LEARNING WITH COUPLED UNCERTAINTY

Anonymous authors
Paper under double-blind review

ABSTRACT

We initiate the study of decision-making under coupled uncertainties. In this problem, a learner has access to ground truth and coarse measurements of outcomes and would like to use them for decision-making. The learner has constrained access to ground truth measurements for only a given fraction of decision outcomes and would like to leverage the cheaper coarse measurements of decision outcomes. We introduce a model where the randomness of the ground and coarse measurements is coupled, and our approach learns their correlation to optimally combine coarse measurements with ground truth and achieve improved performance. This framework unifies several settings, like learning from multi-fidelity data sources and delegating decision-making to AI agents. We provide an upper confidence bounds based algorithm CUUCB for leveraging coupled uncertainties in a multi-armed bandit task, where the covariance structure between coarse measurements and ground truth is unknown. We show theoretically how CUUCB adapts to the underlying covariance structure by deriving instance-dependent and instance-independent regret bounds. We validate our algorithm in two experiments: a task with synthetically generated data, and an LLM benchmarking task. We compare our algorithm to existing UCB variants with access to only ground truth measurements on the constrained fraction of outcomes. In both cases, our algorithm is able to achieve lower regret.

1 INTRODUCTION

Recent advances in artificial intelligence (AI), computational modeling, and the development of foundation models for science are reshaping the way experiments are conducted — accelerating the pace of new discoveries Bommasani et al. (2021); Wang et al. (2023). Predictive tools and language models are increasingly being developed in domains like personalized healthcare Lu et al. (2021); Yang et al. (2022); Thirunavukarasu et al. (2023), polling Tumasjan et al. (2010); Wang et al. (2015), environmental monitoring Shi et al. (2017); Ravuri et al. (2021), and content moderation Gomez et al. (2024) paving the way for new services and more efficiencies in important application areas. Despite these advances, these powerful predictive models are often *black-box* which introduces challenges when one would like to use them for downstream decision-making tasks. For example, language models may be mis-aligned with the underlying human preferences one would like them to emulate. Despite a large body of recent work on uncertainty quantification for black-box machine learning predictions (the exemplar being conformal prediction Vovk et al. (2005); Romano et al. (2019); Angelopoulos et al. (2023b)), existing methods often yield overly conservative estimates or are simply not amenable to online decision-making.

To overcome this challenge, we propose an algorithm for sequential decision-making when one has access to two different types of measurements: (1) coarse measurements (e.g., machine learning or AI-based predictions) which may be cheap and readily available, and (2) ground-truth measurements that may require costly experimentation. The decision-maker does not know how good the coarse measurement is but they would like to minimize their regret over a sequence of decisions while keeping the number of "ground-truth" measurements they query below a desired threshold. To tackle this, we introduce a feedback model where (1) the randomness in the coarse and ground truth measurements is *coupled*, and (2) each time a ground-truth measurement is queried, the corresponding coarse measurement is also observed at negligible cost. This framework captures feedback mechanisms that naturally arise in a variety of empirical settings, including fluid dynamics modeling, climate forecasting and active learning epidemiology systems Niu et al. (2024); Wu et al. (2023; 2022); Li et al. (2022a; 2020).

The key premise of *coupled uncertainty learning* is that if the coarse measurements are highly correlated with the ground-truth measurements, then one can use them to reduce the variance of the estimators obtained from only ground-truth measurements and ultimately to substitute for ground truth measurements altogether. Doing so however, is not straightforward since the covariance structure of the ground truth and coarse measurements may not be known a priori, and consequently, the weight that should be given to the coarse measurements in estimation is unknown. Furthermore, high correlation between ground truth and coarse data is not enough to simply replace one dataset for another since it does not solve the problem of *mis-alignment*—a problem that our experiments on large language models clearly expose.

In this paper, we initiate a study of *learning* with coupled uncertainties in the context of multi-armed bandits. We seek to learn the optimal arm while building estimators that combine ground truth and coarse data to achieve lower regret *with respect to a ground-truth baseline* than existing algorithms. Reflecting the fact that in many applications ground truth feedback can be expensive to collect (e.g., querying human feedback in language-model benchmarking tasks can be costly), we constrain the number of ground truth queries to be an $\alpha(i)$ fraction of the total number of pulls for a given arm i. This ratio can arise from the relative costs of ground truth vs coarse experimentation.

We propose the Coupled Uncertainty Upper Confidence Bounds (CUUCB) algorithm, which can be seen as a variant of the UCB-V algorithm (Audibert et al., 2009), which is a variance adaptive upper confidence bounds algorithm. We show that CUUCB achieves an instance-independent expected regret bound of $\mathcal{O}(\sqrt{C_\alpha T \log T})$ while only requiring approximately $\alpha(i)$ fraction of pulls for arm i being ground-truth samples. Here, $\alpha(i)$ is a user-defined parameter and can be interpreted as the target fraction of pulls of arm i which access ground truth. C_α depends on the covariance structure of the ground-truth and coarse measurements α and has the following structure:

$$C_\alpha = \sum_i \mathrm{Var}^{\mathrm{G}}(i) \left(\frac{(1-\rho(i)^2)}{\alpha(i)} + \rho(i)^2 \right) \tag{1}$$

where $\mathrm{Var}^{\mathrm{G}}(i)$ is the variance of the ground-truth measurements for arm i and $\rho(i) \in [-1, 1]$ is the correlation between ground-truth and coarse measurements for arm i. This can provide improved performance compared to using a variance adaptive algorithm such as UCB-V (Audibert et al., 2009), when there is no access to coarse measurements and the regret bound has the form $\mathcal{O}(\sqrt{C'_\alpha T \log T})$ with $C'_\alpha = \sum_i \frac{\mathrm{Var}^{\mathrm{G}}(i)}{\alpha(i)}$. We can thus expect to see improved performance when $|\rho(i)|$ is large, meaning the ground-truth and coarse measurements have a high correlation, and when $\alpha(i)$ is small, meaning only a small fraction of the measurements are ground-truth.

We verify our findings experimentally both on synthetic data and a LLM benchmarking task in which we observe that our algorithm can overcome mis-alignment in the LLM-as-a-judge paradigm using small amounts of ground-truth data.

2 Related Work

Extensive research since Robbins (1952) has studied the multi-armed bandit problem, with particular emphasis on algorithms using upper confidence bounds (Auer, 2002; Bubeck et al., 2012; Lattimore & Szepesvári, 2020). Within this literature, the most relevant to this paper is a line of work on multi-fidelity bandits (Kandasamy et al., 2016b;a) in which the authors consider bandit problems when multi-fidelity information is available. In such problems, only only one type of feedback (or fidelity level) is available at each time step, and the fidelity levels are known. In contrast our coarse measurements are of unknown quality. By introducing the framework of coupled uncertainty bandits, we highlight new way in which multiple fidelity feedback can be leveraged for decision-making. By collecting *both* low and high fidelity feedback when querying the higher fidelity levels, the measurements become statistically coupled. This allows the learner to learn correlations and improve performance over and beyond existing bandit algorithms that use only high fidelity data.

These problems have also surfaced in the literature on adaptive experiment design for scientific applications of machine learning (Fernández-Godino et al., 2016; Greenhill et al., 2020) where the goal is to fuse data sources with multiple fidelity levels and costs—online— to derive good estimates of desired quantities. Previous works have focused on Bayesian optimization and Bayesian active learning problems, aiming to iteratively optimize or learn a target function by leveraging data across different fidelity levels (Li et al., 2020; Hernandez-Garcia et al., 2023; Li et al., 2022a; Kandasamy et al., 2017). One popular approach involves using Gaussian processes (GPs) for multi-fidelity surrogate modeling, providing well-calibrated uncertainty estimates for acquisition function design (Wu et al., 2020). Other multi-fidelity surrogate modeling approaches utilize neural processes (Wu et al., 2022; 2023) or ordinary differential equations (Li et al., 2022b) as an more flexible and scalable alternative to GPs. None of these lines of work consider the potential effects of leveraging the coupling between data sources to reduce uncertainty.

Our work also relates to the emerging literature on using AI predictions, large language models, or other machine learning based models as substitutes for experimentation in estimation problems. In offline settings, these problems have recently emerged under the umbrella of *prediction-powered inference* (Angelopoulos et al., 2023a; Zrnic & Candès, 2024) which provides a generalization of classic estimators studied in statistics like the augmented inverse propensity weighted (AIPW) estimator from (Robins & Rotnitzky, 1995). These works leverage predictive models to improve the efficiency and accuracy of inference, particularly in settings where labeled data is scarce or expensive to obtain. These techniques have been used to deal with problems of mis-alignment and a lack of human data in the fine tuning of large language models (LLMs) (Zhou et al., 2025). Our paper extends some of these ideas to problems of online learning. Concurrent work (Ji et al., 2025) has also extended these ideas to sequential decision-making settings in an attempt to warm start online learning algorithms with coarse data. Our paper looks at a different problem of extracting statistical signal from coarse data while collecting data under a budget constraint. Furthermore, our theoretical results and derivations gives finite-sample bounds on estimators under a general boundedness assumption that makes the derivations nontrivial. These expand upon the more asymptotic analyses present in the literature on prediction-powered inference and may be of independent interest.

3 The Coupled Uncertainty Framework

We start with formulating the paradigm of decision-making with coupled uncertainty. In this setting, we assume that there is a sequence of contexts $\{c_t\}_{t=1}^{\infty}$ arriving into the system. Given a context, c_t at time t, the system chooses an action i_t from a set of actions $\mathcal{A}$, generating an outcome $y_t(c_t, i_t)$. These contexts and actions might be (patient, treatment) pairs in medical scenarios or (prompt,LLM) pairs in language model triage tasks. Importantly, we assume that outcomes can be measured with different levels of granularity. For simplicity of exposition, we assume that there are two such levels of measurement, one coarse and one ground-truth against which we would like to be benchmarked.

At time t, we formally denote $r_t^C(y)$ as the coarse measurement we receive given the outcome y, and $r_t^G(y)$ as the ground truth one. To further concretize this setup we present a motivating example based on benchmarking large language model that we experimentally validate our algorithm on in Section 6. We remark that many other use cases including personalized medicine, scientific discovery, content moderation, and polling fit into this framework.

Example 3.1 (LLM-as-a-judge). *A number of existing platforms benchmark and rank the performance of large language models (LLMs). This is then used to either keep leaderboards of language model performance (Chiang et al., 2024), or ultimately to triage prompts to optimal models online (Ong et al., 2025), i.e., to perform model routing. The state-of-the-art approach is to allow LLM users to compare and rate the LLM outputs, and match decisions to those made by humans. However, high-quality human-level data is scarce, relative to the categories of the prompts we are interested in and asking people to provide data can be costly. Therefore many researchers make use of language models to facilitate and provide AI feedback. This raises the question of whether the AI feedback is comparable or* aligned *to human preferences and also when and how to make use of it. Such problems fit neatly into our framework: one would like to make use of LLMs as judges of quality (i.e., coarse feedback) and use them to replace human-feedback (i.e., the ground truth). At each round t, given the prompt c_t and the LLM choice i_t, the benchmarking algorithm can either query just the coarse level AI feedback $r_t^C(y_t)$ and save human effort, or query both the ground truth human feedback and the AI feedback as a tuple $[r_t^G(y_t), r_t^C(y_t)]$. Existing LLM-as-a-judge Gu et al. (2025) approaches can easily fall prey to mis-aligned models. As we observe experimentally, incorporating even small amounts of ground truth data (5%) allows us to overcome this problem.*

In many examples, we observe that when the decision makers have access to fine-grained ground truth data $r_t^G(y)$, the coarse level measurement $r_t^C(y)$ with the same input y is cheap to obtain. We therefore assume in our protocol that when the learner opts to observe a ground truth measurement, they receive a two-dimensional tuple $[r_t^{\mathrm{G}}(y), r_t^{\mathrm{C}}(y)]$. Otherwise, they receive a single coarse reward value $r_t^{\mathrm{C}}(y)$. Our framework assumes that the context c arises from some underlying distribution and treats it as a random variable that causes coupling of the ground truth and coarse measurements. Thus, the distribution of $[r_t^{\mathrm{G}}(i,c), r_t^{\mathrm{C}}(i,c)]$ depends only on the learner's action i as c is a random variable. We thus omit the context c in the following discussion since it is understood to be part of the randomness. We remark that a natural extension of our framing would be to consider context-dependent algorithms, though we leave such problems to future work.

In the remainder of the paper we focus on multi-armed bandit problems and demonstrate in our analyses that we can utilize the coupling between the measurements to obtain tighter concentration bounds for the means which ultimately resulting in performance gains.

4 Multi-Armed Bandits with Coupled Uncertainties

In our analysis, we study the multi-armed bandit problem where at time t the learner plays an action $i_t \in \mathcal{A} = [K]$. The learner also has a choice of what type of reward measurement m_t to observe at time t, a coarse and a ground truth measurement ($m_t \in \{\mathrm{C}, \mathrm{G}\}$). We assume the existence of joint measurement vectors $r_t^{\mathrm{C}}, r_t^{\mathrm{G}} \in \mathbb{R}^K$ such that the conditional mean reward values satisfy

$$\mathbb{E}[r_t^{\mathrm{C}}(i)] = \mu^{\mathrm{C}}(i), \qquad \mathbb{E}[r_t^{\mathrm{G}}(i)] = \mu^{\mathrm{G}}(i)$$

In our protocol when the learner opts to observe a ground truth measurement, she receives a two-dimensional tuple $[r_t^{\mathrm{G}}(i_t), r_t^{\mathrm{C}}(i_t)]$. Otherwise she receives a single coarse reward value $r_t^{\mathrm{C}}(i_t)$. We remark that the ground truth and coarse means are not necessarily related to one another, and may in fact encode different orderings of arms. This is a setting which we call *mis-alignment*: the coarse means implicitly encode different reward preferences than the ground truth. We do however, assume that the two measurements are correlated. We denote the *unknown* variance of the coarse and ground truth data $[r^{\mathrm{G}}(i),\ r^{\mathrm{C}}(i)]$ as well as their *unknown* covariance as $\mathrm{Var}^{\mathrm{C}}(i), \mathrm{Var}^{\mathrm{G}}(i)$, and $\mathrm{Cov}(i)$ respectively.

A nonzero covariance between measurements (given an arm) ensures that there is statistical signal to extract from coarse data which is informative for explaining away some of the variation in the ground truth measurements. The correlation can be quantified by the correlation coefficient: $\rho(i) = \frac{\mathrm{Cov}(i)}{\sqrt{\mathrm{Var}^{\mathrm{G}}(i)\mathrm{Var}^{\mathrm{C}}(i)}} \in [-1, 1]$. Given these definitions, we make two key assumptions on these random variables. The first is a common boundedness assumption.

Assumption 4.1. *There exists a* known *value $b > 0$ such that $r^{\mathrm{C}}(i), r^{\mathrm{G}}(i) \in [0, b]$ almost surely, for all $i \in [K]$.*

The second assumption is of knoweldge of an upper bound on the ratio of $\mathrm{Var}^{\mathrm{G}}$ and $\mathrm{Var}^{\mathrm{C}}$.

Assumption 4.2. *There exists a known $\gamma > 0$, such that $\mathrm{Var}^{\mathrm{G}}(i) \leqslant \gamma^2 \mathrm{Var}^{\mathrm{C}}(i)$ for all $i \in [K]$.*

This assumption is also natural, since when $\gamma = 1$, this parameter simply reflects the (natural) setting in which the coarse measurements are "more stochastic" measurements of the ground truth.

Given this setup, we assume the learner seeks to minimize their expected regret when compared to the optimal (ground-truth) arm.

$$\mathrm{Regret}(T) = \sum_{t=1}^{T} \max_{i \in [K]} \mu^{\mathrm{G}}(i) - \mu^{\mathrm{G}}(i_t)$$

Running a no-regret algorithm such as UCB (Lai & Robbins, 1985) and setting the measurement type to ground truth at every time step (i.e. $m_t = \mathrm{G}$ for all $t \in \mathbb{N}$) achieves an expected regret bound of order $\mathcal{O}(\sqrt{KT \log T})$. Unfortunately, this would incur a cost of T expensive ground measurements. In our problem, however, ground truth measurements are comparatively expensive to collect, and the user attempts to limit their usage to a user-specified ratio $\alpha(i)$ of pulls, where $\alpha(i) = 1$ corresponds to an algorithm using only ground truth measurements, and $\alpha(i) = 0$ an algorithm using only coarse measurements.

We define $\mathcal{T}_t^{\mathrm{G}}(i)$ as the set of time-steps $\ell \in [t]$ where $m_\ell = \mathrm{G}$ and $i_\ell = i$ similarly $\mathcal{T}_t^{\mathrm{C}}(i)$ is the set of time-steps $\ell \in [t]$ where $m_\ell = \mathrm{C}$ and $i_\ell = i$. With this notation, we can represent the number of ground and coarse pulls for arm i at time t as $N_t^{\mathrm{G}}(i) = |\mathcal{T}_t^{\mathrm{G}}(i)|$ and $N_t^{\mathrm{C}}(i) = |\mathcal{T}_t^{\mathrm{C}}(i)|$. The total number of pulls for arm i is denoted $N_t(i) = N_t^{\mathrm{G}}(i) + N_t^{\mathrm{C}}(i)$.

To keep the actual ratio $\alpha_t(i) := \frac{N_t^{\mathrm{G}}(i)}{N_t(i)}$ close to the desired $\alpha(i) \in (0,1)$, we develop a simple thresholding rule:

$$m_t = \begin{cases} \mathrm{G} & : \quad N_{t-1}^{\mathrm{G}}(i_t) \leqslant \alpha(i) N_{t-1}(i_t) \\ \mathrm{C} & : \quad \text{otherwise} \end{cases} . \tag{2}$$

For convenience, we define the ratio $x_t(i) := \frac{N_t^{\mathrm{G}}(i)}{N_t^{\mathrm{C}}(i)} = \frac{\alpha_t(i)}{1-\alpha_t(i)}$, i.e., the ratio of ground to coarse pulls for the arm i. Then the target value of $x_t(i)$ is given by $x(i) := \frac{\alpha(i)}{1-\alpha(i)}$. The following result guarantees that this decision rule controls $\alpha_t(i)$ around $\alpha(i)$. The proof is provided in Appendix B.

Lemma 4.1 (Feedback ratio guarantee)**.** *The threshold rule* (2) *guarantees the following bounds on the ratio $\alpha_t(i)$ for each arm $i \in [K]$ and all times t:*

$$\alpha(i)\left(1 - \frac{1}{N_t(i)}\right) \leqslant \alpha_t(i) \leqslant \alpha(i)\left(1 - \frac{1}{N_t(i)}\right) + \frac{1}{N_t(i)}$$

In the next section we combine this rule with statistical estimators for the ground truth means to derive a new multi-armed bandit algorithm: Coupled Uncertainty Upper Confidence Bounds (CUUCB).

5 Coupled Uncertainty Upper Confidence Bounds

In this section we define and analyze the statistical estimators used in the CUUCB algorithm. These estimators exploit the coupling between the ground truth and coarse data and have tighter confidence intervals which scale with the quantity C_α (see (1)) as opposed to the quantity C'_α which would be possible using ground truth measurements alone. As we show, using these estimators in an upper confidence bounds algorithm results in lower regret on the multi-armed bandit problem.

5.1 Unbiased and Reduced Variance Estimators of the Ground Truth Mean

To begin constructing our estimators, recall that at time t, our algorithm CUUCB pulls arm i_t and then selects a measurement type $m_t \in \{\mathrm{C}, \mathrm{G}\}$ after which it observes feedback in the form of either a coarse measurement from arm i_t or a tuple of coarse/ground-truth measurements. Given this feedback, we define empirical estimators of $\mu^{\mathrm{G}}(i)$ and $\mu^{\mathrm{C}}(i)$ at time t given by

$$\mu_t^X(i) := \frac{1}{N_t^X(i)} \sum_{j \in \mathcal{T}_t^X(i)} r_j^X(i), \; X = \mathrm{G}, \mathrm{C}.$$

We also define the estimator for $\mu^{\mathrm{C}}(i)$ which is derived solely from data collected from coupled measurements $\mu_t^{\mathrm{C,G}}(i)$:

$$\mu_t^{\mathrm{C,G}}(i) := \frac{1}{N_t^{\mathrm{G}}(i)} \sum_{j \in \mathcal{T}_t^{\mathrm{G}}(i)} r_j^{\mathrm{C}}(i).$$

Our estimator for the ground truth mean $\mu^{\mathrm{G}}(i)$, linearly combines $\mu_t^{\mathrm{G}}(i)$, $\mu_t^{\mathrm{C,G}}(i)$ and $\mu_t^{\mathrm{C}}(i)$ to for variance reduction as follows:

$$Z_t^\lambda(i) := \mu_t^{\mathrm{G}}(i) - \lambda(\mu_t^{\mathrm{C,G}}(i) - \mu_t^{\mathrm{C}}(i)),$$

for some constant $\lambda \in \mathbb{R}$. Note that $Z_t^\lambda(i)$ is an unbiased estimator of $\mu^{\mathrm{G}}(i)$ for any fixed λ. This estimator is simply the augmented inverse propensity weighted (AIPW) estimator first studied in Robins & Rotnitzky (1995) and more recently in Angelopoulos et al. (2024) which studied the tuning of λ to reduce variance and thus the width of asymptotic confidence intervals. We derive finite-sample confidence intervals in the case of unknown variances and correlations.

Algorithm 1 Coupled Uncertainty Upper Confidence Bounds (CUUCB)

```
Input: $\delta \in (0,1), \alpha \in (0,1)^K, T$.
Initialize estimators $\{\mu_0^X(i), \mathrm{Var}_0^X(i), \mathrm{Cov}_0(i)\}_{i\in[K], X\in\{\mathrm{G},\mathrm{C},(\mathrm{C},\mathrm{G})\}}$
Initialize $N_0^{\mathrm{G}}(i) = N_0^{\mathrm{C}}(i) = 0$ for $i \in [K]$.
Initialize UCB bounds $U_0(i) = b$ for $i \in [K]$.
for $t = 1, 2, \cdots, T$ do
    Select arm $i_t = \arg\max_{i\in[K]} U_{t-1}(i)$.
    Select feedback according to rule (2).
    Update count $N_t^{m_t}(i_t) = N_{t-1}^{m_t}(i_t) + 1$
    Update estimators $\mu_t^{\mathrm{G}}(i_t), \mu_t^{\mathrm{C,G}}(i_t), \mathrm{Var}_t^{\mathrm{G}}(i_t), \mathrm{Var}_t^{\mathrm{C,G}}(i_t)$,$\mu_t^{\mathrm{C}}(i_t), \mathrm{Var}_t^{\mathrm{C}}(i_t)$.
    Compute $Z_t(i_t), \hat{\mathrm{V}}_t^{\lambda_t(i_t)}(i_t)$ and UCB bound $U_t(i_t)$.
end for
```

The main idea is to exploit the coupling between the ground and coarse rewards to choose λ resulting in a Z_t^λ having a lower variance as compared to just using μ_t^{G} as the mean estimator. To see that this is possible, we use the fact that $(\mu_t^{\mathrm{G}}(i) - \lambda\mu_t^{\mathrm{C}}(i))$ and $\lambda\hat{\mu}_t^{\mathrm{C}}(i)$ are independent to obtain the expression $\mathrm{Var}(Z_t^\lambda(i)) = \frac{\mathrm{V}_t^\lambda(i)}{N_t^{\mathrm{G}}(i)}$, where $\mathrm{V}_t^\lambda(i)$ is the following quadratic expression in λ:

$$\mathrm{V}_t^\lambda(i) = (1 + x_t(i))\mathrm{Var}^{\mathrm{C}}(i)\lambda^2 - 2\mathrm{Cov}(i)\lambda + \mathrm{Var}^{\mathrm{G}}(i).$$

One then observes that the choice of λ that minimizes $\mathrm{V}_t^\lambda(i)$ is given by $\lambda_t^*(i) = \frac{(1-\alpha_t(i))\mathrm{Cov}(i)}{\mathrm{Var}^{\mathrm{C}}(i)}$, which results in $Z_t^\lambda(i)$ having variance $\frac{\mathrm{V}_t^*(i)}{N_t^{\mathrm{G}}(i)}$, where

$$\mathrm{V}_t^*(i) = \mathrm{Var}^{\mathrm{G}}(i) - \frac{(1-\alpha_t(i))\mathrm{Cov}(i)^2}{\mathrm{Var}^{\mathrm{C}}(i)} = \mathrm{Var}^{\mathrm{G}}(i)\left(1 - (1-\alpha_t(i))\rho(i)^2\right).$$

Note that by assumption 4.2, we have $|\lambda_t^*(i)| \leqslant (1-\alpha_t(i))\gamma$. We remark that $\mathrm{V}_t^*(i) \leqslant \mathrm{Var}^{\mathrm{G}}(i)$ for all values of α and ρ which means variance reduction is obtained by using the estimator $Z_t^{\lambda_t^*(i)}(i)$ in place of $\mu_t^{\mathrm{G}}(i)$. Through Bennett's concentration inequality we immediately observe how one can exploit this variance reduction to achieve an improved upper confidence bound for $\mu^{\mathrm{G}}(i)$.

Lemma 5.1 (Bennett's Inequality)**.** *Given an arm i and fixing the number of its pulls $N_t^{\mathrm{C}}(i), N_t^{\mathrm{G}}(i)$, consider a fixed $\lambda \in \mathbb{R}$ satisfying $|\lambda| \leqslant (1-\alpha_t(i))\gamma$. Then the following concentration inequality holds with probability at least $1-\delta$:*

$$|Z_t^\lambda(i) - \mu^{\mathrm{G}}(i)| \leqslant \sqrt{\frac{2\mathrm{V}_t^\lambda(i)\log\left(\frac{2}{\delta}\right)}{N_t^{\mathrm{G}}(i)}} + \frac{(1+\gamma)b\log\left(\frac{2}{\delta}\right)}{3N_t^{\mathrm{G}}(i)}.$$

The proof of this follows from a simple application of Bennett's concentration inequality and is supplied in Appendix B.1 for completeness.. Lemma 5.1 gives us a concentration radius with the leading term dependent on $\mathrm{V}_t^\lambda(i)$. Thus, picking $\lambda = \lambda_t^*(i)$ would allow us to obtain the tightest concentration and thus a lower regret algorithm. However, we do not know $\Sigma(i)$ and thus cannot directly compute $\mathrm{V}_t^\lambda(i)$ or $\lambda_t^*(i)$. As we show, one can derive a version of this bound using empirical estimators of these quantities.

5.2 Empirical Estimation of Variances and Covariances

To fully exploit our AIPW estimator without knowledge of the variances or correlations between ground truth and coarse measurements, we introduce empirical estimators of

$\mathrm{Var}^{\mathrm{G}}(i), \mathrm{Var}^{\mathrm{C}}(i), \mathrm{Var}_t^{\mathrm{C,G}}(i)$ and $\mathrm{Cov}(i)$ at time t as:

$$\begin{aligned}
\mathrm{Var}_t^{X}(i) &:= \frac{1}{N_t^{X}(i)} \sum_{j \in \mathcal{T}_t^{X}(i)} (r_j^{X}(i) - \mu_t^{X}(i))^2, \; X = \mathrm{G}, \mathrm{C} \\
\mathrm{Var}_t^{\mathrm{C,G}}(i) &:= \frac{1}{N_t^{\mathrm{G}}(i)} \sum_{j \in \mathcal{T}_t^{\mathrm{G}}(i)} (r_j^{\mathrm{C}}(i) - \mu_t^{\mathrm{C,G}}(i))^2 \\
\mathrm{Cov}_t(i) &:= \frac{1}{N_t^{\mathrm{G}}(i)} \sum_{j \in \mathcal{T}_t^{\mathrm{G}}(i)} (r_j^{\mathrm{G}}(i) - \mu_t^{\mathrm{G}}(i))(r_j^{\mathrm{C}}(i) - \mu_t^{\mathrm{C}}(i)).
\end{aligned}$$

Note that we can only compute the covariance estimator due to the coupling between the two sources of feedback given an outcome. These definitions allow us to define the following empirical estimator for $\mathrm{V}_t^{\lambda}(i)$ as a function of λ:

$$\hat{\mathrm{V}}_t^{\lambda}(i) = (1 + x_t(i))\widehat{\mathrm{Var}}_t^{\mathrm{C}}(i)\lambda^2 - 2\mathrm{Cov}_t(i)\lambda + \mathrm{Var}_t^{\mathrm{G}}(i),$$

where for convenience we define $\widehat{\mathrm{Var}}_t^{\mathrm{C}}(i) := (1 - \alpha_t(i))\mathrm{Var}_t^{\mathrm{C,G}}(i) + \alpha_t(i)\mathrm{Var}_t^{\mathrm{C}}(i)$. The expression for $\hat{\mathrm{V}}_t^{\lambda}$ can be obtained by simply replacing the variances and covariance in the expression for V_t^{λ} with their appropriate empirical estimators. We analyze its concentration around $V_t^{\lambda}(i)$ in Appendix B.2. We also define an empirical estimator of $\lambda_t^*(i)$:

$$\begin{aligned}
\lambda_t(i) &:= \underset{\lambda \in [-(1-\alpha_t(i))\gamma, (1-\alpha_t(i))\gamma]}{\arg\min} \hat{\mathrm{V}}_t^{\lambda}(i) \\
&= (1 - \alpha_t(i))\mathcal{P}_\gamma\left(\frac{\mathrm{Cov}_t(i)}{\widehat{\mathrm{Var}}_t^{\mathrm{C}}(i)}\right) \quad \text{where: } \mathcal{P}_\gamma(x) := \begin{cases} \gamma, \text{ if } x > \gamma \\ x, \text{ if } x \in [-\gamma, \gamma] \\ -\gamma, \text{ if } x < -\gamma \end{cases}.
\end{aligned}$$

Note that the bounds on γ in the minimization follow from assumption 4.2. Plugging-in the estimator $\lambda_t(i)$ into $Z_t^{\lambda}(i)$, we obtain the following estimator for $\mu_t^{\mathrm{G}}(i)$:

$$Z_t(i) := Z_t^{\lambda_t(i)}(i) = \mu_t^{\mathrm{G}}(i) - \lambda_t(i)(\mu_t^{\mathrm{C,G}}(i) - \mu_t^{\mathrm{C}}(i)).$$

The following lemma, which is an empirical analogue to Lemma 5.1 gives us the concentration of $Z_t(i)$ around $\mu^{\mathrm{G}}(i)$ with a confidence bound that scales with the empirical variance.

Lemma 5.2 (Anytime concentration of $Z_t(i)$)**.** *Consider a fixed arm $i \in [K]$ and some $\delta \in (0, 1)$. Define $L_t(i) := \log\left(16\delta^{-1}(N_t(i))^2(N_t(i) + 2)\right)$. Then w.p. at least $1 - \delta$, the following inequalities hold for all times t:*

$$|Z_t(i) - \mu^{\mathrm{G}}(i)| \leqslant \sqrt{\frac{2\hat{\mathrm{V}}_t^{\lambda_t(i)}(i)L_t(i)}{N_t^{\mathrm{G}}(i)}} + \frac{6(1+\gamma)bL_t(i)}{N_t^{\mathrm{G}}(i)} \leqslant \sqrt{\frac{2\mathrm{V}_t^{*}(i)L_t(i)}{N_t^{\mathrm{G}}(i)}} + \frac{7(1+\gamma)bL_t(i)}{N_t^{\mathrm{G}}(i)}.$$

The proof of Lemma 5.2 is deferred to Appendix B.2. The proof follows from showing the concentration of the empirical variance and covariance estimators and a careful covering argument to handle the fact that $\lambda_t(i)$ is itself a stochastic quantity.

We use Lemma 5.2 to define an Upper Confidence Bound on the true mean $\mu^{\mathrm{G}}(i)$ of arm i at time t:

$$U_t(i) := Z_t(i) + \sqrt{\frac{2\hat{\mathrm{V}}_t^{\lambda_t(i)}(i)L_t(i)}{N_t^{\mathrm{G}}(i)}} + \frac{6(1+\gamma)bL_t(i)}{N_t^{\mathrm{G}}(i)} \tag{3}$$

Using this confidence interval in conjunction with the thresholding rule for maintaining the desired ratio $\alpha(i)$ of ground truth to coarse feedback results in the CUUCB algorithm detailed in Algorithm 1.

5.3 Regret Analysis of CUUCB

Given our the estimator and its confidence bounds, we now present instance-dependent and instance-independent expected regret bounds for CUUCB, the proof of which is deferred to Appendix B.4 and follows from our concentration guarantees for our variance reduced estimator.

Theorem 5.3 (Regret Bounds). *Let $\delta = \frac{1}{T}$. Then, the expected regret of* CUUCB *satisfies the instance-dependent bound:*

$$\mathbb{E}[\mathrm{Regret}(T)] \leqslant 128 \sum_{i:\Delta_i>0} \frac{1}{\alpha(i)} \left(\frac{\mathrm{V}^*(i)}{\Delta_i} + b(1+\gamma)\right) \log(2T) + \mathcal{O}(1),$$

where $\Delta_i := \max_{j\in[K]} \mu^{\mathrm{G}}(j) - \mu^{\mathrm{G}}(i)$ is the loss of playing arm i and $\mathrm{V}^*(i) := \mathrm{Var}^{\mathrm{G}}(i) - \frac{(1-\alpha(i))\mathrm{Cov}(i)^2}{\mathrm{Var}^{\mathrm{C}}(i)}$. Moreover, the expected regret also satisfies the instance-independent bound:

$$\mathbb{E}[\mathrm{Regret}(T)] \leqslant 23\sqrt{\left(\sum_i \frac{\mathrm{V}^*(i)}{\alpha(i)}\right) T \log(2T) + \mathcal{O}(\log(T))}$$

In the above result, one can think of $\mathrm{V}^*(i)$ as the limiting value of $\mathrm{V}_t^*(i)$ as the number of samples grows large, resulting in the ratio $\alpha_t(i)$ going to $\alpha(i)$. Lemma 5.3 gives us an instance-independent bound on the regret of order $\mathcal{O}(\sqrt{C_\alpha T \log T})$, where

$$C_\alpha = \sum_i \frac{\mathrm{V}^*(i)}{\alpha(i)} = \sum_i \mathrm{Var}^{\mathrm{G}}(i) \left(\frac{(1-\rho(i)^2)}{\alpha(i)} + \rho(i)^2\right)$$

On the other hand, if we were to use a classical variance adaptive algorithm like UCB-V, which has access to only the $\alpha(i)$ fraction of ground truth measurements but has to play for the same T steps, then the expected regret bound will have the order $\mathcal{O}(\sqrt{C'_\alpha T \log T})$ with $C'_\alpha = \sum_i \frac{\mathrm{Var}^{\mathrm{G}}(i)}{\alpha(i)}$ as shown in Lemma B.3. Thus, we see that C_α has a reduction compared to C'_α due to the $\rho(i)^2$ term. We see that the reduction obtained is greater if $|\rho(i)|$ becomes larger and $\alpha(i)$ becomes smaller, i.e., as the ground and coarse samples become more correlated and as ground samples become more scarce. Thus, CUUCB potentially allows us to exploit the additional coarse samples to obtain a reduction in regret. We look at this behaviour in more detail in the experiments that follow.

6 Experiments

We now evaluate our algorithms and theory on synthetic data and an LLM-benchmarking task. For details on implementation, dataset generation and additional experimental results, see Appendix C.

6.1 Synthetic Experiments

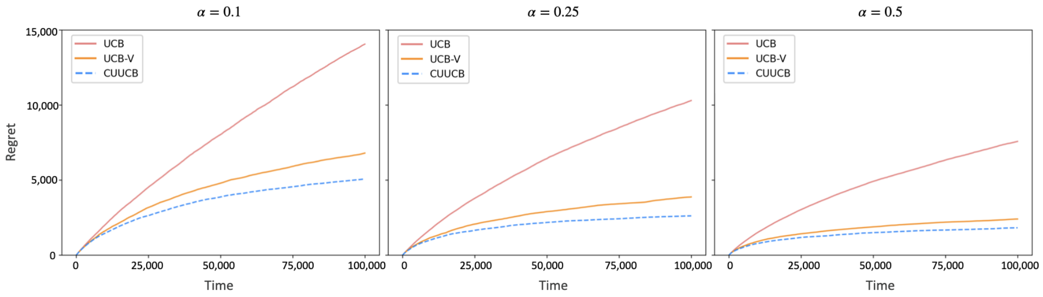

Figure 1: Comparison of CUUCB, UCB and UCB-V on synthetically generated data for fixed value of $\rho = 0.8$ and different values of α with a time horizon of $100,000$ steps, averaged across 40 runs.

In our first experiment, we test CUUCB on synthetically generated data and compare it to UCB and UCB-V. We keep the number of arms as $K = 4$ and generate the coupled ground and coarse data points for all arms i by sampling from a common 2-D jointly Gaussian distribution with correlation set to a parameter ρ. We then clip the values in a desired interval to ensure the data distribution has finite support. We choose a uniform target ratio $\alpha(i) = \alpha$ for all arms i. The algorithms UCB and UCB-V make decisions using only the ground truth samples observed at a given time t, i.e., ground truth samples from time steps $\mathcal{T}_t^{\mathrm{G}}(i)$. Per our theory, we expect CUUCB to outperform the other algorithms if we pick the correlation ρ to be large and set α to be smaller. We compare the algorithms when $\rho = 0.8$ is fixed and we vary α as seen in Figure 1. We run the experiments for a time horizon $T = 100,000$ and for each algorithm, and show the average regret across 40 runs. We observe that while increasing $\alpha(i)$ leads to an overall reduction in regret for all algorithms, the performance improvement obtained by CUUCB over $\mathrm{UCB\text{-}V}$ is largest at smaller α. Additional results in the setting when α is fixed and ρ is varied can be found in Appendix C.4.1.

6.2 LLM-as-a-Judge Experiments

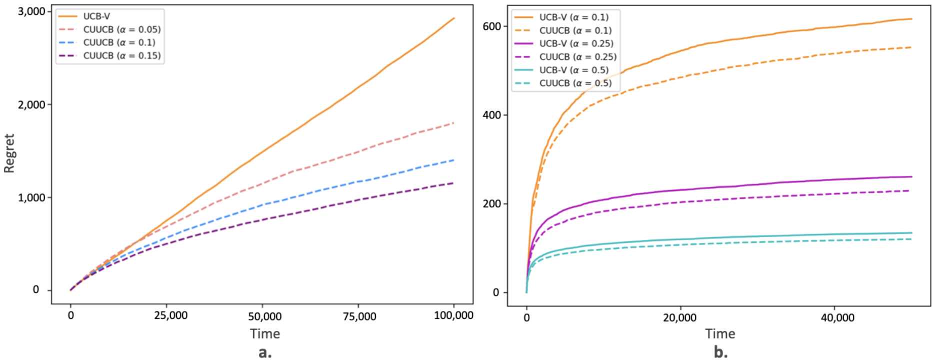

Figure 2: LLM benchmarking regret plots: a. Performance on misaligned Dataset-A of $\mathrm{UCB\text{-}V}$ which makes decisions using only coarse measurements vs. CUUCB for different α with $T = 100,000$, averaged across 40 runs. b. Performance on Dataset-B of $\mathrm{UCB\text{-}V}$ which has access to only ground truth samples vs. CUUCB for different α with $T = 50,000$, averaged across 40 runs.

In our second experiment, we study a task of trying to find the best language model (LLM) amongst a set of large language models for a given set of prompts. Since the idea of a ”best” model can be highly subjective, this task is normally accomplished by asking people to rank or score (prompt, answer) pairs from each LLM to find the best model Chiang et al. (2024). This approach requires potentially massive amounts of human annotations which can be too expensive to try in practice. To overcome this problem, one emerging approach to LLM benchmarking is to simply ask a holdout LLM to supply the scores—an emerging form of learning from AI feedback Zhu et al. (2024) also increasingly called the LLM-as-a-judge paradigm Gu et al. (2025). Crucially—and as we show in our experiments— one cannot simply rely on *only* AI feedback as the AI feedback can be misaligned and fail to correctly capture human preferences. As such, a combination of human and AI feedback is required to accurately benchmark LLMs to human preferences.

This problem fits neatly into our framework: the human annotations take the place of our ground-truth measurements against which the algorithm's performance will ultimately be benchmarked, while the large language model annotations become the coarse measurement.

To simulate this problem we make use of the Nectar dataset of rankings of (prompt,answer) pairs from Zhu et al. (2024). This dataset comprises of 182,954 prompts, each answered by seven LLMs. These answers are then scored by GPT-4 Bubeck et al. (2023). To design this experiment, we extract (prompt, answer) pairs for six chosen models from this dataset and use a pair of LLMs to score the models. We treat the scores given by the more advanced LLM as the ground-truth (i.e., proxies for human labels) and those by the less advanced LLM as coarse feedback. By querying different LLMs, we obtain two datasets–Dataset-A and Dataset-B, consisting of ground-truth and coarse scores of the six models. The dataset generation details can be found in Appendix C.3.

Given this setup, our experiment proceeds in two parts. First, we observe that the ground-truth and coarse scores in Dataset-A are misaligned, i.e., the model with the best mean score differs for the two models. Thus, an attempt to use a MAB algorithm like $\mathrm{UCB\text{-}V}$ with only coarse feedback leads to linear regret when measured using the ground-truth means. We observe that CUUCB is able to overcome the misalignment while using only a small fraction α of ground-truth samples, as shown in Figure 2a. In the second part, we compare regrets of CUUCB and $\mathrm{UCB\text{-}V}$ on Dataset-B while varying the parameter α as shown in Figure 2b. We consistently observe that our algorithm outperforms and that the advantage is largest when the amount of ground truth data is most restricted.

REPRODUCIBILITY STATEMENT

To ensure reproducibility of our theoretical results, detailed proofs are provided in Appendix A and Appendix B. To ensure our experimental results are reproducible, we include experimental details such as algorithm implementations and dataset generation in Appendix C. Additionally, we have submitted all of our code and important datasets as part of the supplementary materials.

REFERENCES

Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023a.

Anastasios N Angelopoulos, Stephen Bates, et al. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023b.

Anastasios N. Angelopoulos, John C. Duchi, and Tijana Zrnic. Ppi++: Efficient prediction-powered inference, 2024.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009. Algorithmic Learning Theory.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

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

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: an open platform for evaluating llms by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24, 2024.

M Giselle Fernández-Godino, Chanyoung Park, Nam-Ho Kim, and Raphael T Haftka. Review of multi-fidelity models. 2016.

Juan Felipe Gomez, Caio Machado, Lucas Monteiro Paes, and Flavio Calmon. Algorithmic arbitrariness in content moderation. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pp. 2234–2253, 2024.

Stewart Greenhill, Santu Rana, Sunil Gupta, Pratibha Vellanki, and Svetha Venkatesh. Bayesian optimization for adaptive experimental design: A review. *IEEE access*, 8:13937–13948, 2020.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025.

Alex Hernandez-Garcia, Nikita Saxena, Moksh Jain, Cheng-Hao Liu, and Yoshua Bengio. Multi-fidelity active learning with gflownets. 2023.

Wenlong Ji, Yihan Pan, Ruihao Zhu, and Lihua Lei. Multi-armed bandits with machine learning-generated surrogate rewards, 2025.

Kirthevasan Kandasamy, Gautam Dasarathy, Junier B Oliva, Jeff Schneider, and Barnabás Póczos. Gaussian process bandit optimisation with multi-fidelity evaluations. *Advances in neural information processing systems*, 29, 2016a.

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

Kirthevasan Kandasamy, Gautam Dasarathy, Jeff Schneider, and Barnabás Póczos. Multi-fidelity bayesian optimisation with continuous approximations. In *International conference on machine learning*, pp. 1799–1808. PMLR, 2017.

T.L Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Shibo Li, Robert M Kirby, and Shandian Zhe. Deep multi-fidelity active learning of high-dimensional outputs. *arXiv preprint arXiv:2012.00901*, 2020.

Shibo Li, Jeff M Phillips, Xin Yu, Robert Kirby, and Shandian Zhe. Batch multi-fidelity active learning with budget constraints. *Advances in Neural Information Processing Systems*, 35:995–1007, 2022a.

Shibo Li, Zheng Wang, Robert Kirby, and Shandian Zhe. Infinite-fidelity coregionalization for physical simulation. *Advances in Neural Information Processing Systems*, 35:25965–25978, 2022b.

Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, Melissa Zhao, Maha Shady, Jana Lipkova, and Faisal Mahmood. Ai-based pathology predicts origins for cancers of unknown primary. *Nature*, 594(7861):106–110, 2021.

Ruijia Niu, Dongxia Wu, Kai Kim, Yian Ma, Duncan Watson-Parris, and Rose Yu. Multi-fidelity residual neural processes for scalable surrogate modeling. In *Forty-first International Conference on Machine Learning*, 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 from preference data. In *The Thirteenth International Conference on Learning Representations*, 2025.

Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, et al. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878):672–677, 2021.

Herbert Robbins. Some aspects of the sequential design of experiments. 1952.

James M. Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Xingjian Shi, Zhihan Gao, Leonard Lausen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Deep learning for precipitation nowcasting: A benchmark and a new model. *Advances in neural information processing systems*, 30, 2017.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8): 1930–1940, 2023.

Andranik Tumasjan, Timm Sprenger, Philipp Sandner, and Isabell Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the international AAAI conference on web and social media*, volume 4, pp. 178–185, 2010.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

Wei Wang, David Rothschild, Sharad Goel, and Andrew Gelman. Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3):980–991, 2015.

Dongxia Wu, Matteo Chinazzi, Alessandro Vespignani, Yi-An Ma, and Rose Yu. Multi-fidelity hierarchical neural processes. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2029–2038, 2022.

Dongxia Wu, Ruijia Niu, Matteo Chinazzi, Yian Ma, and Rose Yu. Disentangled multi-fidelity deep bayesian active learning. In *International Conference on Machine Learning*, pp. 37624–37634. PMLR, 2023.

Jian Wu, Saul Toscano-Palmerin, Peter I Frazier, and Andrew Gordon Wilson. Practical multi-fidelity bayesian optimization for hyperparameter tuning. In *Uncertainty in Artificial Intelligence*, pp. 788–798. PMLR, 2020.

Yuzhe Yang, Yuan Yuan, Guo Zhang, Hao Wang, Ying-Cong Chen, Yingcheng Liu, Christopher G Tarolli, Daniel Crepeau, Jan Bukartyk, Mithri R Junna, et al. Artificial intelligence-enabled detection and assessment of parkinson's disease using nocturnal breathing signals. *Nature medicine*, 28(10):2207–2215, 2022.

Zhaoyi Zhou, Yuda Song, and Andrea Zanette. Accelerating unbiased LLM evaluation via synthetic feedback. In *Forty-second International Conference on Machine Learning*, 2025.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, Karthik Ganesan, Wei-Lin Chiang, Jian Zhang, and Jiantao Jiao. Starling-7b: Improving helpfulness and harmlessness with RLAIF. In *First Conference on Language Modeling*, 2024.

Tijana Zrnic and Emmanuel J Candès. Cross-prediction-powered inference. *Proceedings of the National Academy of Sciences*, 121(15):e2322083121, 2024.

A Supporting Technical Results

Lemma A.1. *[Adapted from Audibert] Let* $\delta \in (0,1)$ *and* $Z_1, \cdots, Z_n$ *be i.i.d. random variables with values in* $[0, b]$*. Define* $\bar{Z}_n = \frac{1}{n}\sum_{i=1}^n Z_i$, $\mathrm{Var}_n = \frac{1}{n}\sum_{i=1}^n \left(Z_i - \bar{Z}_n\right)^2$ *and* $\mathrm{Var} = \mathbb{E}[(Z - \mathbb{E}[Z])^2]$*. With probability at least* $1-\delta$*,*

$$\sqrt{\mathrm{Var}_n} \leqslant \sqrt{\mathrm{Var}} + b\sqrt{\frac{\log(1/\delta)}{2n}}.$$

We also have with probability at least $1-\delta$ *that:*

$$\sqrt{\mathrm{Var}} \leqslant \sqrt{\mathrm{Var}_n} + 1.8b\sqrt{\frac{\log(3/\delta)}{n}}$$

Proof. First, we prove the high probability upper bound. For this, we note that we can bound:

$$\begin{aligned}\mathrm{Var}_n = \frac{1}{n}\sum_{i=1}^n (Z_i - \bar{Z}_n)^2 &= \frac{1}{n}\sum_{i=1}^n (Z_i - \mathbb{E}[Z])^2 - (\bar{Z}_n - \mathbb{E}[Z])^2 \\ &\leqslant \frac{1}{n}\sum_{i=1}^n (Z_i - \mathbb{E}[Z])^2 \end{aligned} \tag{4}$$

Noting that $\{(Z_i - \mathbb{E}[Z])^2\}_{i=1}^n$ are i.i.d. random variables lying almost surely in $[0, b^2]$, we use Bennet's inequality to obtain with probability at least $1-\delta$ that

$$\frac{1}{n}\sum_{i=1}^n (Z_i - \mathbb{E}[Z])^2 - \mathrm{Var} \leqslant \sqrt{\frac{2\mathcal{V}\log(1/\delta)}{n}} + \frac{b^2\log(1/\delta)}{3n}, \tag{5}$$

where $\mathcal{V} = \mathbb{E}[(Z - \mathbb{E}[Z])^4] - \mathrm{Var}^2 \leqslant b^2\mathbb{E}[(Z - \mathbb{E}[Z])^2] = b^2\mathrm{Var}$. Substituting this bound in (5), and combining with (4).], we obtain:

$$\begin{aligned}\mathrm{Var}_n &\leqslant \mathrm{Var} + \sqrt{\frac{2b^2\mathrm{Var}\log(1/\delta)}{n}} + \frac{b^2\log(1/\delta)}{3n}\\ &\leqslant \mathrm{Var} + \sqrt{\frac{2b^2\mathrm{Var}\log(1/\delta)}{n}} + \frac{b^2\log(1/\delta)}{2n} = \left(\sqrt{\mathrm{Var}} + b\sqrt{\frac{\log(1/\delta)}{2n}}\right)^2\end{aligned}$$

Taking square root on both sides above gives us:

$$\sqrt{\mathrm{Var}_n} \leqslant \sqrt{\mathrm{Var}} + b\sqrt{\frac{\log(1/\delta)}{2n}}$$

which proves the upper bound. The lower bound is taken from a result in the appendix of Audibert et al. (2009). □

B Proofs for Coupled Uncertainty UCB

To begin we prove the fact that the thresholding rule (2) maintains the ratio of coarse to ground truth feedback to within the desired range.

Proof of Lemma 4.1. We drop the index i for brevity. At time t let us consider the following:

1. Upper bound: Let us consider the latest time $t_1 < t$ when the decision to pull G was taken. Then we must have had $N_t^{\mathrm{G}} - 1 = N_{t_1}^{\mathrm{G}} \leqslant xN_{t_1}^{\mathrm{C}} \leqslant xN_t^{\mathrm{C}}$. On dividing both sides by N_t, this gives us:
$$\begin{aligned}\alpha_t - \frac{1}{N_t} &\leqslant x(1-\alpha_t)\\ \Rightarrow \alpha_t &\leqslant \frac{x}{(1+x)} + \frac{1}{(1+x)N_t} = \alpha + \frac{1-\alpha}{N_t},\end{aligned}$$
which proves the upper bound.

2. Lower bound: Let us similarly consider the latest time $t_2 < t$ when the decision to pull C was taken. Then we must have had $x(N_t^{\mathrm{C}} - 1) = xN_{t_2}^{\mathrm{C}} \leqslant N_{t_2}^{\mathrm{G}} \leqslant N_t^{\mathrm{G}}$. On dividing both sides by N_t, this gives us:
$$\begin{aligned}x\left(1 - \alpha_t - \frac{1}{N_t}\right) &\leqslant \alpha_t\\ \Rightarrow \alpha_t &\geqslant \frac{x}{(1+x)}\left(1 - \frac{1}{N_t}\right) = \alpha\left(1 - \frac{1}{N_t}\right),\end{aligned}$$
which proves the lower bound.

□

B.1 Proofs for the unbiased mean estimator

In this subsection, we present the proof of the concentration of the mean estimator Z_t^λ.

Proof of Lemma 5.1. We drop the index i for brevity. We apply Bennett's inequality to

$$Z_t^\lambda = \sum_{j \in T_t^{\mathrm{G}}} \frac{(r_j^{\mathrm{G}} - \lambda r_j^{\mathrm{C}})}{N_t^{\mathrm{G}}} + \sum_{j \in T_t^{\mathrm{C}}} \frac{\lambda r_j^{\mathrm{C}}}{N_t^{\mathrm{C}}}$$

by treating it as the sum of the N_t independent random variables in the collection $\mathcal{H} = \left\{\frac{(r_j^{\mathrm{G}} - \lambda r_j^{\mathrm{C}})}{N_t^{\mathrm{G}}}\right\}_{j \in T_t^{\mathrm{G}}} \bigcup \left\{\frac{\lambda r_j^{\mathrm{C}}}{N_t^{\mathrm{C}}}\right\}_{j \in T_t^{\mathrm{C}}}$.

We use the fact that each random variable in $\mathcal{H}$ has a bounded support of length $\leqslant B$, where

$$B = \frac{b(1+|\lambda|)}{N_t^{\mathrm{G}}} + \frac{b|\lambda|}{N_t^{\mathrm{C}}} \leqslant \frac{(1+\gamma)b}{N_t^{\mathrm{G}}},$$

when $|\lambda| \leqslant \frac{\gamma}{(1+x_t)}$. Bennett's inequality then gives us the following concentration result w.p. at least $1-\delta$:

$$|Z_t^\lambda - \mathbb{E}[Z_t^\lambda]| \leqslant \sqrt{2\sum_{X\in\mathcal{H}} \mathrm{Var}(X)\log(2/\delta)} + \frac{B\log(2/\delta)}{3} = \sqrt{2\mathrm{Var}(Z_t^\lambda)\log(2/\delta)} + \frac{B\log(2/\delta)}{3}$$

$$\Rightarrow |Z_t^\lambda - \mu^{\mathrm{G}}| \leqslant \sqrt{\frac{2\mathrm{V}_t^\lambda \log(2/\delta)}{N_t^{\mathrm{G}}}} + \frac{(1+\gamma)b\log(2/\delta)}{3N_t^{\mathrm{G}}}.$$

where we use the fact that $\mathbb{E}[Z_t^\lambda] = \mu^{\mathrm{G}}$. □

B.2 Proofs for the concentration of the empirical variance estimators

In this subsection we present the proofs on the concentration of the empirical variance $\hat{\mathrm{V}}_t^\lambda(i)$.

Lemma B.1 (Concentration of $\hat{\mathrm{V}}_t^\lambda(i)$)**.** *Given an arm i and fixing the number of its pulls $N_t^{\mathrm{C}}(i), N_t^{\mathrm{G}}(i)$, consider a fixed $\lambda \in \mathbb{R}$ satisfying $|\lambda| \leqslant (1-\alpha_t(i))\gamma$. Then the following concentration inequalities hold:*

1. *With probability at least $1-\delta$:*

$$\sqrt{\hat{\mathrm{V}}_t^\lambda(i)} \leqslant \sqrt{\mathrm{V}_t^\lambda(i)} + b(1+\gamma)\sqrt{\frac{\log(2/\delta)}{2N_t^{\mathrm{G}}(i)}}$$

2. *With probability at least $1-\delta$:*

$$\sqrt{\mathrm{V}_t^\lambda(i)} \leqslant \sqrt{\hat{\mathrm{V}}_t^\lambda(i)} + 1.8b(1+\gamma)\sqrt{\frac{\log(6/\delta)}{N_t^{\mathrm{G}}(i)}}$$

Proof of Lemma B.1. We omit the index i for brevity. Note that we can split $\hat{V}_t^\lambda$ into two terms as follows:

$$\frac{\hat{\mathrm{V}}_t^\lambda}{N_t^{\mathrm{G}}} = \frac{\widehat{\mathrm{Var}}_1}{N_t^{\mathrm{G}}} + \frac{\widehat{\mathrm{Var}}_2}{N_t^{\mathrm{C}}},$$

where

$$\widehat{\mathrm{Var}}_1 := \mathrm{Var}_t^{\mathrm{C,G}}\lambda^2 - 2\mathrm{Cov}_t\lambda + \mathrm{Var}_t^{\mathrm{G}} = \frac{1}{N_t^{\mathrm{G}}}\sum_{j\in\mathcal{T}_t^{\mathrm{G}}}(r_j^{\mathrm{G}} - \lambda r_j^{\mathrm{C}})^2$$

$$\widehat{\mathrm{Var}}_2 := \lambda^2\mathrm{Var}_t^{\mathrm{C}}(i) = \frac{1}{N_t^{\mathrm{C}}}\sum_{j\in\mathcal{T}_t^{\mathrm{C}}}(\lambda r_j^{\mathrm{C}})^2$$

We see that $\widehat{\mathrm{Var}}_1, \widehat{\mathrm{Var}}_2$ are empirical estimators of the variances $\mathrm{Var}_1, \mathrm{Var}_2$ (resp.) as defined below:

$$\mathrm{Var}_1 = \mathrm{Var}(r^{\mathrm{G}} - \lambda r^{\mathrm{C}}) = \mathrm{Var}^{\mathrm{C}}\lambda^2 - 2\mathrm{Cov}\lambda + \mathrm{Var}^{\mathrm{G}}$$

$$\mathrm{Var}_2 = \mathrm{Var}(\lambda r^{\mathrm{C}}) = \lambda^2\mathrm{Var}^{\mathrm{C}}.$$

Also note that $\frac{\mathrm{V}_t^\lambda}{N_t^{\mathrm{G}}} = \frac{\mathrm{Var}_1}{N_t^{\mathrm{G}}} + \frac{\mathrm{Var}_2}{N_t^{\mathrm{C}}}$.

We now show the first concentration inequality by considering the following two concentration events:

$$\mathcal{E}_1 = \left\{\sqrt{\widehat{\mathrm{Var}}_1} \leqslant \sqrt{\mathrm{Var}_1} + b(1+|\lambda|)\sqrt{\frac{\log(4/\delta)}{2N_t^{\mathrm{G}}}}\right\}$$

$$\mathcal{E}_2 = \left\{ \sqrt{\widehat{\mathrm{Var}_2}} \leqslant \sqrt{\mathrm{Var}_2} + b|\lambda|\sqrt{\frac{\log(4/\delta)}{2N_t^{\mathrm{C}}}} \right\}$$

From lemma A.1, it follows that each of the above events occurs w.p. at least $1 - \delta/2$, and so we prove our result on $\mathcal{E}_1 \cap \mathcal{E}_2$ which occurs w.p. at least $1 - \delta$. First for the upper bound, we combine the upper bounds in $\mathcal{E}_1, \mathcal{E}_2$ to get:

$$\frac{\hat{V}_t^{\lambda}}{N_t^{\mathrm{G}}} = \frac{\widehat{\mathrm{Var}_1}}{N_t^{\mathrm{G}}} + \frac{\widehat{\mathrm{Var}_2}}{N_t^{\mathrm{C}}}$$

$$\leqslant \frac{1}{N_t^{\mathrm{G}}}\left(\mathrm{Var}_1 + 2w_1\sqrt{\frac{\mathrm{Var}_1}{N_t^{\mathrm{G}}}} + \frac{(w_1)^2}{N_t^{\mathrm{G}}}\right) + \frac{1}{N_t^{\mathrm{C}}}\left(\mathrm{Var}_2 + 2w_2\sqrt{\frac{\mathrm{Var}_2}{N_t^{\mathrm{C}}}} + \frac{(w_2)^2}{N_t^{\mathrm{C}}}\right)$$

$$= \left(\frac{\mathrm{Var}_1}{N_t^{\mathrm{G}}} + \frac{\mathrm{Var}_2}{N_t^{\mathrm{C}}}\right) + 2\left(\frac{w_1}{N_t^{\mathrm{G}}}\right)\sqrt{\frac{\mathrm{Var}_1}{N_t^{\mathrm{G}}}} + 2\left(\frac{w_2}{N_t^{\mathrm{C}}}\right)\sqrt{\frac{\mathrm{Var}_2}{N_t^{\mathrm{C}}}} + \left(\frac{w_1}{N_t^{\mathrm{G}}}\right)^2 + \left(\frac{w_2}{N_t^{\mathrm{C}}}\right)^2$$

$$\leqslant \frac{\mathrm{V}_t^{\lambda}}{N_t^{\mathrm{G}}} + 2\left(\frac{w_1}{N_t^{\mathrm{G}}} + \frac{w_2}{N_t^{\mathrm{C}}}\right)\sqrt{\frac{\mathrm{V}_t^{\lambda}}{N_t^{\mathrm{G}}}} + \left(\frac{w_1}{N_t^{\mathrm{G}}} + \frac{w_2}{N_t^{\mathrm{C}}}\right)^2$$

where $w_1 = 2^{-1/2}b(1+|\lambda|)\sqrt{\log(4/\delta)}$, $w_2 = 2^{-1/2}b|\lambda|\sqrt{\log(4/\delta)}$. Taking sqaure root and multiplying by $\sqrt{N_t^{\mathrm{G}}}$ on both sides gives us:

$$\sqrt{\hat{\mathrm{V}}_t^{\lambda}} \leqslant \sqrt{\mathrm{V}_t^{\lambda}} + \frac{(w_1 + x_t w_2)}{\sqrt{N_t^{\mathrm{G}}}} \leqslant \sqrt{\mathrm{V}_t^{\lambda}} + b(1+\gamma)\sqrt{\frac{\log(4/\delta)}{2N_t^{\mathrm{G}}}},$$

where the last inequality above uses the assumption that $(1 + x_t)|\lambda| \leqslant \gamma$. This proves the first concentration result.

The second concentration result follows from repeating the same steps as above by starting with $\frac{V_t^{\lambda}}{N_t^{\mathrm{G}}} = \frac{\mathrm{Var}_1}{N_t^{\mathrm{G}}} + \frac{\mathrm{Var}_2}{N_t^{\mathrm{C}}}$ and upper bounding each term using the events $\mathcal{E}_1', \mathcal{E}_2'$ as defined below.

$$\mathcal{E}_1' = \left\{ \sqrt{\mathrm{Var}_1} \leqslant \sqrt{\widehat{\mathrm{Var}_1}} + 1.8b(1+|\lambda|)\sqrt{\frac{\log(6/\delta)}{N_t^{\mathrm{G}}}} \right\}$$

$$\mathcal{E}_2' = \left\{ \sqrt{\mathrm{Var}_2} \leqslant \sqrt{\widehat{\mathrm{Var}_2}} + 1.8b|\lambda|\sqrt{\frac{\log(6/\delta)}{N_t^{\mathrm{C}}}} \right\}$$

$\mathcal{E}_1', \mathcal{E}_2'$ each holds w.p. at least $1 - \delta/2$ from Lemma A.1, so the concentration result holds on their intersection w.p. at least $1 - \delta$. □

B.3 Constructing the Coupled Uncertainty Upper Confidence Bound

This section develops the final proofs for the upper confidence bound used in CUUCB.

Lemma B.2 (Variance adaptive concentration of $Z_t^{\lambda}(i)$)**.** *Given an arm i and fixing the number of its pulls $N_t^{\mathrm{C}}(i), N_t^{\mathrm{G}}(i)$, consider a fixed $\lambda \in \mathbb{R}$ satisfying $|\lambda| \leqslant (1 - \alpha_t(i))\gamma$. Then the following concentration inequality holds with probability at least $1 - \delta$:*

$$|Z_t^{\lambda}(i) - \mu^{\mathrm{G}}(i)| \leqslant \sqrt{\frac{2\hat{\mathrm{V}}_t^{\lambda}(i)\log\left(\frac{8}{\delta}\right)}{N_t^{\mathrm{G}}(i)}} + \frac{3(1+\gamma)b\log\left(\frac{8}{\delta}\right)}{N_t^{\mathrm{G}}(i)}$$

Proof of Lemma B.2. We omit the index i for brevity. Consider the following concentration events.

$$\mathcal{E}_1 = \left\{ |Z_t^{\lambda} - \mu^{\mathrm{G}}| \leqslant \sqrt{\frac{2\mathrm{V}_t^{\lambda}\log(8/\delta)}{N_t^{\mathrm{G}}}} + \frac{(1+\gamma)b\log(8/\delta)}{3N_t^{\mathrm{G}}} \right\}$$

$$\mathcal{E}_{2,t} = \left\{ \sqrt{\mathrm{V}_t^{\lambda}} \leqslant \sqrt{\hat{\mathrm{V}}_t^{\lambda}} + 1.8b(1+\gamma)\sqrt{\frac{\log(8/\delta)}{N_t^{\mathrm{G}}}} \right\}$$

From lemma 5.1, we have that $\mathbb{P}(\mathcal{E}_1^c) \leqslant \frac{\delta}{4}$. From lemma B.1, we have that $\mathbb{P}(\mathcal{E}_2^c) \leqslant \frac{3\delta}{4}$. We thus show the concentration result on the event $\mathcal{E}_1 \cap \mathcal{E}_2$ which holds w.p. at least $1 - \delta$.

We denote $l_\delta = \log(8/\delta)$ for simplicity. Combining the two inequalities, we have:

$$|Z_t^\lambda - \mu^{\mathrm{G}}| \leqslant \sqrt{\frac{2\mathrm{V}_t^\lambda l_\delta}{N_t^{\mathrm{G}}}} + \frac{(1+\gamma)bl_\delta}{3N_t^{\mathrm{G}}}$$

$$\leqslant \sqrt{\frac{2\hat{\mathrm{V}}_t^\lambda l_\delta}{N_t^{\mathrm{G}}}} + \frac{1.8\sqrt{2}(1+\gamma)bl_\delta}{N_t^{\mathrm{G}}} + \frac{(1+\gamma)bl_\delta}{3N_t^{\mathrm{G}}} \leqslant \sqrt{\frac{2\hat{\mathrm{V}}_t^\lambda l_\delta}{N_t^{\mathrm{G}}}} + \frac{3(1+\gamma)bl_\delta}{N_t^{\mathrm{G}}},$$

which proves the result. □

Proof of Lemma 5.2. Let us first fix the number of pulls $N_t(i)$ of the arm i. Note that because of our deterministic decision rule for m_t, $N_t^{\mathrm{G}}(i), N_t^{\mathrm{C}}(i)$ are also fixed upon fixing $N_t(i)$. Also, all the estimators for arm i are evaluated using the outcomes of these $N_t(i)$ pulls. In the following steps, we omit the index i for brevity.

Let us define $\gamma' := (1-\alpha_t)\gamma$. Note that our estimator $\lambda_t \in [-\gamma', \gamma']$ by construction. Thus, we define a grid on the interval $[-\gamma', \gamma']$ as follows: Define $\epsilon := \frac{2\gamma'}{N_t}$. We cover the interval $[-\gamma, \gamma]$ using N_t ϵ-width intervals $\{I_j := [\beta_{j-1}, \beta_j]\}_{j=1}^{N_t}$, where $\beta_j := -\gamma' + j\epsilon$.

We define $\delta_n := \frac{\delta}{2n^2}$. Note that $\sum_{n\geqslant 1} \delta_n = \frac{\delta}{2}\sum_{n\geqslant 1}\frac{1}{n^2} \leqslant \delta$.

We have $L_t = \log\left(\frac{8(N_t+2)}{\delta_{N_t}}\right)$. Now, let us consider the following concentration events.

$$\mathcal{E}_{1,N_t} = \left\{ |Z_t^\lambda - \mu^{\mathrm{G}}| \leqslant \sqrt{\frac{2\hat{\mathrm{V}}_t^\lambda L_t}{N_t^{\mathrm{G}}}} + \frac{3(1+\gamma)bL_t}{N_t^{\mathrm{G}}}, \forall \lambda \in \{\beta_0, \beta_1, ..., \beta_{N_t}\} \right\}$$

$$\mathcal{E}_{2,N_t} = \left\{ \sqrt{\hat{\mathrm{V}}_t^{\lambda_t^*}} \leqslant \sqrt{\mathrm{V}_t^{\lambda_t^*}} + b(1+\gamma)\sqrt{\frac{L_t}{2N_t^{\mathrm{G}}}} \right\}$$

From lemma B.2 and a union bound, we have that $\mathbb{P}(\mathcal{E}_{1,N_t}^c) \leqslant \frac{(N_t+1)\delta_{N_t}}{(N_t+2)}$. From lemma B.1 and a union bound, we have that $\mathbb{P}(\mathcal{E}_{2,N_t}^c) \leqslant \frac{\delta_{N_t}}{4(N_t+2)} \leqslant \frac{\delta_{N_t}}{(N_t+2)}$. Now, we consider the intersection event $\mathcal{E}_{N_t} := \bigcap_{i=1}^2 \mathcal{E}_{i,N_t}$. Then a union bound gives us $\mathbb{P}(\mathcal{E}_{N_t}^c) \leqslant \delta_{N_t}$. We prove our bounds on the event $\mathcal{E}_{N_t}$ which occurs w.p. at least $1 - \delta_{N_t}$.

First, we derive some useful inequalities. Given any $\lambda, \lambda' \in [-\gamma', \gamma']$ satisfying $|\lambda - \lambda'| \leqslant \epsilon$,we have:

$$|\hat{\mathrm{V}}_t^\lambda - \hat{\mathrm{V}}_t^{\lambda'}| \leqslant (1+x_t)\widehat{\mathrm{Var}}_t^{\mathrm{C}}(|\lambda| + |\lambda'|)|\lambda - \lambda'| + 2|\lambda - \lambda'||\mathrm{Cov}_t|$$

$$\overset{(i)}{\leqslant} 2(\gamma\widehat{\mathrm{Var}}_t^{\mathrm{C}} + |\mathrm{Cov}_t|)\epsilon \overset{(ii)}{\leqslant} \frac{1}{N_t}\gamma(\gamma+4)b^2, \tag{6}$$

where in step (i) we have used the fact that $|\lambda|, |\lambda'| \leqslant \gamma'$. In step (ii) we have used the fact that $\widehat{\mathrm{Var}}^{\mathrm{C}} \leqslant \frac{b^2}{4}, |\mathrm{Cov}_t| \leqslant b^2$.

Now, consider an arbitrary $\lambda \in [-\gamma', \gamma']$. We derive an upper bound on $|Z_t^\lambda - \mu^{\mathrm{G}}|$. For this, suppose $j \in [N_t]$ is such that $\lambda \in I_j$. Then for $\lambda' = \beta_{j-1}, \beta_j$ we have $|\lambda - \lambda'| \leqslant \epsilon$. Also, for these values of λ', we have from event $\mathcal{E}_{1,N_t}$:

$$|Z_t^{\lambda'} - \mu^{\mathrm{G}}| \leqslant \sqrt{\frac{2\hat{\mathrm{V}}_t^{\lambda'} L_t}{N_t^{\mathrm{G}}}} + \frac{3(1+\gamma)bL_t}{N_t^{\mathrm{G}}} \tag{7}$$

From (6) and using $\sqrt{x+y} \leqslant \sqrt{x} + \sqrt{y}$ for $x, y > 0$, we also have:

$$\sqrt{\hat{\mathrm{V}}_t^{\lambda'}} \leqslant \sqrt{\hat{\mathrm{V}}_t^{\lambda} + \frac{\gamma'(\gamma+4)b^2}{N_t}} \leqslant \sqrt{\hat{\mathrm{V}}_t^{\lambda}} + \sqrt{\frac{\gamma'(\gamma+4)b^2}{N_t}} \leqslant \sqrt{\hat{\mathrm{V}}_t^{\lambda}} + \frac{(\gamma+2)b}{\sqrt{N_t^{\mathrm{G}}}}$$

Using the above bound in (7), we obtain the following for $\lambda' = \beta_{j-1}, \beta_j$

$$\begin{aligned} |Z_t^{\lambda'} - \mu^{\mathrm{G}}| &\leqslant \sqrt{\frac{2\hat{\mathrm{V}}_t^{\lambda} L_t}{N_t^{\mathrm{G}}}} + \frac{(\gamma+2)b\sqrt{2L_t}}{N_t^{\mathrm{G}}} + \frac{3(1+\gamma)bL_t}{N_t^{\mathrm{G}}} \\ &\leqslant \sqrt{\frac{2\hat{\mathrm{V}}_t^{\lambda} L_t}{N_t^{\mathrm{G}}}} + \frac{6(1+\gamma)bL_t}{N_t^{\mathrm{G}}}, \end{aligned} \tag{8}$$

where the last inequality above uses the fact that $L_t \geqslant 1$, which means $\sqrt{L_t} \leqslant L_t$. Now we show the concentration of Z_t^{λ} by noting that $|Z_t^{\lambda} - \mu^{\mathrm{G}}|$ is a convex function of λ and since $\lambda \in I_j$:

$$|Z_t^{\lambda} - \mu^{\mathrm{G}}| \leqslant \max_{\lambda' \in \{\beta_{j-1}, \beta_j\}} |Z_t^{\lambda'} - \mu^{\mathrm{G}}| \leqslant \sqrt{\frac{2\hat{\mathrm{V}}_t^{\lambda} L_t}{N_t^{\mathrm{G}}}} + \frac{6(1+\gamma)bL_t}{N_t^{\mathrm{G}}}, \tag{9}$$

where we obtain the final bound from (8). Since $\lambda \in [-\gamma', \gamma']$ was arbitrary, (9) holds for all $\lambda \in [-\gamma', \gamma']$. In particular, since the random variable $\lambda_t \in [-\gamma', \gamma']$, we have that:

$$|Z_t^{\lambda_t} - \mu^{\mathrm{G}}| = |Z_t - \mu^{\mathrm{G}}| \leqslant \sqrt{\frac{2\hat{\mathrm{V}}_t^{\lambda_t} L_t}{N_t^{\mathrm{G}}}} + \frac{6(1+\gamma)bL_t}{N_t^{\mathrm{G}}} \tag{10}$$

Since $\lambda_t = \arg\min_{\lambda \in [-\gamma', \gamma']} \hat{\mathrm{V}}_t^{\lambda}$ and $\lambda_t^* \in [-\gamma', \gamma']$, we have $\hat{\mathrm{V}}_t^{\lambda_t} \leqslant \hat{\mathrm{V}}_t^{\lambda_t^*}$. Combining this with the bound in event $\mathcal{E}_{2,N_t}$, we obtain:

$$\sqrt{\hat{\mathrm{V}}_t^{\lambda_t}} \leqslant \sqrt{\hat{\mathrm{V}}_t^{\lambda_t^*}} \leqslant \sqrt{\mathrm{V}_t^{\lambda_t^*}} + b(1+\gamma)\sqrt{\frac{L_t}{2N_t^{\mathrm{G}}}}.$$

Plugging the above into (10) and noting $\mathrm{V}_t^{\lambda_t^*} = \mathrm{V}_t^*$ gives us the result in the lemma for a fixed number of pulls $N_t(i)$.

To show the anytime version, note that $\mathbb{P}\left(\bigcap_{n \geqslant 1} \mathcal{E}_n\right) \geqslant 1 - \sum_{n \geqslant 1} \mathbb{P}(\mathcal{E}_n^c) \geqslant 1 - \sum_{n \geqslant 1} \delta_n \geqslant 1 - \delta$. Therefore, the anytime concentration result holds w.p. at least $1 - \delta$. This completes the proof. □

B.4 Proofs for regret of CUUCB

In this subsection we present the proof for Theorem 5.3, which bounds the regret of CUUCB.

Proof of Theorem 5.3. We decompose the pseudo-regret as:

$$\mathrm{Regret}(T) = \sum_{i \in [K]: \Delta_i > 0} N_T(i)\Delta_i,$$

where $\Delta_i = \max_{j \in [K]} \mu^{\mathrm{G}}(j) - \mu^{\mathrm{G}}(i)$ and $N_T(i)$ is the total number of times arm i is pulled. We denote $\mu^{*,\mathrm{G}} := \max_{j \in [K]} \mu^{\mathrm{G}}(j)$. Let i^* be an optimal arm, i.e., an arm satisfying $\mu^{\mathrm{G}}(i^*) = \mu^{*,\mathrm{G}}$.

Let us fix a sub-optimal arm i. We denote by $\mathcal{E}_i$ the event

$$\mathcal{E}_i = \left\{ |Z_t(j) - \mu^{\mathrm{G}}(j)| \leqslant \sqrt{\frac{2\hat{\mathrm{V}}_t^{\lambda_t(j)}(j) L_t(j)}{N_t^{\mathrm{G}}(j)}} + \frac{6(1+\gamma)bL_t(j)}{N_t^{\mathrm{G}}(j)} \leqslant D_t(j), \text{ for } j = i, i^* \right\}$$

where

$$D_t(j) := \sqrt{\frac{2\mathrm{V}_t^*(j) L_t(j)}{N_t^{\mathrm{G}}(j)}} + \frac{7(1+\gamma)bL_t(j)}{N_t^{\mathrm{G}}(j)} \tag{11}$$

From Lemma 5.2, the event $\mathcal{E}_i$ occurs with at least $1 - 2\delta$ probability. On this event, we have that $\mu^{\mathrm{G}}(i^*) \leqslant U_t(i^*)$ for all times t. Thus, the sub-optimal arm i is not pulled at any time t when $U_t(i) < \mu^{\mathrm{G}}(i^*) \leqslant U_t(i^*)$. We will use this upper bound the total number of pulls $N_T(i)$.

On event $\mathcal{E}_i$, we have:

$$U_t(i) \leqslant Z_t(i) + D_t(i) \leqslant \mu^{\mathrm{G}}(i) + 2D_t(i)$$

Thus arm i is not pulled at any time t when $\mu^{\mathrm{G}}(i) + 2D_t(i) < \mu^{*,\mathrm{G}} \iff D_t(i) < \Delta_i/2$. Now, we upper bound $D_t(i)$.

From here on, we drop the index i, which denotes the sub-optimal arm under consideration. From Lemma 4.1, we have $1 - \alpha_t \geqslant (1 - \alpha)\left(1 - \frac{1}{N_t}\right)$. Noting $\mathrm{V}_t^* = \mathrm{Var}^{\mathrm{G}} - (1 - \alpha_t)\frac{\mathrm{Cov}^2}{\mathrm{Var}^{\mathrm{C}}}$, we get

$$\mathrm{V}_t^* \leqslant \mathrm{Var}^{\mathrm{G}} - (1-\alpha)\frac{\mathrm{Cov}^2}{\mathrm{Var}^{\mathrm{C}}} + (1-\alpha)\frac{\mathrm{Cov}^2}{\mathrm{Var}^{\mathrm{C}} N_t} \leqslant \mathrm{V}^* + \frac{\gamma^2 b^2}{4N_t},$$

where the last step above uses $|\mathrm{Cov}| \leqslant \gamma \mathrm{Var}^{\mathrm{C}}$ and $\mathrm{Var}^{\mathrm{C}} \leqslant b^2/4$. Taking square root on both sides in the above inequality, we get:

$$\sqrt{\mathrm{V}_t^*} \leqslant \sqrt{\mathrm{V}^* + \frac{\gamma^2 b^2}{4N_t}} \leqslant \sqrt{\mathrm{V}^*} + \frac{\gamma b}{2\sqrt{N_t}} \leqslant \sqrt{\mathrm{V}^*} + \frac{\gamma b}{2\sqrt{N_t^{\mathrm{G}}}} \tag{12}$$

Substituting (12) into (11), we obtain:

$$\begin{aligned} D_t &\leqslant \sqrt{\frac{2\mathrm{V}^* L_t}{N_t^{\mathrm{G}}}} + \frac{\gamma b\sqrt{L_t/2}}{N_t^{\mathrm{G}}} + \frac{7(1+\gamma)bL_t}{N_t^{\mathrm{G}}} \\ &< \sqrt{\frac{2\mathrm{V}^* L_t}{N_t^{\mathrm{G}}}} + \frac{8(1+\gamma)bL_t}{N_t^{\mathrm{G}}} \end{aligned}$$

where in the last step, we use $L_t \geqslant 1$.

Thus, a sufficient condition for $D_t < \Delta/2$ is:

$$\begin{aligned} &\sqrt{\frac{2\mathrm{V}^* L_t}{N_t^{\mathrm{G}}}} + \frac{8(1+\gamma)bL_t}{N_t^{\mathrm{G}}} \leqslant \frac{\Delta}{2} \\ \Leftarrow\ & \frac{2\mathrm{V}^* L_t}{N_t^{\mathrm{G}}} \leqslant \frac{\Delta^2}{16} \text{ and } \frac{8(1+\gamma)bL_t}{N_t^{\mathrm{G}}} \leqslant \frac{\Delta}{4} \\ \iff\ & \frac{N_t^{\mathrm{G}}}{L_t} \geqslant \max\left\{\frac{32\mathrm{V}^*}{\Delta^2}, \frac{32(1+\gamma)b}{\Delta}\right\} \\ \Leftarrow\ & N_t^{\mathrm{G}} \geqslant \frac{32}{\Delta}\left(\frac{\mathrm{V}^*}{\Delta} + (1+\gamma)b\right)L_t \\ \overset{(i)}{\Leftarrow}\ & \alpha(N_t - 1) \geqslant \frac{128}{\Delta}\left(\frac{\mathrm{V}^*}{\Delta} + (1+\gamma)b\right)\log(2T) \end{aligned}$$

Implication (i) uses Lemma 4.1 which gives us $N_t^{\mathrm{G}} = \alpha_t N_t \geqslant \alpha(1 - N_t^{-1})N_t = \alpha(N_t - 1)$. It also uses the bound $L_t \leqslant \log(16\delta^{-1}(N_t+1)^3) \leqslant 4\log(2T)$ for any $t \in [T]$. Thus if

$$N_t \geqslant \frac{128}{\alpha\Delta}\left(\frac{\mathrm{V}^*}{\Delta} + (1+\gamma)b\right)\log(2T) + 1,$$

then arm i is not pulled. Reintroducing the index i, this means we obtain the following bound on the number of pulls of arm i when $\mathcal{E}_i$ occurs:

$$N_T(i) \leqslant \frac{128}{\alpha(i)\Delta_i}\left(\frac{\mathrm{V}^*(i)}{\Delta_i} + (1+\gamma)b\right)\log(2T) + 2 \tag{13}$$

Thus, we have

$$\begin{aligned}\mathbb{E}[N_T(i)] &= \mathbb{E}[N_T(i)\mathbb{1}\{\mathcal{E}_i\}] + \mathbb{E}[N_T(i)\mathbb{1}\{\mathcal{E}_i^c\}] \\ &\leqslant \frac{128}{\alpha(i)\Delta_i}\left(\frac{\mathrm{V}^*(i)}{\Delta_i} + (1+\gamma)b\right)\log(2T) + 2 + T\mathbb{P}(\mathcal{E}_i^c)\end{aligned}$$

We obtain this inequality by using (13) on event $\mathcal{E}_i$ along with the trivial bound $N_T(i) \leqslant T$ on event $\mathcal{E}_i^c$. Using $\mathbb{P}(\mathcal{E}_i^c) \leqslant 2/T$ gives us:

$$\mathbb{E}[N_T(i)] \leqslant \frac{128}{\alpha(i)\Delta_i}\left(\frac{\mathrm{V}^*(i)}{\Delta_i} + (1+\gamma)b\right)\log(2T) + 4 \tag{14}$$

Using (14), we can bound the expected regret of the algorithm as:

$$\begin{aligned}\mathbb{E}[\text{Regret}(T)] &= \sum_{i\in[K]} \Delta_i \mathbb{E}[N_T(i)] \\ &\leqslant \sum_{i:\Delta_i>0}\left(\frac{128}{\alpha(i)}\left(\frac{\mathrm{V}^*(i)}{\Delta_i} + (1+\gamma)b\right)\log(2T) + 4\Delta_i\right)\end{aligned}$$

This gives us the instance-dependent regret bound. For the instance-independent regret bound, consider some $\hat{\Delta} > 0$. While bounding the expected, we bound the constributions of arms i with $\Delta_i > \hat{\Delta}$ and arms i with $\Delta_i \leqslant \hat{\Delta}$ differently as follows (Note $\Delta_i \leqslant b$ for any i):

$$\begin{aligned}\mathbb{E}[\text{Regret}(T)] &\leqslant \sum_{i:\Delta_i>\hat{\Delta}}\left(\frac{128}{\alpha(i)}\left(\frac{\mathrm{V}^*(i)}{\hat{\Delta}} + (1+\gamma)b\right)\log(2T) + 4b\right) + \sum_{i:\Delta_i\leqslant\hat{\Delta}} \hat{\Delta}\mathbb{E}[N_T(i)] \\ &\leqslant \frac{128\log(2T)}{\hat{\Delta}}\left(\sum_i \frac{\mathrm{V}^*(i)}{\alpha(i)}\right) + \hat{\Delta}T + 132\sum_i \alpha(i)^{-1}(1+\gamma)b\log(2T)\end{aligned}$$

Choosing $\hat{\Delta} = \sqrt{128\left(\sum_i \frac{\mathrm{V}^*(i)}{\alpha(i)}\right)\frac{\log(2T)}{T}}$ to minimize the upper bound gives us:

$$\begin{aligned}\mathbb{E}[\text{Regret}(T)] &\leqslant 2\sqrt{128\left(\sum_i \frac{\mathrm{V}^*(i)}{\alpha(i)}\right)T\log(2T)} + \mathcal{O}(\log(T)) \\ &\leqslant 23\sqrt{\left(\sum_i \frac{\mathrm{V}^*(i)}{\alpha(i)}\right)T\log(2T)} + \mathcal{O}(\log(T))\end{aligned}$$

This completes the proof. □

Lemma B.3 (Regret bound for UCB-V)**.** *When* UCB-V *is implemented with the threshold rule* (2)*, while using only ground-truth samples for decision making, then its expected regret satisfies the following instance-independent bound:*

$$\mathbb{E}[\textit{Regret}(T)] \leqslant \mathcal{O}\left(\sqrt{\left(\sum_i \frac{\mathrm{Var}^{\mathrm{G}}(i)}{\alpha(i)}\right)T\log T}\right)$$

Proof of Lemma B.3. When UCB-V is run using only ground-truth samples, then the estimated upper-confidence bound for arm i at time t by UCB-V(Audibert et al., 2009) is given by (barring constants, which do not affect the asymptotic order of regret):

$$U_t'(i) = \mu_t^{\mathrm{G}}(i) + \sqrt{\frac{2\mathrm{Var}_t^{\mathrm{G}}(i)L_t'(i)}{N_t^{\mathrm{G}}(i)}} + \frac{3bL_t'(i)}{N_t^{\mathrm{G}}(i)},$$

where $L_t'(i)$ is an exploration function typically taken to be growing as $\log t$. We fix the arm i drop the index i in the following analysis. Let us take $\delta_n = \frac{\delta}{2n^2}$ We see by taking $\lambda = 0, \gamma = 0$ in Lemma

B.2 and Lemma B.1, and using union bound that the following concentration inequality holds after N_t^{G} pulls:

$$|Z_t^0 - \mu^{\mathrm{G}}| \leqslant \sqrt{\frac{2\hat{\mathrm{V}}_t^0 \log\left(10/\delta_{N_t^{\mathrm{G}}}\right)}{N_t^{\mathrm{G}}}} + \frac{3b\log\left(10/\delta_{N_t^{\mathrm{G}}}\right)}{N_t^{\mathrm{G}}} \leqslant \sqrt{\frac{2\mathrm{V}_t^0 \log\left(10/\delta_{N_t^{\mathrm{G}}}\right)}{N_t^{\mathrm{G}}}} + \frac{4b\log\left(10/\delta_{N_t^{\mathrm{G}}}\right)}{N_t^{\mathrm{G}}},$$

w.p. at least $1 - \delta_{N_t^{\mathrm{G}}}/5 - 4\delta_{N_t^{\mathrm{G}}}/5 = 1 - \delta_{N_t^{\mathrm{G}}}$. We note that $Z_t^0 = \mu_t^{\mathrm{G}}$, $\hat{\mathrm{V}}_t^0 = \mathrm{Var}_t^{\mathrm{G}}$ and $\mathrm{V}_t^0 = \mathrm{Var}^{\mathrm{G}}$. Since $\sum_n \delta_n \leqslant \delta$, we have that the above inequality holds for all times w.p. at least $1 - \delta$, which gives us the following anytime concentration:

$$|\mu_t^{\mathrm{G}} - \mu^{\mathrm{G}}| \leqslant \sqrt{\frac{2\mathrm{Var}_t^{\mathrm{G}} L_t'}{N_t^{\mathrm{G}}}} + \frac{3bL_t'}{N_t^{\mathrm{G}}} \leqslant \sqrt{\frac{2\mathrm{Var}^{\mathrm{G}} L_t'}{N_t^{\mathrm{G}}}} + \frac{4bL_t'}{N_t^{\mathrm{G}}} \tag{15}$$

Where we have set $L_t' = \log(20(N_t^{\mathrm{G}})^2/\delta)$, and we will show the regret bound setting $\delta = 1/T$. This choice of L_t' matches the $\log t$ growth we want. On this event, we have that

$$U_t(i) \leqslant \mu^{\mathrm{G}}(i) + 2D_t'(i)$$

for $D_t'(i) = \sqrt{\frac{2\mathrm{Var}^{\mathrm{G}}(i)L_t'(i)}{N_t^{\mathrm{G}}(i)}} + \frac{4bL_t'(i)}{N_t^{\mathrm{G}}(i)}$.

Now, we proceed identically as in the regret proof of Lemma 5.3, by considering the event $\mathcal{E}_i$ w.p. at least $1-2/T$, when the anytime-concentration (15) holds for the sub-optimal arm i and some optimal arm i*. On event $\mathcal{E}_i$, we bound the number of pulls after which $U_t(i) < \mu^{\mathrm{G}}(i^*)$, which happens if $D_t'(i) < \Delta_i/2$. This is achieved at times t such that:

$$N_t^{\mathrm{G}}(i) \geqslant \frac{32L_t'(i)}{\Delta_i}\left(\frac{\mathrm{Var}^{\mathrm{G}}(i)}{\Delta_i} + b\right)$$

$$\Leftarrow \alpha(i)(N_t(i) - 1) \geqslant \frac{C\log T}{\Delta_i}\left(\frac{\mathrm{Var}^{\mathrm{G}}(i)}{\Delta_i} + b\right),$$

which follows from Lemma 4.1 and bounding $L_t'(i) \leqslant C\log T$, for an absolute constant C and T sufficiently large. Thus, proceeding as in the derivation of (14),

$$\mathbb{E}[N_T(i)] \leqslant \frac{C}{\alpha(i)\Delta_i}\left(\frac{\mathrm{Var}^{\mathrm{G}}(i)}{\Delta_i} + b\right)\log(T) + 4 \tag{16}$$

Using the same argument as in the derivation of the instance-independent regret bound in the proof of Lemma 5.3, we obtain the following instance-independent bound for UCB-V

$$\mathbb{E}[\text{Regret}(T)] \leqslant 2\sqrt{C\left(\sum_i \frac{\mathrm{Var}^{\mathrm{G}}(i)}{\alpha(i)}\right)T\log(T) + \mathcal{O}(\log(T))}$$

which concludes the proof. □

C Experimental Details

C.1 Implementation of Algorithms

To make the comparison with other UCB variants more straightforward, we implement the upper confidence bound $U_t(i)$ for CUUCB a little differently than described in the theoretical analysis while preserving its main structure. The implemented expressions of $U_t(i)$ for each arm i at time t for each algorithm is given below:

$$\text{UCB}: U_t(i) = \mu_t^{\mathrm{G}}(i) + \sqrt{\frac{\zeta_1 b^2 \log t}{N_t^{\mathrm{G}}(i)}}$$

$$\text{UCB-V} : U_t(i) = \mu_t^{\mathrm{G}}(i) + \sqrt{\frac{2\mathrm{Var}_t^{\mathrm{G}}(i)\log t}{N_t^{\mathrm{G}}(i)}} + \frac{b\zeta_2 \log t}{N_t^{\mathrm{G}}(i)}$$

$$\text{CUUCB} : U_t(i) = Z_t(i) + \sqrt{\frac{2\hat{\mathrm{V}}_t^{\lambda_t(i)}(i)\log t}{N_t^{\mathrm{G}}(i)}} + \frac{b\zeta_3 \log t}{N_t^{\mathrm{G}}(i)}$$

Here $\{\zeta_j\}_{j=1}^3$ are tunable parameters. Moreover, we obtain the final upper confidence bound by taking $\min\{U_t(i), b\}$ in each case. Note that UCB and UCB-V use only ground-truth observations in their confidence bounds. Thus, all algorithms use rule (2) to pick the set of ground-truth measurements and choose arms for all T time steps. Also note that the implementation of UCB-V is different from the one described above in the LLM-as-a-judge experiment on the mis-aligned Dataset-B, where we run UCB-V using only coarse feedback at all time steps. So, in just this experiment, the expression for $U_t(i)$ for UCB-V is given by:

$$U_t(i) = \hat{\mu}_t^{\mathrm{C}}(i) + \sqrt{\frac{2\hat{\mathrm{V}}_t^{\mathrm{C}}(i)\log t}{N_t(i)}} + \frac{b\zeta_2 \log t}{N_t(i)},$$

where

$$\hat{\mu}_t^{\mathrm{C}}(i) = \frac{1}{N_t(i)} \sum_{j\in[t]:i_j=i} r_j^{\mathrm{C}}(i), \qquad \hat{\mathrm{V}}_t^{\mathrm{C}}(i) = \frac{1}{N_t(i)} \sum_{j\in[t]:i_j=i} (r_j^{\mathrm{C}}(i) - \hat{\mu}_t^{\mathrm{C}}(i))^2$$

are empirical mean and variance estimators of all the coarse measurements observed for arm i at time t.

In all of our experiments, we used $\zeta_1 = 1, \zeta_2 = 0.5, \zeta_3 = 0.5$. We also use a uniform target ratio $\alpha(i) = \alpha$ for all arms i in all our experiments.

C.2 Data Generation

We describe the dataset generation methods used for our experiments. Our generated datasets are always a collection of N 2-D vectors for each of the K arms. We refer to N as the size of the dataset. When running algorithms on a generated dataset, the reward pair $[r_t^{\mathrm{G}}(i), r_t^{\mathrm{C}}(i)]$ for an arm i is generated by sampling uniformly from this dataset. Thus, the mean rewards $[\mu^{\mathrm{G}}(i), \mu^{\mathrm{C}}(i)]$ are taken to be the empirical means of the dataset.

C.2.1 Synthetic

For the synthetic experiments, our dataset generation is parameterized by the correlation parameter ρ. The generation of a (ground-truth, coarse) measurement pair for a give arm i proceeds as follows:

1. Sample from a 2-D Gaussian with mean $[\mu_1(i), \mu_2(i)]$ and covariance matrix $C(\rho)$.
2. Clip the values in the interval I.

We set $C(\rho) = \begin{bmatrix} 5 & 4.5\rho \\ 4.5\rho & 4 \end{bmatrix}$.

We fix the means $\mu_1(i)$ for $i = 1, ..., 4$ as $[0.5, 0.7, 0.8, 0.2]$. We then obtain $\mu_2(i)$ as a perturbed version of $\mu_1(i)$ by uniformly sampling in the interval $\mu_1(i) \pm 0.1$ and then clipping in the interval $[0, 1]$. We also set $I = [-4, 6]$.

For each value of ρ, we generate the dataset by sampling $N = 100,000$ datapoints as described above for each arm i.

C.3 LLM-as-a-judge

We make use of the Nectar dataset of rankings of (prompt,answer) pairs from Zhu et al. (2024) for our dataset generation. This dataset comprises of 182,954 prompts, each answered by seven LLMs. These answers are then scored by GPT-4 Bubeck et al. (2023).

To generate our dataset, we extract a subset of the data having (prompt, answer) pairs for six chosen models and use a pair of LLMs to score the models. Crucially, we make use of the same scoring rubric and prompt while querying both LLMs. Details of the chosen models, exact prompt and rubrics used can be found in the code attached in the supplementary material. We treat the scores given by the more advanced LLM as the ground-truth (i.e., proxies for human labels which are costly to collect at scale) and those by the less advanced LLM as coarse feedback.

By this method, we obtain Dataset-A of size 19,089 by querying LLama 3 3b (strong model) and LLama 3 1B (weak model) for each of the six models. We similarly obtain Dataset-B of size 5000 using scores from Gemini 2.5-Flash (strong model) and Gemini 1.5-Flash (weak model). The datasets can be found in the attached code.

C.4 Additional Experimental Results

C.4.1 Synthetic

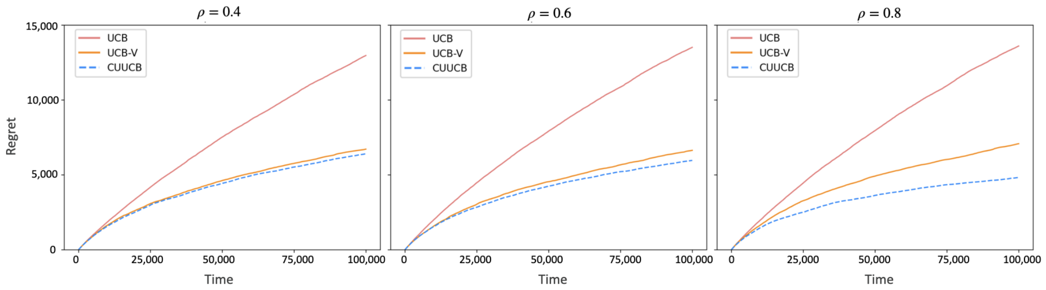

Figure 3: Comparison of CUUCB, UCB and UCB-V on synthetically generated data for a fixed value of $\alpha = 0.1$ and different values of ρ with a time horizon of $100,000$ steps, averaged across 40 runs.

We also run the synthetic experiments with fixed $\alpha = 0.1$ and varying ρ. We see from Figure 3 that increasing ρ leads to an improvement in the performance of CUUCB, which matches our expectations from theory. Since UCB and UCB-V only use ground-truth data, changing ρ does not affect their performance.

C.4.2 LLM-as-a-judge

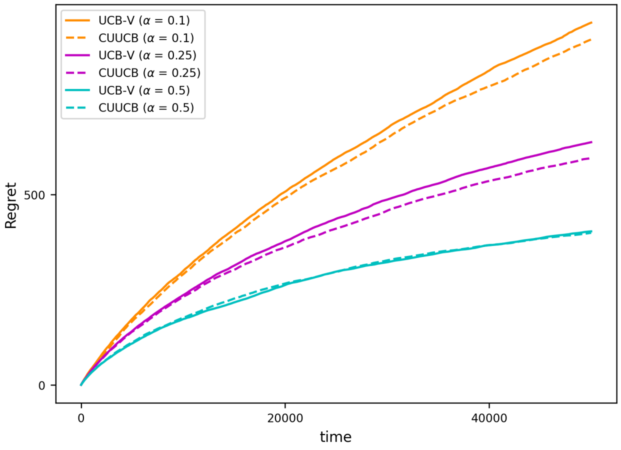

Figure 4: LLM benchmarking regret plots on Dataset-A of UCB-V which has access to only ground truth samples vs. CUUCB for different α, averaged across 40 runs.

We also run an experiment where we compare regrets of CUUCB and UCB-V on Dataset-A while varying the parameter α as shown in Figure 4. We see that our algorithm CUUCB achieves only a small improvement over UCB-V even for small α, while the improvement was more significant when

Dataset-B was used, as seen from Figure 2b. This shows that the performance gains of CUUCB are sensitive to the dataset instance.