Large Language Models as Zero-Shot Conversational Recommenders

Zhankui He*
zhh004@eng.ucsd.edu
University of California, San Diego
La Jolla, California, USA

Zhouhang Xie*
zhx022@ucsd.edu
University of California, San Diego
La Jolla, California, USA

Rahul Jha
rahuljha@netflix.com
Netflix Inc.
Los Gatos, California, USA

Harald Steck
hsteck@netflix.com
Netflix Inc.
Los Gatos, California, USA

Dawen Liang
dliang@netflix.com
Netflix Inc.
Los Gatos, California, USA

Yesu Feng
yfeng@netflix.com
Netflix Inc.
Los Gatos, California, USA

Bodhisattwa Prasad Majumder
bmajumde@eng.ucsd.edu
University of California, San Diego
La Jolla, California, USA

Nathan Kallus
nkallus@netflix.com
Netflix Inc.
Los Gatos, California, USA
Cornell University
New York, New York, USA

Julian McAuley
jmcauley@ucsd.edu
University of California, San Diego
La Jolla, California, USA

ABSTRACT

In this paper, we present empirical studies on conversational recommendation tasks using representative large language models in a zero-shot setting with three primary contributions. **(1) Data:** To gain insights into model behavior in "in-the-wild" conversational recommendation scenarios, we construct a new dataset of recommendation-related conversations by scraping a popular discussion website. This is the largest public real-world conversational recommendation dataset to date. **(2) Evaluation:** On the new dataset and two existing conversational recommendation datasets, we observe that even without fine-tuning, large language models can outperform existing fine-tuned conversational recommendation models. **(3) Analysis:** We propose various probing tasks to investigate the mechanisms behind the remarkable performance of large language models in conversational recommendation. We analyze both the large language models' behaviors and the characteristics of the datasets, providing a holistic understanding of the models' effectiveness, limitations and suggesting directions for the design of future conversational recommenders.

CCS CONCEPTS

• **Information systems** → **Personalization**; • **Computing methodologies** → **Natural language generation**.

KEYWORDS

conversational recommendation, large language model, datasets

*Both authors contributed equally to this research.

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom

ACM ISBN 979-8-4007-0124-5/23/10.
https://doi.org/10.1145/3583780.3614949

ACM Reference Format:
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), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3583780.3614949

1 INTRODUCTION

Conversational recommender systems (CRS) aim to elicit user preferences and offer personalized recommendations by engaging in interactive conversations. In contrast to traditional recommenders that primarily rely on users' actions like clicks or purchases, CRS possesses the potential to: (1) understand not only users' historical actions but also users' (multi-turn) natural-language inputs; (2) provide not only recommended items but also human-like responses for multiple purposes such as preference refinement, knowledgable discussion or recommendation justification. Towards this objective, a typical conversational recommender contains two components [10, 41, 64, 74]: a *generator* to generate natural-language responses and a *recommender* to rank items to meet users' needs.

Recently, significant advancements have shown the remarkable potential of large language models (LLMs)[1], such as ChatGPT [30], in various tasks [4, 6, 51, 71]. This has captured the attention of the recommender systems community to explore the possibility of leveraging LLMs in recommendation or more general personalization tasks [3, 27, 34, 48, 56]. Yet, current efforts generally concentrate on evaluating LLMs in traditional recommendation settings, where only users' past actions like clicks serve as inputs [3, 27, 34, 48]. The conversational recommendation scenario, though involving more natural language interactions, is still in its infancy [16, 63].

[1]We refer to LLMs as the large-sized pre-trained language models with exceptional zero-shot abilities as defined in [71].

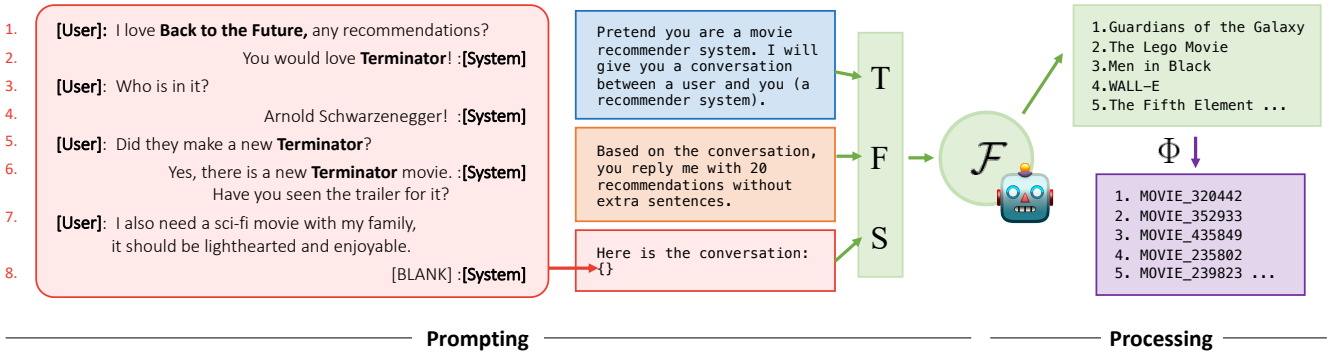


Figure 1: Large Language Models (LLMs) as Zero-Shot Conversational Recommenders (CRS).We introduce a simple prompting strategy to define the task description T, format requirement F and conversation context S for a LLM, denoted as $\mathcal{F}$, we then post-process the generative results into ranked item lists with processor Φ.

In this work, we propose to use *large language models as zero-shot conversational recommenders* and then empirically study the LLMs' [11, 30, 51, 68] recommendation abilities. Our detailed contributions in this study include three key aspects regarding *data*, *evaluation*, and *analysis*.

Data. We construct *Reddit-Movie*, a large-scale conversational recommendation dataset with over 634k naturally occurring recommendation seeking dialogs from users from Reddit[2], a popular discussion forum. Different from existing crowd-sourced conversational recommendation datasets, such as ReDIAL [41] and INSPIRED [22], where workers role-play users and recommenders, the *Reddit-Movie* dataset offers a complementary perspective with conversations where users seek and offer item recommendation in the real world. To the best of our knowledge, this is the largest public conversational recommendation dataset, with 50 times more conversations than ReDIAL.

Evaluation. By evaluating the recommendation performance of LLMs on multiple CRS datasets, we first notice a *repeated item shortcut* in current CRS evaluation protocols. Specifically, there exist "repeated items" in previous evaluation testing samples serving as ground-truth items, which allows the creation of a trivial baseline (e.g., copying the mentioned items from the current conversation history) that outperforms most existing models, leading to spurious conclusions regarding current CRS recommendation abilities. After removing the "repeated items" in training and testing data, we re-evaluate multiple representative conversational recommendation models [10, 41, 64, 74] on ReDIAL, INSPIRED and our Reddit dataset. With this experimental setup, we empirically show that LLMs can outperform existing fine-tuned conversational recommendation models even without fine-tuning.

Analysis. In light of the impressive performance of LLMs as zero-shot CRS, a fundamental question arises: *What accounts for their remarkable performance?* Similar to the approach taken in [53], we posit that LLMs leverage both *content/context knowledge* (e.g., "genre", "actors" and "mood") and *collaborative knowledge* (e.g., "users who like A typically also like B") to make conversational recommendations. We design several probing tasks to uncover the model's workings and the characteristics of the CRS data. Additionally, we present empirical findings that highlight certain limitations of LLMs as zero-shot CRS, despite their effectiveness.

We summarize the key findings of this paper as follows:

- CRS recommendation abilities should be reassessed by eliminating repeated items as ground truth.
- LLMs, as zero-shot conversational recommenders, demonstrate improved performance on established and new datasets over fine-tuned CRS models.
- LLMs primarily use their superior content/context knowledge, rather than their collaborative knowledge, to make recommendations.
- CRS datasets inherently contain a high level of content/context information, making CRS tasks better-suited for LLMs than traditional recommendation tasks.
- LLMs suffer from limitations such as popularity bias and sensitivity to geographical regions.

These findings reveal the unique importance of the superior content/context knowledge in LLMs for CRS tasks, offering great potential to LLMs as an effective approach in CRS; meanwhile, analyses must recognize the challenges in evaluation, datasets, and potential problems (e.g., debiasing) in future CRS design with LLMs.

2 LLMS AS ZERO-SHOT CRS

2.1 Task Formation

Given a user set $\mathcal{U}$, an item set $\mathcal{I}$ and a vocabulary $\mathcal{V}$, a conversation can be denoted as $C = (u_t, s_t, \mathcal{I}_t)_{t=1}^{T}$. That means during the t^{th} turn of the conversation, a speaker $u_t \in \mathcal{U}$ generates an utterance $s_t = (w_i)_{i=1}^{m}$, which is a sequence of words $w_i \in \mathcal{V}$. This utterance s_t also contains a set of mentioned items $\mathcal{I}_t \subset \mathcal{I}$ ($\mathcal{I}_t$ can be an empty set if no items mentioned). Typically, there are two users in the conversation C playing the role of *seeker* and *recommender* respectively. Let us use the 2^{nd} conversation turn in Figure 1 as an example. Here $t = 2$, u_t is `[System]`, s_t is "You would love Terminator !" and $\mathcal{I}_2$ is a set containing the movie `Terminator`.

[2]https://www.reddit.com/

Table 1: Dataset Statistics. We denote a subset of *Reddit-Movie* in 2022 as `base`, and the entire ten-year dataset as `large`.

Dataset	#Conv.	#Turns	#Users	#Items
INSPIRED [22]	999	35,686	999	1,967
ReDIAL [41]	11,348	139,557	764	6,281
Reddit-Movie$^{\text{base}}$	85,052	133,005	10,946	24,326
Reddit-Movie$^{\text{large}}$	634,392	1,669,720	36,247	51,203

Following many CRS papers [10, 41, 64, 74], the *recommender* component of a CRS is specifically designed to optimize the following objective: during the k^{th} turn of a conversation, where u_k is the *recommender*, the recommender takes the conversational context $(u_t, s_t, \mathcal{I}_t)_{t=1}^{k-1}$ as its input, and generate a ranked list of items $\hat{\mathcal{I}}_k$ that best matches the ground-truth items in $\mathcal{I}_k$.

2.2 Framework

Prompting. Our goal is to utilize LLMs as zero-shot conversational recommenders. Specifically, without the need for fine-tuning, we intend to prompt an LLM, denoted as $\mathcal{F}$, using a task description template T, format requirement F, and conversational context S before the k^{th} turn. This process can be formally represented as:

$$\hat{\mathcal{I}}_k = \Phi\left(\mathcal{F}(T, F, S)\right). \tag{1}$$

To better understand this zero-shot recommender, we present an example in Figure 1 with the prompt setup in our experiments.[3]

Models. We consider several popular LLMs $\mathcal{F}$ that exhibit zero-shot prompting abilities in two groups. To try to ensure deterministic results, we set the decoding temperature to 0 for all models.

- **`GPT-3.5-turbo`** [30][4] and **`GPT-4`** [51] from OPENAI with abilities of solving many complex tasks in zero-shot setting [6, 51] but are closed-sourced.
- **BAIZE** [68][5] and **Vicuna** [11], which are representative open-sourced LLMs fine-tuned based on `LLAMA-13B` [61].

Processing. We do not assess model weights or output logits from LLMs. Therefore, we apply a post-processor Φ (e.g., fuzzy matching) to convert a recommendation list in natural language to a ranked list $\hat{\mathcal{I}}_k$. The approach of generating item titles instead of ranking item IDs is referred to as a *generative retrieval* [7, 60] paradigm.

3 DATASET

Ideally, a large-scale dataset with diverse interactions and real-world conversations is needed to evaluate models' ability in conversational recommendation. Existing conversational recommendation datasets are usually crowd-sourced [22, 32, 41, 75] and thus only partially capture realistic conversation dynamics. For example, a crowd worker responded with `"Whatever Whatever I'm open to any suggestion."` when asked about movie preferences in ReDIAL; this happens since crowd workers often do not have a particular preference at the time of completing a task. In contrast, a real user could have a very particular need, as shown in Figure 2.

[3]We leave more prompting techniques such as CoT [66] in future work.
[4]Referred as **`GPT-3.5-t`** hereafter
[5]We use **BAIZE-V2** in https://huggingface.co/project-baize/baize-v2-13b

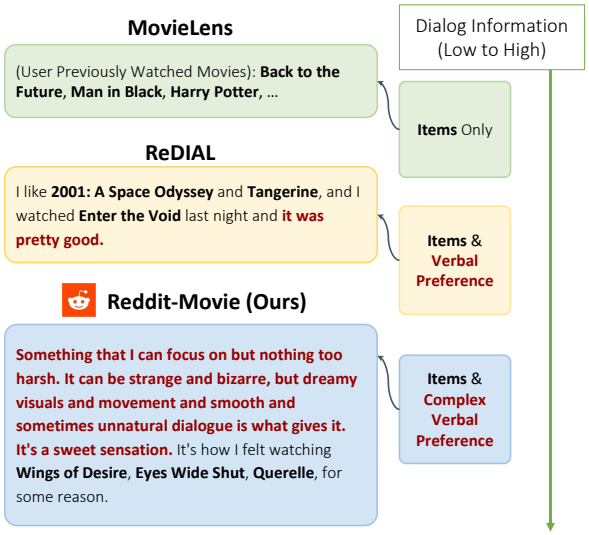


Figure 2: Typical model inputs from a traditional recommendation dataset (MovieLens [21]), an existing CRS dataset (ReDIAL [41]), and our *Reddit-Movie* dataset. The *Reddit-Movie* dataset contains more information in its textual content compared to existing datasets where users often explicitly specify their preference. See Section 5.2 for quantitative analysis.

To complement crowd-sourced CRS datasets, we present the *Reddit-Movie* dataset, the largest-scale conversational movie recommendation dataset to date, with naturally occurring movie recommendation conversations that can be used along with existing crowd-sourced datasets to provide richer perspectives for training and evaluating CRS models. In this work, we conduct our model evaluation and analysis on two commonly used crowd-sourcing datasets: ReDIAL [41] and INSPIRED [22], as well as our newly collected Reddit dataset. We show qualitative examples from the Reddit dataset as in Figure 2 and quantitative analysis in Section 5.2.

Dataset Construction To construct a CRS dataset from Reddit, we process all Reddit posts from 2012 Jan to 2022 Dec from *pushshift.io*[6]. We consider movie recommendation scenarios[7] and extract related posts from five related subreddits: *r/movies*, *r/bestofnetflix*, *r/moviesuggestions*, *r/netflixbestof* and *r/truefilm*. We process the raw data with the pipeline of *conversational recommendation identification*, *movie mention recognition* and *movie entity linking*[8]. In our following evaluation, we use the most recent 9k conversations in *Reddit-Movie*$^{\text{base}}$ from December 2022 as the testing set since these samples occur *after* GPT-3.5-t's release. Meanwhile, GPT-4 [51] also mentioned its pre-training data cut off in Sept. 2021[9]. For other compared models, we use the remaining 76k conversations in *Reddit-Movie*$^{\text{base}}$ dataset for training and validation.

[6]https://pushshift.io/
[7]Other domains like songs, books can potentially be processed in a similar way
[8]Check our evaluation data, LLMs scripts, results and the links of Reddit-Movie datasets in https://github.com/AaronHeee/LLMs-as-Zero-Shot-Conversational-RecSys.
[9]We note that there is a possibility that GPT-4's newest checkpoint might include a small amount of more recent data [51].

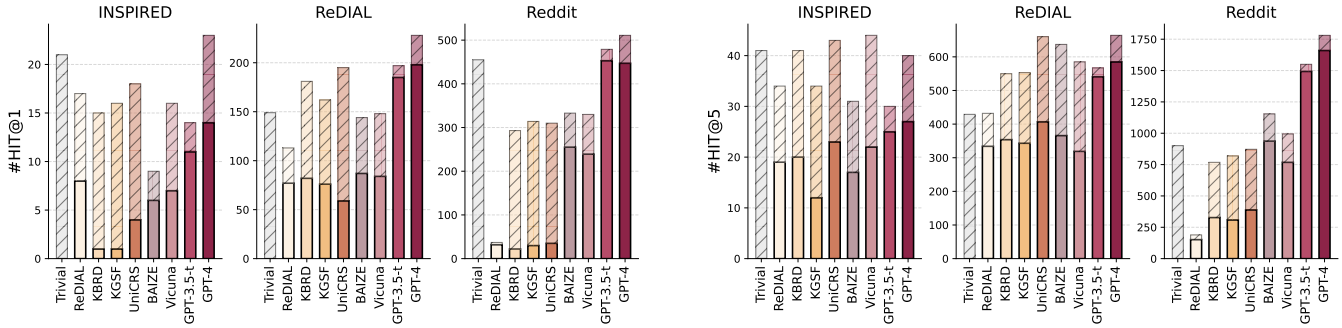


Figure 3: To show the *repeated item shortcut*, we count CRS recommendation hits using the Top-K ranked list $K = \{1, 5\}$. We group the ground-truth hits by repeated items (shaded bars) and new items (not shaded bars). The trivial baseline copies existing items from the current conversation history in chronological order, from the most recent and does not recommend new items.

Discussion. From the statistics in Table 1, we observe: (1) The dataset *Reddit-Movie* stands out as the largest conversational recommendation dataset, encompassing 634,392 conversations and covering 51,203 movies. (2) In comparison to ReDIAL [41] and INSPIRED [22], *Reddit-Movie* contains fewer multi-turn conversations, mainly due to the inherent characteristics of Reddit posts. (3) By examining representative examples depicted in Figure 2, we find that *Reddit-Movie* conversations tend to include more complex and detailed user preference in contrast to ReDIAL, as they originate from real-world conversations on Reddit, enriching the conversational recommendation datasets with a diverse range of discussions.

4 EVALUATION

In this section, we evaluate the proposed LLMs-based frameowrk on ReDIAL [41], INSPIRED [22] and our Reddit datasets. We first explain the evaluation setup and a *repeated item shortcut* of the previous evaluation in Sections 4.1 and 4.2. Then, we re-train models and discuss LLM performance in Section 4.3.

4.1 Evaluation Setup

Repeated *vs.* New Items. Given a conversation $C = (u_t, s_t, \mathcal{I}_t)_{t=1}^{T}$, it is challenging to identify the ground-truth recommended items, i.e., whether the mentioned items $\mathcal{I}_k$ at the $k^{\text{th}} (k \leq T)$ turn are used for recommendation purposes. A common evaluation setup assumes that when u_k is the *recommender*, all items $i \in \mathcal{I}_k$ serve as ground-truth recommended items.

In this work, we further split the items $i \in \mathcal{I}_k$ into two categories: *repeated items* or *new items*. Repeated items are items that have appeared in previous conversation turns, i.e., $\{i \mid \exists t \in [1, k), i \in \mathcal{I}_t\}$; and new items are items not mentioned in previous conversation turns. We explain the details of this categorization in Section 4.2.

Evaluation Protocol. On those three datasets, we evaluate several representative CRS models and several LLMs on their recommendation abilities. For baselines, after re-running the training code provided by the authors, we report the prediction performance using Recall@K [10, 41, 64, 74] (i.e., HIT@K). We consider the means and the standard errors[10] of the metric with $K = \{1, 5\}$.

[10]We show standard errors as error bars in our figures and gray numbers in our tables.

Compared CRS Models. We consider several representative CRS models. For baselines which rely on structured knowledge, we use the entity linking results of ReDIAL and INSPIRED datasets provided by UniCRS [64]. Note that we do not include more works [43, 50, 54] because UniCRS [64] is representative with similar results.

- **ReDIAL** [41]: This model is released along with the ReDIAL dataset with an auto-encoder [58]-based recommender.
- **KBRD** [10]: This model proposes to use the DBPedia [1] to enhance the semantic knowledge of items or entities.
- **KGSF** [74]: This model incorporates two knowledge graphs to enhance the representations of words and entities, and uses the Mutual Information Maximization method to align the semantic spaces of those two knowledge graphs.
- **UniCRS** [64]: This model uses pre-trained language model, DialoGPT [69], with prompt tuning to conduct recommendation and conversation generation tasks respectively.

4.2 Repeated Items Can Be Shortcuts

Current evaluation for conversational recommendation systems does not differentiate between repeated and new items in a conversation. We observed that this evaluation scheme favors systems that optimize for mentioning repeated items. As shown in Figure 3, a trivial baseline that always copies seen items from the conversation history has better performance than most previous models under the standard evaluation scheme. This phenomenon highlights the risk of shortcut learning [18], where a decision rule performs well against certain benchmarks and evaluations but fails to capture the true intent of the system designer. Indeed, the #HIT@1 for the models tested dropped by more than 60% on average when we focus on new item recommendation only, which is unclear from the overall recommendation performance. After manually checking, we observe a typical pattern of repeated items, which is shown in the example conversation in Figure 1. In this conversation, `Terminator` at the 6^{th} turn is used as the ground-truth item. The system repeated this `Terminator` because the system quoted this movie for a content-based discussion during the conversation rather than making recommendations. Given the nature of recommendation conversations between two users, it is more probable that items repeated during a conversation are intended for discussion rather

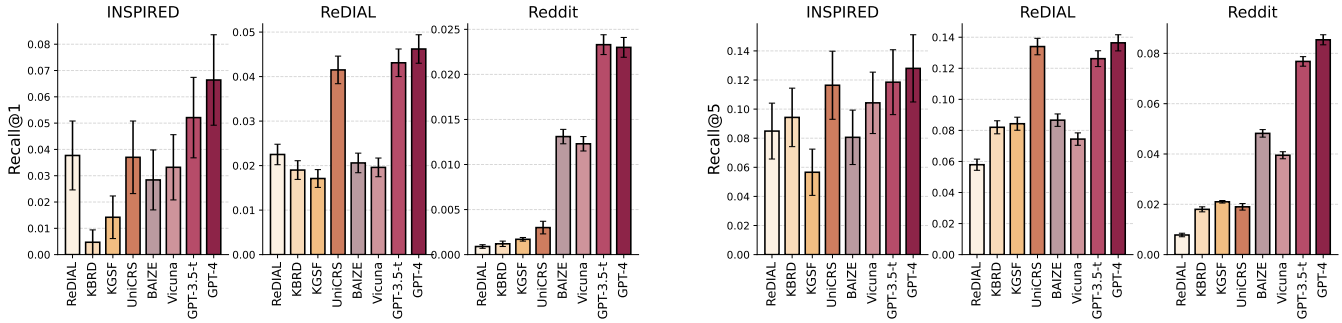


Figure 4: CRS recommendation performance on New Items in terms of Recall@K, with $K = \{1, 5\}$. To exclude the influence of repeated items in CRS evaluation, we remove all repeated items in training and testing datasets and re-train all baselines.

Table 2: Recall@1 results of considering all generated item titles (Φ_0) and only considering in-dataset item titles (Φ_1).

	INSPIRED		ReDIAL		Reddit	
Model	Φ_0	Φ_1	Φ_0	Φ_1	Φ_0	Φ_1
BAIZE	.019 .019	**.028** .011	**.021** .002	**.021** .002	.012 .001	**.013** .008
Vicuna	.028 .011	**.033** .012	**.020** .002	**.020** .002	**.012** .001	**.012** .001
GPT-3.5-t	.047 .015	**.052** .015	.041 .003	**.043** .003	.022 .001	**.023** .001
GPT-4	.062 .017	**.066** .017	.043 .003	**.046** .004	.022 .001	**.023** .001

Table 3: Fraction of Top-K ($K = 20$ in our prompt setup) recommendations (#rec) that can be string matched in the IMDB movie database (%imdb) for the different models, which shows a lower bound of non-hallucinated movie titles.

BAIZE		Vicuna		GPT-3.5-t		GPT-4	
#rec	%imdb	#rec	%imdb	#rec	%imdb	#rec	%imdb
259,333	81.56%	258,984	86.98%	321,048	95.51%	322,323	94.86%

than serving as recommendations. We argue that considering the large portion of repeated items (e.g., more than 15% ground-truth items are repeated items in INSPIRED), it is beneficial to remove repeated items and re-evaluate CRS models to better understand models' recommendation ability. It is worth noting that the repetition patterns have also been investigated in evaluating other recommender systems such as *next-basket* recommendation [40].

4.3 LLMs Performance

Finding 1 - LLMs outperform fine-tuned CRS models in a zero-shot setting. For a comparison between models' abilities to recommend new items to the user in conversation, we re-train existing CRS models on all datasets for new item recommendation only. The evaluation results are as shown in Figure 4. Large language models, although not fine-tuned, have the best performance on all datasets. Meanwhile, the performance of all models is uniformly lower on Reddit compared to the other datasets, potentially due to the large number of items and fewer conversation turns, making recommendation more challenging.

Finding 2 - GPT-based models achieve superior performance than open-sourced LLMs. As shown in Figure 4, large language models consistently outperform other models across all three datasets, while GPT-4 is generally better than GPT-3.5-t. We hypothesize this is due to GPT-4's larger parameter size enables it to retain more correlation information between movie names and user preferences that naturally occurs in the language models' pre-training data. Vicuna and BAIZE, while having comparable performance to prior models on most datasets, have significantly lower performance than its teacher, GPT-3.5-t. This is consistent with previous works' finding that smaller distilled models via imitation learning cannot fully inherit larger models ability on downstream tasks [20].

Finding 3 - LLMs may generate out-of-dataset item titles, but few hallucinated recommendations. We note that language models trained on open-domain data naturally produce items out of the allowed item set during generation. In practice, removing these items improves the models' recommendation performance. Large language models outperform other models (with GPT-4 being the best) consistently regardless of whether these unknown items are removed or not, as shown in Table 2. Meanwhile, Table 3 shows that around 95% generated recommendations from GPT-based models (around 81% from BAIZE and 87% from Vicuna) can be found in IMDB [11] by string matching. Those lower bounds of these matching rates indicate that there are only a few hallucinated item titles in the LLM recommendations in the movie domain.

5 DETAILED ANALYSIS

Observing LLMs' remarkable conversational recommendation performance for zero-shot recommendation, we are interested in *what accounts for their effectiveness* and *what their limitations are*. We aim to answer these questions from both a model and data perspective.

5.1 Knowledge in LLMs

Experiment Setup. Motivated by the probing work of [53], we posit that two types of knowledge in LLMs can be used in CRS:

- **Collaborative knowledge**, which requires the model to match items with similar ones, according to community interactions like "users who like A typically also like B". In

[11] Movie titles in https://datasets.imdbws.com/.

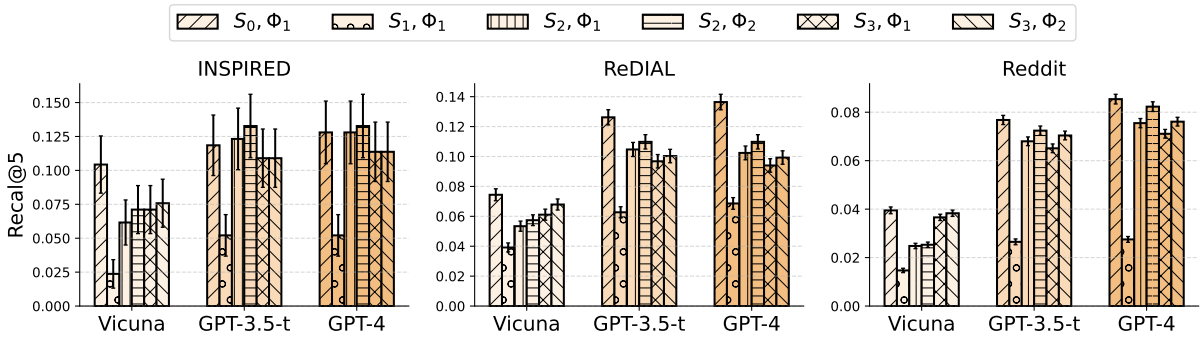


Figure 5: Ablation studies for the research question about the primary knowledge used by LLMs for CRS. Here Φ_1 is the post-processor which only considers in-dataset item titles; Φ_2 is the post-processor based on Φ_1 and further excludes all seen items in conversational context from generated recommendation lists. For inputs like *Original* (S_0) and *ItemOnly* (S_1), LLMs show similar performance with Φ_1 or Φ_2, so we only keep Φ_1 here. We consider Φ_2 because *ItemRemoved* (S_2) and *ItemRandom* (S_3) have no information about already mentioned items, which may cause under-estimated accuracy using Φ_1 compared to *Original*.

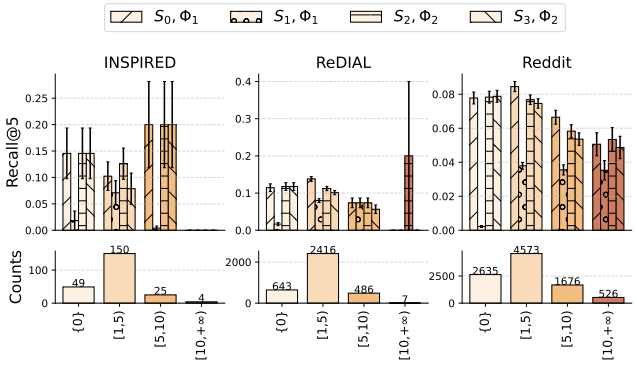


Figure 6: `GPT-3.5-t` Recall@5 results grouped by the occurrences of items in conversation context, and count the conversations per dataset.

our experiments, we define the collaborative knowledge in LLMs as *the ability to make accurate recommendations using item mentions in conversational contexts.*

- **Content/context knowledge**, which requires the model to match recommended items with their content or context information. In our experiments, we define the content/context knowledge in LLMs as *the ability to make accurate recommendations based on all other conversation inputs rather than item mentions,* such as contextual descriptions, mentioned genres, and director names.

To understand how LLMs use these two types of knowledge, given the original conversation context S (Example in Figure 1), we perturb S with three different strategies as follows and subsequently re-query the LLMs. We denote the original as S_0:

- **S_0 (Original):** we use the original conversation context.
- **S_1 (ItemOnly):** we keep mentioned items and remove all natural language descriptions in the conversation context.
- **S_2 (ItemRemoved):** we remove mentioned items and keep other content in the conversation context.

Table 4: To understand the content/context knowledge in LLMs and existing CRS models, we re-train the existing CRS models using the same perturbed conversation context *ItemRemoved* (S_2). We include the results of the representative CRS model UniCRS (denoted as CRS*) as well as a representative text-encoder `BERT-small` [15] (denoted as TextEnc*).

	INSPIRED		ReDIAL		Reddit	
Model	R@1	R@5	R@1	R@5	R@1	R@5
Vicuna	.024 .010	.062 .017	.014 .002	.053 .003	.008 .001	.025 .001
`GPT-3.5-t`	.057 .016	.123 .023	.030 .003	**.105** .005	.018 .001	.068 .002
`GPT-4`	**.062** .017	**.128** .023	**.032** .003	.102 .005	**.019** .001	**.075** .002
CRS*	.039 .011	.087 .014	.015 .002	.058 .003	.001 .000	.008 .001
TextEnc*	.038 .015	.090 .016	.013 .002	.053 .004	.002 .000	.009 .001

- **S_3 (ItemRandom):** we replace the mentioned items in the conversation context with items that are uniformly sampled from the item set $\mathcal{I}$ of this dataset, to eliminate the potential influence of S_2 on the sentence grammar structure.

Finding 4 - LLMs mainly rely on content/context knowledge to make recommendations. Figure 5 shows a drop in performance for most models across various datasets when replacing the original conversation text *Original* (S_0) with other texts, indicating that LLMs leverage both content/context knowledge and collaborative knowledge in recommendation tasks. However, the importance of these knowledge types differs. Our analysis reveals that content/context knowledge is the primary knowledge utilized by LLMs in CRS. When using *ItemOnly* (S_1) as a replacement for *Original*, there is an average performance drop of more than 60% in terms of Recall@5. On the other hand, GPT-based models experience only a minor performance drop of less than 10% on average when using *ItemRemoved* (S_2) or *ItemRandom* (S_3) instead of *Original*. Although the smaller-sized model Vicuna shows a higher performance drop, it is still considerably milder compared to using *ItemOnly*. To accurately reflect the recommendation abilities of LLMs with *ItemRemoved* and *ItemRandom*, we introduce a new post-processor

Table 5: To understand the collaborative knowledge in LLMs and existing CRS models, we re-train the existing CRS models using the same perturbed conversation context *ItemOnly* (S_1). We include the results of the representative CRS model UniCRS (denoted as CRS*) as well as a representative item-based collaborative model FISM [31] (denoted as ItemCF*).

	INSPIRED		ReDIAL		Reddit	
Model	R@1	R@5	R@1	R@5	R@1	R@5
Vicuna	.005 .005	.024 .010	.011 .002	.039 .003	.005 .000	.015 .001
GPT-3.5-t	.024 .010	.052 .015	.021 .002	.063 .004	.007 .001	.026 .001
GPT-4	.014 .008	.052 .015	.025 .002	.069 .004	**.007** .001	**.028** .001
CRS*	.038 .013	.085 .019	.025 .002	.072 .004	.003 .000	.015 .001
ItemCF*	**.042** .012	**.087** .016	**.029** .003	**.088** .004	.004 .001	.018 .001

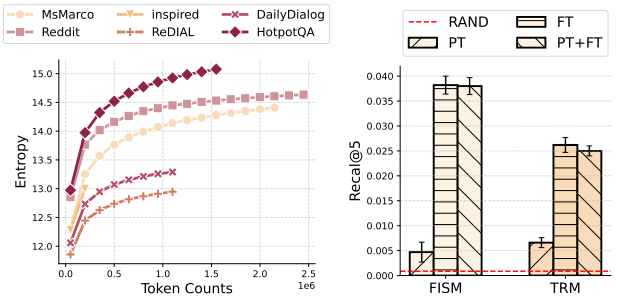
denoted as Φ_2 (describe in the caption of Figure 5). By employing Φ_2, the performance gaps between *Original* and *ItemRemoved* (or *ItemRandom*) are further reduced. Furthermore, Figure 6 demonstrates the consistent and close performance gap between *Original* and *ItemRemoved* (or *ItemRandom*) across different testing samples, which vary in size and the number of item mentions in *Original*.

These results suggest that given a conversation context, LLMs primarily rely on content/context knowledge rather than collaborative knowledge to make recommendations. This behavior interestingly diverges from many traditional recommenders like collaborative filtering [23, 24, 36, 46, 55, 58] or sequential recommenders [25, 33, 59, 73], where user-interacted items are essential.

Finding 5 - GPT-based LLMs possess better content/context knowledge than existing CRS. From Table 4, we observe the superior recommendation performance of GPT-based LLMs against representative conversational recommendation or text-only models on all datasets, showing the remarkable zero-shot abilities in understanding user preference with the textual inputs and generating correct item titles. We conclude that GPT-based LLMs can provide more accurate recommendations than existing trained CRS models in an *ItemRemoved* (S_2) setting, demonstrating better content/context knowledge.

Finding 6 - LLMs generally possess weaker collaborative knowledge than existing CRS. In Table 5, the results from INSPIRED and ReDIAL indicate that LLMs underperform existing representative CRS or ItemCF models by 30% when using only the item-based conversation context *ItemOnly* (S_1). It indicates that LLMs, trained on a general corpus, typically lack the collaborative knowledge exhibited by representative models trained on the target dataset. There are several possible reasons for this weak collaborative knowledge in LLMs. First, the training corpus may not contain sufficient information for LLMs to learn the underlying item similarities. Second, although LLMs may possess some collaborative knowledge, they might not align with the interactions in the target datasets, possibly because the underlying item similarities can be highly dataset- or platform-dependent.

However, in the case of the Reddit dataset, LLMs outperform baselines in both Recall@1 and Recall@5, as shown in Table 5. This outcome could be attributed to the dataset's large number of rarely interacted items, resulting in limited collaborative information. The



(a) Entropy v.s. Token Counts **(b) Pre-training Effectiveness**

Figure 7: The left subfigure shows the entropy of the frequency distribution of 1,2,3-grams with respect to number of words drawn from each dataset (item names excluded) to measure the content/context information across datasets. The right subfigure shows the results of processed Reddit collaborative dataset aligned to ML-25M [21]. RAND denotes random baseline, FT denotes fine tuning on Reddit, PT denotes pre-training on ML-25M, PT+FT means FT after PT.

Reddit dataset contains 12,982 items with no more than 3 mentions as responses. This poses a challenge in correctly ranking these items within the Top-5 or even Top-1 positions. LLMs, which possess at least some understanding of the semantics in item titles, have the chance to outperform baselines trained on datasets containing a large number of cold-start items.

Recent research on LLMs in traditional recommendation systems [27, 34, 48] also observes the challenge of effectively leveraging collaborative information without knowing the target interaction data distribution. Additionally, another study [3] on traditional recommendation systems suggests that LLMs are beneficial in a setting with many cold-start items. Our experimental results support these findings within the context of conversational recommendations.

5.2 Information from CRS Data

Experimental Setup for Finding 7. To understand LLMs in CRS tasks from the data perspective, we first measure the *content/context information* in CRS datasets. Content/context information refers to the amount of information contained in conversations, excluding the item titles, which reasonably challenges existing CRS and favors LLMs according to the findings in Section 5.1. Specifically, we conduct an entropy-based evaluation for each CRS dataset and compare the conversational datasets with several popular conversation and question-answering datasets, namely DailyDialog (chit chat) [45], MsMarco (conversational search) [2], and HotpotQA (question answering). We use *ItemRemoved* (S_2) conversation texts like Section 5.1, and adopt the geometric mean of the entropy distribution of 1,2,3-grams as a surrogate for the amount of information contained in the datasets, following previous work on evaluating information content in text [29]. However, entropy naturally grows with the size of a corpus, and each CRS dataset has a different distribution of words per sentence, sentences per dialog, and corpus size. Thus, it would be unfair to compare entropy between corpus on a

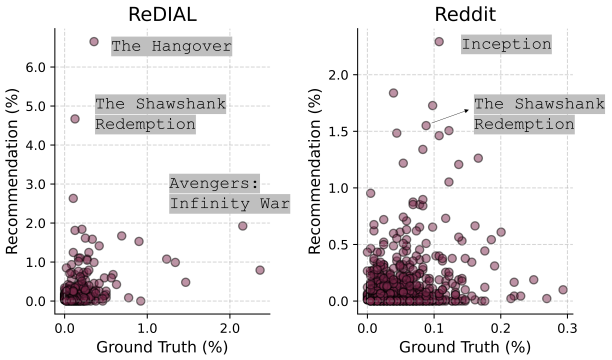


Figure 8: Scatter plots of the frequency of LLMs (GPT-4) generated recommendations and ground-truth items.

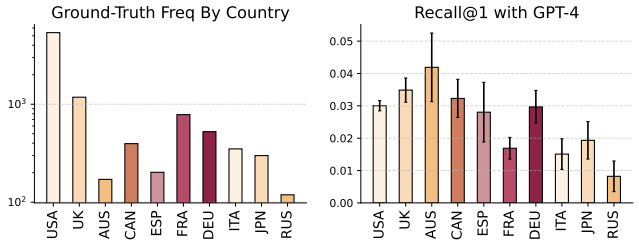


Figure 9: Ground-truth item counts in Reddit by country (in log scale) and the corresponding Recall@1 by country.

per-dialog, per-turn, or per-dataset basis. To ensure a fair comparison, we repeatedly draw increasingly large subsets of texts from each of the datasets, compute the entropy of these subsets, and report the trend of entropy growth with respect to the size of the subsampled text for each CRS dataset.

Finding 7 - Reddit provides more content/context information than the other two CRS datasets. Based on the results in Figure 7a, we observe that the Reddit dataset has the most content/context information among the three conversational recommendation datasets. Those observations are also aligned with the results in Figure 5 and table 4, where LLMs – which possess better content/context knowledge than baselines – can achieve higher relative improvements compared to the other two datasets. Meanwhile, the content/context information in Reddit is close to question answering and conversational search, which is higher than existing conversational recommendation and chit-chat datasets.

Finding 8 - Collaborative information is insufficient for satisfactory recommendations, given the current models. Quantifying the collaborative information in datasets is challenging. Instead of proposing methods to measure collaborative information, we aim to make new observations based on general performance results presented in Figure 4 and recommendation results using only collaborative information in Table 5. Comparing the performance of the best models in Table 5 under an *ItemOnly* (S_1) setting with the performance of the best models in Figure 4 under an *Original* (S_0) setting reveals a significant disparity. For instance, on ReDIAL, the Recall@1 performance is 0.029 for ItemCF* compared to 0.046 for GPT-4, representing a 39.96% decrease. Similarly, for Reddit, the Recall@1 performance is 0.007 compared to 0.023 for GPT-4 both, which is 69.57% lower. We also experimented with other recommender systems, such as transformer-based models [33, 59] to encode the item-only inputs and found similar results. Based on the current performance gap, we find that using the existing models, relying solely on collaborative information, is insufficient to provide satisfactory recommendations. We speculate that either (1) more advanced models or training methods are required to better comprehend the collaborative information in CRS datasets, or (2) the collaborative information in CRS datasets is too limited to support satisfactory recommendations.

Experimental Setup for Finding 9. To understand whether the collaborative information from CRS datasets are aligned with pure interaction datasets, we conduct an experiment on the Reddit dataset. In this experiment, we first process the dataset to link the items to a popular interaction dataset ML-25M [21] [12]. We then experiment with two representative encoders for item-based collaborative filtering based on FISM [31] and Transformer [59] (TRM), respectively. We report the testing results on Reddit, with fine-tuning on Reddit (FT), pre-training on ML-25M (PT), and pre-training on ML-25M then fine-tuning Reddit (PT+FT). Note that since it is a linked dataset with additional processing, the results are not comparable with beforementioned results on Reddit.

Finding 9 - Collaborative information can be dataset- or platform-dependent. From Figure 7b shows that the models solely pre-trained on ML-25M (PT) outperform a random baseline, indicating that the data in CRS may share item similarities with pure interaction data from another platform to some extent. However, Figure 7b also shows a notable performance gap between PT and fine-tuning on Reddit (FT). Additionally, we do not observe further performance improvement when pre-training on ML-25M then fine-tuning on Reddit (PT+FT). These observations indicate that the collaborative information and underlying item similarities, even when utilizing the same items, can be largely influenced by the specific dataset or platform. The finding also may partially explain the inferior zero-shot recommendation performance of LLMs in Table 5 and suggest the necessity of further checking the alignment of collaborative knowledge in LLMs with the target datasets.

5.3 Limitations of LLMs as Zero-shot CRS

Finding 10 - LLM recommendations suffer from popularity bias in CRS. Popularity bias refers to a phenomenon that popular items are recommended even more frequently than their popularity would warrant [8]. Figure 8 shows the popularity bias in LLM recommendations, though it may not be biased to the popular items in the target datasets. On ReDIAL, the most popular movies such as `Avengers: Infinity War` appear around 2% of the time over all ground-truth items; On Reddit, the most popular movies such as `Everything Everywhere All at Once` appears less than 0.3% of the time over ground-truth items. But for the *generated* recommendations from GPT-4 (other LLMs share a similar trend),

[12]We only use items that can be linked to ML-25M in this experiment. Here 63.32% items are linked using the `links.csv` file from ML-25M.

the most popular items such as `The Shawshank Redemption` appear around 5% times on ReDIAL and around 1.5% times on Reddit. Compared to the target datasets, LLMs recommendations are more concentrated on popular items, which may cause further issues like the bias amplification loop [8]. Moreover, the recommended popular items are similar across different datasets, which may reflect the item popularity in the pre-training corpus of LLMs.

Finding 11 - Recommendation performance of LLMs is sensitive to geographical regions. Despite the effectiveness in general, it is unclear whether LLMs can be good recommenders across various cultures and regions. Specifically, pre-trained language models' strong open-domain ability can be attributed to pre-training from massive data [5]. But it also leads to LLMs' sensitivity to data distribution. To investigate LLMs recommendation abilities for various regions, we take test instances from the Reddit dataset and obtain the production region of 7,476 movies from a publicly available movie dataset [13] by exact title matching, then report the Recall@1 for the linked movies grouped by region. We only report regions with more than 300 data points available to ensure enough data to support the result. As shown in Figure 9 the current best model, GPT-4's performance on recommendation is higher for movies produced in English-speaking regions. This could be due to bias in the training data - the left of Figure 9 show item on Reddit forums are dominated by movies from English-speaking regions. Such a result highlights large language model's recommendation performance varies by region and culture and demonstrates the importance of cross-regional analysis and evaluation for language model-based conversational recommendation models.

6 RELATED WORK

Conversational Recommendation. Conversational recommender systems (CRS) aim to understand user preferences and provide personalized recommendations through conversations. Typical traditional CRS setups include template-based CRS [13, 26, 37, 38, 70] and critiquing-based CRS [9, 42, 67]. More recently, as natural language processing has advanced, the community developed "deep" CRS [10, 41, 64] that support interactions in natural language. Aside from collaborative filtering signals, prior work shows that CRS models benefit from various additional information. Examples include knowledge-enhanced models [10, 74] that make use of external knowledge bases [1, 47], review-aware models [49], and session/sequence-based models [43, 76]. Presently, UniCRS [64], a model built on DialoGPT [69] with prompt tuning [4], stands as the state-of-the-art approach on CRS datasets such as ReDIAL [41] and INSPIRED [22]. Currently, by leveraging LLMs, [16] proposes a new CRS pipeline but does not provide quantitative results, and [63] proposes better user simulators to improve evaluation strategies in LLMs. Unlike those papers, we uncover a *repeated item shortcut* in the previous evaluation protocol, and propose a framework where LLMs serve as zero-shot CRS with detailed analyses to support our findings from both model and data perspectives.

Large Language Models. Advances in natural language processing (NLP) show that large language models (LLMs) exhibit strong generalization ability towards unseen tasks and domains [5, 12, 65]. In particular, existing work reveals language models' performance and sample efficiency on downstream tasks can be improved simply through scaling up their parameter sizes [35]. Meanwhile, language models could further generalize to a wide range of unseen tasks by instruction tuning, learning to follow task instructions in natural language [52, 57]. Following these advances, many works successfully deploy large language models to a wide range of downstream tasks such as question answering, numerical reasoning, code generation, and commonsense reasoning without any gradient updates [5, 35, 44, 72]. Recently, there have been various attempts by the recommendation community to leverage large language models for recommendation, this includes both adapting architectures used by large language models [14, 19] and repurposing existing LLMs for recommendation [39, 48, 62]. However, to our best knowledge, we are the first work that provides a systematic quantitative analysis of LLMs' ability on *conversational* recommendation.

7 CONCLUSION AND DISCUSSION

We investigate Large Language Models (LLMs) as zero-shot Conversational Recommendation Systems (CRS). Through our empirical investigation, we initially address a repetition shortcut in previous standard CRS evaluations, which can potentially lead to unreliable conclusions regarding model design. Subsequently, we demonstrate that LLMs as zero-shot CRS surpass all fine-tuned existing CRS models in our experiments. Inspired by their effectiveness, we conduct a comprehensive analysis from both the model and data perspectives to gain insights into the working mechanisms of LLMs, the characteristics of typical CRS tasks, and the limitations of using LLMs as CRS directly. Our experimental evaluations encompass two publicly available datasets, supplemented by our newly-created dataset on movie recommendations collected by scraping a popular discussion website. This dataset is the largest public CRS dataset and ensures more diverse and realistic conversations for CRS research. We also discuss the future directions based on our findings in this section.

On LLMs. Given the remarkable performance even without fine-tuning, LLMs hold great promise as an effective approach for CRS tasks by offering superior content/contextual knowledge. The encouraging performance from the open-sourced LLMs [11, 68] also opens up the opportunities to further improve CRS performance via efficient tuning [3, 28] and collaborative filtering [36] ensembling. Meanwhile, many conventional tasks, such as debiasing [8] and trustworthy [17] need to be revisited in the context of LLMs.

On CRS. Our findings suggest the systematic re-benchmarking of more CRS models to understand their recommendation abilities and the characteristics of CRS tasks comprehensively. Gaining a deeper understanding of CRS tasks also requires new datasets from diverse sources e.g., crowd-sourcing platforms [22, 41], discussion forums, and realistic CRS applications with various domains, languages, and cultures. Meanwhile, our analysis of the information types uncovers the unique importance of the superior content/context knowledge in LLMs for CRS tasks; this distinction also sets CRS tasks apart from traditional recommendation settings and urges us to explore the interconnections between CRS tasks and traditional *recommendation* [21] or *conversational search* [2] tasks.

[13] https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset

REFERENCES

[1] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007+ ASWC 2007, Busan, Korea, November 11-15, 2007. Proceedings*. Springer, 722–735.

[2] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. arXiv:1611.09268 [cs.CL]

[3] 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* (2023).

[4] 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 (2020), 1877–1901.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[6] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* (2023).

[7] Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive Entity Retrieval. In *International Conference on Learning Representations*. https://openreview.net/forum?id=5k8F6UU39V

[8] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2023. Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems* 41, 3 (2023), 1–39.

[9] Li Chen and Pearl Pu. 2012. Critiquing-based recommenders: survey and emerging trends. *User Modeling and User-Adapted Interaction* 22 (2012), 125–150.

[10] 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)*. 1803–1813.

[11] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. https://lmsys.org/blog/2023-03-30-vicuna/

[12] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. *ArXiv* abs/2204.02311 (2022).

[13] 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*. 815–824.

[14] Zeyu Cui, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. M6-Rec: Generative Pretrained Language Models are Open-Ended Recommender Systems. arXiv:2205.08084 [cs.IR]

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

[16] Luke Friedman, Sameer Ahuja, David Allen, Terry Tan, Hakim Sidahmed, Changbo Long, Jun Xie, Gabriel Schubiner, Ajay Patel, Harsh Lara, et al. 2023. Leveraging Large Language Models in Conversational Recommender Systems. *arXiv preprint arXiv:2305.07961* (2023).

[17] Yingqiang Ge, Shuchang Liu, Zuohui Fu, Juntao Tan, Zelong Li, Shuyuan Xu, Yunqi Li, Yikun Xian, and Yongfeng Zhang. 2022. A survey on trustworthy recommender systems. *arXiv preprint arXiv:2207.12515* (2022).

[18] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence* 2 (2020), 665 – 673.

[19] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2022. Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt & Predict Paradigm (P5). In *RecSys '22: Sixteenth ACM Conference on Recommender Systems, Seattle, WA, USA, September 18 - 23, 2022*, Jennifer Golbeck, F. Maxwell Harper, Vanessa Murdock, Michael D. Ekstrand, Bracha Shapira, Justin Basilico, Keld T. Lundgaard, and Even Oldridge (Eds.). ACM, 299–315.

[20] Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. The False Promise of Imitating Proprietary LLMs. arXiv:2305.15717 [cs.CL]

[21] F. Maxwell Harper and Joseph A. Konstan. 2016. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.* 5 (2016), 19:1–19:19.

[22] Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang 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)*. 8142–8152.

[23] Xiangnan He, Zhankui He, Xiaoyu Du, and Tat-Seng Chua. 2018. Adversarial personalized ranking for recommendation. In *The 41st International ACM SIGIR conference on research & development in information retrieval*. 355–364.

[24] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*. 173–182.

[25] Zhankui He, Handong Zhao, Zhe Lin, Zhaowen Wang, Ajinkya Kale, and Julian McAuley. 2021. Locker: Locally constrained self-attentive sequential recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 3088–3092.

[26] 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*. 288–298.

[27] 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* (2023).

[28] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

[29] Harsh Jhamtani, Varun Gangal, Eduard Hovy, Graham Neubig, and Taylor Berg-Kirkpatrick. 2018. Learning to Generate Move-by-Move Commentary for Chess Games from Large-Scale Social Forum Data. In *The 56th Annual Meeting of the Association for Computational Linguistics (ACL)*. Melbourne, Australia.

[30] C Kim Jacob Hilton Jacob Menick Jiayi Weng Juan Felipe Ceron Uribe Liam Fedus Luke Metz Michael Pokorny Rapha Gontijo Lopes Sengjia Zhao John Schulman, Barret Zoph. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI* (2022).

[31] Santosh Kabbur, Xia Ning, and George Karypis. 2013. Fism: factored item similarity models for top-n recommender systems. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 659–667.

[32] Dongyeop Kang, Anusha Balakrishnan, Pararth Shah, Paul A Crook, Y-Lan Boureau, and Jason Weston. 2019. Recommendation as a Communication Game: Self-Supervised Bot-Play for Goal-oriented Dialogue. 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)*. 1951–1961.

[33] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*. IEEE, 197–206.

[34] Wang-Cheng Kang, Jianmo Ni, Nikhil Mehta, Maheswaran Sathiamoorthy, Lichan Hong, Ed Chi, and Derek Zhiyuan Cheng. 2023. Do LLMs Understand User Preferences? Evaluating LLMs On User Rating Prediction. *arXiv preprint arXiv:2305.06474* (2023).

[35] Jared Kaplan, Sam McCandlish, T. J. Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeff Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *ArXiv* abs/2001.08361 (2020).

[36] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.

[37] 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*. 304–312.

[38] Wenqiang Lei, Gangyi Zhang, Xiangnan He, Yisong Miao, Xiang Wang, Liang Chen, and Tat-Seng Chua. 2020. Interactive path reasoning on graph for conversational recommendation. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 2073–2083.

[39] Jinming Li, Wentao Zhang, Tian Wang, Guanglei Xiong, Alan Lu, and Gerard Medioni. 2023. GPT4Rec: A Generative Framework for Personalized Recommendation and User Interests Interpretation. arXiv:2304.03879 [cs.IR]

[40] Ming Li, Sami Jullien, Mozhdeh Ariannezhad, and Maarten de Rijke. 2023. A next basket recommendation reality check. *ACM Transactions on Information Systems* 41, 4 (2023), 1–29.

[41] Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. *Advances in neural information processing systems* 31 (2018).

[42] Shuyang Li, Bodhisattwa Prasad Majumder, and Julian McAuley. 2021. Self-Supervised Bot Play for Conversational Recommendation with Justifications. *arXiv preprint arXiv:2112.05197* (2021).

[43] Shuokai Li, Ruobing Xie, Yongchun Zhu, Xiang Ao, Fuzhen Zhuang, and Qing He. 2022. User-centric conversational recommendation with multi-aspect user modeling. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 223–233.

[44] Yujia Li, David H. Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom, Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de, Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey, Cherepanov, James Molloy, Daniel Jaymin Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de, Freitas, Koray Kavukcuoglu, and Oriol Vinyals. 2022. Competition-level code generation with AlphaCode. *Science* 378 (2022), 1092 – 1097.

[45] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Asian Federation of Natural Language Processing, Taipei, Taiwan, 986–995. https://aclanthology.org/I17-1099

[46] Dawen Liang, Rahul G Krishnan, Matthew D Hoffman, and Tony Jebara. 2018. Variational autoencoders for collaborative filtering. In *Proceedings of the 2018 world wide web conference*. 689–698.

[47] Hugo Liu and Push Singh. 2004. ConceptNet—a practical commonsense reasoning tool-kit. *BT technology journal* 22, 4 (2004), 211–226.

[48] Junling Liu, Chao Liu, Renjie Lv, Kang Zhou, and Yan Zhang. 2023. Is ChatGPT a Good Recommender? A Preliminary Study. arXiv:2304.10149 [cs.IR]

[49] Yu Lu, Junwei Bao, Yan Song, Zichen Ma, Shuguang Cui, Youzheng Wu, and Xiaodong He. 2021. RevCore: Review-Augmented Conversational Recommendation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 1161–1173.

[50] Wenchang Ma, Ryuichi Takanobu, and Minlie Huang. 2021. CR-Walker: Tree-Structured Graph Reasoning and Dialog Acts for Conversational Recommendation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://aclanthology.org/2021.emnlp-main.139

[51] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[52] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*. http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html

[53] Gustavo Penha and Claudia Hauff. 2020. What does bert know about books, movies and music? probing bert for conversational recommendation. In *Proceedings of the 14th ACM Conference on Recommender Systems*. 388–397.

[54] Zhaochun Ren, Zhi Tian, Dongdong Li, Pengjie Ren, Liu Yang, Xin Xin, Huasheng Liang, Maarten de Rijke, and Zhumin Chen. 2022. Variational Reasoning about User Preferences for Conversational Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 165–175.

[55] Steffen Rendle. 2010. Factorization machines. In *2010 IEEE International conference on data mining*. IEEE, 995–1000.

[56] Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. LaMP: When Large Language Models Meet Personalization. *arXiv preprint arXiv:2304.11406* (2023).

[57] Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask Prompted Training Enables Zero-Shot Task Generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net. https://openreview.net/forum?id=9Vrb9D0WI4

[58] Suvash Sedhain, Aditya Krishna Menon, Scott Sanner, and Lexing Xie. 2015. Autorec: Autoencoders meet collaborative filtering. In *Proceedings of the 24th international conference on World Wide Web*. 111–112.

[59] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 1441–1450.

[60] Yi Tay, Vinh Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, et al. 2022. Transformer memory as a differentiable search index. *Advances in Neural Information Processing Systems* 35 (2022), 21831–21843.

[61] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[62] Wenjie Wang, Xinyu Lin, Fuli Feng, Xiangnan He, and Tat-Seng Chua. 2023. Generative Recommendation: Towards Next-generation Recommender Paradigm. arXiv:2304.03516 [cs.IR]

[63] Xiaolei Wang, Xinyu Tang, Wayne Xin Zhao, Jingyuan Wang, and Ji-Rong Wen. 2023. Rethinking the Evaluation for Conversational Recommendation in the Era of Large Language Models. *arXiv preprint arXiv:2305.13112* (2023).

[64] 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*. 1929–1937.

[65] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 24824–24837. https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf

[66] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

[67] Ga Wu, Kai Luo, Scott Sanner, and Harold Soh. 2019. Deep language-based critiquing for recommender systems. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 137–145.

[68] Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196* (2023).

[69] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020. DIALOGPT: Large-Scale Generative Pre-training for Conversational Response Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 270–278.

[70] 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*. 2153–2162.

[71] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).

[72] Qinkai Zheng, Xiao Xia, Xu Zou, Yuxiao Dong, Shan Wang, Yufei Xue, Zihan Wang, Lei Shen, Andi Wang, Yang Li, Teng Su, Zhilin Yang, and Jie Tang. 2023. CodeGeeX: A Pre-Trained Model for Code Generation with Multilingual Evaluations on HumanEval-X. arXiv:2303.17568 [cs.LG]

[73] Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. 2020. S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 1893–1902.

[74] 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*. 1006–1014.

[75] Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. 2020. Towards Topic-Guided Conversational Recommender System. In *Proceedings of the 28th International Conference on Computational Linguistics*. 4128–4139.

[76] Jie Zou, Evangelos Kanoulas, Pengjie Ren, Zhaochun Ren, Aixin Sun, and Cheng Long. 2022. Improving conversational recommender systems via transformer-based sequential modelling. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2319–2324.