Jump Start or False Start? A Theoretical and Empirical Evaluation of LLM-initialized Bandits

Anonymous authors
Paper under double-blind review

Abstract

The recent advancement of Large Language Models (LLMs) offers new opportunities to generate user preference data to warm-start bandits. Recent studies on contextual bandits with LLM initialization (CBLI) have shown that these synthetic priors can significantly lower early regret. However, these findings assume that LLM-generated choices are reasonably aligned with actual user preferences. In this paper, we introduce Noisy-CBLI, a framework for systematically examining how LLM-generated preferences perform when random-response and preference-flipping noise is injected into the synthetic training data. In aligned domains, warm-starting remains effective under moderate preference flipping but loses its advantage at higher corruption levels. Under systematic LLM–human misalignment, LLM-generated priors can lead to higher regret than a cold-start bandit even without injected noise. Random-response noise is empirically milder, degrading performance toward cold-start rather than inducing systematic negative transfer. To explain these behaviors, we derive a general prior-centered confidence bound showing that warm-starting depends on a prior-error term measuring mismatch between the LLM prior and the target reward parameter. Under structured noise assumptions, we obtain closed-form characterizations of how preference flipping and target misalignment affect this term, yielding sufficient conditions under which the warm-start regret bound is tighter than the cold-start bound. We evaluate these results across multiple conjoint datasets and LLMs, using a post hoc estimate of prior error to explain the observed transition between beneficial and harmful LLM warm-starting.

1 Introduction

As a novice attempts to solve a new problem without prior heuristics, the search for a solution often begins as a random walk. This lack of structural guidance mirrors the fundamental challenge in online learning, known as the "cold start" problem. When an agent is initialized *tabula rasa*, without any preconceived notions, the agent faces a vast action space with no means to distinguish between optimal and suboptimal decisions. Consequently, the agent is forced into exploration, often incurring high sample complexity and significant performance penalties before converging to a competitive strategy.

Contextual multi-armed bandits (CBs) have emerged as an essential tool to address this problem in the online learning setting. Here, an agent is tasked with choosing a piece of content for each user in a sequence based on the content's and users' features. The agent receives feedback (e.g., a click) usually instantaneously after choosing the content, and may use this feedback to update itself before choosing the next piece of content. By simultaneously balancing *exploration* (gathering information about user preferences) and *exploitation* (utilizing the information gathered to maximize some reward function), CBs optimize real-time recommendations (Li et al., 2010) and admit a sublinear finite-time regret bound under linear payoff assumptions (Chu et al., 2011). However, when CBs have yet to gather any user data, they perform essentially randomly and thus exhibit linear regret (Li et al., 2010; Auer et al., 2002). Traditional approaches have sought to address this limitation by warm-starting bandits using historical user data or expert knowledge (Zhang et al., 2019).

The recent advancement of Large Language Models (LLMs) offers new opportunities and alternatives, providing built-in knowledge about human preferences (Brown et al., 2020). Alamdari et al. (2024) introduced

the Contextual Bandits with LLM Initialization (CBLI) framework, which prompts an LLM to simulate user preferences to generate a synthetic pre-training dataset for a contextual bandit. By simulating bandit performance using data from a conjoint survey experiment, the authors showed that "jump-starting" a CB with this synthetic data achieved an impressive 14–20% reduction in early regret. This demonstrated that even if the LLM-generated preferences are not perfectly accurate, they can still provide a much better starting point than no prior data. The key limitation of their work is that it focuses on a single domain and does not investigate what statistical mechanisms yield the obtained results.

Previous studies have shown conflicting evidence on whether LLMs can accurately simulate human decision making (Bender et al., 2021; Kosinski, 2024). The original CBLI results implicitly rely on an alignment assumption: that LLM-simulated preferences are reasonably close to human preferences on the target task. Despite the demonstrated benefits of the CBLI framework, its robustness to bias and misalignment in these LLM-generated priors is insufficiently understood. Understanding when the framework may break down is critical before deployment in real-world systems.

In this paper, we present a theoretical and empirical study on LLM-generated priors for bandit algorithms. Consistent with the experimental protocol established in the original CBLI framework (Alamdari et al., 2024), and the broader literature addressing the cold-start problem in personalized recommendation (Li et al., 2010; Zhou & Brunskill, 2016), we situate our work within the domain of recommender systems — a core component of modern digital platforms designed to help users navigate environments characterized by extreme information overload. We specifically focus on contextual bandits (and their sleeping counterparts (Kanade et al., 2009)), as they provide a principled framework to integrate dynamic factors (e.g., time, location) that have been shown to significantly enhance recommendation quality (Panniello et al., 2009).

In summary, our contributions are threefold:

- **Noisy-CBLI framework**: We introduce a novel extension to CBLI where synthetic noise is injected into the LLM-generated preference data before pretraining the bandit. We consider two noise injection strategies: (a) Random Replacement—replacing a certain percentage of LLM-generated responses with random choices, and (b) Preference Flipping—flipping the chosen option in a binary choice for a certain percentage of the responses. This framework allows us to simulate varying levels and types of LLM errors and study their impact.

- **Systematic noise impact evaluation**: We conduct an empirical study across three conjoint datasets and multiple LLMs to measure how noise and misalignment affect cumulative regret. For aligned domains, clean LLM warm starts reduce regret, moderate preference flipping preserves much of the benefit, and higher flip rates eventually reverse the advantage. In contrast, when LLM preferences are weakly aligned or misaligned with human responses, LLM priors can be marginal or harmful even without injected noise, and additional directional corruption can amplify negative transfer. Across these settings, random-response corruption is empirically milder than preference flipping, typically degrading performance toward cold-start rather than inducing systematic harm.

- **Alignment-based theoretical analysis**: We develop a theoretical analysis of CBLI in a sleeping linear contextual bandit model with an LLM-induced prior. The analysis identifies a single prior-error term, $\mathcal{B}_0 = \|\theta_0 - \theta^\star\|_{A_0}$, that captures how the mismatch between the LLM prior and the target reward parameter enters the LinUCB confidence radius. This yields a sufficient condition under which the warm-start regret bound is tighter than the cold-start bound. We connect this theory to the experiments using a post hoc estimate of prior error, empirically showing that smaller estimated prior error is generally associated with beneficial warm-starting, while a larger prior error corresponds to marginal gains or negative transfer.

2 Related Works

Contextual and cold-start bandits. Contextual and non-contextual bandits formalize online personalization under partial feedback, with linear methods such as LinUCB and related regret analyses forming a

standard baseline (Li et al., 2010; Chu et al., 2011; Auer et al., 2002). As discussed above and following previous works and experimentation (Alamdari et al., 2024; Zhou & Brunskill, 2016; Panniello et al., 2009), we situate our work in the domain of recommender systems. Sequence-aware and session-based models capture evolving user preferences over interaction histories but do not fully resolve user and item cold-start issues in deployed systems (Hidasi et al., 2016; Quadrana et al., 2018). Variants of contextual bandits explicitly targeting cold-start recommendation include latent contextual bandits for new-user personalization (Zhou & Brunskill, 2016), and broader overviews of such extensions are given in contextual bandit surveys such as Zhou (2016). Relatedly, settings with stochastic action availability motivate "sleeping" bandit formulations (Kanade et al., 2009).

Warm-starting and transfer in bandits. A common approach to the cold-start problem is to initialize bandit learning with supervised or logged feedback. Robust warm-starting methods analyze regimes where offline and online signals diverge and propose procedures that mitigate harmful initialization (Zhang et al., 2019). Transfer learning for contextual bandits similarly characterizes how source–target similarity governs whether transfer reduces regret or induces negative transfer (Cai et al., 2024). Related multi-task formulations treat transfer learning as shared structure across bandit tasks via hierarchical priors (Hong et al., 2022), while recent work explicitly targets negative transfer under covariate shift in latent and other contextual bandits (Deng et al., 2025).

LLMs in recommendation and sequential decision-making. More recently, LLMs have been used to improve recommenders by reframing recommendation as language modeling/prompting (Geng et al., 2023; Petrov & Macdonald, 2023), and by generating sequential recommendations autoregressively (Volodkevich et al., 2024). LLMs have also been incorporated into contextual bandit pipelines as auxiliary signal providers (Baheri & Alm, 2023), and more broadly positioned as agents for sequential decision-making (Zhang et al., 2023). Conversational recommendation and LLM-centered systems have also shown promise (Gao et al., 2023; Bao et al., 2023), with instruction tuning and alignment methods providing mechanisms by which model outputs may approximate human feedback (Ouyang et al., 2022; Bai et al., 2022) supported by evidence on scaling and capability trends (Brown et al., 2020; OpenAI, 2024).

LLMs as preference simulators and synthetic label generators. Beyond their role as representation learners and controllers, LLMs are increasingly used as simulated participants ("silicon samples") in behavioral, social-science, and preference-elicitation studies (Argyle et al., 2023; Sarstedt et al., 2024). Evidence is mixed: on the one hand, LLMs can match some aggregate patterns in human judgments in specific settings, including theory-of-mind style tasks and certain interactive behaviors (Kosinski, 2024), and have shown promise in jump-starting bandit recommenders with synthetic preference data (Alamdari et al., 2024). On the other hand, multiple evaluations show systematic deviations that undermine naive "drop-in" use, including ordering/labeling artifacts and a tendency toward near-uniform or otherwise distorted response distributions once such artifacts are controlled (Dominguez-Olmedo et al., 2024; Kaiser et al., 2025). These concerns are amplified by the "analytic flexibility" of silicon-sample pipelines: small choices in prompting, sampling, or scoring can substantially change whether a model appears aligned with human data, with no single configuration performing well across evaluation criteria (Cummins, 2025). Broader critiques emphasize that scale and fluent outputs do not guarantee representativeness, transparency, or safety, motivating caution when substituting synthetic respondents for humans (Bender et al., 2021).

Positioning of Our Work. Prior results demonstrate that prompting LLMs for synthetic conjoint choices can meaningfully reduce early regret when used as a warm-start prior, but they largely rely on an implicit alignment assumption between LLM-simulated and human preferences. At the same time, the broader "LLMs as silicon samples" literature reports mixed fidelity and substantial sensitivity to design choices, raising the concern that LLM-generated preference data may contain both unstructured mistakes and systematic deviations from the target population. We develop the noisy-CBLI framework to evaluate the robustness of LLM-generated priors for warm-starting under two types of corruption: random replacement (synonymous with uninformative feedback) and preference flipping (an example of a biased model). We separately formalize systematic misalignment through target shift and empirically delineate regimes in which warm-starting improves performance. Theoretically, we consolidate these effects through a prior-error term characterizing when warm-starting can outperform a cold-start, and we develop an alignment-based diagnostic that explains, post-hoc, the transition from beneficial to harmful initialization.

3 Methodology

In this section, we build on the CBLI framework to study its robustness under noisy and potentially misaligned synthetic priors. We first describe three real-world conjoint datasets used in our study, then formalize the contextual-bandit problem and recap the CBLI jump-start method. Finally, we introduce two noise-injection strategies—random response replacement and preference flipping—that systematically corrupt the synthetic priors, defining the noisy CBLI variants we evaluate under realistic noisy conditions.

3.1 Datasets

We use data collected from three conjoint surveys. In each, respondents' pre-treatment demographics (age, gender, income, ideology, etc.) are recorded, and choices between candidate profiles yield the reward signal for our bandit.

1. **COVID-19 vaccine conjoint (Kriner et al., 2020).** 1,970 American respondents completed a five-task choice-based conjoint survey in July 2020, comparing two hypothetical COVID-19 vaccines described by seven randomized attributes: efficacy, duration of protection, major side-effect rate, minor side effects, FDA approval status, country of origin, and endorser (Kreps et al., 2020). We flatten each respondent's demographic vector and the difference between the two vaccine attribute vectors into user–vaccine feature contexts for LinUCB.

2. **Immigration attitudes conjoint (Hainmueller, 2014).** 1,714 American adults each completed five pairwise choice tasks, selecting which of two hypothetical immigrant applicants they would admit (Hainmueller & Hopkins, 2015). Each immigrant profile was described by nine randomized attributes: education, profession, years of training/experience, reason for migrating, English-language ability, prior U.S. trips, legal entry status, country of origin, and the local industry's percent of foreign-born workers. As before, we concatenate one-hot demographics with the difference in attribute vectors to form user-choice features.

3. **Leisure travel conjoint (Miller, 2023).** In this dataset, roughly 2,100 American adults evaluated ten choice tasks, choosing between three U.S.-based leisure-travel destinations (Miller & Smith, 2024). Destinations are described by six randomized attributes: average July temperature, travel time, attractions, presidential election outcome of the state, recent news coverage, and community sentiments. We reduce each three-way decision to a binary comparison by randomly selecting one of the two unchosen destinations to compare against the chosen destination, resulting in $K = 2$ per round. We additionally evaluated the full three-arm setting ($K = 3$) and found that the regret curves differed by less than 3 percentage points. We flatten each respondent's demographics and these chosen-vs-unchosen attribute differences into user–destination feature vectors.

3.2 Problem Formulation

We set up the problem following the "jump-start" formalized by Alamdari et al. (2024). Each conjoint survey is cast as a *sleeping* contextual bandit (Kanade et al., 2009) over T rounds.

1. **Rounds & Arms.** At round $t \in \{1, \ldots, T\}$, a subset of arms $\mathcal{A}_t$ is presented (e.g., the two vaccines in Dataset 1).

2. **Context–Arm Features.** We embed each respondent's one-hot demographics u_t and the (chosen vs. unchosen) differences of arm attributes into a joint feature vector

$$x_{t,a} = \psi(u_t, a) \in \mathbb{R}^d.$$

3. **Linear Reward Model.** Following standard LinUCB assumptions (Li et al., 2010), we assume:

$$\mathbb{E}[r_t \mid x_{t,a}] = \theta_*^\top x_{t,a}, \quad \theta_* \in \mathbb{R}^d \text{ unknown.}$$

4. **Action Selection (LinUCB).** At each round, choose

$$a_t = \arg\max_{a \in \mathcal{A}_t} \Big[\hat{\theta}_{t-1}^\top x_{t,a} + \alpha \sqrt{x_{t,a}^\top A_{t-1}^{-1} x_{t,a}} \Big],$$

updating $A_t = A_{t-1} + x_{t,a_t} x_{t,a_t}^\top,\ b_t = b_{t-1} + r_t\, x_{t,a_t}$, as in Li et al. (2010); Chu et al. (2011), where the estimate used for action selection is $\hat{\theta}_{t-1} = A_{t-1}^{-1} b_{t-1}$.

5. **Regret.** At round t, let a_t be the arm chosen by LinUCB from among the action set $\mathcal{A}_t$ and

$$a_t^* \;=\; \arg\max_{a \in \mathcal{A}_t} r_t(a)$$

The arm with the highest realized reward among those available. The instantaneous (*sleeping bandit*) regret is the random variable

$$\Delta_t \;=\; r_t\big(a_t^*\big) \;-\; r_t(a_t) \;\in \{0,1\}.$$

The random cumulative regret after T rounds is

$$\widehat{R}(T) \;=\; \sum_{t=1}^{T} \Delta_t.$$

In experiments we plot or report one realization of $\widehat{R}(T)$: its trial-average over $G = 20$ independent seeds. For theoretical comparison we refer to the expected (pseudo-)regret

$$R(T) \;=\; \mathbb{E}\big[\widehat{R}(T)\big],$$

which is a scalar quantity bounded by $\widetilde{O}(\sqrt{T\,d})$ for LinUCB (Li et al., 2010).

6. **Ordinal Rewards.** In our experiments, both the LLM and the users in the conjoint studies are only ever asked to compare two arms in a round. As in (Alamdari et al., 2024), in terms of a contextual bandit model, this means we utilize a feature map $\psi(u, a_1, a_2)$ where u is a collection of user features, and a_1, a_2 are the two arms under consideration. This jointly encodes the user's context represented by u and the features of the arms and then learns a linear model on ψ. In our experiments, we take $\psi(u, a_1, a_2) = \text{flat}(u(f(a_1) - f(a_2))^T)$, where flat represents flattening a matrix of size $n \times m$ as a vector of size nm, and $f(a)$ are the representative features of each arm.

Note that if in Step 1, $\mathcal{A}_t$ contains all arms for all times t, then the problem is a classic linear contextual bandit and not a sleeping bandit.

3.3 CBLI "Jump-Start" Pipeline

We implement the "jump-start" pipeline introduced in Alamdari et al. (2024):

1. **Pre-training on LLM-Generated Priors.** Generate N synthetic context–reward pairs via LLM prompts. Each record is encoded as a feature vector and an observed binary preference in $\{0, 1\}$ where 1 indicates that the first arm in the chosen feature orientation is preferred. We fit LinUCB to these to obtain warm start parameters $\{A_a^{\text{pre}}, b_a^{\text{pre}}\}_{a=1}^{K}$.

2. **Warm-Start Fine-Tuning.** Initialize LinUCB with $\{A_a = A_a^{\text{pre}},\ b_a = b_a^{\text{pre}}\}$ and run for T rounds on the *real* conjoint data. At each round t, only the arms displayed in that task are active; select via the upper-confidence bound and update on the chosen arm.

3. **Cold-Start Baseline.** Repeat the T-round LinUCB procedure on the real data from scratch $(A_a = I_d,\ b_a = 0_d)$ under the same sleeping-bandit constraints to establish a regret baseline.

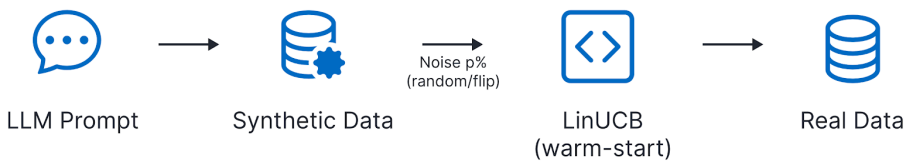

Figure 1: Overview of the CBLI evaluation framework (Noisy-CBLI). An LLM generates synthetic preference data, which is optionally corrupted with random or label-flipping noise at rate p, and used to warm-start a LinUCB bandit that is then fine-tuned on real user data.

3.4 Noise Injection Strategies

To evaluate CBLI's robustness when synthetic priors are imperfect, we corrupt the LLM-generated pre-training labels with two controlled noise schemes, yielding what we refer to as noisy CBLI variants (1). Let p denote the corruption rate (the proportion of synthetic samples to modify). In practical recommender systems, these two schemes correspond to two common failure modes: uninformative feedback and systematic bias. We model uninformative feedback via random response replacement, and systematic bias via preference flipping.

1. **Random Response Replacement.** We uniformly at random select a proportion p of the LLM-generated labels (each an arm index $a \in \{1, \ldots, K\}$) and overwrite each with a new arm drawn uniformly from $\{1, \ldots, K\}$. This simulates uninformative or arbitrary LLM mistakes. At $p = 0$ labels remain intact; at $p = 1$ the entire pre-training set is random.

2. **Preference Flipping.** We randomly choose a proportion p of the synthetic records and invert the original arm choice. For $K = 2$, flipping swaps "A" to "B" (and vice versa). For $K > 2$, we flip by cycling the chosen arm (e.g. $a \mapsto (a \bmod K) + 1$) or by selecting the least-preferred alternative. This introduces systematic bias that directly contradicts the LLM's own judgments. At $p = 1$, every label is inverted.

Once corrupted, each noisy variant (at each noise level p) replaces the original CBLI synthetic dataset. We then run the identical three-stage pipeline from Section 3.3 on every corrupted prior to measure the impact of noise on cumulative regret. In practical recommender systems, these noise models simulate common failure modes such as uninformative feedback and systematic bias, allowing practitioners to gauge how much imperfection in LLM-derived priors can be tolerated before online exploration must take precedence.

3.5 Experimental Protocol and Evaluation

All variants, cold-start LinUCB and CBLI warm-start under each noise scheme and level, are run for $G = 20$ independent trials. At each trial, we execute T rounds of LinUCB on the real conjoint data (Datasets 1-3) under the sleeping-bandit constraint $\mathcal{A}_t$.

Cumulative Regret. We measure performance by cumulative regret $R(T)$, taking the realized reward $r_t \in \{0, 1\}$ to be 1 when the bandit chooses the user's preferred arm. As discussed in Section 3.2, this quantity is assumed to be the realization of a Bernoulli random variable with a probability that is a linear function of the user's and arms' features. For each variant, we report the trial-average regret $\frac{1}{R}\sum_{i=1}^{R} R_i(T)$ and its 95% confidence interval.

Noise Sweep. For each injection strategy (random replacement, preference flipping) and corruption rate $p \in \{0.0, 0.1, \ldots, 0.7\}$, we pre-train LinUCB on the noisy synthetic priors and then fine-tune on the real data. We plot regret curves up to T for each p, comparing warm- vs. cold-start.

4 Theoretical Analysis

In this section we analyze the effect of noisy LLM-generated priors on the performance of LinUCB. We first formalize the warm-start prior induced by Noisy–CBLI, then derive a prior-centered confidence bound in which all pretraining effects enter through a single scalar $\mathcal{B}_0 := \|\theta_0 - \theta^\star\|_{A_0}$. We then make $\mathcal{B}_0$ explicit under preference-flipping noise and target misalignment, and use this to characterize when a warm-start is provably beneficial relative to cold-start.

4.1 Theoretical Problem Setup and Assumptions

We work in the sleeping linear contextual bandit setting described in Section 3.2. At round t an availability set $\mathcal{A}_t$ and feature vectors $\{x_{t,a}\}_{a\in\mathcal{A}_t}$ are revealed; the learner chooses $a_t \in \mathcal{A}_t$ and observes only the reward $r_t := r_t(a_t)$.

We impose the following standard assumptions:

- **Linear realizability.** There exists $\theta^\star \in \mathbb{R}^d$ such that $\mathbb{E}[r_t(a) \mid \mathcal{F}_{t-1}, x_{t,a}] = x_{t,a}^\top \theta^\star$ for all t and $a \in A_t$.
- **Bounded features.** $\|x_{t,a}\|_2 \le 1$ for all t and $a \in \mathcal{A}_t$ (without loss of generality after rescaling).
- **Sub-Gaussian rewards.** Let $\xi_t := r_t(a_t) - x_{t,a_t}^\top \theta^\star$. We assume a martingale-difference and conditional sub-Gaussian condition: $\mathbb{E}[\xi_t \mid \mathcal{F}_{t-1}, \{x_{t,a}\}_{a\in A_t}] = 0$ and $\mathbb{E}[\exp(\lambda\xi_t) \mid \mathcal{F}_{t-1}, \{x_{t,a}\}_{a\in A_t}] \le \exp(\lambda^2\sigma^2/2)$ for all $\lambda \in \mathbb{R}$, for some $\sigma > 0$.

Regret is measured against the best available arm at each round: $a_t^\star \in \arg\max_{a\in A_t} x_{t,a}^\top \theta^\star$, $r_t^\star := x_{t,a_t^\star}^\top \theta^\star$, and instantaneous regret $r_t^\star - x_{t,a_t}^\top \theta^\star$.

4.2 Noisy–CBLI Warm-Start Prior

The Noisy–CBLI warm-start uses an LLM to generate a synthetic conjoint dataset and fits a ridge regression prior to the resulting noisy labels.

Synthetic Responses and flip noise. Let $X \in \mathbb{R}^{n_s\times d}$ denote the synthetic design matrix and $y = X\theta^\star$ the corresponding "clean" mean success probabilities. Let $L \in \{0,1\}^{n_s}$ be Bernoulli labels with $\mathbb{E}[L \mid X] = y$. We inject preference-flipping noise by independently drawing $F_i \sim \text{Bernoulli}(p)$ for some $p \in [0,1]$ and define the the preference flipped response as

$$\tilde{L}_i := (1 - F_i)L_i + F_i(1 - L_i).$$

Then $\mathbb{E}[\tilde{L} \mid X] = (1-2p)\, y + p\,\mathbf{1}$, where $\mathbf{1} \in \mathbb{R}^{n_s}$ is the all-ones vector. We work with the regression proxy

$$\tilde{y} := (1-2p)\, X\theta^\star + p\,\mathbf{1} + \varepsilon, \tag{1}$$

where $\varepsilon := \tilde{L} - \mathbb{E}[\tilde{L} \mid X]$ has mean zero and conditionally σ_s^2-sub-Gaussian components by Hoeffding's lemma. All guarantees are stated for $p < \frac{1}{2}$; if an empirical flip rate $\hat{p} \ge \frac{1}{2}$ arises, one can recode labels via effective rate $p_{\text{eff}} := \min\{\hat{p}, 1-\hat{p}\}$ and apply the bounds with $p_{\text{eff}} < \frac{1}{2}$.

Ridge warm-start and prior error.

Given the synthetic design matrix X and the regression proxy $\tilde{y}$ in equation 1, we construct a ridge prior

$$A_0 := X^\top X + \tau_{\text{pre}} I, \qquad b_0 := X^\top \tilde{y}, \qquad \theta_0 := A_0^{-1} b_0. \tag{2}$$

We write $M := A_0^{-1} X^\top X$ for the corresponding shrinkage operator and define the prior mis-specification in the A_0-geometry as

$$\mathcal{B}_0 := \|\theta_0 - \theta^\star\|_{A_0} := \sqrt{(\theta_0 - \theta^\star)^\top A_0 (\theta_0 - \theta^\star)}.$$

At deployment time, the warm-started LinUCB algorithm initializes $V_0 = A_0$, $\hat{\theta}_0 = \theta_0$ and then updates on the real conjoint bandit stream. The cold-start baseline instead uses $A_0 = I$, $b_0 = 0$ (so $V_0 = I$, $\hat{\theta}_0 = 0$).

4.3 Prior-Centered Confidence Bounds

We first show that the estimation error of warm-started LinUCB admits a confidence bound that is centered at the ridge prior and depends on pretraining only through $\mathcal{B}_0$.

Let:

$$V_t := A_0 + \sum_{s \le t} x_{s,a_s} x_{s,a_s}^\top \quad (\text{so } V_t \succeq A_0)$$

Theorem 1 (Prior-centered confidence inequality)**.** *For any $\delta \in (0,1)$, with probability at least $1-\delta$ the warm-started ridge estimator satisfies, for all $t \ge 0$,*

$$\|\hat{\theta}_t - \theta^\star\|_{V_t} \le \beta_t(\delta) + \mathcal{B}_0, \tag{3}$$

where

$$\beta_t(\delta) := \sigma \sqrt{2 \log \frac{\det(V_t)^{1/2}}{\det(A_0)^{1/2}\delta}}$$

is the usual self-normalized variance term.

The proof, Given in Appendix A.1 follows the decomposition:

$$V_t(\hat{\theta}_t - \theta^\star) = A_0(\theta_0 - \theta^\star) + \sum_{s \le t} \xi_s x_{s,a_s},$$

bounding the first term by $\mathcal{B}_0$ in the A_0-norm and the second term by the self-normalized martingale inequality of Abbasi-Yadkori et al. (2011).

Theorem 1 induces a reward-confidence bound: for any context x,

$$|x^\top(\hat{\theta}_{t-1} - \theta^\star)| \le (\beta_{t-1}(\delta) + \mathcal{B}_0)\sqrt{x^\top V_{t-1}^{-1} x}, \tag{4}$$

so choosing

$$\alpha_t \ge \beta_{t-1}(\delta) + \mathcal{B}_0 \tag{5}$$

ensures that the LinUCB score is optimistic with high probability. Pretraining influences the UCB confidence radius primarily through the prior-error term $\mathcal{B}_0$, with an additional but comparatively mild logarithmic dependence on the design matrix A_0 inside $\beta_t(\delta)$. In practice, $\mathcal{B}_0$ is the dominant quantity governing whether warm-start improves or degrades regret.

4.4 Flip-Noise Bias and Misalignment

We next make $\mathcal{B}_0$ explicit under preference flips and target misalignment.

Substituting equation 1 into equation 2 gives

$$\theta_0 = A_0^{-1} X^\top \tilde{y} = (1-2p) A_0^{-1} X^\top X \theta^\star + p A_0^{-1} X^\top \mathbf{1} + A_0^{-1} X^\top \varepsilon \tag{6}$$

$$= (1-2p) M \theta^\star + p A_0^{-1} X^\top \mathbf{1} + A_0^{-1} X^\top \varepsilon, \tag{7}$$

so that

$$\theta_0 - \theta^\star = ((1-2p)M - I)\theta^\star + p A_0^{-1} X^\top \mathbf{1} + A_0^{-1} X^\top \varepsilon. \tag{8}$$

Taking the A_0-norm and expectation over the pretraining noise ε yields a bias–variance decomposition (Appendix A.2):

$$\mathbb{E}[\mathcal{B}_0^2] \le \|((1-2p)M - I)\theta^\star + p A_0^{-1} X^\top \mathbf{1}\|_{A_0}^2 + \sigma_s^2 \operatorname{tr}(X A_0^{-1} X^\top). \tag{9}$$

To interpret the flip-bias term, we diagonalize the synthetic design. Let $X^\top X = U\Lambda U^\top$ with eigenvalues λ_i and rotated parameter $\theta^\star = U\theta^\star_U$. Because A_0 and M share this eigenbasis,

$$A_0 = U(\Lambda + \tau_{\text{pre}} I)U^\top, \qquad M = U \operatorname{diag}\left(\frac{\lambda_i}{\lambda_i + \tau_{\text{pre}}}\right) U^\top.$$

This also gives the direction-wise form

$$\|((1-2p)M - I)\theta^\star\|_{A_0}^2 = \sum_{i=1}^{d} \frac{(\tau_{\text{pre}} + 2p\lambda_i)^2}{\lambda_i + \tau_{\text{pre}}} (\theta^\star_{U,i})^2. \tag{10}$$

In high-coverage directions where $\lambda_i \gg \tau_{\text{pre}}$, this simplifies to $\frac{(\tau_{\text{pre}}+2p\lambda_i)^2}{\lambda_i+\tau_{\text{pre}}} \approx 4p^2\lambda_i$, so

$$\mathbb{E}[\mathcal{B}_0^2] \approx 4p^2 \|(X^\top X)^{1/2}\theta^\star\|_2^2 + \sigma_s^2 \operatorname{tr}(X A_0^{-1} X^\top), \tag{11}$$

showing that flip-induced bias grows roughly linearly in p (in norm) and is amplified in directions with strong synthetic coverage.

To capture systematic misalignment between LLM-simulated and real preferences, we also consider a target shift

$$\theta^\star_{\text{syn}} = \theta^\star_{\text{real}} + \Delta,$$

where $\theta^\star_{\text{syn}}$ fits the synthetic labels and $\theta^\star_{\text{real}}$ fits the human data. Repeating the above with $\theta^\star_{\text{syn}}$ in place of $\theta^\star$ yields, at $p = 0$,

$$\theta_0 - \theta^\star_{\text{real}} = (M - I)\theta^\star_{\text{real}} + M\Delta, \tag{12}$$

so that a large misalignment vector Δ in well-covered directions can make $\mathcal{B}_0$ large even with no injected flip noise, leading to negative transfer at $p = 0$.

4.5 When Warm-Start is Beneficial

The prior-centered confidence inequality in Theorem 1 yields a direct comparison between warm-start and cold-start through the prior-error term $\mathcal{B}_0$.

Let $R_{\text{warm}}(T)$ and $R_{\text{cold}}(T)$ denote the regret upper bounds obtained by combining Theorem 1 with the standard LinUCB regret analysis for, respectively, a warm-start initialization (A_0, θ_0) and the cold-start initialization $(I, 0)$. Then, up to logarithmic factors,

$$R_{\text{warm}}(T) \lesssim \left(\beta_T^{\text{warm}}(\delta) + \mathcal{B}_0^{\text{warm}}\right)\sqrt{Td\log(\cdot)}, \tag{13}$$

$$R_{\text{cold}}(T) \lesssim \left(\beta_T^{\text{cold}}(\delta) + \mathcal{B}_0^{\text{cold}}\right)\sqrt{Td\log(\cdot)}, \tag{14}$$

where

$$\mathcal{B}_0^{\text{warm}} = \|\theta_0 - \theta^\star\|_{A_0}, \qquad \mathcal{B}_0^{\text{cold}} = \|\theta^\star\|_2.$$

Therefore, if

$$\mathcal{B}_0^{\text{warm}} < \mathcal{B}_0^{\text{cold}}, \tag{15}$$

and the corresponding variance terms $\beta_T^{\text{warm}}(\delta)$ and $\beta_T^{\text{cold}}(\delta)$ are of the same order, then the warm-start regret upper bound is smaller than the cold-start regret upper bound.

The upper bound equations 13 and 14 are a comparison of regret *upper bounds*, not a claim that realized warm-start regret must always be smaller than realized cold-start regret. It isolates the key quantity governing the comparison, the initialization error $\mathcal{B}_0$. This condition states that the bias introduced by the warm-start prior must remain small enough that the potential variance reduction associated with a larger initialization matrix A_0 is not outweighed by systematic prior error. Under the explicit flip-noise and target-misalignment models above, equations 9–11 imply that:

- on aligned tasks ($\Delta \approx 0$), $\mathcal{B}_0^{\text{warm}}$ increases with the flip probability p, implying a bound-level transition point, $p^\star$, at which point the warm-start upper bound can become worse than the cold-start upper bound.
- on misaligned tasks (large Δ), $\mathcal{B}_0^{\text{warm}}$ can exceed $\mathcal{B}_0^{\text{cold}}$ already at $p = 0$, explaining the empirically observed zero-noise warm-start failures.

Section 5.3 shows that the empirical results are broadly consistent with this theoretical picture at the regime level: smaller estimated prior error is generally associated with warm-start gains, while larger estimated prior error is associated with a greater risk of harmful warm-starting. This quantity is used as a post hoc explanatory diagnostic, computed using real conjoint data, rather than as a pre-deployment quantity available before online interaction.

5 Empirical Results & Discussion

5.1 Preference-Flipping Noise on the COVID-19 Vaccine Conjoint Dataset

Figure 2 plots the mean cumulative regret of LinUCB warm-started on GPT-4o priors corrupted by systematic preference flipping at seven noise levels ($p \in \{0.0, 0.1, \ldots, 0.7\}$), together with the uncorrupted baseline ("10k_base") and a cold-start baseline ("Not Pretrained"). Each curve is averaged over $G = 20$ trials, with shaded bands showing the 95% confidence interval. Table 1 reports the percentage reduction in cumulative regret relative to cold-start at horizon T for different synthetic pre-training sizes and flipping rates.

- **Zero noise ($p = 0$).** With uncorrupted GPT-4o labels, warm-start achieves the lowest regret, quickly approaching optimal arm selection and substantially outperforming cold-start. Across all N, Table 1 shows a consistent positive reduction in regret, with larger synthetic datasets mainly tightening confidence intervals and yielding modest additional gains.
- **Low to moderate noise (10–30 percent).** For small to moderate flipping rates, the warm-start curves remain below the cold-start baseline, and the corresponding entries in Table 1 remain positive. Preference-flipping at these levels shifts the regret curves upward but does not eliminate the advantage of pre-training: CBLI still converges faster than cold-start, especially for larger N.
- **High noise (40–70 percent).** Once flipping reaches higher levels, the benefit of pre-training disappears and eventually reverses. Around $p \approx 0.4$, the results enter a transition region. Warm-start and cold-start curves become much closer, with several confidence intervals approaching or overlapping zero. At $p \geq 0.5$, corrupted priors begin to harm performance: warm-started LinUCB exhibits higher regret than cold-start throughout much of the horizon, with the effect most pronounced for larger synthetic datasets where the mis-specified prior is more strongly enforced.

Table 1: Percentage reduction in cumulative regret (%Δ Regret) for the COVID-19 Vaccine dataset using GPT-4o priors. Reported across three synthetic dataset sizes (N). Mean over $G = 20$ seeds $\pm$ 95% CI.

	Pre-training Size (N)		
Noise (p)	**1k**	**3k**	**10k**
0%	7.53 ± 0.83	10.47 ± 0.87	11.16 ± 0.64
10%	6.48 ± 0.81	10.17 ± 0.98	10.25 ± 0.71
20%	5.61 ± 1.09	8.48 ± 0.84	9.74 ± 0.67
30%	4.10 ± 1.10	6.17 ± 1.06	7.23 ± 0.74
40%	2.24 ± 1.49	3.00 ± 1.62	3.07 ± 1.34
50%	-0.36 ± 1.48	-5.12 ± 2.08	-6.10 ± 3.33
60%	-2.79 ± 2.02	-11.66 ± 2.36	-20.91 ± 3.62
70%	-8.67 ± 1.95	-18.64 ± 2.99	-32.29 ± 1.45

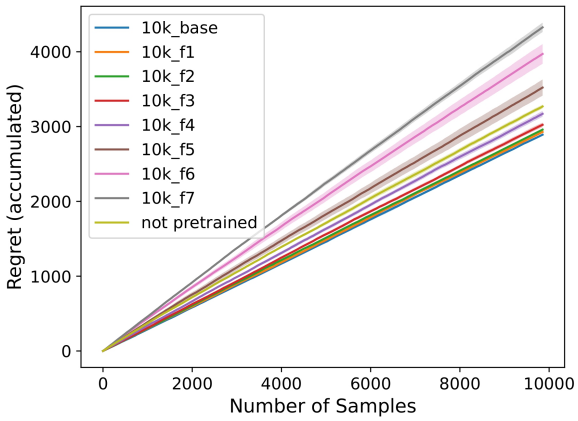

Figure 2: Cumulative regret on the COVID-19 Vaccine dataset under preference-flipping noise. "10k_fX" indicates X times 10% of LLM-generated labels flipped. Shaded regions are 95 percent CI over $G = 20$ runs.

5.2 Random-Response Noise on the COVID-19 Vaccine Conjoint Dataset

We now consider the effects of random noise on synthetic labels. Figure 3 plots the mean cumulative regret of LinUCB warm-started on GPT-4o priors corrupted by random responses at the same noise levels as in Figure 2, together with the cold-start baseline. As before, each curve is averaged over $G = 20$ trials with 95% confidence intervals.

- **Zero noise ($p = 0$).** Without corruption, warm-start again yields the lowest regret, converging rapidly toward optimal recommendations and reproducing the benefits observed in the preference-flipping setting.
- **Low to moderate noise (10–30 percent).** At 10–30 percent random replacements, the warm-start curves shift upward slightly but remain clearly below the cold-start baseline. CBLI continues to reduce cumulative regret relative to cold-start, indicating that a substantial fraction of uninformative labels can be tolerated without losing the gains from pre-training.
- **Moderate to high noise (40–70 percent).** For higher random-response rates, the warm-start curves gradually approach the cold-start curve. Unlike the preference-flipping case, however, we do not observe a regime where random corruption yields consistently higher regret than cold-start. Even at the largest tested p, the warm-start performance is at worst comparable to cold-start and often remains slightly better.

5.3 Noise Effects Across Datasets and Models

We next assess whether the preference-flipping robustness patterns observed in the vaccine domain generalize across tasks and LLMs. For each dataset–model combination, we sweep preference-flipping corruption and compare warm-start to cold-start regret across synthetic-set sizes $N \in \{1\text{k}, 3\text{k}, 10\text{k}\}$. Results are reported in Table 2.

The COVID-19 dataset provides the clearest evidence of LLM–real user alignment. Across all pretraining sizes, with no injected noise, all four models improve over cold-start, with GPT-4o and Llama 3.1 producing the largest gains, GPT-3.5 Turbo somewhat smaller gains, and Qwen 3 the weakest gains. Under moderate flipping ($p = 0.3$), most COVID-19 warm starts remain beneficial, although the model ordering changes: GPT-3.5 Turbo remains especially strong at larger N, while Qwen 3 becomes harmful at $N = 10\text{k}$. At $p = 0.5$, the setting moves into a transition or harmful regime. Several small-prior settings are near zero,

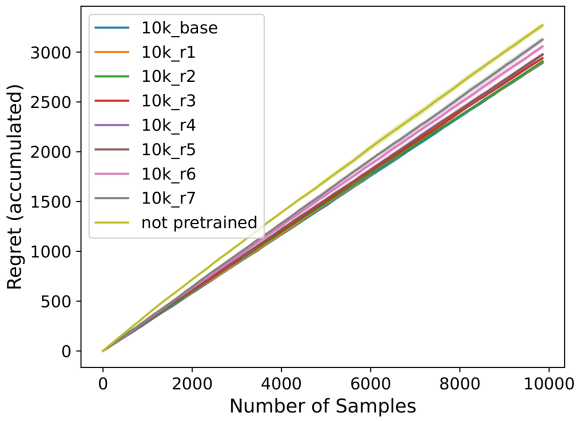

Figure 3: Cumulative regret on the COVID-19 Vaccine dataset under random-response noise.

but larger corrupted priors generally hurt performance, with the strongest negative transfer appearing for Qwen 3 and GPT-3.5 Turbo at $N = 10$k.

The Immigration dataset reflects a weaker baseline-alignment regime. At $p = 0$, GPT-4o gives small positive gains across all N, but the effects are much smaller than in COVID-19. GPT-3.5 Turbo, Llama 3.1, and Qwen 3 generally weaken as N increases, and several confidence intervals overlap zero. This suggests that the LLM priors are not uniformly aligned with real Immigration choices, even before injected corruption. Under preference flipping, performance generally deteriorates across models, especially at larger N. The dominant issue is not only robustness to injected noise, but the weaker alignment of the uncorrupted synthetic prior itself.

The Travel dataset is strongly model-dependent. GPT-4o appears comparatively aligned: it improves over cold-start at $p = 0$ for all N and remains beneficial at larger N under moderate flipping. GPT-3.5 Turbo also improves at $p = 0$, but becomes harmful once preferences are flipped. In contrast, Llama 3.1 and Qwen 3 perform worse than cold-start across all reported Travel settings, including $p = 0$. This shows that baseline alignment is not only dataset-specific but also model-specific: the same conjoint task can yield useful synthetic priors from one LLM and harmful priors from another.

Overall, Table 2 supports the view that CBLI performance is primarily governed by the alignment of the uncorrupted LLM prior, with preference-flipping noise acting as an additional stressor. When the baseline prior is aligned, as in COVID-19, LinUCB can tolerate moderate corruption before the warm-start becomes harmful. When the baseline prior is weak or misaligned, as in much of Immigration and in several Travel model combinations, increasing N or injecting directional noise can amplify negative transfer. In additional random-response sweeps, performance decays gradually toward cold-start as corruption increases, rather than crossing into a systematically harmful regime. We treat this as an empirical robustness observation only; our theory does not provide a formal guarantee for random-response corruption and instead focuses on directional corruption and target misalignment.

We additionally note that these results may also depend on the model checkpoint used to generate the synthetic priors. All OpenAI results in Table 2 use the Sept.–Oct. 2025 access window (Table 5). In preliminary comparisons with an earlier GPT-3.5 Turbo snapshot, we observed larger warm-start gains and cleaner corruption breakpoints. We therefore treat model version as another source of variation in synthetic-prior alignment and report the earlier snapshot results in Appendix Table 7. Similarly, we note that LLMs are susceptible to choice ordering. We observe a difference in behavior when Option A is in position B and vice versa, with the agreement in Llama 3.1 and Qwen 3 models ranging between 53% to 96%. We share these results in Table 8.

Table 2: Percentage reduction in cumulative regret (%Δ Regret) compared to cold-start under preference flipping. Reported across 4 models, 3 datasets, and 3 representative noise levels ($p \in \{0, 0.3, 0.5\}$). Mean over $G = 20$ seeds.

Model	Dataset	Noise (%)	N = 1k	N = 3k	N = 10k
GPT-3.5 Turbo	COVID-19	0	6.04 ± 0.88	7.90 ± 0.70	8.29 ± 0.77
		30	3.97 ± 1.22	8.22 ± 0.54	8.60 ± 0.45
		50	-0.02 ± 1.34	-3.41 ± 0.86	-7.72 ± 1.11
	Immigration	0	0.67 ± 0.50	-0.33 ± 0.73	-0.93 ± 1.39
		30	0.43 ± 1.44	-2.47 ± 2.17	-3.14 ± 1.70
		50	-1.22 ± 1.10	-6.15 ± 2.88	-5.13 ± 3.90
	Travel	0	2.33 ± 0.50	1.10 ± 0.62	2.49 ± 0.45
		30	-2.40 ± 0.66	-1.53 ± 0.51	-1.65 ± 0.51
		50	-2.82 ± 0.53	-1.99 ± 0.39	-2.84 ± 0.90
GPT-4o	COVID-19	0	7.53 ± 0.83	10.47 ± 0.87	11.16 ± 0.64
		30	4.10 ± 1.10	6.17 ± 1.06	7.23 ± 0.74
		50	-0.36 ± 1.48	-5.12 ± 2.08	-6.10 ± 3.33
	Immigration	0	1.74 ± 0.82	0.88 ± 0.67	0.51 ± 1.12
		30	0.99 ± 1.19	-0.42 ± 1.13	-3.15 ± 1.37
		50	-3.34 ± 1.25	-4.00 ± 2.43	-5.39 ± 2.52
	Travel	0	2.54 ± 0.34	1.73 ± 0.31	3.31 ± 0.54
		30	-1.75 ± 0.54	1.33 ± 0.62	2.33 ± 0.55
		50	-1.44 ± 0.77	1.46 ± 0.46	0.48 ± 0.82
Llama 3.1	COVID-19	0	5.84 ± 0.77	10.31 ± 0.83	10.66 ± 0.67
		30	4.26 ± 1.33	6.88 ± 1.01	6.46 ± 1.30
		50	-0.21 ± 1.72	-1.43 ± 1.66	-4.13 ± 2.07
	Immigration	0	0.03 ± 1.15	-1.52 ± 0.95	-2.14 ± 1.13
		30	-0.53 ± 1.69	-4.68 ± 2.30	-5.17 ± 1.27
		50	-1.84 ± 1.39	-6.25 ± 2.70	-6.19 ± 3.55
	Travel	0	-1.90 ± 0.82	-2.53 ± 1.06	-1.21 ± 0.73
		30	-2.96 ± 0.94	-3.31 ± 1.22	-3.94 ± 1.37
		50	-4.06 ± 1.33	-3.81 ± 1.09	-7.36 ± 1.21
Qwen 3	COVID-19	0	2.87 ± 1.07	6.05 ± 0.99	7.18 ± 0.74
		30	0.84 ± 1.50	2.06 ± 1.78	-1.61 ± 1.23
		50	-2.85 ± 1.79	-2.37 ± 1.30	-10.89 ± 2.14
	Immigration	0	1.71 ± 1.18	0.11 ± 0.81	-1.13 ± 1.15
		30	-0.28 ± 0.95	-3.44 ± 2.03	-3.58 ± 2.05
		50	-2.45 ± 1.25	-1.76 ± 2.62	-6.02 ± 3.04
	Travel	0	-1.74 ± 0.60	-2.56 ± 0.86	-1.46 ± 0.95
		30	-1.67 ± 0.85	-2.69 ± 1.33	-1.68 ± 1.17
		50	-1.81 ± 0.85	-1.80 ± 0.78	-1.66 ± 0.95

5.4 Misalignment Analysis

Our theory identifies the prior-error term

$$\mathcal{B}_0 \;=\; \left\|\theta_0 - \theta^\star\right\|_{A_0}, \qquad A_0 = X_{\text{syn}}^\top X_{\text{syn}} + \tau I,$$

as the central quantity governing whether an LLM-generated warm start helps or harms LinUCB. In the prior-centered confidence bound of Section 4, the exploration radius satisfies

$$\|\hat{\theta}_t - \theta^\star\|_{V_t} \;\leq\; \beta_t(\delta) + \mathcal{B}_0,$$

Table 3: Post hoc estimated prior error $\widehat{\mathcal{B}}_0 = \|\theta_0 - \theta_{\text{real}}\|_{A_0}$ for each model and dataset. Intervals are 95% bootstrap confidence intervals over real-data resamples.

Model	**COVID-19**	**Immigration**	**Travel**
GPT-3.5 Turbo	51.0 ± 1.81	57.1 ± 1.79	28.1 ± 1.87
GPT-4o	44.4 ± 1.91	56.2 ± 1.76	26.9 ± 1.83
Llama 3.1	46.2 ± 1.90	59.2 ± 1.70	28.2 ± 1.79
Qwen 3	62.3 ± 1.86	56.9 ± 1.73	28.9 ± 1.73

where $\beta_t(\delta)$ depends on A_0 only through a mild logarithmic term, while $\mathcal{B}_0$ enters additively and therefore controls the bias introduced by the warm-start effect. To connect this analysis to practice, we construct a post hoc alignment diagnostic motivated by $\mathcal{B}_0$. We estimate θ_0 by ridge regression on the synthetic LLM responses. Since $\theta^\star$ is unobserved, we approximate the real preference direction by fitting ridge regression on the real conjoint responses, yielding θ_{real}. The warm-start design matrix is

$$A_0 = X_{\text{syn}}^\top X_{\text{syn}} + \tau I.$$

We define, in the shared feature space used by the bandit,

$$\widehat{\mathcal{B}}_0 = \left\|\theta_0 - \theta_{\text{real}}\right\|_{A_0}.$$

This quantity measures how far the LLM-derived prior lies from the real human-preference fit in the same warm-start geometry that appears in the theoretical prior-error term. To quantify uncertainty, we report 95% bootstrap confidence intervals obtained by resampling the real conjoint dataset 1,000 times, refitting θ_{real}, and recomputing $\widehat{\mathcal{B}}_0$ while holding the synthetic prior θ_0 and design matrix A_0 fixed.

Table 3 reports $\widehat{\mathcal{B}}_0$ for each model–dataset pair. The values support the regime-level interpretation suggested by the theory: model–dataset pairs with smaller estimated prior error generally exhibit stronger warm-start behavior, while larger estimated prior error is associated with weaker gains or negative transfer. We interpret this relationship at the regime level rather than as a calibrated prediction of exact regret.

On COVID-19, the ordering is clearest: GPT-4o has the smallest estimated prior error ($\widehat{\mathcal{B}}_0 \approx 44.4$) and the largest clean-prior regret reduction at $N = 10$k ($\approx 11.2\%$), followed closely by Llama 3.1 ($\widehat{\mathcal{B}}_0 \approx 46.2$, $\approx 10.7\%$). GPT-3.5 Turbo has a larger prior error (≈ 51.0) and a smaller gain ($\approx 8.3\%$), while Qwen 3 has the largest prior error (≈ 62.3) and the weakest clean-prior gain ($\approx 7.2\%$). This supports the interpretation of COVID-19 as an aligned-prior regime.

On Immigration, the estimated prior errors are closer together, and the empirical gains are correspondingly weaker. GPT-4o has the smallest estimated prior error (≈ 56.2) and is the only model with positive clean-prior performance across all reported N, although the $N = 10$k gain is small ($\approx 0.5\%$). Llama 3.1 has the largest estimated prior error (≈ 59.2) and shows negative transfer at larger N. GPT-3.5 Turbo and Qwen 3 fall between these cases, with small or negative effects as pretraining size increases. Thus, Immigration is better interpreted as a weak or partially misaligned baseline regime rather than a robust warm-start success.

Travel illustrates both the usefulness and the limitations of the scalar diagnostic. GPT-4o again has the smallest estimated prior error ($\widehat{\mathcal{B}}_0 \approx 26.9$) and the strongest clean-prior gains at $N = 10$k ($\approx 3.3\%$). Qwen 3 has the largest estimated prior error (≈ 28.9) and is harmful across all reported Travel settings. However, GPT-3.5 Turbo and Llama 3.1 have nearly identical estimated prior errors (28.1 and 28.2) while producing different clean-prior behavior. GPT-3.5 Turbo improves over cold-start, whereas Llama 3.1 underperforms. This indicates that $\widehat{\mathcal{B}}_0$ is a coarse, direction-aggregated diagnostic. Realized regret can also depend on how prior error projects onto the particular comparison contexts and margins encountered during fine-tuning.

The absolute scale of $\widehat{\mathcal{B}}_0$ varies across datasets, reflecting differences in the synthetic design X_{syn} and the A_0-geometry. The values should therefore be compared primarily within a dataset, not across datasets. Moreover, our sufficient condition in Section 4.5 is stated in terms of the theoretical prior-error term, whereas Table 3 reports a retrospective empirical analogue computed using real conjoint data. $\widehat{\mathcal{B}}_0$ is best viewed as a

Table 4: Effect of prior downweighting on COVID-19 GPT-4o warm starts. Values are the percent reduction in cumulative regret (%Δ Regret) relative to the cold start. $\lambda = 1$ is fully CBLI warm starting while $\lambda = 0$ is cold-started LinUCB.

Synthetic prior	$\lambda = 1.0$	$\lambda = 0.75$	$\lambda = 0.5$	$\lambda = 0.25$
10k_base	11.16 ± 0.64	10.99 ± 0.63	11.01 ± 0.63	11.04 ± 0.63
10k_f5	-6.10 ± 3.33	-5.27 ± 2.88	-4.16 ± 2.27	-2.24 ± 1.23
10k_f7	-32.29 ± 1.45	-30.89 ± 1.39	-28.62 ± 1.29	-24.45 ± 1.10

post hoc explanatory diagnostic rather than a pre-deployment test. Taken together, the results support the central mechanism of the paper: LLM priors help when they are close to the real preference parameter in the relevant geometry, and they become weak or harmful when this alignment fails. At the same time, the Travel GPT-3.5/Llama comparison shows that this scalar diagnostic is not a complete predictor of realized regret, since direction, context coverage, and comparison margins can also affect the online trajectory.

5.5 Prior Downweighting as a Lightweight Safeguard

The preceding results and theoretical grounding show how and why LLM warm starts can become harmful if the LLM is misaligned or once synthetic priors are sufficiently directionally corrupted. Existing work offers several related responses: robust warm-starting can reweight auxiliary and bandit feedback under source mismatch, transfer-bandit methods characterize when source information helps or hurts, and conservative-bandit methods maintain performance relative to a trusted baseline (Zhang et al., 2019; Cai et al., 2024; Kazerouni et al., 2017). Our goal here is more limited. We do not develop a new robust warm-starting algorithm. Instead, we test a minimal safeguard that can be applied directly to the CBLI/LinUCB initialization when a prior mismatch is suspected.

After pretraining, we downweight the synthetic prior by interpolating the warm-start sufficient statistics with the cold-start initialization:

$$A_0(\lambda) = (1 - \lambda)I + \lambda A_{\text{pre}}, \qquad b_0(\lambda) = \lambda b_{\text{pre}}, \qquad \lambda \in [0, 1].$$

Here, $\lambda = 0$ recovers the cold-start LinUCB while $\lambda = 1$ uses the full CBLI warm start. The LLM, prompts, dataset, and noise injection methods are unchanged. We report this safeguard on the COVID-19 setting with GPT-4o priors. This is the cleanest diagnostic setting for prior downweighting because the uncorrupted prior gives a clear positive warm-start gain, while preference flipping produces a clear transition to negative transfer. We test whether downweighting preserves value when the prior is aligned, and reduces harm when the prior is directionally corrupted. On Immigration and Travel, many baseline effects are small or overlap zero, making the effect of prior downweighting harder to distinguish from ordinary run-to-run variation. We consider 3 priors: the clean 10k GPT-4o prior, the same prior with 50% preference flips, and the same prior with 70% preference flips. For each prior, we sweep $\lambda \in \{1.0, 0.75, 0.5, 0.25\}$ and report percentage reduction in cumulative regret relative to cold-start LinUCB at $T = 9{,}855$, averaged over 20 seeds.

Table 4 shows the effects of prior downweighting as a simple safeguard. On the clean GPT-4o prior, all tested values of λ retain approximately the same regret reduction, suggesting that moderate downweighting has little cost in this setting. Under 50% preference flipping, downweighting substantially mitigates negative transfer, reducing λ from 1.0 to 0.25 improves performance from -6.10% to -2.24%, recovering about 3.9 percentage points relative to full CBLI. Under 70% flipping, downweighting still reduces harm, improving from -32.29% to -24.45%, but does not recover cold-start performance. These results suggest simple prior shrinkage can mitigate moderate mismatch, while severe misalignment requires either stronger rollback to cold start or adaptive detection and weighting.

6 Conclusion

We examined how noise and underlying preference mismatch affect the usefulness of LLM-generated priors for warm-starting contextual bandits. Across three conjoint datasets and multiple LLMs, we found that warm-start reduces regret only when synthetic preferences track human choices closely. In these aligned settings, random-response corruption is empirically milder, and warm-start remains beneficial under moderate preference-flipping noise before losing its advantage at higher corruption levels. In misaligned settings, warm-start can underperform cold-start even at $p = 0$, with additional corruption further degrading performance.

To explain these observations, we developed a prior-centered analysis in which pretraining affects regret through a single prior-error term, $\mathcal{B}_0 = \|\theta_0 - \theta^\star\|_{A_0}$, and derived sufficient conditions under which the warm-start bound is tighter than the cold-start bound. Empirically, transitions between helpful and harmful behavior align with changes in this prior error: random-response corruption mainly attenuates the informativeness of the synthetic prior, whereas preference-flipping introduces directional bias that can rapidly enlarge effective prior error. Moreover, our post-hoc estimates $\widehat{\mathcal{B}}_0$ support the same regime-level picture: smaller estimated prior error is generally associated with larger regret reductions, while larger prior error corresponds to marginal gains or negative transfer. These findings suggest that LLM-generated priors are most valuable when alignment is high and should be deployed cautiously when synthetic and real preferences may diverge; simple safeguards such as prior downweighting can mitigate moderate mismatch, but do not replace adaptive detection or robust source weighting.

7 Limitations

Our work has provided a systematic analysis of the use of synthetic LLM priors for bandit recommender systems; however, limitations remain. The warm-start procedure depends on a fixed set of prompts yet LLM outputs are highly prompt-sensitive (Pezeshkpour & Hruschka, 2023), so the evaluation may provide an unduly narrow estimate of variance in the prior (Sclar et al., 2024; Errica et al., 2025). We also observe this phenomenon in Table 8, though the differing effects sizes by model may indicate taht these structural sensitivities will decrease as larger and better models are trained. The injected noise follows an independent and identically distributed random-replacement or label-flip model, whereas empirical LLM errors exhibit heteroskedastic and context-correlated structure. Consequently, the corruption sweep may mischaracterize real-world error modes, and future work should consider context-dependent or structured noise (Xia et al., 2020). Our theoretical analysis relies on a high-coverage approximation ($\lambda_i \gg \tau_{\text{pre}}$) and provides an upper-bound-based sufficient condition rather than a matching lower bound. In sparse regimes, the bias-variance trade-off may deviate from our quadratic scaling, and a lower bound is needed to prove failure tolerances more rigorously. Commercial LLMs are governed by evolving safety guardrails that can refuse or reshape responses about sensitive content, altering the effective reward distribution and introducing non-stationarity that violates standard regret assumptions (Pantha et al., 2024). More broadly, model revisions can shift synthetic preference distributions over time: for GPT-3.5 Turbo, we observe materially stronger and cleaner warm-start gains from an earlier snapshot than from the Sept–Oct 2025 access window used for the main results (Appendix B; Table 7). Lastly, LLMs encode demographic and ideological biases from their training data. When such biases manifest in synthetic preferences (Wyllie et al., 2024), they are inherited by the bandit and can persist downstream. These biases may not be immediately observable, so despite potential early-stage regret gains, fairness auditing and bias mitigation remain essential challenges (Gallegos et al., 2024).

8 Future Works

Future work should seek to derive lower regret bounds that formally characterize the "tipping point" trends recorded in our experiments. On the practical side, we envision a lightweight alignment estimator acting as a statistical pre-check to flag potentially harmful priors before they are deployed. Finally, extending the Noisy-CBLI framework beyond linear assumptions to neural bandits would allow for more sophisticated modeling of the context-dependent noise often seen in real-world LLM outputs.

9 Broader Impact Statement

This work investigates the reliability of using Large Language Models to initialize recommender systems. While our primary contribution is technical, establishing robustness thresholds for synthetic priors, we identify several ethical implications regarding the deployment of this technology.

The most significant risk in Noisy-CBLI frameworks is the potential for feedback loops where LLM-encoded biases are transferred to the bandit policy. As noted in our experiments with the Immigration and Vaccine datasets, LLM priors can be opinionated. If a synthetic prior contains demographic or ideological biases (e.g., favoring specific groups in the Immigration task), the warm-started bandit will operationalize this discrimination immediately upon deployment, potentially disadvantaging real users or items before human feedback can correct the policy. Practitioners must audit synthetic priors for fairness, not just regret minimization, before initialization.

Our evaluation includes high-stakes domains, such as public health (COVID-19 vaccination). Deploying warm-started bandits in such settings carries the risk of amplifying hallucinations or medical misinformation inherent in the LLM. Our theoretical analysis provides a safeguard against this by defining a breakdown threshold, offering practitioners principled guidelines detailing when to reject synthetic priors that do not meet strict alignment standards, thereby preventing the deployment of unreliable systems in critical contexts.

This research involved substantial computational resources for generating synthetic data and simulating bandit trajectories across multiple models (Llama, Qwen, GPT families). While the immediate cost is non-negligible, our findings suggest that LLM initialization is detrimental in high-noise or misaligned settings. This insight potentially reduces long-term environmental impact by discouraging the wasteful deployment of generative models in domains where they offer no performance benefit over cold-start algorithms.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/e1d5be1c7f2f456670de3d53c7b54f4a-Paper.pdf.

Parand A. Alamdari, Yanshuai Cao, and Kevin H. Wilson. Jump starting bandits with LLM-generated prior knowledge. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19821–19833, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1107. URL https://aclanthology.org/2024.emnlp-main.1107.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023. doi: 10.1017/pan.2023.2.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 05 2002. doi: 10.1023/A:1013689704352.

Ali Baheri and Cecilia O. Alm. Llms-augmented contextual bandit, 2023. URL https://arxiv.org/abs/2311.02268.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared

Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022. URL https://arxiv.org/abs/2212.08073.

Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. Tallrec: An effective and efficient tuning framework to align large language model with recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, RecSys '23, pp. 1007–1014. ACM, September 2023. doi: 10.1145/3604915.3608857. URL http://dx.doi.org/10.1145/3604915.3608857.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Changxiao Cai, T. Tony Cai, and Hongzhe Li. Transfer learning for contextual multi-armed bandits, 2024. URL https://arxiv.org/abs/2211.12612.

Wei Chu, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, JMLR Workshop and Conference Proceedings, pp. 208–214, 2011.

Jamie Cummins. The threat of analytic flexibility in using large language models to simulate human data: A call to attention, 2025. URL https://arxiv.org/abs/2509.13397.

Mingwei Deng, Ville Kyrki, and Dominik Baumann. Transfer learning in latent contextual bandits with covariate shift through causal transportability, 2025. URL https://arxiv.org/abs/2502.20153.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. Questioning the survey responses of large language models, 2024. URL https://arxiv.org/abs/2306.07951.

Federico Errica, Giuseppe Siracusano, Davide Sanvito, and Roberto Bifulco. What did i do wrong? quantifying llms' sensitivity and consistency to prompt engineering, 2025. URL https://arxiv.org/abs/2406.12334.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and fairness in large language models: A survey, 2024. URL https://arxiv.org/abs/2309.00770.

Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. Chat-rec: Towards interactive and explainable llms-augmented recommender system, 2023. URL https://arxiv.org/abs/2303.14524.

Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5), 2023. URL https://arxiv.org/abs/2203.13366.

Jens Hainmueller. Replication data for: The Hidden American Immigration Consensus: A Conjoint Analysis of Attitudes toward Immigrants, 2014. URL https://doi.org/10.7910/DVN/25505.

Jens Hainmueller and Daniel J. Hopkins. The hidden american immigration consensus: A conjoint analysis of attitudes toward immigrants. *American Journal of Political Science*, 59(3):529–548, 2015. doi: 10.2139/ssrn.2106116. Available at SSRN: https://ssrn.com/abstract=2106116.

Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks, 2016. URL https://arxiv.org/abs/1511.06939.

Joey Hong, Branislav Kveton, Manzil Zaheer, and Mohammad Ghavamzadeh. Hierarchical bayesian bandits, 2022. URL https://arxiv.org/abs/2111.06929.

Carolin Kaiser, Jakob Kaiser, Vladimir Manewitsch, Lea Rau, and Rene Schallner. Simulating human opinions with large language models: Opportunities and challenges for personalized survey data modeling. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*, UMAP Adjunct '25, pp. 82–86, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713996. doi: 10.1145/3708319.3733685. URL https://doi.org/10.1145/3708319.3733685.

Varun Kanade, H Brendan McMahan, and Brent Bryan. Sleeping experts and bandits with stochastic action availability and adversarial rewards. In *Artificial Intelligence and Statistics*, pp. 272–279. PMLR, 2009.

Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi-Yadkori, and Benjamin Van Roy. Conservative contextual linear bandits, 2017. URL https://arxiv.org/abs/1611.06426.

Michal Kosinski. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45), October 2024. ISSN 1091-6490. doi: 10.1073/pnas.2405460121. URL http://dx.doi.org/10.1073/pnas.2405460121.

Sarah Kreps, Sandip Prasad, John S. Brownstein, Yulin Hswen, Brian T. Garibaldi, Baobao Zhang, and Douglas L. Kriner. Factors associated with us adults' likelihood of accepting covid-19 vaccination. *JAMA Network Open*, 3(10):e2025594–e2025594, 10 2020. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2020.25594. URL https://doi.org/10.1001/jamanetworkopen.2020.25594.

Douglas Kriner, Sarah Kreps, John S Brownstein, Yulin Hswen, Baobao Zhang, and Sandip Prasad. Replication Data for: Factors Associated With US Adults' Likelihood of Accepting COVID-19 Vaccination: Evidence From a Survey and Choice-Based Conjoint Analysis, 2020. URL https://doi.org/10.7910/DVN/6BSJYP.

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

David Miller. Replication Data for: (Small D-Democratic) Vacation, All I Ever Wanted?: The Effect of Democratic Backsliding on Leisure Travel in the American States, 2023. URL https://doi.org/10.7910/DVN/KA7DLE.

David Miller and Serena Smith. (small d-democratic) vacation, all i ever wanted? the effect of democratic backsliding on leisure travel in the american states. *Journal of Experimental Political Science*, 12:1–11, 04 2024. doi: 10.1017/XPS.2023.40.

OpenAI. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Umberto Panniello, Alexander Tuzhilin, Michele Gorgoglione, Cosimo Palmisano, and Anto Pedone. Experimental comparison of pre-vs. post-filtering approaches in context-aware recommender systems. In *Proceedings of the third ACM conference on Recommender systems*, pp. 265–268, 2009.

Nishan Pantha, Muthukumaran Ramasubramanian, Iksha Gurung, Manil Maskey, and Rahul Ramachandran. Challenges in guardrailing large language models for science, 2024. URL https://arxiv.org/abs/2411.08181.

Aleksandr V. Petrov and Craig Macdonald. Generative sequential recommendation with gptrec, 2023. URL https://arxiv.org/abs/2306.11114.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions, 2023. URL https://arxiv.org/abs/2308.11483.

Massimo Quadrana, Paolo Cremonesi, and Dietmar Jannach. Sequence-aware recommender systems, 2018. URL https://arxiv.org/abs/1802.08452.

Marko Sarstedt, Susanne J. Adler, Lea Rau, and Bernd Schmitt. Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines. *Psychology & Marketing*, 41(6):1254–1270, 2024. doi: https://doi.org/10.1002/mar.21982. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/mar.21982.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting, 2024. URL https://arxiv.org/abs/2310.11324.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2 edition, 2025. URL https://www.math.uci.edu/~rvershyn/. Second Edition, pre-publication version, November 28, 2025.

Anna Volodkevich, Danil Gusak, Anton Klenitskiy, and Alexey Vasilev. Autoregressive generation strategies for top-k sequential recommendations, 2024. URL https://arxiv.org/abs/2409.17730.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020. URL https://arxiv.org/abs/1910.03771.

Sierra Wyllie, Ilia Shumailov, and Nicolas Papernot. Fairness feedback loops: Training on synthetic data amplifies bias, 2024. URL https://arxiv.org/abs/2403.07857.

Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise, 2020. URL https://arxiv.org/abs/2006.07836.

Chicheng Zhang, Alekh Agarwal, Hal Daumé Iii, John Langford, and Sahand Negahban. Warm-starting contextual bandits: Robustly combining supervised and bandit feedback. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7335–7344. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/zhang19b.html.

Danyang Zhang, Lu Chen, Situo Zhang, Hongshen Xu, Zihan Zhao, and Kai Yu. Large language models are semi-parametric reinforcement learning agents. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 78227–78239. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/f6b22ac37beb5da61efd4882082c9ecd-Paper-Conference.pdf.

Li Zhou. A survey on contextual multi-armed bandits, 2016. URL https://arxiv.org/abs/1508.03326.

Li Zhou and Emma Brunskill. Latent contextual bandits and their application to personalized recommendations for new users, 2016. URL https://arxiv.org/abs/1604.06743.

A Additional Theoretical Details

In this appendix we provide proofs and derivations for the results stated in Section 4. We keep the notation from the main text: A_0, θ_0, and V_t denote the ridge pretraining precision, prior parameter, and cumulative design matrix, respectively, and $\mathcal{B}_0 := \|\theta_0 - \theta^\star\|_{A_0}$ is the prior mis-specification measured in the A_0-Mahalanobis norm.

Notation. For any symmetric positive semi definite matrix $G \in \mathbb{R}^{d\times d}$ and vector $v \in \mathbb{R}^d$ we write

$$\|v\|_G := \sqrt{v^\top G v}, \qquad \langle u, v\rangle_G := u^\top G v.$$

We write $\|\cdot\|_2$ for the Euclidean norm, and $A \succeq B$ for Loewner order on symmetric matrices.

A.1 Proof of Theorem 1

Recall that the warm-started ridge estimator is defined by

$$A_0 := X^\top X + \tau_{\text{pre}} I, \qquad b_0 := X^\top \tilde{y}, \qquad \theta_0 := A_0^{-1} b_0,$$

and the online design matrix is

$$V_t := A_0 + \sum_{s\le t} x_{s,a_s} x_{s,a_s}^\top.$$

The estimator at time t has the usual ridge form

$$\hat{\theta}_t = V_t^{-1}\left(A_0\theta_0 + \sum_{s\le t} r_s x_{s,a_s}\right).$$

Error decomposition. Using $r_s = x_{s,a_s}^\top \theta^\star + \xi_s$, we write

$$\begin{aligned} V_t(\hat{\theta}_t - \theta^\star) &= A_0\theta_0 + \sum_{s\le t} r_s x_{s,a_s} - \left(A_0 + \sum_{s\le t} x_{s,a_s} x_{s,a_s}^\top\right)\theta^\star \\ &= A_0(\theta_0 - \theta^\star) + \sum_{s\le t}\big(r_s - x_{s,a_s}^\top \theta^\star\big) x_{s,a_s} \\ &= A_0(\theta_0 - \theta^\star) + \sum_{s\le t} \xi_s x_{s,a_s}. \end{aligned}$$

Multiplying by $V_t^{-1/2}$ on the left gives

$$V_t^{1/2}(\hat{\theta}_t - \theta^\star) = V_t^{-1/2} A_0(\theta_0 - \theta^\star) + V_t^{-1/2}\sum_{s\le t} \xi_s x_{s,a_s}. \tag{16}$$

Taking Euclidean norms and applying the triangle inequality,

$$\|\hat{\theta}_t - \theta^\star\|_{V_t} = \|V_t^{1/2}(\hat{\theta}_t - \theta^\star)\|_2 \le \underbrace{\|V_t^{-1/2} A_0(\theta_0 - \theta^\star)\|_2}_{\text{prior term}} + \underbrace{\left\| V_t^{-1/2}\sum_{s\le t} \xi_s x_{s,a_s}\right\|_2}_{\text{noise term}}. \tag{17}$$

Bounding the prior term by $\mathcal{B}_0$. We first control the deterministic term. Using the definition of the A_0-norm and the fact that $V_t \succeq A_0$, we have

$$\begin{aligned} \|V_t^{-1/2} A_0(\theta_0 - \theta^\star)\|_2^2 &= (\theta_0 - \theta^\star)^\top A_0 V_t^{-1} A_0(\theta_0 - \theta^\star) \\ &= \|A_0^{1/2}(\theta_0 - \theta^\star)\|^2_{A_0^{1/2} V_t^{-1} A_0^{1/2}}. \end{aligned}$$

Since $V_t = A_0 + \sum_{s \le t} x_{s,a_s} x_{s,a_s}^\top \succeq A_0$, we have $V_t^{-1} \preceq A_0^{-1}$, and hence $A_0^{1/2} V_t^{-1} A_0^{1/2} \preceq I$. Therefore,

$$\|V_t^{-1/2} A_0(\theta_0 - \theta^\star)\|_2^2 \le \|A_0^{1/2}(\theta_0 - \theta^\star)\|_2^2 = \|\theta_0 - \theta^\star\|_{A_0}^2 = \mathcal{B}_0^2,$$

so

$$\|V_t^{-1/2} A_0(\theta_0 - \theta^\star)\|_2 \le \mathcal{B}_0. \tag{18}$$

Bounding the noise term. The second term in equation 17 is the self-normalized noise process

$$\left\| V_t^{-1/2} \sum_{s \le t} \xi_s x_{s,a_s} \right\|_2 = \left\| \sum_{s \le t} \xi_s x_{s,a_s} \right\|_{V_t^{-1}}.$$

Under the sub-Gaussian noise and bounded-feature assumptions, the self-normalized concentration inequality of Abbasi-Yadkori et al. (2011) implies that for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\left\| \sum_{s \le t} \xi_s x_{s,a_s} \right\|_{V_t^{-1}} \le \sigma \sqrt{2 \log \frac{\det(V_t)^{1/2}}{\det(A_0)^{1/2} \delta}} = \beta_t(\delta) \tag{19}$$

simultaneously for all $t \ge 0$.

Combining the two terms. Substituting equation 18 and equation 19 into equation 17 yields, on the same high-probability event and for all $t \ge 0$,

$$\|\hat{\theta}_t - \theta^\star\|_{V_t} \le \beta_t(\delta) + \mathcal{B}_0,$$

which is exactly the prior-centered confidence inequality equation 3. This proves Theorem 1.

Finally, applying equation 3 to a fixed context x gives

$$|x^\top(\hat{\theta}_{t-1} - \theta^\star)| = |\langle \hat{\theta}_{t-1} - \theta^\star, x \rangle| \le \|\hat{\theta}_{t-1} - \theta^\star\|_{V_{t-1}} \cdot \|x\|_{V_{t-1}^{-1}} \le (\beta_{t-1}(\delta) + \mathcal{B}_0)\sqrt{x^\top V_{t-1}^{-1} x},$$

which is the reward-confidence bound equation 4.

A.2 Bias–Variance Decomposition for $\mathcal{B}_0^2$

We now derive the decomposition of $\mathcal{B}_0^2$ and its expectation under the flip-noise model. Recall the regression proxy and ridge prior equation 1–equation 2:

$$\tilde{y} = (1-2p)X\theta^\star + p\mathbf{1} + \varepsilon, \qquad A_0 = X^\top X + \tau_{\text{pre}} I, \qquad \theta_0 = A_0^{-1} X^\top \tilde{y}.$$

Substituting $\tilde{y}$ into θ_0 yields

$$\begin{aligned}
\theta_0 &= A_0^{-1} X^\top \tilde{y} \\
&= A_0^{-1} X^\top \big((1-2p)X\theta^\star + p\mathbf{1} + \varepsilon\big) \\
&= (1-2p) A_0^{-1} X^\top X \theta^\star + p A_0^{-1} X^\top \mathbf{1} + A_0^{-1} X^\top \varepsilon \\
&= (1-2p) M \theta^\star + p A_0^{-1} X^\top \mathbf{1} + A_0^{-1} X^\top \varepsilon,
\end{aligned}$$

where $M := A_0^{-1} X^\top X$. Subtracting $\theta^\star$ and regrouping gives

$$\theta_0 - \theta^\star = \underbrace{((1-2p)M - I)\theta^\star + p A_0^{-1} X^\top \mathbf{1}}_{D} + \underbrace{A_0^{-1} X^\top \varepsilon}_{\text{pretraining noise}}. \tag{20}$$

Define the deterministic component

$$D := ((1-2p)M - I)\theta^\star + p A_0^{-1} X^\top \mathbf{1}.$$

Then the prior error in the A_0-norm satisfies

$$\begin{aligned}\mathcal{B}_0^2 &= \|\theta_0 - \theta^\star\|_{A_0}^2 = \|D + A_0^{-1}X^\top\varepsilon\|_{A_0}^2 \\ &= \|D\|_{A_0}^2 + \|A_0^{-1}X^\top\varepsilon\|_{A_0}^2 + 2\langle D, A_0^{-1}X^\top\varepsilon\rangle_{A_0}.\end{aligned}$$

Using $\|v\|_{A_0}^2 = v^\top A_0 v$ and the definition of the A_0-inner product, we can write the three terms explicitly as

$$\begin{aligned}\|D\|_{A_0}^2 &= D^\top A_0 D, \\ \|A_0^{-1}X^\top\varepsilon\|_{A_0}^2 &= \varepsilon^\top X A_0^{-1} A_0 A_0^{-1} X^\top \varepsilon = \varepsilon^\top X A_0^{-1} X^\top \varepsilon, \\ \langle D, A_0^{-1}X^\top\varepsilon\rangle_{A_0} &= D^\top A_0 A_0^{-1} X^\top \varepsilon = D^\top X^\top \varepsilon.\end{aligned}$$

Taking expectation over pretraining noise. We now take expectation with respect to the pretraining noise ε conditional on X, using the assumptions

$$\mathbb{E}[\varepsilon \mid X] = 0, \qquad \mathbb{E}[\varepsilon\varepsilon^\top \mid X] \preceq \sigma_s^2 I.$$

The cross term has mean zero:

$$\mathbb{E}[\langle D, A_0^{-1}X^\top\varepsilon\rangle_{A_0} \mid X] = D^\top X^\top \mathbb{E}[\varepsilon \mid X] = 0.$$

For the noise quadratic term we use the trace identity $\mathbb{E}[z^\top A z] = \operatorname{tr}(A\,\mathbb{E}[zz^\top])$ to obtain

$$\begin{aligned}\mathbb{E}\big[\|A_0^{-1}X^\top\varepsilon\|_{A_0}^2 \mid X\big] &= \mathbb{E}\big[\varepsilon^\top X A_0^{-1} X^\top \varepsilon \mid X\big] \\ &= \operatorname{tr}\big(X A_0^{-1} X^\top \mathbb{E}[\varepsilon\varepsilon^\top \mid X]\big) \\ &\le \sigma_s^2 \operatorname{tr}\big(X A_0^{-1} X^\top\big).\end{aligned}$$

Combining these pieces yields

$$\begin{aligned}\mathbb{E}[\mathcal{B}_0^2 \mid X] &\le \|D\|_{A_0}^2 + \sigma_s^2 \operatorname{tr}(X A_0^{-1} X^\top) \\ &= \|((1-2p)M - I)\theta^\star + pA_0^{-1}X^\top \mathbf{1}\|_{A_0}^2 + \sigma_s^2 \operatorname{tr}(X A_0^{-1} X^\top),\end{aligned}$$

which is exactly the bound stated in equation 9.

A.3 Eigenbasis Expansion and High-Coverage Approximation

We derive the direction-wise expression equation 10 for the deterministic flip-bias term and its high-coverage approximation.

Joint diagonalization. Let $X^\top X = U\Lambda U^\top$ be the eigendecomposition of the synthetic Gram matrix, with $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$ and U orthogonal. Because

$$A_0 = X^\top X + \tau_{\text{pre}} I = U(\Lambda + \tau_{\text{pre}} I)U^\top,$$

we have

$$A_0^{-1} = U(\Lambda + \tau_{\text{pre}} I)^{-1} U^\top.$$

The shrinkage operator $M := A_0^{-1}X^\top X$ shares the same eigenbasis:

$$M = U \operatorname{diag}\left(\frac{\lambda_i}{\lambda_i + \tau_{\text{pre}}}\right) U^\top.$$

Write $\theta^\star = U\theta_U^\star$ in the eigenbasis. Then

$$((1-2p)M - I)\theta^\star = U \operatorname{diag}\left((1-2p)\frac{\lambda_i}{\lambda_i + \tau_{\text{pre}}} - 1\right)\theta_U^\star.$$

The diagonal entries simplify to

$$(1-2p)\frac{\lambda_i}{\lambda_i+\tau_{\text{pre}}}-1=-\frac{\tau_{\text{pre}}+2p\lambda_i}{\lambda_i+\tau_{\text{pre}}}.$$

Hence the i-th coordinate of $((1-2p)M-I)\theta^\star$ in the U-basis is

$$v_i := -\frac{\tau_{\text{pre}}+2p\lambda_i}{\lambda_i+\tau_{\text{pre}}}\,\theta^\star_{U,i}.$$

Computing the A_0-norm. The A_0-norm of $((1-2p)M-I)\theta^\star$ satisfies

$$\|((1-2p)M-I)\theta^\star\|_{A_0}^2=\sum_{i=1}^{d}(\lambda_i+\tau_{\text{pre}})v_i^2,$$

because A_0 is diagonal with entries $(\lambda_i+\tau_{\text{pre}})$ in the U-basis. Substituting the expression for v_i gives

$$\|((1-2p)M-I)\theta^\star\|_{A_0}^2=\sum_{i=1}^{d}(\lambda_i+\tau_{\text{pre}})\left(\frac{\tau_{\text{pre}}+2p\lambda_i}{\lambda_i+\tau_{\text{pre}}}\right)^2(\theta^\star_{U,i})^2=\sum_{i=1}^{d}\frac{(\tau_{\text{pre}}+2p\lambda_i)^2}{\lambda_i+\tau_{\text{pre}}}(\theta^\star_{U,i})^2,$$

which is exactly equation 10.

High-coverage approximation. In directions where the synthetic design has strong coverage, $\lambda_i \gg \tau_{\text{pre}}$, we have

$$\frac{(\tau_{\text{pre}}+2p\lambda_i)^2}{\lambda_i+\tau_{\text{pre}}}\approx\frac{(2p\lambda_i)^2}{\lambda_i}=4p^2\lambda_i,$$

so

$$\|((1-2p)M-I)\theta^\star\|_{A_0}^2\approx 4p^2\sum_{i=1}^{d}\lambda_i(\theta^\star_{U,i})^2=4p^2\|(X^\top X)^{1/2}\theta^\star\|_2^2.$$

Since $A_0^{1/2}$ and $(X^\top X)^{1/2}$ are comparable in these directions ($\lambda_i \gg \tau_{\text{pre}}$ implies $\lambda_i+\tau_{\text{pre}}\approx\lambda_i$), this yields the norm-level approximation

$$\|((1-2p)M-I)\theta^\star\|_{A_0}\approx 2p\,\|A_0^{1/2}\theta^\star\|,$$

which is the heuristic form used in the main text (cf. equation 11). The exact dependence on $(p,\lambda_i,\tau_{\text{pre}})$ is given by equation 10.

A.4 High-Probability Control of the Pretraining-Noise Term

The expectation bound equation 9 controls the contribution of the pretraining noise ε in $\mathbb{E}[\mathcal{B}_0^2]$. For completeness, we record a high-probability bound on the same quantity; this is not used directly in the main text but may be useful for refined regret bounds.

Recall from equation 20 that the noise component of the prior error is

$$\|A_0^{-1}X^\top\varepsilon\|_{A_0}=\|A_0^{-1/2}X^\top\varepsilon\|_2.$$

Suppose ε has independent, mean-zero components that are σ_s^2-sub-Gaussian. Then $X^\top\varepsilon$ is a sub-Gaussian vector with proxy covariance $\sigma_s^2X^\top X$. A standard concentration inequality for quadratic forms of sub-Gaussian vectors (see, e.g., Vershynin (2025)) implies that, for any $\delta_s\in(0,1)$, with probability at least $1-\delta_s$,

$$\|A_0^{-1/2}X^\top\varepsilon\|_2\le\sigma_s\left(\sqrt{\operatorname{tr}(XA_0^{-1}X^\top)}+\sqrt{2\|XA_0^{-1}X^\top\|_{\text{op}}\log(1/\delta_s)}\right).$$

The leading trace term matches the scale of the variance contribution in equation 9, while the second term inflates this by an operator-norm factor to account for rare large deviations. In regimes where the synthetic design is well-conditioned in the A_0^{-1}-geometry, $\|XA_0^{-1}X^\top\|_{\text{op}}$ is not much larger than $\operatorname{tr}(XA_0^{-1}X^\top)$, so the noise contribution is sharply concentrated around its mean.

B Model Information

B.1 Experimental Details

In our experiments, we train LinUCB (Chu et al., 2011) with a fixed exploration parameter $\alpha = 10$. Data collection was performed using the OpenAI API (OpenAI, 2024) and Hugging Face `transformers` library (Wolf et al., 2020) for open-weights models.

All runs were executed on a 20-core Intel® Core i7-14700F (2.1 GHz), 32 GB DDR5 RAM, and an NVIDIA GeForce RTX 4070 Ti SUPER GPU with 16 GB of dedicated memory. The largest full sweep tested, comprising the `10k` baseline, `10k_x1`...`10k_x7` variants, and a cold-start baseline, each repeated for ten independent rounds, completed in under two wall-clock hours.

B.2 Model Specifications

Table 5 details the specific model revisions used. For open-weights models, we pinpoint the exact snapshot using the Hugging Face commit hash (first 7 characters). The earlier snapshot results in Table 7 correspond to the Jan–Feb 2025 period.

Table 5: Model specifications. OpenAI models are listed by access window; open-weights models include their Git revision ID.

Model	Variant	Checkpoint Path	Revision (Hash)
Llama 3.1	8B Instruct	`meta-llama/llama-3.1-8B-Instruct`	`0e9e39f`
Qwen 3	8B Instruct	`Qwen/Qwen3-8B`	`b968826`
GPT-3.5	Turbo	*Proprietary API*	Sept–Oct 2025
GPT-4o	Omni	*Proprietary API*	Sept–Oct 2025

B.3 Inference Hyperparameters

To ensure fair comparison across model families, we aligned inference parameters as closely as possible. Table 6 details the generation configuration. For OpenAI models, we utilized the default system settings with a fixed temperature. For open-weights models (Llama 3.1, Qwen 3), we utilized the `transformers` library with explicit generation limits to prevent infinite loops in chain-of-thought sequences.

Table 6: Inference parameters.

Parameter	**OpenAI Models** *(GPT-3.5, GPT-4o)*	**Open-Weights Models** *(Llama 3.1, Qwen 3)*
Temperature	0.5	0.5
Max Output Tokens	Model Maximum	4,000
Top_p	1.0	1.0
Frequency Penalty	0.0	0.0
Presence Penalty	0.0	0.0
Stop Sequences	None	None

B.4 Snapshot Sensitivity (GPT3.5 Turbo)

To assess the temporal stability of LLM-generated priors, we consider data generated with GPT3.5 Turbo using an earlier model snapshot (Jan–Feb 2025). All other components remain fixed (settings listed in Table 6). Table 7 reports the resulting δ regret under preference-flipping, in the same format as the main cross-domain summary tables, specifically compared to the results in Table 2. We note that the data in Table 2 is averaged over $G = 20$ seeds compared to the results in Table 7, which are averaged across $G = 10$ seeds.

Table 7: GPT-3.5 Turbo results from older snapshot under preference flipping; percentage reduction in cumulative regret (%Δ Regret) versus a cold-start LinUCB baseline. Mean over $G = 10$ seeds ± 95 % CI.

Dataset	**Noise (%)**	**N = 1k**	**N = 3k**	**N = 10k**
COVID-19	0	17.57 ± 4.29	11.23 ± 2.61	9.45 ± 1.17
COVID-19	30	11.27 ± 3.78	5.91 ± 4.48	7.44 ± 1.34
COVID-19	50	7.74 ± 2.65	1.81 ± 4.20	-8.07 ± 3.45
Immigration	0	13.03 ± 2.37	6.22 ± 1.41	4.03 ± 1.22
Immigration	30	6.49 ± 4.78	1.32 ± 2.83	0.56 ± 1.25
Immigration	50	-2.05 ± 4.16	-15.15 ± 4.12	-17.33 ± 3.50
Travel	0	4.46 ± 4.62	1.45 ± 2.67	4.11 ± 1.01
Travel	30	1.39 ± 5.57	0.84 ± 0.84	0.18 ± 1.08
Travel	50	0.42 ± 2.62	0.02 ± 3.78	-0.51 ± 1.50

Table 8: Prompt-order sensitivity for $N = 1$k synthetic priors. The order-swapped prompt exchanges the textual positions of Option A and Option B while keeping all attributes fixed. Agreement denotes the fraction of synthetic labels that match the original prompt after mapping responses back to the original arm identities. Regret values are percentage reduction in cumulative regret relative to cold-start.

Model	**Dataset**	**Agreement**	**Original Prompt**	**Swapped Prompt**
Qwen 3	COVID-19	93%	2.87 ± 1.07	3.02 ± 0.98
Qwen 3	Immigration	77%	1.71 ± 1.18	2.27 ± 1.33
Qwen 3	Travel	79%	-1.74 ± 0.60	-0.13 ± 0.59
Llama 3.1	COVID-19	96%	5.84 ± 0.77	5.92 ± 0.70
Llama 3.1	Immigration	58%	0.03 ± 1.15	-2.68 ± 1.49
Llama 3.1	Travel	53%	-1.90 ± 0.82	2.11 ± 1.03

Comparing the two snapshots reveals significant performance degradation over time, suggesting a form of "alignment drift." On the Immigration dataset, the earlier snapshot achieved a positive clean-prior gain at $N = 10$k (4.03%), whereas the newer snapshot reported in Table 2 slightly underperforms cold-start (-0.93%), a drop of roughly 5.0 percentage points. On the Travel dataset, the clean-prior gain at $N = 10$k decreases from 4.11% with the earlier snapshot to 2.49% with the newer snapshot. On COVID-19, large-N clean-prior performance remains broadly similar (9.45% versus 8.29% at $N = 10$k), but the earlier snapshot provides a much stronger low-data initialization: at $N = 1$k, regret reduction drops from 17.57% to 6.04% in the newer snapshot. These results indicate that effective alignment is not a static property of a model family (e.g., "GPT-3.5") but is sensitive to specific version updates and other adjustments.

B.5 Prompt-Order Sensitivity

To assess whether the LLM-generated priors are sensitive to prompt presentation order, we run a small prompt-order ablation. For each synthetic conjoint query, we keep the respondent context, candidate attributes, model, decoding parameters, and parsing procedure fixed, but swap the textual order in which the two alternatives are presented: the profile previously shown as Option A is shown as Option B, and vice versa. We map each response back to the original arm identity before constructing the synthetic pretraining label. A disagreement between the original and swapped prompts indicates that the model changed its underlying choice after only the display order was altered.

We run this ablation for Qwen 3 and Llama 3.1 using $N = 1$k synthetic pretraining samples. These models are useful stress cases because the main results show weaker or more model-dependent alignment for these priors. All other parameters match the main experimental protocol (Table 6), and no synthetic noise is injected. Table 8 reports both the agreement rate between the original and order-swapped synthetic labels and the downstream percentage reduction in cumulative regret relative to cold-start.

The order-swap ablation is consistent with the prompt-sensitivity limitation discussed in Section 7. Even with the same decoding temperature and all other inference settings fixed, changing only the textual order

of the two alternatives can alter the synthetic prior. The effect is small in the more aligned COVID-19 setting, where both Qwen 3 and Llama 3.1 show high agreement and nearly unchanged downstream regret. In contrast, agreement falls substantially on Immigration and Travel, especially for Llama 3.1, and the downstream regret can shift by several percentage points. The direction of the shift is not uniform: swapping the option order can either improve or worsen regret. This ablation should not be interpreted as evidence of a simple positional bias, but rather as evidence that weakly aligned LLM priors can be sensitive to prompt formatting even under fixed inference parameters.

B.6 Prompts

The next subsections list the exact prompts used to generate synthetic conjoint responses. Placeholders such as `[User]` and `[Vaccine A]` are replaced at runtime.

B.6.1 COVID-19 Vaccine

Following Alamdari et al. (2024), we reuse their prompt verbatim:

> *Consider you are in the middle of the COVID pandemic, where vaccines are just being produced. Pretend to be the following user: [User]. Now you are given two vaccine choices for COVID. The description of each vaccine is as follows: [Vaccine A] Now the next one: [Vaccine B]. Which one do you take? A or B? Let's think step by step. Print the final answer as [Final Answer] at the end as well.*

B.6.2 Immigration

> *Pretend to be the following user: [User]. You are now evaluating two immigrants applying for admission to the United States. The description of each immigrant is as follows: [Immigrant A] Now the next one: [Immigrant B]. Which immigrant do you admit? A or B? Let's think step by step. Print the final answer as [Final Answer] at the end.*

B.6.3 Travel

> *Consider you are planning a U.S. vacation and some states have recently passed policies that weaken democratic principles. Pretend to be the following user: [User]. Now you are given two locations for vacationing. The description of each location is as follows: [Location A], now the next one: [Location B]. Which location do you visit? A or B? Let's think step by step. Print the final answer as [Final Answer].*