The Whole is Better than the Sum: Using Aggregated Demonstrations in In-Context Learning for Sequential Recommendation

Lei Wang Ee-Peng Lim*
Singapore Management University
{lei.wang.2019, eplim}@smu.edu.sg

Abstract

Large language models (LLMs) have shown excellent performance on various NLP tasks. To use LLMs as strong sequential recommenders, we explore the in-context learning approach to sequential recommendation. We investigate the effects of instruction format, task consistency, demonstration selection, and number of demonstrations. As increasing the number of demonstrations in ICL does not improve accuracy despite using a long prompt, we propose a novel method called LLMSRec-Syn that incorporates multiple demonstration users into one aggregated demonstration. Our experiments on three recommendation datasets show that LLMSRec-Syn outperforms state-of-the-art LLM-based sequential recommendation methods. In some cases, LLMSRec-Syn can perform on par with or even better than supervised learning methods. Our code is publicly available at https://github.com/demoleiwang/LLMSRec_Syn.

1 Introduction

Motivation. Large language models (LLMs) are known to perform well as a zero-shot solution for many natural language processing tasks (Brown et al., 2020; Chowdhery et al., 2022; OpenAI, 2022; Qin et al., 2023). Recently, there are some works that focus on using LLMs to perform recommendation with promising accuracies (Hou et al., 2023; Wang and Lim, 2023; Liu et al., 2023a; Bao et al., 2023; Gao et al., 2023) and to provide explanations (Yang et al., 2023; Wang et al., 2023b). Most of these works developed LLM prompts for zero-shot sequential recommendation.

To investigate whether LLM can serve as a strong zero-shot sequential recommender, Hou et al. (2023) devised a prompt that is filled with historical items in chronological order, candidate items, and instruction to rank the candidate items. Wang and Lim (2023) proposed a three-step prompting method, where LLMs first summarizes the user preference based on the user's past interacted items. It then identifies representative items from the past interacted items that capture the user preference, and finally recommends items among the candidate items which are aligned with the representative items. Among the very few one-shot sequential recommendation works, Liu et al. (2023a) and Hou et al. (2023) explored in-context learning using the test user's second last item as the ground truth next-item and all earlier interacted items as input to create self-demonstrations. Nevertheless, previous experiments have shown that in-context learning (ICL) based sequential recommendation methods perform poorly compared with

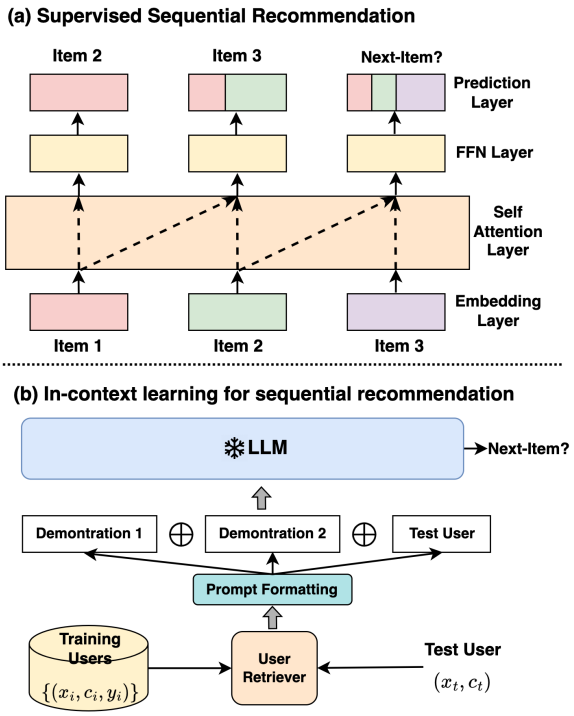

Figure 1: The illustrative comparison of (a) supervised sequential recommendation method and (b) in-context learning based sequential recommendation method.

*Corresponding author.

Findings of the Association for Computational Linguistics: NAACL 2024, pages 876–895

the supervised learning-based methods (e.g., SASRec) due to the complex recommendation task definition (Liu et al., 2023a; Hou et al., 2023; Wang and Lim, 2023). The illustrative comparison of these two methods is shown in Figure 1.

To develop an effective in-context learning approach for LLMs to perform sequential recommendation, we first define the sequential recommendation problem as follows.

Problem definition. We denote each input user instance u_i to be a (x_i, c_i, y_i) tuple where x_i denotes the sequence of past interacted items (excluding y_i) by u_i, c_i denotes the candidate items to be recommended ($|c_i| = M$), and y_i denotes the ground truth next-item which is also the last item interacted by u_i. Note that y_i appears in c_i ($y_i \in c_i$). A LLM-based sequential recommendation method is required to assign a rank $rank(d) \in [1, M]$ to each item d in c_i. Our objective is to ensure that the method ranks y_i, i.e., $rank(y_i)$, as high as possible for all.

The above definition includes c_i as input as it is usually infeasible for LLMs to take all items as input due to limited prompt length. Moreover, having c_i does not introduce bias in the evaluation. The above definition is also adopted in Hou et al. (2023). We also assume a dataset of users' interacted item sequences from which we can construct demonstrations for ICL, and a LLM which is too large for pretraining or finetuning.

Overview of our study. Past works has shown that the effectiveness of ICL in adapting LLMs to new tasks is significantly influenced by instruction wording (Madaan and Yazdanbakhsh, 2022; Yang et al., 2023), label design (Yoo et al., 2022; Wei et al., 2023), selection of demonstrations (Liu et al., 2021; Shi et al., 2022; Zhang et al., 2023b), and number of demonstrations Chen et al. (2023); Zhao et al. (2023). Our study thus begins by systematically investigating how the instruction format, task consistency (between test and demonstration), demonstration selection, and the number of demonstrations affect ICL-based sequential recommendation. Through our preliminary experiments, we obtain four findings including the one that observes degradation of recommendation accuracy when the number of demonstrations increases. As each demonstration takes up significant length, it is also easy for multiple demonstrations to exceed the prompt limit of LLMs. Moreover, as LLMs are known to miss out relevant information in a long input prompt (Liu et al., 2023b), we thus embark on a follow-up study on designing a more efficient ICL scheme based on *aggregated demonstration*.

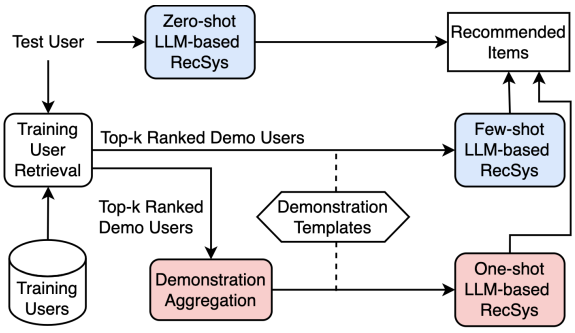

Figure 2: The overall framework of zero-shot, few-shot, and aggregated one-shot LLM-based sequential recommender systems.

Figure 2 shows a comparison of the frameworks for zero-shot, few-shot, and aggregated one-shot LLM-based sequential recommender systems. The key idea in aggregated demonstration is to combine multiple training users into one demonstration. This reduces the repetition of instruction text in the ICL prompt. It also seeks to summarize multiple training users relevant to the test instance in a compact manner. We also develop a novel ICL method using aggregated demonstration for sequential recommendation known as **LLMSRec-Syn**. The length of LLMSRec-Syn prompt increases only gradually with number of demonstration users, LLMSRec-Syn can cope with more relevant information from the demonstration users within a concise input context. We finally show LLMSRec-Syn outperforms other zero-shot and one-shot ICL methods in an extensive set of experiments.

Contribution. Our contributions can be summarized as follows: (1) We systematically explore the ICL approach to sequential recommendation by empirically investigating the effect of instruction format, task consistency, demonstration selection, and number of demonstrations; (2) We propose a new in-context learning method for sequential recommendation called LLMSRec-Syn which leverages on a novel concept of aggregated demonstration; (3) We experiment on three popular recommendation datasets and show that LLMSRec-Syn outperforms previous LLM-based sequential recommendation methods.

2 Related Work

In-Context Learning. Several works show that LLMs can effectively adapt to different NLP and multimodal tasks, including machine trans-

lation (Agrawal et al., 2022), visual question answering (Yang et al., 2022), and foreground segmentation (Zhang et al., 2023b). This adaptation is achieved by learning from a few task-relevant demonstrations, commonly known as in-context learning (ICL) (Brown et al., 2020). Despite the above successes, ICL's performance is still significantly affected by the wording of instructions (Madaan and Yazdanbakhsh, 2022; Yang et al., 2023), label design (Yoo et al., 2022; Wei et al., 2023), demonstration selection (Liu et al., 2021; Shi et al., 2022; Zhang et al., 2023b), and number of demonstrations Chen et al. (2023); Zhao et al. (2023). ICL is much less studied in LLM-based sequential recommendation. As sequential recommendation is distinct from the pretraining tasks of LLMs and also different from the above-mentioned tasks, new designs of demonstration(s) and ICL prompt is necessary.

LLMs for Sequential Recommendation. Early sequential recommendation works adopt techniques such as Markov Chains (Rendle et al., 2010; He and McAuley, 2016) and neural networks (e.g., RNN (Hidasi et al., 2015), CNN (Tang and Wang, 2018), Self-Attention (Kang and McAuley, 2018), and GNN (Chang et al., 2021)). To investigate if LLMs can be used as effective sequential recommenders without training, Hou et al. (2023) formulated sequential recommendation as conditional ranking, employing zero-shot LLM methods to reflect user preferences from past interactions and recency. Wang and Lim (2023) developed a three-step LLM prompting to summarize user preferences, while Hou et al. (2023) and Liu et al. (2023a) introduced a one-shot ICL method that utilizes the previous item interactions of the target user as a demonstration. To address position bias, Hou et al. (2023) proposed to randomize the candidate item order. In this work, we explore using training data demonstrations, not just user own history, and introduce aggregated demonstration for combining relevant users.

3 What Makes In-Context Learning Work for Sequential Recommendation

In this section, we conduct a preliminary empirical study to investigate the role of various aspects of demonstrations. These aspects include the wording of prompts, task consistency between demonstrations and test instances, selection of demonstrations, and number of demonstrations. While previous studies have explored the use of LLM as sequential recommenders in a zero-shot manner (Hou et al., 2023; Wang and Lim, 2023), this is the first study to comprehensively discuss how in-context learning can improve sequential recommendation.

Table 1: Dataset statistics after removing duplicate interactions and users or items with fewer than 5 interactions.

Datasets	ML-1M	LastFM-2K	Games
# Users	6,040	1,143	50,547
# Items	3,706	11,854	16,859
# User-item Interactions	1,000,209	68,436	389,718
Avg. interacted items per user	165.59	59.92	7.71
Avg. interacted users per item	269.88	5.77	23.11

3.1 Experiment Setup

We implement zero-shot, one-shot, and few-shot methods in this study, using three widely used recommendation datasets: the movie rating dataset *MovieLens-1M* (ML-1M) dataset, the category of *Games* from the Amazon Review dataset (McAuley et al., 2015), and the music artist listening dataset *LastFM-2K* (Cantador et al., 2011). The data statistics are summarized in Table 1. Taking into account cost-effectiveness of LLMs, we select 50 data examples from each of the three datasets to carry out all experiments for analysis in Section 3. Following the previous works (Hou et al., 2023; Wang and Lim, 2023), we use a leave-one-out strategy for evaluation, i.e., predicting the last interacted item of each user sequence and using the earlier interacted items as input. For each user sequence, we remove the last item, keeping it aside for testing. The rest of the sequence is used for training and validation. To evaluate the ranking results for each user u_i over a set of candidate items c_i, we adopt the widely used **NDCG@N** (N = 10, 20) as the evaluation metric. For MovieLens-1M and Games, we directly use the candidate sets utilized in an earlier work (Hou et al., 2023). For LastFM, we follow (Hou et al., 2023) and randomly select candidate items from the item universal set for each user sequence. We then insert the ground truth next item into the candidate item set. We use ChatGPT (`GPT-3.5-Turbo`) as the default LLM due to its excellent performance and cost-effectiveness. To ensure the reliability of findings, we repeat each experiment 9 times and report the average results. Without exception, we use ML-1M as an example for discussion.

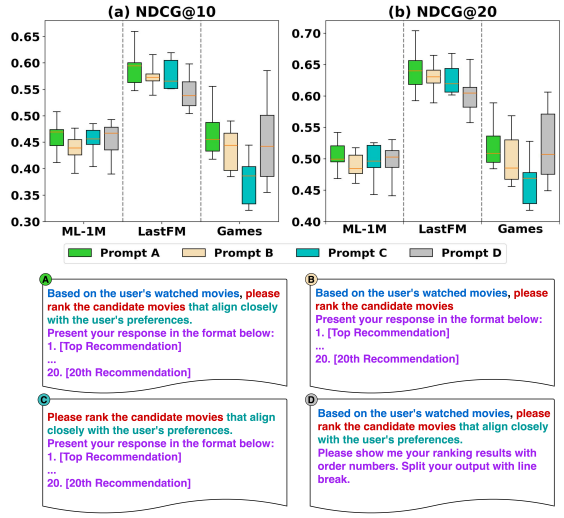

Figure 3: Instruction Format options: (A) Full, (B) w/o preference alignment, (C) w/o watched movie focus, (D) w/o rank result format

3.2 In-Context Learning for Sequential Recommendation

In ICL for sequential recommendation, one or a few training users are used as demonstrations that are included in the LLM prompt. Each demonstration thus includes a training user i's historical item interactions x_i, a set of candidates c_i, and ground truth next-item y_i. We denote the prompt capturing the demonstration user i by $\mathcal{T}(x_i, c_i, y_i)$. The following shows the concatenation of n demonstrations $\mathcal{C}$ which is appended by the instruction prompt for the test user $\mathcal{T}(x_{test}, y_{test})$ for prediction.

$$\mathcal{C} = \mathcal{T}(x_1, c_1, y_1) \oplus \cdots \oplus \mathcal{T}(x_n, c_n, y_n) \quad (1)$$

$$y_{\text{test}} \sim \mathcal{P}_{LLM}(\cdot \mid \mathcal{C} \oplus \mathcal{T}(x_{\text{test}}, c_{\text{test},\cdot})) \quad (2)$$

3.3 Wording of Instructions

LLMs have been found to be sensitive to wording of the prompt (Madaan and Yazdanbakhsh, 2022; Yang et al., 2023). For example, prompts (or instructions) that are semantically similar may yield significantly different results (Kojima et al., 2022; Zhou et al., 2022; Wang et al., 2023a; Zhang et al., 2023a). To examine the impact of instruction wording and exclude the influence of other factors such as demonstration labels and selection, we employ LLM as a zero-shot solver for sequential recommendation.

We discuss four different options for the instruction format to investigate the sensitivity of the LLM to the wording of the instruction. Considering the prompts used in LLM-based zero-shot recommendation models (Hou et al., 2023; Wang and Lim, 2023), we derive instructions with four possible mention components: (a) candidate item ranking, (b) user preference alignment, (c) historical interacted items, and (d) ranked result format. As recommendation is formulated as a ranking task, component (b) is mandatory. The *full* instruction covers all four components. To explore better instructions, we derive other instruction options by leaving out one of the remaining components. We thus have four instruction options: (A) full instruction $\mathcal{T}^{\text{A}}$, (B) full instruction without (b) $\mathcal{T}^{\text{B}}$, (C) full instruction without (c) $\mathcal{T}^{\text{C}}$, and (D) full instruction with (d) replaced by textual result table description $\mathcal{T}^{\text{D}}$ as shown in Figure 3.

As shown in Figure 3, we observe that ChatGPT's performance degrades when the instruction does not make reference to interacted items or user preferences across three datasets. This suggests that explicit inclusion of watched movies or user preferences can improve its ability to leverage the user's historical items effectively. While Instruction (A) shows similar average performance as Instruction (D) on ML-1M and LastFM, the former enjoys a smaller variance and outperforms the latter on LastFm. This suggests that LLM prefers explicit output formats over textual description of output format.

> ***Finding* 1.** For sequential recommendation, ChatGPT prefers explicit mentions of instructions and explicit mentions of interacted items, user preference alignment and ranked result format.

3.4 Task Consistency

LLMs are capable of learning new tasks at test time by understanding the relationship between the input of a demonstration and its corresponding output label (Yoo et al., 2022; Wei et al., 2023). In sequential recommendation, LLM is required to rank the ground truth target item at the top followed by other candidate items. However, in a demonstration example from the training set, we observe only one labeled next item but not the ranking of other candidate items. Hence, when constructing demonstrations for in-context learning, we have to answer the important questions: How to prepare the input-label correspondence for a demonstration to be consistent with the sequential recommendation task?

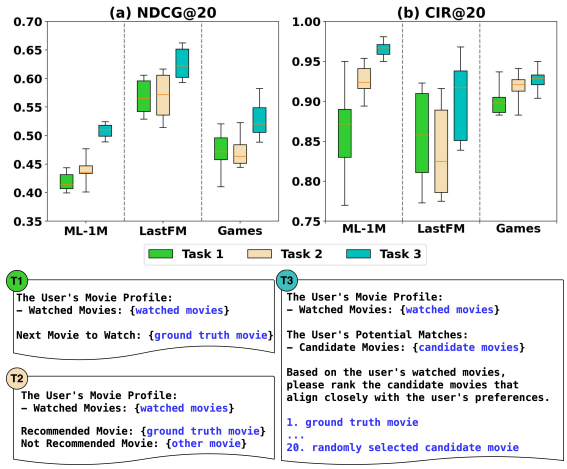

Figure 4: Impact of task consistency between demonstrations and test instances. CIR: Candidate Inclusion Ratio of Demonstration Templates: (T1) Next-Item option; (T2) Contrasting Item Pair option; (T3) Ranked Items option.

To eliminate other factors that may influence the results, such as the number of demonstrations and instructions, we employ instruction (A) as it has proven to be the most effective and robust across three datasets in our previous experiments. We randomly select only one demonstration example for all experiments in this study.

In traditional sequential recommendation, next-item prediction (Song et al., 2021; Petrov and Macdonald, 2022), positive and negative item comparison (Rendle et al., 2012; Kang and McAuley, 2018; Xie et al., 2020), and reranking (Xu et al., 2023) are commonly utilized objectives to train models. Hence, we develop three different prediction tasks for demonstrations for in-context learning. These tasks include: (T1) predicting the next item, (T2) contrasting item pairs, and (T3) ranking candidate items. The prompts corresponding to these prediction tasks are shown in Figure 4. T1 uses the ground truth next-item directly in the demonstration. T2 uses the ground truth next item and another randomly selected item as the positive and negative items respectively. T3 ranks the ground truth next item at the first position and randomly shuffles the remaining candidate items to fill the other positions. Among the task prediction task options, T3 is the only one that aligns closely with the instruction for the test user, i.e., (A).

Figure 4 shows the results of these three tasks across three datasets. T3 consistently outperforms T1 and T2 on all three datasets, suggesting that task consistency between demonstration and test user benefits in-context learning for sequential recommendation. Additionally, As the recommended items may not be found among the provided candidates, we also report *candidate inclusion ratio* (CIR) which measures the proportion of the candidate items that appear in the ranked item results. As shown in Figure 4, we observe that the CIR generally correlates with the NDCG results. The inconsistent demonstration task options (e.g., T1 and T2 coupled with test instruction option (A)) are more likely to cause the LLM to generate non-candidate items in the results. This helps to understand why T3 achieves the best performance.

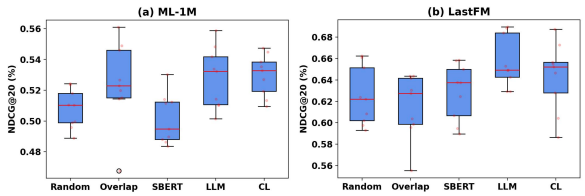

Figure 5: Demonstration selection: (1) random selection; (2) overlapping interacted items; (3) cosine similarity between the SBERT embeddings of interacted item sequences; (4) cosine similarity between the LLM (OpenAI embeddings) of interacted item sequences; (5) cosine similarity using CL embeddings of interacted item sequences.

> ***Finding* 2.** Maintaining task consistency between demonstrations and test users is beneficial for in-context learning in sequential recommendation.

3.5 Selection of Demonstrations

It has been observed that the performance of in-context learning greatly depends on selecting suitable demonstrations (Liu et al., 2021). Utilizing examples that are semantically similar to the test sample can provide more informative and task-relevant knowledge to LLMs. Following Liu et al. (2021), there are several follow-up works (Rubin et al., 2021; Shi et al., 2022; Zhang et al., 2023b; Li et al., 2023) to develop methods for selecting better demonstrations. In this work, we evaluate five different demonstration selection methods to determine their impact to in-context learning for sequential recommendation. These methods include: (1) random selection; (2) overlapping historical items of demonstration user and test user; (3) text similarity scores using Sentence-BERT embedding (Reimers and Gurevych, 2019) (SBERT); (4)

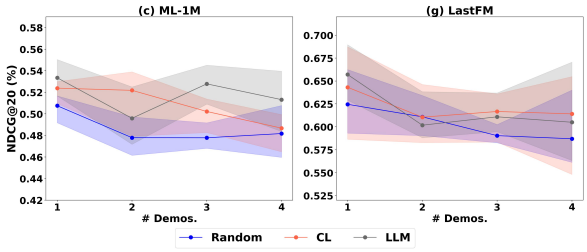

Figure 6: Varying number of demonstrations.

text similarity scores using LLM OpenAI embedding[1] (LLM); and (5) trained retriever using contrastive learning (Xie et al., 2020; Li et al., 2023) (CL). In Option (5), positive examples are obtained by data augmentation applied to the anchor user sequence, while negative examples are randomly selected user item-interaction sequences.

Figure 5 compares the five selection methods on ML-1M and LastFM as they are used in one-shot sequential recommendation. The results show that selection methods (4) and (5) generally outperform the rest. As method (4) appears to be more robust than (5) and it does not require additional training, we thus use that as the default retriever model in the subsequent experiments.

> ***Finding* 3.** Retrieval-based methods are better than random selection, and stronger LLMs can serve as stronger retrievers without any training.

3.6 Number of Demonstrations

When training a model, having more training data examples usually leads to better model performance. However, it is interesting to note that Zhao et al. (2023) discover that increasing the number of demonstrations for in-context learning does not necessarily result in improved performance. Similarly, Chen et al. (2023) finds that using only one demonstration may not perform worse than using more demonstrations. In our case, we evaluate the impact of the number of demonstrations on ML-1M and LastFM using random selection, LLM, and CL demonstration selection methods. We conduct experiments with the number of demonstrations ranging from 1 to 4, as exceeding 4 demonstrations would exceed the input limit of ChatGPT (`GPT-3.5-Turbo`). Figure 6 demonstrates a clear trend of performance decreasing with number of demonstrations.

> ***Finding* 4.** Increasing demonstrations for in-context learning for sequential recommendation would result in performance degradation and exceed the input limit of LLMs.

4 In-Context Learning with Aggregated Demonstrations

Finding 4 suggests LLMs have difficulties coping with multiple demonstrations in sequential recommendation. A similar finding by Liu et al. (2023b) also suggests that the current language models often struggle to utilize information in long input contexts. In particular, their performance tends to significantly degrade when the relevant information is located in the middle of long contexts, also known as the "lost in the middle" phenomenon. The in-context learning prompts for sequential recommendation can easily exceed the prompt length limit of LLM when more than 4 demonstrations are to be accommodated. Such prompts not only suffer from "lost in the Middle", but also incur additional costs of calling LLM APIs.

To address the above challenge, we propose *aggregated demonstration* which combines K ($K > 1$) demonstration users into one for in-context learning. This simple yet effective in-context learning method for sequential recommendation is called **LLMSRec-Syn**. As the prompt length of aggregated demonstration only increases marginally when we increase K, LLMSRec-Syn can accommodate more member demonstration users.

Based on Finding 3, LLMSRec-Syn begins with selecting K demonstration users that are similar to the test user. We use similarity between the LLM embeddings of demonstration and test users. We also follow Finding 1 and adopt instruction template (A) for the test user. Based on Finding 2, we also adopt demonstrate template (T3) for the aggregated demonstration to maintain consistency with the task for test user. Next, we construct the aggregated demonstration's historical item-interactions, candidate items, and the desired ranking of the candidate items from its member demonstrations, as shown in Figure 10 in the Appendix.

Historical item-interactions. Let H denote the historical item-interactions and H is empty initially. We first rank the K selected demonstration users by similarity score. We then add the most recent interacted item from the most similar demonstration

[1] `text-embedding-ada-002` (https://platform.openai.com/docs/models/moderation)

to H. We repeat the same step for the remaining demonstrations in their similarity order. When we run out of most recent interacted items from K selected demonstrations, we continue to add the next recent interacted items of these demonstrations to H until the number of historical items reaches MAX_H.

Candidate Items. Let C denote the candidate items of the aggregated demonstration and C is empty initially. We first gather all the ground truth next items from the K selected demonstrations and add them to C. Next, we randomly add other items from the item pool to C so as to meet the required number of candidate items.

Ranking of Candidate Items. To rank the candidate items in C, we place the ground truth next item of the most similar demonstration at rank 1, followed by that of next similar demonstration until we run out of the ground truth next items of all K selected demonstrations. Next, we assign random ranks to the remaining items in C.

Once the aggregated demonstration is constructed, it is added to the prompt the same way a training user is added as a demonstration. we add it to the corresponding test user and use them as input for the LLM.

$$\mathcal{C}^{\mathbf{agg}} = \mathcal{T}^{\mathrm{A}}(\mathrm{Agg}^{\mathrm{T3}}(x_{\sigma_1}, c_{\sigma_1}, y_{\sigma_1}, \cdots, x_{\sigma_n}, c_{\sigma_n}, y_{\sigma_n})), \quad (3)$$

$$y_{\text{test}} \sim \mathcal{P}_{LLM}\left(\cdot \mid \mathcal{C}^{\mathbf{agg}} \oplus \mathcal{T}^{\mathrm{A}}\left(x_{\text{test}}, c_{\text{test}}\right)\right), \quad (4)$$

where σ_i represents the i^{th} ranked selected users returned by the retrieval model. Finally, the LLM generates a ranked list of candidate items as the recommendation result.

There are several advantages of the proposed LLMSRec-Syn: 1) Standard demonstration only has one ground truth next item in the ranking list. In contrast, the aggregated demonstration includes more next items at high positions in the ranking list. This approach can avoid sparse signals and provide more guidance to LLMs for recommending to the test user; 2) LLMSRec-Syn is less sensitive to the number of demonstrations; 3) Cost of LLMSRec-Syn does not increase much with the number of demonstrations; and 4) LLMSRec-Syn keeps to the prompt length limit of LLMs.

5 Experiments and Results

5.1 Methods for Comparison

To evaluate the performance of LLMSRec-Syn, we conduct an extensive set of experiments on ML-1M, Games, and LastFM-2K datasets. Following Hou et al. (2023), we select 200 data examples from each of the three datasets to carry out all experiments. We use an experiment setup similar to that mentioned in Section 3.1 except that we now uses more LLMs and reports the NDCG@N results where N=5,10, and 20. We compare LLMSRec-Syn with 10 methods categorized into 3 types:

Supervised methods: Most Popular (Recommending items based on their overall popularity among all users in the training data), GRU4Rec (Hidasi et al., 2015) (using GRUs to model user's item sequences), and SASRec (Kang and McAuley, 2018) (employing a self-attention mechanism to learn user preferences from their item sequences).

Zero-shot methods: BM25 (Robertson et al., 2009) (ranking candidate items based on their textual similarity with the test user's interacted items), LLMSeqSim (Harte et al., 2023) (ranking candidate items by semantically similarity using OpenAI embeddings (`text-embedding-ada-002`)), LLMRank-Seq (Hou et al., 2023) (using ChatGPT to rank candidate items with crafted prompts), and LLMSRec (a zero-shot version of the proposed LLMSRec-Syn using the instruction prompt $\mathcal{T}^{\mathrm{A}}$).

One-shot methods: LLMRank-His (Hou et al., 2023) (using historical items of the test user to form a demonstration), LLMSRec-Fixed (using a randomly selected demonstration for all test users), and LLMSRec-Nearest (finding the most similar training user as the demonstration).

As Section 3.6 shows that more than one demonstration in in-context learning for sequential recommendation does not yield better performance, we do not include few-shot methods in this set of experiments. We however will study how many member demonstrations K is ideal for aggregated demonstration (see Section 5.2).

We implement LLMSRec-Syn using three different LLMs, LLaMa2 (Touvron et al., 2023), ChatGPT (OpenAI, 2022) (LLMSRec-Syn), and GPT-4 (OpenAI, 2023) (LLMSRec-Syn-4). For the LLMSRec-Syn-4 experiment, which is shown in the last row of Table 2, we used GPT-4 as the base LLM. For all other experiments, including preliminary studies, in-depth analysis, and method comparisons presented in Table 2, we used the same ChatGPT (`GPT-3.5-Turbo`). To ensure the reliability of our findings, each experiment is conducted 9 times, and the average results are reported. However, we found LLaMa2 unable to follow recommendation instructions and is prone to generating historical

Table 2: Main results. We report NDCG@5, NDCG@10 and NDCG@20 on ML-1M, LastFM-2K and Games. (Best results in each group of methods are **boldfaced** and overall best results are underlined).

Setting	Method	ML-1M			LastFM-2K			Games		
		NDCG@5	NDCG@10	NDCG@20	NDCG@5	NDCG@10	NDCG@20	NDCG@5	NDCG@10	NDCG@20
Supervised	Most Popular	0.3673	0.4623	0.4748	0.4055	0.4205	0.4803	0.2746	0.3905	0.4496
	GRU4Rec	0.7205	0.7494	0.7610	0.3382	0.3971	0.4784	0.6747	0.7002	0.7278
	SASRec	**0.7322**	**0.7595**	**0.7702**	**0.4081**	**0.4680**	**0.5303**	**0.6828**	**0.7189**	**0.7311**
Zero-shot	BM25	0.1314	0.2053	0.3370	0.1215	0.1393	0.3354	0.2285	0.3108	0.4055
	LLMSeqSim	0.3250	0.4037	**0.4723**	0.4090	0.4662	0.5293	**0.4269**	**0.4830**	**0.5360**
	LLMRank-Seq	**0.3344**	0.3882	0.4612	0.5084	0.5545	**0.6070**	0.3063	0.3607	0.4074
	LLMSRec	0.3339	**0.4087**	**0.4723**	**0.5126**	**0.5602**	0.6057	0.4070	0.4555	0.5103
One-shot	LLMRank-His	0.3919	0.4444	0.5074	0.5318	0.5725	0.6212	0.4191	0.4667	0.5206
	LLMSRec-Fixed	0.3590	0.4193	0.4793	0.4961	0.5425	0.5984	0.3744	0.4400	0.4899
	LLMSRec-Nearest	0.3842	0.4382	0.5017	0.5249	0.5697	0.6197	0.3975	0.4388	0.4994
	LLMSRec-Syn	0.4267	0.4813	0.5334	0.5554	0.5918	0.6371	0.4989	0.5334	0.5869
	LLMSRec-Syn-4	**0.5112**	**0.5685**	**0.5936**	**0.6544**	**0.6799**	**0.7017**	**0.5647**	**0.6019**	**0.6277**

interacted items or in-context examples. As a result, we exclude the LLaMa2 results In LLMSRec-Syn, we set the number of member users in the aggregated demonstration as {1,2,3,4,5,6,7} and conduct a brute force search to determine the optimal number for each dataset. We set the number of historical items $MAX_H = 50$ and number of candidate items to 20. We analyse some specific test cases of LLMSRec-Syn-4 in the Appendix 8.2.

5.2 Main Results

The main experiment results are shown in Table 2, from which we obtain the following findings:

ICL one-shot methods with appropriate demonstrations out-perform zero-shot methods. As shown in Table 2, LLMRank-His, LLMSRec-Fixed, and LLMSRec-Nearest using one training user as demonstration outperform LLMRank-Seq on three datasets, except for LLMSRec-Fixed which performs slightly worse than LLMRank-Seq on LastFM-2K. This result suggests that ICL can enhance the LLM's ability to perform a complex task such as sequential recommendation.

Aggregated demonstration, combining multiple member users, allows LLM to effectively gather useful task specific information about the test user within a concise context. Compared to other ICL baselines (i.e., LLMRank-His, LLMSRec-Fixed, and LLMSRec-Nearest), LLMSRec-Syn achieves the superior one-shot performance across all datasets as shown in Table 2. While Figure 6 shows that having more demonstrations may hurt ICL for sequential recommendation, the idea of incorporating multiple demonstration users into an aggregated demonstration enhances the performance of LLMSRec-Syn. These results illustrate the advantage of aggregated demonstration in accommodating multiple training users within a limited prompt length.

LLMSRec-Syn is competitive against supervised methods when the amount of training data is limited. LLMSRec-Syn easily outperforms the simple supervised baseline, Most Popular. While it does not outperform GRU4Rec and SASRec on ML-1M and Games, LLMSRec-Syn surprisingly outperforms all supervised baselines on LastFM-2K. One possible reason is that LastFM-2K has sparse information about items after removing duplicate user-item interactions and users/items with less than 5 interactions, making it challenging to train a good supervised model.

LLMSRec-Syn using more powerful LLMs may outperform supervised methods in the future. With rapid advancement of LLM research, LLMSRec-Syn can be further enhanced when more powerful LLM is used. Our results in Table 2 shows that LLMSRec-Syn-4 significantly outperforms LLMSRec-Syn on all the 3 datasets.

5.3 Analysis of Aggregated Demonstrations

In this section, we study the recommendation performance when varying the settings of aggregated demonstrations. Analysis of ordering of users and label in the aggregated demonstration can be found in the Appendix 8.1.

Impact of number of users in the aggregated demonstration. We evaluate the impact of K (the number of member users) in the aggregated demonstration on LLMSRec-Syn's performance. We empirically vary K from 2 to 7. As shown in Figure 7, an approximate inverted U-shaped relationship exists between K and NDCG@10/20 performance. Initially, as K increases, there is a noticeable performance increase, suggesting that

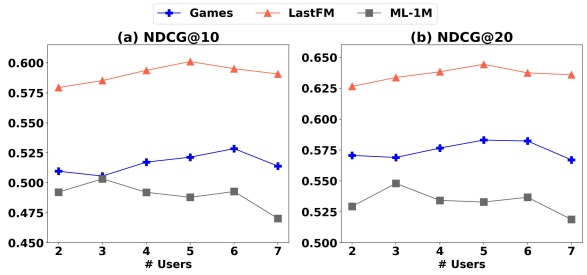

Figure 7: Varying number of users (K) in aggregated demonstration.

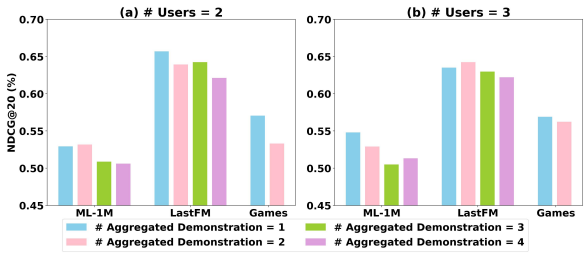

Figure 8: Varying number of aggregated demonstrations each with: (a) 2 member users, and (b) 3 member users.

LLMSRec-Syn benefits from aggregated demonstration. However, beyond some K value, more member users in aggregated demonstration leads to lower performance. This can be explained by more irrelevant training users being incorporated into the aggregated demonstration.

Impact of number of aggregated demonstrations. We evaluate the impact of the number of aggregated demonstrations to LLMSRec-Syn by varying the number of aggregated demonstrations from 1 to 4 such that each demonstration involves 2 users (see Figure 8(a)) and 3 users (see Figure 8(b)). For the Games dataset, experimentation with 3 aggregated demonstrations was not possible due to GPT-3.5-Turbo's input limit. The results show that a single aggregated demonstration outperforms multiple ones, except in the LastFM-2K dataset, where two demonstrations slightly excel.

6 Conclusion

This paper investigates in-context learning (ICL) for LLM-based sequential recommendation. Our study identifies key factors such as instruction format and demonstration selection that influence ICL's effectiveness. We further introduce the LLMSRec-Syn method which utilizes our proposed aggregated demonstration to efficiently incorporate relevant information from multiple training users. Tested on three datasets, LLMSRec-Syn consistently outperforms existing LLM-based sequential recommendation methods. Future work includes a detailed analysis of LLMSRec-Syn's unexpected success compared to some supervised methods and the optimization of aggregated demonstration strategies.

7 Limitations

While this paper considers several factors in applying LLMs to sequential recommendation and proposes a new demonstration concept known as aggregated demonstration, there are still some limitations yet to be addressed. Firstly, the wording of LLMSRec-Syn prompt is manually handcrafted and may not be optimal. This concern is also mentioned in works on prompt optimization (Yang et al., 2023; Deng et al., 2022; Pryzant et al., 2023). However, determining the optimal prompt wording typically requires feedback (such as validation set results (Yang et al., 2023)), carefully designed reward function (Deng et al., 2022), or textual feedback from large language models to iteratively update the initial prompt (Pryzant et al., 2023). Moreover, when a user's historical items are too many, LLMSRec-Syn may still suffer from the issue of long text. Furthermore, the aggregated demonstration method, while mitigating input length constraints, might oversimplify the user preferences, potentially resulting in less personalized recommendations. Moreover, the non-utilization of existing user datasets for pretraining or fine-tuning, due to LLM size constraints, limits the adaptability and fine-tuning of the model to specific recommendation contexts. These limitations highlight the need for further research in optimizing LLMs for complex, dynamic tasks such as sequential recommendation, where user context and historical data play crucial roles.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. In-context examples selection for machine translation. *arXiv preprint arXiv:2212.02437*.

Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. Tallrec: An effective and efficient tuning framework to align large language model with recommendation. *arXiv preprint arXiv:2305.00447*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind

Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Iván Cantador, Peter Brusilovsky, and Tsvi Kuflik. 2011. 2nd workshop on information heterogeneity and fusion in recommender systems (hetrec 2011). In *Proceedings of the 5th ACM conference on Recommender systems*, RecSys 2011, New York, NY, USA. ACM.

Jianxin Chang, Chen Gao, Yu Zheng, Yiqun Hui, Yanan Niu, Yang Song, Depeng Jin, and Yong Li. 2021. Sequential recommendation with graph neural networks. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 378–387.

Jiuhai Chen, Lichang Chen, Chen Zhu, and Tianyi Zhou. 2023. How many demonstrations do you need for in-context learning? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11149–11159.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric P Xing, and Zhiting Hu. 2022. Rlprompt: Optimizing discrete text prompts with reinforcement learning. *arXiv preprint arXiv:2205.12548*.

Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. 2023. Chat-rec: Towards interactive and explainable llms-augmented recommender system. *arXiv preprint arXiv:2303.14524*.

Jesse Harte, Wouter Zorgdrager, Panos Louridas, Asterios Katsifodimos, Dietmar Jannach, and Marios Fragkoulis. 2023. Leveraging large language models for sequential recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 1096–1102.

Ruining He and Julian McAuley. 2016. Fusing similarity models with markov chains for sparse sequential recommendation. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 191–200. IEEE.

Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939*.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2023. Large language models are zero-shot rankers for recommender systems. *arXiv preprint arXiv:2305.08845*.

Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 197–206. IEEE.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. 2023. Unified demonstration retriever for in-context learning. *arXiv preprint arXiv:2305.04320*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Junling Liu, Chao Liu, Renjie Lv, Kang Zhou, and Yan Zhang. 2023a. Is chatgpt a good recommender? a preliminary study. *arXiv preprint arXiv:2304.10149*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023b. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Aman Madaan and Amir Yazdanbakhsh. 2022. Text and patterns: For effective chain of thought, it takes two to tango. *arXiv preprint arXiv:2209.07686*.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

OpenAI. 2022. Introducing chatgpt. https://openai.com/blog/chatgpt.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Aleksandr Petrov and Craig Macdonald. 2022. Effective and efficient training for sequential recommendation using recency sampling. In *Proceedings of the 16th ACM Conference on Recommender Systems*, pages 81–91.

Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with" gradient descent" and beam search. *arXiv preprint arXiv:2305.03495*.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. Bpr: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*.

Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 811–820.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2021. Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633*.

Peng Shi, Rui Zhang, He Bai, and Jimmy Lin. 2022. Xricl: Cross-lingual retrieval-augmented in-context learning for cross-lingual text-to-sql semantic parsing. *arXiv preprint arXiv:2210.13693*.

Wenzhuo Song, Shoujin Wang, Yan Wang, and Shengsheng Wang. 2021. Next-item recommendations in short sessions. In *Proceedings of the 15th ACM Conference on Recommender Systems*, pages 282–291.

Jiaxi Tang and Ke Wang. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 565–573.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lei Wang and Ee-Peng Lim. 2023. Zero-shot next-item recommendation using large pretrained language models. *arXiv preprint arXiv:2304.03153*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023a. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *arXiv preprint arXiv:2305.04091*.

Lei Wang, Songheng Zhang, Yun Wang, Ee-Peng Lim, and Yong Wang. 2023b. Llm4vis: Explainable visualization recommendation using chatgpt. *arXiv preprint arXiv:2310.07652*.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. 2023. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*.

Xu Xie, Fei Sun, Zhaoyang Liu, Shiwen Wu, Jinyang Gao, Bolin Ding, and Bin Cui. 2020. Contrastive learning for sequential recommendation. *arXiv preprint arXiv:2010.14395*.

Yue Xu, Hao Chen, Zefan Wang, Jianwen Yin, Qijie Shen, Dimin Wang, Feiran Huang, Lixiang Lai, Tao Zhuang, Junfeng Ge, et al. 2023. Multi-factor sequential re-ranking with perception-aware diversification. *arXiv preprint arXiv:2305.12420*.

Zhengyi Yang, Jiancan Wu, Yanchen Luo, Jizhi Zhang, Yancheng Yuan, An Zhang, Xiang Wang, and Xiangnan He. 2023. Large language model can interpret latent space of sequential recommender. *arXiv preprint arXiv:2310.20487*.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3081–3089.

Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, and Taeuk Kim. 2022. Ground-truth labels matter: A deeper look into input-label demonstrations. *arXiv preprint arXiv:2205.12685*.

Tianjun Zhang, Xuezhi Wang, Denny Zhou, Dale Schuurmans, and Joseph E Gonzalez. 2023a. Tempera: Test-time prompt editing via reinforcement learning. In *The Eleventh International Conference on Learning Representations*.

Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. 2023b. What makes good examples for visual in-context learning? *arXiv preprint arXiv:2301.13670*.

Fei Zhao, Taotian Pang, Zhen Wu, Zheng Ma, Shujian Huang, and Xinyu Dai. 2023. Dynamic demonstrations controller for in-context learning. *arXiv preprint arXiv:2310.00385*.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*.

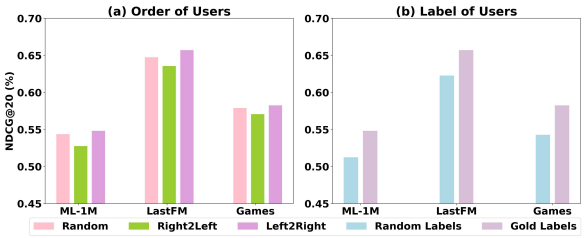

Figure 9: (a) Ordering of member users in the aggregated demonstration. (b) Ground truth vs random next-items in aggregated demonstrations.

Table 3: Further comparison of different tasks with candidates. We have developed two methods with the same information as T3: (1) T1 w/ Candidate (adding candidate items in the T1 prompt) and (2) T2 w/ Candidate (adding candidate items in the T2 prompt).

ML-1M	T1	T1 w/ Cand.	T2	T2 w/ Cand.	T3
NDCG@10	0.3640	0.3766	0.3776	0.3972	0.4584
NDCG@20	0.4193	0.4420	0.4384	0.4510	0.5077

8 Appendix

8.1 More In-Depth Analysis

Impact of user order in aggregated demonstrations. We experiment with 3 possible orders of member users: (i) Random (randomly selects historical items and next-items from the selected users to construct the aggregated demonstration), (ii) Right2Left (the reverse order of demonstration users in constructing an aggregated demonstration in LLMSRec-Syn in Section 4), and (iii) Left2Right (the user order used in the LLMSRec-Syn). Figure 9(a) illustrates that Left2Right and Right2Left are the most and least ideal orders respectively. The performance of Random is naturally sandwiched in between.

Impact of labeled next-items in the aggregated demonstration. According to Min et al. (2022), ground truth labels are not important for in-context learning. To investigate this claim for ICL-based sequential recommendation, we compare LLMSRec-Syn using ground truth next-items in the aggregated demonstration (referred to as "Gold Labels") with that using random non-ground truth next-items (referred to as "Random Labels"). Our results in Figure 9(b) clearly indicate that ground truth next-items are required to yield better performance contradicting the claim by Min et al. (2022). This could possibly be explained by the complexity of sequential recommendation task.

Further comparison of different tasks with candidates. As shown in Table 3, we observed that T1(T2) with candidate items in the prompt performs better than T1(T2). These results support the reviewer's comment that including more information in the prompt will enhance the performance. However, in this more fair comparison, T3 still outperforms T1 with candidate items in the prompt and T2 with candidate items in the prompt.

Table 4: Results of fine-tuned LLaMa2 with LoRA for in-context sequential recommendation. Regular means LLaMa2-LoRA-Regular. Aggregated means LLaMa2-LoRA-Aggregated.

ML-1M	Regular	Aggregated	LLMSRec-Syn
NDCG@10	0.3640	0.3766	0.3776
NDCG@20	0.4193	0.4420	0.4384

Could fine-tuned LLaMa2 improve the performance of in-context sequential recommendation? We initially used a training dataset of 150 data examples to train LLaMa2 with LoRA, which we referred to as LLaMa2-LoRA-Regular. For each training data example in this training dataset, the target output is the ranking of the candidate items for a training user. The input consists of a regular demonstration example, as well as historical items and candidate items from the training user. After training, we evaluated the performance of LLaMa2-LoRA-Regular using the same 50 test users as ChatGPT-based LLMSRec-Syn (0.5283 NDCG@10).

As shown in Table 4, the results showed that LLaMa2-LoRA-Regular achieved a NDCG@10 score of 0.2344. To investigate whether aggregated demonstration helps to train a better model compared to regular demonstrations, we prepared a training dataset using aggregated demonstrations instead of regular demonstrations. We trained LLaMa2 with LoRA using this dataset, which we call LLaMa2-LoRA-Aggregated. LLaMa2-LoRA-Aggregated achieved a NDCG@10 score of 0.3432 on the same test set. Although the initial study indicates that LLaMa-LoRA performs worse than ChatGPT, the fine-tuned LLaMa2-LoRA appears to have the potential to enable in-context learning-based sequential recommendation and aggregated demonstration can help to train a better model.

8.2 Case Study Examples

In this section, we provide comparative examples of one-shot LLMSRec-Syn (Table 5), one-shot LLMSRec-Nearest (Table 6), one-shot LLMSRec-

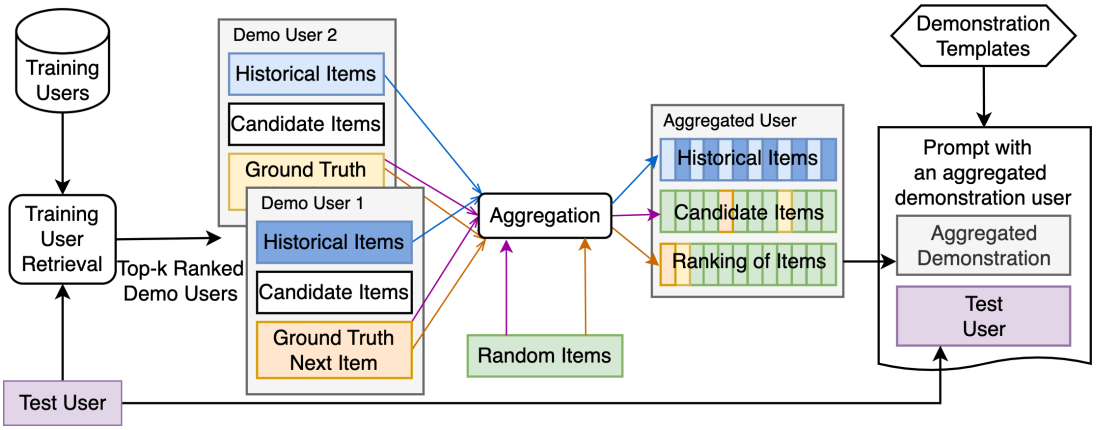

Figure 10: Construction of in-context learning prompt with aggregated demonstration for sequential recommendation.

Fixed (Table 7), and zero-shot LLMSRec (Table 8). Observations show that LLMSRec-Syn ranks the ground truth movie higher than the other methods. Compared to Nearest and Fixed demonstrations, the aggregated demonstration allows the LLM to better identify a user's interests and align the ranking with those interests. Without demonstration, zero-shot LLMSRec relies solely on the LLM's knowledge and performs poorly. This suggests that LLMs can learn from demonstrations to improve in areas where they might not originally be good at.

Table 5: Example of the one-shot LLMSRec-Syn on the ML-1M dataset. Ground truth recommendation is highlighted in Maroon.

Aggregated Demonstration Example:

The User's Movie Profile:
- Watched Movies: ['0. Caddyshack', '1. Glory', '2. A Bug's Life', '3. Star Trek VI: The Undiscovered Country', '4. Indiana Jones and the Last Crusade', '5. The Color of Money', '6. Raging Bull', '7. Edward Scissorhands', '8. Kramer Vs. Kramer', '9. Roger & Me', '10. Romancing the Stone', '11. Full Metal Jacket', '12. The Shining', '13. Easy Rider', '14. Glory', '15. The Color Purple', '16. Die Hard', '17. Who Framed Roger Rabbit?', '18. Ghostbusters', '19. The Right Stuff', '20. No Way Out', '21. The Breakfast Club', '22. Dead Poets Society', '23. One True Thing', '24. Full Metal Jacket', '25. U2: Rattle and Hum', '26. Caddyshack', '27. Jaws', '28. Get Shorty', '29. A Fish Called Wanda', '30. Star Trek IV: The Voyage Home', '31. Brazil', '32. The Fugitive', '33. Willy Wonka and the Chocolate Factory', '34. Ghostbusters', '35. The Blues Brothers', '36. Stripes', '37. Stand by Me', '38. Pleasantville', '39. Bull Durham', '40. Risky Business', '41. Indiana Jones and the Last Crusade', '42. Star Wars: Episode VI - Return of the Jedi', '43. Witness', '44. Brazil', '45. Predator', '46. The Terminator', '47. Grosse Pointe Blank', '48. Contact', '49. The Princess Bride']
The User's Potential Matches:
- Candidate Movies: ['0. Excalibur', '1. The Fox and the Hound', '2. The Cable Guy', '3. Cat on a Hot Tin Roof', '4. Slingshot, The (Kådisbellan)', '5. Institute Benjamenta, or This Dream People Call Human Life', '6. Prefontaine', '7. Star Wars: Episode IV - A New Hope', '8. 8 Seconds', '9. Two Family House', '10. American Graffiti', '11. Seven Days in May', '12. The English Patient', '13. Jurassic Park', '14. Meet the Parents', '15. Weekend at Bernie"s', '16. Kurt & Courtney', '17. Children of a Lesser God', '18. The Haunting', '19. Love and Basketball']
Based on the user's watched movies, please rank the candidate movies that align closely with the user's preferences.
- You ONLY rank the given Candidate Movies.
- You DO NOT generate movies from Watched Movies.
Present your response in the format below:
1. [Top Recommendation (Candidate Movie)]
2. [2nd Recommendation (Candidate Movie)]
...
20. [20th Recommendation (Candidate Movie)]
Answer:
1. Children of a Lesser God
2. Excalibur
3. Jurassic Park
4. American Graffiti
5. Meet the Parents
6. The Haunting
7. Institute Benjamenta, or This Dream People Call Human Life
8. Cat on a Hot Tin Roof
9. Slingshot, The (Kådisbellan)
10. Prefontaine
11. The English Patient
12. Love and Basketball
13. The Fox and the Hound
14. The Cable Guy
15. Two Family House
16. Star Wars: Episode IV - A New Hope
17. Weekend at Bernie"s
18. Seven Days in May
19. 8 Seconds
20. Kurt & Courtney

Test User:

Learn from the above demonstration examples to solve the following test example.
The User's Movie Profile:

- Watched Movies: [‘0. Nikita (La Femme Nikita)’, ‘1. Wallace & Gromit: The Best of Aardman Animation’, ‘2. Unforgiven’, ‘3. Batman: Mask of the Phantasm’, ‘4. The Fugitive’, ‘5. Out of Sight’, ‘6. Twelve Monkeys’, ‘7. GoodFellas’, ‘8. Fight Club’, ‘9. Groundhog Day’, ‘10. A Bug"s Life’, ‘11. Tombstone’, ‘12. Aladdin’, ‘13. Beauty and the Beast’, ‘14. Grosse Pointe Blank’, ‘15. Election’, ‘16. Leaving Las Vegas’, ‘17. Total Recall’, ‘18. A Few Good Men’, ‘19. Pleasantville’, ‘20. Jerry Maguire’, ‘21. Pretty Woman’, ‘22. Contact’, ‘23. True Lies’, ‘24. Waking Ned Devine’, ‘25. Romeo Must Die’, ‘26. Mission: Impossible 2’, ‘27. Mission to Mars’, ‘28. Killer, The (Die xue shuang xiong)’, ‘29. Blade Runner’, ‘30. The Princess Bride’, ‘31. Brazil’, ‘32. Henry V’, ‘33. Amadeus’, ‘34. The Right Stuff’, ‘35. The Terminator’, ‘36. Stand by Me’, ‘37. Back to the Future’, ‘38. This Is Spinal Tap’, ‘39. Gandhi’, ‘40. Star Trek: The Wrath of Khan’, ‘41. Ghostbusters’, ‘42. Mad Max 2 (a.k.a. The Road Warrior)’, ‘43. A Fish Called Wanda’, ‘44. Trading Places’, ‘45. Chariots of Fire’, ‘46. Time Bandits’, ‘47. Who Framed Roger Rabbit?’, ‘48. Indiana Jones and the Last Crusade’, ‘49. Die Hard’]

The User’s Potential Matches:

- Candidate Movies: [‘0. Sorority House Massacre’, ‘1. Arguing the World’, ‘2. Cop Land’, ‘3. The Inkwell’, ‘4. Lovers of the Arctic Circle, The (Los Amantes del Círculo Polar)’, ‘5. Airplane!’, ‘6. Pokémon the Movie 2000’, ‘7. Johnny 100 Pesos’, ‘8. Happy, Texas’, ‘9. Arlington Road’, ‘10. The Prophecy II’, ‘11. Prom Night’, ‘12. The Commitments’, ‘13. Federal Hill’, ‘14. Another Stakeout’, ‘15. The Postman Always Rings Twice’, ‘16. The Crow’, ‘17. Shanghai Surprise’, ‘18. Patriot Games’, ‘19. Modulations’]

Based on the user’s watched movies, please rank the candidate movies that align closely with John’s preferences.

- You ONLY rank the given Candidate Movies.
- You DO NOT generate movies from Watched Movies.

Present your response in the format below: 1. [Top Recommendation (Candidate Movie)] 2. [2nd Recommendation (Candidate Movie)] ... 20. [20th Recommendation (Candidate Movie)]

Answer:

1. The Commitments
2. Cop Land
3. **Airplane!**
4. The Crow
5. Patriot Games
6. The Postman Always Rings Twice
7. Shanghai Surprise
8. Arlington Road
9. Happy, Texas
10. Another Stakeout
11. The Inkwell
12. Modulations
13. Johnny 100 Pesos
14. Sorority House Massacre
15. Pokémon the Movie 2000
16. Lovers of the Arctic Circle, The (Los Amantes del Círculo Polar)
17. The Prophecy II
18. Federal Hill
19. Prom Night
20. Arguing the World

Table 6: Example of the one-sho LLMSRec-Nearest on the ML-1M dataset.

Nearest Demonstration Example:

The User's Movie Profile:
- Watched Movies: ['E.T. the Extra-Terrestrial', 'Gladiator', 'Raiders of the Lost Ark', 'Brazil', 'Aliens', 'Full Metal Jacket', 'The Right Stuff', 'The Terminator', 'Down by Law', 'Blade Runner', 'The Princess Bride', 'Mystery Train', 'Stand by Me', 'Dangerous Liaisons', 'Year of Living Dangerously', 'Poltergeist', 'Crimes and Misdemeanors', 'Never Cry Wolf', 'Mad Max 2 (a.k.a. The Road Warrior)', 'Women on the Verge of a Nervous Breakdown', "Ferris Bueller's Day Off", 'Who Framed Roger Rabbit?', 'Koyaanisqatsi', 'Ghostbusters', 'A Fish Called Wanda', 'The Gods Must Be Crazy', 'Chariots of Fire', 'The Hidden', 'Diva', 'Gandhi', 'F/X', 'Sex, Lies, and Videotape', 'Reds', 'Back to the Future', 'Driving Miss Daisy', 'Body Heat', 'Indiana Jones and the Last Crusade', 'When Harry Met Sally...', 'Batman', 'The Untouchables', 'Predator', 'Risky Business', 'The Blues Brothers', 'Star Trek IV: The Voyage Home', 'U2: Rattle and Hum', 'No Way Out', 'The Color Purple', 'Romancing the Stone', 'The Color of Money', 'Caddyshack']
The User's Potential Matches:
- Candidate Movies: ['Children of a Lesser God', 'Kicking and Screaming', 'Harlem', 'The Good Mother', 'Oscar and Lucinda (a.k.a. Oscar & Lucinda)', 'Meet Me in St. Louis', 'Sinbad and the Eye of the Tiger', 'Gremlins', "Freddy's Dead: The Final Nightmare", 'One Magic Christmas', 'The Scarlet Letter', 'Alaska', 'Red Firecracker, Green Firecracker', 'Bulletproof', 'The Crimson Pirate', 'Clockers', 'The Stepford Wives', 'Twelve Monkeys', 'Bitter Moon', 'Friday the 13th: The Final Chapter']
Based on the user's watched movies, please rank the candidate movies that align closely with the user's preferences.
- You ONLY rank the given Candidate Movies.
- You DO NOT generate movies from Watched Movies.
Present your response in the format below:
1. [Top Recommendation (Candidate Movie)]
2. [2nd Recommendation (Candidate Movie)]
...
20. [20th Recommendation (Candidate Movie)]
Answer:
1. Children of a Lesser God
2. One Magic Christmas
3. Alaska
4. Gremlins
5. The Good Mother
6. Bitter Moon
7. Bulletproof
8. Freddy's Dead: The Final Nightmare
9. Red Firecracker, Green Firecracker
10. The Scarlet Letter
11. Sinbad and the Eye of the Tiger
12. Kicking and Screaming
13. Meet Me in St. Louis
14. Oscar and Lucinda (a.k.a. Oscar & Lucinda)
15. Twelve Monkeys
16. Clockers
17. Harlem
18. The Stepford Wives
19. The Crimson Pirate
20. Friday the 13th: The Final Chapter

Test User:

Learn from the above demonstration examples to solve the following test example.
The User's Movie Profile:

- Watched Movies: [‘0. Nikita (La Femme Nikita)’, ‘1. Wallace & Gromit: The Best of Aardman Animation’, ‘2. Unforgiven’, ‘3. Batman: Mask of the Phantasm’, ‘4. The Fugitive’, ‘5. Out of Sight’, ‘6. Twelve Monkeys’, ‘7. GoodFellas’, ‘8. Fight Club’, ‘9. Groundhog Day’, ‘10. A Bug"s Life’, ‘11. Tombstone’, ‘12. Aladdin’, ‘13. Beauty and the Beast’, ‘14. Grosse Pointe Blank’, ‘15. Election’, ‘16. Leaving Las Vegas’, ‘17. Total Recall’, ‘18. A Few Good Men’, ‘19. Pleasantville’, ‘20. Jerry Maguire’, ‘21. Pretty Woman’, ‘22. Contact’, ‘23. True Lies’, ‘24. Waking Ned Devine’, ‘25. Romeo Must Die’, ‘26. Mission: Impossible 2’, ‘27. Mission to Mars’, ‘28. Killer, The (Die xue shuang xiong)’, ‘29. Blade Runner’, ‘30. The Princess Bride’, ‘31. Brazil’, ‘32. Henry V’, ‘33. Amadeus’, ‘34. The Right Stuff’, ‘35. The Terminator’, ‘36. Stand by Me’, ‘37. Back to the Future’, ‘38. This Is Spinal Tap’, ‘39. Gandhi’, ‘40. Star Trek: The Wrath of Khan’, ‘41. Ghostbusters’, ‘42. Mad Max 2 (a.k.a. The Road Warrior)’, ‘43. A Fish Called Wanda’, ‘44. Trading Places’, ‘45. Chariots of Fire’, ‘46. Time Bandits’, ‘47. Who Framed Roger Rabbit?’, ‘48. Indiana Jones and the Last Crusade’, ‘49. Die Hard’]

The User’s Potential Matches:

- Candidate Movies: [‘0. Sorority House Massacre’, ‘1. Arguing the World’, ‘2. Cop Land’, ‘3. The Inkwell’, ‘4. Lovers of the Arctic Circle, The (Los Amantes del Círculo Polar)’, ‘5. Airplane!’, ‘6. Pokémon the Movie 2000’, ‘7. Johnny 100 Pesos’, ‘8. Happy, Texas’, ‘9. Arlington Road’, ‘10. The Prophecy II’, ‘11. Prom Night’, ‘12. The Commitments’, ‘13. Federal Hill’, ‘14. Another Stakeout’, ‘15. The Postman Always Rings Twice’, ‘16. The Crow’, ‘17. Shanghai Surprise’, ‘18. Patriot Games’, ‘19. Modulations’]

Based on the user’s watched movies, please rank the candidate movies that align closely with John’s preferences.

- You ONLY rank the given Candidate Movies.
- You DO NOT generate movies from Watched Movies.

Present your response in the format below: 1. [Top Recommendation (Candidate Movie)] 2. [2nd Recommendation (Candidate Movie)] ... 20. [20th Recommendation (Candidate Movie)]

Answer:

1. Arlington Road
2. Cop Land
3. The Crow
4. Patriot Games
5. The Postman Always Rings Twice
6. The Commitments
7. Airplane!
8. Another Stakeout
9. Lovers of the Arctic Circle, The (Los Amantes del Círculo Polar)
10. Shanghai Surprise
11. Happy, Texas
12. Modulations
13. The Inkwell
14. Johnny 100 Pesos
15. Sorority House Massacre
16. Arguing the World
17. Prom Night
18. Federal Hill
19. Pokémon the Movie 2000
20. The Prophecy II

Table 7: Example of the one-sho LLMSRec-Fixed on the ML-1M dataset.

Fixed Demonstration Example:

The User's Movie Profile:
- Watched Movies: ['Total Recall', 'Aliens', 'Star Wars: Episode VI - Return of the Jedi', 'E.T. the Extra-Terrestrial', 'Forbidden Planet', 'Brazil', 'Star Trek: First Contact', 'Star Trek: The Wrath of Khan', 'Sneakers', 'Galaxy Quest', 'Contact', 'Village of the Damned', 'Being John Malkovich', 'Waiting for Guffman', 'Clerks', 'American Beauty', 'Toy Story 2', 'Shakespeare in Love', 'Toy Story', 'Flirting With Disaster', 'Smoke Signals', 'Pulp Fiction', 'Erin Brockovich', 'Chicken Run', 'Shanghai Noon', 'Gladiator', 'The Wizard of Oz', 'The Producers', "Singin' in the Rain", 'The Sound of Music', 'Snow White and the Seven Dwarfs', 'Fantasia', 'Sleeping Beauty', 'West Side Story', 'Cinderella', 'The Little Mermaid', 'Holiday Inn', 'James and the Giant Peach', 'Dumbo', 'The Lion King', 'The Nightmare Before Christmas', 'The Rocky Horror Picture Show', 'Oliver!', 'Pocahontas', 'Everyone Says I Love You', 'South Pacific', 'Victor/Victoria', 'Tank Girl', 'His Girl Friday', 'The Maltese Falcon']
The User's Potential Matches:
- Candidate Movies: ['The Shop Around the Corner','Kicking and Screaming','Harlem','The Good Mother','Oscar and Lucinda (a.k.a. Oscar & Lucinda)','Roman Holiday','Sinbad and the Eye of the Tiger','Gremlins', "Freddy's Dead: The Final Nightmare",'One Magic Christmas','The Scarlet Letter','Alaska','Red Firecracker, Green Firecracker','Bulletproof','The Crimson Pirate','Clockers','The Stepford Wives','The Crow','Bitter Moon','Friday the 13th: The Final Chapter']
Based on the user's watched movies, please rank the candidate movies that align closely with the user's preferences.
- You ONLY rank the given Candidate Movies.
- You DO NOT generate movies from Watched Movies.
Present your response in the format below:
1. [Top Recommendation (Candidate Movie)]
2. [2nd Recommendation (Candidate Movie)]
...
20. [20th Recommendation (Candidate Movie)]
Answer:
1. The Shop Around the Corner
2. One Magic Christmas
3. Alaska
4. Gremlins
5. The Good Mother
6. Bitter Moon
7. Bulletproof
8. Freddy's Dead: The Final Nightmare
9. Red Firecracker, Green Firecracker
10. The Scarlet Letter
11. Sinbad and the Eye of the Tiger
12. Kicking and Screaming
13. Roman Holiday
14. Oscar and Lucinda (a.k.a. Oscar & Lucinda)
15. The Crow
16. Clockers
17. Harlem
18. The Stepford Wives
19. The Crimson Pirate
20. Friday the 13th: The Final Chapter

Test User:

Learn from the above demonstration examples to solve the following test example.
The User's Movie Profile:

- Watched Movies: [‘0. Nikita (La Femme Nikita)’, ‘1. Wallace & Gromit: The Best of Aardman Animation’, ‘2. Unforgiven’, ‘3. Batman: Mask of the Phantasm’, ‘4. The Fugitive’, ‘5. Out of Sight’, ‘6. Twelve Monkeys’, ‘7. GoodFellas’, ‘8. Fight Club’, ‘9. Groundhog Day’, ‘10. A Bug"s Life’, ‘11. Tombstone’, ‘12. Aladdin’, ‘13. Beauty and the Beast’, ‘14. Grosse Pointe Blank’, ‘15. Election’, ‘16. Leaving Las Vegas’, ‘17. Total Recall’, ‘18. A Few Good Men’, ‘19. Pleasantville’, ‘20. Jerry Maguire’, ‘21. Pretty Woman’, ‘22. Contact’, ‘23. True Lies’, ‘24. Waking Ned Devine’, ‘25. Romeo Must Die’, ‘26. Mission: Impossible 2’, ‘27. Mission to Mars’, ‘28. Killer, The (Die xue shuang xiong)’, ‘29. Blade Runner’, ‘30. The Princess Bride’, ‘31. Brazil’, ‘32. Henry V’, ‘33. Amadeus’, ‘34. The Right Stuff’, ‘35. The Terminator’, ‘36. Stand by Me’, ‘37. Back to the Future’, ‘38. This Is Spinal Tap’, ‘39. Gandhi’, ‘40. Star Trek: The Wrath of Khan’, ‘41. Ghostbusters’, ‘42. Mad Max 2 (a.k.a. The Road Warrior)’, ‘43. A Fish Called Wanda’, ‘44. Trading Places’, ‘45. Chariots of Fire’, ‘46. Time Bandits’, ‘47. Who Framed Roger Rabbit?’, ‘48. Indiana Jones and the Last Crusade’, ‘49. Die Hard’]

The User’s Potential Matches:

- Candidate Movies: [‘0. Sorority House Massacre’, ‘1. Arguing the World’, ‘2. Cop Land’, ‘3. The Inkwell’, ‘4. Lovers of the Arctic Circle, The (Los Amantes del Círculo Polar)’, ‘5. Airplane!’, ‘6. Pokémon the Movie 2000’, ‘7. Johnny 100 Pesos’, ‘8. Happy, Texas’, ‘9. Arlington Road’, ‘10. The Prophecy II’, ‘11. Prom Night’, ‘12. The Commitments’, ‘13. Federal Hill’, ‘14. Another Stakeout’, ‘15. The Postman Always Rings Twice’, ‘16. The Crow’, ‘17. Shanghai Surprise’, ‘18. Patriot Games’, ‘19. Modulations’]

Based on the user’s watched movies, please rank the candidate movies that align closely with John’s preferences.

- You ONLY rank the given Candidate Movies.
- You DO NOT generate movies from Watched Movies.

Present your response in the format below: 1. [Top Recommendation (Candidate Movie)] 2. [2nd Recommendation (Candidate Movie)] ... 20. [20th Recommendation (Candidate Movie)]

Answer:

1. The Crow
2. Shanghai Surprise
3. The Commitments
4. Arlington Road
5. The Postman Always Rings Twice
6. Lovers of the Arctic Circle, The (Los Amantes del Círculo Polar)
7. Cop Land
8. Happy, Texas
9. Another Stakeout
10. Airplane!
11. Federal Hill
12. Johnny 100 Pesos
13. The Prophecy II
14. Modulations
15. Patriot Games
16. Pokémon the Movie 2000
17. Sorority House Massacre
18. Prom Night
19. Arguing the World
20. The Inkwell

Table 8: Example of the zero-shot LLMSRec on the ML-1M dataset.

Test User:

Learn from the above demonstration examples to solve the following test example.
The User's Movie Profile:
- Watched Movies: ['0. Nikita (La Femme Nikita)', '1. Wallace & Gromit: The Best of Aardman Animation', '2. Unforgiven', '3. Batman: Mask of the Phantasm', '4. The Fugitive', '5. Out of Sight', '6. Twelve Monkeys', '7. GoodFellas', '8. Fight Club', '9. Groundhog Day', '10. A Bug"s Life', '11. Tombstone', '12. Aladdin', '13. Beauty and the Beast', '14. Grosse Pointe Blank', '15. Election', '16. Leaving Las Vegas', '17. Total Recall', '18. A Few Good Men', '19. Pleasantville', '20. Jerry Maguire', '21. Pretty Woman', '22. Contact', '23. True Lies', '24. Waking Ned Devine', '25. Romeo Must Die', '26. Mission: Impossible 2', '27. Mission to Mars', '28. Killer, The (Die xue shuang xiong)', '29. Blade Runner', '30. The Princess Bride', '31. Brazil', '32. Henry V', '33. Amadeus', '34. The Right Stuff', '35. The Terminator', '36. Stand by Me', '37. Back to the Future', '38. This Is Spinal Tap', '39. Gandhi', '40. Star Trek: The Wrath of Khan', '41. Ghostbusters', '42. Mad Max 2 (a.k.a. The Road Warrior)', '43. A Fish Called Wanda', '44. Trading Places', '45. Chariots of Fire', '46. Time Bandits', '47. Who Framed Roger Rabbit?', '48. Indiana Jones and the Last Crusade', '49. Die Hard']
The User's Potential Matches:
- Candidate Movies: ['0. Sorority House Massacre', '1. Arguing the World', '2. Cop Land', '3. The Inkwell', '4. Lovers of the Arctic Circle, The (Los Amantes del Círculo Polar)', '5. Airplane!', '6. Pokémon the Movie 2000', '7. Johnny 100 Pesos', '8. Happy, Texas', '9. Arlington Road', '10. The Prophecy II', '11. Prom Night', '12. The Commitments', '13. Federal Hill', '14. Another Stakeout', '15. The Postman Always Rings Twice', '16. The Crow', '17. Shanghai Surprise', '18. Patriot Games', '19. Modulations']
Based on the user's watched movies, please rank the candidate movies that align closely with John's preferences.
- You ONLY rank the given Candidate Movies.
- You DO NOT generate movies from Watched Movies.
Present your response in the format below: 1. [Top Recommendation (Candidate Movie)] 2. [2nd Recommendation (Candidate Movie)] ... 20. [20th Recommendation (Candidate Movie)]
Answer:
1. The Commitments
2. Patriot Games
3. Arlington Road
4. Cop Land
5. The Crow
6. The Postman Always Rings Twice
7. Shanghai Surprise
8. Another Stakeout
9. Modulations
10. Pokémon the Movie 2000
11. The Inkwell
12. Happy, Texas
13. The Prophecy II
14. Johnny 100 Pesos
15. Lovers of the Arctic Circle, The (Los Amantes del Círculo Polar)
16. Arguing the World
17. Federal Hill
18. Prom Night
19. Sorority House Massacre
20. Airplane!