

An Empirical Study on Zero-Data Bootstrapping for Conversational Recommender Systems

Rohan Surana^{1,*}, Junda Wu^{1,*}, Zhouhang Xie^{1,*},
Yu Xia¹, Nathan Kallus^{2,3}, Julian McAuley¹

¹University of California, San Diego

²Netflix Inc. ³Cornell University

{rsurana, juw069, zhx022, yux078, jmcauley}@ucsd.edu

{nkallus}@netflix.com

Abstract

Conversational Recommender Systems (CRS) typically require domain-specific dialogue data, which is costly, scarce, and often unavailable in new domains. We conduct a systematic empirical study of *zero-data CRS bootstrapping*: generating synthetic conversational supervision from non-conversational signals—item reviews, metadata, and user-item interactions—without any in-domain dialogue corpus. We compare two information-theoretic selection strategies, Jensen-Shannon diversity and Fisher information, across domain signals, model architectures, datasets, and fine-tuning paradigms. Our results show that domain-grounded synthetic data consistently outperforms zero-shot prompting and naïve synthetic baselines; active selection improves data efficiency over random sampling; metadata and collaborative filtering signals each improve selection quality; and, in low-resource settings, synthetic data can outperform scarce real dialogues while further complementing them. These findings establish non-conversational domain signals as a viable path toward building CRS without conversational training data. The code is available at https://anonymous.4open.science/r/zero_data_crs/.

1 Introduction

Conversational Recommender Systems (CRS) aim to deliver personalized recommendations through dialogue (Sun and Zhang, 2018; Lin et al., 2023; Wang et al., 2023a; Xi et al., 2024; Zhang et al., 2018). Building effective CRS has traditionally required large, domain-specific conversational datasets (Wang et al., 2022; Chen et al., 2019; Zhou et al., 2020; Li et al., 2018) (Figure 1(a)), which are scarce due to *high annotation costs* (Soudani et al., 2026; Chen et al., 2023; Aroyo et al., 2023), *privacy concerns* (Reddit, 2023b,a, 2024; Chen et al., 2023), and *domain-specific constraints* (Chen et al., 2019; Zhou et al., 2020; Wang et al., 2022).

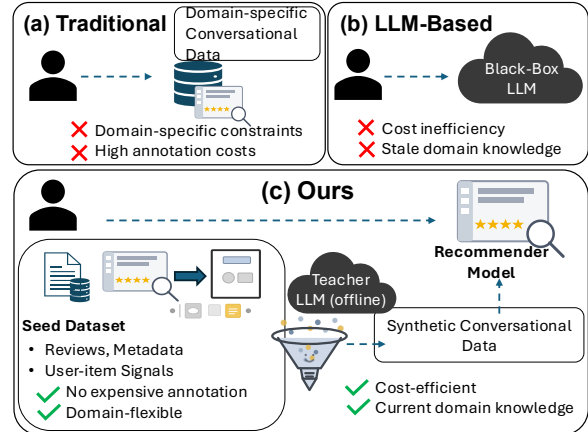


Figure 1: Three CRS paradigms: (a) traditional CRS requiring domain-specific conversational data, (b) direct LLM deployment with inherent limitations, and (c) the zero-data bootstrapping pipeline studied in this work, which synthesizes conversational data from non-conversational seed signals.

While Large Language Models (LLMs) demonstrate promising zero-shot capabilities (He et al., 2023; Feng et al., 2023; Friedman et al., 2023; Wu et al., 2024), direct deployment (Figure 1(b)) faces challenges in *scalability* (Menshaw et al., 2024; Chen et al., 2025), *cost* (Samsi et al., 2023; Chen et al., 2024a), *interpretability* (Singh et al., 2024; Kasneci et al., 2023; Wei Jie et al., 2024), and *privacy* (Yao et al., 2024; Neel and Chang, 2024; Yan et al., 2025; Li et al., 2024b), leading many practical deployments to prefer smaller, internally managed models (Jeong, 2024). Yet fine-tuning these models still requires substantial domain-specific conversational data (Jeong, 2024; Sun et al., 2024; Nguyen et al., 2024), which remains limited (Soudani et al., 2026; Wang et al., 2023c; Ruiz and Sell, 2024).

In parallel, LLMs have been increasingly used as data generators, synthesizing training data for downstream tasks (Zhao et al., 2023; Mysore et al., 2023; Leszczynski et al., 2023). However, existing

augmentation approaches for CRS operate *on top of* an existing conversational corpus (Xu et al., 2025; Liu et al., 2025; Zhao et al., 2023) or require external knowledge graphs (Wang et al., 2022; Chen et al., 2019; Zhou et al., 2020). Active learning offers principled sample selection to maximize the value of costly LLM queries (Du et al., 2025; Liu et al., 2024; Zhang et al., 2024), but its application to CRS data generation from non-conversational signals has not been studied.

This raises a fundamental question: *can LLMs bootstrap effective CRS from non-conversational domain signals alone?* Non-conversational data—item reviews, metadata, and user-item collaborative interactions—is readily available in most domains, yet converting such signals into effective conversational training data involves non-trivial challenges in sample selection, domain grounding, and data efficiency. No systematic investigation exists of whether and how this bootstrapping can work.

In this work, we conduct a comprehensive empirical study addressing this question. We formalize the *zero-data CRS* setting (Figure 1(c)), where no in-domain conversational data is available, and study a three-stage bootstrapping pipeline: (i) active sample selection from non-conversational seed data using information-theoretic criteria, (ii) synthetic conversational data generation via a teacher LLM, and (iii) fine-tuning a target model on the synthesized corpus. We systematically vary domain signals, selection strategies, budgets, model architectures, and fine-tuning paradigms to address four research questions.

Contributions. We summarize our contributions as follows:

- We formalize the *zero-data CRS* setting and establish a controlled experimental framework for studying synthetic data bootstrapping from non-conversational domain signals.
- We provide a systematic empirical study of sample selection strategies for augmenting CRS data generation, characterizing how different selection criteria affect synthetic-data quality and downstream recommendation performance.
- We conduct controlled ablations isolating the effects of domain signals (item metadata, collaborative filtering) on both selection quality and downstream CRS performance across

Method	CRS Data	KG	Zero-Data
UniCRS (Wang et al., 2022)	✓	✓	✗
KGSF (Zhou et al., 2020)	✓	✓	✗
KBRD (Chen et al., 2019)	✓	✓	✗
Xu et al. (2025)	✓	✗	✗
NR-CRS (Liu et al., 2025)	✓	✗	✗
LOT-CRS (Zhao et al., 2023)	✓	✗	✗
This Study	✗	✗	✓

Table 1: Comparison of CRS methods by data requirements. CRS Data: requires in-domain conversational corpus. KG: requires external knowledge graph. Zero-Data: operates without conversational training data.

multiple models, datasets, and fine-tuning paradigms.

Scope. We focus on the strict setting where neither an in-domain conversational corpus nor an external knowledge graph is available. The only assumed resources are non-conversational domain signals such as reviews, metadata, and user-item interactions. Table 1 contrasts this scope with representative CRS and augmentation methods.

2 Zero-Data CRS Bootstrapping

We define the zero-data CRS setting and the bootstrapping pipeline evaluated in this study. Figure 2 provides an overview.

2.1 Setting: Zero-Data CRS

Definition 1 (Zero-Data CRS). *A CRS operates in the zero-data setting when no in-domain conversational dataset $\mathcal{S}_{\text{conv}}$ is available. The only available resources are non-conversational domain signals: item reviews $\{r_i\}$, item metadata $\{m_i\}$, and user-item collaborative interactions $\{u_i\}$.*

This setting captures the practical reality of deploying CRS in new or specialized domains where conversational data has never been collected, requiring conversational supervision to be derived from non-conversational signals alone.

2.2 Pipeline: Select, Generate, Fine-Tune

To study how non-conversational signals can be converted into effective CRS training data, we instantiate a three-stage bootstrapping pipeline (Algorithm 1). Given seed data $\mathcal{D}$, target model π_θ , teacher LLM $\mathcal{M}$, and budget B , the pipeline selects informative samples $\mathcal{R} \subset \mathcal{D}$, generates synthetic conversations $\mathcal{S}$ from $\mathcal{R}$, and fine-tunes π_θ on $\mathcal{S}$ to obtain π_θ^* .

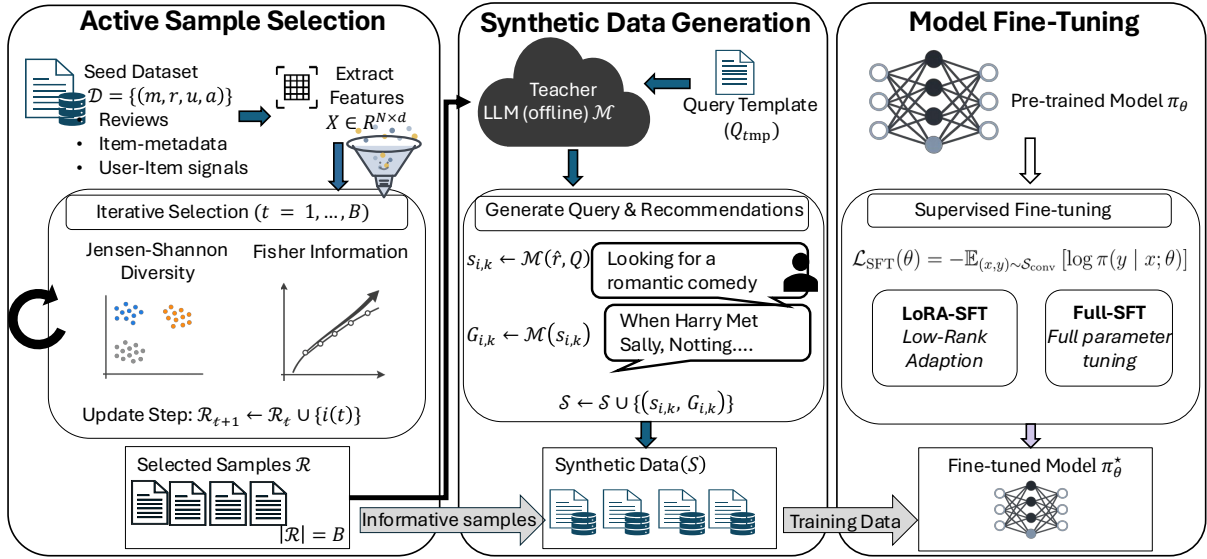


Figure 2: Overview of zero-data CRS bootstrapping. Non-conversational seed signals are encoded, actively selected, converted into synthetic conversations by a teacher LLM, and used to fine-tune the target CRS.

Algorithm 1: Zero-Data CRS Bootstrapping Pipeline

Input : Seed dataset $\mathcal{D} = \{(m_i, r_i, u_i, a_i)\}_{i=1}^N$;
 Language model π_θ ; Black-box LLM $\mathcal{M}$; Query templates $\mathcal{Q}_{\text{tmp}}$;
 Budget B

Output : Fine-tuned model π_θ^*

```

1 Initialize:  $\mathcal{S} \leftarrow \emptyset$ ;
2 foreach  $i \in \mathcal{D}$  do
3    $\mathbf{x}_i \leftarrow \text{embed}(\pi_\theta, i)$ 
4 end
5  $\mathcal{R} \leftarrow \text{select}(\mathcal{D}, \{\mathbf{x}_i\}, B)$  ▷ §2.3
6  $\mathcal{S} \leftarrow \text{generate}(\mathcal{R}, \mathcal{Q}_{\text{tmp}}, \mathcal{M})$  ▷ Alg. 2
7  $\pi_\theta^* \leftarrow \text{fine\_tune}(\pi_\theta, \mathcal{S})$  ▷ §11
8 return  $\pi_\theta^*$ 

```

2.3 Selection: Informative Seed Samples

Given the seed dataset $\mathcal{D} = \{(m_i, r_i, u_i, a_i)\}_{i=1}^N$ with N samples, where m_i denotes item metadata, r_i item reviews, u_i user features derived from collaborative signals, and a_i the item identifier, the goal is to select a subset of size B that maximizes the quality of downstream synthetic data.

We extract dense representations by leveraging π_θ 's last transformer layer hidden states $X \in \mathbb{R}^{N \times d}$, following prior work on LLM-based embeddings (Tang and Yang, 2024; Wang et al., 2024; Li

and Zhou, 2025; Jiang et al., 2024) to prevent sample misspecification (Sugiyama, 2005; Fudenberg et al., 2017; Lin et al., 2025).

Selection precedes generation, so no training targets exist yet and label-dependent criteria such as uncertainty sampling do not apply. We therefore build on two label-free principles from active learning, diversity/coverage and Fisher-information-based optimal design (Sener and Savarese, 2018; Ash et al., 2021; Mukherjee et al., 2024; Kveton et al., 2025), instantiated over X . The two criteria encode competing hypotheses about what makes a seed sample useful for CRS data generation: JS diversity prioritizes distributional coverage, while Fisher information prioritizes parameter-level informativeness under a last-layer approximation.

Definition 2 (JS Diversity Selection). Let p_i denote the soft cluster-membership distribution for sample i , obtained by clustering the embedding space and applying a softmax over distances to cluster centroids. Let $H(p_i)$ be its Shannon entropy. The first sample is initialized as the maximum-entropy item, after which $\bar{p}_t = \frac{1}{|\mathcal{R}_t|} \sum_{j \in \mathcal{R}_t} p_j$ is computed over the non-empty selected set. JS diversity then selects the next sample by

$$i^{(t)} = \arg \max_{i \in \mathcal{D} \setminus \mathcal{R}_t} \left[H(p_i) + \lambda \text{JS}(p_i \parallel \bar{p}_t) \right]. \quad (1)$$

The entropy term favors boundary samples, while the JS term promotes novelty relative to the selected set.

Definition 3 (Fisher Information Selection). *Let $\Lambda_t = (I_d + \sum_{j \in \mathcal{R}_t} \mathbf{x}_j \mathbf{x}_j^\top)^{-1}$ denote the inverse regularized information matrix for the current selected set. Under a last-layer linearization of the fine-tuning objective (Mukherjee et al., 2024; Kveton et al., 2025), Fisher selection chooses the sample with the largest incremental information gain:*

$$i^{(t)} = \arg \max_{i \in \mathcal{D} \setminus \mathcal{R}_t} \log(1 + \mathbf{x}_i^\top \Lambda_t \mathbf{x}_i). \quad (2)$$

This greedy log-determinant objective favors samples that add new parameter-relevant directions beyond those already selected.

Together, the two criteria operationalize complementary notions of sample utility: distributional representativeness and parameter-aligned informativeness (Kirsch and Gal, 2022). Fisher selection corresponds to greedy D-optimal design (Pukelsheim, 2006), while JS selection is motivated by submodular coverage objectives (Nemhauser et al., 1978). Both criteria are applied iteratively until $|\mathcal{R}| = B$; construction and update details are provided in Section A.

2.4 Pipeline: From Domain Signals to CRS Training Data

Domain Signal Encoding.

Definition 4 (Domain Signals). *We distinguish three types of non-conversational domain signals available for seed sample representation:*

- **Semantic signals** (r_i): item review text, capturing fine-grained user opinions and item characteristics;
- **Structural signals** (m_i): item metadata including descriptions, genres, and categories, providing categorical and descriptive context;
- **Relational signals** (u_i): user-item collaborative filtering features derived from interaction histories, encoding user preference patterns.

Each signal type is encoded into the embedding space by concatenating it with the seed input before extracting representations via π_θ . This study systematically investigates how each signal type, and their combinations, affects both selection quality and downstream CRS performance.

Synthetic Conversation Generation. After active sample selection, the selected samples $\mathcal{R}$ are converted into synthetic conversational data using

Algorithm 2: Synthetic conversational data generation from actively selected non-conversational seed items.

Input : Selected samples $\mathcal{R} \subset \mathcal{D}$; Query templates $\mathcal{Q}_{\text{tmp}}$; Teacher LLM $\mathcal{M}$; Queries per item K

Output : Synthetic dataset $\mathcal{S} = \{(s, G)\}$, where G contains pseudo-target recommendations

```

1 Initialize:  $\mathcal{S} \leftarrow \emptyset$ ;
2 foreach  $i \in \mathcal{R}$  do
3   Sample 5 style templates  $Q \subset \mathcal{Q}_{\text{tmp}}$ ;
4   for  $k = 1$  to  $K$  do
5     Sample 3 reviews  $\hat{r} \subset r_i$ ;
6      $s_{i,k} \leftarrow \mathcal{M}(\hat{r}, Q) \triangleright$  Generate query;
7      $G_{i,k} \leftarrow \mathcal{M}(s_{i,k}) \triangleright$  Generate 20
      recommendation targets;
8      $\mathcal{S} \leftarrow \mathcal{S} \cup \{(s_{i,k}, G_{i,k})\}$ ;
9   end
10 end
11 return  $\mathcal{S}$ 

```

a teacher LLM. Let $\mathcal{Q}_{\text{tmp}}$ be a collection of style templates sourced from Reddit movie recommendation discussions, used only to control the style and tone of generated queries (see Box 4).

For each selected item $i \in \mathcal{R}$, we sample five style templates $Q \subset \mathcal{Q}_{\text{tmp}}$. For each of K synthetic queries, we sample three reviews $\hat{r} \subset r_i$ and prompt a teacher LLM $\mathcal{M}$ to generate a synthetic query $s_{i,k}$ (Box 3). Following He et al. (2023), we then prompt the LLM with each query to produce 20 pseudo-target recommendations, $G_{i,k} = \{g_1, \dots, g_{20}\}$ (Box 2). Each pair $(s_{i,k}, G_{i,k})$ is added to the synthetic dataset $\mathcal{S}$ for fine-tuning the target language model π_θ . For B selected items and K queries per item, this yields BK training pairs at a cost of $2BK$ offline teacher calls; selection itself requires one embedding pass and no additional teacher calls. The full procedure is detailed in Algorithm 2, with representative examples in Box 1.

Target Model Fine-Tuning. We adapt π_θ to the conversational recommendation domain using supervised fine-tuning on $\mathcal{S}$, evaluating both LoRA and Full-SFT following standard practice (Chen et al., 2024b; Li et al., 2024a; Dong et al., 2024; Zhang et al., 2026). Training details and hyperparameters are provided in Section E.1.

3 Experimental Setup

3.1 Benchmarks and Seed Data

We evaluate on two standard CRS benchmarks of different scales: ReDial (Li et al., 2018) and INSPIRED (Hayati et al., 2020). This pairing enables studying bootstrapping effectiveness across both moderate- and low-resource evaluation settings. Our non-conversational seed data is drawn from Amazon Reviews ’23 (Hou et al., 2026) (Movies & TV category). Dataset statistics are reported in Table 2.

	ReDial	INSPIRED
Total dialogues	10,006	1,001
Test dialogues	2,001	200

Table 2: Dataset statistics for the ReDial and INSPIRED evaluation benchmarks.

3.2 Backbone and Teacher Models

We study instruction-tuned LLMs as backbone models: Llama3.2-3B-Instruct (Grattafiori et al., 2024), Qwen2.5-1.5B-Instruct (Qwen et al., 2025), and Qwen3-4B (Yang et al., 2025) (RQ1 only). This range of model sizes enables studying how model capacity interacts with synthetic data quality.

For fine-tuning, we consider two paradigms: (1) LoRA (Hu et al., 2022), parameter-efficient fine-tuning using Low-Rank Adaptation, and (2) Full-SFT, where all model parameters are optimized. The teacher LLM for synthetic data generation is GPT-4o (OpenAI et al., 2024). Implementation details are provided in Section E.1.

3.3 Baselines and Selection Conditions

To isolate the effect of domain-grounded synthetic data, we compare against baselines operating under the same zero-data constraint:

Zero-Shot: Models are directly prompted for conversational recommendation without fine-tuning.

GPT-Generated: Models fine-tuned on synthetic data generated by naively prompting GPT-4o (OpenAI et al., 2024) without domain-specific seed data or active selection. **Popularity:** Recommends items based on frequency in the dataset. **NBCRS** (Xie et al., 2024): A neighborhood-based CRS evaluated both on raw seed data and on synthesized conversational data.

Existing CRS augmentation methods (Xu et al., 2025; Liu et al., 2025) require an in-domain con-

versational corpus and are therefore not applicable in the zero-data setting.

For active selection, we compare **Random** uniform sampling; semantic-only **JS/Fisher** selection using review embeddings; **JS+Meta/Fisher+Meta**, which add item metadata; and **JS+CF/Fisher+CF**, which add collaborative filtering (CF) signals from user interaction histories. We evaluate with Recall@ k ($k \in \{1, 5, 10, 20\}$) and NDCG@ k (Section C); unless otherwise stated, RQ1–RQ3 train exclusively on synthetic data, while RQ4 studies combinations of synthetic and in-domain data.

4 Main Results

We organize the study around four questions: whether domain-grounded synthetic data improves zero-data CRS (RQ1), whether active selection improves data efficiency (RQ2), how metadata and collaborative signals affect selection (RQ3), and how synthetic data compares with scarce real dialogues (RQ4).

4.1 RQ1: Domain-Grounded Synthetic Data Enables Zero-Data CRS

We first ask whether a CRS can be trained without any in-domain conversational corpus by converting non-conversational domain signals into synthetic conversational supervision.

We compare popularity ranking, zero-shot prompting, naïve GPT-generated training data, and our domain-grounded synthetic data across Qwen2.5-1.5B, Qwen3-4B, Llama3-3B, and NBCRS. For LLM backbones, we evaluate both LoRA and Full-SFT when applicable; for NBCRS, we compare raw seed data against the synthesized conversational data. Table 3 reports Recall@ k on ReDial and INSPIRED.

Finding 1: Domain-grounded synthetic data consistently outperforms zero-shot and naïve generation. Models trained on domain-grounded synthetic data outperform both zero-shot baselines and models trained on GPT-generated data (produced without domain-specific seed signals) across backbone architectures and benchmarks. On INSPIRED with Qwen2.5-1.5B, SFT-tuned models achieve +207.8% Recall@1 improvement over zero-shot, compared to only +18.8% for GPT-generated data. This indicates that incorporating domain-specific signals during synthetic data generation significantly improves recommendation ac-

Model	Setting	ReDial				INSPIRED			
		R@1	R@5	R@10	R@20	R@1	R@5	R@10	R@20
Popularity	Seed Dataset	0.09 _{0.04}	0.24 _{0.06}	0.33 _{0.07}	0.34 _{0.07}	0.03 _{0.03}	0.03 _{0.03}	0.55 _{0.13}	0.70 _{0.15}
	Synth Data	0.56 _{0.09}	1.87 _{0.16}	2.46 _{0.19}	3.72 _{0.23}	0.30 _{0.10}	0.85 _{0.16}	1.77 _{0.23}	2.35 _{0.26}
Qwen2.5-1.5B	Zero-shot	2.39 _{0.19}	6.96 _{0.31}	8.15 _{0.33}	9.51 _{0.36}	0.64 _{0.14}	1.43 _{0.21}	2.01 _{0.24}	2.35 _{0.26}
	GPT-Generated	2.56 _{0.19}	7.63 _{0.32}	9.29 _{0.35}	10.71 _{0.38}	0.76 _{0.17}	1.89 _{0.26}	2.34 _{0.29}	3.29 _{0.35}
	Synth-LoRA	3.39 _{0.22}	9.27 _{0.35}	11.13 _{0.38}	12.83 _{0.41}	1.17 _{0.21}	3.21 _{0.34}	4.43 _{0.40}	5.82 _{0.46}
	Synth-SFT	3.36 _{0.22}	9.38 _{0.36}	11.40 _{0.39}	13.29 _{0.42}	1.97 _{0.27}	4.58 _{0.41}	6.09 _{0.47}	7.90 _{0.52}
Qwen3-4B	Zero-shot	7.09 _{0.31}	19.58 _{0.48}	22.70 _{0.51}	26.87 _{0.54}	5.42 _{0.40}	10.20 _{0.53}	12.03 _{0.57}	13.80 _{0.60}
	GPT-Generated	7.45 _{0.32}	22.19 _{0.51}	25.62 _{0.53}	28.74 _{0.55}	5.94 _{0.46}	11.12 _{0.61}	12.86 _{0.65}	14.03 _{0.68}
	Synth-LoRA	8.39 _{0.34}	23.72 _{0.52}	26.26 _{0.54}	28.79 _{0.55}	5.82 _{0.46}	10.48 _{0.60}	12.07 _{0.63}	12.86 _{0.65}
	Synth-SFT	7.73 _{0.33}	22.81 _{0.51}	26.32 _{0.54}	29.80 _{0.56}	6.01 _{0.46}	11.35 _{0.62}	13.12 _{0.66}	14.67 _{0.69}
Llama3-3B	Zero-shot	1.57 _{0.15}	5.56 _{0.28}	9.07 _{0.35}	13.26 _{0.41}	2.04 _{0.25}	5.00 _{0.38}	6.94 _{0.44}	8.89 _{0.50}
	GPT-Generated	1.54 _{0.15}	5.86 _{0.29}	8.96 _{0.35}	13.47 _{0.42}	1.97 _{0.27}	5.03 _{0.43}	7.07 _{0.50}	9.34 _{0.57}
	Synth-LoRA	1.53 _{0.15}	5.71 _{0.28}	8.79 _{0.35}	13.35 _{0.42}	1.97 _{0.27}	5.07 _{0.43}	7.26 _{0.50}	9.42 _{0.57}
	Synth-SFT	1.91 _{0.17}	6.54 _{0.30}	9.72 _{0.36}	14.42 _{0.43}	2.61 _{0.31}	5.75 _{0.45}	7.83 _{0.52}	9.98 _{0.58}
NBCRS	Seed Dataset	0.00 _{0.00}	0.39 _{0.28}	0.78 _{0.39}	1.17 _{0.48}	0.00 _{0.00}	0.00 _{0.00}	0.00 _{0.00}	0.00 _{0.00}
	Synth Data	2.54 _{0.70}	11.52 _{1.41}	19.53 _{1.75}	27.34 _{1.97}	4.55 _{2.23}	9.09 _{3.08}	13.64 _{3.68}	23.86 _{4.57}

Table 3: RQ1 Recall (%) results on ReDial and INSPIRED. Small gray values denote standard errors. Shaded entries denote LLM synthetic fine-tuning variants; bold marks the best Recall in each model block and dataset.

curacy beyond what generic LLM generation provides.

Finding 2: Smaller models benefit disproportionately from synthetic fine-tuning. The relative gains from domain-grounded synthetic data are most pronounced for smaller models. Qwen2.5-1.5B sees Recall@1 improvements of +41.8% (LoRA) and +40.6% (SFT) on ReDial, while the larger Qwen3-4B achieves +18.3% and +9.0% respectively, suggesting that smaller models benefit more from domain-specific fine-tuning data.

Finding 3: Full-SFT generally outperforms LoRA, particularly for smaller models. Across most configurations, Full-SFT yields stronger results than LoRA-based fine-tuning, with the gap being especially pronounced for smaller models (e.g., Qwen2.5-1.5B on INSPIRED: +207.8% vs. +82.8% for Recall@1). For Llama3-3B, LoRA fine-tuning sometimes underperforms zero-shot, while Full-SFT consistently improves. This suggests that when training data is synthetically generated, full parameter updates are often important for adapting the model sufficiently.

Finding 4: The bootstrapping pipeline generalizes beyond LLM-based CRS. NBCRS (Xie et al., 2024), a traditional neighborhood-based CRS, achieves substantially higher performance when trained on synthesized conversational data compared to raw seed data (e.g., Recall@5 improves from 0.39 to 11.52 on ReDial). This confirms that the data transformation—from non-

conversational signals to conversational format—provides value independent of the downstream model architecture.

Synthetic-data quality. We separately assess the generated corpus itself: human ratings of sampled query-recommendation pairs average close to 4 on a 1–5 scale across naturalness, coherence, relevance, and diversity, and both corpora follow the requested output format. Notably, the naïve corpus matches the item catalog *more* often than the grounded corpus while performing worse downstream, indicating that catalog overlap does not track the usefulness of the supervision. Full human-evaluation and title-validity results are reported in Section D.

Takeaway 1. The decisive factor is not simply using an LLM to generate more text, but grounding generation in domain-specific seed signals. This supports zero-data bootstrapping as a practical alternative when conversational logs are unavailable.

4.2 RQ2: Active Selection Improves Data Efficiency

Synthetic data generation incurs teacher-LLM cost, so the next question is whether active selection can identify more useful seed items under fixed generation budgets.

We compare Jensen-Shannon (JS) divergence, Fisher information, random sampling, and a popularity heuristic that selects items with the highest review counts. We fine-tune Llama3-3B and

Qwen2.5-1.5B on data generated from each selected subset and evaluate Recall@ k ($k \in \{1, 5\}$) on ReDial and INSPIRED.

Figure 3 and Figure 7 compare performance across budgets.

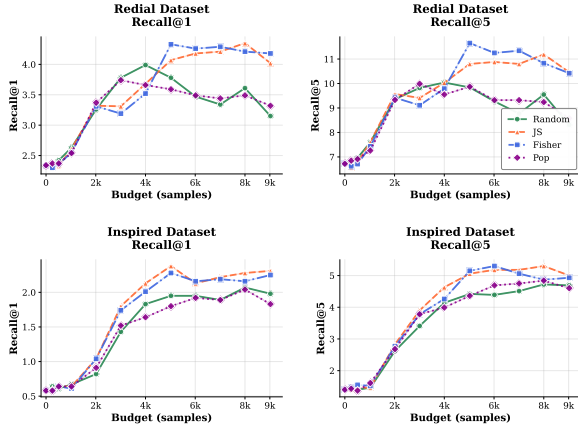


Figure 3: RQ2: Qwen2.5-1.5B performance across budgets on ReDial and INSPIRED. Active selection strategies (JS, Fisher) consistently outperform random sampling and popularity-based selection.

Finding 5: Both active strategies outperform random sampling and popularity-based selection, but the margin depends on budget and signal. JS and Fisher consistently identify more beneficial seed items than random or popularity-based selection, reducing teacher-LLM calls while improving performance at similar budgets. Popularity, which prioritizes high-review-count items, often underperforms even random sampling, consistent with prior observations that random sampling can be a strong baseline in text-generation tasks (Perlitz et al., 2023).

Finding 6: JS and Fisher exhibit complementary strengths. JS prioritizes distributional coverage, while Fisher targets parameter-level information gain aligned with the fine-tuning objective (Kirsch and Gal, 2022). Their relative advantage varies by dataset and budget, confirming that they capture different facets of sample informativeness and that domain popularity alone is not a reliable proxy. The same ordering holds under NDCG at an equal teacher-call budget (Section C), so active selection improves the rank position of correct items and not only their presence in the list. We further analyze the popularity profile and recommendation coverage of active selection in Section B.

Takeaway 2. For zero-data CRS, seed selection should be treated as part of the data-generation algorithm rather than as a preprocessing detail. Information-theoretic criteria make teacher queries more efficient than simple item-frequency heuristics.

4.3 RQ3: Domain Signals Improve Active Selection

Non-conversational domains provide multiple signal types. We ask whether metadata and collaborative filtering signals improve selection beyond review text alone. We isolate domain-signal effects by comparing semantic-only strategies (JS, Fisher), metadata-aware strategies (JS+Meta, Fisher+Meta), collaborative-filtering-aware strategies (JS+CF, Fisher+CF), and Random. Unless otherwise noted, this analysis uses Llama3 across budget levels on ReDial and INSPIRED.

Metadata Provides Structural Coverage. Figure 4 shows Recall@1 and Recall@5 across budget levels for both datasets using Llama3.

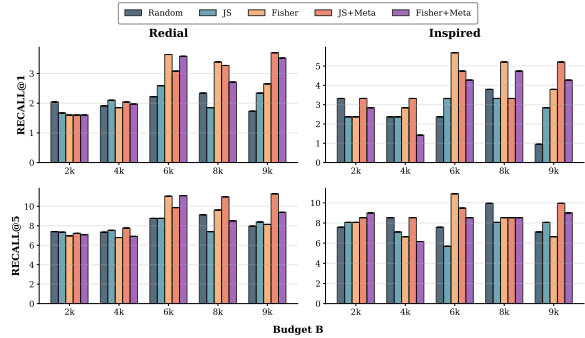


Figure 4: RQ3: Metadata-aware selection improves Recall@1 and Recall@5 over semantic-only and random selection, especially at larger budgets.

Finding 7: Metadata-aware selection consistently improves performance, with gains most pronounced at larger budgets. Methods incorporating item metadata outperform semantic-only and random baselines across both datasets. The cumulative advantage grows with budget, suggesting that metadata enables active selection to sustain informative sample identification over longer selection horizons. This confirms that structural domain signals (descriptions, genres, categories) provide context beyond what review text alone captures.

Collaborative Signals Add Preference Structure. Figure 5 presents results across budget levels for both datasets.

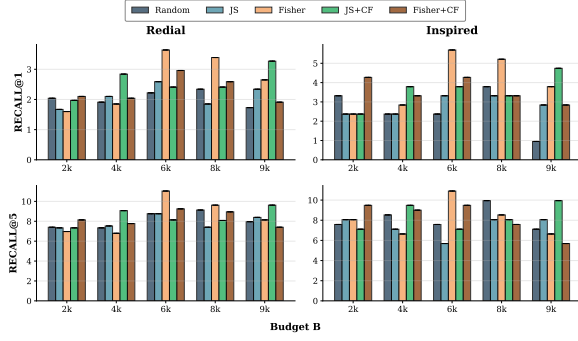


Figure 5: RQ3: Collaborative filtering signals provide complementary gains to semantic and metadata-based selection.

Finding 8: Collaborative filtering signals provide additional gains orthogonal to semantic features. Collaborative-aware methods consistently demonstrate improved recommendation accuracy over baselines, with noticeable gains at larger budgets. This suggests that relational signals, which encode user preference patterns from interaction histories, enrich the selection process by emphasizing realistic, user-centric interactions that complement the content-based information in reviews and metadata.

Takeaway 3. Domain signals should be used not only as prompt context for generation, but also as selection signals. Metadata offers an easily available structural signal, while collaborative histories add preference information when such logs exist.

4.4 RQ4: Synthetic Data Complements Scarce Real Data

Finally, we ask whether synthetic data substitutes for or complements scarce real conversational data. We compare models trained on **ReD** (ReDial only), **INS** (INSPIRED only), **Synth** (synthetic data only), **R+O** (ReDial + our synthetic data), and **INS+O** (INSPIRED + our synthetic data). Figure 6 compares these training mixtures.

Finding 9: In low-resource settings, synthetic data alone outperforms real in-domain data. On INSPIRED (~1k dialogues), the model trained solely on synthetic data (**Synth**) outperforms the model trained on the original INSPIRED training set (**INS**), showing that domain-grounded synthetic data generated from non-conversational signals can surpass actual in-domain conversational data when the latter is scarce.

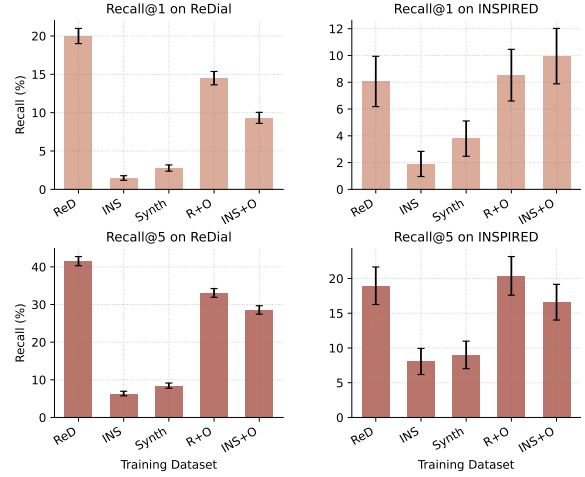


Figure 6: RQ4: Synthetic data is most beneficial when real conversational data is scarce and becomes less helpful when in-domain dialogue coverage is already large.

Finding 10: Combining synthetic and real data yields further gains, confirming complementarity. On INSPIRED, training with synthetic data alongside scarce or related real dialogues (**INS+O** and **R+O**) substantially improves over **INS** alone. On ReDial, **ReD** remains competitive and **R+O** slightly reduces performance, suggesting that synthetic augmentation is most valuable when real conversational coverage is limited.

Takeaway 4. Synthetic bootstrapping is most useful precisely in the low-resource setting that motivates zero-data CRS. When real conversational coverage is already broad, synthetic data should be used more selectively or filtered more aggressively.

5 Related Work

Conversational Recommender Systems. CRS supports interactive recommendation, often with mixed initiative between user and system (Christakopoulou et al., 2016; Li et al., 2018; Chen et al., 2019; Lei et al., 2020; He et al., 2022; Zhang et al., 2022; Xia et al., 2023). Recent work improves CRS by integrating recommendation and conversation modules, incorporating language models, or directly prompting and fine-tuning LLMs for recommendation (Wang et al., 2022; Xie et al., 2024; Friedman et al., 2023; Yang and Chen, 2024; He et al., 2023; Feng et al., 2023; Xi et al., 2024). Our study targets a stricter setting: fine-tuning LLM-based CRS when *no* in-domain conversational data is available, using only non-conversational domain signals.

Data Augmentation for CRS. Data scarcity is a central challenge for CRS, where cold-start issues are common and large-scale conversational interactions are difficult to obtain (Chae et al., 2020; Mysore et al., 2023; Leszczynski et al., 2023; Wang et al., 2023b). Recent augmentation methods use LLMs to simulate conversations involving long-tail items (Zhao et al., 2023), generate synthetic narrative queries from user-item histories (Mysore et al., 2023), synthesize music-domain conversations from item collections (Leszczynski et al., 2023), or enrich existing dialogues with retrieved items, hard negatives, and neighboring conversations (Jandaghi et al., 2024; Xu et al., 2025; Liu et al., 2025; Wang et al., 2023b). These methods still operate *on top of* existing in-domain conversations. LLM-REDIAL (Liang et al., 2024) instead verbalizes user histories into a large multi-domain synthetic corpus, complementary to our focus on which non-conversational signals and selection strategies yield the most useful supervision under a fixed teacher budget. Our zero-data setting bootstraps conversational training data entirely from reviews, metadata, and collaborative interactions, without any pre-existing conversational corpus.

Active Learning. Active learning has been used in recommender systems and NLP to maximize knowledge transfer from LLMs while minimizing costly queries (Du et al., 2025; Zhang et al., 2024; Liu et al., 2024; Palo et al., 2024; Li et al., 2024c; Xia et al., 2025). Prior work selects informative user sessions for teacher annotation (Du et al., 2025), generates targeted examples from student errors (Liu et al., 2024), or guides query selection with model explanations and teacher corrections (Zhang et al., 2024). We apply established information-theoretic criteria (Jensen-Shannon divergence and Fisher information) to a different problem: selecting non-conversational seed instances for synthetic CRS data generation. Our contribution is not a new active learning criterion, but a systematic empirical study of how active selection interacts with CRS-specific domain signals in a zero-data pipeline.

6 Conclusion

We presented a systematic empirical study of zero-data CRS bootstrapping: generating conversational supervision from non-conversational domain signals via LLMs. Across models, benchmarks, and

fine-tuning paradigms, domain-grounded synthetic data outperforms zero-shot and naïve synthetic baselines; active selection improves data efficiency; metadata and collaborative signals strengthen selection; and synthetic data is especially effective when real dialogues are scarce. These findings show that non-conversational signals, when systematically selected and transformed, offer a practical path to building CRS without any conversational training data.

LLM Usage: We used LLMs solely for grammar refinement and wording edits in drafting parts of this paper.

Limitations

We evaluate exclusively on the movie domain. The selection and bootstrapping mechanics are domain-general, but transfer to domains such as music or e-commerce would require domain-specific seed corpora, generation templates, and evaluation benchmarks; establishing cross-domain generalization is therefore left to future work. The bootstrapping pipeline relies on LLM-generated synthetic data, which may contain hallucinations or stylistic artifacts that propagate to downstream models. Finally, the quality of generated dialogues depends on prompt design. While recent work on prompt optimization could potentially improve generation quality, we do not explore this direction.

References

- Lora Aroyo, Alex Taylor, Mark Díaz, Christopher Homan, Alicia Parrish, Gregory Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. 2023. [Dices dataset: Diversity in conversational ai evaluation for safety](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 53330–53342. Curran Associates, Inc.
- Jordan T. Ash, Surbhi Goel, Akshay Krishnamurthy, and Sham Kakade. 2021. [Gone fishing: Neural active learning with fisher embeddings](#). *Preprint*, arXiv:2106.09675.
- Dong-Kyu Chae, Jihoo Kim, Duen Horng Chau, and Sang-Wook Kim. 2020. [Ar-cf: Augmenting virtual users and items in collaborative filtering for addressing cold-start problems](#). In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’20, page 1251–1260, New York, NY, USA. Association for Computing Machinery.
- Lingjiao Chen, Matei Zaharia, and James Zou. 2024a. [FrugalGPT: How to use large language models while](#)

- reducing cost and improving performance. *Transactions on Machine Learning Research*. Featured Certification.
- Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023. [PLACES: Prompting language models for social conversation synthesis](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 844–868, Dubrovnik, Croatia. Association for Computational Linguistics.
- Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. 2019. [Towards knowledge-based recommender dialog system](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1803–1813, Hong Kong, China. Association for Computational Linguistics.
- Xiang Chen, Chaoyang Gao, Chunyang Chen, Guangbei Zhang, and Yong Liu. 2025. [An empirical study on challenges for llm application developers](#). *ACM Trans. Softw. Eng. Methodol.*, 34(7).
- Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024b. Self-play fine-tuning converts weak language models to strong language models. In *Proceedings of the 41st International Conference on Machine Learning, ICML’24*. JMLR.org.
- Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. [Towards conversational recommender systems](#). In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’16*, page 815–824, New York, NY, USA. Association for Computing Machinery.
- Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2024. [How abilities in large language models are affected by supervised fine-tuning data composition](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 177–198, Bangkok, Thailand. Association for Computational Linguistics.
- Yingpeng Du, Zhu Sun, Ziyang Wang, Haoyan Chua, Jie Zhang, and Yew-Soon Ong. 2025. [Active large language model-based knowledge distillation for session-based recommendation](#). In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI’25/IAAI’25/EAAI’25*. AAAI Press.
- Yue Feng, Shuchang Liu, Zhenghai Xue, Qingpeng Cai, Lantao Hu, Peng Jiang, Kun Gai, and Fei Sun. 2023. [A large language model enhanced conversational recommender system](#). *Preprint*, arXiv:2308.06212.
- Luke Friedman, Sameer Ahuja, David Allen, Zhenning Tan, Hakim Sidahmed, Changbo Long, Jun Xie, Gabriel Schubiner, Ajay Patel, Harsh Lara, Brian Chu, Zexi Chen, and Manoj Tiwari. 2023. [Leveraging large language models in conversational recommender systems](#). *Preprint*, arXiv:2305.07961.
- Drew Fudenberg, Gleb Romanyuk, and Philipp Strack. 2017. [Active learning with a misspecified prior](#). *Theoretical Economics*, 12(3):1155–1189.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Shirley Anugrah Hayati, Dongyeop Kang, Qingxi-aoyang Zhu, Weiyan Shi, and Zhou Yu. 2020. [INSPIRED: Toward sociable recommendation dialog systems](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8142–8152, Online. Association for Computational Linguistics.
- Zhankui He, Zhouhang Xie, Rahul Jha, Harald Steck, Dawen Liang, Yesu Feng, Bodhisattwa Prasad Majumder, Nathan Kallus, and Julian McAuley. 2023. [Large language models as zero-shot conversational recommenders](#). In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM ’23*, page 720–730, New York, NY, USA. Association for Computing Machinery.
- Zhankui He, Handong Zhao, Tong Yu, Sungchul Kim, Fan Du, and Julian McAuley. 2022. [Bundle mcr: Towards conversational bundle recommendation](#). In *Proceedings of the 16th ACM Conference on Recommender Systems, RecSys ’22*, page 288–298, New York, NY, USA. Association for Computing Machinery.
- Yupeng Hou, Jiacheng Li, Xiangjun Fu, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2026. [Bridging language and items for retrieval and recommendation: Benchmarking LLMs as semantic encoders](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3251–3265, San Diego, California, United States. Association for Computational Linguistics.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.

- Abiodun M. Ikotun, Absalom E. Ezugwu, Laith Abualigah, Belal Abuhaija, and Jia Heming. 2023. [K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data](#). *Information Sciences*, 622:178–210.
- Pegah Jandaghi, Xianghai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. 2024. [Faithful persona-based conversational dataset generation with large language models](#). In *Proceedings of the 6th Workshop on NLP for Conversational AI (NLP4ConvAI 2024)*, pages 114–139, Bangkok, Thailand. Association for Computational Linguistics.
- Cheonsu Jeong. 2024. Domain-specialized llm: Financial fine-tuning and utilization method using mistral 7b. *Journal of Intelligence and Information Systems*, 30(1):93–120.
- Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. 2024. [Scaling sentence embeddings with large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3182–3196, Miami, Florida, USA. Association for Computational Linguistics.
- Enkelejda Kasneci, Kathrin Sessler, Stefan Küchermann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, and 4 others. 2023. [Chatgpt for good? on opportunities and challenges of large language models for education](#). *Learning and Individual Differences*, 103:102274.
- Andreas Kirsch and Yarin Gal. 2022. [Unifying approaches in active learning and active sampling via fisher information and information-theoretic quantities](#). *Preprint*, arXiv:2208.00549.
- Branislav Kveton, Xintong Li, Julian McAuley, Ryan Rossi, Jingbo Shang, Junda Wu, and Tong Yu. 2025. [Active learning for direct preference optimization](#). *Preprint*, arXiv:2503.01076.
- Wenqiang Lei, Xiangnan He, Yisong Miao, Qingyun Wu, Richang Hong, Min-Yen Kan, and Tat-Seng Chua. 2020. [Estimation-action-reflection: Towards deep interaction between conversational and recommender systems](#). In *Proceedings of the 13th International Conference on Web Search and Data Mining, WSDM '20*, page 304–312, New York, NY, USA. Association for Computing Machinery.
- Megan Leszczynski, Shu Zhang, Ravi Ganti, Krisztian Balog, Filip Radlinski, Fernando Pereira, and Arun Tejasvi Chaganty. 2023. [Talk the walk: Synthetic data generation for conversational music recommendation](#). *Preprint*, arXiv:2301.11489.
- Jiaxiang Li, Siliang Zeng, Hoi To Wai, Chenliang Li, Alfredo Garcia, and Mingyi Hong. 2024a. [Getting more juice out of the SFT data: Reward learning from human demonstration improves SFT for LLM alignment](#). In *ICML 2024 Workshop on Theoretical Foundations of Foundation Models*.
- Qinbin Li, Junyuan Hong, Chulin Xie, Jeffrey Tan, Rachel Xin, Junyi Hou, Xavier Yin, Zhun Wang, Dan Hendrycks, Zhangyang Wang, Bo Li, Bingsheng He, and Dawn Song. 2024b. [Llm-pbe: Assessing data privacy in large language models](#). *Proc. VLDB Endow.*, 17(11):3201–3214.
- Quan Li, Tianxiang Zhao, Lingwei Chen, Junjie Xu, and Suhang Wang. 2024c. [Enhancing graph neural networks with limited labeled data by actively distilling knowledge from large language models](#). *Preprint*, arXiv:2407.13989.
- Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. [Towards deep conversational recommendations](#). In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.
- Ziyue Li and Tianyi Zhou. 2025. [Your mixture-of-experts LLM is secretly an embedding model for free](#). In *The Thirteenth International Conference on Learning Representations*.
- Tingting Liang, Chenxin Jin, Lingzhi Wang, Wenqi Fan, Congying Xia, Kai Chen, and Yuyu Yin. 2024. [LLM-REDIAL: A large-scale dataset for conversational recommender systems created from user behaviors with LLMs](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8926–8939, Bangkok, Thailand. Association for Computational Linguistics.
- Allen Lin, Ziwei Zhu, Jianling Wang, and James Caverlee. 2023. [Enhancing user personalization in conversational recommenders](#). In *Proceedings of the ACM Web Conference 2023, WWW '23*, page 770–778, New York, NY, USA. Association for Computing Machinery.
- J. Lin. 1991. [Divergence measures based on the shannon entropy](#). *IEEE Transactions on Information Theory*, 37(1):145–151.
- Yong Lin, Chen Liu, Chenlu Ye, Qing Lian, Yuan Yao, and Tong Zhang. 2025. [Optimal sample selection through uncertainty estimation and its application in deep learning](#). *Journal of Machine Learning Research*, 26(128):1–47.
- Chengyuan Liu, Fubang Zhao, Kun Kuang, Yangyang Kang, Zhuoren Jiang, Changlong Sun, and Fei Wu. 2024. [Evolving knowledge distillation with large language models and active learning](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6717–6731, Torino, Italia. ELRA and ICCL.
- Yuanxing Liu, Jiahuan Pei, Wei-Nan Zhang, Ming Li, Wanxiang Che, and Maarten de Rijke. 2025. [Augmentation with neighboring information for conversational recommendation](#). *ACM Trans. Inf. Syst.*, 43(3).

- Ahmed Menshawy, Zeeshan Nawaz, and Mahmoud Fahmy. 2024. [Navigating challenges and technical debt in large language models deployment](#). In *Proceedings of the 4th Workshop on Machine Learning and Systems*, EuroMLSys '24, page 192–199, New York, NY, USA. Association for Computing Machinery.
- Subhojyoti Mukherjee, Anusha Lalitha, Kousha Kalantari, Aniket Anand Deshmukh, Ge Liu, Yifei Ma, and Branislav Kveton. 2024. [Optimal design for human preference elicitation](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Sheshera Mysore, Andrew McCallum, and Hamed Zamani. 2023. [Large language model augmented narrative driven recommendations](#). In *Proceedings of the 17th ACM Conference on Recommender Systems*, RecSys '23, page 777–783, New York, NY, USA. Association for Computing Machinery.
- Seth Neel and Peter Chang. 2024. [Privacy issues in large language models: A survey](#). *Preprint*, arXiv:2312.06717.
- George Nemhauser, Laurence Wolsey, and M. Fisher. 1978. [An analysis of approximations for maximizing submodular set functions—i](#). *Mathematical Programming*, 14:265–294.
- Zoey Nguyen, Anthony Annunziata, Vinh Luong, Sang Dinh, Quynh Le, Anh Hai Ha, Chanh Le, Hong An Phan, Shruti Raghavan, and Christopher Nguyen. 2024. [Enhancing q&a with domain-specific fine-tuning and iterative reasoning: A comparative study](#). *Preprint*, arXiv:2404.11792.
- OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. [Gpt-4o system card](#). *Preprint*, arXiv:2410.21276.
- Flavio Di Palo, Prateek Singhi, and Bilal H Fadlallah. 2024. [Performance-guided LLM knowledge distillation for efficient text classification at scale](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3675–3687, Miami, Florida, USA. Association for Computational Linguistics.
- Yotam Perlitz, Ariel Gera, Michal Shmueli-Scheuer, Dafna Sheinwald, Noam Slonim, and Liat Ein-Dor. 2023. [Active learning for natural language generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9862–9877, Singapore. Association for Computational Linguistics.
- Kaare Brandt Petersen, Michael Syskind Pedersen, et al. 2008. The matrix cookbook. *Technical University of Denmark*, 7(15):510.
- Friedrich Pukelsheim. 2006. *Optimal design of experiments*. SIAM.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Reddit. 2023a. [Reddit data api update: Changes to pushshift access](#). Accessed: 2025-03-30.
- Reddit. 2023b. [An update regarding reddit’s api](#). Accessed: 2025-03-30.
- Reddit. 2024. [Publishing our public content policy and introducing a new community for researchers](#). Accessed: 2025-03-30.
- Daniel C. Ruiz and John Sell. 2024. [Fine-tuning and evaluating open-source large language models for the army domain](#). *Preprint*, arXiv:2410.20297.
- Siddharth Samsi, Dan Zhao, Joseph McDonald, Baolin Li, Adam Michaleas, Michael Jones, William Bergeron, Jeremy Kepner, Devesh Tiwari, and Vijay Gadepally. 2023. [From words to watts: Benchmarking the energy costs of large language model inference](#). *Preprint*, arXiv:2310.03003.
- Ozan Sener and Silvio Savarese. 2018. [Active learning for convolutional neural networks: A core-set approach](#). In *International Conference on Learning Representations*.
- C. E. Shannon. 1948. [A mathematical theory of communication](#). *Bell System Technical Journal*, 27(3):379–423.
- Jack Sherman and Winifred J. Morrison. 1950. [Adjustment of an inverse matrix corresponding to a change in one element of a given matrix](#). *The Annals of Mathematical Statistics*, 21(1):124–127.
- Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. 2024. [Rethinking interpretability in the era of large language models](#). *Preprint*, arXiv:2402.01761.
- Heydar Soudani, Roxana Petcu, Evangelos Kanoulas, and Faegheh Hasibi. 2026. [A survey on recent advances in conversational data generation](#). *ACM Comput. Surv.*, 58(10).
- Masashi Sugiyama. 2005. [Active learning for misspecified models](#). In *Advances in Neural Information Processing Systems*, volume 18. MIT Press.
- Jianwei Sun, Chaoyang Mei, Linlin Wei, Kaiyu Zheng, Na Liu, Ming Cui, and Tianyi Li. 2024. [Dial-insight: Fine-tuning large language models with high-quality domain-specific data preventing capability collapse](#). *Preprint*, arXiv:2403.09167.

- Yueming Sun and Yi Zhang. 2018. [Conversational recommender system](#). In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '18*, page 235–244, New York, NY, USA. Association for Computing Machinery.
- Yixuan Tang and Yi Yang. 2024. [Pooling and attention: What are effective designs for llm-based embedding models?](#) *Preprint*, arXiv:2409.02727.
- Tomasz Walkowiak and Mateusz Gniewkowski. 2019. [Evaluation of vector embedding models in clustering of text documents](#). In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1304–1311, Varna, Bulgaria. INCOMA Ltd.
- Jian Wang, Yi Cheng, Dongding Lin, Chak Leong, and Wenjie Li. 2023a. [Target-oriented proactive dialogue systems with personalization: Problem formulation and dataset curation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1132–1143, Singapore. Association for Computational Linguistics.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. [Improving text embeddings with large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, Bangkok, Thailand. Association for Computational Linguistics.
- Xi Wang, Hossein Rahmani, Jiqun Liu, and Emine Yilmaz. 2023b. [Improving conversational recommendation systems via bias analysis and language-model-enhanced data augmentation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3609–3622, Singapore. Association for Computational Linguistics.
- Xiaolei Wang, Kun Zhou, Xinyu Tang, Wayne Xin Zhao, Fan Pan, Zhao Cao, and Ji-Rong Wen. 2023c. [Improving conversational recommendation systems via counterfactual data simulation](#). In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '23*, page 2398–2408, New York, NY, USA. Association for Computing Machinery.
- Xiaolei Wang, Kun Zhou, Ji-Rong Wen, and Wayne Xin Zhao. 2022. [Towards unified conversational recommender systems via knowledge-enhanced prompt learning](#). In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '22*, page 1929–1937, New York, NY, USA. Association for Computing Machinery.
- Yeo Wei Jie, Ranjan Satapathy, Rick Goh, and Erik Cambria. 2024. [How interpretable are reasoning explanations from prompting large language models?](#) In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2148–2164, Mexico City, Mexico. Association for Computational Linguistics.
- Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, Hui Xiong, and Enhong Chen. 2024. [A survey on large language models for recommendation](#). *World Wide Web*, 27(5).
- Yunjia Xi, Weiwen Liu, Jianghao Lin, Bo Chen, Ruiming Tang, Weinan Zhang, and Yong Yu. 2024. [Memocrs: Memory-enhanced sequential conversational recommender systems with large language models](#). In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM '24*, page 2585–2595, New York, NY, USA. Association for Computing Machinery.
- Yu Xia, Subhojyoti Mukherjee, Zhouhang Xie, Junda Wu, Xintong Li, Ryan Aponte, Hanjia Lyu, Joe Barrow, Hongjie Chen, Franck Dernoncourt, Branislav Kveton, Tong Yu, Ruiyi Zhang, Jiuxiang Gu, Nisreen K. Ahmed, Yu Wang, Xiang Chen, Hanieh Deilamsalehy, Sungchul Kim, and 15 others. 2025. [From selection to generation: A survey of LLM-based active learning](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14552–14569, Vienna, Austria. Association for Computational Linguistics.
- Yu Xia, Junda Wu, Tong Yu, Sungchul Kim, Ryan A. Rossi, and Shuai Li. 2023. [User-regulation deconfounded conversational recommender system with bandit feedback](#). In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '23*, page 2694–2704, New York, NY, USA. Association for Computing Machinery.
- Zhouhang Xie, Junda Wu, Hyunsik Jeon, Zhankui He, Harald Steck, Rahul Jha, Dawen Liang, Nathan Kallus, and Julian McAuley. 2024. [Neighborhood-based collaborative filtering for conversational recommendation](#). In *Proceedings of the 18th ACM Conference on Recommender Systems, RecSys '24*, page 1045–1050, New York, NY, USA. Association for Computing Machinery.
- Haozhe Xu, Xiaohua Wang, Changze Lv, and Xiaoqing Zheng. 2025. [Beyond single labels: Improving conversational recommendation through LLM-powered data augmentation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15573–15590, Vienna, Austria. Association for Computational Linguistics.
- Biwei Yan, Kun Li, Minghui Xu, Yueyan Dong, Yue Zhang, Zhaochun Ren, and Yiuzhen Cheng. 2025. [On protecting the data privacy of large language models \(llms\) and llm agents: A literature review](#). *High-Confidence Computing*, 5(2):100300.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41

others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.

Ting Yang and Li Chen. 2024. [Unleashing the retrieval potential of large language models in conversational recommender systems](#). In *Proceedings of the 18th ACM Conference on Recommender Systems, RecSys '24*, page 43–52, New York, NY, USA. Association for Computing Machinery.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. [A survey on large language model \(llm\) security and privacy: The good, the bad, and the ugly](#). *High-Confidence Computing*, 4(2):100211.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Guoyin Wang, and Fei Wu. 2026. [Instruction tuning for large language models: A survey](#). *ACM Comput. Surv.*, 58(7).

Yifei Zhang, Bo Pan, Chen Ling, Yuntong Hu, and Liang Zhao. 2024. [ELAD: Explanation-guided large language models active distillation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4463–4475, Bangkok, Thailand. Association for Computational Linguistics.

Yiming Zhang, Lingfei Wu, Qi Shen, Yitong Pang, Zhihua Wei, Fangli Xu, Bo Long, and Jian Pei. 2022. [Multiple choice questions based multi-interest policy learning for conversational recommendation](#). In *Proceedings of the ACM Web Conference 2022, WWW '22*, page 2153–2162, New York, NY, USA. Association for Computing Machinery.

Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. 2018. [Towards conversational search and recommendation: System ask, user respond](#). In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM '18*, page 177–186, New York, NY, USA. Association for Computing Machinery.

Zipeng Zhao, Kun Zhou, Xiaolei Wang, Wayne Xin Zhao, Fan Pan, Zhao Cao, and Ji-Rong Wen. 2023. [Alleviating the long-tail problem in conversational recommender systems](#). In *Proceedings of the 17th ACM Conference on Recommender Systems, RecSys '23*, page 374–385, New York, NY, USA. Association for Computing Machinery.

Kun Zhou, Wayne Xin Zhao, Shuqing Bian, Yuanhang Zhou, Ji-Rong Wen, and Jingsong Yu. 2020. [Improving conversational recommender systems via knowledge graph based semantic fusion](#). In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '20*, page 1006–1014, New York, NY, USA. Association for Computing Machinery.

A Selection Criteria Details

This appendix provides the construction details for the two active selection criteria defined in Section 2.3.

JS Diversity. We partition the embedding space using K -means clustering to obtain cluster centers $C = \{\mathbf{c}_1, \dots, \mathbf{c}_K\} \subset \mathbb{R}^d$. For each embedding $\mathbf{x}_i$, we compute its Euclidean distance to each cluster center (Walkowiak and Gniewkowski, 2019; Ikotun et al., 2023), $d(\mathbf{x}_i, \mathbf{c}_j) = \|\mathbf{x}_i - \mathbf{c}_j\|$, and construct a soft cluster-membership distribution $p_i = \{p_{ij}\}_{j=1}^K$ via softmax:

$$p_{ij} = \frac{\exp(-d(\mathbf{x}_i, \mathbf{c}_j))}{\sum_{k=1}^K \exp(-d(\mathbf{x}_i, \mathbf{c}_k))}, \quad j = 1, \dots, K.$$

High entropy $H(p_i)$ indicates that $\mathbf{x}_i$ lies close to multiple centroids, while low entropy indicates strong association with a single cluster. We initialize $\mathcal{R}_0 = \{i^{(1)}\}$ with the maximum-entropy sample (Shannon, 1948):

$$i^{(1)} = \arg \max_{i \in \mathcal{D}} H(p_i),$$

$$H(p_i) = - \sum_{j=1}^K p_{ij} \log p_{ij}.$$

Subsequent samples are selected by Equation (1). The entropy term favors boundary samples, while the JS term (Lin, 1991) promotes novelty relative to the selected set. We update $\mathcal{R}_{t+1} = \mathcal{R}_t \cup \{i^{(t)}\}$ and repeat until $|\mathcal{R}| = B$.

Fisher Information. Under a last-layer linearization of the fine-tuning objective (Mukherjee et al., 2024; Kveton et al., 2025), greedy log-determinant selection approximates expected information gain (Kirsch and Gal, 2022). We initialize $\Sigma_0 = I_d$ and $\Lambda_0 = I_d$. Given selected set $\mathcal{R}_t$, the regularized information matrix is $\Sigma_t = I_d + \sum_{j \in \mathcal{R}_t} \mathbf{x}_j \mathbf{x}_j^\top$, $\Lambda_t = \Sigma_t^{-1}$. The next sample is selected by Equation (2), corresponding to the largest incremental information gain (Pukelsheim, 2006; Ash et al., 2021; Petersen et al., 2008). After selecting $i^{(t)}$, we update $\mathcal{R}_{t+1} = \mathcal{R}_t \cup \{i^{(t)}\}$ and update Λ_t using the rank-1 Sherman–Morrison formula (Sherman and Morrison, 1950):

$$\Lambda_{t+1} \leftarrow \Lambda_t - \frac{(\Lambda_t \mathbf{x}_{i^{(t)}})(\Lambda_t \mathbf{x}_{i^{(t)}})^\top}{1 + \mathbf{x}_{i^{(t)}}^\top \Lambda_t \mathbf{x}_{i^{(t)}}}.$$

The Fisher criterion is equivalent to greedy D-optimal design (Pukelsheim, 2006): each step

maximizes $\log \det(\Sigma_{t+1}) - \log \det(\Sigma_t)$, reducing the volume of the confidence ellipsoid for model parameters under the last-layer approximation. The JS criterion is motivated by submodular coverage objectives (Nemhauser et al., 1978). Together, JS and Fisher represent complementary notions of sample utility: distributional coverage and parameter-aligned informativeness.

B Popularity Bias and Coverage Analysis

We further characterize the active-selection results in RQ2 by analyzing the popularity profile of selected seed items and the coverage of the resulting recommendations. These diagnostics show that information-theoretic selection identifies useful seeds across the item distribution and produces broader recommendation coverage.

B.1 Selected-Item Popularity

We partition seed items by Amazon review count into *head* (top 20%), *mid* (next 30%), and *tail* (bottom 50%) buckets. For budgets $B \in \{1k, 2k, 3k\}$, we report selected-bucket proportions, the Gini coefficient of the selected-item distribution, and the Spearman correlation ρ between item popularity and selection probability.

B	Method	Head	Mid	Tail	Gini	ρ
1k	Random	20.6	32.5	46.9	.706	+.002
	JS	22.7	32.6	44.7	.677	+.042
	Fisher	14.8	23.3	61.9	.787	−.172
2k	Random	20.7	31.5	47.9	.702	+.001
	JS	22.7	32.8	44.5	.684	+.065
	Fisher	18.2	26.1	55.8	.763	−.154
3k	Random	20.4	32.0	47.6	.717	+.006
	JS	22.5	32.2	45.3	.700	+.076
	Fisher	18.4	27.8	53.8	.763	−.145

Table 4: Popularity composition of selected samples at budgets $B \in \{1k, 2k, 3k\}$ (Qwen2.5-1.5B). Head/Mid/Tail denote selected-item percentages by Amazon review-count bucket. Lower Gini indicates more equitable coverage; ρ is the Spearman correlation between item popularity and selection probability.

Table 4 shows a clear distinction between popularity-based selection and the active criteria. Fisher has a strong long-tail orientation: tail items account for 53.8–61.9% of selected samples, and selection probability is negatively correlated with review-count popularity across all budgets ($\rho = -.172$ to $-.145$). JS remains closer to the seed-pool distribution while maintaining only a mild positive popularity correlation ($\rho = +.042$

to $+.076$). Thus, the active strategies select useful seeds from broader regions of the item space, with Fisher especially highlighting informative long-tail items. Because these statistics are computed from the selected set alone, before any teacher calls, they offer a low-cost check on selection quality: a concentrated or low-coverage selection can be corrected by falling back to uniform-random sampling, which RQ2 shows remains a usable baseline at every budget.

B.2 Recommendation Coverage

We next evaluate how these selection patterns translate into model outputs. We bucket test-set ground-truth items by evaluation frequency and report Recall@5 for each bucket using Qwen2.5-1.5B. We also report Coverage@ K , the number of unique items appearing in top- K recommendations across all test conversations.

Method	ReDial R@5			INSPIRED R@5		
	Head	Mid	Tail	Head	Mid	Tail
Random	9.44	5.58	3.70	5.47	3.83	3.06
JS	11.37	7.44	3.89	5.89	4.94	3.06
Fisher	12.10	7.54	4.44	6.25	4.78	2.64

Table 5: Stratified Recall@5 by item popularity bucket on ReDial and INSPIRED (Qwen2.5-1.5B). Items are partitioned into head, mid, and tail buckets by evaluation-set frequency.

Method	ReDial		INSPIRED	
	Cov@5	Cov@10	Cov@5	Cov@10
Random	1030	1407	502	696
JS	1164	1571	569	802
Fisher	1188	1570	547	777

Table 6: Coverage@ K : number of unique items in top- K recommendations across all test conversations (Qwen2.5-1.5B). Higher values indicate greater recommendation diversity.

On ReDial, both active strategies improve Recall@5 over Random across head, mid, and tail buckets (Table 5); Fisher raises tail Recall@5 from 3.70 to 4.44 while also improving head Recall@5 from 9.44 to 12.10. On INSPIRED, JS improves head and mid performance while maintaining tail Recall@5, and Fisher gives the strongest head performance among the three methods. Coverage shows the most consistent pattern: JS and Fisher recommend substantially more unique items than Random on both datasets (Table 6). Overall, active

selection improves both recommendation accuracy and coverage, reinforcing the RQ2 conclusion that information-theoretic seed selection is more effective than random sampling.

C Ranking Quality (NDCG)

Table 7 reports NDCG for the RQ2 selection strategies at the full generation budget (Qwen2.5-1.5B). We omit NDCG@1, which is identical to Recall@1 when each evaluation instance has a single ground-truth item.

Method	ReDial			INSPIRED		
	@5	@10	@20	@5	@10	@20
Random	5.82	6.53	7.03	3.41	3.85	4.27
Popularity	6.05	6.75	7.27	3.28	3.73	4.10
JS	7.40	8.13	8.62	3.76	4.30	4.71
Fisher	7.47	8.19	8.70	3.68	4.25	4.59

Table 7: NDCG@ k (%) at the full generation budget (Qwen2.5-1.5B). Both active criteria outperform random and popularity-based selection at every cutoff on both benchmarks.

NDCG tracks Recall closely: both active criteria improve over random and popularity selection at every cutoff on both benchmarks, and their advantage varies with budget in the same way as the Recall curves in Figure 3. Active selection therefore improves the rank position of correct items, not only their presence in the list.

D Synthetic Data Quality

Because the pipeline depends on LLM generation, we assess the generated corpus directly along two axes: human judgments of query-recommendation quality, and automatic validation of output format and item validity.

D.1 Human Evaluation

Two graduate students independently rated 100 randomly sampled synthetic query-recommendation pairs on a 1-5 scale along four dimensions: naturalness (does the query read like a real user request), coherence (is the request internally consistent), relevance (do the recommendations match the request), and diversity (is the list varied rather than repetitive). Agreement within one point is 83-90% across dimensions.

All four dimensions average close to 4, and averaging the two annotators, 84% of the 400 example-dimension scores are at least 3.5. Relevance is

Dimension	Mean score
Naturalness	3.92
Coherence	3.94
Relevance	3.79
Diversity	4.18

Table 8: Human evaluation of 100 sampled synthetic query-recommendation pairs (1-5 scale, two annotators). Averaging the two annotators, 84% of the 400 example-dimension scores are at least 3.5.

the lowest-scoring dimension, consistent with the teacher occasionally returning loosely related items in the 20-item list.

D.2 Title and Format Validation

We normalize every generated title (lowercasing, removing release years and punctuation) and check it against a verified catalog of 9,944 titles compiled from ReDial, INSPIRED, and the seed data.

Corpus	In-catalog	Exactly 20	Dup. lists
Domain-grounded	76.5	97.3	2.6
Naïve GPT	92.3	100.0	1.3

Table 9: Format and item validity (%) of the generated corpora. In-catalog: generated titles matching the verified catalog. Exactly 20: lists with the requested number of items. Dup. lists: lists containing a repeated title.

Both corpora follow the requested format almost always. Exact matching is a lower bound on validity, since long-tail, international, and alternatively formatted titles are absent from this small catalog. Notably, the naïve corpus matches the catalog *more* often yet gains far less downstream (Table 3): catalog overlap rewards popular titles, not useful ones.

E Prompts and Implementation Details

E.1 Implementation Details

All experiments are implemented using PyTorch and conducted on NVIDIA A6000 GPUs with 48GB memory. Fine-tuning is a one-time cost of under four hours per run on a single A6000; at inference the fine-tuned model serves requests without further teacher calls. For training stability, we use the Adam optimizer coupled with a warm-up learning rate scheduler. The batch size is set to 32, with gradient accumulation steps of 8.

Training configurations are tailored to each research question: for RQ1 (synthetic data comparison), we use a learning rate of 5×10^{-6} and train

for 5 epochs with early stopping to ensure convergence; for RQ2 (active-selection budget curves), we use a learning rate of 5×10^{-6} and train for 1 epoch; for RQ3 (domain-signal comparisons), we use a learning rate of 5×10^{-6} and train for 1 epoch; for RQ4 (real/synthetic training mixtures), we use a learning rate of 1×10^{-5} and train for 3 epochs. We synthesize conversational datasets following the procedure outlined in Section 2.4. Synthetic data is generated using OpenAI’s GPT-4o API (OpenAI et al., 2024), with a generation temperature of 0.8. We fix $\lambda = 0.5$ (Equation (1)) in all experiments.

E.2 Experiments

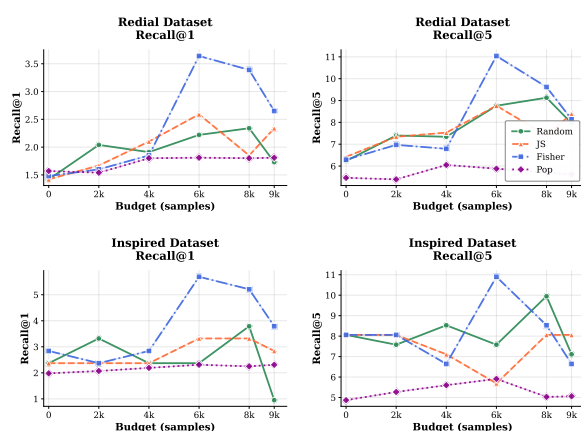


Figure 7: RQ2: Llama3-3B performance across budgets on ReDial and INSPIRED. Active selection strategies (JS, Fisher) consistently outperform random sampling and popularity-based selection.

Sample generated user-query and assistant response

User Query 1

Movies with a unique twist on classic tales? I’m interested in films that take a well-known story and give it a fresh spin. I recently watched a Hercules movie that wasn’t your typical mythological tale—it had humor and The Rock was pretty awesome, even if his acting chops are still a work in progress. Can you suggest other films like that? Maybe ones that add humor or a cool new perspective to a classic narrative?

Assistant Response: *Shrek; Enchanted; Into the Woods; The Princess Bride; Stardust; 10 Things I Hate About You; Warm Bodies; Clueless; Ella Enchanted; Malefi-*

cent; The Lego Movie; Tangled; John Tucker Must Die; Easy A; The Many Adventures of Winnie the Pooh; Hook; Ever After: A Cinderella Story; The Nightmare Before Christmas; Pan; The Brothers Grimm.

User Query 2

Looking for a quirky, whimsical movie with a great ensemble cast and a bit of adventure. I’ve heard “The Brothers Bloom” might fit the bill with its delightful, offbeat charm and diverse settings. If you enjoy films like “Dirty Rotten Scoundrels” and appreciate a touch of the macabre mixed with a fun romp, this might be the one! Would love to know if there are other similar movies out there that capture this unique blend of humor and adventure. Any recommendations?

Assistant Response: *The Grand Budapest Hotel; The Royal Tenenbaums; Moonrise Kingdom; Snatch; Amélie; Big Fish; Lock, Stock and Two Smoking Barrels; The Adventures of Tintin; In Bruges; The Life Aquatic with Steve Zissou; The Darjeeling Limited; Midnight in Paris; Scott Pilgrim vs. the World; The Men Who Stare at Goats; Kiss Kiss Bang Bang; The Fantastic Mr. Fox; Delicatessen; O Brother, Where Art Thou?; Catch Me If You Can; The Imaginarium of Doctor Parnassus.*

User Query 3

My partner and I watched a film recently and thoroughly enjoyed it. However, I’m curious about movies where not much happens plot-wise, but dialogue and arguments drive the narrative. Any recommendations for films with this kind of storytelling approach? Also, any thoughts on movies that critique or play with horror fan stereotypes?

Assistant Response: *My Dinner with Andre; Before Sunrise; The Breakfast Club; Clerks; Dazed and Confused; The Big Chill; Slacker; Richard Linklater’s Waking Life; Coffee and Cigarettes; Frances Ha; Annie Hall; The Squid and the Whale; Glengarry Glen Ross; Tape; Carnage; What We Do in the Shadows; The Cabin in the Woods; Scream; Tucker and Dale vs. Evil; Behind the Mask: The Rise of Leslie Vernon.*

Movie Recommendations Prompt

Pretend you are a movie recommender system. I will give you a conversation between a user and you (a recommender system). Based on the conversation, you reply with 20 recommendations without extra sentences.

Here is the conversation: {conv}

Query Generation Prompt

You will be given multiple product reviews. Your task is to convert the sentiment, issues, or features mentioned in these reviews into one distinct query or question. I want you to style it similar to the following queries.

Sample Queries: {sample_queries_text}

Now, convert the following product reviews into one distinct query with a similar style.

Input Review: {product_reviews_text}

Sample Query Templates (Q_{tmp})

The following are example queries from Q_{tmp} , sourced from Reddit movie recommendation discussions. These templates guide the style and tone of synthetic query generation.

“Sick with Covid, dealing with some serious cabin fever. Any good carefree films that might lighten the mood?”

“Movies about the friends made along the way. I’m interested in movies where the main character continually meets and befriends new interesting characters. Maybe the main character is on a journey. Maybe the relationships with the new characters both begin and end during the story. Something like ‘Into the Wild’ would be perfect.”