RHOPE: LLM-AUGMENTED ADAPTIVE EXPERIMENTAL DESIGN WITH COMPLEX CONTEXTUAL INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Experimental design allocates limited measurement effort across candidate treatments to learn their expected outcomes. Adaptive designs refine allocations as evidence accumulates, improving estimation and treatment selection. Contextual information can inform these decisions, but exploiting unstructured or high-dimensional context remains challenging. We use large language models (LLMs) to transform complex context into proxy outcomes and propose Residual-Horizon Optimization for Prediction-Powered Experimentation (RHOPE), a principled Bayesian adaptive experimentation framework that combines true outcomes with LLM-generated proxy outcomes and formulates finite-batch design as a posterior Markov decision process (MDP). Our methodology uses proxy outcomes in two ways: sample augmentation improves estimation precision, while observing proxies before each allocation decision provides advance information. Synthetic experiments and semi-synthetic experiments using Genomics of Drug Sensitivity in Cancer (GDSC) data suggest that estimation precision drives most gains, with smaller advance-information gains, and highlight the importance of proxy informativeness and moment-estimation accuracy.

1 INTRODUCTION

Experimental design allocates experimental units among candidate treatments to estimate their mean outcomes under resource constraints. Unlike static designs, adaptive designs revise allocations as outcomes accumulate, improving estimation and treatment selection within a fixed budget. Adaptive experimental design is used in clinical trials, online platforms, recommender systems, and scientific screening (Villar et al., 2015; Burtini et al., 2015; Wu et al., 2018; Pérez et al., 2020; Liang & Bojinov, 2023). We study fixed-budget simple-regret best-arm identification (BAI): researchers sequentially collect data to identify the arm with the highest expected outcome. A fixed budget reflects a predetermined sample size or horizon; simple regret measures the terminal recommendation's expected-outcome shortfall, not merely whether the best arm is identified (Audibert et al., 2010; Bubeck et al., 2011; Russo, 2016; Zhao et al., 2023; Komiyama et al., 2024).

Prior work on fixed-budget simple-regret BAI has developed methods from both theoretical and algorithmic perspectives (Che & Namkoong, 2023; Che et al., 2024; Komiyama et al., 2024; Kato, 2025). The broader BAI literature considers objectives and settings beyond fixed-budget simple regret (Audibert et al., 2010; Karnin et al., 2013; Russo, 2016; Jourdan et al., 2022; Kato et al., 2022). Despite this substantial literature, two main issues remain insufficiently discussed.

The first issue concerns context. Contextual information captures unit heterogeneity and can inform experimental design (Kato & Ariu, 2026), but its use remains understudied in fixed-budget simple-regret BAI. Other BAI-related work estimates conditional moments of potential outcomes from experimental data (Kato et al., 2020; 2022). Under limited budgets, this approach scales poorly and is brittle with unstructured or high-dimensional context, such as text, images, genomic profiles, and multimodal records (Yang et al., 2013; Rindtorff et al., 2019; Dhawan et al., 2024).

The second issue is objective mismatch. Most of the BAI literature relies on heuristic or mimic algorithms (Karnin et al., 2013; Russo, 2016; Jourdan et al., 2022; Kato et al., 2022), whose allocation decisions are driven by surrogate objectives based on confidence bounds, posterior misidentification

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.

probabilities, or asymptotic error bounds. Our final objective, by contrast, is simple regret. Even when these algorithms attain optimal regret or error rates, unfavorable constants can impair finite-budget performance. In a one-armed bandit experiment, Adusumilli (2025) shows that Thompson sampling, as a heuristic algorithm, can incur twice the optimal regret despite having a favorable rate. Thus, optimizing a surrogate objective can weaken adaptive experimental design in practice.

To address these issues, we propose Residual-Horizon Optimization for Prediction-Powered Experimentation (RHOPE). Inspired by prediction-powered inference (Angelopoulos et al., 2023; 2026) and its Bayesian counterpart (Hofer et al., 2024), it uses LLMs as general-purpose generative models to turn rich context into auxiliary proxy outcomes, avoiding conditional-moment estimation from scarce experimental data. Knowledge encoded in LLMs can also support cold-start experiments. Building on Che & Namkoong (2023), RHOPE formulates a posterior Markov decision process (MDP) and applies RHO to directly optimize simple regret. Proxies contribute information for arm-mean estimation and an early look at upcoming-batch outcomes before allocation.

Contributions. Our contributions are threefold. *First,* we introduce an adaptive experimentation framework for fixed-budget simple-regret BAI that uses LLM-generated proxy outcomes without requiring a high-dimensional response model to be estimated from scarce experimental data. *Second,* we establish value comparisons for proxy-based estimation and advance information under the Gaussian sequential experiment. Experiments suggest that improved arm-mean estimation drives most gains, with smaller advance-information benefits. When moments are estimated, strong proxy–outcome correlation can offset the performance losses caused by moment-estimation error. *Third,* we show that, with known moments and a flat prior, the finite-batch working-posterior mean for each arm has an exact PPI++ representation (Angelopoulos et al., 2023; 2026). This algebraic connection applies arm by arm in a setting with multiple potential outcomes (De Bartolomeis et al., 2025; Shankar & Fiterau, 2025), connecting PPI++ to adaptive experimental design.

2 Related Work

Adaptive experimental design and bandits. Adaptive experimental design is often modeled as a multi-armed bandit problem. Depending on the objective, adaptive designs can be broadly categorized into variance-minimizing designs (Neyman, 1934; Zhao, 2024; Kato et al., 2020), cumulative-regret bandits (Lai & Robbins, 1985; Auer et al., 2002; Russo et al., 2018; Lattimore & Szepesvári, 2020), and pure-exploration problems such as BAI (Audibert et al., 2010; Bubeck et al., 2011; Russo, 2016; Kato & Ariu, 2026). A recent line of work instead uses diffusion limits to directly optimize experiment-level risk (Kuang & Wager, 2024; Fan & Glynn, 2021; Adusumilli, 2025; Che & Namkoong, 2023).

Best-arm identification. BAI has fixed-confidence (Garivier & Kaufmann, 2016; Kaufmann et al., 2016; Kato & Ariu, 2026) and fixed-budget settings. Within the fixed-budget setting, common objectives are the probability of error (Audibert et al., 2010; Karnin et al., 2013; Carpentier & Locatelli, 2016; Kato et al., 2022) and simple regret (Audibert et al., 2010; Bubeck et al., 2011; Russo, 2016; Zhao et al., 2023; Kato, 2025).

Prediction-powered inference. Prediction-powered inference (PPI) uses predictions from machine learning systems to improve statistical inference (Angelopoulos et al., 2023). Subsequent work develops power-tuned variants and multi-predictor extensions (Angelopoulos et al., 2026; Ji et al., 2025a; Miao et al., 2025; Shan et al., 2025; Carlson & Dell, 2025; Xu et al., 2025; Kluger et al., 2025; van der Laan & van der Laan, 2026). Bayesian PPI and Bayes-assisted frequentist PPI provide Bayesian and hybrid formulations of prediction-powered inference, respectively (Hofer et al., 2024; Cortinovis & Caron, 2025). Recent studies extend PPI ideas to causal inference (De Bartolomeis et al., 2025; Shankar & Fiterau, 2025) and to adaptive experimentation (Ji et al., 2025b; Ma et al., 2026; Ao et al., 2026).

3 Problem Setup

Following the setup of BAI in multi-armed bandits, we seek the highest-mean arm among $[K] := \{1, \dots, K\}$ over T batches. Batch $t = 0, \dots, T-1$ contains $b_t n$ units, where n is the total budget,

$b_t > 0$ is fixed and known, $\sum_{t=0}^{T-1} b_t = 1$, and $b_t n$ is an integer. We focus on this finite-batch structure for two reasons. First, updating the allocation policy entails nontrivial operational costs and engineering difficulties, particularly in medical and platform experiments; consequently, the number of reallocations is often finite (Bakshy et al., 2018; Chow & Chang, 2008). Second, under the finite-batch structure, each batch's outcome aggregates can be approximated by Gaussian signals without imposing parametric assumptions on individual-outcome distributions. This representation underpins the MDP formulation in Section 4.

For unit j in batch t, context $X_{t,j}$ may include high-dimensional covariates and unstructured information such as images and text. Let $Y_{t,j,a}$ be its true potential outcome under arm a. Querying an LLM with $X_{t,j}$ and arm information produces the proxy $\tilde{Y}_{t,j,a}$. We model these outcomes as

$$Y_{t,j,a} = h_a/\sqrt{n} + \epsilon_{t,j,a}, \qquad \tilde{Y}_{t,j,a} = \tilde{h}_a/\sqrt{n} + \tilde{\epsilon}_{t,j,a}. \tag{1}$$

For the centered noise terms $\epsilon_{t,j,a}$ and $\tilde{\epsilon}_{t,j,a}$, we define $\mathrm{Var}(\epsilon_{t,j,a}) = s_a^2$, $\mathrm{Var}(\tilde{\epsilon}_{t,j,a}) = \tilde{s}_a^2$, and $\mathrm{Cov}(\epsilon_{t,j,a}, \tilde{\epsilon}_{t,j,a}) = \varrho_a s_a \tilde{s}_a$, where ϱ_a is the arm-specific correlation between the true outcome and the proxy outcome. We assume $s_a^2 > 0$, $\tilde{s}_a^2 > 0$, and $|\varrho_a| < 1$ for every arm a, excluding degenerate noise distributions. The $1/\sqrt{n}$ local scaling and the treatment of variances and covariances as known are standard in diffusion-limit analyses (Kuang & Wager, 2024; Adusumilli, 2025). In practice, unknown outcome variances and covariances can instead be estimated from an initial batch; we discuss the impact of estimation error through the ablation experiments in Appendix D.

At the beginning of batch t, the researcher observes all contexts in the batch and generates $\tilde{Y}_{t,j,a}$ from $X_{t,j}$ for every unit–arm pair. After observing these proxy outcomes, the researcher chooses an allocation policy $\pi_t = (\pi_{t,1}, \ldots, \pi_{t,K}) \in \Delta_K$, where $\Delta_K := \{p \in [0,1]^K : \sum_a p_a = 1\}$, and assigns units in accordance with it. Let $A_{t,j}$ denote the treatment arm assigned to unit j; only the corresponding true potential outcome $Y_{t,j,A_{t,j}}$ is observed.

Throughout, we adopt a Bayesian perspective and treat the unknown local parameters $h = (h_1, \ldots, h_K)$ and $\tilde{h} = (\tilde{h}_1, \ldots, \tilde{h}_K)$ as random, with joint prior $(h, \tilde{h}) \sim \nu$. Let $a^*(h) \in \arg\max_a h_a$ be the best arm at the local scale, let $\hat{a}_T$ be the terminal recommendation, and let $\pi = (\pi_0, \ldots, \pi_{T-1})$ denote the complete adaptive allocation policy. Our final optimization target is Bayes simple regret,

$$\mathfrak{R}_T(\pi, \nu) = \mathbb{E}^{\pi}_{(h,\tilde{h})\sim\nu}\left[h_{a^*(h)} - h_{\hat{a}_T}\right]. \tag{2}$$

Intuitively, the proposed algorithm should therefore perform well on average over problem instances drawn from this joint prior.

4 LLM-Augmented Adaptive Experimental Design

We combine orthogonal mean estimation, a Gaussian posterior MDP, and RHO to obtain tractable batch-level allocation policies.

4.1 Orthogonal Decomposition for Mean Estimation

For each arm a and batch t, we estimate h_a and $\tilde{h}_a$ through the orthogonal decomposition

$$\theta_a^{(1)} := h_a - \lambda_a \tilde{h}_a, \qquad \theta_a^{(2)} := \tilde{h}_a. \tag{3}$$

The arm-wise projection coefficient is $\lambda_a := \mathrm{Cov}(\epsilon_{t,j,a}, \tilde{\epsilon}_{t,j,a})/\mathrm{Var}(\tilde{\epsilon}_{t,j,a}) = \varrho_a s_a/\tilde{s}_a$. Since $h_a = \theta_a^{(1)} + \lambda_a \theta_a^{(2)}$ and $\tilde{h}_a = \theta_a^{(2)}$, these transformed parameters determine the original means and are the targets of our estimation and posterior updates.

To estimate $\theta_a^{(1)}$ and $\theta_a^{(2)}$, define $\xi_{t,j,a} := \mathbf{1}\{A_{t,j} = a\}$. The corresponding aggregate estimators are

$$\hat{J}_{t,a} := \frac{1}{b_t\sqrt{n}} \sum_{j=1}^{b_t n} \tilde{Y}_{t,j,a}, \qquad \hat{G}_{t,a} := \frac{1}{b_t\sqrt{n}} \sum_{j=1}^{b_t n} \xi_{t,j,a}\left(Y_{t,j,a} - \lambda_a \tilde{Y}_{t,j,a}\right). \tag{4}$$

Write $\hat{J}_t = (\hat{J}_{t,a})_{a=1}^K$ and $\hat{G}_t = (\hat{G}_{t,a})_{a=1}^K$ for the all-arm vectors. Let $s_{a,\mathrm{res}}^2 := s_a^2 - \lambda_a^2 \tilde{s}_a^2 = s_a^2(1-\varrho_a^2)$. For fixed local parameters and a fixed allocation vector π_t, as $n \to \infty$, the central limit theorem gives

$$\hat{J}_{t,a} \xrightarrow{d} \mathcal{N}\left(\theta_a^{(2)}, \tilde{s}_a^2/b_t\right), \qquad \hat{G}_{t,a} \xrightarrow{d} \mathcal{N}\left(\pi_{t,a}\theta_a^{(1)}, \pi_{t,a} s_{a,\mathrm{res}}^2/b_t\right).$$

Under the stated moment conditions, these aggregates admit Gaussian approximations without Gaussian individual outcomes, following the diffusion-limit perspective of Kuang & Wager (2024); Fan & Glynn (2021); Adusumilli (2025); Che & Namkoong (2023).

Why use the orthogonal decomposition? First, the normalized aggregates correspond to efficient generalized method of moments (GMM) estimators.

Proposition 1 (GMM form of the orthogonal aggregates)**.** *Fix an arm a and a batch t with fixed allocation probability $\pi_{t,a} > 0$. Consider the GMM moments constructed from the labeled true outcome, the labeled proxy outcome, and the all-unit proxy outcome:*

$$\left(\frac{\xi_{t,j,a}}{\pi_{t,a}}\{\sqrt{n}\, Y_{t,j,a} - \theta_a^{(1)} - \lambda_a\theta_a^{(2)}\},\ \frac{\xi_{t,j,a}}{\pi_{t,a}}\{\sqrt{n}\,\tilde{Y}_{t,j,a} - \theta_a^{(2)}\},\ \sqrt{n}\,\tilde{Y}_{t,j,a} - \theta_a^{(2)}\right).$$

Under the optimal weighting matrix, the GMM estimators satisfy $\hat{\theta}_{t,a}^{(2)} = \hat{J}_{t,a} + o_p(1)$ and $\hat{\pi}_{t,a}\hat{\theta}_{t,a}^{(1)} = \hat{G}_{t,a}$, where $\hat{\pi}_{t,a} := (b_t n)^{-1}\sum_j \xi_{t,j,a}$ is the realized allocation proportion. On the event $\hat{\pi}_{t,a} > 0$, whose probability tends to one, the normalized orthogonal aggregates $(\hat{G}_{t,a}/\hat{\pi}_{t,a}, \hat{J}_{t,a})$ are asymptotically equivalent to the efficient GMM estimators $(\hat{\theta}_{t,a}^{(1)}, \hat{\theta}_{t,a}^{(2)})$.

The proof is in Appendix B.1. These efficient estimators facilitate downstream allocation decisions.

Second, from a computational perspective, conditional on fixed local parameters and for a fixed allocation probability $\pi_{t,a}$, the two aggregate estimators $(\hat{G}_{t,a}, \hat{J}_{t,a})$ are uncorrelated and, by the joint central limit theorem, asymptotically independent. This motivates the independent priors on $\theta_a^{(1)}$ and $\theta_a^{(2)}$ adopted in Section 4.2. Together with the Gaussian likelihood, the independent-prior assumption decouples the estimation of $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(2)}$, providing substantial computational convenience for the Bayesian posterior updates introduced in Section 4.2. A similar factorization of the posteriors for the prediction and correction components is used in Bayesian PPI (Hofer et al., 2024).

4.2 Bayesian Posterior MDP

The arm-wise Gaussian approximations in Section 4.1 motivate the following Gaussian sequential experiment and its posterior updates.

Definition 1 (Gaussian sequential experiment)**.** *A Gaussian sequential experiment, conditional on fixed local parameters, is characterized by observations $(J_0, G_0, \ldots, J_{T-1}, G_{T-1})$ satisfying*

$$J_t \sim \mathcal{N}\left(\boldsymbol{\theta}^{(2)}, \mathrm{diag}\left\{\tilde{s}_a^2/b_t\right\}_{a=1}^K\right), \tag{5}$$

and

$$G_t \mid J_0, G_0, \ldots, J_{t-1}, G_{t-1}, J_t \sim \mathcal{N}\left(\pi_t \odot \boldsymbol{\theta}^{(1)}, \mathrm{diag}\left\{\pi_{t,a} s_{a,\mathrm{res}}^2/b_t\right\}_{a=1}^K\right). \tag{6}$$

The proxy signal J_t is independent of the preceding observations, π_t is evaluated on the displayed history, and $\odot$ denotes elementwise multiplication.

For the aggregate-based policies in Assumption 1, the next theorem establishes the sequential Gaussian limit.

Theorem 1 (Gaussian sequential experiment limit)**.** *Under Assumption 1, for any fixed T and any fixed local parameter pair $(h, \tilde{h})$,*

$$(\hat{J}_0, \hat{G}_0, \ldots, \hat{J}_{T-1}, \hat{G}_{T-1}) \xrightarrow{d} (J_0, G_0, \ldots, J_{T-1}, G_{T-1}), \tag{7}$$

where the limiting process is the Gaussian sequential experiment defined in Definition 1.

Assumption 1 and the proof are given in Appendices B.2 and B.3, respectively. This limit motivates analyzing the Gaussian sequential experiment as the asymptotic framework underlying Algorithms 1–2; their finite-batch posterior updates use realized allocation proportions.

For the Gaussian experiment, we specify the joint prior ν through independent conjugate priors

$$\theta_a^{(\ell)} \sim \mathcal{N}\left(\mu_{0,a}^{(\ell)}, (\sigma_{0,a}^{(\ell)})^2\right), \qquad a \in [K], \quad \ell \in \{1,2\}.$$

The prior variances are finite and positive. This within- and across-arm independence mirrors the orthogonal decomposition and yields two scalar posterior updates per arm, simplifying computation and optimization.

Suppose that the posterior means and standard deviations upon completion of batch $t-1$ are $(\boldsymbol{\mu}_t^{(1)}, \boldsymbol{\sigma}_t^{(1)}, \boldsymbol{\mu}_t^{(2)}, \boldsymbol{\sigma}_t^{(2)})$. Given the Gaussian signals J_t followed by G_t, the standard posterior updates are, for each arm a,

$$(\sigma_{t+1,a}^{(2)})^{-2} = (\sigma_{t,a}^{(2)})^{-2} + \frac{b_t}{\tilde{s}_a^2}, \qquad \mu_{t+1,a}^{(2)} = (\sigma_{t+1,a}^{(2)})^2 \left[(\sigma_{t,a}^{(2)})^{-2}\mu_{t,a}^{(2)} + \frac{b_t}{\tilde{s}_a^2} J_{t,a}\right]. \tag{8}$$

$$(\sigma_{t+1,a}^{(1)})^{-2} = (\sigma_{t,a}^{(1)})^{-2} + \frac{\pi_{t,a} b_t}{s_{a,\mathrm{res}}^2}, \qquad \mu_{t+1,a}^{(1)} = (\sigma_{t+1,a}^{(1)})^2 \left[(\sigma_{t,a}^{(1)})^{-2}\mu_{t,a}^{(1)} + \frac{b_t}{s_{a,\mathrm{res}}^2} G_{t,a}\right]. \tag{9}$$

The posterior update induces the transition between the pre-LLM state x_t^- and post-LLM state x_t^+:

$$\begin{aligned} x_t^- = (\boldsymbol{\mu}_t^{(1)}, \boldsymbol{\sigma}_t^{(1)}, \boldsymbol{\mu}_t^{(2)}, \boldsymbol{\sigma}_t^{(2)}) &\xrightarrow{J_t} x_t^+ = (\boldsymbol{\mu}_t^{(1)}, \boldsymbol{\sigma}_t^{(1)}, \boldsymbol{\mu}_{t+1}^{(2)}, \boldsymbol{\sigma}_{t+1}^{(2)}) \\ &\xrightarrow{G_t} x_{t+1}^- = (\boldsymbol{\mu}_{t+1}^{(1)}, \boldsymbol{\sigma}_{t+1}^{(1)}, \boldsymbol{\mu}_{t+1}^{(2)}, \boldsymbol{\sigma}_{t+1}^{(2)}). \end{aligned}$$

Because $h_a = \theta_a^{(1)} + \lambda_a \theta_a^{(2)}$ and the two posterior components remain conditionally independent given the observations, after batch $t-1$ the local-scale mean h_a has a normal posterior distribution with mean and variance

$$\mu_{t,a} := \mu_{t,a}^{(1)} + \lambda_a \mu_{t,a}^{(2)}, \qquad \sigma_{t,a}^2 := (\sigma_{t,a}^{(1)})^2 + \lambda_a^2 (\sigma_{t,a}^{(2)})^2.$$

At the terminal stage, minimizing the Bayes simple regret in Equation 2 is equivalent to recommending the arm with the highest posterior mean. Under this default recommendation rule, optimizing the allocation policy is therefore equivalent to solving the following problem, where $V_t^{\pi,-}$ and $V_t^{\pi,+}$ distinguish values before and after observing J_t, respectively.[1]

$$\max_{\pi_0,\ldots,\pi_{T-1}} \left\{ V_0^{\pi,-}(\boldsymbol{\mu}_0^{(1)}, \boldsymbol{\sigma}_0^{(1)}, \boldsymbol{\mu}_0^{(2)}, \boldsymbol{\sigma}_0^{(2)}) := \mathbb{E}^\pi \left[\max_{a\in[K]} \left\{ \mu_{T,a}^{(1)} + \lambda_a \mu_{T,a}^{(2)} \right\} \right] \right\}. \tag{10}$$

To express Equation 10 in terms of each batch's allocation policy, we reparameterize the Gaussian posterior updates, connecting the current states x_t^- and x_t^+ to the terminal state x_T^-.

Lemma 1 (Posterior random-walk representation)**.** *For every batch t and arm a, the posterior means satisfy*

$$\mu_{t+1,a}^{(2)} = \mu_{t,a}^{(2)} + \sqrt{\frac{(\sigma_{t,a}^{(2)})^4 b_t}{\tilde{s}_a^2 + (\sigma_{t,a}^{(2)})^2 b_t}} Z_{t,a}^J, \qquad \mu_{t+1,a}^{(1)} = \mu_{t,a}^{(1)} + \sqrt{\frac{(\sigma_{t,a}^{(1)})^4 \pi_{t,a} b_t}{s_{a,\mathrm{res}}^2 + (\sigma_{t,a}^{(1)})^2 \pi_{t,a} b_t}} Z_{t,a}^G, \tag{11}$$

where $Z_{t,a}^J$ and $Z_{t,a}^G$ are independent standard normal innovations across arms, batches, and signal types.

From a post-LLM state, the value under policy π is

$$V_t^{\pi,+}(x_t^+) := \mathbb{E}^\pi \left[\max_{a\in[K]} \left\{ \mu_{T,a}^{(1)} + \lambda_a \mu_{T,a}^{(2)} \right\} \mid x_t^+ \right]. \tag{12}$$

[1] Here $\mathbb{E}^\pi$ denotes expectation over the random variables $J_0, G_0, \ldots, J_{T-1}, G_{T-1}$ under the prior predictive law.

Lemma 1 expands this value in future Gaussian innovations; the full expression is given in Appendix B.4.

We have now transformed the Bayes simple regret objective in Equation 2 into a dynamic-programming problem in Equation 12. Exact dynamic programming is nevertheless intractable because the state and action spaces are continuous. Residual horizon optimization (RHO) addresses this challenge by reducing the remaining-horizon policy optimization to a tractable single-simplex problem in each batch (Che & Namkoong, 2023).

4.3 Residual Horizon Optimization

In batch t, RHO jointly optimizes $(\pi_t, \ldots, \pi_{T-1})$ using only x_t^+, implements the current allocation, and replans after the next posterior update. The Gaussian posterior structure reduces this remaining-horizon optimization to a single simplex.

Lemma 2 (RHO reduction to one simplex)**.** *Let $B_t := \sum_{v=t}^{T-1} b_v$ for $t = 0, \ldots, T$, with an empty sum equal to zero. If the sequence of future allocation policies is optimized using only the currently available information x_t^+, then the RHO optimization is equivalent to the single-simplex maximization problem*

$$\rho_t \in \arg\max_{\bar{\rho} \in \Delta_K} \mathbb{E}_{Z^G, Z^J} \left[\max_{a \in [K]} \left\{ \begin{array}{l} \mu_{t,a}^{(1)} + \sqrt{\dfrac{(\sigma_{t,a}^{(1)})^4 \bar{\rho}_a B_t}{s_{a,\mathrm{res}}^2 + (\sigma_{t,a}^{(1)})^2 \bar{\rho}_a B_t}} Z_a^G \\ \quad + \lambda_a \mu_{t+1,a}^{(2)} + \lambda_a \sqrt{\dfrac{(\sigma_{t+1,a}^{(2)})^4 B_{t+1}}{\tilde{s}_a^2 + (\sigma_{t+1,a}^{(2)})^2 B_{t+1}}} Z_a^J \end{array} \right\} \right]. \tag{13}$$

Here Z_a^G and Z_a^J are independent standard normal random variables.

We approximate the expectation in Equation 13 by Monte Carlo and use gradient-based methods, such as SGD or Adam, to obtain a numerical allocation policy ρ_t. The complete finite-batch procedure is given in Algorithm 1 (Appendix F). Relative to the RHO formulation of Che & Namkoong (2023), proxy outcomes provide advance information about the upcoming batch (hence the use of $\mu_{t+1,a}^{(2)}$ rather than $\mu_{t,a}^{(2)}$) and reduce posterior uncertainty about arm means, supporting better allocation policies; Section 5.2 provides a detailed theoretical analysis.

5 Theoretical Analysis

We provide a theoretical analysis of RHOPE, examining its flat-prior connection to PPI++ (Section 5.1) and the two benefits of LLM-based sample augmentation: improved estimation precision and advance information for upcoming-batch allocation (Section 5.2).

5.1 Relationship with PPI++

PPI++ combines labeled outcomes with predictions through a tuned correction (Angelopoulos et al., 2026). We relate its mean-estimation formula to our finite-batch working-posterior mean, accounting for pooled predictions and realized sample counts.

Theorem 2 (Flat-prior PPI++ representation)**.** *Consider the finite-batch Gaussian working-posterior updates in Algorithms 1 and 2, with fixed known moments, $\tilde{s}_a^2 > 0$, $s_{a,\mathrm{res}}^2 > 0$, and zero initial precisions. Fix $t \in \{1, \ldots, T\}$ and an arm a such that $\sum_{v=0}^{t-1} \hat{\pi}_{v,a} b_v > 0$. Let $\bar{Y}_{0:t-1,a}^{\mathrm{obs}}$ and $\bar{\tilde{Y}}_{0:t-1,a}^{\mathrm{obs}}$ be the true-outcome and proxy sample means among units assigned to arm a in batches $0, \ldots, t-1$, normalized by their actual count, and let $\bar{\tilde{Y}}_{0:t-1,a}^{\mathrm{all}}$ be the corresponding all-unit proxy mean. Then the working-posterior mean $\mu_{n,t,a} := \mu_{n,t,a}^{(1)} + \lambda_a \mu_{n,t,a}^{(2)}$ for h_a satisfies*

$$\mu_{n,t,a} = \sqrt{n} \left(\bar{Y}_{0:t-1,a}^{\mathrm{obs}} - \lambda_a \bar{\tilde{Y}}_{0:t-1,a}^{\mathrm{obs}} \right) + \lambda_a \sqrt{n}\, \bar{\tilde{Y}}_{0:t-1,a}^{\mathrm{all}}. \tag{14}$$

Up to the $\sqrt{n}$ scaling, this is the pooled-sample PPI++ formula, with the tuning-coefficient conversion given in the proof.

Theorem 2 establishes an exact algebraic identity, without transferring i.i.d.-sampling guarantees to adaptive data. With a nonflat prior, the orthogonal aggregates are combined with prior information through Gaussian posterior updates. Whereas conventional PPI centers on a single outcome, our BAI setting has one potential outcome per arm, so the connection holds arm by arm. Under the theorem's conditions, proxy outcomes reduce working-posterior variance for the same realized allocation, as shown in the proof. This connection extends PPI++ from statistical inference to adaptive experimental design.

5.2 Benefits of LLM-Generated Proxy Outcomes

Within the Gaussian sequential experiment, we compare three methods to isolate the estimation and advance-information benefits of proxy outcomes.

RHO. RHO (Che & Namkoong, 2023) uses only true outcomes, without proxy outcomes or their contextual information.

RHOPE-Delayed. RHOPE-Delayed adds proxy outcomes and their contextual information to RHO but uses the current batch's proxies only after allocation. This delayed-information ablation isolates the value of observing proxies before allocation. The complete procedure is given in Algorithm 2.

RHOPE. RHOPE is our proposed method. In addition to using proxy outcomes and their contextual information, it uses the current batch's proxy outcomes when optimizing the current allocation policy. The complete procedure is given in Algorithm 1.

Under the RHO approximation, denote the fixed-allocation expected returns of RHO, RHOPE-Delayed, and RHOPE by $\widetilde{V}_t^{\mathrm{RHO}}(\bar{\rho}, x_t^-)$, $\widetilde{V}_t^{\mathrm{RHOPE\text{-}Delayed}}(\bar{\rho}, x_t^-)$, and $\widetilde{V}_t^{\mathrm{RHOPE}}(\bar{\rho}, x_t^+)$, respectively. Here $\bar{\rho} \in \Delta_K$ is fixed, and x_t^- and x_t^+ are the pre-LLM and post-LLM states.

We compare returns at a common pre-LLM state, initializing RHO with the same marginal posterior for h as the two proxy-based methods.

Theorem 3 (RHOPE return comparison)**.** *For any fixed pre-LLM state x_t^-,*

$$\max_{\bar{\rho}\in\Delta_K} \widetilde{V}_t^{\mathrm{RHO}}(\bar{\rho}, x_t^-) \le \max_{\bar{\rho}\in\Delta_K} \widetilde{V}_t^{\mathrm{RHOPE\text{-}Delayed}}(\bar{\rho}, x_t^-) \le \mathbb{E}_{J_t}\left[\max_{\bar{\rho}\in\Delta_K} \widetilde{V}_t^{\mathrm{RHOPE}}(\bar{\rho}, x_t^+) \mid x_t^-\right]. \tag{15}$$

Theorem 3 (proved in Appendix B.8) identifies two benefits of sample augmentation, which supplies proxy information for every arm, including those receiving no units. First, combining proxies with true outcomes improves estimation precision by reducing posterior uncertainty about arm means. Second, observing current-batch proxies before allocation provides advance information for better allocation decisions.

The advance-information benefit appears not only in expected returns under the RHO approximation but also in the optimal MDP value. Let $V_t^{\mathrm{RHOPE\text{-}Delayed}}(x_t^-)$ and $V_t^{\mathrm{RHOPE}}(x_t^+)$ denote the optimal MDP values under the information structures of RHOPE-Delayed and RHOPE, respectively, not the returns of their RHO policies.

Theorem 4 (MDP value of advance information)**.** *For any fixed pre-LLM state x_t^-,*

$$\mathbb{E}_{J_t}\left[V_t^{\mathrm{RHOPE}}(x_t^+) \mid x_t^-\right] \ge V_t^{\mathrm{RHOPE\text{-}Delayed}}(x_t^-). \tag{16}$$

The proof is in Appendix B.9. Notably, the two information channels need not contribute equally. Experiments in Section 6 suggest that estimation precision drives most observed gains, with smaller gains from advance information. Estimated-moment experiments also assess how proxy informativeness and moment-estimation error affect these gains.

6 Experiments

Synthetic and GDSC-based semi-synthetic experiments evaluate finite-sample decision gains and their dependence on proxy informativeness and moment estimation.

Table 1: Final relative simple regret (%) for (a) Bernoulli and (b) Gumbel with $\sigma^2 = 1$; uniform is 100 in every column, and lower is better. Boldface marks column minima; dashes denote non-LLM results unchanged across ϱ.

(a) Bernoulli setting

	$b_t n = 100$			$b_t n = 1000$		
Policy / ϱ	0.2	0.4	0.6	0.2	0.4	0.6
Without LLM						
uniform	100.00	–	–	100.00	–	–
TS	79.78	–	–	77.74	–	–
TTTS	78.64	–	–	76.89	–	–
Sequential Halving	77.17	–	–	75.00	–	–
TS-SPAS	108.89	–	–	107.31	–	–
myopic	76.87	–	–	74.11	–	–
RHO	74.48	–	–	71.37	–	–
With LLM						
uniform-LLM	100.45	95.48	83.46	94.18	87.73	79.15
TS-LLM	77.51	73.78	63.58	73.91	69.16	59.96
TTTS-LLM	77.33	70.36	62.85	75.93	70.28	61.37
MLA-UCB	97.09	92.35	75.62	92.45	86.39	77.24
myopic-LLM	74.93	70.43	61.64	74.39	68.79	59.23
RHOPE-DELAYED	71.68	**66.43**	59.37	69.95	65.86	57.89
RHOPE	**71.43**	67.05	**57.88**	**69.82**	**65.47**	**56.66**

(b) Gumbel setting ($\sigma^2 = 1$)

	$b_t n = 100$			$b_t n = 1000$		
Policy / ϱ	0.2	0.4	0.6	0.2	0.4	0.6
Without LLM						
uniform	100.00	–	–	100.00	–	–
TS	63.76	–	–	70.18	–	–
TTTS	66.48	–	–	71.85	–	–
Sequential Halving	64.90	–	–	69.35	–	–
TS-SPAS	125.62	–	–	125.24	–	–
myopic	63.82	–	–	68.75	–	–
RHO	61.63	–	–	65.06	–	–
With LLM						
uniform-LLM	97.04	89.79	76.82	96.23	88.72	76.41
TS-LLM	62.59	56.32	49.40	66.91	63.62	51.97
TTTS-LLM	65.57	59.76	46.89	68.17	62.11	52.56
MLA-UCB	91.15	89.08	73.42	98.73	89.78	74.20
myopic-LLM	61.97	55.60	47.95	67.72	61.42	50.87
RHOPE-DELAYED	60.02	**51.36**	44.94	**62.83**	58.43	49.24
RHOPE	**58.34**	52.91	**44.45**	63.11	**56.23**	**48.45**

6.1 SYNTHETIC DATA EXPERIMENTS

The main synthetic experiments pair true and proxy outcomes with Bernoulli or Gumbel marginals, using equal-sized batches of $b_t n \in \{100, 1000\}$ and Pearson correlation $\varrho \in \{0.2, 0.4, 0.6\}$. The main Gumbel variance is $\sigma^2 = 1$; policies use known variances and covariances. Appendix C specifies the distributions and parameter priors and reports the more complex Bernoulli–Gumbel setting and Gumbel variance ablations with $\sigma^2 \in \{1/3, 3\}$.

We compare RHOPE with uniform allocation, Thompson sampling (TS), top-two Thompson sampling (TTTS) (Russo, 2016), MLA-UCB (Ji et al., 2025b), sequential halving (Karnin et al., 2013; Zhao et al., 2023), TS-SPAS (Kato, 2025), myopic one-step optimization, RHO (Che & Namkoong, 2023), and LLM-assisted counterparts where available. Uniform allocation provides the 100%

Table 2: GDSC relative simple regret (%) with known moments. Proxies are unaugmented LLM outputs at $v = 0$ and synthetically augmented at $v > 0$. Uniform is 100; lower is better. Boldface marks column minima; dashes denote unchanged non-LLM results.

	$v = 0$			$v = 2$			$v = 4.5$		
Policy / shots	0	4	16	0	4	16	0	4	16
Without LLM									
uniform	100.00	–	–	–	–	–	–	–	–
TS	71.21	–	–	–	–	–	–	–	–
TTTS	67.59	–	–	–	–	–	–	–	–
Sequential Halving	74.85	–	–	–	–	–	–	–	–
TS-SPAS	171.68	–	–	–	–	–	–	–	–
myopic	65.76	–	–	–	–	–	–	–	–
RHO	64.51	–	–	–	–	–	–	–	–
With LLM									
uniform-LLM	95.91	95.82	94.24	93.76	89.84	90.36	87.21	82.90	82.92
TS-LLM	69.87	68.91	67.20	66.58	62.25	61.47	60.06	55.11	54.31
TTTS-LLM	67.79	66.46	63.50	64.77	62.16	60.85	58.21	54.67	53.08
MLA-UCB	71.82	73.89	73.05	71.16	66.76	64.90	60.19	54.38	54.17
myopic-LLM	65.66	64.63	65.98	63.42	60.00	59.74	58.07	51.92	51.64
RHOPE-DELAYED	**64.02**	**63.06**	63.35	**61.18**	**58.37**	58.73	58.31	52.68	**51.38**
RHOPE	64.36	64.07	**62.36**	62.79	58.82	**58.30**	**57.79**	**51.41**	51.65

simple-regret reference in each setting. Table 1 presents the main results; Appendix C.1 provides experimental details and Appendix D presents ablation experiments with estimated moments.

Across Tables 1 and 3, either RHOPE or RHOPE-DELAYED attains the lowest relative simple regret in every setting, showing that directly optimizing Bayes simple regret with informative proxies can outperform heuristic algorithms. Their similar performance suggests that estimation precision drives most gains, with a modest additional benefit from advance information. LLM-assisted methods generally improve on their non-LLM counterparts, increasingly so at larger ϱ. These findings are consistent with Section 5.

6.2 SEMI-SYNTHETIC DATA EXPERIMENT: GDSC DATASET

The GDSC-based benchmark uses seven cancer drugs and 896 cell lines with 18,522 covariates (Yang et al., 2013). DeepSeek V4 Flash generates response scores from selected features using zero-shot, 4-shot, and 16-shot prompts. To vary proxy informativeness, we evaluate unaugmented LLM outputs at $v = 0$ and add a synthetic signal correlated with simulated patient-level outcome noise at $v \in \{2, 4.5\}$, keeping true outcomes unchanged. As v increases, proxy outcomes become more strongly correlated with true outcomes and hence more informative (Appendix C.2).

With known moments, the two RHOPE variants attain the lowest point estimates across all nine conditions (Table 2). With estimated moments, their advantage over RHO largely disappears at $v = 0$, but both outperform RHO under controlled augmentation ($v > 0$) (Table 7, Appendix D). This pattern suggests that weak proxy–outcome correlation offers limited precision gains, which may not offset the effects of moment-estimation error on posterior updates and allocation. Stronger correlation can compensate for these losses, improving decisions despite estimated moments. Practical gains therefore depend jointly on proxy informativeness and moment-estimation accuracy.

7 CONCLUSION

We introduced RHOPE, a framework incorporating contextual LLM-generated proxy outcomes into posterior updating and allocation for finite-batch BAI. Our analysis distinguishes estimation and advance-information benefits, while experiments suggest estimation precision drives most gains. Both synthetic and GDSC-based semi-synthetic experiments highlight the importance of proxy informativeness and moment-estimation accuracy.

AI USE STATEMENT

DeepSeek V4 Flash was used as part of the experimental procedure to generate the auxiliary proxy outcomes evaluated in the semi-synthetic experiment. Generative AI tools were also used to assist with writing proofs, writing code, generating figures, reviewing methodological choices, interpreting results, translating and polishing text, and organizing references. The authors manually reviewed and verified all citations, proofs, code, data processing, and reported results. The authors take full responsibility for the final content of this work, including all text, claims, code, results, figures, and other artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

The assumptions and complete proofs for the theoretical results are provided in Appendix B. Appendix C documents the experimental settings, data construction, prompting procedure, benchmark policies, and additional results. The complete finite-batch implementations of RHOPE and RHOPE-DELAYED are specified in Appendix F. Appendix D reports ablations with estimated moments for both synthetic and semi-synthetic experiments, estimating the required outcome moments from the first batch.

REFERENCES

Karun Adusumilli. Risk and optimal policies in bandit experiments. *Econometrica*, 93(3):1003–1029, 2025. doi: 10.3982/ECTA21075.

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

Anastasios N. Angelopoulos, John C. Duchi, and Tijana Zrnic. PPI++: Efficient prediction-powered inference. *The Annals of Applied Statistics*, 2026. To appear.

Ruicheng Ao, Hongyu Chen, Siyang Gao, Hanwei Li, and David Simchi-Levi. Designing service systems from textual evidence. *arXiv preprint arXiv:2603.10400*, 2026.

Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *Proceedings of the 23rd Conference on Learning Theory*, pp. 41–53. Omnipress, 2010.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2–3):235–256, 2002.

Eytan Bakshy, Lili Dworkin, Brian Karrer, Konstantin Kashin, Benjamin Letham, Ashwin Murthy, and Shaun Singh. AE: A domain-agnostic platform for adaptive experimentation. In *Neural Information Processing Systems Workshop on Systems for Machine Learning*, pp. 1–8, 2018.

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412(19):1832–1852, 2011. doi: 10.1016/j.tcs.2010.12.059.

Giuseppe Burtini, Jason Loeppky, and Ramon Lawrence. A survey of online experiment design with the stochastic multi-armed bandit. *arXiv preprint arXiv:1510.00757*, 2015.

Jacob Carlson and Melissa Dell. A unifying framework for robust and efficient inference with unstructured data. *arXiv preprint arXiv:2505.00282*, 2025.

Alexandra Carpentier and Andrea Locatelli. Tight (lower) bounds for the fixed budget best arm identification bandit problem. In *Proceedings of the 29th Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 590–604. PMLR, 2016.

Ethan Che and Hongseok Namkoong. Adaptive experimentation at scale: A computational framework for flexible batches. *arXiv preprint arXiv:2303.11582*, 2023.

Ethan Che, Daniel R. Jiang, Hongseok Namkoong, and Jimmy Wang. Optimization-driven adaptive experimentation. *arXiv preprint arXiv:2408.04570*, 2024.

Shein-Chung Chow and Mark Chang. Adaptive design methods in clinical trials: A review. *Orphanet Journal of Rare Diseases*, 3:11, 2008. doi: 10.1186/1750-1172-3-11.

Stefano Cortinovis and François Caron. FAB-PPI: Frequentist, assisted by bayes, prediction-powered inference. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 11328–11356. PMLR, 2025.

Piersilvio De Bartolomeis, Javier Abad, Guanbo Wang, Konstantin Donhauser, Raymond M. Duch, Fanny Yang, and Issa J. Dahabreh. Efficient randomized experiments using foundation models. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Nikita Dhawan, Leonardo Cotta, Karen Ullrich, Rahul G. Krishnan, and Chris J. Maddison. End-to-end causal effect estimation from unstructured natural language data. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2455.

Lin Fan and Peter W. Glynn. Diffusion approximations for thompson sampling in the small gap regime. *arXiv preprint arXiv:2105.09232*, 2021. v6, revised April 2026.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Proceedings of the 29th Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 998–1027. PMLR, 2016.

R. Alex Hofer, Joshua Maynez, Bhuwan Dhingra, Adam Fisch, Amir Globerson, and William W. Cohen. Bayesian prediction-powered inference. *arXiv preprint arXiv:2405.06034*, 2024. doi: 10.48550/arXiv.2405.06034.

Wenlong Ji, Lihua Lei, and Tijana Zrnic. Predictions as surrogates: Revisiting surrogate outcomes in the age of AI. *arXiv preprint arXiv:2501.09731*, 2025a.

Wenlong Ji, Yihan Pan, Ruihao Zhu, and Lihua Lei. Multi-armed bandits with machine learning-generated surrogate rewards. *arXiv preprint arXiv:2506.16658*, 2025b.

Marc Jourdan, Rémy Degenne, Dorian Baudry, Rianne de Heide, and Emilie Kaufmann. Top two algorithms revisited. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26791–26803. Curran Associates, Inc., 2022.

Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1238–1246. PMLR, 2013.

Masahiro Kato. Minimax and bayes optimal best-arm identification. *arXiv preprint arXiv:2506.24007*, 2025.

Masahiro Kato and Kaito Ariu. The role of contextual information in best arm identification. *Journal of Machine Learning Research*, 27:1–61, 2026.

Masahiro Kato, Takuya Ishihara, Junya Honda, and Yusuke Narita. Efficient adaptive experimental design for average treatment effect estimation. *arXiv preprint arXiv:2002.05308*, 2020.

Masahiro Kato, Masaaki Imaizumi, Takuya Ishihara, and Toru Kitagawa. Best arm identification with contextual information under a small gap. *arXiv preprint arXiv:2209.07330*, 2022.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016.

Dan M. Kluger, Kerri Lu, Tijana Zrnic, Sherrie Wang, and Stephen Bates. Prediction-powered inference with imputed covariates and nonuniform sampling. *arXiv preprint arXiv:2501.18577*, 2025.

Junpei Komiyama, Kaito Ariu, Masahiro Kato, and Chao Qin. Rate-optimal bayesian simple regret in best arm identification. *Mathematics of Operations Research*, 49(3):1629–1646, 2024. doi: 10.1287/moor.2022.0011.

Xu Kuang and Stefan Wager. Weak signal asymptotics for sequentially randomized experiments. *Management Science*, 70(10):7024–7041, 2024. doi: 10.1287/mnsc.2023.4964.

Tze Leung 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. doi: 10.1017/9781108571401.

Biyonka Liang and Iavor Bojinov. An experimental design for anytime-valid causal inference on multi-armed bandits. *arXiv preprint arXiv:2311.05794*, 2023.

Tianyi Ma, Hanzhang Qin, Ruihao Zhu, and Jierui Zuo. Best-arm identification with generative proxy. *arXiv preprint arXiv:2607.06879*, 2026.

Jiacheng Miao, Xinran Miao, Yixuan Wu, Jiwei Zhao, and Qiongshi Lu. Assumption-lean and data-adaptive post-prediction inference. *Journal of Machine Learning Research*, 26(179):1–31, 2025.

Jerzy Neyman. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4): 558–606, 1934. doi: 10.1111/j.2397-2335.1934.tb04184.x.

Adrià Pérez, Pablo Herrera-Nieto, Stefan Doerr, and Gianni De Fabritiis. AdaptiveBandit: A multi-armed bandit framework for adaptive sampling in molecular simulations. *Journal of Chemical Theory and Computation*, 16(7):4685–4693, 2020. doi: 10.1021/acs.jctc.0c00205.

Niklas T. Rindtorff, MingYu Lu, Nisarg A. Patel, Huahua Zheng, and Alexander D'Amour. A biologically plausible benchmark for contextual bandit algorithms in precision oncology using in vitro data. *arXiv preprint arXiv:1911.04389*, 2019.

Daniel Russo. Simple bayesian algorithms for best arm identification. In *Proceedings of the 29th Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 1417–1418. PMLR, 2016.

Daniel Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on thompson sampling. *Foundations and Trends in Machine Learning*, 11(1):1–96, 2018.

Jiawei Shan, Yiming Dong, and Jiwei Zhao. SADA: Safe and adaptive inference with multiple black-box predictions. *arXiv preprint arXiv:2509.21707*, 2025.

Shiv Shankar and Madalina Fiterau. Augmenting randomized controlled trials with foundation models as synthetic units. In *Proceedings of the 16th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 1–8. ACM, 2025. doi: 10.1145/3765612.3767256.

James R. Swisher, Sheldon H. Jacobson, and Enver Yücesan. Discrete-event simulation optimization using ranking, selection, and multiple comparison procedures: A survey. *ACM Transactions on Modeling and Computer Simulation*, 13(2):134–154, 2003. doi: 10.1145/858481.858484.

Lars van der Laan and Mark van der Laan. Calibeating prediction-powered inference. *arXiv preprint arXiv:2604.21260*, 2026.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, 1996.

Sofía S. Villar, Jack Bowden, and James Wason. Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges. *Statistical Science*, 30(2):199–215, 2015.

Qingyun Wu, Naveen Iyer, and Hongning Wang. Learning contextual bandits in a non-stationary environment. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 495–504, 2018.

Zichun Xu, Daniela Witten, and Ali Shojaie. A unified framework for semiparametrically efficient semi-supervised learning. *arXiv preprint arXiv:2502.17741*, 2025.

Wanjuan Yang, Jorge Soares, Patricia Greninger, Elena J. Edelman, Howard Lightfoot, Simon Forbes, Nidhi Bindal, Dave Beare, James A. Smith, I. Richard Thompson, et al. Genomics of drug sensitivity in cancer (GDSC): A resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research*, 41(D1):D955–D961, 2013.

Jinglong Zhao. Adaptive neyman allocation. In *Proceedings of the 25th ACM Conference on Economics and Computation*, pp. 776. Association for Computing Machinery, 2024. doi: 10.1145/3670865.3673535.

Yao Zhao, Connor Stephens, Csaba Szepesvári, and Kwang-Sung Jun. Revisiting simple regret: Fast rates for returning a good arm. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 42110–42158, 2023.

A Extended Related Work

Adaptive experimental design and bandits. Adaptive experimental design is often modeled as a multi-armed bandit problem because both involve sequential sampling under partial feedback. However, the appropriate design and algorithm can differ substantially across optimization objectives. Variance-minimizing designs allocate samples to reduce uncertainty in target estimators (Neyman, 1934; Zhao, 2024; Kato et al., 2020). Cumulative-regret bandits trade off exploration and exploitation during the sampling process (Lai & Robbins, 1985; Auer et al., 2002; Russo et al., 2018; Lattimore & Szepesvári, 2020). Pure-exploration objectives, including BAI, evaluate the final recommendation, not the cumulative outcomes collected during sampling (Audibert et al., 2010; Bubeck et al., 2011; Russo, 2016; Kato & Ariu, 2026). Most algorithms in these areas are driven by confidence bounds, posterior sampling, or elimination heuristics.

A recent line of work instead uses diffusion limits to directly optimize experiment-level risk. Kuang & Wager (2024) study weak-signal asymptotics for sequentially randomized experiments, and Fan & Glynn (2021) derive diffusion approximations for Thompson sampling. Adusumilli (2025) formulates bandit experiments as Bayes and minimax risk problems in a diffusion regime, defines asymptotic Bayes risk as expected cumulative regret, and characterizes the minimal Bayes risk through the associated partial differential equation, the continuous-time limit of the dynamic program. Che & Namkoong (2023) approximate batched adaptive experiments by Gaussian sequential experiments and use residual horizon optimization (RHO) to optimize the allocation policy. Our algorithm follows this line: rather than using a classical bandit heuristic, it formulates a Gaussian posterior MDP and optimizes the Bayes simple regret objective.

Best-arm identification. BAI is a pure-exploration problem that asks researchers to identify the arm with the largest mean. It is closely related to ranking-and-selection problems in simulation optimization (Swisher et al., 2003). BAI is broadly divided into fixed-confidence and fixed-budget settings. In the fixed-confidence setting, the goal is to stop as quickly as possible while controlling the probability of error (Garivier & Kaufmann, 2016; Kaufmann et al., 2016; Kato & Ariu, 2026). In the fixed-budget setting, common objectives are to minimize either the probability of error (Audibert et al., 2010; Karnin et al., 2013; Carpentier & Locatelli, 2016; Kato et al., 2022) or simple regret (Audibert et al., 2010; Bubeck et al., 2011; Russo, 2016; Zhao et al., 2023; Kato, 2025). The probability-of-error criterion considers only whether the selected arm is optimal, whereas simple regret also accounts for the expected-outcome gap between the selected and optimal arms. This paper studies fixed-budget simple-regret BAI.

Prediction-powered inference. Prediction-powered inference (PPI) uses predictions from machine learning systems to improve statistical inference when labels are expensive and unlabeled covariates are abundant (Angelopoulos et al., 2023). PPI++ improves efficiency by tuning the coefficient of a prediction-based correction (Angelopoulos et al., 2026). Recent work studies additional variants and multi-predictor extensions (Ji et al., 2025a; Miao et al., 2025; Shan et al., 2025; Carlson & Dell, 2025; Xu et al., 2025; Kluger et al., 2025; van der Laan & van der Laan, 2026). Bayesian and Bayes-assisted frequentist formulations have also been developed (Hofer et al., 2024; Cortinovis & Caron, 2025). Recent studies extend PPI ideas to causal inference (De Bartolomeis et al., 2025; Shankar & Fiterau, 2025). Adaptive experimentation is naturally a missing-data problem: each sampled unit reveals only the outcome under the assigned arm. Thus, although our objective is decision

quality rather than efficient inference alone, PPI ideas can also be used in adaptive experimentation (Ji et al., 2025b; Ma et al., 2026; Ao et al., 2026). We use prediction-powered aggregates in posterior MDP state transitions and, on this basis, construct an optimization framework for adaptive experimentation with Bayes simple regret as its objective.

B PROOFS

B.1 PROOF OF PROPOSITION 1

Fix the local parameters, one arm, and one batch, and suppress the arm and batch indices. Let the batch size be $M = b_t n$, let $S_i = \mathbf{1}\{A_i = a\}$, let $\pi = \Pr(S_i = 1)$, and write

$$R_i = \sqrt{n}\, Y_i, \qquad \tilde{R}_i = \sqrt{n}\, \tilde{Y}_i.$$

Then $\mathbb{E}[R_i] = h$, $\mathbb{E}[\tilde{R}_i] = \tilde{h}$, $\mathrm{Var}(R_i) = ns^2$, $\mathrm{Var}(\tilde{R}_i) = n\tilde{s}^2$, and $\mathrm{Cov}(R_i, \tilde{R}_i) = nc$, where $c := \varrho s\tilde{s}$. The factor n appears because the GMM moments use the local-scale variables $\sqrt{n}\, Y_i$ and $\sqrt{n}\, \tilde{Y}_i$. This is a fixed-π, one-batch GMM calculation; under randomized assignment, S_i is independent of $(R_i, \tilde{R}_i)$. The raw GMM construction uses the three moments

$$\begin{pmatrix} \frac{S_i}{\pi}(R_i - h) \\ \frac{S_i}{\pi}(\tilde{R}_i - \tilde{h}) \\ \tilde{R}_i - \tilde{h} \end{pmatrix}.$$

They correspond respectively to the labeled true outcome, the labeled proxy outcome, and the all-unit proxy outcome. Their covariance matrix is

$$F = n \begin{pmatrix} \pi^{-1}s^2 & \pi^{-1}c & c \\ \pi^{-1}c & \pi^{-1}\tilde{s}^2 & \tilde{s}^2 \\ c & \tilde{s}^2 & \tilde{s}^2 \end{pmatrix}.$$

For $0 < \pi < 1$, the nondegeneracy conditions $s^2 > 0$, $\tilde{s}^2 > 0$, and $|\varrho| < 1$ make F positive definite. We first consider this case. The optimally weighted GMM criterion uses weighting matrix F^{-1}:

$$Q(h, \tilde{h}) = \bar{m}(h, \tilde{h})^\top F^{-1} \bar{m}(h, \tilde{h}),$$

where $\bar{m}(h, \tilde{h})$ is the sample average of the displayed moment vector. The common factor n in F only rescales the optimal GMM criterion and does not change its minimizer.

Set $\lambda = c/\tilde{s}^2 = \varrho s/\tilde{s}$, and define the transformed parameters

$$\theta^{(1)} = h - \lambda\tilde{h}, \qquad \theta^{(2)} = \tilde{h}.$$

Because GMM with the optimal weighting matrix is invariant to nonsingular linear transformations of the moment vector, we apply

$$L^{-1} = \begin{pmatrix} 1 & -\lambda & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix}.$$

Then

$$L^{-1} \begin{pmatrix} \frac{S_i}{\pi}(R_i - h) \\ \frac{S_i}{\pi}(\tilde{R}_i - \tilde{h}) \\ \tilde{R}_i - \tilde{h} \end{pmatrix} = \begin{pmatrix} \frac{S_i}{\pi}\{R_i - \lambda\tilde{R}_i - \theta^{(1)}\} \\ \left(\frac{S_i}{\pi} - 1\right)(\tilde{R}_i - \theta^{(2)}) \\ \tilde{R}_i - \theta^{(2)} \end{pmatrix}.$$

The first transformed moment is the residual score. It is orthogonal to the proxy score because

$$\mathrm{Cov}(R_i - \lambda\tilde{R}_i, \tilde{R}_i) = n\left(c - \lambda\tilde{s}^2\right) = 0.$$

Since S_i is independent of the outcomes, the transformed covariance matrix is diagonal:

$$L^{-1}F(L^{-1})^\top = n \begin{pmatrix} \pi^{-1}s_{\mathrm{res}}^2 & 0 & 0 \\ 0 & (\pi^{-1} - 1)\tilde{s}^2 & 0 \\ 0 & 0 & \tilde{s}^2 \end{pmatrix}, \qquad s_{\mathrm{res}}^2 = s^2 - \lambda^2\tilde{s}^2.$$

Thus the optimal weighting matrix in the transformed coordinates simply weights three mutually orthogonal components. The second transformed moment, $(S_i/\pi - 1)(\tilde{R}_i - \theta^{(2)})$, has expectation zero for every value of $(\theta^{(1)}, \theta^{(2)})$. Its population derivative with respect to $\theta^{(2)}$ is also zero because

$$-\mathbb{E}\left[\frac{S_i}{\pi} - 1\right] = 0.$$

It therefore contributes no first-order information about the target means. Write $N = \sum_{i=1}^{M} S_i$, $\hat{\pi} = N/M$, and $\hat{J} = M^{-1}\sum_{i=1}^{M} \tilde{R}_i$. Because $\pi > 0$, $\Pr(N = 0) = (1-\pi)^M \to 0$; the following estimator calculations are on the event $N > 0$. More explicitly, if $\bar{D}_M$ denotes the sample Jacobian of the transformed sample moments with respect to $(\theta^{(1)}, \theta^{(2)})$, then

$$\bar{D}_M = \begin{pmatrix} -\hat{\pi}/\pi & 0 \\ 0 & -(\hat{\pi}/\pi - 1) \\ 0 & -1 \end{pmatrix} \xrightarrow{p} D := \begin{pmatrix} -1 & 0 \\ 0 & 0 \\ 0 & -1 \end{pmatrix}.$$

The optimally weighted GMM first-order condition is

$$\bar{D}_M^{\top}\{L^{-1}F(L^{-1})^{\top}\}^{-1}\bar{m}_L(\theta^{(1)}, \theta^{(2)}) = 0,$$

where $\bar{m}_L$ is the sample average of the transformed moment vector. The first component of this condition gives the exact identity

$$\hat{\theta}^{(1)} = \frac{1}{N}\sum_{i:S_i=1}(R_i - \lambda\tilde{R}_i).$$

The second component gives

$$\hat{\theta}^{(2)} - \hat{J} = \frac{\hat{\pi}/\pi - 1}{\pi^{-1} - 1}\frac{1}{M}\sum_{i=1}^{M}\left(\frac{S_i}{\pi} - 1\right)(\tilde{R}_i - \hat{\theta}^{(2)}).$$

Subtracting and adding $\hat{J}$ inside the final factor and solving for $\hat{\theta}^{(2)} - \hat{J}$ yields

$$\hat{\theta}^{(2)} - \hat{J} = \frac{\hat{\pi}/\pi - 1}{(\pi^{-1} - 1) + (\hat{\pi}/\pi - 1)^2}\frac{1}{M}\sum_{i=1}^{M}\left(\frac{S_i}{\pi} - 1\right)(\tilde{R}_i - \hat{J}).$$

Thus the finite-sample GMM estimator need not equal $\hat{J}$. To bound their difference, note that $\hat{\pi}/\pi - 1 = O_p(M^{-1/2})$ and

$$\begin{aligned}&\frac{1}{M}\sum_{i=1}^{M}\left(\frac{S_i}{\pi} - 1\right)(\tilde{R}_i - \hat{J}) \\ &\quad= \frac{1}{M}\sum_{i=1}^{M}\left(\frac{S_i}{\pi} - 1\right)(\tilde{R}_i - \theta^{(2)}) - \left(\frac{\hat{\pi}}{\pi} - 1\right)(\hat{J} - \theta^{(2)}) = O_p(1).\end{aligned}$$

Indeed, independence of assignment and outcomes implies that the first term on the right has mean zero and variance $n(\pi^{-1} - 1)\tilde{s}^2/M$, while $\mathrm{Var}(\hat{J}) = n\tilde{s}^2/M$. Both variances are bounded because $M = b_t n$. The denominator in the exact difference formula converges in probability to $\pi^{-1} - 1 > 0$, so

$$\hat{\theta}^{(2)} - \hat{J} = O_p(M^{-1/2}) = o_p(1).$$

Returning to the notation of batch t, these relations imply

$$\hat{\theta}^{(2)}_{t,a} = \frac{1}{b_t\sqrt{n}}\sum_{j=1}^{b_t n}\tilde{Y}_{t,j,a} + o_p(1) = \hat{J}_{t,a} + o_p(1),$$

and

$$\hat{\pi}_{t,a}\hat{\theta}^{(1)}_{t,a} = \frac{1}{b_t n}\sum_{j=1}^{b_t n}\xi_{t,j,a}\sqrt{n}\left(Y_{t,j,a} - \lambda_a\tilde{Y}_{t,j,a}\right) = \frac{1}{b_t\sqrt{n}}\sum_{j=1}^{b_t n}\xi_{t,j,a}\left(Y_{t,j,a} - \lambda_a\tilde{Y}_{t,j,a}\right) = \hat{G}_{t,a}.$$

Consequently, on $\hat{\pi}_{t,a} > 0$,

$$\left(\frac{\hat{G}_{t,a}}{\hat{\pi}_{t,a}}, \hat{J}_{t,a}\right) - \left(\hat{\theta}^{(1)}_{t,a}, \hat{\theta}^{(2)}_{t,a}\right) \xrightarrow{p} (0,0).$$

Thus the normalized orthogonal aggregates are asymptotically equivalent to the optimally weighted GMM estimators and inherit their asymptotic efficiency. If $\pi = 1$, the labeled and all-unit proxy moments are identical. Removing the duplicate moment leaves two exactly identified equations, which give $\hat{\theta}^{(2)} = \hat{J}$ and the same residual-mean estimator of $\theta^{(1)}$, with $\hat{\pi} = 1$. The same conclusion therefore holds in this case, with exact equality. This proves the proposition.

B.2 Regularity Conditions for Theorem 1

Theorem 1 concerns the aggregate-based policy class specified below. It establishes the Gaussian limit used as the asymptotic framework for Algorithms 1–2. The finite-batch implementations, whose posterior updates also use realized allocation proportions, are specified separately in Appendix F.

Assumption 1 (Regularity for the finite-batch experiment)**.** *Conditional on fixed local parameters, for every batch t and unit j, let $\xi_{t,j} = (\xi_{t,j,1}, \ldots, \xi_{t,j,K})$, $\epsilon_{t,j} = (\epsilon_{t,j,1}, \ldots, \epsilon_{t,j,K})$, and $\tilde{\epsilon}_{t,j} = (\tilde{\epsilon}_{t,j,1}, \ldots, \tilde{\epsilon}_{t,j,K})$. The noise-vector pairs $(\epsilon_{t,j}, \tilde{\epsilon}_{t,j})$ are identically distributed and independent across units and batches. The following conditions hold.*

1. (Ignorability) *Given the aggregate history available before physical assignment, the assignment vectors are conditionally independent categorical draws with probability vector π_t, and the collection of assignment vectors is conditionally independent of all true and proxy outcomes in batch t:*

$$\{\xi_{t,j} : 1 \le j \le b_t n\} \perp \{Y_{t,j,a}, \tilde{Y}_{t,j,a} : 1 \le j \le b_t n,\ a \in [K]\} \mid \hat{J}_0, \hat{G}_0, \ldots, \hat{J}_{t-1}, \hat{G}_{t-1}, \hat{J}_t.$$

2. (Moment restriction) *There exists a constant $C > 0$ such that $\mathbb{E}\|\epsilon_{t,j}\|_2^4 \le C$ and $\mathbb{E}\|\tilde{\epsilon}_{t,j}\|_2^4 \le C$. With $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_K)$, the joint covariance structure satisfies $\mathrm{Cov}(\tilde{\epsilon}_{t,j}) = \mathrm{diag}\{\tilde{s}_a^2\}_{a=1}^K$ and $\mathrm{Cov}(\tilde{\epsilon}_{t,j}, \epsilon_{t,j} - \Lambda\tilde{\epsilon}_{t,j}) = 0$.*

3. (Sufficiency) *The policy depends on the experimental data only through the aggregate estimators:*

$$\pi_t = \pi_t(\hat{J}_0, \hat{G}_0, \ldots, \hat{J}_{t-1}, \hat{G}_{t-1}, \hat{J}_t).$$

4. (Continuity) *The set of discontinuity points of $\pi_t(\cdot)$ has measure zero under the distribution of $(J_0, G_0, \ldots, J_{t-1}, G_{t-1}, J_t)$.*

Ignorability specifies the randomization mechanism: given the aggregate history used by the policy, physical arm assignment follows the chosen allocation probabilities and is independent of the realized potential outcomes. The joint orthogonality condition extends the arm-wise decomposition in Section 4.1 to the full vectors: it rules out residual cross-covariance between the all-arm proxy signal and the all-arm residual signal. Sufficiency states that the adaptive allocation rule uses the aggregate estimators rather than the unit-level outcomes. The moment and continuity restrictions are standard regularity conditions needed to pass from batch-level central limit approximations to a sequential weak-convergence result.

B.3 Proof of Theorem 1

For probability laws μ and ν on $\mathbb{R}^m$, define the bounded-Lipschitz distance by

$$d_{\mathrm{BL}}(\mu, \nu) := \sup_{\varphi \in \mathcal{F}_{\mathrm{BL}}} \left| \int \varphi \, d\mu - \int \varphi \, d\nu \right|, \qquad \mathcal{F}_{\mathrm{BL}} := \{\varphi : \mathbb{R}^m \to \mathbb{R} : \|\varphi\|_\infty \le 1,\ \mathrm{Lip}(\varphi) \le 1\},$$

where

$$\mathrm{Lip}(\varphi) := \sup_{x \ne y} \frac{|\varphi(x) - \varphi(y)|}{\|x - y\|_2}.$$

For random vectors U and V, we write $d_{\mathrm{BL}}(U, V)$ as shorthand for $d_{\mathrm{BL}}(\mathcal{L}(U), \mathcal{L}(V))$. This distance metrizes weak convergence on every finite-dimensional Euclidean space.

The proof follows the induction structure of Che & Namkoong (2023, Appendix B.1), but the one-batch argument must be modified. In our experiment, the current allocation is selected after observing the current all-unit proxy aggregate estimator $\hat{J}_t$. Consequently, conditioning on $\hat{J}_t$ need not leave the current residuals independent, even though the residuals and proxy outcomes are orthogonal. We therefore prove a joint functional central limit theorem for processes representing $\hat{J}_t$ and $\hat{G}_t$, indexed by all possible allocation vectors.

Let $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_K)$, and, within a fixed batch, write

$$V_j := \tilde{\epsilon}_{t,j}, \qquad R_j := \epsilon_{t,j} - \Lambda \tilde{\epsilon}_{t,j}.$$

Assumption 1 and the definition of λ_a imply

$$\mathrm{Cov}(V_j) = \mathrm{diag}\{\tilde{s}_a^2\}_{a=1}^K, \qquad \mathrm{Cov}(V_j, R_j) = 0, \qquad \mathrm{Var}(R_{j,a}) = s_{a,\mathrm{res}}^2.$$

One-batch joint limit. Fix batch t, and consider deterministic pre-batch histories $r_{0:t-1}^{(n)} \to r_{0:t-1}$; for $t = 0$, the history is absent. We show that the finite-batch transition from $r_{0:t-1}^{(n)}$ converges to the Gaussian sequential-experiment transition from $r_{0:t-1}$, provided that π_t is continuous at $(r_{0:t-1}, J_t)$ with probability one.

Set $m = b_t n$. By the conditional categorical randomization in Assumption 1, the assignments can be represented using i.i.d. $U_j \sim \mathrm{Unif}(0,1)$, independent of the current-batch noise vectors. For $p \in \Delta_K$, define

$$c_0(p) := 0, \qquad c_a(p) := \sum_{\ell=1}^{a} p_\ell, \qquad \xi_{j,a}(p) := \mathbf{1}\{c_{a-1}(p) < U_j \le c_a(p)\}.$$

Also define

$$\mathbb{V}_m := \frac{1}{\sqrt{m}} \sum_{j=1}^{m} V_j, \qquad \mathbb{W}_m(q) := \frac{1}{\sqrt{m}} \sum_{j=1}^{m} R_j \mathbf{1}\{U_j \le q\}, \quad 0 \le q \le 1,$$

and $F_m(q) := m^{-1} \sum_{j=1}^{m} \mathbf{1}\{U_j \le q\}$.

To justify the required joint functional limit, consider the class consisting of the coordinate functions $(v, r, u) \mapsto v_a$ and the threshold-indexed functions $(v, r, u) \mapsto r_a \mathbf{1}\{u \le q\}$, where $a \in [K]$ and $q \in [0,1]$. This is a VC-type class with an L^2 envelope under Assumption 1; hence it is Donsker (van der Vaart & Wellner, 1996, Sections 2.5–2.6). Its empirical-process limit, together with the uniform Glivenko–Cantelli theorem for F_m, gives

$$\big(\mathbb{V}_m, \mathbb{W}_m(\cdot), F_m(\cdot)\big) \xrightarrow{d} \big(\mathbb{V}, \mathbb{W}(\cdot), q \mapsto q\big)$$

in $\mathbb{R}^K \times \ell^\infty([0,1])^K \times \ell^\infty([0,1])$. The limiting variables satisfy

$$\mathbb{V} \sim \mathcal{N}\left(0, \mathrm{diag}\{\tilde{s}_a^2\}_{a=1}^K\right),$$

and $\mathbb{W}$ is a continuous mean-zero Gaussian process with

$$\mathrm{Cov}(\mathbb{W}(q), \mathbb{W}(q')) = (q \wedge q')\mathrm{Cov}(R_j), \qquad \mathrm{Cov}(\mathbb{V}, \mathbb{W}(q)) = q\mathrm{Cov}(V_j, R_j) = 0.$$

Because the joint limit is Gaussian, $\mathbb{V}$ and the entire process $\mathbb{W}$ are independent. The Glivenko–Cantelli part of the same display also gives

$$\sup_{0 \le q \le 1} |F_m(q) - q| \xrightarrow{p} 0.$$

For the actual allocation

$$p_m := \pi_t(r_{0:t-1}^{(n)}, \hat{J}_t),$$

where the pre-batch-history argument is omitted when $t = 0$, the definitions of the aggregate estimators give the exact representations

$$\hat{J}_t = \boldsymbol{\theta}^{(2)} + \frac{1}{\sqrt{b_t}} \mathbb{V}_m$$

and

$$\hat{G}_{t,a} = \frac{N_{t,a}}{m}\theta_a^{(1)} + \frac{1}{\sqrt{b_t}}\left\{\mathbb{W}_{m,a}(c_a(p_m)) - \mathbb{W}_{m,a}(c_{a-1}(p_m))\right\},$$

where

$$\frac{N_{t,a}}{m} = F_m(c_a(p_m)) - F_m(c_{a-1}(p_m)).$$

Uniformly over $p_m \in \Delta_K$,

$$\left|\frac{N_{t,a}}{m} - p_{m,a}\right| \le 2 \sup_{0\le q\le 1} |F_m(q) - q| \xrightarrow{p} 0.$$

It follows from the functional central limit theorem, the assumed continuity of π_t, and the extended continuous mapping theorem that

$$\left(\hat{J}_t, p_m, \mathbb{W}_m(\cdot)\right) \xrightarrow{d} \left(J_t, \pi_t(r_{0:t-1}, J_t), \mathbb{W}(\cdot)\right),$$

where

$$J_t = \boldsymbol{\theta}^{(2)} + \frac{1}{\sqrt{b_t}}\mathbb{V}.$$

Because $\mathbb{W}$ has continuous sample paths, evaluation at the allocation-dependent interval endpoints is continuous. Hence

$$(\hat{J}_t, \hat{G}_t) \xrightarrow{d} (J_t, G_t)$$

under the corresponding deterministic histories, where, writing $p = \pi_t(r_{0:t-1}, J_t)$,

$$G_{t,a} = p_a\theta_a^{(1)} + \frac{1}{\sqrt{b_t}}\left\{\mathbb{W}_a(c_a(p)) - \mathbb{W}_a(c_{a-1}(p))\right\}.$$

It remains to verify that this is exactly the transition in Definition 1. Conditional on J_t, the vector p is fixed and $\mathbb{W}$ remains independent of J_t. For any $a, b \in [K]$, the covariance of the two Gaussian increments in the preceding display is

$$\begin{aligned}
&\mathrm{Cov}\Big(\mathbb{W}_a(c_a(p)) - \mathbb{W}_a(c_{a-1}(p)), \mathbb{W}_b(c_b(p)) - \mathbb{W}_b(c_{b-1}(p))\Big) \\
&\quad = \mathrm{Cov}(R_{j,a}, R_{j,b})\,\big|(c_{a-1}(p), c_a(p)] \cap (c_{b-1}(p), c_b(p)]\big| \\
&\quad = \mathbf{1}\{a = b\}\, p_a s_{a,\mathrm{res}}^2.
\end{aligned}$$

Therefore

$$G_t \mid r_{0:t-1}, J_t \sim \mathcal{N}\left(p \odot \boldsymbol{\theta}^{(1)}, \mathrm{diag}\left\{\frac{p_a s_{a,\mathrm{res}}^2}{b_t}\right\}_{a=1}^{K}\right).$$

This calculation also covers $p_a = 0$: the corresponding interval is empty and the Gaussian increment is identically zero. This proves the one-batch joint limit without conditioning the residual sample on the realized value of $\hat{J}_t$.

Induction over batches. For every fixed $t = 0, \ldots, T-1$, write

$$H_t^{(n)} := (\hat{J}_0, \hat{G}_0, \ldots, \hat{J}_t, \hat{G}_t), \qquad H_t := (J_0, G_0, \ldots, J_t, G_t).$$

We prove $H_t^{(n)} \xrightarrow{d} H_t$ by induction. For $t = 0$, the one-batch joint limit applies with no previous history. Assumption 1 ensures that π_0 is continuous at J_0 with probability one, so the base case follows.

Suppose that $H_t^{(n)} \xrightarrow{d} H_t$. To prove convergence through batch $t+1$, fix a bounded Lipschitz function f satisfying

$$\sup_{x\ne y} \frac{|f(x) - f(y)|}{\|x - y\|_2} \le 1, \qquad \sup_{x,y} |f(x) - f(y)| \le 1.$$

It is enough to consider this normalization because multiplying f by a positive constant and subtracting a constant do not affect whether the difference of expectations converges to zero.

For a deterministic history $r \in \mathbb{R}^{2K(t+1)}$, let $P_{n,t+1}(r,\cdot)$ denote the finite-batch adaptive-experiment transition law of $(\hat{J}_{t+1}, \hat{G}_{t+1})$ when the previous history supplied to the policy is fixed at r, and let $P_{t+1}(r,\cdot)$ denote the corresponding Gaussian transition law. Define

$$g_n(r) := \int f(r,z)\, P_{n,t+1}(r, dz), \qquad g(r) := \int f(r,z)\, P_{t+1}(r, dz).$$

This transition-kernel definition avoids conditioning on a probability-zero event of the form $H_t^{(n)} = r$.

Let D_{t+1} be the set of discontinuity points of π_{t+1}, and let Q_{t+1}^J be the Gaussian law of J_{t+1}. Define

$$\mathcal{C}_t := \left\{ r : \; Q_{t+1}^J(\{j : (r,j) \in D_{t+1}\}) = 0 \right\}.$$

Assumption 1 and Fubini's theorem imply

$$\Pr(H_t \in \mathcal{C}_t) = 1.$$

The one-batch joint limit established above gives the following *continuous convergence* of the transition kernels: for every deterministic sequence $r^{(n)} \to r$ with $r \in \mathcal{C}_t$,

$$P_{n,t+1}(r^{(n)}, \cdot) \Rightarrow P_{t+1}(r, \cdot).$$

Indeed, this is precisely the preceding one-batch argument with $r_{0:t}^{(n)} = r^{(n)}$ and $r_{0:t} = r$. Moreover,

$$\begin{aligned} |g_n(r^{(n)}) - g(r)| \le & \int |f(r^{(n)}, z) - f(r,z)|\, P_{n,t+1}(r^{(n)}, dz) \\ & + \left| \int f(r,z)\, P_{n,t+1}(r^{(n)}, dz) - \int f(r,z)\, P_{t+1}(r, dz) \right|. \end{aligned}$$

The first term is at most $\|r^{(n)} - r\|_2$ by the Lipschitz property of f, and the second converges to zero because $z \mapsto f(r,z)$ is bounded and continuous and the transition laws converge weakly. Hence

$$r^{(n)} \to r \in \mathcal{C}_t \quad \Longrightarrow \quad g_n(r^{(n)}) \to g(r). \tag{17}$$

Appendix B.1 of Che & Namkoong (2023) obtains the induction step by proving uniform convergence of its one-step conditional expectations. Here, the implication in equation 17, which follows from the joint functional limit, is sufficient. Because $\mathbb{R}^{2K(t+1)}$ is a Polish space, the Skorohod representation theorem applied to the induction hypothesis guarantees the existence of random vectors $H_t^{(n),*}$ and H_t^* on a common probability space such that

$$H_t^{(n),*} \overset{d}{=} H_t^{(n)}, \qquad H_t^* \overset{d}{=} H_t, \qquad H_t^{(n),*} \longrightarrow H_t^* \quad \text{almost surely.}$$

The equalities in distribution are essential: the coupled variables need not be the original histories, but every expectation below is unchanged by replacing the original histories with these copies.

Let Ω_0 be the event on which $H_t^{(n),*} \to H_t^*$ and $H_t^* \in \mathcal{C}_t$. The preceding displays give $\Pr(\Omega_0) = 1$. For each $\omega \in \Omega_0$, apply equation 17 to the deterministic sequence $r^{(n)} = H_t^{(n),*}(\omega)$ and limit $r = H_t^*(\omega)$. This establishes that

$$g_n(H_t^{(n),*}) \longrightarrow g(H_t^*) \quad \text{almost surely.}$$

Because f is bounded, g_n and g are bounded uniformly in n. Dominated convergence and the equalities in distribution therefore give

$$\mathbb{E}[g_n(H_t^{(n)})] = \mathbb{E}[g_n(H_t^{(n),*})] \longrightarrow \mathbb{E}[g(H_t^*)] = \mathbb{E}[g(H_t)].$$

By the tower property, the first and last expectations are respectively

$$\mathbb{E}[f(H_t^{(n)}, \hat{J}_{t+1}, \hat{G}_{t+1})] \quad \text{and} \quad \mathbb{E}[f(H_t, J_{t+1}, G_{t+1})].$$

Thus the expectations converge for every bounded Lipschitz f, proving $H_{t+1}^{(n)} \xrightarrow{d} H_{t+1}$. This completes the induction step. Since T is fixed, induction gives

$$(\hat{J}_0, \hat{G}_0, \ldots, \hat{J}_{T-1}, \hat{G}_{T-1}) \xrightarrow{d} (J_0, G_0, \ldots, J_{T-1}, G_{T-1}),$$

which is the claimed Gaussian sequential experiment limit.

B.4 Proof of Lemma 1

Consider a scalar Gaussian prior and likelihood:

$$\theta \sim \mathcal{N}(\mu, \sigma^2), \qquad Y \mid \theta \sim \mathcal{N}(\theta, v).$$

The posterior precision and mean are

$$\sigma_+^{-2} = \sigma^{-2} + v^{-1}, \qquad \mu_+ = \sigma_+^2(\sigma^{-2}\mu + v^{-1}Y).$$

Under the prior predictive distribution,

$$Y = \theta + \eta, \qquad \theta = \mu + \sigma Z_1, \qquad \eta = \sqrt{v} Z_2,$$

with independent standard normals Z_1, Z_2. Hence

$$\mu_+ - \mu = \sigma_+^2 v^{-1}(Y - \mu) = \sigma_+^2 v^{-1}(\sigma Z_1 + \sqrt{v} Z_2).$$

Its mean is zero and its variance is

$$(\sigma_+^2 v^{-1})^2(\sigma^2 + v) = \left(\frac{\sigma^2}{\sigma^2 + v}\right)^2 (\sigma^2 + v) = \frac{\sigma^4}{\sigma^2 + v}.$$

Therefore

$$\mu_+ = \mu + \sqrt{\frac{\sigma^4}{\sigma^2 + v}}\, Z$$

for a standard normal Z. Substituting $v = \tilde{s}_a^2 / b_t$ for the proxy signal gives the displayed recursion from $\mu_{t,a}^{(2)}$ to $\mu_{t+1,a}^{(2)}$. Let $\mathcal{F}_t^-$ be the sigma-field generated by all observations before J_t, and let $\mathcal{F}_t^+ := \sigma(\mathcal{F}_t^-, J_t)$; the allocation π_t is $\mathcal{F}_t^+$-measurable. For the residual component, when $\pi_{t,a} > 0$, the conditional likelihood $G_{t,a} \mid \theta_a^{(1)}, \mathcal{F}_t^+ \sim \mathcal{N}(\pi_{t,a}\theta_a^{(1)}, \pi_{t,a} s_{a,\mathrm{res}}^2 / b_t)$ is equivalent to $G_{t,a}/\pi_{t,a} \mid \theta_a^{(1)}, \mathcal{F}_t^+ \sim \mathcal{N}(\theta_a^{(1)}, s_{a,\mathrm{res}}^2/(\pi_{t,a} b_t))$. Substituting this variance gives the second formula. When $\pi_{t,a} = 0$, $G_{t,a} = 0$ almost surely and its likelihood does not depend on $\theta_a^{(1)}$. The posterior mean remains $\mu_{t,a}^{(1)}$, and the random-walk coefficient is zero. Thus the second formula also holds at the boundary $\pi_{t,a} = 0$.

It remains to justify the joint independence claimed in the lemma when the allocation is adaptive. Conditional on $\mathcal{F}_t^-$, the vector of standardized proxy posterior-mean increments is multivariate standard normal with identity covariance, and its conditional law does not depend on $\mathcal{F}_t^-$. It is therefore independent of $\mathcal{F}_t^-$, with independent coordinates across arms. Likewise, conditional on $\mathcal{F}_t^+$, the standardized residual posterior-mean increments on coordinates with $\pi_{t,a} > 0$ are independent standard normals. Complete this vector on coordinates with $\pi_{t,a} = 0$ using independent standard normals; their coefficients are zero, so they do not change the posterior-mean process. The resulting K-dimensional vector has conditional law $\mathcal{N}(0, I_K)$, which does not depend on $\mathcal{F}_t^+$. It is therefore independent of the observed history, including the current proxy innovations. Applying this argument successively over $t = 0, \ldots, T-1$ proves mutual independence across arms, batches, and signal types.

Iterating the recursions in Lemma 1 gives the full expansion of Equation 12:

$$V_t^{\pi,+}(x_t^+) = \mathbb{E}^\pi\left[\max_{a\in[K]} \left\{ \begin{array}{l} \mu_{t,a}^{(1)} + \displaystyle\sum_{v=t}^{T-1} \sqrt{\frac{(\sigma_{v,a}^{(1)})^4 \pi_{v,a} b_v}{s_{a,\mathrm{res}}^2 + (\sigma_{v,a}^{(1)})^2 \pi_{v,a} b_v}}\, Z_{v,a}^G \\ \qquad + \lambda_a \mu_{t+1,a}^{(2)} + \lambda_a \displaystyle\sum_{v=t+1}^{T-1} \sqrt{\frac{(\sigma_{v,a}^{(2)})^4 b_v}{\tilde{s}_a^2 + (\sigma_{v,a}^{(2)})^2 b_v}}\, Z_{v,a}^J \end{array} \right\} \,\middle|\, x_t^+ \right].$$

B.5 Proof of Lemma 2

Fix $t \in \{0, \ldots, T-1\}$ and a post-LLM state x_t^+. Under RHO, the future allocation sequence $(\pi_t, \ldots, \pi_{T-1})$ is selected using only x_t^+. Hence, conditional on x_t^+, it is a deterministic open-loop sequence. For each arm a, define its budget-weighted average allocation by

$$\bar{\rho}_a = \frac{1}{B_t} \sum_{v=t}^{T-1} b_v \pi_{v,a}.$$

Because every π_v belongs to Δ_K, $\bar{\rho}_a \geq 0$ and

$$\sum_{a=1}^{K} \bar{\rho}_a = \frac{1}{B_t} \sum_{v=t}^{T-1} b_v \sum_{a=1}^{K} \pi_{v,a} = 1.$$

Thus $\bar{\rho} \in \Delta_K$. For arm a, the residual posterior precision after the full future allocation sequence is

$$(\sigma_{T,a}^{(1)})^{-2} = (\sigma_{t,a}^{(1)})^{-2} + \sum_{v=t}^{T-1} \frac{b_v \pi_{v,a}}{s_{a,\mathrm{res}}^2}.$$

By the definition of $\bar{\rho}$, this is

$$(\sigma_{T,a}^{(1)})^{-2} = (\sigma_{t,a}^{(1)})^{-2} + \frac{B_t \bar{\rho}_a}{s_{a,\mathrm{res}}^2}.$$

By Lemma 1, the ex-ante variance of the terminal residual posterior mean equals the current residual posterior variance minus the terminal residual posterior variance:

$$(\sigma_{t,a}^{(1)})^2 - \left[(\sigma_{t,a}^{(1)})^{-2} + \frac{B_t \bar{\rho}_a}{s_{a,\mathrm{res}}^2}\right]^{-1} = \frac{(\sigma_{t,a}^{(1)})^4 \bar{\rho}_a B_t}{s_{a,\mathrm{res}}^2 + (\sigma_{t,a}^{(1)})^2 \bar{\rho}_a B_t}.$$

Therefore, conditional on x_t^+, the terminal residual posterior mean has the representation

$$\mu_{T,a}^{(1)} = \mu_{t,a}^{(1)} + \sqrt{\frac{(\sigma_{t,a}^{(1)})^4 \bar{\rho}_a B_t}{s_{a,\mathrm{res}}^2 + (\sigma_{t,a}^{(1)})^2 \bar{\rho}_a B_t}} Z_a^G.$$

The current proxy signal J_t has already been absorbed into x_t^+, so only batches $t+1, \ldots, T-1$ contribute future proxy uncertainty. These future proxy signals do not depend on the allocation sequence. The terminal proxy posterior precision is therefore

$$(\sigma_{T,a}^{(2)})^{-2} = (\sigma_{t+1,a}^{(2)})^{-2} + \frac{B_{t+1}}{\tilde{s}_a^2},$$

and the same normal posterior calculation gives

$$\mu_{T,a}^{(2)} = \mu_{t+1,a}^{(2)} + \sqrt{\frac{(\sigma_{t+1,a}^{(2)})^4 B_{t+1}}{\tilde{s}_a^2 + (\sigma_{t+1,a}^{(2)})^2 B_{t+1}}} Z_a^J.$$

The random variables Z_a^G and Z_a^J are mutually independent standard normals across arms and signal types by the diagonal Gaussian transition covariances and the independence of the residual and proxy posterior components. Combining the two displays gives the terminal posterior mean of the target arm value,

$$\begin{aligned}\mu_{T,a}^{(1)} + \lambda_a \mu_{T,a}^{(2)} = \mu_{t,a}^{(1)} + \lambda_a \mu_{t+1,a}^{(2)} \\ + \sqrt{\frac{(\sigma_{t,a}^{(1)})^4 \bar{\rho}_a B_t}{s_{a,\mathrm{res}}^2 + (\sigma_{t,a}^{(1)})^2 \bar{\rho}_a B_t}} Z_a^G \\ + \lambda_a \sqrt{\frac{(\sigma_{t+1,a}^{(2)})^4 B_{t+1}}{\tilde{s}_a^2 + (\sigma_{t+1,a}^{(2)})^2 B_{t+1}}} Z_a^J.\end{aligned}$$

Thus every open-loop sequence selected from x_t^+ induces a terminal predictive distribution that depends on that sequence only through $\bar{\rho}$. Conversely, for every $\bar{\rho} \in \Delta_K$, the constant sequence $\pi_v = \bar{\rho}$ for $v = t, \ldots, T-1$ realizes the same budget-weighted average allocation. Hence the set of terminal predictive distributions attainable by jointly optimizing the $T-t$ future allocation policies is exactly the set indexed by $\bar{\rho} \in \Delta_K$. Taking the expectation of the terminal posterior maximum over this set yields the single-simplex optimization problem displayed in Lemma 2.

B.6 Proof of Theorem 2

We use the finite-batch updates in Algorithms 1 and 2. Their ordering differs, but both incorporate $\hat{J}_v$ and $\hat{G}_v$ by the end of batch v. Throughout this proof, the moments and λ_a are fixed and known. By the definition of the realized allocation proportion,

$$n \sum_{v=0}^{t-1} \hat{\pi}_{v,a} b_v = \sum_{v=0}^{t-1} \sum_{j=1}^{b_v n} \xi_{v,j,a}.$$

Thus the three means in the theorem are

$$\bar{Y}^{\mathrm{obs}}_{0:t-1,a} := \frac{\sum_{v=0}^{t-1} \sum_{j=1}^{b_v n} \xi_{v,j,a} Y_{v,j,a}}{n \sum_{v=0}^{t-1} \hat{\pi}_{v,a} b_v},$$
$$\bar{\tilde{Y}}^{\mathrm{obs}}_{0:t-1,a} := \frac{\sum_{v=0}^{t-1} \sum_{j=1}^{b_v n} \xi_{v,j,a} \tilde{Y}_{v,j,a}}{n \sum_{v=0}^{t-1} \hat{\pi}_{v,a} b_v},$$
$$\bar{\tilde{Y}}^{\mathrm{all}}_{0:t-1,a} := \frac{\sum_{v=0}^{t-1} \sum_{j=1}^{b_v n} \tilde{Y}_{v,j,a}}{n \sum_{v=0}^{t-1} b_v}.$$

The condition in the theorem requires at least one observed outcome for arm a, ensuring that the residual working posterior is proper.

Accumulating the finite-batch updates. A flat prior has zero initial precision and zero initial precision-weighted mean. Summing the residual precision and precision-weighted-mean increments in the algorithms gives

$$(\sigma^{(1)}_{n,t,a})^{-2} = \frac{\sum_{v=0}^{t-1} b_v \hat{\pi}_{v,a}}{s^2_{a,\mathrm{res}}},$$
$$(\sigma^{(1)}_{n,t,a})^{-2} \mu^{(1)}_{n,t,a} = \frac{\sum_{v=0}^{t-1} b_v \hat{G}_{v,a}}{s^2_{a,\mathrm{res}}}.$$

Consequently, the definition of $\hat{G}_{v,a}$ yields

$$\begin{aligned}
\mu^{(1)}_{n,t,a} &= \frac{\sum_{v=0}^{t-1} b_v \hat{G}_{v,a}}{\sum_{v=0}^{t-1} \hat{\pi}_{v,a} b_v} \\
&= \frac{\sum_{v=0}^{t-1} \sum_{j=1}^{b_v n} \xi_{v,j,a} (Y_{v,j,a} - \lambda_a \tilde{Y}_{v,j,a})}{\sqrt{n} \sum_{v=0}^{t-1} \hat{\pi}_{v,a} b_v} \\
&= \sqrt{n} \left(\bar{Y}^{\mathrm{obs}}_{0:t-1,a} - \lambda_a \bar{\tilde{Y}}^{\mathrm{obs}}_{0:t-1,a} \right).
\end{aligned}$$

Similarly, summing the proxy update increments gives

$$(\sigma^{(2)}_{n,t,a})^{-2} = \frac{\sum_{v=0}^{t-1} b_v}{\tilde{s}^2_a},$$
$$(\sigma^{(2)}_{n,t,a})^{-2} \mu^{(2)}_{n,t,a} = \frac{\sum_{v=0}^{t-1} b_v \hat{J}_{v,a}}{\tilde{s}^2_a}.$$

Using the definition of $\hat{J}_{v,a}$, we obtain

$$\mu^{(2)}_{n,t,a} = \frac{\sum_{v=0}^{t-1} b_v \hat{J}_{v,a}}{\sum_{v=0}^{t-1} b_v} = \sqrt{n}\, \bar{\tilde{Y}}^{\mathrm{all}}_{0:t-1,a}.$$

Recombining the two posterior means gives

$$\mu_{n,t,a} = \mu^{(1)}_{n,t,a} + \lambda_a \mu^{(2)}_{n,t,a} = \sqrt{n} \left(\bar{Y}^{\mathrm{obs}}_{0:t-1,a} - \lambda_a \bar{\tilde{Y}}^{\mathrm{obs}}_{0:t-1,a} \right) + \lambda_a \sqrt{n}\, \bar{\tilde{Y}}^{\mathrm{all}}_{0:t-1,a}.$$

This is an exact identity for the finite-batch working-posterior updates, using realized allocation proportions throughout; no replacement of $\hat{\pi}_{v,a}$ by $\pi_{v,a}$ is made.

Relation to the usual PPI++ formula. For this comparison, write $m := n\sum_{v=0}^{t-1}\hat{\pi}_{v,a}b_v > 0$ and $u := n\sum_{v=0}^{t-1}(1-\hat{\pi}_{v,a})b_v$ for the numbers of units with and without an observed true outcome for arm a, respectively. If $u > 0$, define the unlabeled-only proxy mean by

$$\bar{\tilde{Y}}^{\text{unobs}}_{0:t-1,a} := \frac{1}{u}\sum_{v=0}^{t-1}\sum_{j=1}^{b_v n}(1-\xi_{v,j,a})\tilde{Y}_{v,j,a}.$$

The pooled proxy mean decomposes exactly as

$$\bar{\tilde{Y}}^{\text{all}}_{0:t-1,a} = \frac{m}{m+u}\bar{\tilde{Y}}^{\text{obs}}_{0:t-1,a} + \frac{u}{m+u}\bar{\tilde{Y}}^{\text{unobs}}_{0:t-1,a}.$$

Substitution into the posterior-mean identity gives

$$\frac{\mu_{n,t,a}}{\sqrt{n}} = \bar{Y}^{\text{obs}}_{0:t-1,a} + \frac{u}{m+u}\lambda_a\left(\bar{\tilde{Y}}^{\text{unobs}}_{0:t-1,a} - \bar{\tilde{Y}}^{\text{obs}}_{0:t-1,a}\right).$$

This is the usual disjoint-sample PPI++ mean-estimation formula (Angelopoulos et al., 2026), with tuning coefficient $u\lambda_a/(m+u)$, rather than λ_a. If $u = 0$, the all-unit and observed proxy means coincide, so the identity reduces to $\mu_{n,t,a}/\sqrt{n} = \bar{Y}^{\text{obs}}_{0:t-1,a}$ without requiring an unlabeled-only mean.

To relate this coefficient to oracle tuning, consider separately a static i.i.d. design with fixed sample sizes $m, u > 0$, independent labeled and unlabeled samples, and known moments. For any fixed $c \in \mathbb{R}$,

$$\begin{aligned}&\operatorname{Var}\left[\bar{Y}^{\text{obs}}_{0:t-1,a} + c\left(\bar{\tilde{Y}}^{\text{unobs}}_{0:t-1,a} - \bar{\tilde{Y}}^{\text{obs}}_{0:t-1,a}\right)\right]\\ &\quad= \frac{s_a^2}{m} - \frac{2c\lambda_a\tilde{s}_a^2}{m} + c^2\tilde{s}_a^2\left(\frac{1}{m}+\frac{1}{u}\right).\end{aligned}$$

This strictly convex quadratic is minimized at $c = u\lambda_a/(m+u)$. Thus the coefficient is the unconstrained oracle PPI++ coefficient in this static comparison. Under adaptive allocation, the pooled-to-unlabeled identity remains exact for every realized dataset, but this calculation does not establish oracle optimality, unbiasedness, efficiency, or confidence-interval validity for adaptively collected data. Nor does fixed known-moment tuning imply equality with a PPI++ implementation that estimates its tuning coefficient from data.

Working-posterior variance comparison. The same update calculation verifies the variance reduction invoked after the theorem. Define $\sigma^2_{n,t,a} := (\sigma^{(1)}_{n,t,a})^2 + \lambda_a^2(\sigma^{(2)}_{n,t,a})^2$, using the factorization of the working posterior. Then

$$\sigma^2_{n,t,a} = \frac{s^2_{a,\text{res}}}{\sum_{v=0}^{t-1}\hat{\pi}_{v,a}b_v} + \lambda_a^2\frac{\tilde{s}_a^2}{\sum_{v=0}^{t-1}b_v}.$$

With a flat prior, known s_a^2, and the same realized allocation, the true-outcome-only Gaussian working-posterior variance is

$$\sigma^2_{n,t,a,\text{obs}} = \frac{s_a^2}{\sum_{v=0}^{t-1}\hat{\pi}_{v,a}b_v}.$$

Using $s^2_{a,\text{res}} = s_a^2 - \lambda_a^2\tilde{s}_a^2$ and $\lambda_a^2\tilde{s}_a^2 = \varrho_a^2 s_a^2$ gives

$$\sigma^2_{n,t,a,\text{obs}} - \sigma^2_{n,t,a} = \varrho_a^2 s_a^2\left(\frac{1}{\sum_{v=0}^{t-1}\hat{\pi}_{v,a}b_v} - \frac{1}{\sum_{v=0}^{t-1}b_v}\right) \geq 0,$$

because $\sum_{v=0}^{t-1}b_v \geq \sum_{v=0}^{t-1}\hat{\pi}_{v,a}b_v$. Equality holds if $\varrho_a = 0$ or all units in batches $0,\ldots,t-1$ were assigned to arm a; otherwise, the variance reduction is strict. This compares variances assigned by the two working posteriors under the same realized allocation, not the frequentist variances of their estimators under adaptive sampling.

B.7 LEMMA FOR THEOREM 3

Lemma 3 (Sample augmentation under a fixed RHO allocation)**.** *For any fixed pre-LLM state x_t^- and allocation $\bar{\rho} \in \Delta_K$,*

$$\widetilde{V}_t^{\text{RHOPE-Delayed}}(\bar{\rho}, x_t^-) \geq \widetilde{V}_t^{\text{RHO}}(\bar{\rho}, x_t^-).$$

Proof. Fix an arm a and an allocation $\bar{\rho}$. For the proof, define

$$m_{t,a} := \mu_{t,a}^{(1)} + \lambda_a \mu_{t,a}^{(2)}, \qquad q_a := B_t \bar{\rho}_a,$$
$$P_a := (\sigma_{t,a}^{(1)})^2, \qquad Q_a := \lambda_a^2 (\sigma_{t,a}^{(2)})^2, \qquad \tau_{t,a}^2 := P_a + Q_a,$$
$$r_a := s_{a,\text{res}}^2, \qquad u_a := \lambda_a^2 \tilde{s}_a^2.$$

Thus $s_a^2 = r_a + u_a$. We first consider the nondegenerate case $P_a, Q_a, r_a, u_a > 0$; the case $\lambda_a = 0$, for which $Q_a = u_a = 0$, follows by continuity of the Gaussian posterior variances and the expected maximum in these variance parameters.

Under RHO, the terminal posterior variance of h_a after allocating the normalized budget q_a to arm a is

$$v_a^{\text{RHO}} = \left[(\tau_{t,a}^2)^{-1} + \frac{q_a}{r_a + u_a} \right]^{-1}.$$

Under RHOPE-DELAYED, the residual component receives physical precision q_a/r_a, while the proxy component receives all-unit precision B_t/u_a. Its terminal posterior variance for h_a is therefore

$$v_a^{\text{RHOPE-Delayed}} = \left(P_a^{-1} + \frac{q_a}{r_a} \right)^{-1} + \left(Q_a^{-1} + \frac{B_t}{u_a} \right)^{-1}.$$

If $q_a = 0$, then

$$v_a^{\text{RHOPE-Delayed}} = P_a + \left(Q_a^{-1} + \frac{B_t}{u_a} \right)^{-1} \leq P_a + Q_a = v_a^{\text{RHO}}.$$

Suppose henceforth that $q_a > 0$. Since $B_t \geq q_a$,

$$\left(Q_a^{-1} + \frac{B_t}{u_a} \right)^{-1} \leq \left(Q_a^{-1} + \frac{q_a}{u_a} \right)^{-1}.$$

Consider an experiment that observes the residual and proxy components separately with the same normalized information budget q_a:

$$Y_1 = \theta_a^{(1)} + \eta_1, \qquad Y_2 = \lambda_a \theta_a^{(2)} + \eta_2,$$

where $\eta_1 \sim \mathcal{N}(0, r_a/q_a)$ and $\eta_2 \sim \mathcal{N}(0, u_a/q_a)$ are independent of each other and of the parameters. The corresponding posterior variance is

$$\text{Var}(h_a \mid Y_1, Y_2) = \left(P_a^{-1} + \frac{q_a}{r_a} \right)^{-1} + \left(Q_a^{-1} + \frac{q_a}{u_a} \right)^{-1}.$$

The RHO observation is the sum

$$Y_1 + Y_2 = h_a + \eta_1 + \eta_2,$$

whose noise variance is $(r_a + u_a)/q_a$, and hence

$$\text{Var}(h_a \mid Y_1 + Y_2) = v_a^{\text{RHO}}.$$

Because $Y_1 + Y_2$ is a measurable function of (Y_1, Y_2), the law of total variance, conditional on $Y_1 + Y_2$, gives

$$\begin{aligned} \text{Var}(h_a \mid Y_1 + Y_2) = \mathbb{E}[\text{Var}(h_a \mid Y_1, Y_2) \mid Y_1 + Y_2] \\ + \text{Var}(\mathbb{E}[h_a \mid Y_1, Y_2] \mid Y_1 + Y_2). \end{aligned}$$

The Gaussian conditional variances in this display are deterministic, and the second term is nonnegative. Therefore

$$\text{Var}(h_a \mid Y_1, Y_2) \leq \text{Var}(h_a \mid Y_1 + Y_2).$$

Together with the previous precision inequality, this proves

$$v_a^{\text{RHOPE-Delayed}} \le v_a^{\text{RHO}}.$$

Under a fixed RHO allocation, conditional on the current state, the terminal posterior mean has variance equal to the reduction in Gaussian posterior variance. Consequently, the two fixed-allocation returns in Section 5 can be written as

$$\widetilde{V}_t^{\text{RHO}}(\bar{\rho}, x_t^-) = \mathbb{E}\left[\max_{a\in[K]} \left\{m_{t,a} + \Gamma_a^{\text{RHO}}\right\}\right],$$

and

$$\widetilde{V}_t^{\text{RHOPE-Delayed}}(\bar{\rho}, x_t^-) = \mathbb{E}\left[\max_{a\in[K]} \left\{m_{t,a} + \Gamma_a^{\text{RHOPE-Delayed}}\right\}\right],$$

where the innovations are independent across arms and satisfy

$$\Gamma_a^{\text{RHO}} \sim \mathcal{N}\left(0, \tau_{t,a}^2 - v_a^{\text{RHO}}\right), \qquad \Gamma_a^{\text{RHOPE-Delayed}} \sim \mathcal{N}\left(0, \tau_{t,a}^2 - v_a^{\text{RHOPE-Delayed}}\right).$$

Since $v_a^{\text{RHOPE-Delayed}} \le v_a^{\text{RHO}}$, construct a coupling explicitly as follows. Let the ζ_a's be independent across arms and independent of Γ^{RHO}, with

$$\zeta_a \sim \mathcal{N}\left(0, v_a^{\text{RHO}} - v_a^{\text{RHOPE-Delayed}}\right).$$

Then

$$\Gamma_a^{\text{RHOPE-Delayed}} \overset{d}{=} \Gamma_a^{\text{RHO}} + \zeta_a,$$

jointly across arms. The map $z \mapsto \max_{a\in[K]}\{m_{t,a} + z_a\}$ is convex. Jensen's inequality then gives

$$\mathbb{E}_\zeta\left[\max_{a\in[K]} \left\{m_{t,a} + \Gamma_a^{\text{RHO}} + \zeta_a\right\} \,\middle|\, \Gamma^{\text{RHO}}\right] \ge \max_{a\in[K]} \left\{m_{t,a} + \Gamma_a^{\text{RHO}}\right\}.$$

Taking expectation over Γ^{RHO} proves the lemma. □

B.8 Proof of Theorem 3

Proof. For a fixed $\bar{\rho}$, the expected returns introduced in Section 5 are

$$\widetilde{V}_t^{\text{RHO}}(\bar{\rho}, x_t^-) = \mathbb{E}\left[\max_{a\in[K]} \left\{\mu_{t,a}^{(1)} + \lambda_a \mu_{t,a}^{(2)} + \sqrt{\frac{\left((\sigma_{t,a}^{(1)})^2 + \lambda_a^2(\sigma_{t,a}^{(2)})^2\right)^2 \bar{\rho}_a B_t}{s_a^2 + \left((\sigma_{t,a}^{(1)})^2 + \lambda_a^2(\sigma_{t,a}^{(2)})^2\right)\bar{\rho}_a B_t}} Z_a\right\}\right].$$

$$\widetilde{V}_t^{\text{RHOPE-Delayed}}(\bar{\rho}, x_t^-) = \mathbb{E}\left[\max_{a\in[K]} \left\{\begin{array}{l}\mu_{t,a}^{(1)} + \lambda_a \mu_{t,a}^{(2)} + \sqrt{\dfrac{(\sigma_{t,a}^{(1)})^4 \bar{\rho}_a B_t}{s_{a,\text{res}}^2 + (\sigma_{t,a}^{(1)})^2 \bar{\rho}_a B_t}} Z_a^G \\ \qquad + \lambda_a \sqrt{\dfrac{(\sigma_{t,a}^{(2)})^4 B_t}{\tilde{s}_a^2 + (\sigma_{t,a}^{(2)})^2 B_t}} Z_a^J\end{array}\right\}\right].$$

$$\widetilde{V}_t^{\text{RHOPE}}(\bar{\rho}, x_t^+) = \mathbb{E}\left[\max_{a\in[K]} \left\{\begin{array}{l}\mu_{t,a}^{(1)} + \lambda_a \mu_{t+1,a}^{(2)} + \sqrt{\dfrac{(\sigma_{t,a}^{(1)})^4 \bar{\rho}_a B_t}{s_{a,\text{res}}^2 + (\sigma_{t,a}^{(1)})^2 \bar{\rho}_a B_t}} Z_a^G \\ \qquad + \lambda_a \sqrt{\dfrac{(\sigma_{t+1,a}^{(2)})^4 B_{t+1}}{\tilde{s}_a^2 + (\sigma_{t+1,a}^{(2)})^2 B_{t+1}}} Z_a^J\end{array}\right\}\right].$$

All Gaussian innovations in these expressions are independent standard normals across arms and signal types. The RHO comparator starts from the same marginal posterior of h; it ignores future proxies, not any information already encoded in that common starting state. Lemma 3 holds for every fixed $\bar{\rho} \in \Delta_K$. Maximizing its two sides over $\bar{\rho}$ yields the first inequality in Theorem 3.

For the second inequality, define

$$F(\bar{\rho}, J_t) := \widetilde{V}_t^{\mathrm{RHOPE}}\big(\bar{\rho}, x_t^+(J_t)\big).$$

We first show that, for every fixed $\bar{\rho}$,

$$\mathbb{E}_{J_t}\left[F(\bar{\rho}, J_t) \mid x_t^-\right] = \widetilde{V}_t^{\text{RHOPE-Delayed}}(\bar{\rho}, x_t^-). \tag{18}$$

For this purpose, let

$$L_{t,a} := (\sigma_{t,a}^{(2)})^2, \qquad d_a := \tilde{s}_a^2.$$

Conditional on x_t^-, the current proxy update satisfies

$$\mu_{t+1,a}^{(2)} \overset{d}{=} \mu_{t,a}^{(2)} + \sqrt{\frac{L_{t,a}^2 b_t}{d_a + L_{t,a} b_t}}\, Z_{t,a}^{J,0},$$

where $Z_{t,a}^{J,0}$ is standard normal, and

$$L_{t+1,a} := (\sigma_{t+1,a}^{(2)})^2 = \frac{L_{t,a} d_a}{d_a + L_{t,a} b_t}.$$

The current proxy innovation and the future proxy innovation in the RHOPE return are independent. Their combined variance, before multiplying by λ_a^2, is

$$\begin{aligned}&\frac{L_{t,a}^2 b_t}{d_a + L_{t,a} b_t} + \frac{L_{t+1,a}^2 B_{t+1}}{d_a + L_{t+1,a} B_{t+1}} \\ &\quad = \frac{L_{t,a}^2 B_t}{d_a + L_{t,a} B_t},\end{aligned}$$

where the equality follows by substituting the displayed expression for $L_{t+1,a}$ and using $B_t = b_t + B_{t+1}$. Therefore,

$$\begin{aligned}&\lambda_a \mu_{t+1,a}^{(2)} + \lambda_a \sqrt{\frac{(\sigma_{t+1,a}^{(2)})^4 B_{t+1}}{\tilde{s}_a^2 + (\sigma_{t+1,a}^{(2)})^2 B_{t+1}}}\, Z_a^{J,+} \\ &\overset{d}{=} \lambda_a \mu_{t,a}^{(2)} + \lambda_a \sqrt{\frac{(\sigma_{t,a}^{(2)})^4 B_t}{\tilde{s}_a^2 + (\sigma_{t,a}^{(2)})^2 B_t}}\, Z_a^{J},\end{aligned}$$

where $Z_a^{J,+}$ is independent of $Z_{t,a}^{J,0}$, and Z_a^J is a standard normal variable. This equality holds jointly across arms. The residual innovation in $F(\bar{\rho}, J_t)$ is independent of J_t, so the last display identifies the joint distribution inside the expected maximum with that in $\widetilde{V}_t^{\text{RHOPE-Delayed}}(\bar{\rho}, x_t^-)$. This proves equation 18.

For each J_t, the map $\bar{\rho} \mapsto F(\bar{\rho}, J_t)$ is continuous on the compact simplex Δ_K. Hence the maximum over Δ_K is attained; moreover, the resulting maximum is measurable in J_t by the measurable maximum theorem.

For every realization of J_t and every fixed $\bar{\rho}$,

$$\max_{\bar{\rho}' \in \Delta_K} F(\bar{\rho}', J_t) \geq F(\bar{\rho}, J_t).$$

Taking the conditional expectation given x_t^- and then maximizing the right-hand side over $\bar{\rho}$ gives

$$\begin{aligned}\mathbb{E}_{J_t}\left[\max_{\bar{\rho}' \in \Delta_K} F(\bar{\rho}', J_t) \,\middle|\, x_t^-\right] &\geq \max_{\bar{\rho} \in \Delta_K} \mathbb{E}_{J_t}\left[F(\bar{\rho}, J_t) \mid x_t^-\right] \\ &= \max_{\bar{\rho} \in \Delta_K} \widetilde{V}_t^{\text{RHOPE-Delayed}}(\bar{\rho}, x_t^-),\end{aligned}$$

where the equality follows from equation 18. Substituting the definition of F yields the second inequality in the theorem. □

B.9 Proof of Theorem 4

Proof. Let $\Pi_t^{\text{RHOPE-Delayed}}$ be the class of admissible policies whose allocation at each remaining batch $v \geq t$ is measurable with respect to $\mathcal{F}_v^-$, the sigma-field generated by observations before J_v. Let Π_t^{RHOPE} allow allocations measurable with respect to $\mathcal{F}_v^+ = \sigma(\mathcal{F}_v^-, J_v)$. The posterior states are sufficient for future prediction, so the optimal values over these history-dependent policy classes equal the corresponding MDP values. Every $\pi \in \Pi_t^{\text{RHOPE-Delayed}}$ induces a policy $\pi^\uparrow \in \Pi_t^{\text{RHOPE}}$: at batch t, $\pi^\uparrow$ ignores the realized J_t and uses the allocation selected by π; at every subsequent batch it likewise uses the pre-LLM history and follows π's decision rule. Thus the terminal posterior state under $\pi^\uparrow$ has the same conditional distribution, given x_t^-, as the terminal posterior state under π.

Fix any $\pi \in \Pi_t^{\text{RHOPE-Delayed}}$ and its embedded policy $\pi^\uparrow$. For every realized J_t, optimality of the advance-information value function gives

$$V_t^{\text{RHOPE}}(x_t^+) \geq V_t^{\pi^\uparrow,+}(x_t^+),$$

where $V_t^{\pi^\uparrow,+}$ is the expected terminal posterior maximum when the policy $\pi^\uparrow$ is followed from x_t^+. Taking conditional expectation over J_t and using the tower property yields

$$\begin{aligned}\mathbb{E}_{J_t}\left[V_t^{\text{RHOPE}}(x_t^+) \mid x_t^-\right] &\geq \mathbb{E}_{J_t}\left[V_t^{\pi^\uparrow,+}(x_t^+) \mid x_t^-\right] \\ &= V_t^{\pi,-}(x_t^-).\end{aligned}$$

The equality follows from the equality in conditional distribution of the terminal posterior states under $\pi^\uparrow$ and π. Since the last display holds for every $\pi \in \Pi_t^{\text{RHOPE-Delayed}}$, taking the supremum over this class gives

$$\mathbb{E}_{J_t}\left[V_t^{\text{RHOPE}}(x_t^+) \mid x_t^-\right] \geq \sup_{\pi \in \Pi_t^{\text{RHOPE-Delayed}}} V_t^{\pi,-}(x_t^-) = V_t^{\text{RHOPE-Delayed}}(x_t^-),$$

which is the desired inequality. This argument avoids interchanging an expectation and a supremum over policies. □

C Additional Experimental Details

C.1 Synthetic Data Experiments

C.1.1 Details of Synthetic Data Experiments

Every synthetic experiment has $K = 10$ arms and $T = 10$ equal-sized batches. The common batch size is $b_t n \in \{100, 1000\}$, so the total number of experimental units is $10 b_t n$. Conditional on the arm-level parameters, outcome pairs are independent across batches, units, and arms. The proxy outcomes are generated synthetically to emulate LLM information; these experiments do not call an LLM. In all three data-generating processes (DGPs), ϱ denotes the Pearson correlation between the true outcome $Y_{t,j,a}$ and the proxy outcome $\tilde{Y}_{t,j,a}$ for the same unit and arm; the main correlation grid is $\varrho \in \{0.2, 0.4, 0.6\}$.

Bernoulli DGP. For each arm a, we independently draw $p_a \sim \text{Beta}(b_t n, b_t n)$. The true and proxy outcomes have the same Bernoulli marginal distribution:

$$Y_{t,j,a} \mid p_a \sim \text{Bernoulli}(p_a), \qquad \tilde{Y}_{t,j,a} \mid p_a \sim \text{Bernoulli}(p_a).$$

We first draw $Y_{t,j,a}$ and then generate $\tilde{Y}_{t,j,a}$ according to

$$\Pr(\tilde{Y}_{t,j,a} = 1 \mid Y_{t,j,a} = 1, p_a) = p_a + \varrho(1 - p_a), \quad \Pr(\tilde{Y}_{t,j,a} = 1 \mid Y_{t,j,a} = 0, p_a) = p_a(1 - \varrho).$$

This construction preserves both Bernoulli marginals and gives $\text{Corr}(Y_{t,j,a}, \tilde{Y}_{t,j,a} \mid p_a) = \varrho$. Thus the true and proxy outcomes share the same arm means and arm ranking.

Gumbel DGP. For each arm a, we independently draw the location parameter ℓ_a from a Gamma distribution with shape $b_t n$ and scale $1/(b_t n)$. Let $\beta = \sqrt{6\sigma^2}/\pi$, let γ_{E} denote the Euler–Mascheroni constant, and define $g(u) = -\log\{-\log(u)\}$, with Φ denoting the standard normal cumulative distribution function. For independent standard-normal variables $Z_{t,j,a}$ and $E_{t,j,a}$, set

$$W_{t,j,a} = r_\varrho Z_{t,j,a} + \sqrt{1 - r_\varrho^2}\, E_{t,j,a},$$
$$Y_{t,j,a} = \ell_a + \beta g(\Phi(Z_{t,j,a}))\,, \qquad \tilde{Y}_{t,j,a} = \ell_a + \beta g(\Phi(W_{t,j,a}))\,.$$

The latent Gaussian correlation r_ϱ is numerically calibrated to target Pearson correlation ϱ after the Gumbel transformations. Both outcomes have marginal distribution $\mathrm{Gumbel}(\ell_a, \beta)$, mean $\ell_a + \gamma_{\mathrm{E}}\beta$, and variance σ^2. The main grid uses $\sigma^2 = 1$. We additionally consider $\sigma^2 \in \{1/3, 3\}$ at $b_t n = 100$ as a variance ablation.

Bernoulli–Gumbel DGP. This mixed DGP allows the true potential outcomes and proxy outcomes to have different marginal distributions and different arm rankings. Independently for each arm, we draw

$$p_a \sim \mathrm{Beta}(b_t n, b_t n), \qquad \ell_a \sim \mathrm{Gamma}(\text{shape} = b_t n, \text{scale} = 1/(b_t n)), \qquad p_a \perp \ell_a.$$

Using independent standard-normal variables $Z_{t,j,a}$ and $E_{t,j,a}$, we then set

$$Y_{t,j,a} = \mathbf{1}\{Z_{t,j,a} > \Phi^{-1}(1 - p_a)\}\,,$$
$$W_{t,j,a} = r_{\varrho,a} Z_{t,j,a} + \sqrt{1 - r_{\varrho,a}^2}\, E_{t,j,a},$$
$$\tilde{Y}_{t,j,a} = \ell_a + \beta g(\Phi(W_{t,j,a}))\,,$$

where $\beta = \sqrt{6}/\pi$. Because the Bernoulli threshold depends on p_a, the latent Gaussian correlation $r_{\varrho,a}$ is numerically calibrated separately for each arm to target $\mathrm{Corr}(Y_{t,j,a}, \tilde{Y}_{t,j,a} \mid p_a, \ell_a) = \varrho$. The resulting marginal distributions are $\mathrm{Bernoulli}(p_a)$ and $\mathrm{Gumbel}(\ell_a, \beta)$, with means p_a and $\ell_a + \gamma_{\mathrm{E}}\beta$, respectively. Because p_a and ℓ_a are drawn independently, the ranking implied by the proxy outcomes need not match the true arm ranking.

Experimental protocol. The first batch uses balanced uniform assignment for every method, with arm counts differing by at most one. For the optimization-based policies, TS, and TTTS, including their LLM-assisted versions, subsequent posterior updates use realized allocation proportions. Uniform and uniform-LLM also use realized allocation proportions in every posterior update, while keeping their allocation probabilities uniform. Under the flat prior, uniform's terminal posterior mean on the outcome scale equals the observed sample mean for each arm.

For each configuration and each policy, we average results over 5,000 independent replications, using random seed 42 and 12 parallel workers. In one replication, we generate the true potential outcomes and synthetic proxy outcomes prescribed by the corresponding DGP and run the policy over all ten batches. Posterior-based policies update their working posteriors after each batch and recommend the arm with the largest terminal posterior mean. Sequential Halving, TS-SPAS, and MLA-UCB use their respective recommendation rules described below. The DGP-specific variances and covariances are treated as known by policies that require them. At every batch, relative simple regret is 100 times the ratio of the policy's mean simple regret across replications to that of uniform allocation in the same configuration and batch. Thus, uniform is 100 by construction, and smaller values are better.

For RHO, both RHOPE variants, and the two myopic policies, we optimize the allocation on the simplex through a direct softmax parameterization and Adam. We use 1,024 Sobol Monte Carlo draws, 300 optimization steps, learning rate 0.005, three restarts, a simplex floor of 10^{-6}, and early stopping with patience 40 and minimum improvement 10^{-5}. Both components of the LLM-assisted working posterior have zero initial precision, corresponding to flat priors. The TS-SPAS implementation uses a two-stage screening-and-allocation procedure; its finite-support saddle solver has support size 32, 64 Monte Carlo draws, 12 inner steps, 12 outer steps, and tolerance 0.005.

Benchmark policies. The synthetic experiments compare the following policies.

- **uniform** uses equal allocation probabilities across arms and the true-outcome-only posterior.
- **TS** draws arm means from the true-outcome-only posterior and allocates according to the resulting Thompson-sampling decision.
- **TTTS** applies top-two Thompson sampling with top-two mixing probability 0.5, again using the true-outcome-only posterior.
- **myopic** chooses the next-batch allocation by optimizing a one-step objective under the true-outcome-only posterior, without explicitly planning over all remaining batches.
- **RHO** uses the true-outcome-only posterior and the residual horizon optimization objective of Che & Namkoong (2023) to select the allocation.
- **Sequential Halving** is a non-LLM active-set elimination method. It halves the active set over batch stages and retains at least two active arms before recommending the active arm with the largest sample mean.
- **TS-SPAS** is the non-LLM implementation of the two-stage screening plus saddle-point allocation approach of Kato (2025). It first screens arms and then computes a second-stage allocation and terminal decision rule on the retained set.
- **MLA-UCB** uses the machine learning-assisted (MLA) mean estimator of Ji et al. (2025b) and an upper-confidence-bound allocation rule based on true and proxy outcomes. Its terminal recommendation maximizes the MLA mean estimate.

The policies uniform-LLM, TS-LLM, TTTS-LLM, myopic-LLM, and RHOPE are, respectively, the LLM-assisted versions of uniform, TS, TTTS, myopic, and RHO. They combine proxy outcomes with observed true outcomes through finite-batch working-posterior versions of the updates in Section 4.2, with the allocation-proportion conventions above; the difference from their non-LLM counterparts is therefore the information used for posterior updates and the terminal recommendation, and, where applicable, allocation.

C.1.2 Additional Synthetic Results

Table 3: Supplementary final relative simple regret results for the Bernoulli–Gumbel synthetic setting. Each entry is a percentage relative to uniform, which is 100 in every column; lower is better. Boldface marks the lowest value in each column. Dashes omit values for non-LLM methods that are unchanged across ϱ.

	$b_t n = 100$			$b_t n = 1000$		
Policy / ϱ	0.2	0.4	0.6	0.2	0.4	0.6
Without LLM						
uniform	100.00	–	–	100.00	–	–
TS	79.75	–	–	82.64	–	–
TTTS	79.08	–	–	80.00	–	–
Sequential Halving	76.12	–	–	79.37	–	–
TS-SPAS	112.11	–	–	113.85	–	–
myopic	74.55	–	–	76.68	–	–
RHO	74.66	–	–	73.05	–	–
With LLM						
uniform-LLM	100.84	95.00	85.98	99.06	93.26	82.52
TS-LLM	79.67	71.62	64.32	78.88	72.94	62.64
TTTS-LLM	76.87	71.86	60.62	80.46	72.09	64.75
MLA-UCB	97.30	91.54	82.17	94.56	88.51	80.45
myopic-LLM	73.05	69.64	61.91	75.32	69.21	60.65
RHOPE-Delayed	**72.19**	65.76	58.11	**71.99**	**66.62**	58.32
RHOPE	73.21	**64.96**	**57.33**	73.89	66.69	**57.78**

Table 3 reports the main-grid results for the Bernoulli–Gumbel setting. Table 4 reports the Gumbel outcome-variance ablations for (a) $\sigma^2 = 1/3$ and (b) $\sigma^2 = 3$. Together with Table 1 in the main

text, these tables give the final, tenth-batch relative simple regret for the synthetic configurations considered.

Figures 1–4 show how relative simple regret evolves across the ten batches. Each curve is normalized by uniform in the same configuration and batch, so values below 100 indicate lower simple regret than uniform at that point in the experiment. In the main-grid figures, each image contains Bernoulli, Gumbel, and Bernoulli–Gumbel panels. The variance-ablation figures contain the Gumbel setting only.

Table 4: Gumbel outcome-variance ablations for (a) $\sigma^2 = 1/3$ and (b) $\sigma^2 = 3$. Entries are final simple regret percentages relative to uniform; lower is better. Boldface marks column minima; dashes denote non-LLM results unchanged across ϱ.

(a) $\sigma^2 = 1/3$

Policy / ϱ	0.2	0.4	0.6
Without LLM			
uniform	100.00	–	–
TS	51.50	–	–
TTTS	51.85	–	–
Sequential Halving	54.99	–	–
TS-SPAS	166.67	–	–
myopic	50.64	–	–
RHO	46.83	–	–
With LLM			
uniform-LLM	92.02	85.97	70.05
TS-LLM	50.97	44.61	36.49
TTTS-LLM	49.24	42.30	33.19
MLA-UCB	93.36	81.31	66.99
myopic-LLM	46.20	41.46	33.63
RHOPE-DELAYED	43.85	42.32	33.06
RHOPE	**43.17**	**40.95**	**32.29**

(b) $\sigma^2 = 3$

Policy / ϱ	0.2	0.4	0.6
Without LLM			
uniform	100.00	–	–
TS	81.46	–	–
TTTS	81.27	–	–
Sequential Halving	74.69	–	–
TS-SPAS	111.86	–	–
myopic	76.37	–	–
RHO	75.12	–	–
With LLM			
uniform-LLM	96.23	93.08	83.30
TS-LLM	78.77	72.98	64.04
TTTS-LLM	79.94	75.48	65.53
MLA-UCB	94.44	89.10	79.95
myopic-LLM	75.66	72.03	63.78
RHOPE-DELAYED	**73.20**	68.71	60.97
RHOPE	73.44	**68.61**	**60.50**

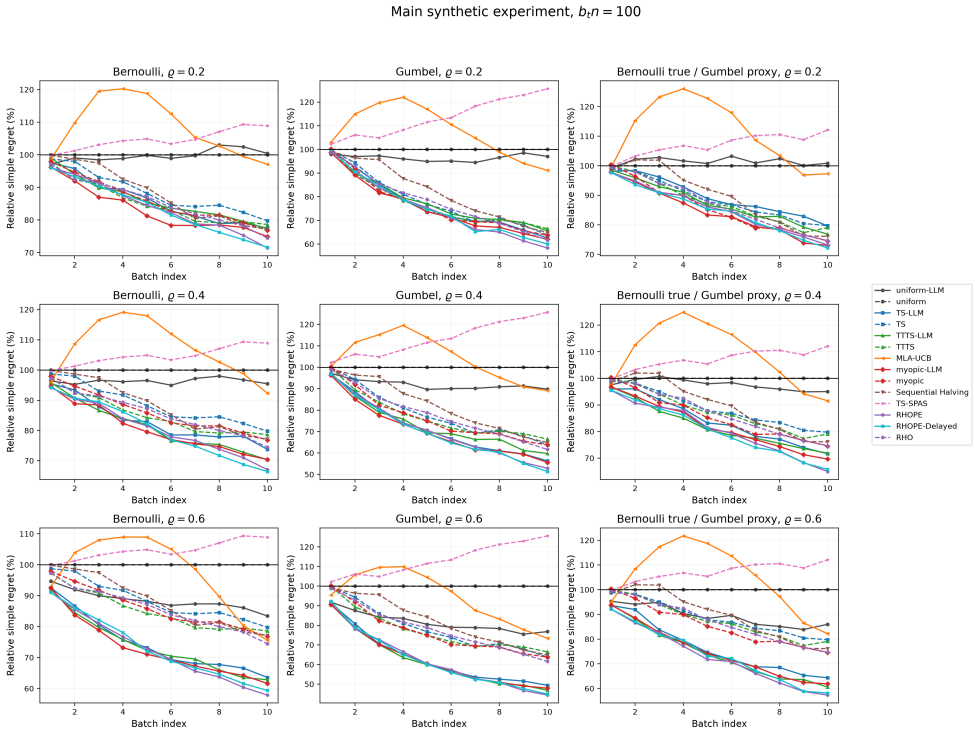

Figure 1: Batch-level relative simple regret trajectories in the main synthetic experiments with $b_t n = 100$. From top to bottom, the rows have $\varrho = 0.2, 0.4, 0.6$; the columns report Bernoulli, Gumbel, and Bernoulli–Gumbel results, respectively.

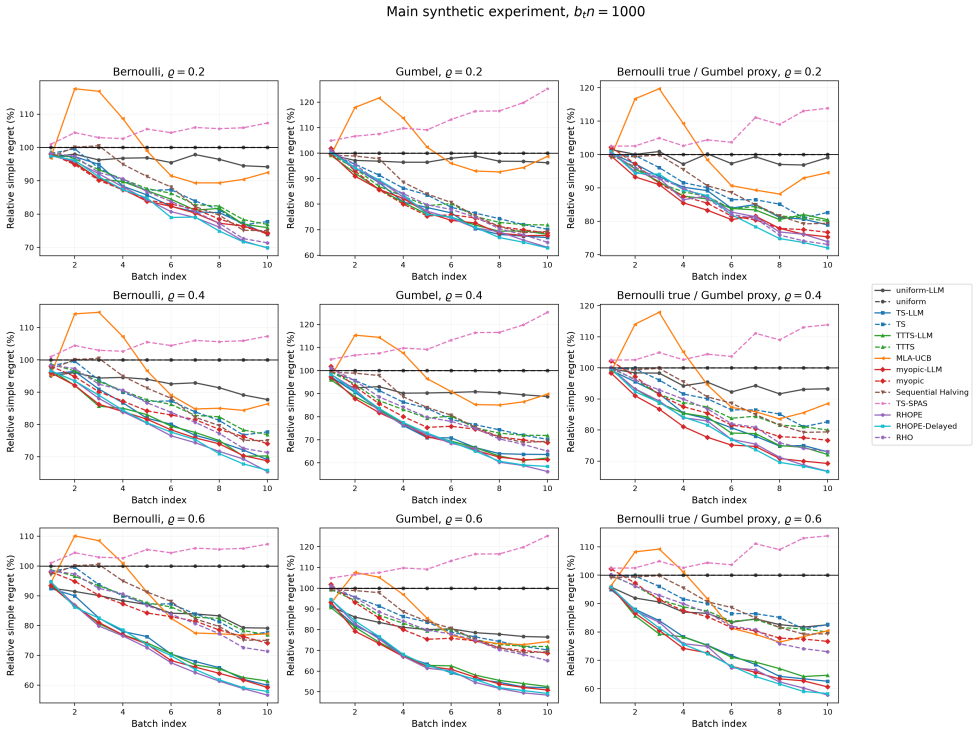

Figure 2: Batch-level relative simple regret trajectories in the main synthetic experiments with $b_t n = 1000$. From top to bottom, the rows have $\varrho = 0.2, 0.4, 0.6$; the columns report Bernoulli, Gumbel, and Bernoulli–Gumbel results, respectively.

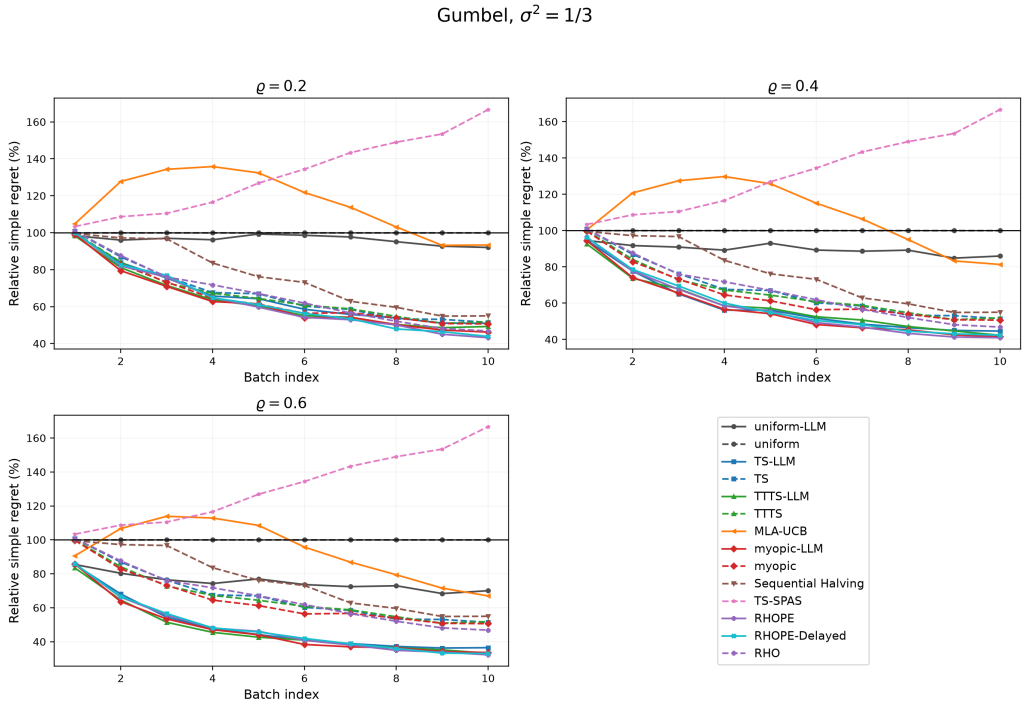

Figure 3: Batch-level relative simple regret trajectories in the Gumbel variance-ablation experiment with $b_t n = 100$ and $\sigma^2 = 1/3$. The three trajectory panels correspond to $\varrho = 0.2, 0.4, 0.6$; the lower-right panel contains the shared legend.

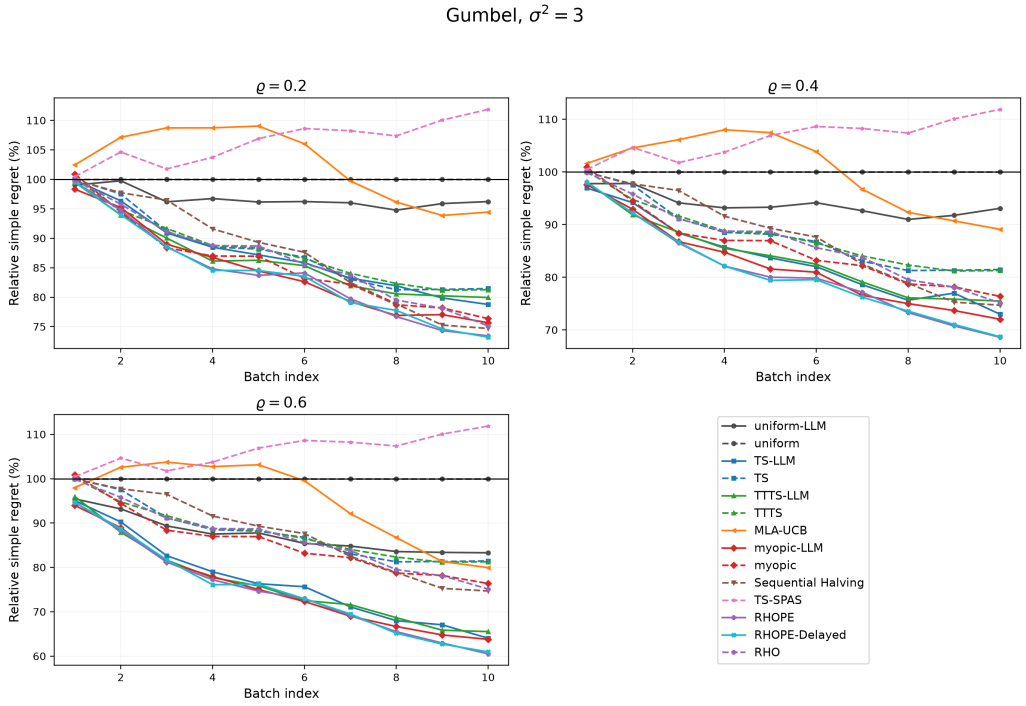

Figure 4: Batch-level relative simple regret trajectories in the Gumbel variance-ablation experiment with $b_t n = 100$ and $\sigma^2 = 3$. The three trajectory panels correspond to $\varrho = 0.2, 0.4, 0.6$; the lower-right panel contains the shared legend.

C.2 Semi-Synthetic Data Experiments

C.2.1 Details of Semi-Synthetic Data Experiments

The semi-synthetic experiment starts from the GDSC response matrix and its cell-line covariates. The source data contain 896 cell lines, complete cell-line-level potential outcomes for seven cancer drugs, and 18,522 features per cell line, including mutation indicators, copy-number alterations, phenotype and tissue covariates, and expression features. The raw response score is smaller for a cell line that is more sensitive to the drug. We reserve a fixed calibration pool of 32 cell lines and exclude all of them from the experiment. The remaining 864 cell lines form the fixed semi-synthetic instance, with the seven drugs defining $K = 7$ treatment arms. The zero-shot prompt uses none of the calibration profiles, while the 4-shot and 16-shot prompts use nested subsets of this pool.

For each experimental cell line c, synthetic patient i, and treatment arm a, we construct the true raw potential outcome as

$$Y^{\mathrm{raw}}_{c,i,a} = Y^{\mathrm{raw}}_{c,a} + \eta_{c,i,a},$$

where $Y^{\mathrm{raw}}_{c,a}$ is the observed cell-line-level raw response in GDSC and $\eta_{c,i,a} \sim \mathcal{N}(0, 9)$ is the synthetic patient noise. The adaptive policies operate on outcomes $Y_{c,i,a} = -Y^{\mathrm{raw}}_{c,i,a}$, so a larger outcome corresponds to greater drug sensitivity. Thus, the patient-level variation is synthetic while the cell-line profiles and the seven treatment-specific baseline potential outcomes remain anchored in the real dataset.

We use DeepSeek V4 Flash (API model "deepseek-v4-flash") to produce LLM-generated proxy outcomes. Because directly providing all 18,522 covariates is impractical, the prompt uses 785 selected mutation, copy-number, phenotype, and tissue features; expression features are excluded. Nonzero selected features are converted into a natural-language tumor-profile description, while unlisted mutation and copy-number indicators are explicitly declared to be zero. The system prompt describes the benchmark's standardized continuous response scale, the names, targets, and pathways of the seven drugs, and the required seven-key JSON schema. The zero-shot condition contains no calibration examples. The 4-shot and 16-shot conditions embed complete profile-and-response examples from the nested calibration subsets in the system prompt. Generation uses temperature 1, a maximum of 256 output tokens, JSON mode, and disabled thinking mode. The following is an actual target user prompt shared across the three prompting conditions; the tumor-profile block identifies the target cell line, the feature blocks describe its selected contextual information, the all-unlisted-zero statement specifies the omitted binary indicators, and the task block requests the seven raw response scores.

Tumor profile:
- Profile ID: profile_000000
- Active mutation indicators:
- Coding mutation or alteration indicators are active for: MLL2, HRAS, TET2.
- Active copy-number alteration indicators:
- Copy-number loss indicators are active for: cna_pancan77 (CASP1);
cna_pancan78 (CASP1); cna_pancan79; cna_pancan80 (SDHD);
cna_pancan230 (SYNCRIP, ZNF292); cna_pancan334; cna_pancan335;
cna_pancan336; cna_pancan337; cna_pancan338; cna_pancan339 (ATRX);
cna_pancan340; cna_pancan341; cna_pancan342; cna_pancan343.
- Phenotype and tissue descriptors:
- msi: 1.
- haematopoietic neoplasm other descriptor is active.
- suspension descriptor is active.
All mutation and copy-number indicators not listed above have value 0.
Task:
Estimate the raw treatment-response scores for this tumor profile under the seven treatment candidates listed in the system instructions. Lower raw response means stronger drug sensitivity. Return only JSON with exactly the required raw_responses keys.

For every experimental profile and each shot condition, we retain 100 parsed LLM-generated proxy outcome vectors, each covering the seven treatments. Within each replication, we sample ten distinct vectors without replacement from the profile-specific pool and randomly pair them with that profile's ten synthetic patients. Sampling from this bank propagates variation across LLM-generated proxy outcomes into the experimental replications.

Crossing the three shot counts $s \in \{0, 4, 16\}$ with $v \in \{0, 2, 4.5\}$ gives the nine experimental conditions. The LLM bank is generated once per shot count; changing v acts on the proxy after sampling from the bank, without changing the prompt or the true outcomes. Suppressing the shot index, let $\widetilde{Y}^{(0)}_{c,i,a}$ denote the reward-scale proxy outcome drawn from this bank and write $U_{c,i,a} = -\eta_{c,i,a}/3$. Draw $V_{c,i,a} \sim \mathcal{N}(0,1)$ independently of $U_{c,i,a}$ and the LLM bank, and construct

$$\widetilde{Y}^{(v)}_{c,i,a} = \widetilde{Y}^{(0)}_{c,i,a} + \sqrt{v}\left(\sqrt{\frac{v}{9}}\,U_{c,i,a} + \sqrt{1-\frac{v}{9}}\,V_{c,i,a}\right).$$

The same v is used for every arm. This construction leaves each true outcome and its patient-noise variance of 9 unchanged while increasing both the proxy variance and proxy–outcome covariance by v. In particular, $v = 0$ recovers the unaugmented LLM outcomes; for $v > 0$, the LLM outcomes are augmented by a synthetic patient-level shared signal. For shot count s, let $A_a = \mathrm{Var}(Y_{c,i,a})$, $B_{a,s} = \mathrm{Var}(\widetilde{Y}^{(0)}_{c,i,a})$, and $C_{a,s} = \mathrm{Cov}(Y_{c,i,a}, \widetilde{Y}^{(0)}_{c,i,a})$, with the moments taken over the fixed experimental profiles, their LLM proxy bank, and patient noise. The resulting arm-wise correlation is

$$\mathrm{Corr}\left(Y_{c,i,a}, \widetilde{Y}^{(v)}_{c,i,a}\right) = \frac{C_{a,s} + v}{\sqrt{A_a(B_{a,s} + v)}}.$$

Thus, a common v does not impose a common correlation across arms: their baseline variances and covariances remain different. Table 5 reports the correlations implied by the known moments in all nine conditions. The mean across arms rises from 0.066–0.105 at $v = 0$, to 0.286–0.404 at $v = 2$, and to 0.488–0.612 at $v = 4.5$. These are design-moment correlations, not sample correlations from one replication. True outcomes, batch membership, proxy-bank selections, the Gaussian draws $V_{c,i,a}$, and policy random seeds are coupled across the three v conditions.

Table 5: Arm-wise proxy–outcome correlations in the nine semi-synthetic conditions. The shared v increases covariance and proxy variance by v for every arm. Arms 1–7 are cisplatin, vismodegib, palbociclib, trametinib, dabrafenib, afatinib, and olaparib, respectively. Means are over the seven arms; entries are rounded independently.

Shots	v	Arm 1	Arm 2	Arm 3	Arm 4	Arm 5	Arm 6	Arm 7	Mean
0	0	0.001	0.000	-0.003	0.199	0.189	0.079	0.000	0.066
0	2	0.212	0.280	0.233	0.352	0.361	0.313	0.252	0.286
0	4.5	0.419	0.519	0.448	0.505	0.531	0.518	0.479	0.488
4	0	0.005	-0.002	-0.032	0.235	0.268	0.079	0.014	0.081
4	2	0.395	0.417	0.313	0.416	0.448	0.389	0.355	0.391
4	4.5	0.628	0.655	0.546	0.577	0.618	0.601	0.595	0.603
16	0	0.024	0.014	0.038	0.258	0.283	0.097	0.018	0.105
16	2	0.397	0.435	0.363	0.423	0.462	0.374	0.372	0.404
16	4.5	0.629	0.667	0.585	0.577	0.630	0.584	0.610	0.612

Each of the 864 cell lines contributes one patient to each of $T = 10$ batches, yielding 864 patients per batch and 8,640 patients in total. We evaluate each LLM-assisted policy in 10,000 replications for every shot-by-v condition using seed 42; non-LLM policy results are reused because their inputs are invariant to shot count and v. As in the synthetic experiments, the first batch is uniformly assigned. Posterior-based policies recommend the arm with the largest terminal posterior mean; the other baselines retain the recommendation rules described above. Subsequent optimization-based updates use realized allocation proportions.

D Ablation Experiments with Estimated Moments

The known-moment conditions in Sections 6.1 and 6.2 supply policies with the true arm-level variances and covariances. Here we report the complete ablations with estimated moments, including the unaugmented-proxy GDSC results discussed in Section 6.2. These moments are instead estimated from experimental data. The first batch remains a forced-uniform batch, so neither the unknown moments nor their estimates affect its allocation. After observing that batch, we construct arm-specific plug-in estimates, recompute the first posterior update with those estimates, and hold them fixed for all subsequent posterior updates and allocation decisions.

Specifically, let $\mathcal{I}_{0,a} = \{j : A_{0,j} = a\}$ and $N_{0,a} = |\mathcal{I}_{0,a}| \geq 2$. The barred quantities below are the sample means over the indicated assigned or all-unit samples. We estimate the true-outcome variance from the assigned observations, the proxy variance from all units in the batch, and their covariance from the paired assigned observations:

$$\hat{s}^2_{a,Y} = \frac{1}{N_{0,a}-1} \sum_{j \in \mathcal{I}_{0,a}} \left(Y_{0,j,a} - \bar{Y}^{\text{obs}}_{0,a}\right)^2,$$

$$\hat{s}^2_{a,\tilde{Y}} = \frac{1}{b_0 n - 1} \sum_{j=1}^{b_0 n} \left(\tilde{Y}_{0,j,a} - \bar{\tilde{Y}}^{\text{all}}_{0,a}\right)^2,$$

$$\hat{c}_{a,Y\tilde{Y}} = \frac{1}{N_{0,a}-1} \sum_{j \in \mathcal{I}_{0,a}} \left(Y_{0,j,a} - \bar{Y}^{\text{obs}}_{0,a}\right)\left(\tilde{Y}_{0,j,a} - \bar{\tilde{Y}}^{\text{obs}}_{0,a}\right).$$

We then replace each displayed variance estimate by its maximum with 10^{-6}, retaining the same notation. To ensure a valid plug-in covariance matrix, we project the covariance so that the implied correlation has an absolute value of at most 0.999. Denoting the projected covariance by $\hat{c}^{\text{proj}}_{a,Y\tilde{Y}}$, the orthogonal-decomposition parameters are

$$\hat{\lambda}_a = \frac{\hat{c}^{\text{proj}}_{a,Y\tilde{Y}}}{\hat{s}^2_{a,\tilde{Y}}}, \qquad \hat{s}^2_{a,\text{res}} = \max\left\{\hat{s}^2_{a,Y} - \frac{(\hat{c}^{\text{proj}}_{a,Y\tilde{Y}})^2}{\hat{s}^2_{a,\tilde{Y}}}, 10^{-6}\right\}.$$

The retained non-LLM policies require only $\hat{s}^2_{a,Y}$. MLA-UCB, Sequential Halving, and TS-SPAS do not use the oracle moments replaced in this exercise, so this ablation reports the remaining 11 policies. All relative simple regret values continue to use uniform as the 100% reference.

D.1 Synthetic Experiments

The ablation uses the three main DGPs, $K = 10$, $T = 10$, $b_t n \in \{100, 1000\}$, and $\varrho \in \{0.2, 0.4, 0.6\}$. Each policy–configuration pair uses 5,000 replications. Table 6 reports the terminal results; no oracle moments are used in the recomputed first-batch posterior update or in subsequent batches.

Table 6: Final relative simple regret in the synthetic experiments with moments estimated from the first batch: (a) Bernoulli, (b) Gumbel with $\sigma^2 = 1$, and (c) Bernoulli–Gumbel settings. Each entry is a percentage relative to uniform; lower is better. Boldface marks the lowest value in each column when that value is displayed. Dashes omit values for non-LLM methods that are unchanged across ϱ.

(a) Bernoulli setting

	$b_t n = 100$			$b_t n = 1000$		
Policy / ϱ	0.2	0.4	0.6	0.2	0.4	0.6
Without LLM						
uniform	100.00	–	–	100.00	–	–
TS	78.57	–	–	77.59	–	–
TTTS	80.24	–	–	74.73	–	–
myopic	77.06	–	–	73.42	–	–
RHO	**74.27**	–	–	70.23	–	–
With LLM						
uniform-LLM	104.40	97.92	86.72	93.74	88.16	79.96
TS-LLM	80.99	77.07	65.83	77.19	70.61	61.10
TTTS-LLM	82.25	74.58	**65.63**	74.77	70.47	62.32
myopic-LLM	81.28	77.65	74.17	73.71	69.49	59.36
RHOPE-DELAYED	77.45	75.96	71.00	**68.28**	**63.75**	57.94
RHOPE	77.85	76.67	69.30	70.93	64.91	**55.41**

(b) Gumbel setting ($\sigma^2 = 1$)

	$b_t n = 100$			$b_t n = 1000$		
Policy / ϱ	0.2	0.4	0.6	0.2	0.4	0.6
Without LLM						
uniform	100.00	–	–	100.00	–	–
TS	67.62	–	–	69.51	–	–
TTTS	71.66	–	–	72.68	–	–
myopic	68.16	–	–	69.01	–	–
RHO	**66.88**	–	–	64.64	–	–
With LLM						
uniform-LLM	104.05	97.84	85.70	96.90	89.56	77.44
TS-LLM	72.08	67.74	61.01	69.52	64.88	53.81
TTTS-LLM	72.94	67.01	**60.04**	69.65	64.64	54.31
myopic-LLM	73.71	74.54	69.11	66.55	61.10	53.53
RHOPE-DELAYED	71.03	72.09	69.07	63.94	**59.33**	51.33
RHOPE	70.60	69.22	68.23	**62.94**	59.60	**50.46**

Table 6 (continued)
(c) Bernoulli–Gumbel setting

	$b_t n = 100$			$b_t n = 1000$		
Policy / ϱ	0.2	0.4	0.6	0.2	0.4	0.6
Without LLM						
uniform	100.00	–	–	100.00	–	–
TS	79.36	–	–	83.72	–	–
TTTS	78.27	–	–	80.40	–	–
myopic	**76.21**	–	–	75.48	–	–
RHO	76.47	–	–	74.01	–	–
With LLM						
uniform-LLM	105.86	100.30	88.91	99.42	94.31	81.89
TS-LLM	83.04	77.57	66.78	79.93	75.48	63.78
TTTS-LLM	80.25	75.69	**64.93**	78.70	71.74	64.13
myopic-LLM	79.48	77.23	70.93	75.89	72.27	60.12
RHOPE-DELAYED	77.95	**73.40**	68.90	73.01	**67.82**	59.07
RHOPE	79.01	73.81	69.48	**71.89**	68.96	**57.92**

Comparing Table 6 with the known-moment results in Tables 1 and 3 reveals a clear role for batch size. When $b_t n = 1000$, the forced-uniform first batch provides approximately 100 observed outcomes per arm, enabling sufficiently accurate plug-in estimates of the arm-level variances and covariances. Consequently, either RHOPE-DELAYED or RHOPE achieves the best performance in every setting, as in the known-moment experiments. When $b_t n = 100$, however, the first batch provides only approximately 10 observed outcomes per arm, making the variance and covariance estimates substantially less accurate; other methods then often outperform the two RHOPE variants. This comparison indicates that strong practical performance from our methods when the required moments are unknown depends on a finite-batch structure with sufficiently large batches. This is precisely the regime targeted in Section 3. As explained there, updating the allocation policy can entail nontrivial operational costs and engineering difficulties, particularly in medical and platform experiments, so the number of reallocations is often finite. In the target applications, this finite number of reallocations is paired with many units in each batch, allowing data from the initial uniform batch to support sufficiently accurate estimates of the required variances and covariances for subsequent allocation decisions.

D.2 SEMI-SYNTHETIC EXPERIMENT

We apply the same first-batch estimation procedure to the semi-synthetic GDSC experiment across all nine combinations of $v \in \{0, 2, 4.5\}$ and zero-shot, 4-shot, or 16-shot prompting. The fixed instance has seven arms, ten batches, and 864 patients per batch; each condition uses 10,000 replications. The first batch therefore provides approximately 123 assigned true outcomes per arm and all 864 proxy outcomes for each arm. Non-LLM results are identical across the nine conditions and are shown once. Table 7 reports the terminal relative simple regret of the 11 policies retained in this ablation.

Table 7: Final relative simple regret (%) in the semi-synthetic GDSC experiment with moments estimated from the first batch. The $v = 0$ conditions use unaugmented LLM proxies; $v > 0$ conditions use controlled synthetic proxy augmentation. Uniform is 100 in every condition; lower values are better. Boldface marks the lowest value in each column when displayed, and dashes denote non-LLM results unchanged across v and prompting conditions.

	$v = 0$			$v = 2$			$v = 4.5$		
Policy / shots	0	4	16	0	4	16	0	4	16
Without LLM									
uniform	100.00	–	–	–	–	–	–	–	–
TS	70.40	–	–	–	–	–	–	–	–
TTTS	67.40	–	–	–	–	–	–	–	–
myopic	64.62	–	–	–	–	–	–	–	–
RHO	**64.02**	–	–	–	–	–	–	–	–
With LLM									
uniform-LLM	97.05	96.74	96.06	94.24	90.80	91.60	88.47	83.66	83.88
TS-LLM	69.49	69.21	69.26	66.89	64.70	63.14	60.73	55.08	53.59
TTTS-LLM	66.80	66.46	65.97	64.86	61.93	61.19	59.42	55.05	51.62
myopic-LLM	67.08	66.65	67.03	64.84	60.70	59.57	57.05	52.49	53.66
RHOPE-DELAYED	64.35	64.62	64.55	62.88	59.11	60.03	58.28	**51.94**	52.04
RHOPE	64.52	**63.95**	64.51	**62.32**	**58.85**	**59.27**	**56.88**	52.59	**51.58**

Comparing Table 7 with the known-moment results in Table 2 illustrates a trade-off between the information supplied by proxy outcomes and the error introduced by estimating the required moments. Replacing known variances and covariances with first-batch estimates can distort the proxy correction and posterior uncertainty, thereby impairing subsequent allocation decisions. At $v = 0$, Table 5 shows that the within-arm correlations between true and proxy outcomes are weak overall under all three prompting conditions: many are close to zero, and the mean across arms ranges from 0.066 to 0.105. The resulting gains in estimation precision are limited and can be insufficient to offset the losses from moment estimation. Consistent with this interpretation, the advantage of RHOPE and RHOPE-DELAYED over RHO largely disappears at $v = 0$: neither method improves on RHO under zero-shot or 16-shot prompting, while RHOPE has only a small advantage under 4-shot prompting.

In contrast, at $v = 2$ and $v = 4.5$, the within-arm proxy–outcome correlations are substantially larger, with means across arms of 0.286–0.404 and 0.488–0.612, respectively (Table 5). More informative proxies offer greater potential gains in estimation precision. In these six conditions, both RHOPE and RHOPE-DELAYED outperform RHO despite using estimated moments, and one of the two attains the lowest relative simple regret among all evaluated methods in every condition. These results suggest that the stronger proxy information can compensate for the losses associated with moment estimation. Thus, practical gains depend not only on the accuracy of the moment estimates, but also on the strength of the within-arm proxy–outcome correlation. These comparisons concern point estimates, not claims of statistically significant differences among the leading methods.

Taken together, the synthetic and semi-synthetic experiments with estimated moments show that moment-estimation errors can adversely affect our methods and suggest two complementary ways to mitigate their impact in practice. First, larger batches, especially a larger initial uniform batch, can provide a more reliable basis for estimating the required variances and covariances. In our implementation, these estimates are computed from the first batch and then held fixed, so increasing the first-batch sample size directly increases the number of observations available per arm for moment estimation, reducing sampling error and its subsequent influence on posterior updates and allocation decisions. Second, improving the quality of proxy outcomes, in the sense of strengthening their within-arm correlation with true outcomes, can increase the gains in arm-mean estimation precision and help offset the performance losses associated with imperfect moment estimates. The first approach reduces the estimation error itself, whereas the second strengthens the information benefit available despite that error. These results therefore motivate both sufficiently large initial batches and informative proxy outcomes when applying our methods with unknown moments.

E Robustness to Proxy Covariance Misspecification

E.1 A Parsimonious Working Model

The diagonal proxy covariance and the orthogonality between the full proxy and residual vectors in Assumption 1 define a parsimonious working model for posterior updating and allocation. This model requires $3K$ scalar moments: the true-outcome variance, proxy variance, and their covariance for each arm. Without these restrictions, a general model would require the full proxy covariance matrix, the full proxy–true-outcome cross-covariance matrix, and the K true-outcome variances, totaling $K(K+1)/2 + K^2 + K = 3K(K+1)/2$ moments. For the seven-arm semi-synthetic experiment, this increases the number from 21 to 84.

Estimating a covariance involving the true outcome of arm b requires units actually assigned to that arm. An initial uniform batch of 100 units across ten arms, for example, supplies only about ten paired observations per arm. Estimating many cross-covariances from such limited samples can be noisy, and strong cross-arm proxy correlations can make covariance inversion unstable. These errors propagate into posterior uncertainty and allocation optimization, potentially impairing decisions and exacerbating cold-start difficulties. The working model therefore balances flexibility against estimation and computational demands.

The theory in Sections 4.2 and 5.2 establishes guarantees under the stated model assumptions. The synthetic DGPs generate outcome pairs independently across arms conditional on their parameters, satisfy the covariance restrictions, and demonstrate the benefits of our methods in Section 6.1. The semi-synthetic experiments instead examine decision performance when the batch proxy covariance departs from the diagonal working model, as diagnosed below.

E.2 Batch Proxy Correlation Diagnostics

We compute correlations between the actual batch proxy aggregates $\hat{J}_{t,a}$, conditional on the fixed experimental profiles and proxy banks. Each of the ten batches contains exactly one patient from each of the $C = 864$ profiles, so $b_t n = C$. For each profile, its patient's unaugmented seven-arm proxy vector is marginally uniform over the 100 stored joint draws, with independent sampling across profiles. The variance-v augmentation is integrated analytically according to Appendix C.2.

Let $\tilde{\boldsymbol{Y}}_{t,c}$ denote the proxy vector, including this augmentation, for the patient from profile c in batch t, and let $W = C^{-1}\sum_{c=1}^{C} \mathrm{Cov}(\tilde{\boldsymbol{Y}}_{t,c})$. The marginal batch covariance and cross-arm correlations are therefore

$$\mathrm{Cov}(\hat{J}_t) = \frac{1}{b_t^2 n}\sum_{c=1}^{C} \mathrm{Cov}(\tilde{\boldsymbol{Y}}_{t,c}) = \frac{W}{b_t},$$

$$\mathrm{Corr}(\hat{J}_{t,a}, \hat{J}_{t,b}) = \frac{W_{ab}}{\sqrt{W_{aa}W_{bb}}}, \qquad a \neq b.$$

Because every batch contains the same fixed profiles, differences between profile means contribute no sampling covariance to $\hat{J}_t$. Thus, the relevant calculation averages profile-specific covariance matrices before standardization, rather than pooling unit-level observations. Sampling proxy vectors without replacement across batches does not alter this single-batch marginal calculation. This diagnostic concerns only the proxy covariance block; it does not assess proxy–residual orthogonality or dependence across batches.

Table 8 reports the maximum absolute off-diagonal entry of each batch proxy correlation matrix. At $v = 0$, these values are 0.779, 0.622, and 0.511 for zero-shot, 4-shot, and 16-shot prompts; across all nine conditions, they range from 0.100 to 0.779. The nonzero cross-arm correlations show that the actual batch proxy covariance is not diagonal.

E.3 Decision Performance Under Misspecification

Despite this cross-arm dependence in the batch proxy signals, either RHOPE or RHOPE-Delayed attains the lowest final relative simple regret point estimate in all nine known-moment conditions

Table 8: Batch proxy–proxy correlations in the nine semi-synthetic conditions. Each entry is the maximum absolute correlation $\max_{a \neq b} |\operatorname{Corr}(\hat{J}_{t,a}, \hat{J}_{t,b})|$, computed analytically under the balanced sampling design with fixed profiles and proxy banks, including the variance-v augmentation.

Shots	v	Batch proxy–proxy
0	0	0.779
0	2	0.479
0	4.5	0.324
4	0	0.622
4	2	0.214
4	4.5	0.117
16	0	0.511
16	2	0.181
16	4.5	0.100

(Table 2). RHOPE itself achieves 51.41%–64.36%, compared with 64.51% for RHO. With moments estimated from the first batch, one of the two variants is best in seven of nine conditions, including all six with $v > 0$ (Table 7); RHO performs better in the zero-shot and 16-shot conditions at $v = 0$.

Together, these results provide empirical evidence that the decision benefits of our methods persist under deviations from the diagonal proxy covariance model in the settings studied. They support robustness of decision performance, not exact posterior calibration or universal superiority under arbitrary dependence. Comparisons refer to point estimates, without asserting statistical significance for small differences among the leading methods.

F ALGORITHMS

The subscript n distinguishes finite-batch working-posterior quantities from their Gaussian-limit counterparts. Algorithms 1–2 implement allocation rules derived within the Gaussian sequential experiment, which provides their asymptotic framework. In that model, $\pi_{t,a}$ determines the residual posterior precision increment. In the finite-batch updates below, $\pi_t = \rho_t$ is the chosen allocation, whereas $\hat{\pi}_{t,a}$ is the realized proportion used to account for the actual sample count. The vectors Z^J, Z^G have mutually independent standard-normal coordinates. Ties in the terminal recommendation are broken by choosing the smallest arm index.

The displayed optimization steps assume proper current posteriors. For the flat-prior experiments, both initial precisions and precision-weighted means are zero. We replace the first optimization and categorical assignment by balanced uniform assignment, with at least one observation per arm, and accumulate the displayed update increments in precision form. Optimization starts after this initialization, when the working posteriors are proper. Experiments with estimated moments additionally use the first-batch estimates in Appendix D.

Algorithm 1: RHOPE

Input. Total budget n; batch fractions $b_0, \dots, b_{T-1}$; variances $s_a^2, \tilde{s}_a^2$ and correlations ϱ_a, $a \in [K]$; prior means and standard deviations $(\boldsymbol{\mu}_0^{(1)}, \boldsymbol{\sigma}_0^{(1)}, \boldsymbol{\mu}_0^{(2)}, \boldsymbol{\sigma}_0^{(2)})$.

Initialize. Set the initial MDP state $(\boldsymbol{\mu}_{n,0}^{(1)}, \boldsymbol{\sigma}_{n,0}^{(1)}, \boldsymbol{\mu}_{n,0}^{(2)}, \boldsymbol{\sigma}_{n,0}^{(2)}) = (\boldsymbol{\mu}_0^{(1)}, \boldsymbol{\sigma}_0^{(1)}, \boldsymbol{\mu}_0^{(2)}, \boldsymbol{\sigma}_0^{(2)})$.

1. For $t = 0, \dots, T-1$:

   (a) For every $a \in [K]$, form the all-unit proxy aggregate estimator $\hat{J}_{t,a} = \frac{1}{b_t\sqrt{n}} \sum_{j=1}^{b_t n} \tilde{Y}_{t,j,a}$, and update the proxy posterior by

   $$(\sigma_{n,t+1,a}^{(2)})^{-2} = (\sigma_{n,t,a}^{(2)})^{-2} + \frac{b_t}{\tilde{s}_a^2},$$
   $$\mu_{n,t+1,a}^{(2)} = (\sigma_{n,t+1,a}^{(2)})^2 \left[(\sigma_{n,t,a}^{(2)})^{-2} \mu_{n,t,a}^{(2)} + \frac{b_t}{\tilde{s}_a^2} \hat{J}_{t,a} \right].$$

   (b) Let $B_t := \sum_{v=t}^{T-1} b_v$. Solve

   $$\rho_t \in \arg\max_{\bar{\rho} \in \Delta_K} \mathbb{E}_{Z^G, Z^J} \left[\max_{a \in [K]} \left\{ \mu_{n,t,a}^{(1)} + \sqrt{\frac{(\sigma_{n,t,a}^{(1)})^4 \bar{\rho}_a B_t}{s_{a,\mathrm{res}}^2 + (\sigma_{n,t,a}^{(1)})^2 \bar{\rho}_a B_t}} Z_a^G \right.\right.$$
   $$\left.\left. + \lambda_a \mu_{n,t+1,a}^{(2)} + \lambda_a \sqrt{\frac{(\sigma_{n,t+1,a}^{(2)})^4 B_{t+1}}{\tilde{s}_a^2 + (\sigma_{n,t+1,a}^{(2)})^2 B_{t+1}}} Z_a^J \right\} \right].$$

   (c) For every $j = 1, \dots, b_t n$, assign $A_{t,j} \sim \mathrm{Categorical}(\rho_t)$ and set $\xi_{t,j,a} = \mathbf{1}\{A_{t,j} = a\}$ for $a \in [K]$.

   (d) For every $a \in [K]$, update the residual posterior with the allocation-weighted residual aggregate estimator $\hat{G}_{t,a} = \frac{1}{b_t\sqrt{n}} \sum_{j=1}^{b_t n} \xi_{t,j,a} \left(Y_{t,j,a} - \lambda_a \tilde{Y}_{t,j,a} \right)$:

   $$(\sigma_{n,t+1,a}^{(1)})^{-2} = (\sigma_{n,t,a}^{(1)})^{-2} + \frac{\hat{\pi}_{t,a} b_t}{s_{a,\mathrm{res}}^2},$$
   $$\mu_{n,t+1,a}^{(1)} = (\sigma_{n,t+1,a}^{(1)})^2 \left[(\sigma_{n,t,a}^{(1)})^{-2} \mu_{n,t,a}^{(1)} + \frac{b_t}{s_{a,\mathrm{res}}^2} \hat{G}_{t,a} \right].$$

   Here $\hat{\pi}_{t,a} = \frac{1}{b_t n} \sum_{j=1}^{b_t n} \xi_{t,j,a}$ is the realized allocation proportion for arm a in batch t.

2. Return

   $$\hat{a}_T \in \arg\max_{a \in [K]} \left\{ \mu_{n,T,a}^{(1)} + \lambda_a \mu_{n,T,a}^{(2)} \right\}.$$

Algorithm 2: RHOPE-DELAYED

Input. Total budget n; batch fractions $b_0, \dots, b_{T-1}$; variances $s_a^2, \tilde{s}_a^2$ and correlations ϱ_a, $a \in [K]$; prior means and standard deviations $(\boldsymbol{\mu}_0^{(1)}, \boldsymbol{\sigma}_0^{(1)}, \boldsymbol{\mu}_0^{(2)}, \boldsymbol{\sigma}_0^{(2)})$.

Initialize. Set the initial MDP state $(\boldsymbol{\mu}_{n,0}^{(1)}, \boldsymbol{\sigma}_{n,0}^{(1)}, \boldsymbol{\mu}_{n,0}^{(2)}, \boldsymbol{\sigma}_{n,0}^{(2)}) = (\boldsymbol{\mu}_0^{(1)}, \boldsymbol{\sigma}_0^{(1)}, \boldsymbol{\mu}_0^{(2)}, \boldsymbol{\sigma}_0^{(2)})$.

1. For $t = 0, \dots, T-1$:

   (a) Let $B_t := \sum_{v=t}^{T-1} b_v$. Before using any proxy outcomes from batch t, solve

   $$\rho_t \in \arg\max_{\bar{\rho} \in \Delta_K} \mathbb{E}_{Z^G, Z^J}\left[\max_{a \in [K]} \left\{ \mu_{n,t,a}^{(1)} + \lambda_a \mu_{n,t,a}^{(2)} + \sqrt{\frac{(\sigma_{n,t,a}^{(1)})^4 \bar{\rho}_a B_t}{s_{a,\mathrm{res}}^2 + (\sigma_{n,t,a}^{(1)})^2 \bar{\rho}_a B_t}} Z_a^G + \lambda_a \sqrt{\frac{(\sigma_{n,t,a}^{(2)})^4 B_t}{\tilde{s}_a^2 + (\sigma_{n,t,a}^{(2)})^2 B_t}} Z_a^J \right\} \right].$$

   (b) For every $j = 1, \dots, b_t n$, assign $A_{t,j} \sim \mathrm{Categorical}(\rho_t)$ and set $\xi_{t,j,a} = \mathbf{1}\{A_{t,j} = a\}$ for $a \in [K]$.

   (c) Once the allocation is fixed, incorporate the current batch's all-arm proxy outcomes. For every $a \in [K]$, form $\hat{J}_{t,a} = \frac{1}{b_t \sqrt{n}} \sum_{j=1}^{b_t n} \tilde{Y}_{t,j,a}$ and update the proxy posterior by

   $$(\sigma_{n,t+1,a}^{(2)})^{-2} = (\sigma_{n,t,a}^{(2)})^{-2} + \frac{b_t}{\tilde{s}_a^2},$$
   $$\mu_{n,t+1,a}^{(2)} = (\sigma_{n,t+1,a}^{(2)})^2 \left[(\sigma_{n,t,a}^{(2)})^{-2} \mu_{n,t,a}^{(2)} + \frac{b_t}{\tilde{s}_a^2} \hat{J}_{t,a} \right].$$

   (d) For every $a \in [K]$, use the assigned true outcomes and their paired proxies to form the residual aggregate estimator $\hat{G}_{t,a} = \frac{1}{b_t \sqrt{n}} \sum_{j=1}^{b_t n} \xi_{t,j,a} \left(Y_{t,j,a} - \lambda_a \tilde{Y}_{t,j,a} \right)$ and update the residual posterior by

   $$(\sigma_{n,t+1,a}^{(1)})^{-2} = (\sigma_{n,t,a}^{(1)})^{-2} + \frac{\hat{\pi}_{t,a} b_t}{s_{a,\mathrm{res}}^2},$$
   $$\mu_{n,t+1,a}^{(1)} = (\sigma_{n,t+1,a}^{(1)})^2 \left[(\sigma_{n,t,a}^{(1)})^{-2} \mu_{n,t,a}^{(1)} + \frac{b_t}{s_{a,\mathrm{res}}^2} \hat{G}_{t,a} \right].$$

   Here $\hat{\pi}_{t,a} = \frac{1}{b_t n} \sum_{j=1}^{b_t n} \xi_{t,j,a}$ is the realized allocation proportion.

2. Return

   $$\hat{a}_T \in \arg\max_{a \in [K]} \left\{ \mu_{n,T,a}^{(1)} + \lambda_a \mu_{n,T,a}^{(2)} \right\}.$$

G NOTATION

Symbol	Meaning
Indices and experimental design	
$[K]$	set of treatment arms $\{1, \dots, K\}$
t, a, j	batch, treatment-arm, and within-batch unit indices, respectively
K	number of treatment arms
T	number of experimental batches
n	total sample budget
$b_t,\ b_t n$	batch-t budget fraction and batch-t sample size
Δ_K	probability simplex over the K treatment arms
$\pi_t,\ \pi_{t,a}$	batch-t allocation vector and its arm-a probability; $\pi_t(\cdot)$ denotes the history-dependent rule producing that vector
$\hat{\pi}_{t,a}$	realized allocation proportion for arm a in batch t
$\pi = (\pi_0, \dots, \pi_{T-1})$	complete adaptive allocation policy
$X_{t,j}$	context of unit j in batch t
$A_{t,j},\ \xi_{t,j,a}$	assigned arm and indicator $\mathbf{1}\{A_{t,j} = a\}$
Outcome model and orthogonal decomposition	
$Y_{t,j,a},\ \tilde{Y}_{t,j,a}$	true potential outcome and proxy outcome for unit j, batch t, and arm a
$\epsilon_{t,j,a},\ \tilde{\epsilon}_{t,j,a}$	noise terms in the true and proxy outcome models
$h_a,\ \tilde{h}_a$	local-scale mean parameters of the true and proxy outcomes for arm a
$s_a^2,\ \tilde{s}_a^2,\ \varrho_a$	true-outcome noise variance, proxy noise variance, and their correlation
λ_a	projection coefficient $\varrho_a s_a / \tilde{s}_a$
$\theta_a^{(1)},\ \theta_a^{(2)}$	residual local parameter $h_a - \lambda_a \tilde{h}_a$ and proxy local parameter $\tilde{h}_a$
$\boldsymbol{\theta}^{(\ell)}$	all-arm parameter vector $(\theta_a^{(\ell)})_{a=1}^K$, $\ell \in \{1, 2\}$
$s_{a,\mathrm{res}}^2$	residual-noise variance $s_a^2 - \lambda_a^2 \tilde{s}_a^2 = s_a^2(1 - \varrho_a^2)$
Λ	diagonal matrix $\mathrm{diag}(\lambda_1, \dots, \lambda_K)$
Decision objective and aggregate signals	
$a^*(h),\ \hat{a}_T$	best arm under h and terminal recommended arm
ν	joint prior distribution of $(h, \tilde{h})$
$\mathfrak{R}_T(\pi, \nu)$	Bayes simple regret at the local scale
$\hat{J}_t,\ J_t$	finite-batch all-unit proxy aggregate estimator and its Gaussian-limit signal
$\hat{G}_t,\ G_t$	finite-batch allocation-weighted residual aggregate estimator and its Gaussian-limit signal
$\hat{\theta}_{t,a}^{(1)},\ \hat{\theta}_{t,a}^{(2)}$	efficient GMM estimators of the transformed parameters
$\bar{Y}_{0:t-1,a}^{\mathrm{obs}},\ \bar{\tilde{Y}}_{0:t-1,a}^{\mathrm{obs}},\ \bar{\tilde{Y}}_{0:t-1,a}^{\mathrm{all}}$	true-outcome and proxy sample means among units assigned to arm a, normalized by their actual count, and the all-unit proxy sample mean
$d_{\mathrm{BL}},\ \mathcal{F}_{\mathrm{BL}}$	bounded-Lipschitz distance and its normalized test-function class
Posterior MDP and residual horizon optimization	
$\mu_{t,a},\ \sigma_{t,a}^2$	posterior mean and variance of h_a after batches $0, \dots, t-1$; $t = 0$ denotes the prior
$\mu_{t,a}^{(\ell)},\ \sigma_{t,a}^{(\ell)}$	Gaussian posterior mean and standard deviation of $\theta_a^{(\ell)}$ after batches $0, \dots, t-1$, $\ell \in \{1, 2\}$
$\mu_{n,t,a}^{(\ell)},\ \sigma_{n,t,a}^{(\ell)}$	finite-batch working-posterior counterparts in Algorithms 1 and 2
$\mu_{n,t,a},\ \sigma_{n,t,a}^2$	finite-batch working-posterior mean and variance of h_a

Symbol	Meaning
$x_t^-,\ x_t^+$	pre-LLM state before J_t and post-LLM state after J_t, before G_t
$V_t^{\pi,-},\ V_t^{\pi,+}$	expected terminal maximum posterior mean under π, before and after observing J_t
$V_t^{\text{RHOPE-Delayed}}, V_t^{\text{RHOPE}}$	optimal MDP values under the delayed and advance-information structures
$\widetilde{V}_t^{\text{RHO}}, \widetilde{V}_t^{\text{RHOPE-Delayed}}, \widetilde{V}_t^{\text{RHOPE}}$	fixed-allocation expected returns under the RHO approximation for RHO, RHOPE-DELAYED, and RHOPE
B_t	remaining budget fraction $\sum_{v=t}^{T-1} b_v$, with the empty sum at $t = T$ equal to zero
$\bar{\rho},\ \rho_t$	candidate budget-weighted average allocation over the remaining horizon (equivalently, a constant allocation vector) and selected RHO allocation vector in batch t
$Z^J,\ Z^G$	independent standard-normal innovation vectors for the proxy and residual components
$\odot,\ \text{diag}$	elementwise vector multiplication and diagonal-matrix construction
Experimental quantities	
ϱ	common within-arm proxy–outcome correlation in the synthetic experiments, distinct from the allocation vector ρ_t
σ^2	Gumbel outcome variance in the synthetic experiments, distinct from posterior variances $\sigma_{t,a}^2$
$s \in \{0, 4, 16\},\ v \in \{0, 2, 4.5\}$	shot count and synthetic proxy-augmentation variance in the semi-synthetic experiments
$C,\ c,\ i$	number of experimental profiles, profile index, and patient index in the semi-synthetic experiments
W	average profile-specific proxy covariance matrix; the batch proxy covariance is W/b_t