

# Evaluating Large Language Models as Generative User Simulators for Conversational Recommendation

Se-eun Yoon   Zhankui He   Jessica Maria Echterhoff   Julian McAuley

University of California, San Diego

{seeuny, zhh004, jechterh, jmcauley}@ucsd.edu

## Abstract

Synthetic users are cost-effective proxies for real users in the evaluation of conversational recommender systems. Large language models show promise in simulating human-like behavior, raising the question of their ability to represent a diverse population of users. We introduce a new protocol to measure the degree to which language models can accurately emulate human behavior in conversational recommendation. This protocol is comprised of five tasks, each designed to evaluate a key property that a synthetic user should exhibit: choosing which items to talk about, expressing binary preferences, expressing open-ended preferences, requesting recommendations, and giving feedback. Through evaluation of baseline simulators, we demonstrate these tasks effectively reveal deviations of language models from human behavior, and offer insights on how to reduce the deviations with model selection and prompting strategies.<sup>1</sup>

## 1 Introduction

In everyday life, recommendations are often sought through conversations: we ask others for advice on which movies to watch, appliances to buy, or restaurants to explore. Such experience is what conversational recommendation systems (CRSs) seek to provide, by developing autonomous agents that could chat with users, understand their needs, and provide well-tailored recommendations. A core challenge that hinders the advancement of the field is evaluation (Gao et al., 2021). While an ideal approach would involve comprehensive testing with real user interactions, the associated costs and risks drive studies towards proxy methods, which are limited in representing real user evaluation. Offline evaluation restricts evaluation to non-interactive modes, allowing only single-turn assessments (Li et al., 2018; Moon et al., 2019; He et al., 2023).

<sup>1</sup>We release our code and datasets at <https://github.com/granelle/naacl24-user-sim>.

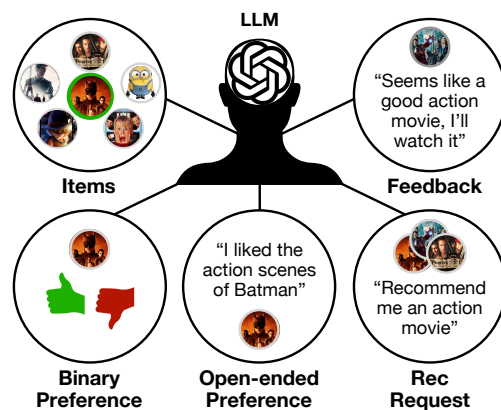


Figure 1: To be successful user simulators for conversational recommendation, representing a population of users, LLMs must fulfill a variety of tasks.

To enable interactive evaluation, studies have introduced synthetic users. However, they are overly simplified representations of human users, being restricted to binary responses (e.g., yes or no) (Christakopoulou et al., 2016; Lei et al., 2020a) or holding ‘target’ items as if users and agents are playing guessing games (Sun and Zhang, 2018; Lei et al., 2020b; Guo et al., 2018). Other line of work restrict interactions to predetermined rules and templates (Zhang and Balog, 2020; Zhang et al., 2022). Essentially, these user simulators suffer from an inherent constraint: they are static (i.e., confined to a finite set of actions), not generative.

Recently, large language models (LLMs) have demonstrated impressive proficiency in conversational tasks (Pan et al., 2023; Zhao et al., 2023), motivating a growing number of works to explore their capacity to simulate human behavior (Park et al., 2023; Argyle et al., 2023; Aher et al., 2023; Gao et al., 2023; Momennejad et al., 2023). Agents simulated by LLMs are *generative*; conditioned upon profiles and memories, these agents exhibit emergent behaviors that appear believable (Park et al., 2023; Qian et al., 2023; Gao et al., 2023). Studies have also explored the use of LLMs as user

simulators for recommender systems (Wang et al., 2023a,b). An important question in each of these studies is to evaluate how closely these simulators represent humans in the task. While there are automatic evaluation protocols for replicating general human behavior (Aher et al., 2023; Momennejad et al., 2023), no protocol exists in the context of recommendation.

In conversational recommendation, the requirements of user simulators are distinct from general-purpose human simulators. The goal is to simulate a *population* of users, each with distinct *preferences*, in a way that these preferences collectively reflect the characteristics of human preferences. Real user preferences are highly granular and diverse, shaped by each individual’s particular set of traits, interaction history, and circumstances. Such uniqueness is reflected in conversational utterances among individuals, each mentioning distinct items, expressing various preferences, and making highly personalized recommendation requests. There are also population-level patterns in preferences, such as users preferring some items over others. Protocols in other domains are unsuitable for evaluating the requirements in conversational recommendation, since they do not consider the behaviors driven by personal preferences (Momennejad et al., 2023; Aher et al., 2023). Although some work considers the uniqueness of simulated individuals, evaluations are limited to manual case studies (Wang et al., 2023a; Park et al., 2023).

We propose a new evaluation protocol for measuring the extent to which LLM-based simulators can represent users in conversational recommendation. This objective poses new challenges: First, we lack data that maps inputs to ground-truth outputs, such as demographics to survey outcomes (Sanurkar et al., 2023). Second, our outcomes are free-form text, unlike previous work where behavioral outcomes are choices or numerical values for ground truth comparison (Aher et al., 2023). Third, the infinite possibilities of conversational trajectories make the concept of ‘ground truth’ increasingly ambiguous as conversations unfold.

We tackle these challenges by decomposing evaluation into five independent tasks, each measuring a key property that a user simulator should exhibit. Each task prompts a simulator and stores the outcomes of a population of simulators. The outcomes can then be compared to the human data we curate from four different platforms. The tasks by them-

selves do not guarantee simulators to be perfect representations of human users, but rather, capture *distortions* in simulators, which is the systematic difference from humans (Aher et al., 2023).

We demonstrate the effectiveness of these tasks by applying them to baseline simulators and revealing the distortions present in these simulators. We observe that simulators tend to favor mentioning popular items, correlate little with human preferences, exhibit lack of personalization in requests, and occasionally give incoherent feedback. We also identify methods to reduce the gap between simulators and humans, indicating that our evaluation protocol can guide future research in developing more realistic user simulators.

Our work is summarized as follows:

- We propose the first evaluation protocol for LLM-based user simulation in conversational recommendation. Our protocol allows automatic and reproducible evaluation through five tasks and real user datasets.
- By running our tasks, we show how simulators could differ from real users. Discrepancies include low item diversity, low correlation with human preference, lack of request personalization, and incoherent feedback.
- We show that the gaps can be reduced through prompting and model selection strategies.

## 2 Evaluation Tasks

Here we first introduce our tasks and later explain the execution of these tasks in Section 3.

(T1) **ItemsTalk**: *Choosing items to talk about.* Often when users talk about recommendations, they mention items. Contexts may vary: to request similar items, to express preference on certain items, or to simply chat about an item (Li et al., 2018). We compare the distributions of items mentioned by simulators and real users.

(T2) **BinPref**: *Expressing binary preference.* Binary questions, such as ‘Did you enjoy this movie?’ are commonly observed in conversations (Li et al., 2018). While answers need not be binary, we fix the answers to binary in this task and examine how well simulators reflect human preferences.

(T3) **OpenPref**: *Expressing open-ended preference.* Open-ended utterances allow users to express detailed preferences, such as appreciating the cast of a movie while finding the plot uneventful (Xia

Task	Baselines	Datasets	Example prompt
(T1) ItemsTalk	DI, IH	ReDial, Reddit, IMDB	A person mentions <i>Concussion</i> (2015) and <i>Jerry Maguire</i> (1996) in a conversation about movies and proceeds to mention 2 more. What would these 2 movies be?
(T2) BinPref	DI, DI + PP	MovieLens	Pretend to be Ms. Guzman. You watched the movie <i>Whiplash</i> (2014). Did you like the movie? Answer Yes or No. Don't say anything else.
(T3) OpenPref	DI, DI + PP	IMDB	Pretend to be Mr. Li. You watched the movie <i>The Bellboy</i> (1960). What are your thoughts on this movie? Answer should not exceed 809 characters.
(T4) RecRequest	Vanilla LLM	Reddit	Generate a movie recommendation request. Include the following movies in your text: <i>Oldboy</i> (2003), <i>Memento</i> (2000). Length of the request is approximately 374 characters.
(T5) Feedback	Vanilla LLM	Reddit	In the following conversation ... If the recommendation is coherent to your request, answer Accept. If the recommendation is incoherent to your request, answer Reject.

Table 1: Tasks overview. Prompts are partially displayed. See A.2 for full prompt descriptions. (DI: Demographic Information, IH: Interaction History, PP: Pickiness Personality)

et al., 2023). We examine whether simulators can express preferences on aspects of items (e.g., cast and plot), and whether the aspects and preferences are similar to those expressed by real users.

**(T4) RecRequest: Requesting recommendations.** A need for recommendation is verbalized through requests. Requests can range from something general, such as ‘Recommend me a good movie,’ to a more personalized demand, such as ‘Recommend me a movie that involves a lawyer or a magician but does not contain action scenes.’ While related to preferences, requests stem from immediate demand, such as being in a mood for a certain movie. Given the vastness of tastes and circumstances, a wide variety of requests may emerge (He et al., 2023). We investigate whether requests generated by simulators are as diverse as those of real users.

**(T5) Feedback: Giving feedback.** To evaluate CRSs, simulators should be able to provide final feedback of whether the recommendation was successful (Wang et al., 2023b). (Real users may or may not provide explicit feedback, but they have a general impression of whether the recommendation was satisfactory.) Particularly, if recommendations and explanations are relevant to one’s requests and preferences, one should be likely to accept the recommendation. If irrelevant, one should reject them. We examine whether simulators can exhibit such coherent patterns of feedback generation.

### 3 Methods

Our protocol treats the design choices of simulators as a black box. We only require that simulators should accept free-form natural language as input

and generate language as output. As noted in the introduction, we consider the population of users, since recommender systems are tested against a large group of users. Importantly, our focus is not to replicate a fixed pool of users, but to generate a *new* group of users whose behavior characteristics resemble those of human users. Tasks should be zero-shot; simulators should not be trained or conditioned on our tasks, nor be informed about our evaluation metrics. This is to avoid simulators fitting to the tasks instead of performing well in generic situations.

**Datasets** We use real-world datasets to compare simulator outputs to human output. Dataset statistics are summarized in A.1. **ReDial** (Li et al., 2018) consists of multi-turn conversations, where one person plays the role of a movie seeker, and the other as a recommender. We use the seeker side of this dataset.<sup>2</sup> **Reddit** (He et al., 2023) consists of conversations in Reddit communities on movie recommendations. Users post requests for recommendations and other users comment on this post with movies, sometimes with explanations. **MovieLens** (Harper and Konstan, 2015) is a movie ratings dataset consisting of 25M ratings. **IMDB** is a movie review dataset from IMDB,<sup>3</sup> aggregated per user. Each task uses different dataset(s), since the datasets are heterogeneous and are not applicable to every task (see Table 1). We only include movies up to the year 2021 to ensure that LLMs

<sup>2</sup>The Mechanical Turk workers may differ from real users with genuine incentives. However, we assume the person playing as seeker bears insignificant difference from the real seeker, as the role requires less effort and expertise.

<sup>3</sup><https://www.imdb.com/>

have knowledge about the movies.

**Baselines** We use prompt-based simulators as baselines, using OpenAI (OpenAI, 2021) models *gpt-3.5-turbo*, *gpt-4*, and *text-davinci-003*. The baselines can be used in any task, but for each task, we select the baselines that give the best insights (see Table 1). **Vanilla LLM** runs without any specialized prompts designed to induce variability in outputs. Instead, it relies solely on the inherent variability of LLM outputs.<sup>4</sup> **DI** (Demographic Information) is a method used by Aher et al. (2023) to simulate gender and racial diversity. We follow their method and sample from titles Mr. and Ms.,<sup>5</sup> and 500 most common surnames across five racial groups.<sup>6</sup> **DI + PP** (Pickiness Personality) adds a personality trait to demographic information, that is, pickiness toward movies. For each simulator, we randomly sample one of three pickiness levels: not picky, moderately picky, and extremely picky. **IH** (Interaction History) samples a set of interactions from a real user and prompts to act like this user. The interaction may be a subset of mentioned items (ReDial), mentioned items with timestamps (Reddit), or reviews of items (IMDB).

**Execution and evaluation** For each task, we create a population of simulators, each given a task-specific prompt. Example prompts are in Table 1 and all the prompts are in A.2.

*ItemsTalk* prompts the simulator to mention a certain number of items. Each prompt uses a single dataset entry to determine the number of items to mention and interaction history (for the IH baseline); the number of prompts equals the number of dataset entries. For evaluation, we compare the distribution of mentioned items between the simulator and the dataset (items in prompt are removed). The diversity of distribution is summarized by entropy:  $H(X) = -\sum_{i=1}^n p(x_i) \log p(x_i)$ , where  $p(x_i)$  is the probability that an item  $x_i$  is mentioned.

*BinPref* prompts the simulator to act as if one has interacted with an item, and asks whether one has positive opinions on it. We sample two groups of 200 movies from MovieLens: frequent ( $\geq 5000$  ratings) and infrequent ( $\leq 500$  ratings). This is to observe if the simulators reflect human preferences

<sup>4</sup>The degree of variability can be adjusted by the temperature parameter, but we use the default temperature values across all LLMs.

<sup>5</sup>There are more titles, but we use these two for simplicity.

<sup>6</sup>[https://www.census.gov/topics/population/genealogy/data/2010\\_surnames.html](https://www.census.gov/topics/population/genealogy/data/2010_surnames.html)

better on frequent movies. The distribution of average rating ranges from 1 to 5. For each movie, we run 100 simulators to output a binary preference and get the proportion of ‘Yes’ answers (i.e., positive rate). We compute the Pearson correlation coefficient between the average rating and positive rate.

*OpenPref* prompts the simulator to assume one has interacted with a certain item, and asks one’s thoughts on it. Each prompt uses a review from IMDB to obtain target response length. After getting a collection of responses, we conduct aspect-based sentiment analysis with PyABSA (Yang et al., 2023). We compare the aspect and sentiment distributions of humans and simulators.

*RecRequest* prompts the simulator to generate a recommendation request containing a set of items. The reason we include the set of items is to evaluate only the capability to generate requests, and not items included in the request. In each prompt, the items and target length are determined by a real user request in the Reddit dataset; we obtain the same number of requests. We compare the diversity and granularity of synthetic and real requests by computing type-token ratio, word embeddings with Word2Vec (Mikolov et al., 2013), and sentence (request) embeddings with SBERT (Reimers and Gurevych, 2019). We use the cosine diversity of embeddings (Anderson et al., 2020):

$$1 - \frac{1}{N} \sum_{i=1}^N \frac{\vec{s}_i \cdot \hat{\mu}}{\|\vec{s}_i\| \cdot \|\hat{\mu}\|}$$

where  $N$  is the number of embeddings,  $\vec{s}_i$  is the  $i^{th}$  embedding, and  $\hat{\mu}$  is the centroid  $\hat{\mu} = \sum_i \vec{s}_i / N$ .

*Feedback* prompts the simulator to provide feedback to request-recommendation pairs. For each request in the Reddit dataset, we sample: (1) a comment from this request (positive recommendation) (2) a random comment (negative recommendation). We formulate two sub-tasks. *Accept/reject*: simulators should reject negative recommendations. *Comparison*: simulators should prefer positive recommendations over negative ones.

## 4 Experiments

We summarize our experiment results as follows.

**Finding 1: Simulators mention less diverse items compared to real users.** Our first task, *ItemsTalk*, reveals that the distribution of items mentioned by simulators is heavily skewed toward popular items,

Table 2: Entropy of mentioned items. Simulators yield lower entropy, indicating lower diversity. Prompting with interaction history enhances diversity.

Generator	IMDB	Reddit	ReDial
Human	<b>12.61</b>	<b>11.73</b>	<b>9.71</b>
Demographic information			
gpt-3.5	4.79	3.97	4.00
gpt-4	5.29	4.78	4.18
text-davinci	6.42	6.69	6.66
Interaction history			
gpt-3.5	7.96	7.14	7.68
gpt-4	8.59	9.50	9.03
text-davinci	10.79	9.97	8.63

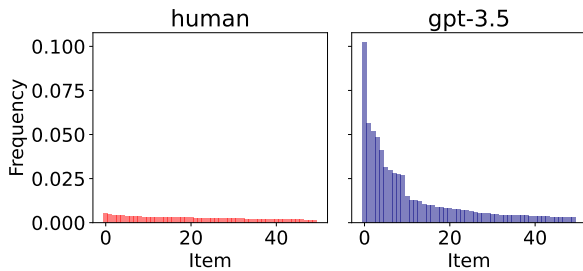


Figure 2: Distribution of mentioned items (Reddit+IH). Items are sorted in descending frequency. Humans mention more diverse items (left) than simulators (right).

in contrast to a more even distribution of items mentioned by humans (Figure 2). We observe the trend across all baselines and datasets, quantified by entropy (Table 2).

**Finding 2: Prompting with interaction history enhances item diversity.** Comfortingly, prompting with interaction history yields much higher diversity than prompting with demographic information, and we even observe cases (gpt-4 and text-davinci-003) prompted with interaction history from Reddit and IMDB slightly exceed the diversity of humans in ReDial. This suggests that interaction history (‘trigger’ items) is a strong condition for generating diverse items.

**Finding 3: Simulators may poorly represent real user preferences.** Our next task, *BinPref*, captures simulators failing to represent human preferences. In Figure 3, we sort items in decreasing average rating and plot the positive rate (proportion of simulators that answered ‘yes’ to whether they liked the movie). Positive rates remain constant regardless of human preferences, except gpt-4 + DI + PP, where the positive rate decreases as average rating decreases. Unexpectedly, higher item frequency (how well known is an item, measured by number

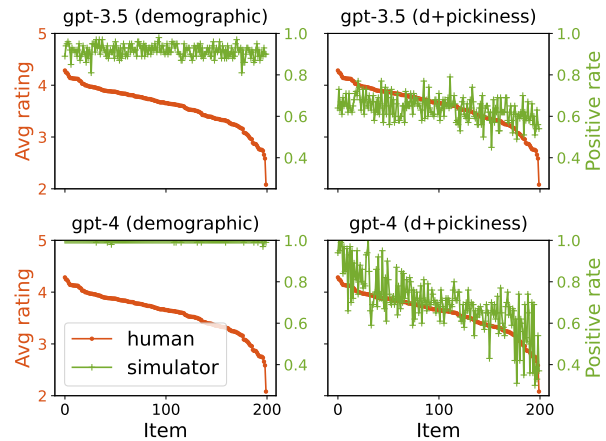


Figure 3: How well do simulators reflect human preferences? Most fail, except gpt-4 with pickiness (bottom right). The units for ratings and positive rates are different but included in the same plot to compare trends.

Table 3: Correlation coefficient between human and simulator preferences. Higher correlation is better, showing the effect of providing pickiness personality. ‘Undefined’ indicate undefined correlation; all simulators responded ‘yes’. P-values are less than 0.05.

Generator	Frequent items	Infrequent items
Demographic information		
gpt-3.5	0.18	0.12
gpt-4	0.24	0.53
text-davinci	Undefined	0.29
Demographic information + Pickiness		
gpt-3.5	0.45	0.36
gpt-4	0.75	0.76
text-davinci	0.49	0.64

of ratings) does not necessarily lead to better preference alignment (Table 3), despite LLMs likely being more exposed to these items during training. We show all the results in A.3.2.

**Finding 4: Adding pickiness personality improves preference alignment.** Endowing simulators with varying levels of pickiness not only diversifies preferences but also improves correlation, sometimes yielding strong correlation (Table 3). This suggests that picky simulators can successfully discern low-rated movies. Without pickiness, correlations are low to moderate: simulators tend to be consistently optimistic—in one case (text-davinci-003 + DI), all the answers were ‘yes’.

**Finding 5: Simulators express preferences differently from real users. Model choice and prompting may mitigate the difference.** *OpenPref* reveals how simulators express preferences in

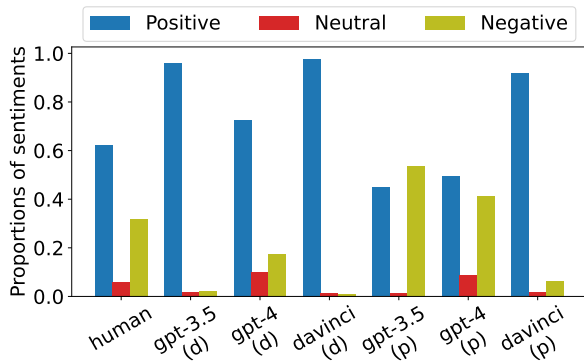


Figure 4: Sentiments in open-ended responses.

Table 4: Aspects and sentiments in open-ended responses. Humans have low number of aspects but high aspect entropy and sentiment entropy.

Generator	# aspects	Aspect entropy	Sentiment entropy
Human	85	<b>5.85</b>	1.19
Demographic information			
gpt-3.5	71	4.86	0.29
gpt-4	97	5.57	1.11
text-davinci	194	5.63	0.18
Demographic information + Pickiness			
gpt-3.5	101	5.20	1.09
gpt-4	97	5.59	<b>1.34</b>
text-davinci	<b>232</b>	5.47	0.48

a way different from humans (Table 4). First, simulators generate more sentiment-associated aspects than humans. Real users often express opinions in subtler ways, such as a movie being ‘suitable for background noise,’ rather than simply praising or criticizing explicit aspects (e.g., cast and plot) of the movie. Second, simulators have lower *aspect entropy*, even though they have more aspects. This indicates that it is predictable which aspects they will mention, e.g., mentioning the same aspects repeatedly. Finally, simulators are biased towards positive sentiment, resulting in low sentiment entropy, unless prompted to behave as picky users (Table 4 and Figure 4). The simulator closest to humans is gpt-4 + DI + PP, with similar aspect and sentiment statistics. Therefore, model choice and prompting strategies (e.g., adding pickiness) may enhance realism in simulator preference and expressions of preference.

**Finding 6: Simulators struggle to generate a diverse pool of personalized requests.** *RecRequest* reveals that simulators generate less personalized requests than real users. The request diversity, mea-

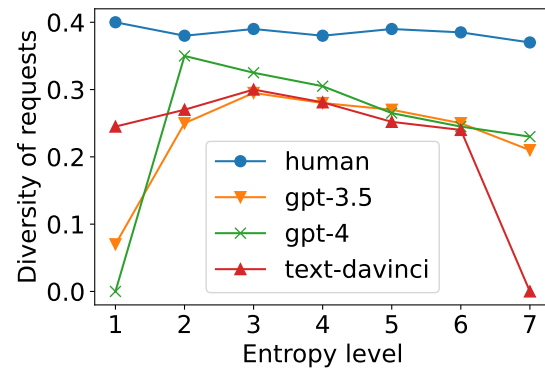


Figure 5: Diversity of requests per entropy level. Simulator requests are less diverse across all entropy levels.

- **Human:** ‘Movies about alcoholism’, ‘Space movies?’, ‘Movies about redemption’, ‘Inspirational movies’, ‘Good biography movies’, ‘Impactful endings?’, ‘Growth mindset versus Fixed Mindset’, ‘Rock climbing movies’, ‘Movies about nihilism’
- **gpt-3.5-turbo:** ‘Movie recommendation?’, ‘Need movie suggestions’, ‘Need movie recs!!’, ‘Need movie recommendations’, ‘Movie recs?’, ‘Movie recommendations?’
- **gpt-4:** ‘Got recs?’
- **text-davinci-003:** ‘Recommend a movie’, ‘Cheerful movies?’, ‘Recommend me!’

Figure 6: Low-entropy requests generated by humans and simulators.

sured by cosine diversity of sentence embeddings, are lower than humans across all models, with gpt-4 (the most diverse) generating 23% less diverse requests than humans (Table 6). Simulators have lower diversity across all entropy levels (taking words as variables), particularly in the lowest entropy level (Figure 5). For better understanding, we show low-entropy requests in Figure 6, where we see simulators struggling to make specific requests, while humans are specific even when text is short.

We also measure the diversity of words and word embeddings (Table 6). Interestingly, simulators have lower *word diversity* and higher *word embedding diversity*: although simulators generate a semantically diverse range of vocabulary, they tend to *reuse the same words*. Upon manual inspection, we find that expressions such as ‘gripping,’ ‘mind-bending,’ ‘compelling,’ ‘keeps me on the edge of my seat’ are repeatedly used. These

Human requests	gpt-3.5-turbo requests
Movies showing <b>extreme loneliness or depression</b> . I have watched <b>Taxi Driver (1976)</b> and <b>Joker (2019)</b> and would like to see more similar movies showing loneliness or depression.	Looking for a <b>gripping psychological thriller</b> similar to <b>Taxi Driver (1976)</b> or <b>Joker (2019)</b> ? Seeking a movie that delves into the mind of <b>complex characters</b> ?
Movies about <b>conspiracies, lies, and finding the truth</b> like <b>Memento (2000)</b> and <b>Fight Club (1999)</b> ?. Especially ones that have <b>big plot twists</b> .	Looking for <b>mind-bending thrillers</b> like <b>Fight Club (1999)</b> and <b>Memento (2000)</b> . Any suggestions? Need <b>gripping plots</b> that <b>leave me questioning reality</b> !

Table 5: Examples of recommendation requests. Even when humans and simulators include the same movies (orange), simulators tend to produce more general requests (green), repeating same expressions for different requests.

Table 6: Diversity of requests: word diversity (type-token ratio) and embedding (cosine) diversities. Simulators reuse the same words across different requests, generating less personalized requests.

Generator	Word	Word emb.	Sentence emb.
Human	<b>0.65</b>	0.427	<b>0.391</b>
gpt-3.5	0.50	<b>0.433</b>	0.295
gpt-4	0.61	0.436	0.300
text-davinci	0.49	0.418	0.288

Table 7: Feedback coherence (proportion of coherent feedback). Simulators are often coherent, but there is room for improvement.

	Generator	Prop. coherent	Prop. neither
Items only	gpt-3.5	0.8264	0.0087
	gpt-4	0.9096	0.0120
	text-davinci	0.8108	0
Items + explain	gpt-3.5	0.8039	0.0049
	gpt-4	0.9047	0
	text-davinci	0.6567	0

expressions can be applied to a general range of movies. Humans, in contrast, tend to express finer-grained preferences, asking for movies that meet a specific criterion. We show in Table 5 how the requests differ, even when they contain the same set of movies. For example, many movies other than *Taxi Driver (1976)* or *Joker (2019)* can be ‘gripping psychological thrillers,’ but a more limited set of movies deal with ‘extreme loneliness or depression.’ Therefore, the low diversity of requests seems to arise from their *generality*; LLMs make generic requests, while humans are free to be more personal, and hence, diverse in the population level.<sup>7</sup>

**Finding 7: While simulators often give coherent feedback, there is room for improvement.** For

<sup>7</sup>We tried various prompts to generate more diverse requests, but there was no significant difference in the results.

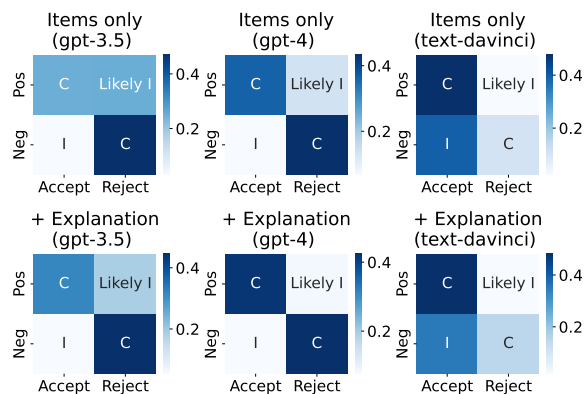


Figure 7: Feedback coherence (accept/reject task). ‘I’ stands for incoherent; ‘C’ stands for coherent. Recommendations contain only items (left column) or items with explanations (right column).

the accept/reject task, feedback is *coherent* when the simulator rejects negative or accepts positive recommendation. Feedback is *incoherent* when the simulator accepts negative recommendation. The case where the simulator rejects positive recommendation is controversial; a user may still reject a relevant recommendation for reasons external to the request. We leave this case as *likely incoherent* and focus evaluation on clearer cases. As in Figure 7, simulators are overall coherent, but sometimes give incoherent feedback, its proportion ranging from 3% (gpt-3.5-turbo) to 35% (text-davinci-003). Particularly, text-davinci-003 is biased towards optimistic feedback, i.e., accepting even if the recommendation is irrelevant. We also see that providing explanations encourages simulators accept relevant recommendation.

For the comparison task, feedback is coherent when the simulator prefers positive recommendation over negative recommendation, and incoherent otherwise. We exclude cases where simulators respond ‘neither,’ and compute its proportion. We see in Table 7 that simulators are more often coherent than incoherent, and the proportion of ‘neither’

- **Request:** I'm in a state of life rn that i really want/need movies with the Main Character being a loner or alone in general, he doesn't have to be alone the entire movie, but just portrayal of a good loner main character.
- **Recommendation:** Nightcrawler (2014)
- **Feedback from simulator:** Reject. Nightcrawler is not about a loner main character, but rather about a character who becomes involved in the underground world of crime journalism.

Figure 8: Example feedback from user simulator (gpt-3.5-turbo), given human request and recommendation.

is negligible. Variations exist; coherence ranges from 65% (text-davinci-003) to 90% (gpt-4). Interestingly, with explanations, simulators become slightly less coherent. A possible reason is that explanations add persuasiveness to negative recommendations as well as positive, making it slightly trickier for simulators to distinguish the two.

**Finding 8: Simulators may not capture subtle nuances in requests, and thus reject relevant recommendations.** Finally, we ask simulators *why* they have given such feedback. We manually inspect the responses and find that the reasons are rather compelling, based on accurate factual knowledge. Incoherence often arise due to missing subtle nuance of what users are asking for. For instance, in Figure 8, the simulator rejects *Nightcrawler (2014)* because the movie is not ‘about’ a loner main character, while the user asks for a movie ‘with’ a loner main character. More examples are in A.3.3. We leave a more thorough analysis of the classification of incoherent cases as future work.

## 5 Related Work

**User simulation for recommendation** Early work in conversational recommendation formulate bandit problems to efficiently update traditional models, focusing on item selection and expecting binary preference answers from synthetic users (Christakopoulou et al., 2016). Recent work considers more realistic conversations with flexibility in natural language, but still confines users to binary or multi-choice responses (Lei et al., 2020a,b). For evaluation, often a set of target items is predefined per user and the user rejects a recommendation un-

less the target item is mentioned, and a model is regarded superior if fewer turns are used to reach the target item (Guo et al., 2018; Sun and Zhang, 2018; Lei et al., 2020a,b). Alternatively, agenda-based simulation use a state diagram of actions, and recommendation is successful when the conversation reaches the ‘complete’ state (Zhang and Balog, 2020; Zhang et al., 2022). However, user actions follow a fixed set of rules and utterance templates, which is unlikely with real users. Generative simulators, powered by LLMs, effectively avoids this problem, demonstrating more realistic conversation capabilities (Wang et al., 2023a,b; Zhang et al., 2023). However, it is still uncertain how realistic these simulators are compared to real users.

**LLMs as human proxies** There is increasing effort to substitute expensive human experiments with LLMs. In conversational recommendation, Wang et al. (2023b) prompts ChatGPT with target items, which results in users giving ‘hints’ toward specified items. Wang et al. (2023a) goes beyond conversation and introduces a simulation environment where users not only chat about recommendations, but also browse websites, search for items, and share opinions on social media. Non-verbal user behavior in recommendation is explored in Zhang et al. (2023). In other domains, Owoi-cho et al. (2023) and Wang et al. (2023c) explore LLMs as user simulators in conversational search; Hämäläinen et al. (2023) evaluates LLMs for generating synthetic user experience data in HCI; Aher et al. (2023) evaluates LLMs for replicating human behavior in social science experiments. Others create simulation environments where LLM agents interact with each other and generate realistic behaviors (Park et al., 2023; Gao et al., 2023; Qian et al., 2023). No work, to the best of our knowledge, introduces protocols for synthetic users in conversational recommendation.

**LLMs for recommendation** Recent papers explore the use LLMs for recommendation (Hou et al., 2023; Li et al., 2023; Kang et al., 2023; Fan et al., 2023; Chen, 2023), but these work focus on LLMs as *recommenders*, not recommendation *seekers*, and are therefore orthogonal to our work.

## 6 Conclusion

We introduce a new protocol for evaluating LLMs as user simulators for conversational recommendation. We design five evaluation tasks, where each



task addresses an essential property for simulators to be realistic user proxies. By running the tasks on simulators, we show how the tasks effectively reveal discrepancies of simulators from real users. Our work aims to set a benchmark for evaluating simulators automatically, with future plans to enhance their realism.

## Limitations

Our tasks provide necessary conditions, not sufficient conditions, for simulators to represent a group of real users. More tasks could be added to evaluate more properties, such as asking questions about recommendations, or dealing with evolving items (that are not in the training corpus of LLMs).

While our approach is domain-agnostic, the datasets used in this paper are limited to movies. Different domains (e.g., e-commerce) may require domain-specific tasks and may produce different results. An important avenue for future work is to collect CRS datasets in various domains—most existing datasets are on movies or media content.

Our observations on baseline simulators may not represent all possible simulators. In particular, we use OpenAI models, default temperature values, and simple prompt-based baselines. More analysis could be made with open-source models, various hyperparameters, and advanced simulators.

## Ethics Statement

Our paper is primarily centered on CRS applications, but has broader implications in making artificial intelligence agents more closely aligned to real humans. As simulators exhibit more realistic behavior, the risks of misuse, deception, and over-reliance on these simulators may arise. A possible way to mitigate these risks is to implement distinct markers on simulators. For instance, simulators should truthfully disclose that they are not humans. Finally, we stress that, while simulators are valuable tools for pre-deployment testing, they cannot fully replace human interactions. Ultimately, real user experiments are needed for final testing.

We use scientific artifacts. Redial is licensed under the CC BY 4.0 License. MovieLens states that the data may be used for research purposes under a set of conditions (e.g., citation), which this paper meets. We ask direct permission from the authors of the Reddit dataset. IMDB is available for non-commercial usage. PyABSA is under the MIT License. Word2Vec and SBERT are under the

Apache License, Version 2.0.

## Acknowledgements

This research was supported by a research grant from Cisco Systems, Inc.

## References

- Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *ICML*.
- Ashton Anderson, Lucas Maystre, Ian Anderson, Rishabh Mehrotra, and Mounia Lalmas. 2020. Algorithmic effects on the diversity of consumption on spotify. In *WWW*.
- Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.
- Zheng Chen. 2023. Palr: Personalization aware llms for recommendation. *arXiv preprint arXiv:2305.07622*.
- Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards conversational recommender systems. In *KDD*.
- Wenqi Fan, Zihuai Zhao, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Jiliang Tang, and Qing Li. 2023. Recommender systems in the era of large language models (llms). *arXiv preprint arXiv:2307.02046*.
- Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. 2023. S3: Social-network simulation system with large language model-empowered agents. *arXiv preprint arXiv:2307.14984*.
- Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. 2021. Advances and challenges in conversational recommender systems: A survey. *AI Open*.
- Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesaro, and Rogerio Feris. 2018. Dialog-based interactive image retrieval. *NeurIPS*.
- Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating large language models in generating synthetic hci research data: a case study. In *CHI*.
- F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. *ACM TITS*.
- 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 *CIKM*.

- 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, 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*.
- Wenqiang Lei, Xiangnan He, Yisong Miao, Qingyun Wu, Richang Hong, Min-Yen Kan, and Tat-Seng Chua. 2020a. Estimation-action-reflection: Towards deep interaction between conversational and recommender systems. In *WSDM*.
- Wenqiang Lei, Gangyi Zhang, Xiangnan He, Yisong Miao, Xiang Wang, Liang Chen, and Tat-Seng Chua. 2020b. Interactive path reasoning on graph for conversational recommendation. In *KDD*.
- 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 preprint arXiv:2304.03879*.
- Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. In *NeurIPS*.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Ida Momennejad, Hosein Hasanbeig, Felipe Vieira, Hiteshi Sharma, Robert Osazuwa Ness, Nebojsa Jojic, Hamid Palangi, and Jonathan Larson. 2023. Evaluating cognitive maps and planning in large language models with cogeval. *arXiv preprint arXiv:2309.15129*.
- Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *ACL*.
- OpenAI. 2021. [About openai](#).
- Paul Owoicho, Ivan Sekulic, Mohammad Aliannejadi, Jeffrey Dalton, and Fabio Crestani. 2023. Exploiting simulated user feedback for conversational search: Ranking, rewriting, and beyond. In *SIGIR*.
- Wenbo Pan, Qiguang Chen, Xiao Xu, Wanxiang Che, and Libo Qin. 2023. A preliminary evaluation of chatgpt for zero-shot dialogue understanding. *arXiv preprint arXiv:2304.04256*.
- Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *UIST*.
- Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software development. *arXiv preprint arXiv:2307.07924*.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *EMNLP*.
- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cino Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *ICML*.
- Yueming Sun and Yi Zhang. 2018. Conversational recommender system. In *SIGIR*.
- Lei Wang, Jingsen Zhang, Xu Chen, Yankai Lin, Ruihua Song, Wayne Xin Zhao, and Ji-Rong Wen. 2023a. Recagent: A novel simulation paradigm for recommender systems. *arXiv preprint arXiv:2306.02552*.
- Xiaolei Wang, Xinyu Tang, Wayne Xin Zhao, Jingyuan Wang, and Ji-Rong Wen. 2023b. Rethinking the evaluation for conversational recommendation in the era of large language models. In *EMNLP*.
- Zhenduo Wang, Zhichao Xu, Qingyao Ai, and Vivek Srikumar. 2023c. An in-depth investigation of user response simulation for conversational search. *arXiv preprint arXiv:2304.07944*.
- Yu Xia, Junda Wu, Tong Yu, Sungchul Kim, Ryan A Rossi, and Shuai Li. 2023. User-regulation deconfounded conversational recommender system with bandit feedback. In *KDD*.
- Heng Yang, Chen Zhang, and Ke Li. 2023. Pyabsa: A modularized framework for reproducible aspect-based sentiment analysis. In *CIKM*.
- Junjie Zhang, Yupeng Hou, Ruobing Xie, Wenqi Sun, Julian McAuley, Wayne Xin Zhao, Leyu Lin, and Ji-Rong Wen. 2023. Agentcf: Collaborative learning with autonomous language agents for recommender systems. *arXiv preprint arXiv:2310.09233*.
- Shuo Zhang and Krisztian Balog. 2020. Evaluating conversational recommender systems via user simulation. In *KDD*.
- Shuo Zhang, Mu-Chun Wang, and Krisztian Balog. 2022. Analyzing and simulating user utterance reformulation in conversational recommender systems. In *SIGIR*.
- Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. 2023. Is chatgpt equipped with emotional dialogue capabilities? *arXiv preprint arXiv:2304.09582*.

## A Appendix

### A.1 Dataset statistics

The **ReDial** dataset contains 11,348 conversations and 6,925 movies. The dataset was collected until 2018, and hence, movies up to this year are mentioned. 1,309 movies are used for *ItemsTalk*.

The original **Reddit** dataset contains 634,392 conversations, 1,669,720 turns, 36,247 users, and 5,1203 movies. We process this dataset in the following way: remove posts after 2021, remove comments without movie mentions, remove requests that are not about movies, and sample one head comment for each request. The resulting dataset has 23,167 requests, each with one comment. 9,974 movies are used for *ItemsTalk*.

The **MovieLens** dataset contains ratings of 62,000 movies rated by 162,000 users. We sample 200 movies rated by more than 5000 users (frequent movies) and 200 movies rated by less than 500 and more than 50 users (infrequent movies). While we also sample 300 movies without considering rating frequency, we observe that due to the long-tail distribution of frequencies, this random sampling result in movies with significantly lower frequencies (median: 5, mode: 1 appears 49/300 times). We nonetheless show the results.

The **IMDB** dataset originally consists of 1,083 users and 22,918 reviews. Each of 928 users has at least 11 movie reviews. 8,138 movies are used for *ItemsTalk*.

### A.2 Prompts

Here we provide the full prompts for all the tasks.

#### A.2.1 ItemsTalk

For the demographic information (DI) simulator, we randomly sample a prefix and a surname and ask to generate the movies one wants to talk about as this person.

Pretend to be {prefix} {surname}.  
You decide to talk about {target\_num} movies. What would these {target\_num} movies be? Reply as a list of <Title (yyyy)>. Say nothing else.

The interaction history (IH) simulator has slight variations according to the dataset where the interaction histories are sampled.

IMDB: 10 movies and the corresponding review titles from each user.

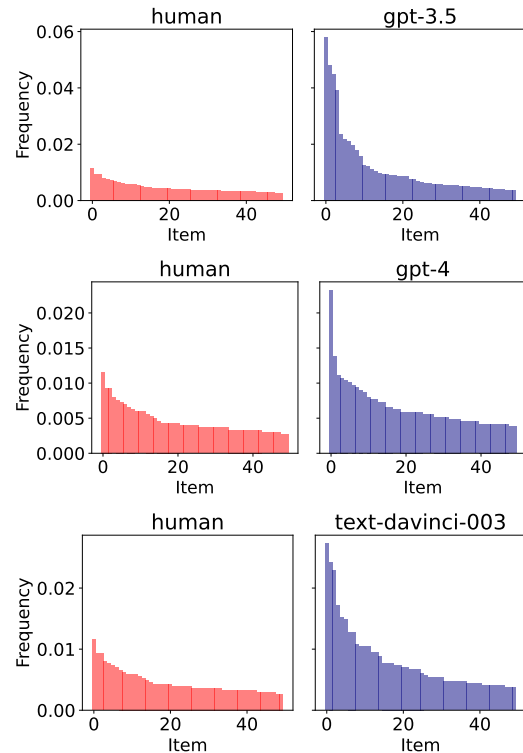


Figure 9: ItemsTalk results for ReDial. Simulators are prompted with interaction history per ReDial conversation.

A person leaves the following remarks on movies...

{movie 1}: {review title 1}

...

{movie N}: {review title N}

and proceeds to talk about {target\_num} more movies. What would these {target\_num} movies be? Reply as a list of <Title (yyyy)>. Say nothing else.

Reddit: one movie and the UTC time it was mentioned, from each request.

At UTC time {time}, a person starts to talk about the movies {movies} and proceeds to talk about {target\_num} more. What would these {target\_num} movies be? Reply as a list of <Title (yyyy)>. Say nothing else.

ReDial: 2 movies from the seeker side of each conversation.

A person mentions {movies} in a conversation about movies and proceeds to mention {target\_num} more. What would these {target\_num} movies be? Reply

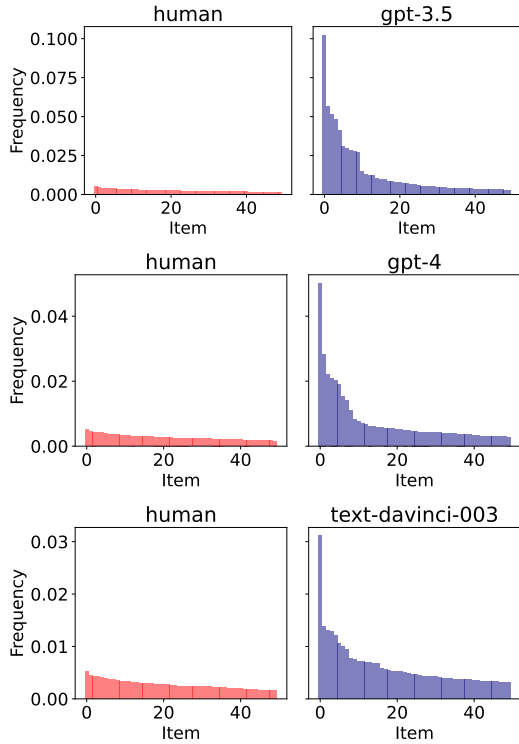


Figure 10: ItemsTalk results for Reddit. Simulators are prompted with interaction history per Reddit request.

as a list of <Title (yyyy)>. Say nothing else.

Target number of items is the number of total movies in the entry, minus the number of movies used as interaction history.

### A.2.2 BinPref

We randomly sample a prefix and a surname for the demographic information (DI) simulator.

Pretend to be {prefix} {surname}. You watched the movie {movie}. Did you like the movie? Answer Yes or No. Don't say anything else.

We add a pickiness trait (DI + PP) by randomly selecting among the three levels of pickiness: extremely picky, moderately picky, and not picky.

Pretend to be {prefix} {surname}. You are {pickiness} about movies. You watched the movie {movie}. Did you like the movie? Answer Yes or No. Don't say anything else.

### A.2.3 OpenPref

Similar to BinPref, we prompt the DI simulator

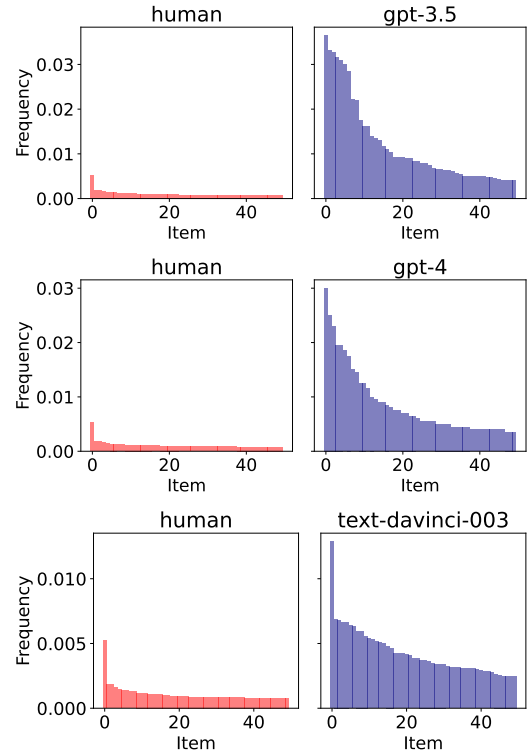


Figure 11: ItemsTalk results for IMDB. Simulators are prompted with interaction history per IMDB user.

Pretend to be {prefix} {surname}. You watched the movie {movie}. What are your thoughts on this movie? Answer should not exceed {review\_len} characters.

and the DI + PP simulator

Pretend to be {prefix} {surname}. You are {pickiness} about movies. You watched the movie {movie}. What are your thoughts on this movie? Answer should not exceed {review\_len} characters.

Target review length is determined by the review length in the processed IMDB baseline, where one review is sampled per user.

### A.2.4 RecRequest

From each Reddit request, we use the movie names mentioned and the length of the request to prompt the following.

Generate a movie recommendation request. Include (but do not request) the following movies in your text: {movies}. Make sure the length of the request is approximately {target\_len} characters.

Table 8: Correlation coefficient between human and simulator preferences. Movies are randomly sampled across all frequency levels, resulting in a sample with movies with very low frequencies.

Demographic information	
gpt-3.5	0.27
gpt-4	0.29
Demographic information + Pickiness	
gpt-3.5	0.32
gpt-4	0.44

### A.2.5 Feedback

For the accept/reject task, we sample one positive and one negative response (recommendation) per request and use the following prompt for each pair.

In the following conversation, a USER asks for movie recommendations. Your task is to act like the USER by giving the following responses to the AGENT’s recommendation: If the recommendation is coherent to your request, answer Accept. If the recommendation is incoherent to your request, answer Reject. Simply answer Accept or Reject.

USER: {request}

AGENT: {response}

USER (answer Accept or Reject):

For the comparison task, we use both the positive and negative responses of a request in a single prompt. To eliminate position bias, we randomly choose assign the responses to the AGENTs. That is, the positive response is assigned to either AGENT 1 or AGENT 2 with equal probability.

A USER asks for movie recommendation. AGENT 1 and AGENT 2 gives recommendations. Your task is to choose the AGENT that gives better recommendations. Simply answer AGENT 1 or AGENT 2. You HAVE to choose one.

USER: {request}

AGENT 1’s response: {response}

AGENT 2’s response: {response}

Which response is better?

(Reply AGENT 1 or AGENT 2)

For cases (20 cases for each configuration) where we ask to provide reasons, we add the following to the prompt.

Provide a short reason (less than 40 words) for your response.

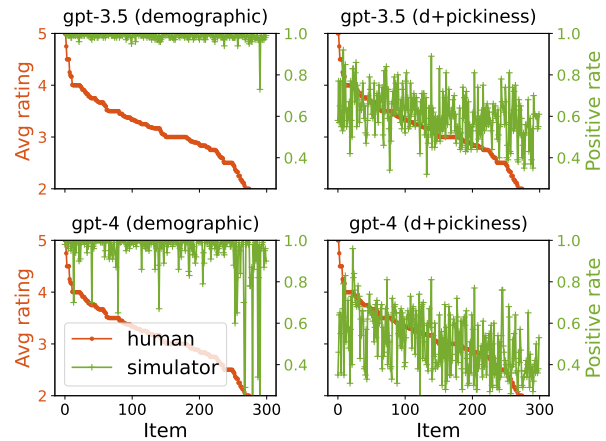


Figure 12: Preference trends of humans and simulators. See Table 8 for correlation coefficients.

## A.3 More results

### A.3.1 Results from ItemsTalk

We compare the distributions of items mentioned by simulators and real users. We plot human and simulator distributions side-by-side per dataset: R-Dial (Figure 9), Reddit (Figure 10), and IMDB (Figure 11). Note that the three human distributions within each figure is the same, but scaled differently according to the scale of different language models.

### A.3.2 Results from BinPref

We compare the trends of average rating and positive rate (proportion of simulators that answered that they liked the movie). Results for items randomly sampled regardless of frequency levels are shown in Table 8 and Figure 12.<sup>8</sup> However, these movies often have very low frequency values (median: 5, mode: 1 appears 49/300 times); the average ratings of these movies may not reflect true user preferences. Plots for frequent items are shown in Figure 13 and infrequent items in Figure 14.

### A.3.3 More Feedback examples

We provide more examples generated by simulators, when asked to provide a reason for their feedback. All examples are from gpt-3.5-turbo.

In our first example, The human user asks for movies, and possibly even shows. However, the simulator rejects the relevant recommendation, saying that it is a show not a movie:

<sup>8</sup>These experiments were conducted after we received our paper reviews. OpenAI’s configurations might have been updated or modified in the meantime. We did not obtain results for text-davinci-003, since by the time we ran these experiments, the model was deprecated.

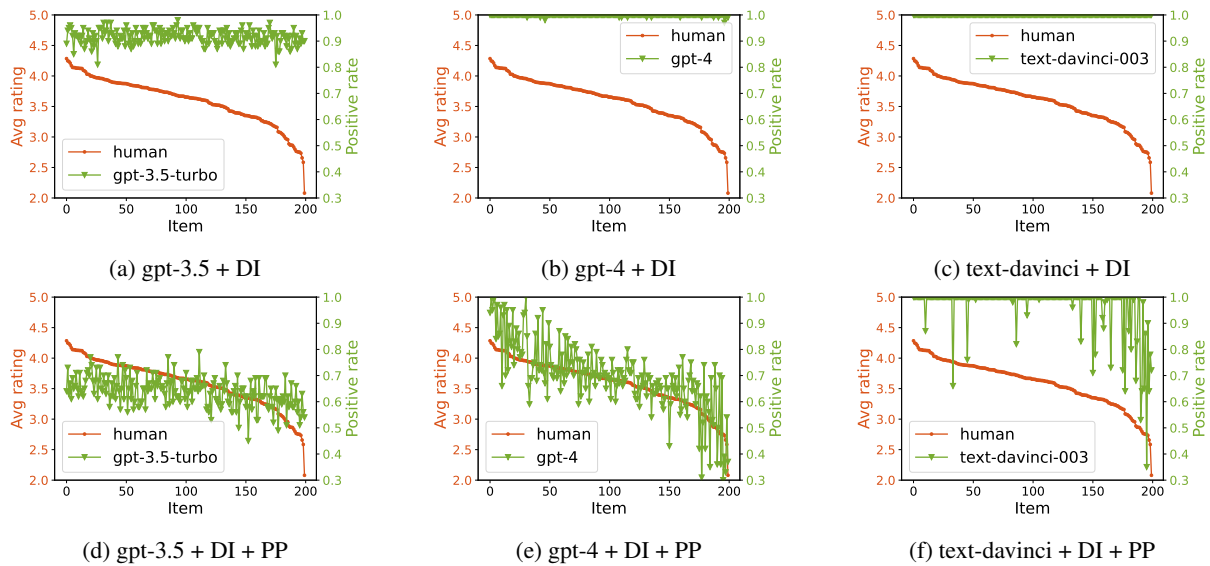


Figure 13: Preference trends of frequent movies

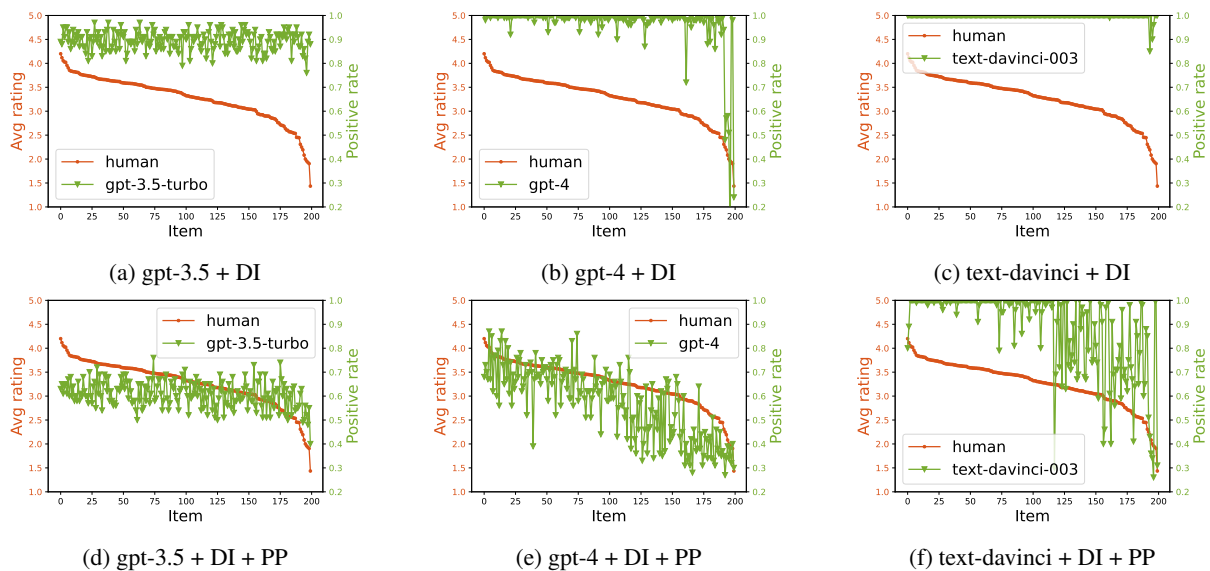


Figure 14: Preference trends of infrequent movies

- **Request:** Looking for a Murder Mystery Serial Killer movie. Recently watched the Bone Collector, Taking Lives and Copy Cat, if you have any recommendations, (even ones that were just fine, or even sub-par) I'd love to see them! Or even a show like Criminal Minds, where each episode is it's own little movie.
- **Positive Rec:** Not movie, but show called : La Mante
- **Feedback:** Reject. La Mante is recommended as a show, not a movie.

In another example, a list of relevant movies is recommended, while including the one that the

human user has already watched. The simulator rejects the recommendation just because it includes this movie. (We alter certain words that could be perceived as offensive, although their original usage was intended positively.)

- **Request:** What are the most disturbing up movies on UK Netflix right now!. In the past there have been some great films on UK Netflix to mess with you, (antichrist, I saw the devil, the original oldboy, Ichi the killer, the human centipede, requiem for a dream etc.) But it seems there aren't any good ones to find that I haven't seen and that are still streaming! Most of these have been taken off now

and I'm after something new but along similar lines... does anyone have any suggestions? Thanks in advance!

- **Positive Rec:** It Follows (2014), Victoria (2015), Antichrist (2009)
- **Feedback:** Reject. Reason: Already mentioned and seen Antichrist in the past. Looking for something new but along similar lines.

Finally, a user wants to be cheered up, wanting to be inspired for a better future. The simulator understood the mission, but chooses the negative recommendation (*Love (2005)*, a romance movie) over the positive one (*Trolls (2016)*, an animation about happiness). The reason is unclear.

- **Request:** Movie about positively changing in the future. Hi, I got into some trouble recently with the law and in general have been reckless in my drug use. I fear the social and legal repercussions. I'd like to watch something that reminds me good times can still be had in the future. I'd just like something to cheer me up, no comedies please.
- **Positive Rec:** Trolls (2016)
- **Negative Rec:** Love (2005)
- **Feedback:** Love (2005). Love (2005) seems like a better recommendation as it may provide the user with a positive outlook on the future and remind them that good times can still be experienced even in difficult situations.