Extended Abstract

Motivation Recommender systems enable platforms like streaming services, online shopping, and social media to suggest relevant content using past interaction data. However, when a new user arrives, the system faces the cold-start problem: it has little preference data, so popular-item exploitation may fail for niche users, while pure exploration may produce irrelevant recommendations. We study cold-start recommendation as a finite-horizon contextual bandit problem, aiming to maximize reward in the first few interactions where user data is scarce. We compare greedy, Bayesian, Meta-RL, and constrained bandit policies on MovieLens-1M to evaluate exploration strategies.

Method We evaluate six policies in a shared cold-start contextual bandit environment. A random policy serves as a simple lower-bound baseline. Greedy Collaborative Filtering (CF) serves as a strong exploitation baseline. Greedy CF trains an SVD matrix-factorization model on warm-user ratings, then updates a cold user's latent preference vector and always selects the item with the highest predicted rating.

We compare Greedy CF against four exploration-based methods. Neural Linear Thompson Sampling maintains a Bayesian linear model over fixed sentence-transformer item embeddings and samples a plausible user preference vector at each step. Hybrid Thompson Sampling instead replaces content embeddings with the SVD latent factors used by Greedy CF. RL^2 trains a recurrent policy over warm-user episodes so that its hidden state summarizes the current user's interaction history and learns an exploration strategy directly from data. Finally, the Constrained Bandit uses a LinUCB-style policy, but only explores when a lower-confidence estimate remains in a safe non-personalized baseline; otherwise, it falls back to the baseline recommendation.

Implementation We evaluate all methods on MovieLens-1M, keeping the users with at least 25 ratings and sorting each user's history chronologically to simulate cold-start recommendation. Users are split into warm and cold sets. Warm users are used for training priors and model parameters, while cold users are held out for evaluation. Each movie is represented by a 402-dimension embedding from title, release year, and genres using `all-MiniLM-L6-v2`, while user metadata is encoded using gender, age group, and occupation. We evaluate under two protocols: a ranking setting where policies order a fixed pool of 20 items, and a selection setting where policies choose 20 recommendations from a larger cold-user history. Metrics used are R@1, R@5, NDCG@5, and cumulative reward.

Results At $T = 20$, Greedy CF performed best under both protocols, achieving NDCG@5 of 0.882 in the ranking setting and 0.896 in the selection setting. For exploration methods, RL^2 achieved the strongest early ranking performance in the selection setting (R@1 of 4.48, NDCG@5 of 0.892), while Hybrid TS achieved the highest cumulative reward among exploration methods (84.13 vs. 85.51 for Greedy CF). Long-horizon results show that exploration gains compound over time, with the cumulative reward gap between Greedy CF and Hybrid TS decreasing from 2.2% at $T = 20$ to 0.6% at $T = 100$.

Discussion The main takeaway is that exploration is not inherently better suited for cold-start recommendation. When the warm-user population provides a strong collaborative prior, greedy exploitation remains a strong method. Exploration becomes useful when that prior is weak, noisy, or mismatched, and our results suggest that feature representation also matters since Hybrid TS consistently outperforms content-based NLTS.

Conclusion Our results show that with a strong collaborative prior, greedy exploitation is surprisingly difficult to beat at the standard $T = 20$ cold-start horizon. Hybrid Thompson Sampling comes closest, especially at longer horizons as the Bayesian posterior sharpens. Ablation studies confirm that exploration methods are able to outperform greedy baselines when the prior is weakened through reduced data, noise, or population mismatch. When rich collaborative data exists, the marginal value of exploration is small at short horizons. When the prior is weak or miscalibrated, Bayesian approaches provide meaningful gains.

Bandits for Cold-Start: Exploration vs. Exploitation in Recommendation

Eva Casto
Department of Computer Science
Stanford University
ecasto@stanford.edu

Anastasiya Masalava
Department of Computer Science
Stanford University
masalava@stanford.edu

Ela Sigin
Department of Computer Science
Stanford University
elasigin@stanford.edu

Abstract

The cold-start problem is a central challenge to recommender systems because there is scarce information on new user preferences, making it difficult to balance accurate early recommendations with exploration for preference learning. We study cold-start recommendation as a finite-horizon contextual bandit problem on MovieLens-1M, evaluating greedy collaborative filtering, Neural Linear Thompson Sampling, Hybrid Thompson Sampling, Meta-RL via RL^2, and a constrained bandit under ranking and selection protocols. At the $T = 20$ horizon, Greedy CF performs the best with an NDCG@5 of 0.882 in the ranking setting and 0.896 in selection. Among exploration methods, Hybrid TS obtains the highest cumulative reward of 84.13, still beneath Greedy CF. However, long-horizon and ablation results show that exploration becomes more advantageous when the prior is weak, noisy, or mismatched, revealing cases where strong Greedy CF is less effective.

1 Introduction

Recommender systems are information filtering tools that enable platforms like streaming services, online shopping, and social media to suggest relevant content to users. These systems rely on past interaction data – including ratings, clicks, watch history, or purchases – to infer what a user is likely to prefer. Accurate recommendations are important because they help to increase user satisfaction, thereby increasing engagement and revenue.

A central challenge in recommendation is the cold-start problem. When a new user arrives, the system has little to no activity or preference data for that user, making it difficult to personalize recommendations. A purely exploitative system may recommend popular items, but this can fail for users with niche preferences. A purely exploratory system may collect useful information quickly, but this has the risk of recommending irrelevant items. Thus, cold-start recommendation creates an exploration-exploitation tradeoff, learning user preferences quickly while still keeping early recommendations relevant and useful.

We study cold-start recommendation as a finite-horizon contextual bandit problem. At each step, the policy recommends an item, observes a rating as reward, then updates its knowledge about the user's preferences. The goal is to maximize reward over the first few interactions, where user-specific data is scarce. This is especially challenging because early recommendations must satisfy the user now and help the system learn what to recommend next.

We compare several strategies for cold-start recommendation on MovieLens-1M. We implement a random baseline, a greedy collaborative-filtering baseline, Neural Linear Thompson Sampling, Hybrid Thompson Sampling, Meta-RL via RL^2, and a constrained bandit with a relevance guarantee. Our work aims to make the following contributions:

1. A two-protocol evaluation framework that separates ordering ability (ranking) from discovery ability (selection), revealing when exploration is preferred over exploitation.
2. An empirical comparison of greedy, Bayesian, meta-RL, and constrained bandit policies, showing that strong collaborative priors make greedy exploitation surprisingly difficult to beat at short horizons.
3. An ablation analysis identifying the conditions under which exploration methods gain a clear advantage over greedy exploitation.

2 Related Work

Prior work on cold-start recommendation has explored how systems can make useful recommendations when little or no user history is available. A foundational line of work in this area involves recommendation via contextual bandits, where the system observes user or item features, recommends an item, and receives feedback for that action. Li et al. (2010) introduced LinUCB in the setting of personalized news recommendation and helped establish contextual bandits as a standard framework for exploration-exploitation in recommender systems. However, standard LinUCB begins with an uninformative prior – the algorithm starting with no knowledge of the user – and therefore may spend many early interactions exploring, which is not ideal in cold-start settings.

Neural and warm-started bandits extend upon the idea of linear bandits, allowing for richer representations. Riquelme et al. (2018) ran a large benchmark of Bayesian neural network methods for Thompson sampling, discovering that a frozen pretrained network with Bayesian linear regression on the last layer alone outperforms more advanced posteriors. This motivates our Neural Linear Thompson Sampling model, which uses fixed sentence-transformer item embeddings and maintains a Bayesian linear posterior over each cold user's preference vector. In addition, neural contextual bandit work has studied scalable exploration in recommender systems. Zhu and Roy (2023) propose a neural contextual bandit architecture for recommendation that combines deep representation learning with Thompson sampling. Su et al. (2024) present another relevant finding, showing that neural linear bandits can support industrial-scale recommender exploration, but not in a cold-start setting. These works demonstrate the usefulness of neural bandits, but they do not isolate the cold-start window, which is our main evaluation target.

An alternative approach that has been studied is cold-start recommendation through meta-learning. MeLU (Lee et al., 2019) uses model-agnostic meta-learning to estimate new user preferences from a small number of items, making it very relevant to cold-start recommendation. NICF (Zou et al., 2020) is a closely related work to our Meta-RL implementation, since it also treats cold-start recommendation as an interactive sequential decision problem and learns an exploration policy from user feedback rather than relying on a fixed bandit rule. However, NICF represents the exploration policy with stacked self-attention and trains it with Q-learning, while our RL^2 approach modernizes this by using a recurrent exploration policy over warm-user episodes.

Finally, prior work in conservative or safe exploration motivates our constrained bandit method. Conservative contextual linear bandits require a learning algorithm to remain competitive with a baseline policy while still exploring. Kazerouni et al. (2017) formalize this idea for contextual linear bandits and propose a conservative LinUCB algorithm that maintains performance above a fixed percentage of a baseline policy. This directly motivates our constrained bandit method, which uses a baseline policy as a relevance floor so that exploration is only taken when it is estimated to be sufficiently safe.

3 Methods

We evaluate six policies in a shared contextual bandit framework: a **Random** lower-bound baseline, a **Greedy Collaborative Filtering** exploitation baseline, **Neural Linear Thompson Sampling**, **Hybrid Thompson Sampling**, **Meta-RL** (RL^2), and a **Constrained Bandit**.

3.1 Random Baseline

The random baseline selects uniformly at random from a set of items not yet chosen to recommend to a user. This baseline does not require any training and serves as a simple lower-bound reference for other policies.

3.2 Greedy Collaborative Filtering Baseline

Greedy Collaborative Filtering is our strong exploitation baseline. Despite performing no exploration, it benefits from a very well-calibrated SVD prior trained on warm users and adapts rapidly via ridge regression (Koren et al., 2009).

In this approach, we construct a warm-user rating matrix $R \in \mathbb{R}^{n_{\text{warm}} \times n_{\text{items}}}$, subtract per-item mean ratings from the observed entries, and apply an SVD to obtain item latent factors $Q \in \mathbb{R}^{n_{\text{items}} \times k}$ ($k = 50$) and per-item mean ratings $\boldsymbol{\mu}$. The global prior θ_0 is set to the mean warm-user vector in latent space.

During the evaluation stage, for each cold user, the preference vector θ_t is updated after every observed rating via ridge regression. Specifically:

$$\theta_t \;=\; \arg\min_{\theta} \sum_{(a,\,r)\,\in\, s_t} \left(q_a^\top \theta + \mu_a - r\right)^2 \;+\; \lambda \left\|\theta - \theta_0\right\|^2, \tag{1}$$

where s_t is the set of (item, rating) pairs observed so far for the current user, $q_a \in \mathbb{R}^k$ is the latent factor of item a, μ_a is the mean rating of item a over warm users, r is the observed rating, and $\lambda > 0$ is the ridge regularisation coefficient.

At each step the Greedy Collaborative Filtering baseline selects the highest-scoring candidate:

$$a_t \;=\; \arg\max_{a}\; q_a^\top \theta_t + \mu_a. \tag{2}$$

3.3 Neural Linear Thompson Sampling (NLTS)

Neural Linear Thompson Sampling (NLTS) addresses the core challenge of the cold-start problem: early in the episode for a cold user, the agent has little prior knowledge about the user's preferences and must decide whether to select its current best estimate (exploitation) or explore the uncertainty (exploration). Greedy CF ignores this uncertainty entirely, while NLTS quantifies it through a Bayesian posterior and uses it to drive exploration.

We model reward as linear in the item embedding, $r = \phi(x)^\top \theta + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, and maintain a Gaussian posterior over the user preference vector θ. At the beginning, the prior (μ_0, Λ_0) is initialised by pooling all warm-user ratings into a single ridge regression, giving an informed starting point. After observing a rating r_t for item a_t, the posterior (μ_t, Λ_t) is updated using the following formula:

$$\Lambda_t = \Lambda_{t-1} + \tfrac{1}{\sigma^2}\phi_t \phi_t^\top, \tag{3}$$

$$b_t = b_{t-1} + \tfrac{1}{\sigma^2} r_t \phi_t, \tag{4}$$

$$\mu_t = \Lambda_t^{-1} b_t, \tag{5}$$

where $b_t \in \mathbb{R}^d$ is the information vector, initialised as $b_0 = \Lambda_0 \mu_0$.

At each step, the policy samples $\tilde{\theta}_t \sim \mathcal{N}(\mu_{t-1}, \Lambda_{t-1}^{-1})$ and selects $a_t = \arg\max_a\, \phi(a)^\top \tilde{\theta}_t$. At the beginning, the posterior is broad and the agent explores freely; as the agent accumulates more ratings, the policy starts to exploit what it has learned (Riquelme et al., 2018; Thompson, 1933).

3.4 Hybrid Thompson Sampling (HTS)

Hybrid Thompson Sampling follows the same Bayesian framework as NLTS, but it replaces the raw content embeddings $\phi(x) \in \mathbb{R}^d$ with the k-dimensional ($k = 50$) SVD latent factors $q_a \in \mathbb{R}^k$ learned

by Greedy CF. This change allows HTS to explore the action space while having the same expressive feature space as the greedy baseline.

The prior (μ_0, Λ_0) is initialised from the empirical covariance of warm-user latent vectors: $\Lambda_0 = \lambda \cdot \Sigma_{\text{warm}}^{-1}$, where Σ_{warm} captures the directions of greatest user variability. The posterior update and action selection follow Equations 3–5. At each step the policy samples $\tilde{\theta}_t \sim \mathcal{N}(\mu_{t-1}, \Lambda_{t-1}^{-1})$ and selects $a_t = \arg\max_a \; q_a^\top \tilde{\theta}_t$.

3.5 Meta-RL via RL^2

The key idea of the Meta-RL via RL^2 is to train a recurrent policy across the warm-user episodes so that the LSTM hidden state h_t serves as an in-context task posterior. This hidden state accumulates the knowledge of the current user's preferences without requiring any parameter updates at test time (Duan et al., 2016).

Architecture At each step t, the policy observes the previous (item, reward) pair as input:

$$x_t = \left[\, \phi(a_{t-1}) \parallel r_{t-1} \,\right] \in \mathbb{R}^{d+1}, \qquad x_0 = \mathbf{0}, \tag{6}$$

and updates its belief via an LSTM and a bilinear scoring rule:

$$h_t = \text{LSTM}(x_t,\, h_{t-1}), \qquad \text{score}(a) = \phi(a)^\top W h_t, \tag{7}$$

where $W \in \mathbb{R}^{d \times d_h}$ is a learned matrix and d_h is the LSTM hidden dimension.

Training The policy is trained using REINFORCE across warm-user episodes, where each user corresponds to one task. To reward high-rated recommendations at the very beginning, which is important for the cold-start problem, the returns are shaped using the DCG-inspired discount, which penalizes the policy for deferring good recommendations to later steps:

$$G_t = \frac{r_t}{\log_2(t+2)}, \tag{8}$$

Test time For each cold user at test time, the hidden state is set to $h_0 = \mathbf{0}$ and actions are selected greedily using $a_t = \arg\max_a \; \text{score}(a)$.

3.6 Constrained Bandit with Relevance Guarantee

While Thompson Sampling and RL^2 are allowed to explore freely, unrestricted exploration during cold start risks recommending irrelevant items when users are least patient for poor suggestions. The Constrained Bandit with Relevance Guarantee addresses this by utilizing a LinUCB-style optimistic policy with a hard relevance floor (Li et al., 2010). This strategy allows the agent to explore only when the proposed action is likely to remain competitive with a safe non-personalised baseline.

Action selection The policy maintains a Bayesian linear estimate $\hat{\theta}_t$ of the current user's preference vector. At each step it proposes the most optimistic candidate:

$$a_t^\star = \arg\max_a \left(\phi(a)^\top \hat{\theta}_t \;+\; \beta \sqrt{\phi(a)^\top A_t^{-1} \phi(a)} \right), \tag{9}$$

where A_t is the precision matrix and $\beta > 0$ controls the degree of optimism.

Safety constraint Before choosing the $a_t^\star$, the policy validates a cumulative relevance constraint against a non-personalised baseline with expected score μ_{base}:

$$R_{1:t-1} + \text{LCB}(a_t^\star) \;\geq\; \alpha \left(B_{1:t-1} + \mu_{\text{base}}(a_{\text{base}}) \right), \tag{10}$$

In this formula, the $R_{1:t-1}$ is the cumulative reward collected so far, $B_{1:t-1}$ is the expected reward the baseline would have collected, and $\text{LCB}(a_t^\star) = \phi(a_t^\star)^\top \hat{\theta}_t - \beta_{\text{safe}} \sqrt{\phi(a_t^\star)^\top A_t^{-1} \phi(a_t^\star)}$. The parameter $\alpha \in [0, 1]$ controls the trade-off: smaller α allows more exploration, while larger α makes recommendations closer to the baseline.

Fallback If the safety constraint (Equation 10) is violated, the policy falls back to the action that the baseline would have taken (a_{base}). The approach guarantees that the total reward never falls below an α-fraction of what the non-personalised baseline would have achieved, which serves as a formal relevance floor (Kazerouni et al., 2017).

4 Experimental Setup

4.1 Dataset

We evaluate our methods on the MovieLens-1M dataset (Harper and Konstan, 2015), a common benchmark for recommender systems. We retain only users with at least 25 ratings to ensure that each cold user has sufficient interactions to construct a meaningful evaluation sequence of $T = 20$ steps with held-out ratings for ground-truth reward. Our final dataset contains 991,077 ratings across 5,624 users and 3,702 items, with an average of 176.2 ratings per user.

Each user's history is ordered chronologically to simulate the cold-start problem. Users are divided into **warm** (70%, 3,936 users) and **cold** (30%, 1,688 users) sets. Warm users are used only for training priors and model parameters, while cold users are held out for evaluation. The agent has no prior knowledge of the cold users' preferences, which simulates cold-start conditions and ensures fair evaluation.

4.2 Feature Representations

Item features. Each movie is represented by a 402-dimensional embedding $\phi(x) \in \mathbb{R}^{402}$ obtained by encoding its title, release year, and genres using the `all-MiniLM-L6-v2` sentence transformer (Reimers and Gurevych, 2019; Wang et al., 2020). All movie embeddings are ℓ_2-normalized to ensure that the scoring rule $\phi(a)^\top \theta$ reflects preference alignment rather than embedding scale.

User features. Each user is represented by a 30-dimensional one-hot vector $\psi(u) \in \mathbb{R}^{30}$, which encodes the user's gender, age bracket, and occupation. User features are used only by the Constrained Bandit, while other methods rely only on the item embeddings.

4.3 Evaluation Protocols and Metrics

Ranking protocol ($|A| = 20$). At evaluation time, the policy selects the order in which to recommend a fixed number of items ($T = 20$) drawn from the first 20 interactions of a cold user. Since every policy has to select all 20 items, the cumulative reward is identical between all methods. This protocol measures how well a policy chooses high-rated items early in the episode, but naturally penalizes exploration since there is no benefit to trying uncertain items.

Selection protocol ($|A| \approx 176$). At evaluation time, the agent selects $T = 20$ items from the cold user's full rating history ($\sim$176 items on average). In this protocol, the cumulative reward varies across policies and exploration has a genuine value since the agent must discover which items are good from a much larger candidate pool.

Metrics **R@1** is an average rating of the first recommendation, measuring how well the policy exploits what it knows during the first step. **R@5** is an average rating over the first five recommendations, measuring early-episode quality. **NDCG@5** is a normalized discounted cumulative gain at rank 5, measuring how well high-rated items are concentrated at the top of the recommendation list. **Cumulative reward** is the total rating collected across all $T = 20$ steps. Under the ranking protocol it is fixed across methods; under the selection protocol it varies.

4.4 Ablation Studies

To identify conditions where exploration methods outperform greedy exploitation, we stress-test the prior along three dimensions.

Warm-user fraction: we vary the fraction of warm users for prior training from 2% to 70%. A smaller warm-user fraction results in a weaker, less-calibrated prior which should favor exploration-based methods over greedy exploitation.

Reward noise: we add Gaussian noise $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ to observed ratings, sweeping σ from 0.0 to 1.5. This approach reduces the signal available per observation, testing whether Thompson Sampling's uncertainty handling provides an actual advantage under stochastic feedback.

Wrong-population prior: we train the prior on a mismatched warm population (users aged 50+) and evaluate on a different population (users aged under 25). This tests whether exploration methods are more robust to prior miscalibration than greedy exploitation.

5 Results

We evaluate all six methods under both the ranking and selection protocols at $T = 20$, then extend the selection protocol to $T = 100$ to test whether exploration gains compound over longer horizons. All reported metrics are averaged across 1,688 cold users (894 for the longer horizon experiment, which requires users with at least 100 ratings).

5.1 Quantitative Evaluation

Table 1 presents the main results under both evaluation protocols at $T = 20$.

Ranking Protocol ($\lvert A\rvert = 20$)				**Selection Protocol** ($\lvert A\rvert \approx 176$)				
Method	R@1	R@5	NDCG@5	Method	R@1	R@5	NDCG@5	Cum. Reward
Greedy CF	**4.40**	**4.27**	**0.882**	Greedy CF	**4.49**	**4.43**	**0.896**	**85.51**
Constrained	**4.40**	4.22	0.874	Constrained	**4.49**	4.34	0.883	83.23
RL^2	4.34	4.21	0.870	RL^2	4.48	4.40	0.892	83.61
Hybrid TS	4.26	4.21	0.866	Hybrid TS	4.29	4.29	0.865	84.13
NLTS	4.18	4.11	0.847	NLTS	4.20	4.17	0.846	81.47
Random	3.79	3.76	0.772	Random	3.73	3.71	0.770	74.28

Table 1: Results under both evaluation protocols at $T = 20$ (see Section 4.3 for protocol details). In the ranking protocol cumulative reward is fixed (75.43); in the selection protocol it varies.

5.1.1 Ranking protocol

Under the ranking protocol, where all items must be recommended and only ordering matters, Greedy CF achieves the highest scores across all metrics. This is expected: with a fixed pool of 20 items and no opportunity to discover new ones, exploration offers no benefit. The Constrained Bandit ties Greedy CF on R@1 (4.40), which is expected since at the first step both methods default to the same baseline recommendation 37.8% of the time, effectively defaulting to greedy behavior. RL^2 ranks third, slightly ahead of Hybrid TS, while NLTS trails due to its use of content embeddings rather than SVD factors.

5.1.2 Selection protocol

When the candidate pool expands to the full item catalog ($\sim$176 items), methods must actively identify good items from a much larger set. Greedy CF still leads in cumulative reward (85.51), but the gap narrows. Hybrid TS (84.13) is the closest exploration method, trailing by only 1.6%. RL^2 performs surprisingly well on early metrics (R@1 = 4.48, NDCG@5 = 0.892), ranking second on NDCG@5 and third on R@1. This suggests that the meta-learned policy produces strong initial recommendations, though it accumulates less total reward (83.61) than Hybrid TS over 20 steps. NLTS again underperforms Hybrid TS, confirming that the SVD latent space provides a stronger prior than raw content embeddings for this dataset.

5.1.3 Longer horizon experiment

To test whether exploration gains compound as the Bayesian posterior sharpens, we extend the selection protocol from $T = 20$ to $T = 100$ for the 894 cold users with at least 100 ratings. Figure 1 shows the per-step mean reward for all six methods. Greedy CF dominates during the first $\sim$30 steps, but Hybrid TS increasingly matches or exceeds it at later steps. Over the full 100 steps, Hybrid

TS outperforms Greedy CF at 34 individual steps. This trend accelerates in the later portion of the episode: Hybrid TS wins 10 of the first 50 steps but 24 of the last 50, with 8 out of 10 wins in steps 81–90.

Figure 2 shows cumulative reward as a function of horizon length T. The relative gap between Greedy CF and Hybrid TS shrinks from 2.2% at $T = 20$ to 0.6% at $T = 100$. However, the cumulative reward curves do not cross because Greedy CF's early advantage accumulates. This indicates that exploration does provide compounding returns over longer horizons, but overcoming a strong greedy prior requires either more steps or a weaker prior.

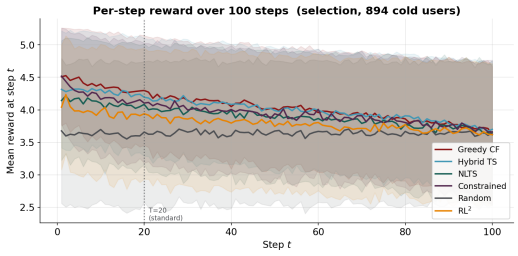


Figure 1: Per-step mean reward over 100 steps (selection protocol, 894 cold users). Shaded regions show ± 1 SD; dashed line marks $T = 20$.

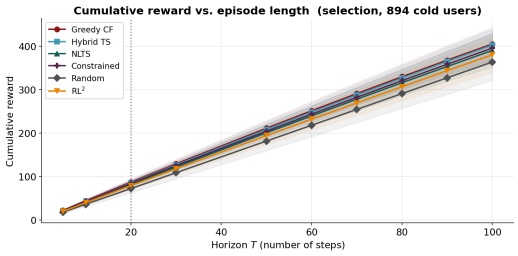


Figure 2: Cumulative reward vs. horizon T (selection protocol, 894 cold users). The Greedy CF – Hybrid TS gap narrows from 2.2% at $T = 20$ to 0.6% at $T = 100$.

5.2 Qualitative Analysis and Ablation Studies

5.2.1 When does exploration win?

To understand why Greedy CF remained competitive with exploration-based methods, we stress-tested the learned prior by changing the fraction of warm-users, increasing reward noise, and training the prior on a mismatched user population.

Warm-user fraction. Figure 3 shows the NDCG@5 gap between Greedy CF and Hybrid TS as the warm-user fraction changes. Positive values indicate that Greedy CF outperforms Hybrid TS, and negative values mean that Hybrid TS outperforms. When the warm-user prior is very weak at only 2% of users, Hybrid TS wins slightly over Greedy CF. However, once the prior is trained on at least 5% of warm users, Greedy CF stays on top and this gap grows. This suggests that exploration is most useful when the collaborative prior is very weak but loses its advantage when the prior is improved.

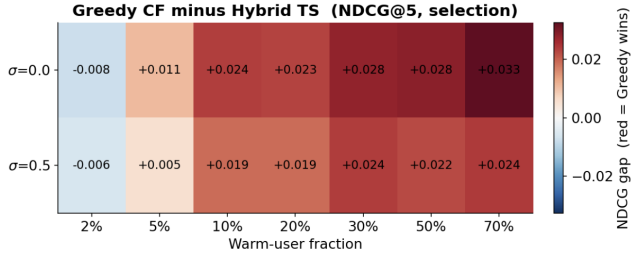


Figure 3: Warm-user fraction ablation under the selection protocol, recording difference in NDCG@5 between Greedy CF and Hybrid TS.

Reward noise. Figure 4 indicates that increasing reward noise causes a decrease in NDCG@5 for all methods under both protocols. Greedy CF remains strongest, and among the exploration methods Hybrid TS consistently performs closest to it. This suggests that reward noise weakens all methods but does not make exploration outperform the strong collaborative prior.

Population mismatch. Figure 5 shows that under the matched prior, Greedy CF performs best for both young and senior users, confirming that the collaborative prior is strong when the warm

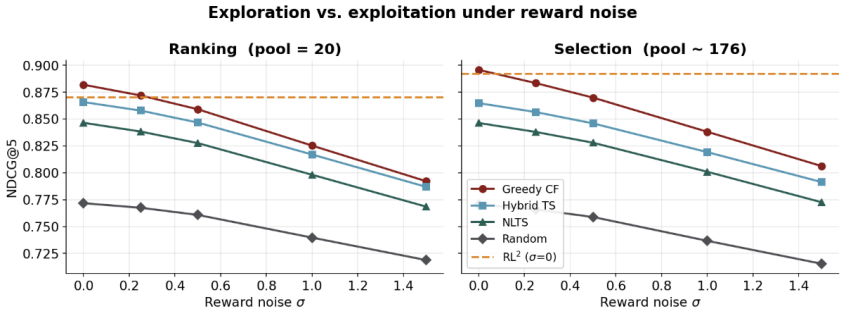


Figure 4: Reward noise ablation under the ranking and selection protocols between Greedy CF, Hybrid TS, NLTS, and Random. We show the original $\sigma = 0$ performance of RL^2 as a reference point.

and cold populations are aligned. However, when trained on the wrong/opposite population prior, Greedy CF performs similarly or worse than exploration methods, losing its advantage. This shows that exploration becomes more valuable when the prior is actively miscalibrated.

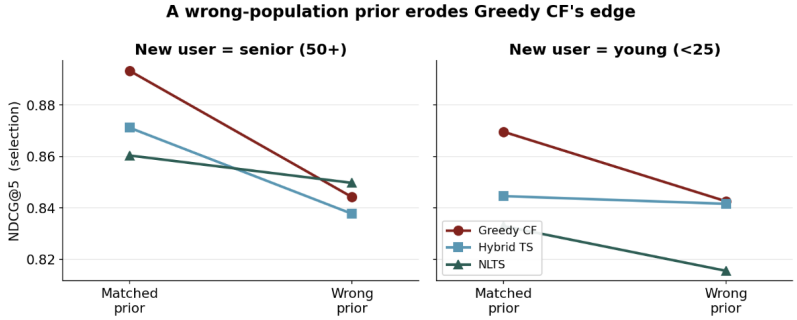


Figure 5: Comparison of NDCG@5 for the matched prior and wrong prior between methods, among senior and young cold users.

5.2.2 Which users benefit from exploration?

We wanted to empirically study how well each method handles mainstream versus niche interest users. To investigate where exploration is most useful, we split cold users into quartiles based on a niche score,

$$\text{niche}(u) = 1 - \rho\big(\text{item\_means}, r_u\big),$$

where ρ is the Spearman correlation between warm-user item means and the user's observed ratings. Quartile 1 contains mainstream users whose tastes align with what warm users rated highly, while Quartile 4 contains niche users whose liked items have low or inverse rank under warm-user popularity. Thus, we hypothesized that Greedy CF's popularity-based prior would work well for Q1 users and be misleading for Q4 users.

Figure 6 shows that Greedy CF holds its advantage overall, with all exploration methods ultimately remaining below Greedy CF (indicated by 0.0 on the y-axis), but the gap narrows as users have increasingly niche interests. This verifies our claim that Greedy CF commits to more popular items.

This trend also explains the behavior of individual exploration methods. RL^2 is the most robust to niche taste, with its gap to Greedy CF narrowing the most percentage-wise and approaching near-parity in Q4. RL^2 appears to generalize across taste profiles because it learned to explore adaptively during meta-training, not relying on a fixed prior. Random also closes the gap significantly because it does not use the popularity prior at all, making it bad for mainstream users and not any worse for niche users. Finally, NLTS closes its gap more than Hybrid TS, suggesting that content embeddings can be more informative for users whose taste does not match the collaborative warm-user population.

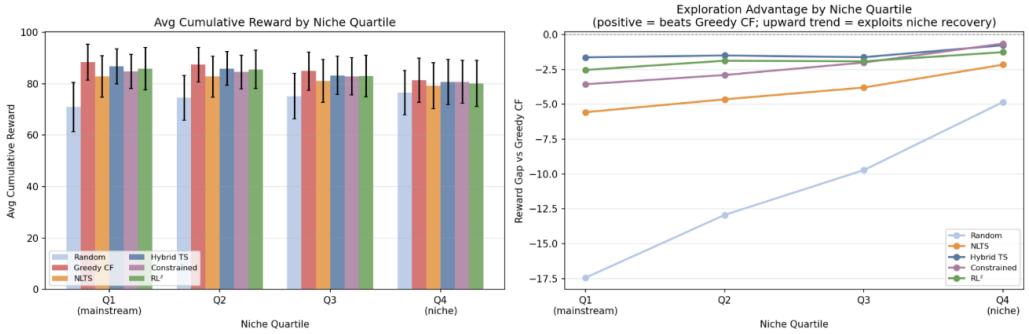


Figure 6: Prior quality case study by niche quartile. To the left compares average cumulative reward by niche quartile. To the right is exploration advantage by niche quartile, shown in terms of the reward gap with greedy CF.

6 Discussion

6.1 The Strength of the SVD Prior

The most consistent finding across our experiments is that Greedy CF is difficult to beat when trained on a large, well-matched warm population. With 3,936 warm users on MovieLens-1M, the SVD prior captures the dominant preference patterns well enough that greedy exploitation is near-optimal for most cold users. This aligns with prior findings that simple matrix factorization with good priors remains a strong baseline in recommendation (Koren et al., 2009).

6.2 When Exploration Helps

Our ablation studies (Section 5.2.1) show that exploration methods gain a clear advantage when the prior degrades. Reducing the warm-user fraction, injecting reward noise, or training on a mismatched population all narrow or reverse the gap between Greedy CF and the Thompson Sampling variants. The longer horizon experiment provides further evidence: as the episode extends to 100 steps, Hybrid TS wins an increasing share of individual steps, particularly in the second half of the episode. This suggests that the Bayesian posterior does sharpen with more observations, and the value of early exploration compounds over time.

6.3 Feature Representation Matters

NLTS consistently underperforms Hybrid TS despite sharing the same Bayesian update rule. The key difference is the feature space: NLTS uses 402-dimensional content embeddings from a sentence transformer, while Hybrid TS operates in the 50-dimensional SVD latent space. The SVD factors, learned from warm-user ratings, directly encode collaborative preference structure. Content embeddings capture semantic similarity between items but do not reflect how users actually rate them. This gap explains why NLTS ranks below Greedy CF even under the selection protocol, where exploration should be most beneficial.

6.4 RL^2 and the Prior Gap

RL^2 shows strong early performance in the selection protocol (R@1 = 4.48, NDCG@5 = 0.892), nearly matching Greedy CF on first-step metrics. The meta-learned LSTM appears to internalize a useful recommendation strategy from warm-user episodes. However, RL^2 cannot access the SVD prior and operates over content embeddings, which limits its cumulative performance. In the longer horizon experiment at $T = 100$, RL^2 ranks fifth out of six methods in cumulative reward (380.36, Figure 2). This suggests that the meta-learned exploration policy, while effective at producing good initial recommendations, cannot substitute for the structured collaborative prior that the Bayesian methods use.

6.5 Constrained Exploration as a Practical Middle Ground

The Constrained Bandit achieves strong early metrics (R@1 = 4.49, tied with Greedy CF) while providing a formal relevance guarantee. Its fallback rate of 37.8% means it defaults to the safe baseline in over a third of steps. This makes it a practical choice for settings where user patience is limited and poor recommendations carry high cost, such as onboarding flows. The tradeoff is a modest reduction in cumulative reward compared to unconstrained methods.

6.6 Limitations

Our study evaluates on a single dataset (MovieLens-1M) with simulated cold-start conditions. Real cold-start settings involve additional challenges such as implicit feedback, non-stationary preferences, and much larger item catalogs. The ranking protocol uses a fixed pool of 20 items, which limits its ability to capture the benefits of exploration. Additionally, our RL^2 implementation uses content embeddings rather than SVD factors, which introduces a confound when comparing it to methods that benefit from the collaborative prior.

7 Conclusion

We studied exploration strategies for cold-start recommendation by comparing six methods on MovieLens-1M under two evaluation protocols. Our results show that when a strong collaborative prior is available, greedy exploitation is surprisingly difficult to beat at the standard $T = 20$ horizon. Hybrid Thompson Sampling comes closest, and its advantage grows over longer horizons as the Bayesian posterior sharpens. At $T = 100$, the relative gap between Greedy CF and Hybrid TS narrows from 2.2% to 0.6%, and Hybrid TS wins 24 of the last 50 individual steps. Ablation studies confirm that exploration methods outperform greedy baselines when the prior is weakened through reduced data, noise, or population mismatch.

These findings suggest that the choice of exploration strategy should depend on the quality of the available prior. When rich collaborative data exists, the marginal value of exploration is small at short horizons. When the prior is weak or miscalibrated, Thompson Sampling and similar Bayesian approaches provide meaningful gains. Future work could extend this analysis to implicit feedback datasets, incorporate side information from large language models to improve cold-start priors, and evaluate in online settings where user engagement metrics replace ground-truth ratings.

8 Team Contributions

- **Anastasiya Masalava:** problem formulation, dataset construction and cleaning, Meta-RL via RL^2 implementation and evaluation, prior quality case studies, report writing.
- **Eva Casto:** random baseline, implementation and evaluation for Greedy CF and Constrained Bandit, report writing.
- **Ela Sigin:** NLTS and HTS implementation, robustness analysis (warm-user fraction, reward noise, population mismatch ablations), longer horizon evaluation, report writing.

AI Tools Disclosure: We used Claude and ChatGPT as an aid tool for writing assistance (LaTeX formatting, grammar correction, phrasing) and for minor code development (data loading and plotting). All core algorithm implementation was developed independently. Experimental design, result analysis, and paper writing were our own work, with AI tools used only for polish and boilerplate.

Changes from Proposal: Our original proposal focused on four methods: Greedy CF, Neural Linear Thompson Sampling, Meta-RL via RL^2, and Constrained Bandit. During the project, we added Hybrid Thompson Sampling to isolate the effect of the feature space (SVD vs. content embeddings) on Thompson Sampling performance. Finally, we introduced the longer horizon experiment ($T = 100$) and prior quality ablations, which were not in the original proposal but proved essential for understanding when exploration provides genuine benefit.

References

Yan Duan, John Schulman, Xi Chen, Peter L. Bartlett, Ilya Sutskever, and Pieter Abbeel. 2016. RL^2: Fast Reinforcement Learning via Slow Reinforcement Learning. *arXiv preprint arXiv:1611.02779* (2016).

F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 5, 4, Article 19 (Dec. 2015), 19 pages. doi:10.1145/2827872

Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi-Yadkori, and Benjamin Van Roy. 2017. Conservative Contextual Linear Bandits. arXiv:1611.06426 [cs.LG]

Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *Computer* 42, 8 (2009), 30–37.

Hoyeop Lee, Jinbae Im, Seongwon Jang, Hyunjae Cho, and Sehee Chung. 2019. MeLU: Meta-Learned User Preference Estimator for Cold-Start Recommendation. arXiv:1908.00413 [cs.IR]

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. 2010. A Contextual-Bandit Approach to Personalized News Article Recommendation. arXiv:1003.0146 [cs.LG]

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

Carlos Riquelme, George Tucker, and Jasper Snoek. 2018. Deep Bayesian Bandits Showdown: An Empirical Comparison of Bayesian Deep Networks for Thompson Sampling. arXiv:1802.09127 [stat.ML]

Yi Su, Haokai Lu, Yujing Li, Liang Liu, Shuchao Bi, Ed H. Chi, and Minmin Chen. 2024. Multi-Task Neural Linear Bandit for Exploration in Recommender Systems. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 5723–5730. doi:10.1145/3637528.3671649

William R. Thompson. 1933. On the likelihood that one unknown probability exceeds another. *Biometrika* 25 (1933), 285–294.

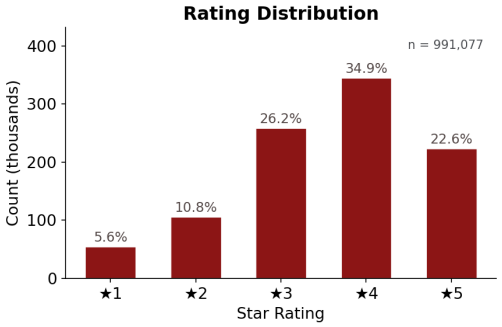
Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. *arXiv* (2020). doi:10.48550/arxiv.2002.10957

Zheqing Zhu and Benjamin Van Roy. 2023. Scalable Neural Contextual Bandit for Recommender Systems. arXiv:2306.14834 [cs.IR]

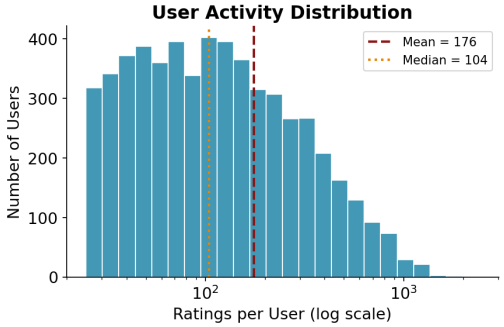
Lixin Zou, Long Xia, Yulong Gu, Xiangyu Zhao, Weidong Liu, Jimmy Xiangji Huang, and Dawei Yin. 2020. Neural Interactive Collaborative Filtering. arXiv:2007.02095 [cs.IR]

A Dataset Analysis

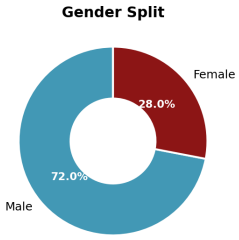
Figure 7 summarizes the key statistics of our filtered MovieLens-1M dataset. The rating distribution (Figure 7a) is positively skewed, with 4-star ratings being the most common (34.9%) and fewer than 17% having less than 3 stars. The user activity distribution (Figure 7b) is right-skewed, with a mean of 175 ratings, which suggests that this dataset naturally supports our short-horizon ($T = 20$) and long-horizon ($T = 100$) experiments. Demographically, the dataset has more male users (72%, Figure 7c) and is concentrated in the 25–34 age bracket (1,980 users, Figure 7d), with college and graduate students forming the largest group (717 users, Figure 7f). This demographic skew toward younger, educated users is directly relevant to our population mismatch ablation (Section 5.2.1), where we train the prior on users aged 50+ and evaluate on users aged under 25. Finally, the warm/cold split (Figure 7e) shows that 3,936 warm users contribute 694K ratings for prior training, while 1,688 cold users with 297K ratings are held out for evaluation. Overall, the dataset provides a well-calibrated prior training set, a realistic cold-start evaluation set, and enough demographic variation to support population mismatch stress-testing.



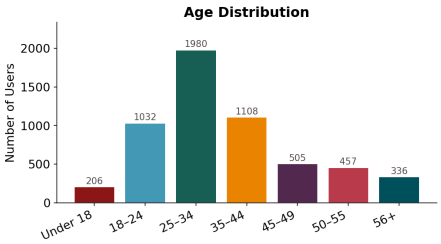
(a) Rating distribution.



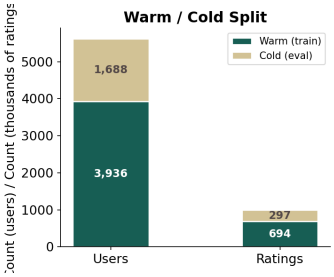
(b) User activity distribution (log scale).



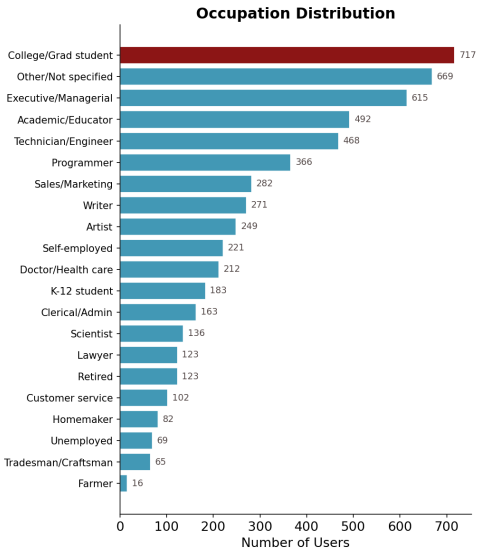
(c) Gender split.



(d) Age distribution.



(e) Warm/cold user split.



(f) Occupation distribution.

Figure 7: Dataset statistics for the filtered MovieLens-1M dataset (5,624 users, 3,702 items, 991,077 ratings).