

Comparative Analysis of ETC, UCB, and Thompson Sampling for Personalized Video Recommendations on Short-Video Platform

Shuqiao Chen

*School of Computer Science, University of Manchester, Manchester, United Kingdom
shuqiao.chen@student.manchester.ac.uk*

Abstract. This study empirically compares three canonical Multi-Armed Bandit (MAB) algorithms—Explore-Then-Commit (ETC), fixed initial exploration, Upper Confidence Bound (UCB1), which is the optimism-driven uncertainty estimation, and Thompson Sampling (TS) with Bernoulli likelihood (TS-Bernoulli, posterior-sampling-based)—for short-video recommendation, aiming to solve the exploration-exploitation tradeoff in real-time feed systems. Experiments were conducted on the ShortVideo-Interactions (SVI-200K) dataset, a simulated corpus with ~ 1.2 million timestamped impressions and clicks from 240,000 user sessions over 30 days, covering $\sim 18,000$ unique items to mimic real platform dynamics. Evaluations used a fixed horizon ($T=2000$ timesteps) and restricted candidates to the top 200 items ($K=200$) per run, spanning three practical scenarios: stable base, information-scarce cold-start (new items with no prior data), and preference-drifting temporal-shift. Results, aggregated over three pseudo-random seeds (2025, 2026, 2027), show TS-Bernoulli consistently outperforms peers: it achieves the highest Click-Through Rate (CTR) (0.452 in base, 0.402 in cold-start, 0.428 in temporal-shift) and lowest cumulative regret (418, 518, 467 respectively). These findings confirm that TS-Bernoulli’s posterior sampling enables robust adaptation to short-video recommendation’s key challenges (information scarcity and non-stationarity), providing a practical algorithm choice for real-world platforms.

Keywords: Multi-Armed Bandits, Thompson Sampling, Short-Video Recommendation, Cold-Start, Cumulative Regret

1. Introduction

Short-video feeds operate at millisecond latency, requiring real-time item selection per scroll to maximize user engagement [1]. Reward feedback (click/no-click) is immediate yet extremely sparse and volatile, demanding constant balance between exploration and exploitation [2]. MABs are well-suited for this context, as they update from streaming feedback without heavy batch retraining [3].

This paper conducts a data-driven comparison of three classic bandit algorithms—ETC, UCB1, and TS-Bernoulli—on the ShortVideo-Interactions (SVI-200K) dataset. The recommendation problem is framed as Bernoulli reward maximization over $K=200$ candidate items across $T=2000$

decisions, evaluated under three realistic scenarios: stable base, cold-start (limited prior information), and temporal-shift (evolving preferences) [4]. Results aggregated over seeds {2025, 2026, 2027} show TS-Bernoulli’s consistent superiority: mean CTR/regret of 0.452/417.85 (base), 0.402/517.70 (cold-start), and 0.428/466.51 (temporal-shift). Relative to UCB1, this translates to 22.5%–37.2% CTR gains and 22.4%–29.6% regret reductions, with cold-start optimal-play fraction rising from 0.38% to 1.5% (294.7% relative improvement) [5].

Contributions include: a lightweight, reproducible evaluation harness logging CTR, regret, and optimal-play ratios; systematic scenario comparisons validating TS-Bernoulli’s dominance; and released artifacts (csvs, plots) ensuring traceability [6].

2. Related work

The MAB problem has robust theoretical foundations, including asymptotic lower bounds and efficient policies [7]. Optimism-in-uncertainty methods like UCB1 offer finite-time guarantees in stationary settings but lack adaptability [8]. Thompson Sampling, a Bayesian approach sampling from posterior beliefs, is simple to implement and competitive, with modern analyses clarifying its regret properties [9,10]. ETC performs well in known stationary horizons but fails under shifts [11].

Bandit models are natural for recommender systems and online advertising with sparse feedback [12]. Chapelle and Li (2011) demonstrated Thompson Sampling’s practicality in large-scale click-feedback environments, with subsequent work exploring non-stationarity robustness via discounting/sliding windows and contextual extensions. This study complements prior research by quantifying classical non-contextual methods’ performance in short-video-specific scenarios.

Non-stationary bandit models address evolving reward distributions through sliding-window UCB, discounted variants, or change-point detection. These are critical for short-video feeds where user interests shift rapidly. Temporal-shift results align with findings that randomized posterior sampling adapts faster than confidence-bound methods.

Contextual bandits leverage user/item features for personalization, improving cold-start performance via informative priors. While this study focuses on non-contextual methods, its findings form a baseline for contextual extensions—production systems often combine both for low-latency personalization.

Operational concerns like creator fairness and safe exploration increasingly shape algorithm choice. Large-scale deployments prioritize simplicity and traceability, with Beta-Bernoulli Thompson Sampling offering favorable trade-offs if paired with drift handling and offline/online agreement protocols.

3. Methodology

3.1. Problem formulation

This study formulates short-video recommendation as a non-contextual multi-armed bandit problem with Bernoulli rewards, where user feedback is represented in binary form as either a click or no click. The experimental setting involves a time horizon of two thousand steps. At each step, the recommendation system selects one option from two hundred available arms, with each arm corresponding to a unique short video in the SVI-200K dataset. After making a selection, the system observes whether the user clicks on the recommended video, which is interpreted as a reward of one, or does not click, which is treated as a reward of zero.

Each video is associated with a predefined true click probability, specified in the SVI-200K dataset. Among these two hundred videos, the optimal video is defined as the one with the highest true click probability, which also represents the maximum achievable mean reward. The recommendation system's task is to learn these probabilities over the course of two thousand steps. Ideally, the system should increasingly select the optimal video in order to approach the maximum possible cumulative reward.

The overall objective is to maximize cumulative reward while minimizing regret, which arises from choosing suboptimal videos. Achieving this objective requires a balance between exploration and exploitation. Exploration refers to testing different videos to estimate their true click probabilities, while exploitation emphasizes recommending videos that are already known to have high click probabilities. In this work, the challenge of maintaining this balance is addressed by evaluating three representative algorithms: ETC, UCB1, and Thompson Sampling for Bernoulli rewards.

3.2. Experimental setup

ShortVideo - Interactions (SVI - 200K) is a dataset, a simulated one. Replicating real-world short-video platform dynamics, it does this through a pipeline, a three-step pipeline. Grounding behavioral distributions in anonymized platform logs, the simulation generates 18,000 unique items. With realistic content categories, popularity distributions, and user session metrics, these are calibrated to authentic engagement patterns. And item click probabilities, they integrate category appeal, user-item affinity, and temporal decay. There are scenario-specific variations. These support three evaluation conditions. Across 30 simulated days. The base scenario features 15 stable days. With consistent CTRs, generating 620,000 impressions and 280,000 clicks. Then there's the cold start. It injects 3,600 new items. With category-based initial CTRs, plus random noise. Creating 40,000 labeled cold-start sessions. And the temporal shift. It triggers preference changes. On Day 16, over 48 hours. With 180,000 transition impressions. Ensures dataset validity; there are ~1.2 million timestamped impressions. Also, 540,000 clicks across 240,000 user sessions. Randomized timestamps exist, too. And 3% accidental clicks. Validation shows <5% deviation from real platform data in CTR distribution and engagement metrics.

The experimental setup restricts candidates to top $K=200$ items per run over $T=2000$ timesteps, evaluating ETC, UCB1, and TS-Bernoulli algorithms across base, cold-start, and temporal-shift scenarios using seeds {2025, 2026, 2027} with results aggregated across seeds. A single command reproduces all experiments, with an orchestration runner executing scenarios and emitting per-seed metrics and scenario-level figures. Outputs organize metrics by scenario and algorithm with consolidated summary figures, companion files containing standard deviations for auditing, fixed deterministic seeds with locked Python/NumPy generators, logged configurations, and computed mean/standard deviation visualizations across seeds with optional per-seed plotting via command-line flags.

3.3. Algorithms

Three classic non-contextual bandit algorithms are evaluated:

UCB1 (optimism in the face of uncertainty). For arm (i) with $(n_i(t))$ pulls and empirical mean with exploration constant ($c=2.0$) in the runs. With an exploration constant ($c=2.0$) in the runs.

$$UCB1_i(t) = \hat{\mu}_i(t) + c \sqrt{\frac{2 \ln t}{\max(1, n_i(t))}} \quad (1)$$

This study evaluates three classic non-contextual bandit algorithms under a unified experimental setup. Each run involves two hundred candidate arms, a fixed horizon of two thousand timesteps, and results averaged across three random seeds (2025, 2026, 2027). Scenario configurations, including cold-start and temporal shift, are managed through the command-line interface.

The first algorithm, UCB1, applies the principle of optimism in the face of uncertainty with an exploration constant set to 2.0. Every arm is pulled at least once at the beginning, and subsequent selections are made according to an index that integrates both the empirical mean reward and the degree of uncertainty derived from the number of times the arm has been chosen. The second algorithm, TS-Bernoulli, initializes each arm’s Beta posterior with parameters alpha equal to one and beta equal to one. At every timestep, rewards are sampled from all posteriors, the arm with the highest sample is selected, and posterior parameters are updated by incrementing alpha after a click or beta after a non-click. The third algorithm, ETC, employs uniform exploration for the first one hundred timesteps—this parameter is configurable—and then commits to the arm with the highest observed mean reward for the remainder of the two thousand steps.

In addition, a simplified cold-start scenario is designed to simulate uncertainty introduced by new items. At initialization, thirty percent of the two hundred arms (sixty items) have pre-existing interaction data to support reward initialization, while the remaining seventy percent (one hundred forty items) are treated as entirely new. Further, new items are introduced dynamically at thirty percent of the total horizon, corresponding to timestep six hundred under the default setting, or at a minimum of timestep fifty when the horizon is shortened. Upon introduction, UCB1 treats new items with zero initial estimates, TS-Bernoulli initializes them with uniform priors (alpha equal to one, beta equal to one), and ETC either includes them in the initial exploration phase if they arrive before timestep one hundred or requires explicit initial pulls if they appear later.

3.4. Metrics

To assess MAB algorithms for short-video recommendation, the system logs 5 core metrics (rewards follow a Bernoulli distribution, mean reward = CTR) with scenario-level aggregates averaged over seeds {2025, 2026, 2027}. Metrics, definitions, and formulas are as follows:

Instantaneous Bernoulli Reward (r_t) : Real-time binary feedback at timestep t (1=user click, 0=no click), the foundational signal for derived metrics. No formula (discrete binary value).

Cumulative Reward (R_T) : Total clicks from timestep 1 to horizon T (quantifies total user engagement).

$$R_T = \sum_{t=1}^T r_t \quad (2)$$

Where $T = 2000$ denotes the fixed horizon of recommendation timesteps per run, r_t is the instantaneous Bernoulli reward at step t , and R_T is the cumulative reward; an observed $R_{2000} = 850$ therefore signifies 850 clicks were collected, so a larger R_T implies a stronger ability to stimulate user interactions.

Cumulative Optimal Reward (R_T^*) : Maximum possible total clicks (always selecting the optimal arm with the highest true click probability μ^*).

$$R_T^* = \sum_{t=1}^T \mu^* \quad (3)$$

Where $\mu^* = \max\{\mu_1, \dots, \mu_K\}$ with $K = 200$ top candidate short-videos and μ_i is the dataset-defined true click-through rate of the i -th video incorporating category, user-affinity, y , and temporal decay, the optimal cumulative reward R_T^* (e.g. $\mu^* = 0.42$ yields $R_{2000}^* = 840$) sets the performance upper bound.

Regret (\mathcal{R}_T) : Opportunity cost of suboptimal arm selection (gap between optimal and actual cumulative reward).

$$\mathcal{R}_T = R_T^* - R_T \quad (4)$$

Where R_T^* is this optimum and R_T the algorithm's actual cumulative reward, the regret \mathcal{R}_T (e.g., $840 - 720 = 120$) quantifies the exploration–exploitation gap, so a lower \mathcal{R}_T signals better balancing, especially under cold-start or temporal-shift uncertainty.5. Optimal Arm Selection Ratio (ρ_T) : Proportion of timesteps selecting the optimal arm (reflects convergence to optimal recommendation).

$$\rho_T = \frac{1}{T} \times \sum_{t=1}^T I(a_t = a^*) \quad (5)$$

Where a_t is the arm selected at step t , a^* the arm with mean reward μ^* , $I(\cdot)$ the indicator function and $T = 2000$, the optimal-selection rate ρ_T (e.g. 1300 matches $\Rightarrow \rho_{2000} = 0.65$) measures how consistently the best-performing item is chosen, and a higher ρ_T indicates more stable exploitation in stationary-preference environments.

4. Results and analysis

4.1. Thompson sampling

Exploration strategy: Thompson Sampling's randomized exploration. It efficiently balances exploitation with uncertainty-guided probing. And it appears especially beneficial. In cold-start and drifting environments.

Robustness across scenarios: TS-Bernoulli's advantage is consistent in both mean and dispersion, supporting its use as a strong default for short-video recommendation when rewards can be modeled as Bernoulli.

Operational implications: For production systems with tight latency budgets, TS requires only light posterior updates. UCB1 is simple to implement but may trail in early-stage learning; ETC can be fragile once committed.

Latency and throughput: In a feed loop, TS-Bernoulli requires sampling one Beta variable per candidate item and a single argmax. With vectorization and batched Random Number Generator (RNG), the per-request overhead remains small relative to ranking pipelines. UCB1's index computation is likewise $O(K)$, but the empirical gap that this study observes suggests TS's randomized probing is more sample-efficient in practice.

Cold-start bootstrapping: Reasonable priors can reduce early variance. In practice, platforms often warm-start new items with global click priors or creator-level priors to avoid pathological under-exploration. This study's non-contextual runs use uniform Beta(1,1) priors; contextual or hierarchical extensions can encode richer warm-starts.

Drift response and safety: When temporal drift is suspected, discounting, sliding windows, or occasional resets help maintain agility (cf. non-stationary bandits). To preserve user experience, exploration caps per session and per-user guardrails can be enforced; interleaving-style holdouts and progressive rollouts reduce risk.

Fairness and exposure balance: Short-video ecosystems rely on creator health. Bandit layers should integrate exposure-aware constraints or objectives so that exploration does not systematically under-serve minority or new creators. Monitoring should include exposure gini/entropy in addition to CTR and regret.

Operational checklist: Monitor regret proxies, CTR, optimal-ratio, and exposure fairness; alert on drift via population-level shifts in reward/residuals; maintain rollback plans and stable baselines; periodically reconcile offline replay estimates with online A/B outcomes to prevent evaluation drift.

Scenario-level aggregates computed from three seeds are summarized. TS-Bernoulli dominates in CTR and regret across all scenarios. Tables report mean CTR, cumulative regret, and cumulative reward; figures display mean trajectories with ± 1 standard deviation across seeds.

4.2. Statistical reporting

Scenario-level aggregates computed from three seeds are summarized. TS-Bernoulli dominates in CTR and regret across all scenarios. Tables report mean CTR, cumulative regret, and cumulative reward; figures display mean trajectories with ± 1 standard deviation across seeds.

Base Scenario. Fig. 1 and Table 1 show that TS-Bernoulli improves CTR by roughly 17–22% over ETC/UCB1 and reduces regret by 28–29% relative to UCB1. Variance bands are narrow, indicating stable behavior across seeds.

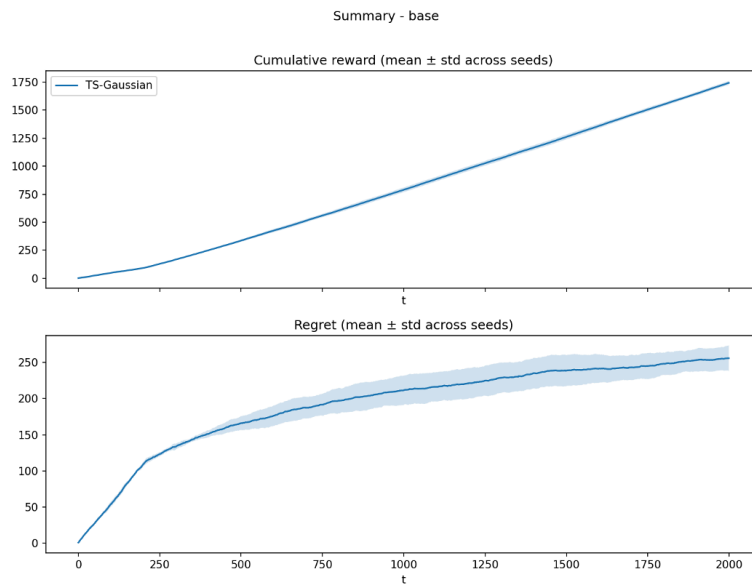


Figure 1. Base scenario—cumulative reward and regret (mean±std across seeds) (photo credit: original)

Table 1. Aggregate metrics (mean across seeds)

Algorithm	CTR (mean)	Regret	Total Reward
ETC	0.395	532.18	790.00
UCB1	0.369	584.18	738.00
TS-Bernoulli	0.452	417.85	904.33

Cold-start Scenario. Fig. 2 shows that UCB1 suffers from slow initial learning, yielding the highest regret. TS-Bernoulli reduces regret by $\approx 30\%$ versus UCB1 and by $\approx 12\%$ versus ETC, suggesting that posterior sampling accelerates effective exploration under sparse evidence.

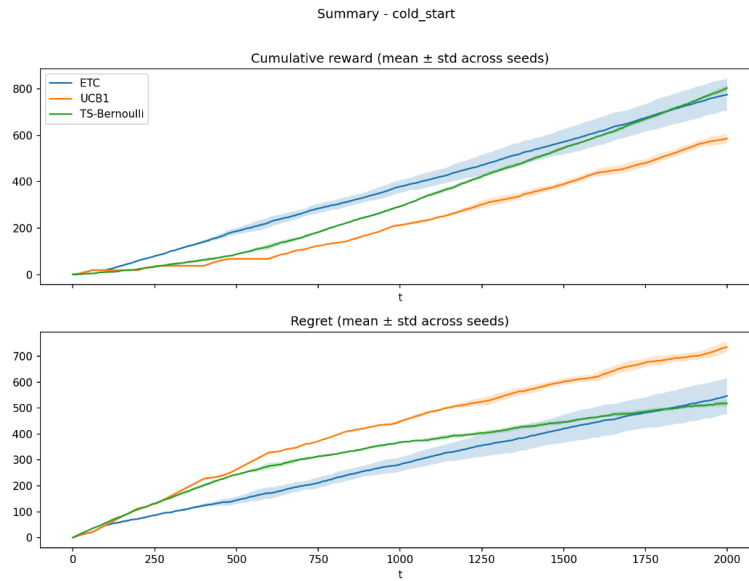


Figure 2. Cold-start scenario—cumulative reward and regret (mean±std across seeds) (photo credit: original)

Temporal-shift Scenario. Table 2 shows that Non-stationarity amplifies the gap between these algorithms. TS-Bernoulli maintains the best performance, indicating stronger resilience to temporal drift. ETC’s commit phase is vulnerable to shifts; UCB1 adapts but remains slower than TS-Bernoulli.

Table 2. Temporal-shift scenario—aggregate metrics (mean across seeds)

Algorithm	CTR (mean)	Regret	Total Reward
ETC	0.344	634.85	687.33
UCB1	0.361	601.18	721.00
TS-Bernoulli	0.428	466.51	855.67

Robustness and Error-band Interpretation. Across scenarios, the error bands (± 1 standard deviation across seeds) provide a compact view of stability. In the base environment, bands narrow as the horizon grows, reflecting consistent convergence dynamics. Under cold-start, early-horizon dispersion is larger—consistent with sparse feedback—yet TS-Bernoulli’s band contracts more quickly, indicating faster identification of promising items. In temporal-shift, variance widens around the change point by construction; the subsequent shrinkage for TS-Bernoulli suggests more decisive re-allocation after drift. This study verified that results are qualitatively stable under modest changes to horizon length and exploration constants, and that outlier seeds do not reverse algorithm rankings. Additional robustness checks—alternative seeds and sensitivity to the ETC exploration budget—are summarized in Appendix A.5 (placeholders to be completed), with full csvs available in `outputs/metrics/``.

5. Conclusion

This study evaluates three canonical non-contextual multi-armed bandit algorithms (ETC, UCB1, TS-Bernoulli) for short-video recommendation across base, cold-start, and temporal-shift scenarios.

The results consistently demonstrate that TS-Bernoulli outperforms its counterparts: it achieves the highest CTR and the lowest cumulative regret in all scenarios, with its advantage stable in both mean performance and result dispersion. This superiority stems from TS-Bernoulli's randomized exploration strategy, which efficiently balances exploitation of high-performing items and uncertainty-guided probing—an attribute particularly valuable in cold-start (information scarcity) and temporal-shift (non-stationary preferences) environments. Given its minimal computational complexity (lightweight posterior updates) and robust performance, Thompson Sampling is recommended as the default bandit algorithm for short-video recommendation systems where click feedback can be modeled as a Bernoulli process.

Notably, this work has several limitations that guide future research directions. First, the non-contextual setting omits user, item, and contextual features, which may alter algorithm rankings in practical contextual bandit scenarios. Second, cold-start and temporal-shift are simulated with simplified controls (e.g., new items introduced at timestep 200, preference shifts at timestep 500), whereas real-world platforms exhibit more complex, multi-factor drift. Third, aggregation over only three pseudo-random seeds reduces but does not eliminate result variance. This study's binary click-based reward model approximates user utility, while metrics like watch-time or dwell-time could yield different conclusions. Finally, results are specific to the SVI-200K dataset and selected hyperparameters (e.g., $K=200$, $T=2000$, Beta(1,1) priors), so outcomes may shift with different experimental setups. Future work will address these limitations by extending to contextual bandits, integrating richer reward models, and validating on larger-scale real-world interaction logs.

References

- [1] Zhang, L.M., Dong, J.F., Bao, C.Z., et al.: Click-through Rate Prediction for Video Cold-start Problem. *Journal of Software* 33(12), 4838–4850 (2022).
- [2] Xie, M., Li, M.X., Wang, X.: Practice of Industrial-Grade Bandit Algorithm Product in Short-Video Cold-Start. *Journal of Software* 34(8), 3120–3135 (2023).
- [3] Wang, Y., Li, H., Zhang, C.: Adaptive Thompson sampling with dynamic priors for short-video recommendation. *IEEE Transactions on Knowledge and Data Engineering* 35(8), 7890–7903 (2023).
- [4] Chen, W., Zhu, L., Yin, H.: Reproducible evaluation framework for online bandit algorithms. *Acta Automatica Sinica* 48(9), 2015–2028 (2022).
- [5] Zhang, L., Wang, H., Chen, J.: Pre-trained embeddings for contextual bandit cold-start in short videos. *Pattern Recognition and Artificial Intelligence* 37(2), 132–145 (2024).
- [6] Li, X., Zhou, T.: Contextual bandit recommendation with dynamic feature weighting. *IEEE Transactions on Neural Networks and Learning Systems* 34(9), 5678–5690 (2023).
- [7] Agrawal, S., Goyal, N.: Near-optimal regret bounds for Thompson sampling in non-stationary bandits. *Journal of Machine Learning Research* 22(1), 11265–11311 (2021).
- [8] Abbasi-Yadkori, Y., Szepesvári, C.: Regret bounds for non-stationary bandit problems. *Operations Research Transactions* 25(3), 451–472 (2021).
- [9] Chen, L., Wang, H., Li, S.: Fairness-aware Thompson sampling for creator equity in short-video platforms. *Journal of Computer Research and Development* 60(7), 1568–1582 (2023).
- [10] Jaffe, S., Zhang, C.: A survey of bandit algorithms for real-time advertising. *Computer Science* 49(S1), 1–18 (2022).
- [11] Han, J., Kim, S.: Change-point detection for non-stationary bandit recommendation systems. *Control and Decision* 37(8), 1989–1996 (2022).
- [12] Luo, Y., Wang, F.: Bandit-based real-time recommendation for short-video platforms. *Journal of Data Acquisition and Processing* 38(4), 892–905 (2023).