proxy feedback ranks action 1 strictly above action 2 in both environments, whereas the clean objective reverses this preference under ν_1.

The underlying parameter vectors (with vanishing coordinates for dimensions $k > 2$) are explicitly parameterized by

$$\begin{aligned} \theta_p^\star &= \tfrac{1}{2}e_1 + (\tfrac{1}{2} - \Delta)e_2, & & \\ \theta_{c,0}^\star &= \theta_p^\star, & \theta_{b,0}^\star &= 0, \\ \theta_{c,1}^\star &= \tfrac{1}{2}e_1 + (\tfrac{1}{2} + \Delta)e_2, & \theta_{b,1}^\star &= -2\Delta e_2. \end{aligned}$$

Both environments are linearly realizable, and $\nu_0 \in \mathcal{M}_0$ by construction. Since $\Delta \le 1/4$, the Euclidean norms satisfy $\|\theta^\star_{c,i}\|_2 \le \sqrt{(1/2)^2 + (3/4)^2} = \sqrt{13/16} < 1$ for $i \in \{0,1\}$, $\|\theta^\star_p\|_2 < 1$, and $\|\theta^\star_{b,1}\|_2 = 2\Delta \le 1/2$. All action features have unit norm, and rewards are supported in $[0,1]$. The zero-mean clean reward noises are conditionally $1/2$-sub-Gaussian, the proxy noise is identically zero, and the discrepancy noise satisfies $\xi^b_t = -\xi^c_t$. Both instances therefore belong to the normalized realizable class $\mathcal{M}$. In particular, vanishing mean bias in ν_0 does not imply identical sample-path realizations of clean and proxy signals.

Let $\mathbb{P}_i$ denote the probability measure induced on the observation history under ν_i. Define the expected clean regret and the number of pulls of action 2 by

$$\mathcal{R}_i = \mathbb{E}_i[\mathrm{Reg}_c(T)], \qquad N_2 = \sum_{t=1}^{T} \mathbf{1}\{A_t = 2\}.$$

Since action 2 is suboptimal only in ν_0 with gap Δ, and action 1 is suboptimal in ν_1 with the identical gap Δ, the expected clean regrets satisfy the exact identities:

$$\mathcal{R}_0 = \Delta \mathbb{E}_0[N_2], \qquad \mathcal{R}_1 = \Delta\big(T - \mathbb{E}_1[N_2]\big). \tag{77}$$

Step 2: Information Acquisition Incurs Regret on the Unbiased Instance. By construction, the two instances differ only in the clean-reward distribution of action 2. Let $C_2 = \sum_{t=1}^{T} M_t \mathbf{1}\{A_t = 2\}$ denote the total number of audits allocated to action 2. Invoking Lemma 5, the Kullback–Leibler divergence between the trajectory laws admits the exact decomposition:

$$D_{\mathrm{KL}}(\mathbb{P}_0 \,\|\, \mathbb{P}_1) = \mathbb{E}_0[C_2]\ \mathrm{kl}\left(\frac{1}{2} - \Delta, \frac{1}{2} + \Delta\right). \tag{78}$$

Since both $\mathbf{1}\{A_t = 2\}$ and I_t are $\mathcal{G}_t$-measurable, the tower property of conditional expectation combined with equation 76 implies

$$\begin{aligned}
\mathbb{E}_0[C_2] &= \sum_{t=1}^{T} \mathbb{E}_0[\mathbb{E}_0[M_t \mathbf{1}\{A_t = 2\} \mid \mathcal{G}_t]] \\
&= \sum_{t=1}^{T} \mathbb{E}_0[\mathbf{1}\{A_t = 2\}\, \mathbb{E}_0[M_t \mid \mathcal{G}_t]] \\
&= p\, \mathbb{E}_0\left[\sum_{t=1}^{T} I_t \mathbf{1}\{A_t = 2\}\right] \le p\, \mathbb{E}_0[N_2].
\end{aligned} \tag{79}$$

Substituting equation 79 and the relative entropy bound from Lemma 4 into equation 78 yields

$$D_{\mathrm{KL}}(\mathbb{P}_0 \,\|\, \mathbb{P}_1) \le 16 p \Delta^2 \mathbb{E}_0[N_2] = 16 p \Delta \mathcal{R}_0. \tag{80}$$

Inequality equation 80 encapsulates the fundamental statistical cost of uniform auditing: acquiring an informative clean observation requires pulling the suboptimal arm in the unbiased environment, revealing ground-truth feedback with probability at most p, and incurring an instantaneous clean regret of Δ. Any budget exhaustion can only strictly reduce the number of collected clean samples.

To make the resulting trade-off between the two environments explicit, Pinsker's inequality and Lemma 4 ensure that

$$\mathbb{E}_1[N_2] - \mathbb{E}_0[N_2] \le T \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\mathbb{P}_0 \,\|\, \mathbb{P}_1)} \le T \sqrt{8 p \Delta \mathcal{R}_0}.$$

Substituting this difference into the regret identity for $\mathcal{R}_1$ in equation 77 yields

$$\mathcal{R}_1 \ge \Delta T - \mathcal{R}_0 - \Delta T \sqrt{8 p \Delta \mathcal{R}_0}, \tag{81}$$

which concludes Step 2.

Step 3: robustness requires the two instances to be distinguished. Choose

$$\Delta = \frac{4G}{T}.$$

The assumption $0 < G \le T/16$ ensures $0 < \Delta \le 1/4$, so both constructed instances remain in $\mathcal{M}$. The uniform expected-regret guarantee equation 75 applies to both, giving

$$\mathcal{R}_0 \le G = \frac{\Delta T}{4}, \qquad \mathcal{R}_1 \le G = \frac{\Delta T}{4}.$$

By equation 77, this forces

$$\mathbb{E}_0[N_2] \le \frac{T}{4}, \qquad \mathbb{E}_1[N_2] \ge \frac{3T}{4}.$$

Thus the expected number of action-2 pulls must differ by at least $T/2$ between the two instances. Since $N_2 \in [0, T]$, the testing inequality implies

$$\frac{T}{2} \le \mathbb{E}_1[N_2] - \mathbb{E}_0[N_2] \le T\,\mathrm{TV}(P_0, P_1).$$

Consequently, $\mathrm{TV}(P_0, P_1) \ge 1/2$, and Pinsker's inequality, as proved in Lemma 4, gives

$$D_{\mathrm{KL}}(P_0 \| P_1) \ge 2\,\mathrm{TV}(P_0, P_1)^2 \ge \frac{1}{2}. \tag{82}$$

Step 4: Concluding the Zero-Bias Lower Bound. Combining the informational upper bound equation 80 with the required distinguishability bound equation 82 yields

$$\frac{1}{2} \le 16p\Delta\mathcal{R}_0 \implies \mathcal{R}_0 \ge \frac{1}{32p\Delta}.$$

Substituting the audit probability $p = B/T$ and the assigned gap $\Delta = 4G/T$ into this inequality gives

$$\mathcal{R}_0 \ge \frac{T^2}{128BG}.$$

Since the constructed unbiased instance ν_0 belongs to the zero-bias subclass $\mathcal{M}_0$ by definition, we conclude that

$$\sup_{\nu \in \mathcal{M}_0} \mathbb{E}_\nu[\mathrm{Reg}_c(T)] \ge \mathcal{R}_0 \ge \frac{T^2}{128BG}. \tag{83}$$

This completes the proof of Theorem 2 (ii) for any ambient dimension $d \ge 2$. □

Scope and Rate Interpretation. Theorem 2 (ii) characterizes a worst-case trade-off over the zero-bias subclass rather than an instance-dependent lower bound for every unbiased environment. Specifically, the constructed unbiased instance depends on the robust target G through the assigned gap $\Delta = 4G/T$. The technical restriction $G \le T/16$ guarantees that $\Delta \le 1/4$, which simultaneously ensures the parameter admissibility of both instances within $\mathcal{M}$ and forces sufficient separation between the action-selection frequencies across environments.

To see the operational consequence of this trade-off, suppose the learner guarantees the minimax-optimal robust rate matching Theorem 1, namely $G = CT/\sqrt{B}$ for some universal constant $C > 0$. Whenever $B \ge 256C^2$, the condition $G \le T/16$ is satisfied, and equation 83 yields

$$\sup_{\nu \in \mathcal{M}_0} \mathbb{E}_\nu[\mathrm{Reg}_c(T)] \ge \frac{T}{128C\sqrt{B}}.$$

Consequently, any uniform-auditing policy achieving worst-case safety against arbitrary bias cannot adapt to the benign $\widetilde{\mathcal{O}}(\sqrt{T})$ rate in zero-bias environments when $B \ll T$.

The protocol-specific step in this derivation is the marginal sampling bound in equation 79. An adaptive auditing policy that allocates higher audit probabilities to informative contexts or actions is not constrained by $\mathbb{E}[M_t \mid \mathcal{G}_t] \le B/T$. The lower bound in Theorem 2 (ii) thus exposes the structural limitation of non-adaptive audit allocation, without ruling out more graceful adaptation for adaptive schemes such as PACE.

C.4 Adaptation under a known bias direction

Context and Motivation. Theorem 2(ii) establishes that in the worst case over unstructured d-dimensional linear bias, uniform auditing cannot adapt to benign, zero-bias environments without sacrificing safety. However, in many real-world LLM-as-a-judge deployments, the potential discrepancy often concentrates along known nuisance directions. A prominent example is *verbosity bias* or *formatting preference* (Zheng et al., 2023; Shi et al., 2025), where the judge's score is inflated roughly proportionally to output length or structural markers: the feature direction $v \in \mathbb{R}^d$ is known *a priori*, while only the scalar distortion scale $\gamma \in \mathbb{R}$ remains unknown.

This motivates examining whether structural priors over the bias geometry can circumvent the adaptation barrier of Theorem 2(ii). Below, we formalize this structured setting and show that an explicit hypothesis-testing mechanism enables free adaptation when the judge is indeed unbiased ($\gamma = 0$). We emphasize that this relies on a zero-tested modification of the decision rule, rather than the default PACE-U action rule.

Model and algorithm. Let $v \in \mathbb{R}^d$ be known and write

$$z_t(a) = \phi(x_t, a)^\top v, \qquad \theta_p^\star = \theta_c^\star + \gamma v, \qquad |\gamma| \le S.$$

Assume that, for every candidate action,

$$\|\phi(x_t, a)\|_2 \le 1, \qquad 0 < \kappa \le |z_t(a)| \le 1,$$

and that both conditional reward means lie in $[0, 1]$. Let $\|\theta_j^\star\|_2 \le S_j$ for $j \in \{c, p\}$. As in PACE-U, the clean and proxy noises ξ_t^c, ξ_t^p, and their difference $\xi_t^b = \xi_t^p - \xi_t^c$, are conditionally centered and $\sigma_c, \sigma_p, \sigma_b$-sub-Gaussian, respectively. The two rewards need not be independent.

Use the capped uniform audit rule $M_t = \mathbf{1}\{N_{t-1} < B\} Z_t$, where $N_t = \sum_{s \le t} M_s$ and the independent coin Z_t is Bernoulli with probability $p = B/T$. The action is chosen before this coin or either current reward is observed. On an audited round,

$$Y_t = \frac{R_t^p - R_t^c}{z_t(A_t)} = \gamma + \frac{\xi_t^b}{z_t(A_t)}.$$

For $n \ge 1$, let $\widehat{\gamma}_n$ be the mean of the first n audited values Y_t. Fix $\delta \in (0, 1)$, and set

$$a_\delta = \frac{\sigma_b}{\kappa}\sqrt{2 \log \frac{8T}{\delta}}, \qquad C_0 = [-S, S], \qquad C_n = \left[\widehat{\gamma}_n - \frac{a_\delta}{\sqrt{n}}, \widehat{\gamma}_n + \frac{a_\delta}{\sqrt{n}}\right] \cap [-S, S].$$

If an intersection is empty, set $C_n = [-S, S]$. Let c_n and w_n be its midpoint and half-width.

Use the pre-round, unweighted ridge estimates

$$V_t^p = I_d + \sum_{s<t} \phi_s \phi_s^\top, \qquad \widehat{\theta}_t^p = (V_t^p)^{-1} \sum_{s<t} \phi_s R_s^p,$$
$$V_t^c = I_d + \sum_{s<t} M_s \phi_s \phi_s^\top, \quad \widehat{\theta}_t^c = (V_t^c)^{-1} \sum_{s<t} M_s \phi_s R_s^c, \qquad \phi_s = \phi(x_s, A_s).$$

Put $n_p = T$, $n_c = B$, and, for $j \in \{c, p\}$, define

$$D_j = d \log(1 + n_j/d), \qquad \alpha_j = \max\left\{\frac{1}{2}, S_j + \sigma_j \sqrt{D_j + 2\log(4/\delta)}\right\}, \qquad r_t^j(a) = \alpha_j \|\phi(x_t, a)\|_{(V_t^j)^{-1}}.$$

Write $[u]_0^1 = \min\{1, \max\{0, u\}\}$, and form

$$U_t^c(a) = [\phi(x_t, a)^\top \widehat{\theta}_t^c + r_t^c(a)]_0^1,$$
$$U_t^p(a) = [\phi(x_t, a)^\top \widehat{\theta}_t^p + r_t^p(a)]_0^1,$$
$$U_t^{\mathrm{cal}}(a) = \left[\phi(x_t, a)^\top \widehat{\theta}_t^p - z_t(a) c_{N_{t-1}} + r_t^p(a) + |z_t(a)| w_{N_{t-1}}\right]_0^1.$$

The zero-tested rule selects a maximizer, with fixed tie-breaking, of

$$S_t(a) = \begin{cases} \min\{U_t^c(a), U_t^p(a)\}, & 0 \in C_{N_{t-1}}, \\ \min\{U_t^c(a), U_t^{\mathrm{cal}}(a)\}, & 0 \notin C_{N_{t-1}}. \end{cases}$$

Theorem 3 (Known-direction adaptation)**.** *Under the assumptions above, let*

$$G_p = \alpha_p\sqrt{2TD_p}, \qquad K_\delta = \max\{S, \sqrt{2}\,a_\delta\}, \qquad h_B = \sum_{k=1}^{B}\frac{1}{k}, \qquad \rho = 1 - e^{-1}.$$

For $1 \le B \le T$, the zero-tested rule uses at most B clean labels and, with probability at least $1-\delta$,

$$\mathrm{Reg}_c(T) \le \min\left\{T,\ 2G_p + 4K_\delta\left[\sqrt{\frac{T\{h_B + \log(4/\delta)\}}{\rho p}} + \frac{T}{\sqrt{B+1}}\right]\right\}.$$

If $\gamma = 0$, then for every $0 \le B \le T$, with probability at least $1-\delta$,

$$\mathrm{Reg}_c(T) \le \min\{T, 2G_p\}.$$

Thus bounded parameter and noise envelopes give $\widetilde{O}(d\sqrt{T})$ regret when $\gamma = 0$, and $\widetilde{O}(d\sqrt{T} + (S + \sigma_b/\kappa)T/\sqrt{B})$ regret for $B \ge 1$.

Proof. Expose each audit coin before the current reward pair. This leaves the joint law unchanged, since the coin is independent of that pair and is not used to choose the action. At an audited time, $Y_t - \gamma$ is therefore conditionally σ_b/κ-sub-Gaussian, even for adaptive contexts and actions.

For completeness, set $r = \sigma_b/\kappa$. For fixed $u \in \mathbb{R}$,

$$L_t(u) = \exp\left\{u\sum_{s\le t} M_s(Y_s - \gamma) - \frac{u^2r^2}{2}N_t\right\}$$

is a nonnegative supermartingale. Stop it at the n-th audit or at T, and apply Markov's inequality with $u = a_\delta/(r^2\sqrt{n})$ and its negative. For every $n \le T$,

$$\Pr\left(n \text{ audits occur and } |\widehat{\gamma}_n - \gamma| > \frac{a_\delta}{\sqrt{n}}\right) \le 2e^{-a_\delta^2/(2r^2)} = \frac{\delta}{4T}.$$

When $r = 0$, each audited ratio equals γ almost surely and the same conclusion holds. A union bound gives an event $\mathcal{E}_\gamma$, of probability at least $1 - \delta/4$, on which $\gamma \in C_n$ for every attained count. On this event,

$$|\gamma - c_n| \le w_n, \qquad w_0 = S, \qquad w_n \le \min\{S, a_\delta/\sqrt{n}\} \le \frac{K_\delta}{\sqrt{n+1}} \quad (n \ge 1).$$

The last bound also holds for $n = 0$. If $0 \in C_n$, it follows that $|\gamma| \le 2w_n$.

The self-normalized ridge bound, applied separately to the proxy observations and the audited clean observations, gives events $\mathcal{E}_p, \mathcal{E}_c$, each of probability at least $1 - \delta/4$, on which, simultaneously for all rounds and actions,

$$|\phi(x_t, a)^\top(\widehat{\theta}_t^j - \theta_j^\star)| \le r_t^j(a), \qquad j \in \{c, p\}.$$

The audited design is predictable after the independent coin is exposed. It has at most B nonzero feature vectors, which gives the stated value of D_c. If $u_t = \|\phi_t\|^2_{(V_t^p)^{-1}}$, then $u_t \le 1$; the determinant lemma and $u \le 2\log(1+u)$ on $[0,1]$ yield

$$\sum_{t=1}^{T} u_t \le 2D_p, \qquad \sum_{t=1}^{T} r_t^p(A_t) \le G_p.$$

Work on $\mathcal{E}_\gamma \cap \mathcal{E}_p \cap \mathcal{E}_c$. Abbreviate the clean mean by $\mu_t(a)$. On a round using the raw proxy, the proxy mean differs from $\mu_t(a)$ by at most $|\gamma|$. Confidence and clipping imply

$$\mu_t(a) - |\gamma| \le S_t(a) \le \mu_t(a) + |\gamma| + 2r_t^p(a).$$

Comparing a clean-optimal action with the selected maximizer gives

$$\Delta_t \le 2r_t^p(A_t) + 2|\gamma| \le 2r_t^p(A_t) + 4w_{N_{t-1}}.$$

On a corrected round, the center of $U_t^{\mathrm{cal}}(a)$ differs from $\mu_t(a)$ by at most $r_t^p(a) + |z_t(a)|w_{N_{t-1}}$. Both entries of S_t are then optimistic for the clean mean, and

$$\Delta_t \le 2r_t^p(A_t) + 2|z_t(A_t)|w_{N_{t-1}} \le 2r_t^p(A_t) + 2w_{N_{t-1}}.$$

Consequently,

$$\mathrm{Reg}_c(T) \le 2G_p + 4\sum_{t=1}^{T} w_{N_{t-1}}. \tag{*}$$

It remains to bound the last sum under the hard cap. Let $I_t = \mathbf{1}\{N_{t-1} < B\}$ and $q_t = I_t/(N_{t-1}+1)$. Since $1 - e^{-q} \ge \rho q$ for $q \in [0, 1]$, conditional on the history before the coin,

$$\mathbb{E}[e^{-M_t q_t}] = 1 - pI_t(1 - e^{-q_t}) \le e^{-\rho p q_t}.$$

Thus $\exp\{\rho p \sum_{s\le t} q_s - \sum_{s\le t} M_s q_s\}$ is a nonnegative supermartingale. With probability at least $1 - \delta/4$,

$$\rho p \sum_{t=1}^{T} q_t \le \sum_{t=1}^{T} M_t q_t + \log(4/\delta) = h_{N_T} + \log(4/\delta) \le h_B + \log(4/\delta).$$

The equality holds pathwise: the k-th audit contributes $1/k$. On rounds before the cap, $w_{N_{t-1}} \le K_\delta\sqrt{q_t}$; Cauchy–Schwarz bounds their sum by

$$K_\delta\sqrt{\frac{T\{h_B + \log(4/\delta)\}}{\rho p}}.$$

After the cap, $w_{N_{t-1}} = w_B \le K_\delta/\sqrt{B+1}$, and there are at most T such rounds. Substitution into (*) proves the first claim on the intersection of four events, whose probability is at least $1 - \delta$. The additional bound by T follows because clean conditional means lie in $[0, 1]$.

If $\gamma = 0$, the confidence event keeps 0 in every attained C_n, so every round uses the raw branch. Its displayed regret bound then reduces to $\Delta_t \le 2r_t^p(A_t)$. This also holds when $B = 0$, since C_0 contains zero. Summing proves the second claim. Finally, $h_B \le 1 + \log B$ and $p = B/T$ give the stated rates. □

Discussion: Escaping the Robustness-Adaptation Dilemma. Theorem 3 offers two significant theoretical insights:

- **Bypassing the Minimax Trade-Off:** When $\gamma = 0$, the learner attains $\widetilde{\mathcal{O}}(d\sqrt{T})$ regret independently of the budget B (even for $B = 0$). This sharply contrasts with Theorem 2(ii), where achieving worst-case safety over all d-dimensional biases necessarily forces $\widetilde{\Omega}(T/\sqrt{B})$ regret on zero-bias instances. The reason this lower bound is bypassed is structural: isolating the unknown bias to a 1D scalar parameter γ alongside a bounded-margin condition $|z_t(a)| \ge \kappa$ allows the learner to verify the null hypothesis $\gamma = 0$ at an $\mathcal{O}(1/\sqrt{n})$ rate purely from passive audits, without requiring deliberate suboptimal explorations that sacrifice clean regret.
- **Additive vs. Multiplicative Penalty:** When $\gamma \ne 0$, the regret scales as $\widetilde{\mathcal{O}}(d\sqrt{T} + (S + \sigma_b/\kappa)T/\sqrt{B})$. Crucially, the $T/\sqrt{B}$ label-scarcity penalty is entirely decoupled from the ambient context dimension d, reflecting the fact that clean labels are only needed to identify the 1D bias magnitude γ, while the d-dimensional representation learning is carried entirely by the full proxy stream T.

D Comparisons with Alternative Feedback Models

D.1 Provable Gains over Budgeted Human-Only Learning

In this section, we compare the feedback model introduced in this work with standard budgeted online learning paradigms, with a particular focus on the human-only audit setting and models with

limited verification. Our primary objective is to isolate and quantify the statistical benefit conferred by paired proxy observations under a hard audit budget.

We begin by examining the human-only regime, where no auxiliary proxy signal is available. First, we establish a refined upper bound for PACE-U that explicitly depends on the sub-Gaussian scale of the paired-difference noise in Section D.1.1. Second, we prove a minimax lower bound for any human-only learning policy operating over a common normalized instance class, allowing for arbitrary adaptive audit schedules in Section D.1.3. Finally, by comparing these bounds, we exhibit a class of problems where PACE-U achieves strictly lower worst-case regret than all possible human-only learners in Section D.1.4. This separation demonstrates that paired proxy observations can yield a provable informational advantage that cannot be matched by clean evaluations alone.

D.1.1 Feedback Models and a Common Instance Class

Budgeted clean-feedback learning. In the budgeted observation framework of Efroni et al. (2021), unqueried rounds yield no reward feedback. For linear contextual bandits under a constant query budget $B \le T$, their CBM-OFUL algorithm achieves $\widetilde{\mathcal{O}}(dT/\sqrt{B})$ regret. To focus on the feedback models themselves rather than specific algorithmic designs, we consider a generic *human-only* learner. This learner can adaptively choose actions and up to B clean audits, but has no access to proxy feedback. In contrast, PACE-U operates with the identical audit budget while observing the proxy reward at every round, thereby acquiring paired proxy–reference feedback upon auditing.

Limited verification. Rangi et al. (2022) study multi-armed bandits with corrupted feedback, where a learner can verify the true (uncorrupted) reward in up to B rounds. In their model, rounds that are not verified still provide an unverified reward signal. This feedback structure is fundamentally different from the human-only regime, where uninspected rounds provide no feedback at all. Furthermore, their corruption is chosen adversarially, whereas our LLM proxy exhibits a stationary, feature-dependent bias. We clarify that the theoretical separation established below compares PACE-U to human-only learning; it is not meant as an empirical or theoretical dominance over verification algorithms that exploit corrupted observations.

A common normalized instance class. To establish a theoretical comparison between PACE-U and human-only learners, both regimes must be evaluated over the same collection of environments. To this end, we define $\mathcal{C}_\tau$ as the normalized class of linear conditional-mean problems satisfying

$$\|\phi_t(a)\|_2 \le 1, \qquad \mu_t^c(a), \mu_t^p(a) \in [0,1], \qquad S_c, S_p, S_b \le 1, \qquad \sigma_c, \sigma_p \le 1, \tag{84}$$

where the underlying parameter vectors satisfy $\theta_b^\star = \theta_p^\star - \theta_c^\star$. The parameter $\tau \ge 0$ denotes a valid upper bound on the conditional sub-Gaussian scale of the paired-difference noise $\xi_t^b = \xi_t^p - \xi_t^c$.

We clarify two key aspects of this definition:

- **Role of τ:** The parameter τ quantifies the stochastic noise in the paired differences rather than the systematic mean bias. In particular, the extreme case $\tau = 0$ corresponds to noiseless paired observations (where observation noise completely cancels out), which remains valid under Theorem 1 even when the systematic bias $\theta_b^\star$ is large.
- **Boundedness of rewards:** Only the conditional mean rewards are restricted to $[0,1]$, while realized observations may be unbounded. Specifically, the lower bound constructed in the next subsection uses Gaussian noise. Therefore, the resulting separation holds for sub-Gaussian feedback rather than almost-surely bounded realizations.

D.1.2 A Noise-Sensitive Regret Bound for PACE-U

Under PACE-U, the algorithm employs unit ridge regularization ($\lambda_c = \lambda_p = \lambda_b = 1$), unweighted regression on audited observations, and independent Bernoulli auditing with parameter $p = B/T$ until the budget B is exhausted. We instantiate the confidence radii of Theorem 1 using the valid paired-difference noise scale τ:

$$\ell = \log(5/\delta), \qquad L_c = L_b = d\log(1 + B/d), \qquad L_p = d\log(1 + T/d),$$

$$\alpha_j = \max\{1/2,\, S_j + \sigma_j\sqrt{L_j + 2\ell}\}, \quad j \in \{c,p\}, \qquad \alpha_b = \max\{1/2,\, S_b + \tau\sqrt{L_c + 2\ell}\}.$$

Proposition 1 (Noise-sensitive regret bound). *Fix $d, T \geq 1$, $1 \leq B \leq T$, and $\delta \in (0,1)$. For every instance in $\mathcal{C}_\tau$,* PACE-U *satisfies, with probability at least $1-\delta$,*

$$\mathrm{Reg}_c(T) \leq \min\left\{T,\ 20L\min\left\{\frac{dT}{\sqrt{B}},\ d\sqrt{T}+(\tau d+\sqrt{d})\frac{T}{\sqrt{B}}\right\}\right\}, \tag{85}$$

where $L = \log\frac{5(T+1)}{\delta}$, and the universal constant 20 is independent of all problem parameters. Furthermore, running PACE-U *with confidence parameter $\delta_T = (T+1)^{-2}$ and $L_T = \log(5(T+1)^3)$ yields the expected regret bound*

$$\mathbb{E}[\mathrm{Reg}_c(T)] \leq \min\left\{T,\ 20L_T\min\left\{\frac{dT}{\sqrt{B}},\ d\sqrt{T}+(\tau d+\sqrt{d})\frac{T}{\sqrt{B}}\right\}+\frac{T}{(T+1)^2}\right\}. \tag{86}$$

Proof. We depart from the explicit high-probability guarantee established in the proof of Theorem 1. Recall the notation

$$p=\frac{B}{T},\qquad \kappa=1-e^{-1},\qquad Q_B=\frac{2L_c+\ell}{\kappa p},\qquad A_c=\sqrt{TQ_B},\qquad A_p=\sqrt{2TL_p},$$

and the post-exhaustion horizon bound

$$H_B=\begin{cases}\min\{T,\ \lceil T\sqrt{2\ell/B}\rceil\}, & 1\leq B<T,\\ 0, & B=T.\end{cases}$$

Conditioned on the favorable event $\mathcal{E}_C\cap\mathcal{E}_W\cap\mathcal{E}_K$ (which holds with probability at least $1-\delta$), the cumulative regret admits the finite-sample upper bound

$$\mathrm{Reg}_c(T)\leq\min\{T,\ H_B+2\min\{\alpha_cA_c,\ \alpha_pA_p+\alpha_bA_c\}\}. \tag{87}$$

We now bound the direct-clean branch and calibrated-proxy branch separately under the normalized class $\mathcal{C}_\tau$.

Since $L\geq 1$, $d\geq 1$, and

$$\ell\leq L\leq dL,\qquad L_c=L_b\leq L_p\leq dL,$$

we have $L_j+2\ell\leq 3dL$ for every $j\in\{c,p,b\}$. Together with the normalized parameter and noise bounds, this gives

$$\alpha_c,\alpha_p\leq 1+\sqrt{3dL}\leq 3\sqrt{dL},\qquad \alpha_b\leq 1+\tau\sqrt{3dL}. \tag{88}$$

Defining the numerical constant $c_0=\sqrt{3/\kappa}$, the cumulative design widths satisfy

$$A_c\leq c_0T\sqrt{\frac{dL}{B}},\qquad A_p\leq\sqrt{2dTL},\qquad H_B\leq(1+\sqrt{2})T\sqrt{\frac{L}{B}}, \tag{89}$$

where the bound on H_B holds deterministically for $B=T$ and follows from $\lceil x\rceil\leq 1+x$ alongside $T\sqrt{L/B}\geq 1$ for $B<T$.

Direct-clean branch. Substituting the bounds from equation 88 and equation 89 into the direct-clean term yields

$$\begin{aligned}H_B+2\alpha_cA_c&\leq(1+\sqrt{2})T\sqrt{\frac{L}{B}}+6c_0\frac{dLT}{\sqrt{B}}\\&\leq(1+\sqrt{2}+6c_0)L\frac{dT}{\sqrt{B}}.\end{aligned}$$

Calibrated-proxy branch. Similarly, expanding the proxy-calibrated term gives

$$\begin{aligned}H_B+2\alpha_pA_p+2\alpha_bA_c&\leq(1+\sqrt{2})T\sqrt{\frac{L}{B}}+6\sqrt{2}dL\sqrt{T}+2c_0T\sqrt{\frac{dL}{B}}+2\sqrt{3}c_0\tau\frac{dLT}{\sqrt{B}}\\&\leq L\left[6\sqrt{2}d\sqrt{T}+2\sqrt{3}c_0\tau\frac{dT}{\sqrt{B}}+(2c_0+1+\sqrt{2})\frac{\sqrt{d}T}{\sqrt{B}}\right],\end{aligned}$$

where we used $\sqrt{L} \le L$ and $d \ge 1$.

Since $\kappa = 1 - e^{-1} > 1/2$, we have $c_0 = \sqrt{3/\kappa} < \sqrt{6} < 2.45$. By direct numerical evaluation,

$$1 + \sqrt{2} + 6c_0 < 17.2 < 20, \qquad \max\{6\sqrt{2},\, 2\sqrt{3}c_0,\, 2c_0 + 1 + \sqrt{2}\} < 10 < 20.$$

Taking the minimum over both branches and bounding the leading numerical prefactors by 20 establishes the high-probability inequality equation 85.

Expected regret bound. Let U_T denote the deterministic bound in the right-hand side of equation 85 evaluated at $L = L_T$. Since the conditional mean rewards are normalized in $[0, 1]$, the clean regret is deterministically bounded by T almost surely. Integrating across the sample space yields

$$\mathbb{E}[\mathrm{Reg}_c(T)] \le (1 - \delta_T)U_T + \delta_T \cdot T \le U_T + \frac{T}{(T+1)^2}.$$

Together with the trivial bound $\mathbb{E}[\mathrm{Reg}_c(T)] \le T$, this concludes the proof of equation 86. □

Interpretation and sufficient audit budget. In the calibrated-proxy branch, the statistical penalty from clean observation noise is reduced from $dT/\sqrt{B}$ to $\tau dT/\sqrt{B}$, accompanied by a $d\sqrt{T}$ cost for estimating the proxy mean across all T rounds. The secondary term $\sqrt{d}T/\sqrt{B}$ captures the regularized estimation of the bias vector alongside post-exhaustion regret, and remains present even when $\tau = 0$.

Whenever the audit budget satisfies

$$B \ge T\big(\tau + d^{-1/2}\big)^2, \tag{90}$$

we have $(\tau d + \sqrt{d})T/\sqrt{B} \le d\sqrt{T}$, recovering the unconstrained rate $\widetilde{\mathcal{O}}(d\sqrt{T})$. In the low-noise regime where $\tau \le d^{-1/2}$, setting $B = \min\{T, \lceil 4T/d \rceil\}$ is sufficient: either condition equation 90 is satisfied, or $B = T$, in which case the direct-clean branch independently matches the $\widetilde{\mathcal{O}}(d\sqrt{T})$ guarantee.

D.1.3 A Minimax Lower Bound for Human-Only Learning

A *human-only learner* sequentially selects actions A_t and audit indicators $M_t \in \{0, 1\}$ based solely on the observed contexts, the history of previously queried clean rewards, and private internal randomness, without access to realized proxy scores. The audit decision M_t is determined prior to revealing the potential clean label R_t^c, and the query budget is strictly respected: $\sum_{t=1}^{T} M_t \le B$ almost surely. The querying policy may be fully adaptive, and the learner may have prior knowledge of the parameter envelopes, the proxy conditional mean, and the underlying noise coupling.

To support the lower-bound construction with continuous observations below, we first formalize the information accumulation under Gaussian observation noise.

Lemma 6 (Gaussian history divergence and testing inequality)**.** *Consider two stationary instances over a shared finite action set $\mathcal{A}$, with clean rewards distributed as $\mathcal{N}(\mu_v(a), 1)$ and $\mathcal{N}(\mu_w(a), 1)$, respectively. The observation noise is independent across rounds and independent of the learner's decisions. For any human-only learner, let P_v and P_w denote the probability measures induced on the complete interaction history. Then*

$$D_{\mathrm{KL}}(P_v \| P_w) = \frac{1}{2}\mathbb{E}_v\left[\sum_{t=1}^{T} M_t\big(\mu_v(A_t) - \mu_w(A_t)\big)^2\right]. \tag{91}$$

Furthermore, for any measurable event E, the testing error satisfies

$$P_v(E^c) + P_w(E) \ge 1 - \mathrm{TV}(P_v, P_w) \ge 1 - \sqrt{\frac{1}{2}D_{\mathrm{KL}}(P_v \| P_w)}. \tag{92}$$

Proof. Let the observation at round t be represented by the tuple (A_t, M_t, Y_t), where $Y_t = R_t^c$ when $M_t = 1$ and $Y_t = \perp$ (an uninformative missing symbol) when $M_t = 0$. Let $\mathcal{H}_t = (A_s, M_s, Y_s)_{s=1}^{t}$

denote the history up to round t, and let U denote the learner's private randomness, which has an identical instance-independent distribution under both P_v and P_w.

Conditioned on $(U, \mathcal{H}_{t-1})$, the conditional distributions of the action A_t and audit decision M_t are governed entirely by the policy and are therefore identical under P_v and P_w. Consequently, their terms cancel out in the log-likelihood ratio. When $M_t = 0$, $Y_t = \perp$ deterministically under both measures, contributing zero relative entropy. When $M_t = 1$, the observation R_t^c follows a standard Gaussian distribution with mean $\mu_i(A_t)$ under P_i ($i \in \{v, w\}$). By the chain rule for relative entropy,

$$D_{\mathrm{KL}}(P_v \| P_w) = \sum_{t=1}^{T} \mathbb{E}_v \big[D_{\mathrm{KL}}\big(P_v(\cdot \mid U, \mathcal{H}_{t-1}, A_t, M_t) \,\|\, P_w(\cdot \mid U, \mathcal{H}_{t-1}, A_t, M_t)\big)\big].$$

Evaluating the Gaussian relative entropy yields

$$D_{\mathrm{KL}}\big(\mathcal{N}(\mu_v(A_t), 1) \,\|\, \mathcal{N}(\mu_w(A_t), 1)\big) = \frac{1}{2}\big(\mu_v(A_t) - \mu_w(A_t)\big)^2,$$

which, combined with the indicator M_t, yields identity equation 91.

The first inequality in equation 92 follows from the definition of total variation distance: $P_v(E^c) + P_w(E) = 1 - (P_v(E) - P_w(E)) \geq 1 - \mathrm{TV}(P_v, P_w)$. The second inequality is Pinsker's inequality, $\mathrm{TV}(P_v, P_w) \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(P_v \| P_w)}$. □

Theorem 4 (Human-Only Lower Bound with Noiseless Paired Differences). *Let $d \geq 2$ and $(d-1)^2 \leq B \leq T$ be integers. For any human-only learner constrained to at most B clean reference queries almost surely, there exists an environment instance in $\mathcal{C}_0$ such that*

$$\mathbb{E}[\mathrm{Reg}_c(T)] \geq \frac{dT}{32\sqrt{B}}. \tag{93}$$

The underlying instance is characterized by a static context, an action space of cardinality 2^{d-1}, and conditionally standard Gaussian marginal noise. Furthermore, the proxy reward mean satisfies $\mu_p^*(s) \equiv 1/2$ identically across all candidate actions $s \in \mathcal{A}$. The lower bound persists even if both the uninformative proxy mean and the coupled noise mechanism are fully disclosed to the learner a priori.

Proof. **Step 1: Construction of the normalized shared-noise family.** Set $m = d - 1$ and configure the coordinate-wise gap parameter as $\gamma = \sqrt{m}/(8\sqrt{B})$. For each hypercube vertex $v \in \{-1, +1\}^m$, we define a common discrete action set $\mathcal{A} = \{-1, +1\}^m$, mapped through the canonical feature embedding

$$\phi(s) = \left(\frac{1}{\sqrt{2}}, \frac{s}{\sqrt{2m}}\right)^\top \in \mathbb{R}^{m+1}.$$

Corresponding to each latent configuration v, we specify the clean, proxy, and bias parameter vectors by

$$\theta_c^v = \left(\frac{1}{\sqrt{2}}, \gamma v\right)^\top, \qquad \theta_p = \left(\frac{1}{\sqrt{2}}, \mathbf{0}_m\right)^\top, \qquad \theta_b^v = (0, -\gamma v)^\top.$$

Let $\{\eta_t\}_{t=1}^T \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ be an exogenous noise sequence independent of the filtration generated by the learner's sampling policy. At each round $t \in [T]$, the paired feedback signals are governed by the coupled mechanism

$$R_t^c = \phi(A_t)^\top \theta_c^v + \eta_t, \qquad R_t^p = \frac{1}{2} + \eta_t. \tag{94}$$

Under this specification, the conditional expectation of the clean reward evaluates to $\mu_v^c(s) = 1/2 + \gamma s^\top v/\sqrt{2m}$. Crucially, the realized paired discrepancy under equation 94 satisfies

$$R_t^p - R_t^c = \phi(A_t)^\top \theta_b^v$$

almost surely across all trajectories, which establishes an exact noiseless difference structure ($\tau = 0$) despite the unit marginal variance of each individual feedback channel.

Under the dimensional budget condition $B \geq m^2$, admissibility is directly verified:

$$\|\phi(s)\|_2^2 = 1, \qquad \|\theta_p\|_2^2 = \frac{1}{2}, \qquad \|\theta_b^v\|_2^2 = m\gamma^2 = \frac{m^2}{64B} \leq \frac{1}{64}, \qquad \|\theta_c^v\|_2^2 = \frac{1}{2} + m\gamma^2 \leq \frac{33}{64} < 1.$$

Furthermore, the Cauchy–Schwarz inequality guarantees the uniform mean-reward containment

$$\left|\mu_v^c(s) - \frac{1}{2}\right| \leq \gamma\sqrt{\frac{m}{2}} = \frac{m}{8\sqrt{2B}} \leq \frac{1}{8\sqrt{2}} < \frac{1}{2},$$

ensuring that $\mu_v^c(s) \in [0,1]$ globally for all $s \in \mathcal{A}$. Because the additive residuals are conditionally standard normal (hence 1-sub-Gaussian), every constructed model in this family strictly belongs to $\mathcal{C}_0$.

Step 2: Information divergence between adjacent environments. Consider an arbitrary adaptive human-only policy, and let P_v denote the probability law induced over the observable trajectory in environment v. For an arbitrary coordinate $i \in [m]$, let $v^{(i)}$ denote the sign vector differing from v solely at index i, with $v_i^{(i)} = -v_i$. For any arm pull $s \in \mathcal{A}$, the expected mean shift and its induced Fisher information satisfy

$$\mu_v^c(s) - \mu_{v^{(i)}}^c(s) = \frac{2\gamma}{\sqrt{2m}} s_i v_i, \qquad \frac{1}{2}\left(\mu_v^c(s) - \mu_{v^{(i)}}^c(s)\right)^2 = \frac{\gamma^2}{m}.$$

Because this squared discrepancy is invariant to the choice of action s, invoking Lemma 6 alongside the hard audit budget constraint $\sum_{t=1}^T M_t \leq B$ implies that for any adaptive querying strategy,

$$D_{\mathrm{KL}}(P_v \,\|\, P_{v^{(i)}}) = \frac{\gamma^2}{m}\mathbb{E}_v\left[\sum_{t=1}^T M_t\right] \leq \frac{B\gamma^2}{m} = \frac{1}{64}. \tag{95}$$

By Pinsker's inequality, the total variation distance is bounded by $\mathrm{TV}(P_v, P_{v^{(i)}}) \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(P_v \,\|\, P_{v^{(i)}})} \leq \frac{1}{8\sqrt{2}} \leq \frac{1}{8}$.

Step 3: Accumulation of coordinate-wise errors and regret bound. Let $A_{t,i} \in \{-1, +1\}$ denote the i-th coordinate of the arm selected at round t. For adjacent instances $(v, v^{(i)})$, consider the decision event $E = \{A_{t,i} = v_i\}$. By the Le Cam two-point testing lemma, the joint probability of error satisfies

$$P_v(A_{t,i} \neq v_i) + P_{v^{(i)}}(A_{t,i} \neq v_i^{(i)}) = P_v(E^c) + P_{v^{(i)}}(E) \geq 1 - \mathrm{TV}(P_v, P_{v^{(i)}}) \geq \frac{7}{8}.$$

Partitioning the 2^m hypercube vertices into 2^{m-1} disjoint adjacent pairs along coordinate i and taking the uniform average over the family yields

$$\frac{1}{2^m} \sum_{v \in \{-1,+1\}^m} P_v(A_{t,i} \neq v_i) \geq \frac{7}{16}. \tag{96}$$

In instance v, the unique optimal action is $s^* = v$, which induces the instantaneous regret gap

$$\mu_v^c(v) - \mu_v^c(s) = \frac{\gamma}{\sqrt{2m}} \sum_{i=1}^m (1 - s_i v_i) = \frac{2\gamma}{\sqrt{2m}} \sum_{i=1}^m \mathbf{1}\{s_i \neq v_i\}.$$

Accumulating this instantaneous shortfall over rounds $t \in [T]$ and coordinates $i \in [m]$, and applying Fubini's theorem in conjunction with equation 96, the average cumulative regret is lower bounded by

$$\begin{aligned} \frac{1}{2^m} \sum_{v \in \{-1,+1\}^m} \mathbb{E}_v[\mathrm{Reg}_c(T)] &\geq \frac{2\gamma}{\sqrt{2m}} \cdot mT \cdot \frac{7}{16} = \frac{7mT}{64\sqrt{2B}} \\ &\geq \frac{7dT}{128\sqrt{2B}} \geq \frac{dT}{32\sqrt{B}}, \end{aligned}$$

where the penultimate step follows from $m = d - 1 \geq d/2$ for all $d \geq 2$, and the final inequality uses $7 \geq 4\sqrt{2}$. Since the average regret over this finite family exceeds the threshold, there exists at least one worst-case instance $v^\star \in \{-1, +1\}^m$ achieving the stated lower bound. Finally, a priori knowledge of θ_p or the noise coupling does not alter the information-theoretic lower bound, as both remain invariant across all worlds and the realized proxy trajectories remain unobservable to the human-only learner. □

Necessary budget for human-only learning. In the regime of Theorem 4, achieving a human-only clean regret of at most $K_T d\sqrt{T}$ requires

$$\frac{dT}{32\sqrt{B}} \le K_T d\sqrt{T}, \qquad \text{or equivalently} \qquad B \ge \frac{T}{1024 K_T^2}.$$

When $K_T = \mathrm{polylog}(T)$, this imposes an audit budget of $B = \widetilde{\Omega}(T)$. In contrast, when $\tau \le d^{-1/2}$ is known, PACE-U attains the classical $\widetilde{\mathcal{O}}(d\sqrt{T})$ rate using an audit budget of only $B = \mathcal{O}(T/d)$ within the comparison regime. Thus, paired feedback yields an $\mathcal{O}(d)$ reduction in required audits up to logarithmic factors.

D.1.4 A Strict Regret Separation under a Common Audit Budget

Corollary 1 (Separation via Growing Dimension)**.** *For any integer $d \ge 2$, consider the scaling $T = d^4$ and $B = d^3$. Then, PACE-U calibrated with known $\tau = 0$ and failure tolerance $\delta_T = (T+1)^{-2}$ achieves*

$$\sup_{\nu \in \mathcal{C}_0} \mathbb{E}_\nu^{\text{PACE-U}}[\mathrm{Reg}_c(T)] \le 40 L_T d^3 + 1 = \widetilde{\mathcal{O}}(T^{3/4}). \tag{97}$$

In contrast, for the class $\Pi_{\text{human}}(B)$ of all human-only learners subject to the budget constraint $\sum_{t=1}^T M_t \le B$ almost surely,

$$\inf_{\pi \in \Pi_{\text{human}}(B)} \sup_{\nu \in \mathcal{C}_0} \mathbb{E}_\nu^\pi[\mathrm{Reg}_c(T)] \ge \frac{d^{7/2}}{32} = \frac{T^{7/8}}{32}. \tag{98}$$

Consequently, the ratio of the worst-case regret of PACE-U to the minimax regret of any human-only learner vanishes asymptotically as $d \to \infty$.

Proof. Direct substitution confirms that $(d-1)^2 \le B \le T$ for all $d \ge 2$. Under $\tau = 0$, evaluating the constituent terms yields

$$d\sqrt{T} = d^3, \qquad \sqrt{d}\,\frac{T}{\sqrt{B}} = d^3, \qquad \frac{dT}{\sqrt{B}} = d^{7/2}.$$

Applying Proposition 1 along with $T/(T+1)^2 \le 1$ establishes the upper bound equation 97. Meanwhile, Theorem 4 directly delivers the matching lower bound equation 98 over the same instance class $\mathcal{C}_0$.

Taking the ratio of the two bounds gives

$$\frac{\sup_{\nu \in \mathcal{C}_0} \mathbb{E}_\nu^{\text{PACE-U}}[\mathrm{Reg}_c(T)]}{\inf_{\pi \in \Pi_{\text{human}}(B)} \sup_{\nu \in \mathcal{C}_0} \mathbb{E}_\nu^\pi[\mathrm{Reg}_c(T)]} \le \frac{40 L_T d^3 + 1}{d^{7/2}/32} = \frac{1280 L_T}{\sqrt{d}} + \frac{32}{d^{7/2}} \longrightarrow 0 \quad \text{as } d \to \infty,$$

where the final convergence holds because $L_T = \log(5(d^4+1)^3) = \mathcal{O}(\log d)$. □

This separation formalizes that paired observations offer an information advantage through variance reduction rather than ranking accuracy. In our lower-bound construction, the proxy mean provides zero ranking signal across actions. However, because the proxy and clean feedback share realization noise, subtracting the two yields an exact, noiseless measurement of the linear bias. In contrast, a human-only learner must estimate the clean parameters from noisy reference labels alone, incurring a statistical penalty that cannot be bypassed even with prior knowledge of the coupling structure.

Importantly, this result establishes a minimax separation over a specific problem class ($\mathcal{C}_0$) rather than a pointwise dominance across all environments. The asymptotic polynomial gap arises in the growing-dimension regime with an exponential action space (2^{d-1} arms), and relies on conditionally Gaussian noise.

Finally, this separation is fully consistent with Theorem 2 (ii). While the robustness–adaptation trade-off applies to general, noisy discrepancies, here the learner is equipped with the additional knowledge that discrepancy noise vanishes ($\tau = 0$). This structural assumption restricts the envelope of adversarial instances and allows paired auditing to eliminate estimation variance, achieving provably lower regret than querying clean labels in isolation.

D.2 A Cost-Matched Comparison with Multi-Fidelity Feedback

Multi-fidelity (MF) bandits associate each arm with several observation sources of different costs and accuracies. At each query, the learner selects an arm–fidelity pair, pays for that fidelity, and receives one sample from the selected source. Cheaper fidelities can have biased means relative to the high-fidelity target; standard formulations often supply a known bound on this discrepancy (Kandasamy et al., 2016; Poiani et al., 2022; 2024). The learner can use inexpensive evaluations to eliminate implausible arms before spending on the expensive source. This tradeoff has been analyzed under distinct objectives: capital-based, cost-weighted regret (Kandasamy et al., 2016), alternative cost-budgeted regret (Wang et al., 2023), and the total cost of best-arm identification (Poiani et al., 2022; 2024).

Our feedback interface fundamentally departs from the canonical multi-fidelity (MF) bandit framework in both its cost structure and its information interface (Kandasamy et al., 2016; Wang et al., 2023). First, regarding cost, every actual decision automatically yields a proxy reward R_t^p at zero charged feedback cost, whereas an audit reveals the clean feedback R_t^c at an additional audit cost; in contrast, standard MF models charge a strictly positive cost even for low-fidelity queries. Second, regarding information, an audit additionally reveals R_t^c for the same outcome while retaining R_t^p, thereby providing access to the paired difference $D_t = R_t^p - R_t^c$; conversely, standard MF querying is exclusive—requesting high fidelity returns only the high-fidelity evaluation rather than a paired realization. Crucially, even when the proxy reward remains biased, pairing enables variance reduction whenever this discrepancy exhibits low noise. Since standard MF objectives (e.g., cost-weighted or capital regret) differ from our clean regret metric, we unify both interfaces under the same horizon T, charged budget Λ, and cumulative clean-regret objective to enable a rigorous theoretical comparison.

D.2.1 Common environment, feedback rules, and cost

We adopt the normalized problem class $\mathcal{C}_\tau$ and notation introduced in Appendix D.1. At round t, the learner observes the context x_t and action set $\mathcal{A}(x_t)$, and subsequently selects an action A_t. For brevity, we denote $\phi_t(a) = \phi(x_t, a)$ and $\phi_t = \phi_t(A_t)$. The potential feedback associated with the chosen action is governed by:

$$\begin{aligned} R_t^c &= \phi_t^\top \theta_c^\star + \xi_t^c, & R_t^p &= \phi_t^\top \theta_p^\star + \xi_t^p, \\ D_t &= R_t^p - R_t^c = \phi_t^\top \theta_b^\star + \xi_t^b, & \theta_b^\star &= \theta_p^\star - \theta_c^\star, \qquad \xi_t^b = \xi_t^p - \xi_t^c. \end{aligned} \tag{99}$$

Specifically, we assume $\|\phi_t(a)\|_2 \le 1$, $\|\theta_j^\star\|_2 \le 1$ for all $j \in \{c, p, b\}$, and that the conditional means satisfy $\mu_c^\star(x_t, a), \mu_p^\star(x_t, a) \in [0, 1]$. The conditional sub-Gaussian noise scales satisfy $\sigma_c, \sigma_p \le 1$ and $\sigma_b \le \tau$. Here, τ is a known, valid upper bound on the stochastic noise of the *paired difference*, rather than the systematic mean bias. Realized reward observations are not required to be bounded. Both interfaces are evaluated under the cumulative clean pseudo-regret:

$$\mathrm{Reg}_c(T) = \sum_{t=1}^{T} \left[\max_{a \in \mathcal{A}(x_t)} \mu_c^\star(x_t, a) - \mu_c^\star(x_t, A_t) \right]. \tag{100}$$

To facilitate a minimax comparison under a common environment class, let $\mathcal{C}_\tau^{\mathrm{fix}}(d, T) \subseteq \mathcal{C}_\tau$ denote the subclass of instances characterized by deterministic, exogenous sequences of contexts, action sets, and features. Conditional on the chosen action, a fresh potential feedback pair (R_t^p, R_t^c) is drawn at each round from an instance-specific kernel under both interfaces, independently of past outcomes and the learner's internal randomization. While this restricted subclass suffices to establish our minimax lower bounds, we emphasize that the upper bound guarantees for PACE-U hold more broadly over the full class $\mathcal{C}_\tau$.

Known fidelity-discrepancy bounds. In the two-fidelity case, classical MF models assume a known upper bound on the difference between the low- and high-fidelity means (Kandasamy et al., 2016; Wang et al., 2023). Within our common linear model, this condition defines a subclass: for a known $\zeta \in (0, 1]$, let $\mathcal{C}_{\tau,\zeta}^{\mathrm{fix}}(d, T) \subseteq \mathcal{C}_\tau^{\mathrm{fix}}(d, T)$ consist of the instances satisfying

$$\left| \mu_p^\star(x_t, a) - \mu_c^\star(x_t, a) \right| = \left| \phi_t(a)^\top \theta_b^\star \right| \le \zeta, \qquad \text{for all } t \text{ and } a \in \mathcal{A}(x_t). \tag{101}$$

When this subclass is used, ζ is public information available to both learners. Our normalization already implies the envelope 1; the class $\mathcal{C}_\tau^{\mathrm{fix}}(d,T)$ does not require a tighter envelope and is therefore broader with respect to this mean-discrepancy restriction. The PACE-U upper bound remains valid on every such subclass. Moreover, the hard family in Theorem 5 satisfies the same known envelope $\zeta = 1/4$ for every hidden instance, as verified in equation 111. Its information-theoretic lower bound holds even when this envelope is supplied in advance. Consequently, both separation results below remain valid when $\zeta = 1/4$ is given to both learners; they do not rely on withholding this standard MF information. Here ζ bounds the mean bias, whereas τ bounds the noise in the paired difference.

Let c_L and c_H denote the known unit costs of low- and high-fidelity feedback, respectively, with $0 < c_L < c_H$. To ensure a fair comparison over T sequential decisions, both learners operate under the same cumulative feedback budget Λ. We restrict Λ to $Tc_L \le \Lambda \le Tc_H$, where Tc_L represents the baseline budget required to query low fidelity at every round, while Tc_H permits querying high fidelity unconditionally. Within this budget, the maximum numbers of affordable clean audits (or high-fidelity queries) under the respective feedback interfaces are given by:

$$B = \left\lfloor \frac{\Lambda}{c_H} \right\rfloor, \qquad Q = \left\lfloor \frac{\Lambda - Tc_L}{c_H - c_L} \right\rfloor. \tag{102}$$

Here, B denotes the audit capacity when proxy signals are free, whereas Q represents the number of rounds where the learner can upgrade from low to high fidelity under the cost differential $c_H - c_L$.The two observation models operate under distinct revelation and charging protocols:

- **Always-observed proxy and paired audit (Our interface):** At each round t, the learner automatically observes the proxy feedback R_t^p at zero charged feedback cost. In addition, the learner may selectively audit the current outcome to reveal R_t^c at cost c_H, obtaining the *paired* realization (R_t^p, R_t^c). Denoting the audit decision by $M_t \in \{0,1\}$, the policy must satisfy the hard budget constraint $\sum_{t=1}^T M_t \le B$ almost surely. PACE-U serves as one admissible policy under this protocol, employing independent Bernoulli audit decisions that are conditionally independent of current rewards.
- **Exclusive multi-fidelity query (Canonical MF interface):** At each round t, the learner must purchase exactly one evaluation source for the selected action: querying the low-fidelity source costs c_L and reveals only R_t^p, whereas querying the high-fidelity source costs c_H and reveals only R_t^c. Crucially, the action and fidelity source must be selected simultaneously before observing any feedback from the current round. The policy may be adaptive, randomized, and aware of all environment parameters, public features, prices, the noise bound τ, and any known discrepancy envelope, subject only to the hard budget constraint that total charged expenditure does not exceed Λ almost surely.

We emphasize that assigning a zero price to the proxy feedback is a principled modeling abstraction of the *charged feedback resource* in our setting, rather than a claim that generating proxy scores incurs zero physical or computational overhead. In practical LLM deployments, evaluation via an automated LLM judge is orders of magnitude cheaper than acquiring gold reference feedback—such as expert human annotation or exhaustive verification. Because the per-query inference cost of the proxy is virtually negligible compared to that of a clean audit, our theoretical model normalizes this marginal proxy cost to zero to isolate the binding bottleneck of the audit budget. Both learners encounter the identical sequence of T contexts and are evaluated by the cumulative clean pseudo-regret in equation 100; the sole distinction lies in their feedback revelation mechanisms and accounting rules.

For $\mathcal{C} \subseteq \mathcal{C}_\tau^{\mathrm{fix}}(d,T)$, write

$$\begin{aligned} \mathfrak{R}_{\mathrm{prox}}(\mathcal{C};T,\Lambda) &= \inf_{\pi \in \Pi_{\mathrm{prox}}(\Lambda)} \sup_{\nu \in \mathcal{C}} \mathbb{E}_\nu^\pi[\mathrm{Reg}_c(T)], \\ \mathfrak{R}_{\mathrm{MF}}(\mathcal{C};T,\Lambda) &= \inf_{\pi \in \Pi_{\mathrm{MF}}(\Lambda)} \sup_{\nu \in \mathcal{C}} \mathbb{E}_\nu^\pi[\mathrm{Reg}_c(T)]. \end{aligned} \tag{103}$$

The infima include every policy obeying the respective observation rule, not only a particular upper-confidence algorithm.

Proposition 2 (Cost Identity and Protocol Inclusion)**.** *Any admissible exclusive multi-fidelity policy executes at most Q high-fidelity queries almost surely, satisfying $Q \le B \le T$. Furthermore, for every common instance class $\mathcal{C} \subseteq \mathcal{C}_\tau^{\mathrm{fix}}(d,T)$,*

$$\mathfrak{R}_{\mathrm{prox}}(\mathcal{C};T,\Lambda) \le \mathfrak{R}_{\mathrm{MF}}(\mathcal{C};T,\Lambda). \tag{104}$$

Proof. Let an arbitrary exclusive multi-fidelity policy execute N_H high-fidelity queries and $T - N_H$ low-fidelity queries over the T-round horizon. The total charged cost incurred by this policy is

$$c_H N_H + c_L(T - N_H) = Tc_L + (c_H - c_L)N_H.$$

Because the policy must satisfy the budget constraint $Tc_L + (c_H - c_L)N_H \le \Lambda$ almost surely, it follows directly from the definition of Q that $N_H \le \lfloor \frac{\Lambda - Tc_L}{c_H - c_L} \rfloor = Q$ almost surely. Furthermore, under the standing condition $\Lambda \le Tc_H$, straightforward algebraic rearrangement yields:

$$\frac{\Lambda - Tc_L}{c_H - c_L} \le \frac{\Lambda}{c_H} \quad \Longleftrightarrow \quad c_H(\Lambda - Tc_L) \le \Lambda(c_H - c_L) \quad \Longleftrightarrow \quad \Lambda \le Tc_H.$$

Taking floor operations on both sides establishes $Q \le B \le T$.

Next, we establish the minimax regret inequality equation 104 via a policy simulation argument. Specifically, any admissible exclusive policy $\pi_{\mathrm{MF}} \in \Pi_{\mathrm{MF}}(\Lambda)$ can be simulated by an proxy policy $\pi_{\mathrm{prox}} \in \Pi_{\mathrm{prox}}(\Lambda)$ that maintains an internal, virtual interaction history $\mathcal{H}_t^{\mathrm{vir}}$. At round t:

- If π_{MF} issues a low-fidelity query, π_{add} executes the same action, observes the freely available proxy feedback R_t^p, and feeds R_t^p to $\mathcal{H}_t^{\mathrm{vir}}$ at zero charged feedback cost.
- If π_{MF} issues a high-fidelity query, π_{add} executes an audit on the selected action, observes the paired feedback (R_t^p, R_t^c), and feeds only R_t^c into $\mathcal{H}_t^{\mathrm{vir}}$ while discarding R_t^p. This step incurs an audit cost of c_H.

Using the identical internal randomization stream, the joint distribution of selected actions and reported feedback under π_{prox} perfectly matches that of π_{MF} at every round $t \in [T]$. Moreover, the total audit expenditure incurred by π_{prox} satisfies:

$$c_H N_H \le c_H Q \le c_H B \le \Lambda,$$

which guarantees that π_{prox} is strictly budget-feasible ($\pi_{\mathrm{prox}} \in \Pi_{\mathrm{prox}}(\Lambda)$). Consequently, on every environment instance $\nu \in \mathcal{C}$, the cumulative clean regret of π_{prox} is identically distributed to that of π_{MF}, implying:

$$\mathbb{E}_\nu^{\pi_{\mathrm{prox}}}[\mathrm{Reg}_c(T)] = \mathbb{E}_\nu^{\pi_{\mathrm{MF}}}[\mathrm{Reg}_c(T)].$$

Taking the supremum over $\nu \in \mathcal{C}$ followed by the infimum over all admissible policies yields:

$$\mathfrak{R}_{\mathrm{prox}}(\mathcal{C}; T, \Lambda) = \inf_{\pi \in \Pi_{\mathrm{prox}}(\Lambda)} \sup_{\nu \in \mathcal{C}} \mathbb{E}_\nu^{\pi}[\mathrm{Reg}_c(T)] \le \inf_{\pi_{\mathrm{MF}} \in \Pi_{\mathrm{MF}}(\Lambda)} \sup_{\nu \in \mathcal{C}} \mathbb{E}_\nu^{\pi_{\mathrm{MF}}}[\mathrm{Reg}_c(T)] = \mathfrak{R}_{\mathrm{MF}}(\mathcal{C}; T, \Lambda),$$

which completes the proof. □

D.2.2 A LOWER BOUND AT THE AVAILABLE HIGH-QUERY COUNT

The noise-sensitive regret bound for PACE-U established in Appendix D.1, configured with the hard audit budget B, guarantees:

$$\sup_{\nu \in \mathcal{C}_\tau^{\mathrm{fix}}(d,T)} \mathbb{E}_\nu^{\text{PACE-U}}[\mathrm{Reg}_c(T)] \le 20 L_T \min\left\{ \frac{dT}{\sqrt{B}},\ d\sqrt{T} + \frac{(\tau d + \sqrt{d})T}{\sqrt{B}} \right\} + \frac{T}{(T+1)^2}, \quad (105)$$

where $L_T = \log\big(5(T+1)^3\big)$. We now contrast this *achievable upper bound* against an information-theoretic lower bound that holds for any admissible exclusive multi-fidelity policy under its effective high-fidelity query cap Q.

Theorem 5 (Cost-Matched Exclusive-Feedback Lower Bound)**.** *Let $d \ge 2$, $0 \le \tau \le \sqrt{2}$, and suppose the effective high-query budget satisfies:*

$$(d-1)^2 \le Q \le B \le T. \quad (106)$$

Then, the minimax clean regret of any exclusive multi-fidelity policy is lower bounded by:

$$\mathfrak{R}_{\mathrm{MF}}\big(\mathcal{C}_\tau^{\mathrm{fix}}(d,T); T, \Lambda\big) \ge \frac{dT}{32\sqrt{Q}}. \quad (107)$$

This bound permits arbitrary adaptive, history-dependent selections of candidate actions and query fidelities. Furthermore, the hard instance exhibits a strictly positive, non-vanishing proxy mean bias and non-degenerate paired-difference noise whenever $\tau > 0$. The same bound holds on $\mathcal{C}_{\tau,1/4}^{\mathrm{fix}}(d,T)$ when the discrepancy envelope $\zeta = 1/4$ is known to the learner.

Proof. We establish the lower bound via a multi-hypothesis testing construction over an ensemble of 2^m environments indexed by the vertices of a hypercube.

Step 1: Construction and admissibility of the hard instance family. Set $m = d-1$ and define the coordinate-wise signal gap $\gamma = \sqrt{m}/(8\sqrt{Q})$. We index instances by $v \in \{-1,+1\}^m$ and provide the learner with 2^m actions corresponding to the hypercube vertices $s \in \{-1,+1\}^m$ at every round. The context-action feature mapping is static and defined as:

$$\phi(s) = \left(\frac{1}{\sqrt{2}}, \frac{s}{\sqrt{2m}}\right)^\top \in \mathbb{R}^d. \tag{108}$$

For each configuration v, we specify the underlying parameter vectors for the clean reward, proxy reward, and discrepancy as:

$$\theta^\star_{c,v} = \left(\frac{1}{\sqrt{2}}, \gamma v\right)^\top, \quad \theta^\star_p = \left(\frac{5\sqrt{2}}{8}, 0_m\right)^\top, \quad \theta^\star_{b,v} = \theta^\star_p - \theta^\star_{c,v} = \left(\frac{\sqrt{2}}{8}, -\gamma v\right)^\top. \tag{109}$$

The induced conditional expected rewards evaluate to:

$$\mu^\star_{c,v}(s) = \phi(s)^\top \theta^\star_{c,v} = \frac{1}{2} + \frac{\gamma}{\sqrt{2m}} s^\top v, \qquad \mu^\star_p(s) = \phi(s)^\top \theta^\star_p = \frac{5}{8}. \tag{110}$$

We verify that these parameters strictly comply with the normalization envelopes of $\mathcal{C}^{\mathrm{fix}}_\tau(d,T)$. First, $\|\phi(s)\|_2^2 = \frac{1}{2} + \frac{m}{2m} = 1$. Since $Q \geq m^2$, we have $m\gamma^2 = \frac{m^2}{64Q} \leq \frac{1}{64}$, ensuring:

$$\|\theta^\star_{c,v}\|_2^2 = \frac{1}{2} + m\gamma^2 \leq \frac{33}{64} < 1, \quad \|\theta^\star_p\|_2^2 = \frac{25}{32} < 1, \quad \|\theta^\star_{b,v}\|_2^2 = \frac{1}{32} + m\gamma^2 \leq \frac{3}{64} < 1.$$

Moreover, applying the Cauchy–Schwarz inequality gives $|\mu^\star_{c,v}(s) - 1/2| \leq \gamma\sqrt{m/2} \leq 1/(8\sqrt{2}) < 1/2$, which guarantees that $\mu^\star_{c,v}(s) \in [0,1]$ for all s. The conditional mean proxy bias is strictly positive and non-vanishing across the entire action space:

$$0 < \frac{1 - 1/\sqrt{2}}{8} \leq \mu^\star_p(s) - \mu^\star_{c,v}(s) \leq \frac{1 + 1/\sqrt{2}}{8} < \frac{1}{4}. \tag{111}$$

The learner may be fully informed of the entire family structure, the proxy mean, and this discrepancy envelope a priori, with only the latent sign vector v remaining unknown.

Step 2: Coupled observation noise and marginal unfaithfulness. At round t, conditioned on the chosen action A_t, the environment generates potential clean and proxy feedback by coupling two independent standard Gaussian variables $G_t, W_t \sim \mathcal{N}(0,1)$:

$$\begin{aligned} R^c_t &= \mu^\star_{c,v}(A_t) + \sqrt{1 - \tau^2/4}\, G_t - \frac{\tau}{2} W_t, \\ R^p_t &= \frac{5}{8} + \sqrt{1 - \tau^2/4}\, G_t + \frac{\tau}{2} W_t. \end{aligned} \tag{112}$$

Because $\tau \leq \sqrt{2}$, both marginal noises have unit variance ($\mathrm{Var}(R^c_t) = \mathrm{Var}(R^p_t) = 1$), and the paired realization residual satisfies $D_t - \phi(A_t)^\top \theta^\star_{b,v} = \tau W_t$, confirming that the discrepancy noise is conditionally τ-sub-Gaussian.

Crucially, under equation 112, the proxy observation $R^p_t \sim \mathcal{N}(5/8, 1)$ is *statistically uninformative*: its distribution is entirely identical across all actions s and all latent instances v.

Step 3: Information divergence under the high-query budget. Let π be an arbitrary exclusive multi-fidelity policy, and let P^π_v denote the probability law induced over the observable interaction history under instance v (including selected actions, query fidelity choices, and revealed rewards). For any coordinate $i \in [m]$, let $v^{(i)}$ denote the sign vector differing from v solely at index i, with $v^{(i)}_i = -v_i$.

Conditional on selecting action s, the expected clean rewards between instances v and $v^{(i)}$ differ by:

$$\mu^\star_{c,v}(s) - \mu^\star_{c,v^{(i)}}(s) = \frac{2\gamma}{\sqrt{2m}} s_i v_i.$$

Under an exclusive query at round t:

- A *low-fidelity query* reveals only $R_t^p \sim \mathcal{N}(5/8, 1)$, whose conditional Kullback–Leibler (KL) divergence between v and $v^{(i)}$ vanishes identically ($D_{\mathrm{KL}} = 0$).
- A *high-fidelity query* reveals $R_t^c \sim \mathcal{N}(\mu^\star_{c,v}(A_t), 1)$, contributing a conditional KL divergence of $\frac{1}{2}(\mu^\star_{c,v}(s) - \mu^\star_{c,v^{(i)}}(s))^2 = \frac{\gamma^2}{m}$.

Because the policy's action and fidelity selection kernels are identical under both instances, they contribute zero to the log-likelihood ratio. Applying the chain rule for relative entropy across the T rounds yields:

$$D_{\mathrm{KL}}(P_v^\pi \,\|\, P_{v^{(i)}}^\pi) = \frac{\gamma^2}{m}\mathbb{E}_v^\pi[N_H] \le \frac{Q\gamma^2}{m} = \frac{1}{64}, \tag{113}$$

where N_H is the total number of high-fidelity queries executed, satisfying $N_H \le Q$ almost surely by Proposition 2. Pinsker's inequality then bounds the total variation distance between the full trajectory laws:

$$\mathrm{TV}(P_v^\pi, P_{v^{(i)}}^\pi) \le \sqrt{\frac{1}{2}D_{\mathrm{KL}}(P_v^\pi \,\|\, P_{v^{(i)}}^\pi)} \le \frac{1}{\sqrt{128}} \le \frac{1}{8}.$$

Step 4: Testing error and cumulative clean regret. Under instance v, the unique optimal action is $s^\star = v$. Testing the latent sign v_i using the i-th coordinate of the chosen action $A_{t,i}$ yields an instantaneous error event $E = \{A_{t,i} = v_i\}$. By the Le Cam two-point testing lemma:

$$\mathbb{P}_v^\pi(A_{t,i} \ne v_i) + \mathbb{P}_{v^{(i)}}^\pi(A_{t,i} \ne v_i^{(i)}) = \mathbb{P}_v^\pi(E^c) + \mathbb{P}_{v^{(i)}}^\pi(E) \ge 1 - \mathrm{TV}(P_v^\pi, P_{v^{(i)}}^\pi) \ge \frac{7}{8}. \tag{114}$$

Summing equation 114 over all 2^{m-1} disjoint adjacent pairs along coordinate i and averaging over a uniform prior on $v \in \{-1, +1\}^m$ implies that the coordinate-wise error probability satisfies:

$$\frac{1}{2^m}\sum_{v\in\{-1,+1\}^m} \mathbb{P}_v^\pi(A_{t,i} \ne v_i) \ge \frac{7}{16}.$$

For any chosen action s, the instantaneous suboptimality gap with respect to the clean objective is:

$$\mu^\star_{c,v}(v) - \mu^\star_{c,v}(s) = \frac{\gamma}{\sqrt{2m}}\sum_{i=1}^m (1 - s_i v_i) = \frac{2\gamma}{\sqrt{2m}}\sum_{i=1}^m \mathbf{1}\{s_i \ne v_i\}.$$

Accumulating this instantaneous shortfall over rounds $t \in [T]$ and coordinates $i \in [m]$, and substituting $\gamma = \frac{\sqrt{m}}{8\sqrt{Q}}$, the average cumulative clean regret over the ensemble is lower bounded by:

$$\frac{1}{2^m}\sum_{v\in\{-1,+1\}^m} \mathbb{E}_v^\pi[\mathrm{Reg}_c(T)] \ge \frac{2\gamma}{\sqrt{2m}} \cdot mT \cdot \frac{7}{16} = \frac{7mT}{64\sqrt{2Q}} \ge \frac{dT}{32\sqrt{Q}},$$

where the final inequality holds because $m = d - 1 \ge d/2$ for all $d \ge 2$, and $7/(64\sqrt{2}) > 1/32$. Since the maximum is at least the average, the worst-case regret over $\mathcal{C}_\tau^{\mathrm{fix}}(d, T)$ obeys the same lower bound, completing the proof. □

Corollary 2 (Quantitative comparison)**.** *Under the conditions of Theorem 5, let U_{PACE} denote the supremum of PACE-U's expected regret on $\mathcal{C}_\tau^{\mathrm{fix}}(d, T)$. Then*

$$\frac{U_{\mathrm{PACE}}}{\mathfrak{R}_{\mathrm{MF}}(\mathcal{C}_\tau^{\mathrm{fix}}(d,T); T, \Lambda)} \le 640 L_T \min\left\{\sqrt{\frac{Q}{B}}, \sqrt{\frac{Q}{T}} + \left(\tau + \frac{1}{\sqrt{d}}\right)\sqrt{\frac{Q}{B}}\right\} + \frac{32\sqrt{Q}}{d(T+1)^2}. \tag{115}$$

Proof. Divide equation 105 by equation 107 and simplify each of the two branches. □

D.2.3 Disentangling Cost-Structure and Paired-Information Effects

The regret bound in equation 105 encompasses two distinct mechanisms reflected by its two constituent branches:

- **Direct clean estimation (Branch 1):** Achieves an advantage purely through *cost-structure asymmetries* ($B \geq Q$), even when proxy and clean observation noises are entirely independent.
- **Calibrated proxy estimation (Branch 2):** Exploits *paired noise cancellation*, yielding a strict statistical advantage even in regimes where the high-query budget gap closes ($Q/B \to 1$).

Below, we formalize these two independent sources of separation through concrete scaling regimes.

Separation 1: Cost-structure advantage without paired variance reduction. We first demonstrate that even when the proxy feedback is completely uninformative and its noise is statistically independent of the clean noise, our interface strictly outperforms the exclusive MF interface solely because free proxy signals spare budget for additional clean audits.

Corollary 3 (Cost-Structure Separation under Independent Feedback Noise)**.** *For any integer $n \geq 2$, consider the problem scaling:*

$$d = 2, \qquad T = n^2, \qquad c_L = 1, \qquad c_H = 2, \qquad \Lambda = n^2 + n, \qquad \tau = \sqrt{2}. \tag{116}$$

Under these parameters, the query budgets evaluate to $Q = n$ and $B = n(n+1)/2$, while the cumulative regret guarantees satisfy:

$$\mathfrak{R}_{\mathrm{MF}}\big(\mathcal{C}^{\mathrm{fix}}_{\sqrt{2}}(2, n^2);\, n^2, \Lambda\big) \geq \frac{T^{3/4}}{16}, \qquad U_{\mathrm{PACE}} \leq 40\sqrt{2} L_T \sqrt{T} + 1. \tag{117}$$

Consequently, the ratio $U_{\mathrm{PACE}}/\mathfrak{R}_{\mathrm{MF}} \to 0$ asymptotically as $n \to \infty$.

Proof. The budget expressions for Q and B follow by direct substitution into equation 102. The lower bound is an immediate consequence of Theorem 5. For the upper bound, evaluating the direct-clean branch (the first term) of equation 105 using $d = 2$ and $B = n(n+1)/2 \geq T/2$ immediately yields $U_{\mathrm{PACE}} \leq 40\sqrt{2} L_T \sqrt{T} + 1$.

Crucially, setting $\tau = \sqrt{2}$ implies that the cross-covariance between the clean and proxy noises in equation 112 is $1 - \tau^2/2 = 0$. Under bivariate Gaussianity, zero covariance implies statistical independence. Furthermore, the proxy conditional mean is static by equation 110, providing zero signal regarding the identity of the optimal arm. Thus, this strict asymptotic separation is certified entirely by the budget asymmetry (namely $B = \Theta(T)$ versus $Q = \Theta(\sqrt{T})$), without relying on any variance reduction from paired differences. □

Separation 2: Paired-information advantage with matched query counts. To isolate the intrinsic value of *pairing* when the budget discrepancy is negligible, we introduce an auxiliary *independent-sample* interface as an intermediate baseline. Under this interface, the learner observes free proxy feedback at every round and is additionally permitted up to B clean queries (the decision to query round t may depend on the observed proxy). However, each clean query yields an *independent* draw from the clean reward marginal, rather than being paired with that round's proxy realization. While this model removes the proxy cost handicap of standard MF bandits, it denies access to the paired difference $D_t = R^p_t - R^c_t$. Let $\mathfrak{R}_{\mathrm{ind}}(\mathcal{C}; T, B)$ denote its minimax clean regret.

Proposition 3 (Minimax Lower Bound under Independent Querying)**.** *Let $d \geq 2$, $0 \leq \tau \leq \sqrt{2}$, and suppose $(d-1)^2 \leq B \leq T$. Then:*

$$\mathfrak{R}_{\mathrm{ind}}\big(\mathcal{C}^{\mathrm{fix}}_{\tau}(d, T);\, T, B\big) \geq \frac{dT}{32\sqrt{B}}. \tag{118}$$

Proof. We employ the identical hard instance family specified in equation 109–equation 112, with Q replaced by B in the definition of γ. Because the proxy reward distribution is identical across all

candidate actions and all latent instances, free proxy observations contribute zero relative entropy to the trajectory law. Each independently queried clean label contributes at most γ^2/m conditional relative entropy. Since the policy executes at most B total clean queries, the bound on the total KL divergence in equation 113 and the subsequent Le Cam testing argument in equation 114 apply directly with B substituting for Q, establishing equation 118. □

Equipped with this control interface, we now establish a strict separation driven entirely by the low noise of paired discrepancies in a regime where the budget gap vanishes ($Q/B \to 1$).

Corollary 4 (Low-Noise Separation with Asymptotically Matched Query Capacities)**.** *For any dimension $d \geq 2$, configure the horizon, budget, and price schedules as:*

$$T = d^4, \qquad c_H = 1, \qquad c_L = d^{-2}, \qquad \Lambda = d^3, \qquad \tau = d^{-1/2}. \tag{119}$$

Under this scaling, the affordable clean evaluation counts satisfy:

$$B = d^3, \qquad Q = d^3 - d^2 + d - 1, \qquad \textit{such that} \quad \frac{Q}{B} \longrightarrow 1 \quad (\textit{as } d \to \infty). \tag{120}$$

Nevertheless, the minimax regret metrics diverge asymptotically:

$$U_{\mathrm{PACE}} \leq 60 L_T d^3 + 1, \qquad \mathfrak{R}_{\mathrm{MF}}\big(\mathcal{C}_\tau^{\mathrm{fix}}(d,T); T, \Lambda\big) \geq \frac{d^{7/2}}{32}, \qquad \mathfrak{R}_{\mathrm{ind}}\big(\mathcal{C}_\tau^{\mathrm{fix}}(d,T); T, B\big) \geq \frac{d^{7/2}}{32}. \tag{121}$$

In particular, both ratios $U_{\mathrm{PACE}}/\mathfrak{R}_{\mathrm{MF}}$ and $U_{\mathrm{PACE}}/\mathfrak{R}_{\mathrm{ind}}$ scale as $\mathcal{O}(\log d/\sqrt{d})$, vanishing to zero as $d \to \infty$.

Proof. Direct algebra yields $(\Lambda - Tc_L)/(c_H - c_L) = (d^3 - d^2)/(1 - d^{-2}) = d^4/(d+1) = d^3 - d^2 + d - 1 + \frac{1}{d+1}$. Taking the floor establishes the formula for Q in equation 120. The lower bound for $\mathfrak{R}_{\mathrm{MF}}$ follows directly from Theorem 5 using $Q \leq B = d^3$, while the lower bound for $\mathfrak{R}_{\mathrm{ind}}$ follows from Proposition 3.

For the upper bound on PACE-U, we evaluate the proxy-calibrated branch (the second term) in equation 105. Under the assigned scaling, we have $d\sqrt{T} = d^3$ and $(\tau d + \sqrt{d})T/\sqrt{B} = (d^{1/2} + d^{1/2})d^4/d^{3/2} = 2d^3$. Summing these yields $d\sqrt{T} + (\tau d + \sqrt{d})T/\sqrt{B} = 3d^3$, which upon multiplication by the prefactor $20L_T$ establishes $U_{\mathrm{PACE}} \leq 60 L_T d^3 + 1$.

In contrast, the direct clean estimation branch alone would scale as $\mathcal{O}(dT/\sqrt{B}) = \mathcal{O}(d^{7/2})$, matching the lower bounds. The polynomial rate improvement from $\tilde{\mathcal{O}}(d^{7/2})$ down to $\tilde{\mathcal{O}}(d^3)$ is thus enabled strictly by calibrating the abundant proxy stream via low-noise paired differences ($\tau \ll 1$). Because the auxiliary independent interface permits an identical query budget ($B = d^3$) but cannot cancel realization noise via pairing, it remains bounded from below by $\Omega(d^{7/2})$. This confirms that paired observation confers a fundamental informational advantage beyond budget accounting. □

Concluding discussion on comparative scope. The theoretical separations established above evaluate fixed-horizon policies under a shared feedback budget Λ and our task-relevant clean regret $\mathrm{Reg}_c(T)$; they are not meant as a critique of canonical cost-weighted regret objectives natively studied in multi-fidelity literature. Together, the two constructions clarify the dual benefits of our interface:

1. Corollary 3 isolates the *pure cost benefit*—namely, that providing automated proxy evaluations at negligible cost spares budget for significantly more clean verifications ($B \gg Q$).
2. Corollary 4 isolates the *pure information benefit*—showing that even when budget parity is asymptotically enforced ($Q/B \to 1$), observing simultaneous paired realizations enables noise cancellation and variance reduction that independent multi-source querying cannot replicate.

We note that the scaling in Corollary 4 employs a vanishing cost ratio $c_L/c_H = d^{-2}$; this is structurally necessary because with fixed, non-vanishing unit costs c_L, a budget of $\Lambda = d^3$ would be strictly insufficient to cover the mandatory baseline cost $Tc_L = \Theta(d^4)$ required to sustain T rounds of interaction in an exclusive query regime.

D.3 Comparison with Bandits with Biased Offline Data

Both learning with biased offline data and learning with an online proxy address the fundamental challenge of exploiting auxiliary feedback whose mean deviates from the true target reward. Despite this common theme, the two paradigms operate under fundamentally different feedback availability and information structures.

In the biased offline data formulation (Cheung & Lyu, 2024; Zhang et al., 2025), the learner receives auxiliary samples prior to interaction, while every subsequent online trial reveals the ground-truth target reward. In our setting, the feedback structure is reversed: the inexpensive proxy R_t^p is observed at every round, whereas the clean target reward R_t^c is observed only when an audit is purchased, subject to a hard budget $\sum_{t=1}^{T} M_t \le B$. Consequently, discarding the proxy reduces the problem to budgeted clean learning rather than full-feedback bandit optimization. Crucially, each audit provides a paired realization (R_t^p, R_t^c), allowing the learner to directly regress the signed, feature-dependent discrepancy $\theta_b^\star = \theta_p^\star - \theta_c^\star$ via the difference $D_t = R_t^p - R_t^c$.

Statistically, both paradigms reveal that auxiliary feedback benefits learning through a delicate interplay between systematic discrepancy, residual noise, and feature coverage. In our model, a large systematic bias $\theta_b^\star$ can coexist with small discrepancy noise σ_b, so that paired differences enable precise calibration even when the uncalibrated proxy is systematically misleading. However, such variance reduction remains contingent on the audit coverage over the feature space (reflected in the design matrix V_t^b), paralleling the role of offline design coverage identified in Zhang et al. (2025).

Finally, the statistical barrier that governs our robustness–adaptation trade-off mirrors the lower-bound mechanism in Cheung & Lyu (2024). There, distinguishing whether offline data is benign requires target pulls; in our framework, instances with identical proxy rewards require clean audits to identify the optimal action. Under a uniform audit rate B/T, playing an informative action reveals clean feedback with low probability, demonstrating that the ability to calibrate a biased proxy does not eliminate the fundamental statistical cost of identifying its bias. A natural unification would initialize proxy estimation using historical data, while reserving online audits to calibrate the proxy during deployment.

D.4 Comparison with Online Proxies and Additional Ground-Truth Feedback

Two closely related approaches combine online auxiliary feedback with scarce ground-truth information: CUUCB for coupled measurements (Mirza et al., 2026) and CESA-LinUCB for contextual source reliability (Ghasemi & Crowley, 2026). We compare their observation models and theoretical characterizations with the realizable model analyzed for PACE-U.

D.4.1 Coupled measurements: CUUCB

Setting and guarantee. Mirza et al. (2026) study stochastic K-armed bandits with coarse feedback on every pull and paired ground-truth feedback on selected pulls. The means may be arbitrarily misaligned. Their sampling rule tracks prescribed per-arm fractions $\alpha_i \in (0, 1)$, and CUUCB learns the covariance to tune a control-variate estimator. Under bounded measurements and their variance-ratio assumption, Theorem 5.3 gives gap-dependent and gap-free bounds. Using v_i^c for the clean variance and ρ_i for the correlation, its effective variance is

$$V_i^\star = v_i^c\big[1 - (1-\alpha_i)\rho_i^2\big], \qquad \frac{V_i^\star}{\alpha_i} = v_i^c\left(\frac{1-\rho_i^2}{\alpha_i} + \rho_i^2\right).$$

Making the sampling-fraction dependence of the remainder explicit, the gap-free guarantee has the form

$$\mathbb{E}[\mathrm{Reg}_c^{\mathrm{CUUCB}}(T)] \le \widetilde{O}\left(\sqrt{T\sum_{i=1}^{K}\frac{V_i^\star}{\alpha_i}} + r_{\max}(1+\gamma)\sum_{i=1}^{K}\frac{1}{\alpha_i}\right),$$

where measurements lie in $[0, r_{\max}]$ and their assumption is $v_i^c \le \gamma^2 v_i^p$, with v_i^p the coarse variance. The remainder matters when the fractions shrink with T.

Relation to our model. Paired feedback and correlation-based gains are shared features of the two approaches. Our setting additionally models the observed context–action features $\phi_t(a)$ and shares $\theta_c^\star$, $\theta_p^\star$, and $\theta_b^\star$ across them. The benchmark is the clean-optimal action for each current context. Consequently, learning requires controlling prediction uncertainty in the directions of the observed features, and our analysis allows predictable context and action-set sequences without an i.i.d. context assumption.

PACE-U audits independently with probability B/T while budget remains and enforces $\sum_{t=1}^{T} M_t \leq B$ almost surely. A common CUUCB fraction also controls total queries up to an $O(K)$ rounding term, by its Lemma 4.1. Thus the budget conventions are closely related. Our analysis specifically handles the interaction between a global stopping cap and the adaptively generated clean, proxy, and bias design matrices.

Different uses of coupling. Our calibrated predictor subtracts an estimated signed bias learned from

$$D_t = R_t^p - R_t^c = \phi_t^\top \theta_b^\star + \xi_t^b, \qquad \xi_t^b = \xi_t^p - \xi_t^c.$$

Its confidence radius uses a known valid sub-Gaussian envelope σ_b for this residual. This exposes a favorable regime where the paired difference is substantially less noisy than either label. CUUCB's covariance adaptation can exploit either sign of correlation; our fixed subtraction does not generally attain its optimal scalar mixing variance. The theoretical extension we study is calibration through shared linear features under a hard audit budget, with explicit control of both reward and bias estimation.

D.4.2 Contextual source reliability: CESA-LinUCB

Setting and guarantee. Ghasemi & Crowley (2026) consider multiple evaluators with context-dependent bias and learn which sources are reliable. Reliability means that an evaluator's mean bias is at most ϵ_{tol} in magnitude for every action at that context. Independent audits arrive with probability p_{aud} and return binary ground-truth probes used to train reliability labels. CESA-LinUCB uses the learned trust weights in reward regression. Under its stated assumptions, including realizable trust classifiers and informative audit labels, Theorem 4 reports

$$\mathcal{R}_T^{\mathrm{latent}} \leq \widetilde{O}\left(\sqrt{\frac{T d_{\mathrm{VC}}}{p_{\mathrm{aud}}}} + d_{\mathrm{ctx}}\sqrt{T}\right) + \epsilon_{\mathrm{tol}} T.$$

Here d_{ctx} denotes the dimension called d in that paper, d_{VC} is the trust-class complexity, and the suppressed factors include the inverse square of its audit-fidelity margin. Its analysis uses i.i.d. contexts. The parameter ϵ_{tol} is the allowed mean bias within a trust region. Their Theorem 1 also states an impossibility result for learning without audits.

Numerical calibration and structural assumptions. PACE-U receives a clean numerical label whose conditional expectation equals $\phi_t^\top \theta_c^\star$. A binary good/bad probe does not by itself imply this expectation identity. The numerical label makes D_t observable and supports estimation of the magnitude and sign of the discrepancy. Accordingly, PACE-U can correct a proxy that remains systematically biased across contexts, provided the shared linear model holds and the audits are informative in the relevant feature directions. It requires no classification of that proxy into trustworthy and untrustworthy regions.

This capability relies on exact linear realizability of the clean and proxy means, bounded parameters, and valid noise envelopes. CESA-LinUCB instead places structure on the trust regions and can describe bias functions outside our linear class. Our absence of an $\epsilon_{\mathrm{tol}} T$ term therefore reflects a different model and correction mechanism; it does not establish a stronger guarantee for its full bias class. Moreover, σ_b measures centered difference noise, whereas ϵ_{tol} measures tolerated mean error. These quantities cannot be identified.

Audit counts and interpretation of rates. Setting $p_{\mathrm{aud}} = B/T$ gives B audits in expectation and rewrites the trust-learning term as $T\sqrt{d_{\mathrm{VC}}/B}$. This substitution exhibits a familiar dependence on scarce supervision. It does not impose an almost-sure cap on the number of audits or analyze stopping at that cap. Thus a $T/\sqrt{B}$ dependence alone does not distinguish our work.

D.4.3 What our theoretical characterization addresses

Our upper bound separates the uncertainty of direct clean estimation from that of proxy estimation plus bias calibration. Under the normalized parameter bounds, $\sigma_c, \sigma_p \le 1$, and a known valid envelope $\sigma_b \le \tau$, the noise-sensitive guarantee used in the preceding comparisons gives, for $1 \le B \le T$,

$$\mathbb{E}[\mathrm{Reg}_c(T)] \le \widetilde{O}\left(\min\left\{T,\ \frac{dT}{\sqrt{B}},\ d\sqrt{T} + \frac{(\tau d + \sqrt{d})T}{\sqrt{B}}\right\}\right).$$

The proxy term uses all online observations, while the calibration term depends on the audited design and paired residual noise. The remaining $\sqrt{d}\,T/\sqrt{B}$ term must be retained even when τ is small. These guarantees concern numerical calibration with the stated valid envelopes. They neither estimate an optimal covariance-based mixing coefficient nor require the mean bias $\theta_b^\star$ to vanish.

Our lower bounds address two additional questions over the full normalized realizable class, without imposing a small paired-residual envelope. The hard-budget minimax bound is $\Omega(T/\sqrt{B+1})$, already in dimension two and even for arbitrary adaptive auditing. It identifies the worst-case dependence on the clean-label budget at fixed dimension. The robustness–adaptation bound concerns the more specific capped uniform-audit protocol: for $d \ge 2$, a worst-case guarantee G implies regret $\Omega(T^2/(BG))$ on some zero-bias instance, in the stated regime $0 < G \le T/16$. This limitation concerns certification of an unknown benign proxy and does not extend automatically to arbitrary adaptive auditing.

Together, these results characterize shared linear calibration, a pathwise audit budget, and the limits of adaptation under uniform auditing. They complement covariance-adaptive estimation and contextual trust learning. A possible further direction is to combine learned covariance weighting with linear calibration, or to combine trust selection across evaluators with numerical correction within each selected source; either combination would require its own analysis.