Robustness of LLM-Initialized Bandits for Recommendation Under Noisy Priors

Adam Bayley
Queen's University
Kingston, ON, Canada

Kevin H. Wilson
RBC Borealis
Toronto, ON, Canada

Yanshuai Cao
RBC Borealis
Toronto, ON, Canada

Raquel Aoki
RBC Borealis
Vancouver, BC, Canada

Xiaodan Zhu
Queen's University
Kingston, ON, Canada

Abstract

Contextual bandits have proven effective for building personalized recommender systems, yet they suffer from the cold-start problem when little user interaction data is available. Recent work has shown that Large Language Models (LLMs) can help address this by simulating user preferences to warm-start bandits—a method known as Contextual Bandits with LLM Initialization (CBLI). While CBLI reduces early regret, it is unclear how robust the approach is to inaccuracies in LLM-generated preferences. In this paper, we extend the CBLI framework to systematically evaluate its sensitivity to noisy LLM priors. We inject both random and label-flipping noise into the synthetic training data and measure how these affect cumulative regret across three tasks generated from conjoint-survey datasets. Our results show that CBLI is robust to random corruption but exhibits clear breakdown thresholds under preference-flipping: warm-starting remains effective up to 30% corruption, loses its advantage around 40%, and degrades performance beyond 50%. We further observe diminishing returns with larger synthetic datasets: beyond a point, more data can reinforce bias rather than improve performance under noisy conditions. These findings offer practical insights for deploying LLM-assisted decision systems in real-world recommendation scenarios.

CCS Concepts

• **Information systems** → **Recommender systems**.

Keywords

Recommendation Systems, Contextual Bandits, Large Language Models

ACM Reference Format:
Adam Bayley, Kevin H. Wilson, Yanshuai Cao, Raquel Aoki, and Xiaodan Zhu. 2025. Robustness of LLM-Initialized Bandits for Recommendation Under Noisy Priors. In *Proceedings of Workshop on Generative AI for Recommender Systems and Personalization (GenAIRecP '25).* ACM, New York, NY, USA, 7 pages.

1 INTRODUCTION

Personalization in recommender systems presents significant challenges due to the inherently dynamic and context-dependent nature of user preferences. Factors such as time, location, and device substantially influence user relevance. Accordingly, context-aware approaches that integrate these factors have been demonstrated to significantly enhance recommendation quality and user engagement [26].

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) essentially 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 [21] and admit a sublinear finite-time regret bound under linear payoff assumptions [8]. However, when CBs have yet to gather any user data, they perform essentially randomly and thus exhibit linear regret [2, 21]. Traditional approaches have sought to address this limitation by warm-starting bandits using historical user data or expert knowledge [37].

Large Language Models (LLMs) offer a promising alternative by providing built-in knowledge about human preferences [6]. Alamdari et al. [1] 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, they showed that "jump-starting" a CB with this synthetic data achieved a 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.

Motivation. Previous studies have shown conflicting evidence on whether LLMs can accurately simulate human decision making [5, 18]. Despite the demonstrated benefits of the CBLI framework, its robustness to these potential biases is insufficiently understood. Understanding when the framework may break down is critical before deployment in real world systems.

Our Contributions. In this paper, we systematically evaluate the robustness of the CBLI framework to inaccuracies in LLM-generated priors. Our contributions are as follows:

- **Noisy-CBLI Framework**: We introduce an extension to CBLI where synthetic noise is injected into the LLM generated preference data. 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 datasets (including the original vaccine-preference data) to measure how different types and levels of noise affect cumulative regret. We compare conditions across noise injection strategies—random replacement and preference flipping—as well as noise intensities and task domains. Our results identify clear thresholds for systematic label corruption: warm-start gains persist up to 30% preference-flipping, vanish around 40%, and reverse beyond 50%, where synthetic priors become harmful. In contrast, bandits remain resilient to random replacement noise: regret increases gradually, and warm-start continues to match or outperform cold-start even at the highest tested corruption levels. However, we also observe diminishing returns with larger synthetic datasets: while confidence intervals shrink, the marginal benefit flattens and reverses as noise increases.

2 RELATED WORK

Contextual Bandits. Sequential recommender systems model user preferences over time by leveraging historical sequences of item-user interactions. These models operate under the assumption that prior user behavior provides a predictive signal for future engagement [14, 30] However, training these systems online, at scale, and in cold-start scenarios have been an important concern [22].

To address these limitations, contextual bandit algorithms (CBs) [21, 39], were introduced as a principled framework for online recommendation under partial feedback. Li et al. [21] introduced LinUCB, a foundational CB algorithm first utilized for personalized news recommendation, and for which theoretical regret analysis appears in [8]. In this setting, the CB utilizes contextual features to estimate expected rewards and select content accordingly, thereby enabling online learning under cold-start conditions. This formulation established a practical framework for integrating exploration-exploitation tradeoffs into real-time recommendations.

LLMs and CBs. Subsequent work extended this paradigm by incorporating large language models (LLMs) into both recommendation and bandit architectures. For instance, LLMs have been used to encode high-dimensional textual input into dense arm representations to improve reward estimation [3]. Recommendation tasks have been reframed as conditional text generation in prompt-based frameworks, supporting zero-shot generalization across domains [11]. Subsequent work has advanced this perspective by casting the recommendation process as autoregressive modeling over item sequences [28, 32]. Other works have focused on using long-context transformers for next-item prediction directly from unstructured user-item histories [11].

Beyond modeling capabilities, alignment methods such as reinforcement learning from human feedback (RLHF) and Constitutional AI enable preference-tuned LLMs to approximate human reward functions, providing stable, low-variance synthetic signals for initializing or evaluating interactive policies [4, 25].

Warm-starting bandits with auxiliary or synthetic datasets. While contextual bandits provide a framework for real-time personalization, their early performance can be suboptimal due to limited initial data [21, 38]. To mitigate this, a growing body of work explores warm-starting strategies using auxiliary and synthetic datasets, as well as transfer learning [1, 15]. Pre-training on expert-labeled or cross-domain data has been shown to reduce initial regret, though success often hinges on alignment between the pre-training and target distributions [7, 37].

In the LLM context, approaches such as Universal Language Model Fine-tuning (ULMFiT) apply general-domain pre-training followed by task-specific fine-tuning, achieving faster convergence and improved generalization despite domain mismatch. Similarly, controlled-noise fine-tuning, as in NoisyTune, injects perturbations during adaptation to improve robustness to distributional shift in neural bandit models [33]. Recent work has shown that LLMs can mimic key aspects of human cognition, including multi-step reasoning, deliberation, and action selection. The ReAct framework [36] combines reasoning traces with actions, enabling LLMs to interact with environments in a human-like fashion.

Building on this, prompt-based methods such as CBLI simulate user preference data for synthetic pre-training, achieving a 14–20% reduction in early-stage regret relative to standard baselines [1].

Despite these advances, key challenges remain. Simulated user behavior may diverge from real-world interaction patterns, particularly under strategic, long-horizon, or feedback-driven conditions. Ensuring alignment between synthetic and deployed environments, and developing methods for online correction or adaptation, remains an open area of research.

Positioning of this work. This work builds directly on the CBLI framework, which demonstrated the feasibility of using LLM-generated synthetic preferences to warm-start contextual bandits. While prior studies have shown that such synthetic priors can reduce early regret, the conditions under which this advantage holds remain poorly understood. In particular, it is unclear how the quality and distributional characteristics of the generated data influence downstream performance. We address this gap by systematically varying the type and magnitude of noise injected into LLM-generated pre-training datasets and evaluating their effect on cumulative regret. Experiments are conducted across three real-world conjoint-survey datasets, allowing us to identify the noise thresholds at which LLM-based warm starts cease to be beneficial. Our findings offer practical guidance for deploying synthetic-data-driven initialization in bandit-based recommender systems under distributional uncertainty.

3 METHOD

In this section, we describe three real-world conjoint datasets we use in our study. We 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—to systematically corrupt synthetic priors and assess CBLI's robustness 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 [20].** 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 [19]. 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 [12].** 1,714 American adults each completed five pairwise choice tasks, selecting which of two hypothetical immigrant applicants they would admit [13]. 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 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 [23].** In this dataset, roughly 2,000 American adults evaluated ten choice tasks, choosing between three U.S.-based leisure-travel destinations [24]. 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 mathematical problem following the "jump-start" formalized by Alamdari et al. [1]. Each conjoint survey is cast as a *sleeping* contextual bandit [17] over T rounds.

(1) **Rounds & Arms.** At round $t \in \{1, \dots, T\}$, a subset of arms $\mathcal{A}_t$ is presented (e.g., the two vaccines in Dataset 1 or three destinations in Dataset 3).

(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 [21], 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 [8, 21].

(5) **Regret.** At round t, let a_t be the arm chosen by LinUCB 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 regret is the random variable

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

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 $G = 10$ independent seeds. For theoretical comparison we refer to the expected (pseudo-)regret

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

which is the scalar quantity bounded by $\widetilde{O}(\sqrt{T\,d})$ for LinUCB [21].

(6) **Ordinal Rewards.** The LLM is only ever asked to compare two arms, so its outputs yield a pairwise preference rather than an absolute score. As shown in [1], summing these binary comparisons recovers the correct ranking of arms by success probability, even though the true reward magnitudes are not preserved.

3.3 CBLI "Jump-Start" Pipeline

We implement the "jump-start" pipeline introduced in [1]:

(1) **Pre-training on LLM-Generated Priors.** Generate N synthetic context–reward pairs via LLM prompts. 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.

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. Let p denote the corruption rate (the proportion of synthetic samples to modify).

(1) **Random Response Replacement.** We uniformly at random select a proportion p of the LLM-generated labels (each an arm index $a \in \{1, \dots, K\}$) and overwrite each with a new arm drawn uniformly from $\{1, \dots, K\}$. 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 $G = 10$ 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 reward $r_t \in \{0, 1\}$ to be 1 when the bandit chooses the correct arm. 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 $p \in \{0.0, 0.1, \dots, 0.7\}$, 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 RESULTS

4.1 Effect of Preference-Flipping Noise on COVID-19 Vaccine Conjoint

Figure 1 plots the mean cumulative regret of LinUCB warm-started on GPT-3.5-Turbo priors corrupted by systematic preference flipping at seven noise levels (0.1 to 0.7), together with the uncorrupted baseline ("10k_base") and a cold-start baseline ("Not Pretrained"). Each curve is averaged over $G = 10$ trials, with shaded bands showing the 95% confidence interval.

- **Zero noise (10k_base).** With no flips, warm-start achieves the lowest regret, quickly approaching optimal arm selection and substantially outperforming cold-start. This aligns with the results from [1], where uncorrupted LLM-generated priors can help alleviate the cold start problem in contextual bandit scenarios.
- **Low to moderate noise (10–30%).** Even when 10 percent or 20 percent of the synthetic labels are flipped (curves "10k_f1" and "10k_f2"), warm-start still outperforms cold-start, with only a 4-6 pp upward shift in regret. At 30 percent flips ("10k_f3"), the margin narrows but remains positive.
- **High noise (40–70%).** Once flipping reaches 40 percent ("10k_f4"), the regret advantage essentially vanishes: the warm-start curve overlaps "Not Pretrained." At 50–70 percent flips ("10k_f5"–"10k_f7"), corrupted priors begin to harm early performance, yielding higher regret than cold-start throughout.

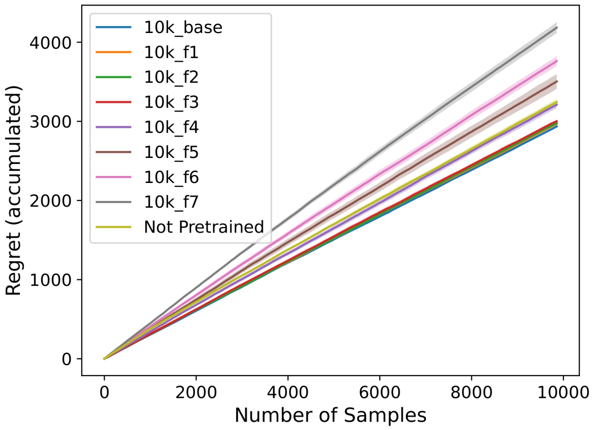

Figure 1: Cumulative regret on the COVID-19 Vaccine dataset under preference-flipping noise. "10k_fX" indicates X times 10% of LLM-generated labels flipped. Shaded regions are 95 percent CI over $G = 10$ runs.

4.2 Effect of Random Response Noise on COVID-19 Vaccine Conjoint

Figure 2 plots the mean cumulative regret of LinUCB warm-started on GPT-3.5-Turbo priors corrupted by random responses with the same baselines and noise levels in Figure 1. Each curve is averaged over $G = 10$ trials, with shaded bands showing the 95% confidence interval.

- **Zero noise (10k_base).** Without corruption, warm-start yields the lowest regret, converging to optimal recommendations.
- **Low noise (10–30 percent).** At 10 percent and 20 percent random replacements ("10k_r1" and "10k_r2"), regret rises only slightly yet remains below cold-start. Even 30 percent noise ("10k_r3") still maintains a clear warm-start advantage.
- **Moderate to high noise (40–70 percent).** From 40 percent onward ("10k_r4"–"10k_r7"), regret gradually approaches the cold-start curve. However, unlike systematic flips, we observe random corruption never produces regret worse than cold-start. These results imply that when errors in LLM-generated priors are random, a recommender can keep using its warm-start model without harm, only falling back to full exploration if corruption shows systematic bias.

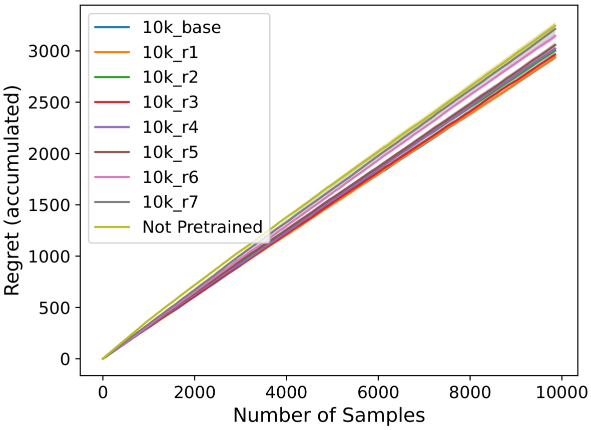

Figure 2: Cumulative regret on the COVID-19 Vaccine dataset under random-response noise.

Table 1: Percentage reduction in cumulative regret (%Δ Regret) versus a cold-start LinUCB baseline when using preference-flipping. 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

4.3 Comparison of Noise Effects on Other Datasets

Table 1 reports the percentage change in cumulative regret relative to a cold−start LinUCB based on the breakpoints found earlier. The results reveal three general broad patterns:

(1) *Benefit persists through* ≤ 30% *noise.* With clean priors, the COVID−19 and Immigration tasks achieve +17.6 ± 4.3 pp and +13.0 ± 2.4 pp at N=1k, respectively, while Travel shows a statistically non-zero +4.5 pp (p = 0.03). Even at 30 % corruption warm−start remains advantageous in every cell.

(2) *Performance collapses at* 50 % *flips.* Once half the synthetic labels are wrong, at least one pre-train size per domain turns negative (−8.1 ± 3.5 pp on COVID−19, −17.3 ± 3.5 pp on Immigration, −0.5 ± 1.5 pp on Travel), indicating that heavily biased priors can harm learning more than cold starts.

(3) *More synthetic data yields diminishing and sometimes adverse returns.* Confidence intervals tighten as N grows (e.g. from ±4.3 pp at 1k to ±1.2 pp at 10k on clean COVID−19), but the mean gains plateau and eventually reverse under high noise because additional corrupted samples reinforce the bias.

A cross-dataset contrast is instructive: Travel exhibits uniformly smaller deltas than COVID−19 and Immigration. Each Travel respondent contributes ten logged interactions (versus five in the other surveys), so the cold-start baseline is warmer *a priori*, leaving less head-room for synthetic jump-starts. That COVID−19 and Immigration follow nearly identical degradation curves—despite different content domains, which suggests the thresholds observed here capture a general noise-sensitivity of warm-started LinUCB rather than task-specific quirks.

Implications for deployed recommenders. Practitioners can safely leverage LLM-generated priors when they are expected to be at least ∼70 % accurate; below that, the synthetic head-start quickly erodes and may even back-fire, especially if large synthetic corpora are used. Systems with abundant early feedback (analogous to the Travel dataset) derive only marginal benefit, whereas sparse-interaction settings stand to gain the most from modestly reliable warm-starts.

5 CONCLUSION

We conducted a systematic study of how noise in LLM-generated priors affects CBLI's warm-start advantage over cold-start LinUCB. Across three real-world conjoint tasks, warm-start retains a lower cumulative regret than cold-start up to ≈ 30% corruption. Beyond ≈ 40% noise, this benefit diminishes rapidly, and at high corruption levels ≥ 50%, noisy priors can degrade performance relative to learning from scratch. Moreover, random response replacements are easier for LinUCB to overcome than systematic preference flips. However, adding more synthetic data does not always help: as noise rises, larger synthetic datasets can amplify harmful biases, yielding diminishing or even negative returns. These findings imply that AI-augmented recommenders should leverage LLM-generated priors when they are moderately reliable, but revert to pure exploration once synthetic feedback becomes excessively noisy.

6 LIMITATIONS

Our approach demonstrates strong robustness in the CBLI method to noise; however, it has limitations. The warm-start procedure depends on a fixed set of prompts, yet LLM outputs are highly prompt-sensitive: minor wording changes, added context, or altered arm order [29] can materially shift the synthetic labels, so the evaluation may provide an unduly narrow estimate of variance in the prior [9, 31]. 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 [35]. 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 [27]. Lastly, LLMs encode demographic and ideological biases from their training data. When such biases manifest in synthetic preferences [34], 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 [10].

References

[1] Parand A. Alamdari, Yanshuai Cao, and Kevin H. Wilson. 2024. Jump Starting Bandits with LLM-Generated Prior Knowledge. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 19821–19833. doi:10.18653/v1/2024.emnlp-main.1107

[2] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. 2002. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning* 47 (05 2002), 235–256. doi:10.1023/A:1013689704352

[3] Ali Baheri and Cecilia O. Alm. 2023. LLMs-augmented Contextual Bandit. arXiv:2311.02268 [cs.LG] https://arxiv.org/abs/2311.02268

[4] 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. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073 [cs.CL] https://arxiv.org/abs/2212.08073

[5] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. doi:10.1145/3442188.3445922

[6] 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. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL] https://arxiv.org/abs/2005.14165

[7] Changxiao Cai, T. Tony Cai, and Hongzhe Li. 2024. Transfer Learning for Contextual Multi-armed Bandits. arXiv:2211.12612 [stat.ML] https://arxiv.org/abs/2211.12612

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

[9] Federico Errica, Giuseppe Siracusano, Davide Sanvito, and Roberto Bifulco. 2025. What Did I Do Wrong? Quantifying LLMs' Sensitivity and Consistency to Prompt Engineering. arXiv:2406.12334 [cs.LG] https://arxiv.org/abs/2406.12334

[10] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. arXiv:2309.00770 [cs.CL] https://arxiv.org/abs/2309.00770

[11] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2023. Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt Predict Paradigm (P5). arXiv:2203.13366 [cs.IR] https://arxiv.org/abs/2203.13366

[12] Jens Hainmueller. 2014. Replication data for: The Hidden American Immigration Consensus: A Conjoint Analysis of Attitudes toward Immigrants. doi:10.7910/DVN/25505

[13] Jens Hainmueller and Daniel J. Hopkins. 2015. The Hidden American Immigration Consensus: A Conjoint Analysis of Attitudes Toward Immigrants. *American Journal of Political Science* 59, 3 (2015), 529–548. doi:10.2139/ssrn.2106116 Available at SSRN: https://ssrn.com/abstract=2106116.

[14] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2016. Session-based Recommendations with Recurrent Neural Networks. arXiv:1511.06939 [cs.LG] https://arxiv.org/abs/1511.06939

[15] Manisha Jangid and Rakesh Kumar. 2023. Mitigating Cold Start Problem in Recommendation Systems via Transfer Learning Approach. In *2023 International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC)*. 1–6. doi:10.1109/ISACC56298.2023.10084296

[16] Sandhini Agarwal Lama Ahmad Ilge Akkaya Florencia Leoni Aleman Diogo Almeida Janko Altenschmidt Sam Altman Shyamal Anadkat et al. 2023. Josh Achiam, Steven Adler. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL] https://arxiv.org/abs/2303.08774

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

[18] Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences* 121, 45 (Oct. 2024). doi:10.1073/pnas.2405460121

[19] Sarah Kreps, Sandip Prasad, John S. Brownstein, Yulin Hswen, Brian T. Garibaldi, Baobao Zhang, and Douglas L. Kriner. 2020. Factors Associated With US Adults' Likelihood of Accepting COVID-19 Vaccination. *JAMA Network Open* 3, 10 (10 2020), e2025594–e2025594. doi:10.1001/jamanetworkopen.2020.25594

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

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

[22] Zhuoran Liu, Leqi Zou, Xuan Zou, Caihua Wang, Biao Zhang, Da Tang, Bolin Zhu, Yijie Zhu, Peng Wu, Ke Wang, and Youlong Cheng. 2022. Monolith: Real Time Recommendation System With Collisionless Embedding Table. arXiv:2209.07663 [cs.IR] https://arxiv.org/abs/2209.07663

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

[24] David Miller and Serena Smith. 2024. (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 (04 2024), 1–11. doi:10.1017/XPS.2023.40

[25] 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. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155 [cs.CL] https://arxiv.org/abs/2203.02155

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

[27] Nishan Pantha, Muthukumaran Ramasubramanian, Iksha Gurung, Manil Maskey, and Rahul Ramachandran. 2024. Challenges in Guardrailing Large Language Models for Science. arXiv:2411.08181 [cs.AI] https://arxiv.org/abs/2411.08181

[28] Aleksandr V. Petrov and Craig Macdonald. 2023. Generative Sequential Recommendation with GPTRec. arXiv:2306.11114 [cs.IR] https://arxiv.org/abs/2306.11114

[29] Pouya Pezeshkpour and Estevam Hruschka. 2023. Large Language Models Sensitivity to The Order of Options in Multiple-Choice Questions. arXiv:2308.11483 [cs.CL] https://arxiv.org/abs/2308.11483

[30] Massimo Quadrana, Paolo Cremonesi, and Dietmar Jannach. 2018. Sequence-Aware Recommender Systems. arXiv:1802.08452 [cs.IR] https://arxiv.org/abs/1802.08452

[31] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. arXiv:2310.11324 [cs.CL] https://arxiv.org/abs/2310.11324

[32] Anna Volodkevich, Danil Gusak, Anton Klenitskiy, and Alexey Vasilev. 2024. Autoregressive Generation Strategies for Top-K Sequential Recommendations. arXiv:2409.17730 [cs.IR] https://arxiv.org/abs/2409.17730

[33] Chuhan Wu, Fangzhao Wu, Tao Qi, Yongfeng Huang, and Xing Xie. 2022. NoisyTune: A Little Noise Can Help You Finetune Pretrained Language Models Better. arXiv:2202.12024 [cs.CL] https://arxiv.org/abs/2202.12024

[34] Sierra Wyllie, Ilia Shumailov, and Nicolas Papernot. 2024. Fairness Feedback Loops: Training on Synthetic Data Amplifies Bias. arXiv:2403.07857 [cs.LG] https://arxiv.org/abs/2403.07857

[35] Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. 2020. Part-dependent Label Noise: Towards Instance-dependent Label Noise. arXiv:2006.07836 [cs.LG] https://arxiv.org/abs/2006.07836

[36] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. arXiv:2210.03629 [cs.CL] https://arxiv.org/abs/2210.03629

[37] Chicheng Zhang, Alekh Agarwal, Hal Daumé Iii, John Langford, and Sahand Negahban. 2019. Warm-starting Contextual Bandits: Robustly Combining Supervised and Bandit Feedback. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 7335–7344. https://proceedings.mlr.press/v97/zhang19b.html

[38] Li Zhou. 2016. A Survey on Contextual Multi-armed Bandits. arXiv:1508.03326 [cs.LG] https://arxiv.org/abs/1508.03326

[39] Li Zhou and Emma Brunskill. 2016. Latent Contextual Bandits and their Application to Personalized Recommendations for New Users. arXiv:1604.06743 [cs.LG] https://arxiv.org/abs/1604.06743

A Experimental Details

In our experiments, we train LinUCB [8] with $\alpha = 10$. We use ChatGPT [16] to collect the data for our experiments using the API provided by OpenAI.

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 24 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 three wall-clock hours.

A.1 OpenAI gpt-3.5-turbo API Settings

Table 2 lists the exact parameters sent to the OpenAI `ChatCompletion` endpoint; we override only `temperature`, leaving all other fields at their default values.

Parameter	Value
temperature	0.5
top_p	1.0
n	1
stream	False
stop	None
max_tokens	None
presence_penalty	0.0
frequency_penalty	0.0
logit_bias	{}
user	None

Table 2: ChatCompletion parameters.

A.2 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.

A.2.1 COVID-19 Vaccine. Following Alamdari et al. [1], 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.*

A.2.2 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.*

A.2.3 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].*

accepted 18 June 2025