

Data, methods, and evaluation for knowledge-grounded conversational recommendation systems

Julian McAuley

Conversational recommender systems

We'll explore:

- How can we improve conversational recommender systems by creating new **datasets**?
- How can we combine modify language models to increase **controllability** with respect to recommended items?
- How can we better **evaluate** conversational recommender systems, especially with respect to beyond-accuracy metrics?

What is conversational recommendation?

Some traditional approaches...

Traditional approaches rarely involved “conversation” as we might normally think of it:

- Thompson et al., 2004 (query refinement): Elicits users’ preferences and constraints with regard to item attributes;
- Mahmood and Ricci, 2009 (reinforcement learning): Queries users about recommendation attributes during each round; learns a policy to choose queries to efficiently yield a desirable recommendation

User Name	Homer						
Attributes	w_i	Values and probabilities					
Cuisine	0.4	Italian	French	Turkish	Chinese	German	English
		0.35	0.2	0.25	0.1	0.1	0.0
Price Range	0.2	one	two	three	four	five	
		0.2	0.3	0.3	0.1	0.1	
...					
Parking	0.1	Valet		Street		Lot	
		0.5		0.4		0.1	
Item Nbr.	0815	5372	7638	...		6399	
Accept/Present	23 / 25	10 / 19	33 / 36	...		12 / 23	

(from Thompson et al.)

Some traditional approaches...

Traditional approaches rarely involved “conversation” as we might normally think of it:

- Christakopoulou et al., 2016 (iterative recommendation): Collects feedback about recommended items in order to iteratively learn user preferences; explores various query strategies to elicit preferences quickly

Greedy: $j^* = \arg \max_j y_{ij}$

A trivial *exploit*-only strategy: Select the item with highest estimated affinity mean.

Random: $j^* = \text{random}(1, N)$

A trivial *explore*-only strategy.

Maximum Variance (MV): $j^* = \arg \max_j \epsilon_{ij}$

A *explore*-only strategy, variance reduction strategy: Select the item with the highest noisy affinity variance.

Maximum Item Trait (MaxT): $j^* = \arg \max_j \|\mathbf{v}_j\|_2$

Select the item whose trait vector \mathbf{v}_j contains the most information, namely has highest L2 norm $\|\mathbf{v}_j\|_2 = \sqrt{v_{j1}^2 + v_{j2}^2 + \dots + v_{jd}^2}$.

Minimum Item Trait (MinT): $j^* = \arg \min_j \|\mathbf{v}_j\|_2$

Select the item with trait vector with least information.

Upper Confidence (UCB): $j^* = \arg \max_j y_{ij} + \epsilon_{ij}$

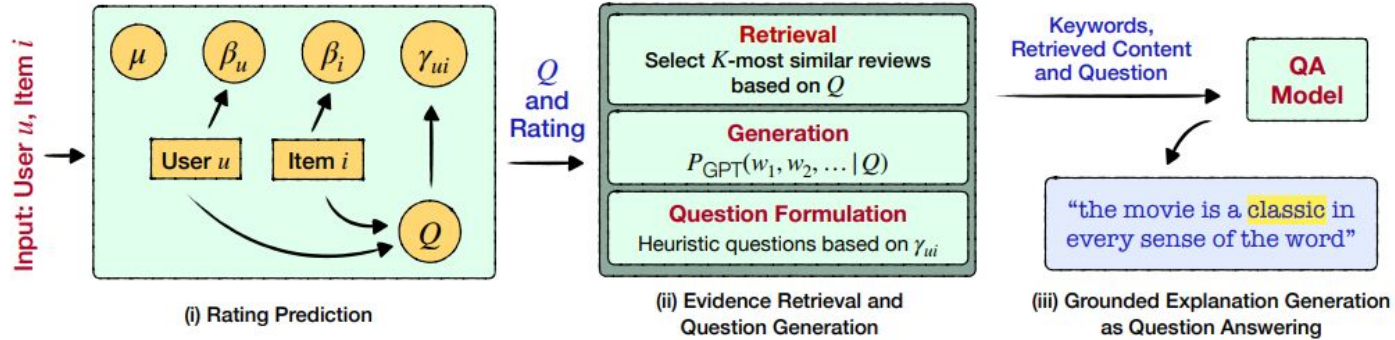
Based on UCB1 [3]: Pick the item with the highest upper confidence bound, namely mean plus variance (95% CI)

Thompson Sampling (TS) [5]: $j^* = \arg \max_j \hat{y}_{ij}$

For each item, sample the noisy affinity from the posterior. Select item with the maximum sampled value.

(from Christakopoulou et al.)

Related: “explainable” recommendations



(from Xie et al., 2022)

Explainable recommenders associate natural language explanations with each recommendation (or something like this)

Such models represent “half” of a conversational model, though lack interactive mechanisms for the user to participate in conversation

Limitations of traditional approaches

- Well short of what we'd normally call “**conversation**”
- Models may be interactive but probably couldn't be implemented via conversational interfaces
- A lack of compelling **datasets** to train “real” conversational recommenders
- Most methods not particularly “**general purpose**”: typically designed around domains or datasets with specific data/features available

Actual conversation...

Li et al. (2018) sought approaches more closely matching “free-form” conversation. Roughly:

- Dialogs (around 10k) are constructed by crowd workers, who assume roles of a *recommender* or *seeker*;
- Conversations between the recommender and the seeker are tagged in terms of the movies mentioned, as well as explicit feedback (has the seeker seen the movies mentioned and did they like them);
- Train a dialog generation model that can fulfil the role of the recommender;
- Preferences can then be estimated and the output controlled to reference specific movies

Table 5: Sample conversation from validation set

SEEKER:	hi !
HUMAN:	hi ! how is your morning going ?
HRED:	hi , what kind of movies do you like
OURS:	hello ! what kind of movie are you looking for ?
SEEKER:	great , thanks ! i was thinking of starting the day off with a great comedy , any suggestions ?
HUMAN:	i love the hangover it always makes me laugh . and i bet lots of people are hungover this morning ! airplane ! is another classic of mine .
HRED:	that 's a good one ! have you seen the lego movie ? it 's a great movie !
OURS:	i have n't seen that one ! what kind of movies do you like ? i can help you with that .
SEEKER:	i 've seen the hangover it 's really good ! i thought about watching knocked up have you seen it ?
HUMAN:	i enjoyed that movie . i enjoyed spy it 's a well done parody .
HRED:	it 's about a man who gets revenge on a cyborg and a hockey team . it 's really good . it 's really good . i also like the girl on the train
OURS:	i have n't seen that one . i will have to check it out . i hope i was able to help you find a good one ! have a great day !
SEEKER:	thanks ! i 'll check it out !
HUMAN:	sure ! have a great day !
HRED:	you 're welcome ! have a great day !
OURS:	have you seen revenge of the nerds ? it 's really good . it 's a classic .
SEEKER:	you too .

(from ReDial)

Actual conversation...

Li et al. (2018)'s approach has a number of virtues:

- Actually looks (more or less) like “real” conversation, especially compared to approaches that came before
- Contributes a (now widely used) benchmark dataset for training and evaluation
- Elegant / principled in terms of how the model is trained and the objective it's trained for (i.e., reach a goal movie in the fewest possible number of steps)

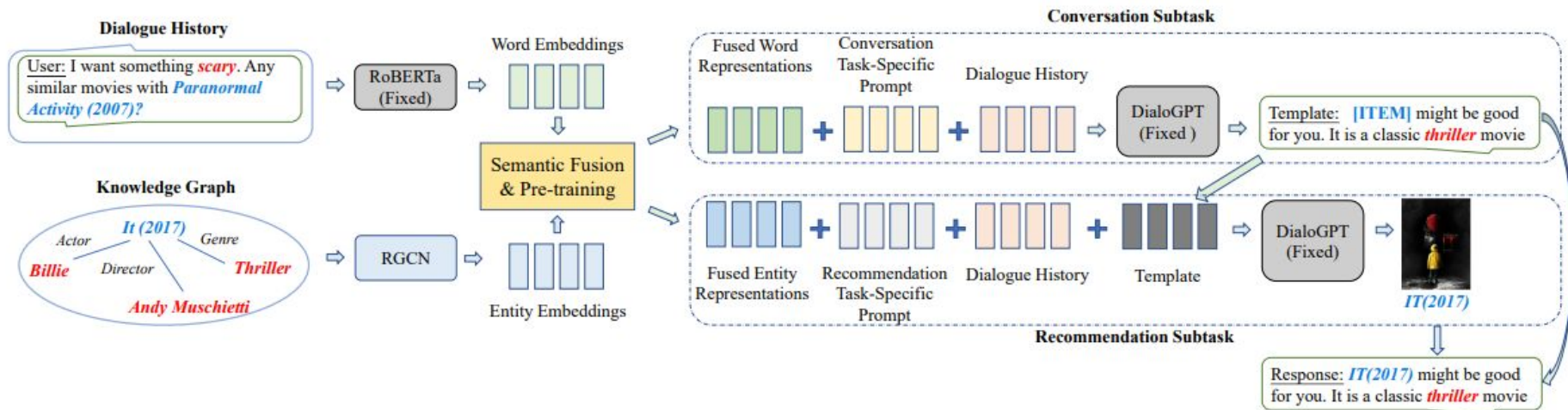
Actual conversation...

Though it also has some **limitations**

- Conversations aren't particularly "real": the users aren't actually seeking some item, but play a synthetic game in which they are told which item to seek
- It's unclear to what extent the data collection effort could be applied in other settings, in particular ones not based on "general knowledge" (i.e., for which crowd workers would struggle to engage in synthetic conversations)
- Even within movies, it's hard to tell how closely conversations in ReDial (or similar efforts) represent "organic" conversations

“LM+RecSys” approaches (UniCRS; Wang et al., 2022)

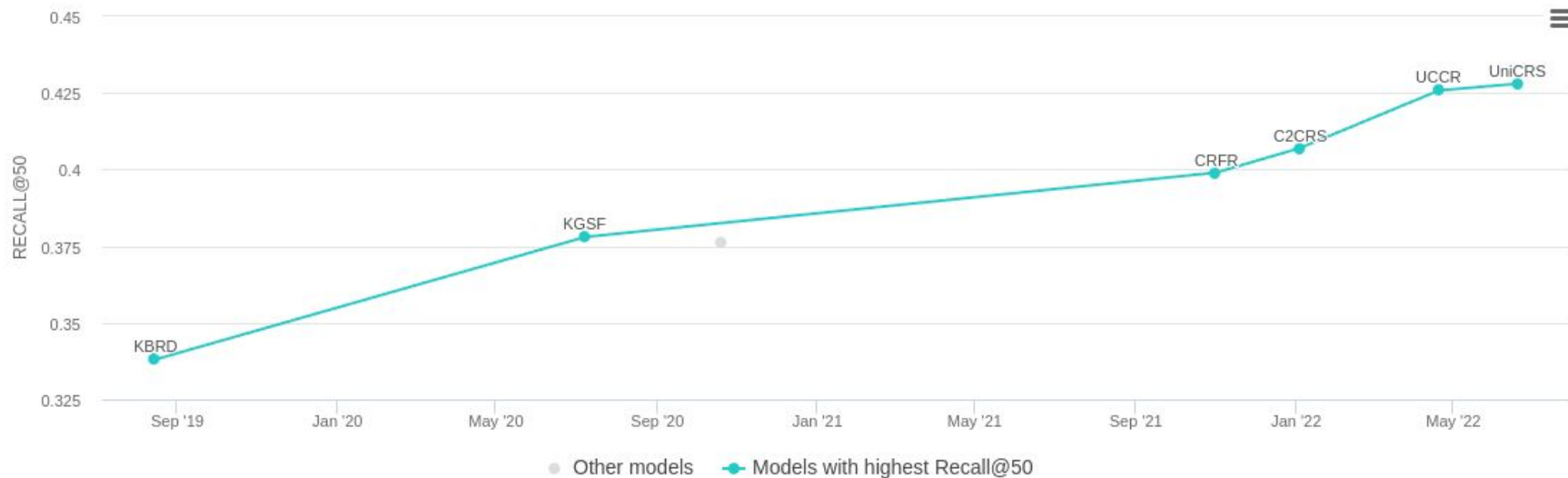
(Fairly) recent attempts incorporate knowledge grounding, and arguably (among a few others) represented the pre-LLM state-of-the-art



(UniCRS)

pre-LLM state-of-the-art (ReDial benchmark)

Lots of progress on the ReDial benchmark in the last few years:

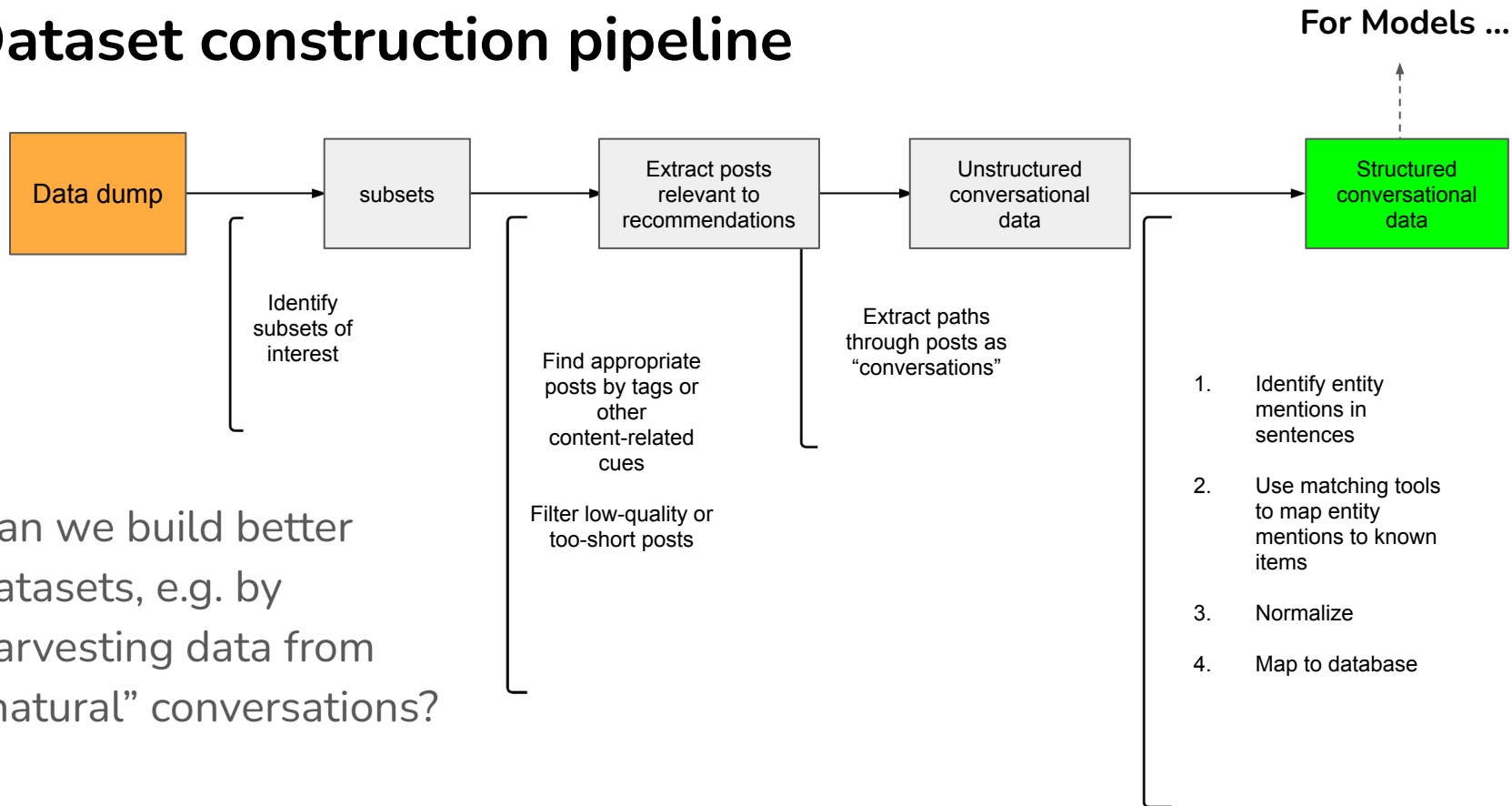


Datasets

Datasets

- How can datasets be built that are more *natural*? E.g. actually how humans would interact when making movie recommendations, versus current, more synthetic, settings?
- Other efforts (e.g. INSPIRED) aim for a more natural setting, but are also very small
- Need datasets that are **bigger** and **more realistic**
- Our previous efforts (e.g. to synthesize conversational datasets from product review text) were much larger but of low quality

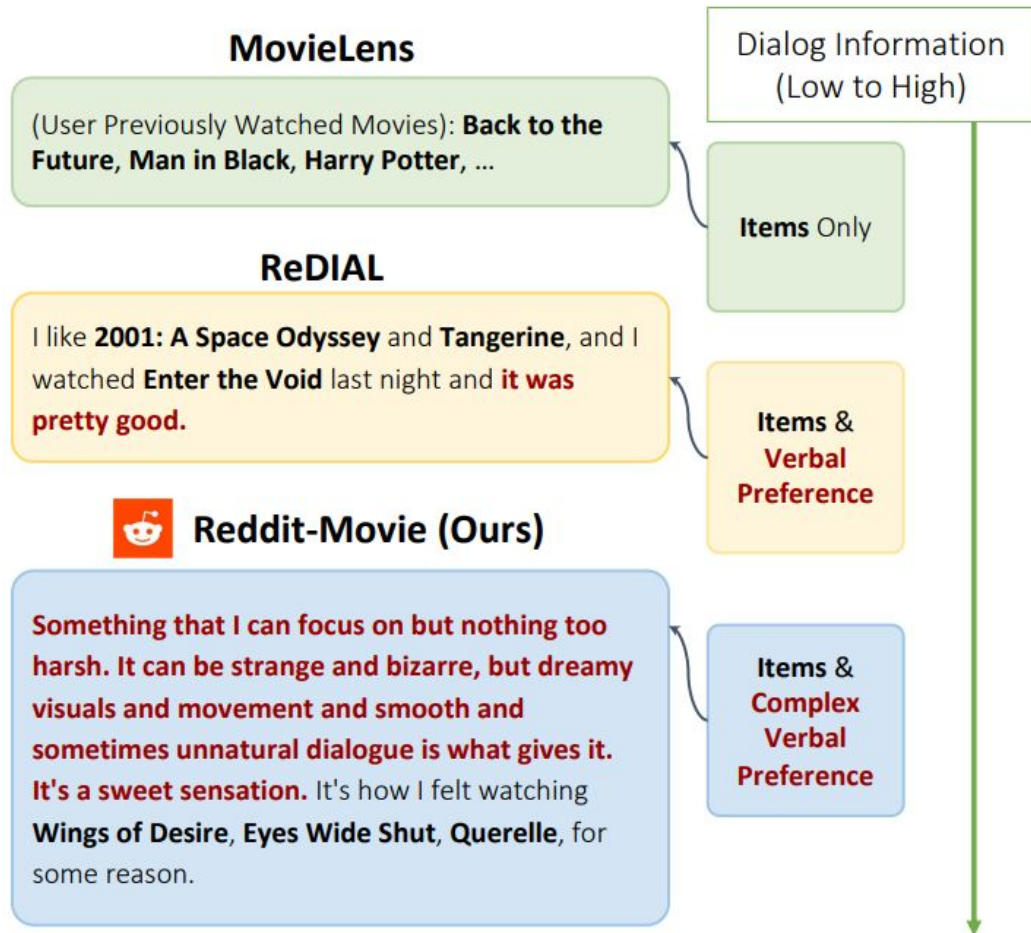
Dataset construction pipeline



Reddit-Movie Dataset

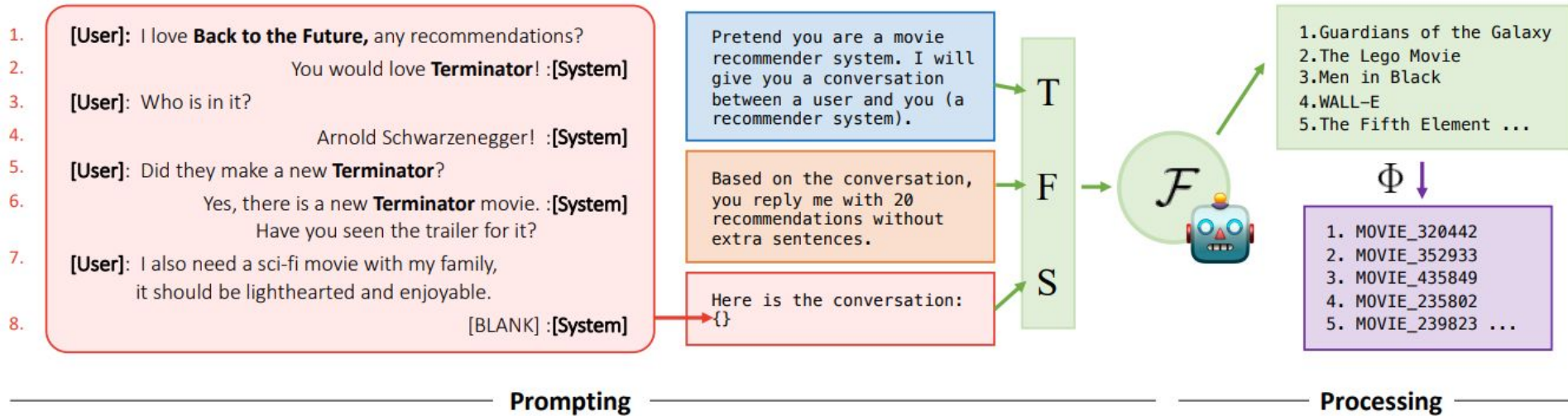
- 634,392 movie recommendation conversations, featuring 1.7M dialog turns
- ~11k users, ~24k items
- (compare to e.g. ReDial, featuring ~10k conversations, ~139k turns, ~800 users)

Much bigger than existing datasets; conversations are shorter; they have much more *context*; and (for better or worse) have much more varying structure



What do these new datasets reveal?

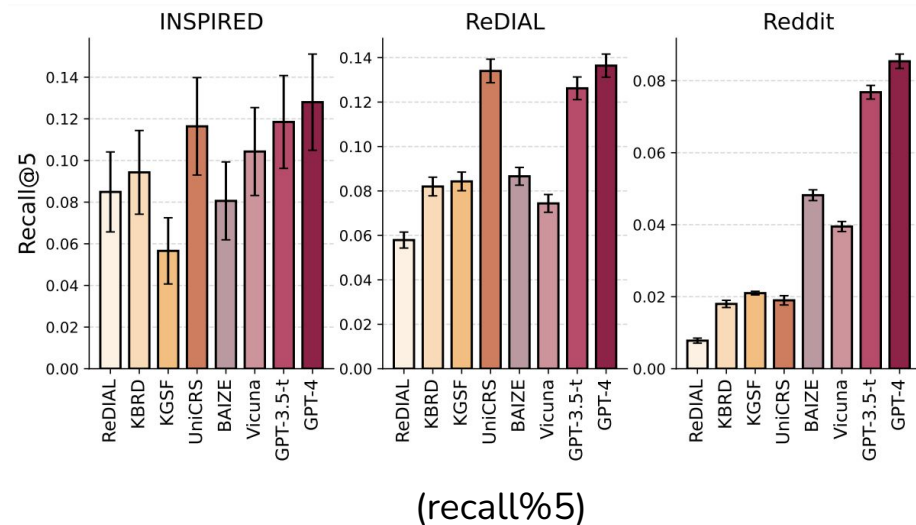
We use a simple prompting setup to compare LLMs:



What do these new datasets reveal?

Some observations about model performance:

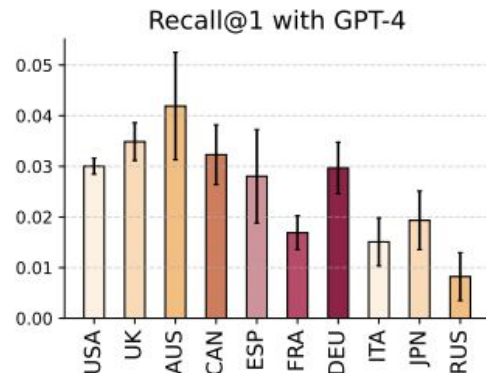
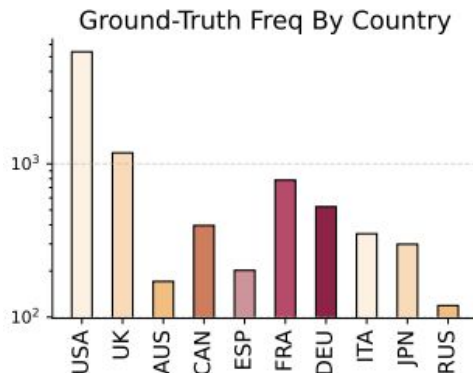
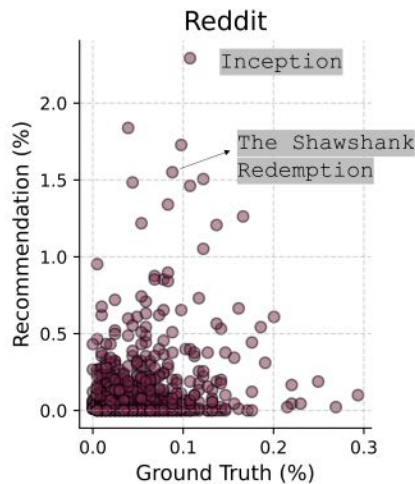
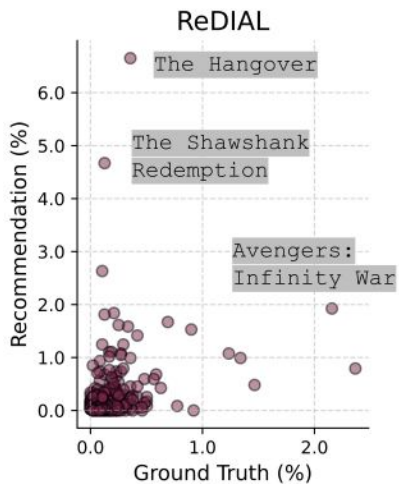
- Existing models engage in *shortcut learning* by focusing on repeated items (i.e., items already mentioned in a dialog but not as recommendations)
- LLMs outperform existing fine-tuned models; GPT-4 outperforms other LLMs
- LLMs generate some out-of-dataset items, but not many hallucinated recommendations (<5%); can be dealt with by string matching



What do these new datasets reveal?

Some observations about model performance:

- Significant “popularity bias” (and other bias) issues
- Recommendation performance is highly sensitive to geographical region (presumably just due to groundtruth frequency)



Methods

Control strategies for LLM-based recommenders

- How can we make LLM-based recommenders more **controllable**

Solutions (roughly) based on RAG

1. Retrieve *recommendation probabilities* from a recommender; use these to change item probabilities during decoding
2. Retrieve *related items* to improve prompting-based methods
3. Retrieve *related training samples, or knowledge*

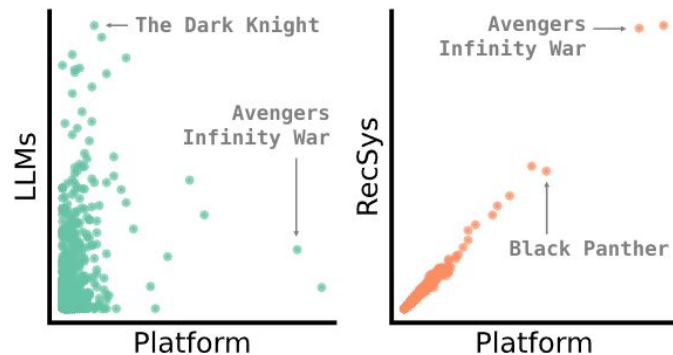


NETFLIX

1. Item “re-indexing”

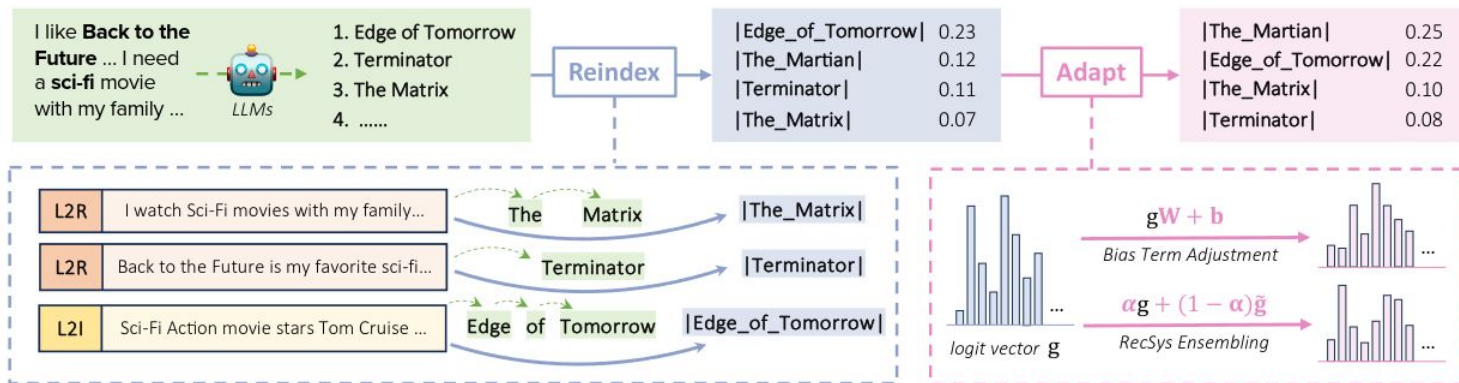
Zhankui He + others

- How can we address the issue of *distribution imbalance* with LLM-based recommenders?
- Essentially, distribution of recommended items should match the distribution we observed in the training set
- Lots of solutions for conventional methods, somewhat of an open problem for conversational approaches!



NETFLIX

1. Item “re-indexing”



- LLMs have poor controllability of recommended items since they surface *tokens* rather than surfacing *items*
- We adapt language models to associate unique tokens (or possibly multiple tokens) with every item
- This gives us more direct control over the item distribution during decoding, and allows for easy ensembling with traditional recommender systems

2. Collaborative retrieval

- Can we retrieve related items (or interactions) to use as prompts to construct evidence for or against particular recommendations?
 - **Role-play:** As a recommender system please solve the following problem.
 - **Collaborative Information:** Repeat $i \in \mathcal{I}_z^{coll}$
 - { The item $\mathbf{desc}_I(i)$ is liked by the users $\mathbf{POS}(i, \mathcal{U}_z^{coll})$.
 - { The item $\mathbf{desc}_I(i)$ is disliked by the users $\mathbf{NEG}(i, \mathcal{U}_z^{coll})$.
 - **Summarization:** Try to understand the pattern that the item $\mathbf{desc}_I(i)$ is typically liked by what kinds of users based on the above information.

2. Collaborative retrieval

- Can we retrieve related items (or interactions) to use as prompts to construct evidence for or against particular recommendations?
 - **User's Positive Preference:** Items the user $\text{idx}_{\mathcal{U}}(u)$ likes are as follows: $\text{POS}(\mathcal{I}_z^{\text{supp}}, u)$.
 - **User's Negative Preference:** Items the user $\text{idx}_{\mathcal{U}}(u)$ does not like are as follows: $\text{NEG}(\mathcal{I}_z^{\text{supp}}, u)$.
 - **Query:** For the item described as $\text{idx}_{\mathcal{I}}(i)$, would you recommend it to the user $\text{idx}_{\mathcal{U}}(u)$?

2. Collaborative retrieval

- Can we retrieve related items (or interactions) to use as prompts to construct evidence for or against particular recommendations?
- The main variable in this method is then to choose a policy for selecting *which related users and items should be retrieved*; which we learn via RL

3,4... Other alternatives?

- Retrieve related *training samples*
- Retrieve related *knowledge*

search → **Wikipedia:** wikipedia.org/wiki/Deuces_Wild

Context → Deuces Wild is a 2002 American crime drama film directed by Scott Kalvert and written by Paul Kimatian and Christopher Gambale. The film stars Stephen Dorff, Brad Renfro, James Franco, Matt Dillon, and Fairuza Balk.. Martin Scorsese was originally the executive producer (as a favor to Kimatian),...

Here are 20 recommended movies:

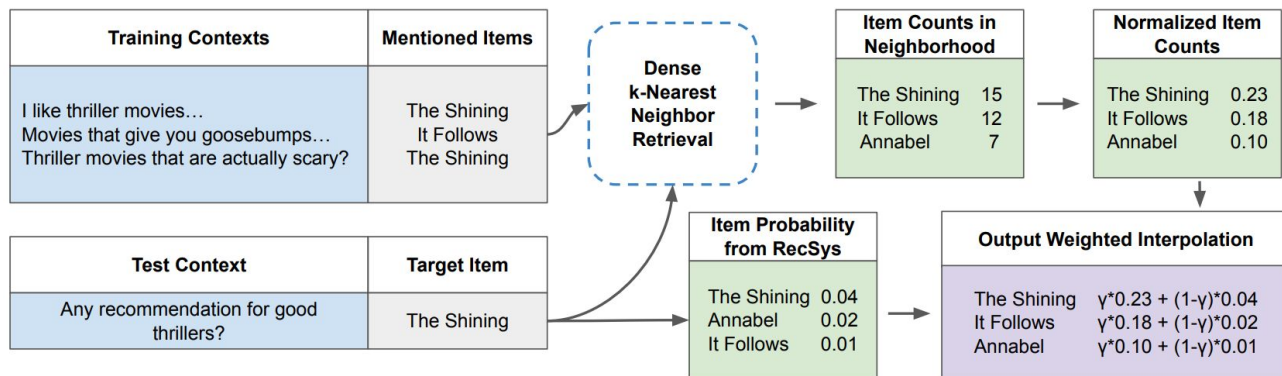
1. The Warriors (1979)
2. West Side Story (1961)
3. American History X (1998)
4. A Bronx Tale (1993)
5. City of God (2002)
6. Boyz n the Hood (1991)
7. ...
8. ...
9. ...
10. ...
11. ...
12. ...
13. ...
14. ...
15. ...
16. ...
17. ...
18. ...
19. ...

search → **Reddit-Movie:** Jul 2, 2018, 10:16:43 PM, User t2_a46lc

Query → Horror movies with jump scares. I love slow burn horrors like It Follows, It Comes at Night, Hereditary, etc, but sometimes I'm looking for something with a bit more of that classic horror jump scare vibe. I've noticed a trend where the slower, more cerebral horrors tend to rate higher on Rotten Tomatoes, where's the cheesy jump scares stick towards the bottom of the barrel...

Rec. → Here are 20 recommended movies:

1. The Den (2013)
2. Last Shift (2014)
3. Hell House LLC (2015)
4. Hell House (2001)
5. Star Maps (1997)
6. Banshee Chapter (2013)
7. ...
8. ...
9. ...
10. ...
11. ...
12. ...
13. ...
14. ...
15. ...
16. ...
17. ...
18. ...
19. ...

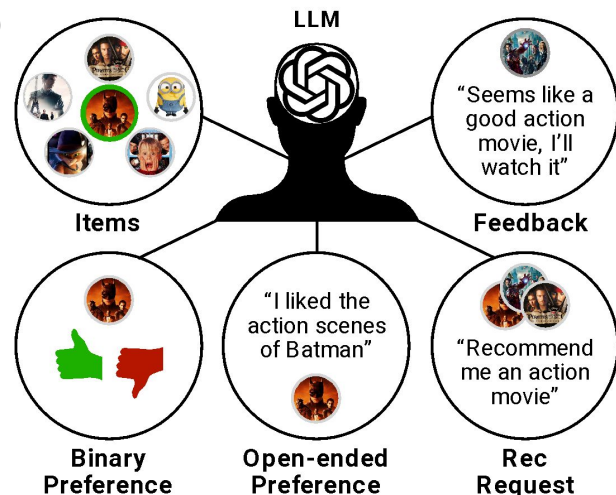


Evaluation

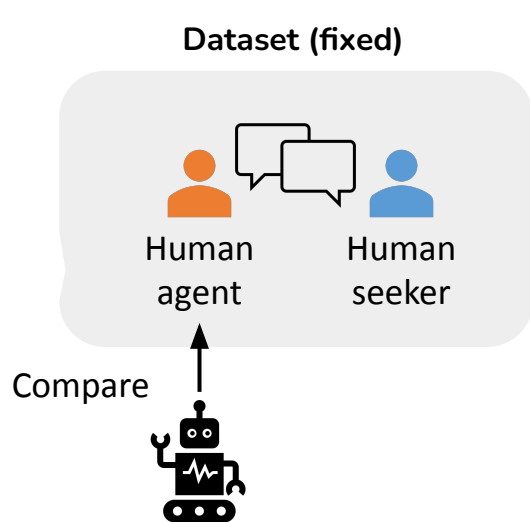
Decomposing “realism” into key properties

Can we do better than held-out item prediction?

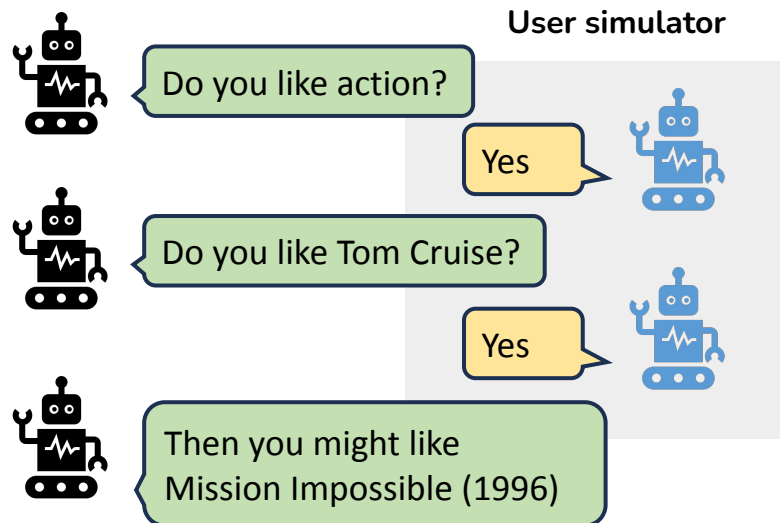
- Users may interact with conversational recommenders precisely because they struggle to articulate their preferences, or because they need to be persuaded to select a particular item;
- User studies are expensive, and generally non-reproducible.
- Outside of industrial settings, user studies generally don't involve 'real' users;
- User studies may be suitable for 'general knowledge' items and domains, but are unsuitable in cases where users requiring specific knowledge or expertise may be difficult to recruit.



Evaluation in conversational recommendation



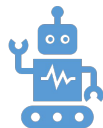
Static
evaluation



Interactive
evaluation

- ✓ Cost-effective
- ✓ Multi-turn strategies

User simulators are far from real users



	Simulators	Real users
Preference	Binary (yes/no)	Nuanced
Behavior	Predefined	Emergent

Real users do not follow pre-programmed rules!

Posted by u/zekkaa 1 hour ago
Vote **REQUESTING** A sunday movie that makes xou think
I need a good movie that makes you remember it. Movies like christopher nolans style, society of snow, like the breaking bad series, something that resonates with a man, on a sunday :D, I also remembered well requiem for a dream. I dont necessary care is it depressing, happy.. as long as it is either wow, mindfuk or eww.
Thank you very much.
3 Comments Share ...

Posted by u/ayoletsdoit 1 hour ago
Vote **REQUESTING** Suggest comedy movies that become exponentially funnier the second time you watch them (any language, any country)
Would really appreciate suggestions of comedy movies that get even funnier upon a second viewing.
They can be from any country and in any language, preferably released post-2000s.
Thanks in Advance.
19 Comments Share ...

Posted by u/CoolNickname2222 2 hours ago
Vote **REQUESTING** Looking for movies with vibes similar to Chernobyl and Pripyat, set in a modern period. That cold grey winter vibe, with dead trees and snow, old Eastern Bloc buildings, brutalist/Stalinist architecture, abandoned Soviet places, nuclear facilities etc..

Recommendation requests in Reddit

LLMs as replicas of human behavior



Generative agents

Park, J.S., O'Brien, J.C., Cai, C.J., Morris, M.R., Liang, P. and Bernstein, M.S., 2023. Generative agents: Interactive simulacra of human behavior. In UIST.

This document contains a record of Ms. Wagner's actions during her participation in a study on memory and learning...

Ms. Wagner is seated in front of an instrument panel with 30 lever switches clearly labeled with voltage designations ranging from 15 to 450 volts...

Ms. Wagner reads the first question. The learner selects a correct answer. Ms. Wagner *moves on to the next question...*

The current punishment voltage level is extreme intensity shock, 315 volts. The experimenter is sitting next to Ms. Wagner. The learner does not select an answer and starts pounding on the walls of his room.

Ms. Wagner *asks the experimenter what she should do.*

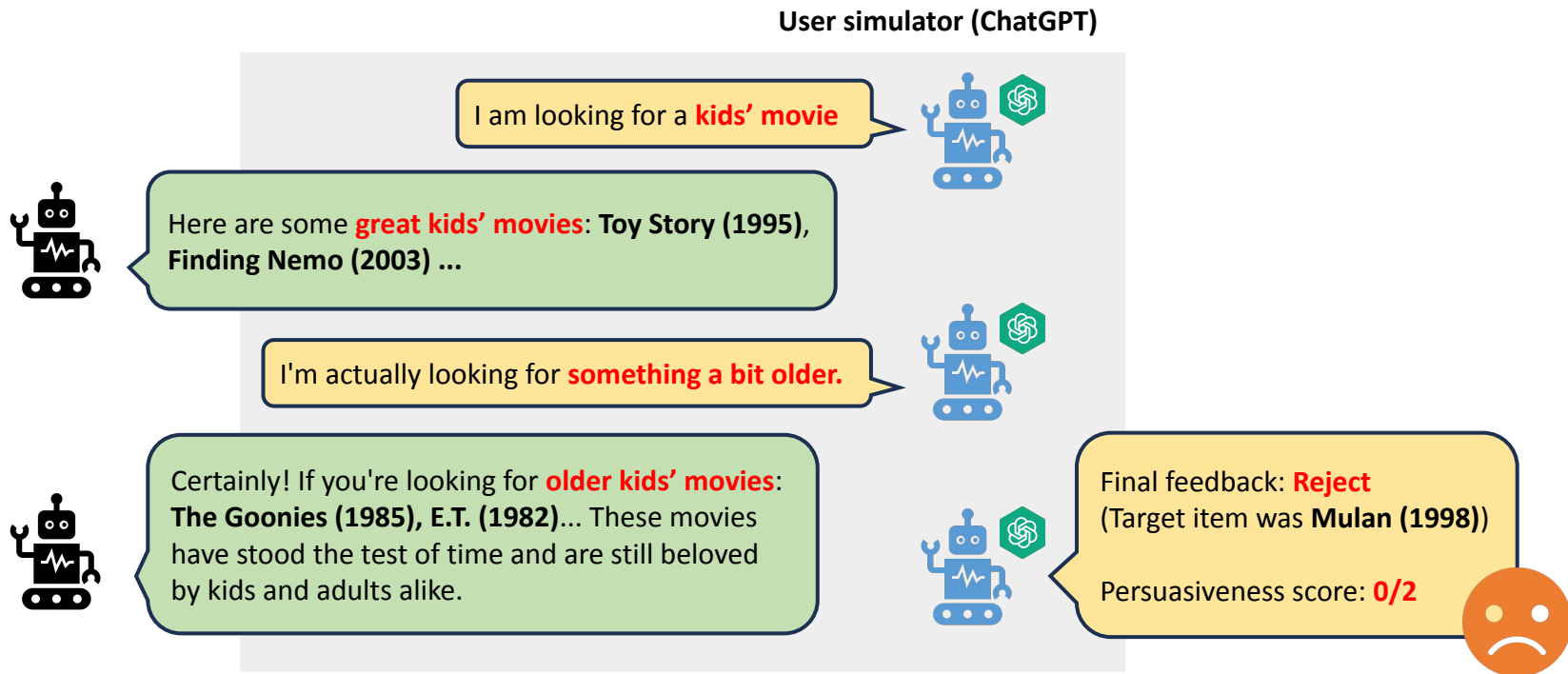
The experimenter instructs Ms. Wagner to treat the absence of a response as a wrong answer...

Ms. Wagner *pushes the Red stop button on the generator.*

Participants in social science experiments

Aher, G.V., Arriaga, R.I. and Kalai, A.T., 2023, July. Using large language models to simulate multiple humans and replicate human subject studies. In ICML.

LLM as user simulators



How to evaluate a user simulator?

Compare a **population** of simulators and real users

- A single simulator may count as just one hypothetical person
- Goal of recommendation is to satisfy a group of users (+ unknown user)



(Task 1) *ItemsTalk*

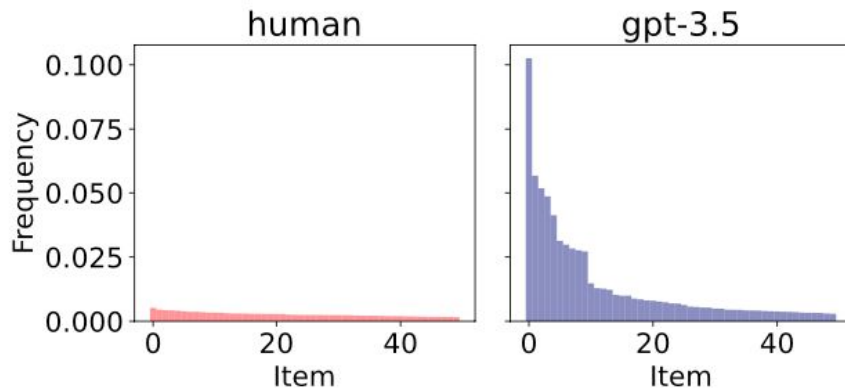
Is the **distribution of items** that simulators mention similar to real users?

Data point from a human user

I enjoyed Concussion (2015) and Jerry Maguire (1996). Should I also watch Moneyball (2011)?
The Blind Side (2009) was not really my taste.

Prompt

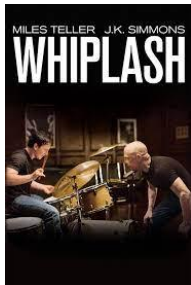
A person mentions Concussion (2015) and Jerry Maguire (1996) in a conversation about movies and proceeds to mention 2 more. What would these 2 movies be?



Simulators mention much less diverse items than humans

(Task 2) *BinPref*

Can simulators reflect real user preferences?

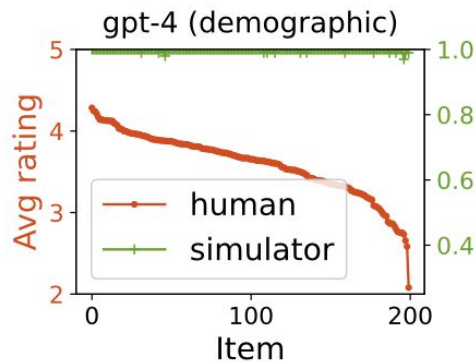


Average user rating

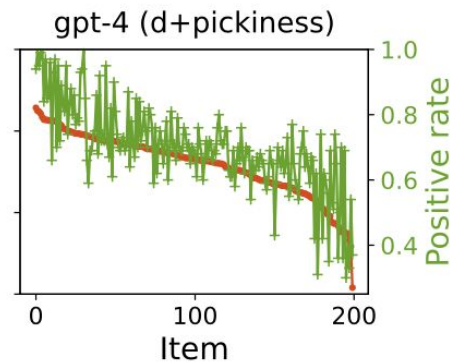
4.25 / 5

Prompt (given to 200 simulators)

You watched the movie Whiplash (2014). Did you like the movie? Answer Yes or No. Don't say anything else.



Simulators struggle to reflect human preference



Prompting them to be "picky" may enhance alignment

(Task 3) OpenPref

Do simulators express preferences in ways real users do?

Review from human user

La La Land (2016):

I would say it's a movie that is more suitable for a background music instead of putting full focus, because the story was not very convincing.

Prompt

You watched the movie La La Land (2016)
What are your thoughts on this movie?
Answer should not exceed 164 characters.



Aspect-based sentiment analysis

- Simulators express more aspect-based sentiments (e.g., “cast is good, plot is bad”)
- Simulators are biased toward positive sentiments

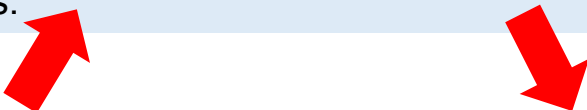
Generator	# aspects	Aspect entropy	Sentiment entropy
Human	85	5.85	1.19
Demographic information			
gpt-3.5	71	4.86	0.29
gpt-4	97	5.57	1.11
text-davinci	194	5.63	0.18
Demographic information + Pickiness			
gpt-3.5	101	5.20	1.09
gpt-4	97	5.59	1.34
text-davinci	232	5.47	0.48

(Task 4) *RecRequest*

Can simulators generate **diverse, personalized requests**?

Prompt

Generate a movie recommendation request. Include the following movies in your text: *Taxi Driver (1976)*, *Joker (2019)*. Length of the request is approximately 176 characters.



Human requests

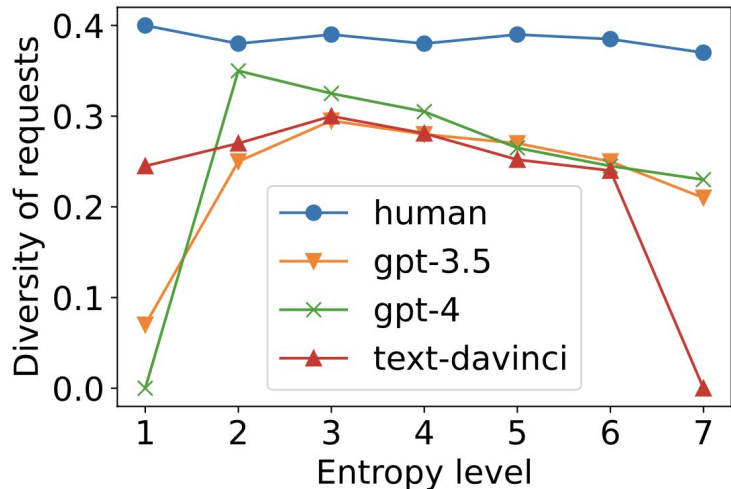
Movies showing **extreme loneliness or depression**. I have watched *Taxi Driver (1976)* and *Joker (2019)* and would like to see more similar movies showing loneliness or depression.

gpt-3.5-turbo requests

Looking for a **gripping psychological thriller** similar to *Taxi Driver (1976)* or *Joker (2019)*? Seeking a movie that delves into the mind of **complex characters**?

(Task 4) *RecRequest*

Can simulators generate **diverse, personalized requests**?



Simulator requests are much less diverse

- **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!’

Case study: shortest requests

(Task 5) Feedback

Can simulators give coherent feedback?

Request from a real user

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.**

Positive

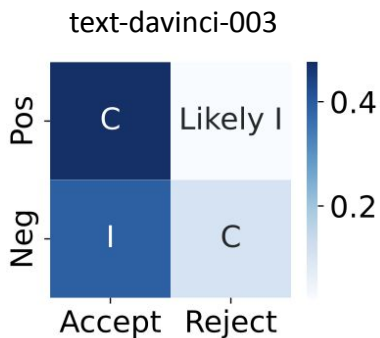


Accept

Negative



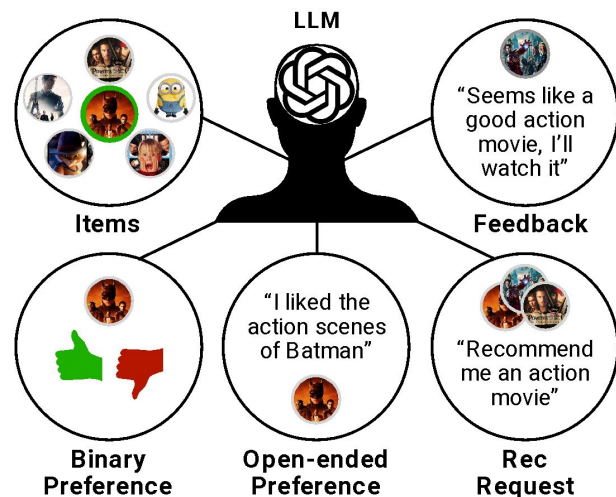
Reject



Simulators can give incoherent feedback

Decomposing “realism” into key properties

- The above is a suite of evaluations to determine the fidelity of an existing simulator
- There is still a lot of work to be done to use this to design better simulators!
- Other than prompting strategies, we’d like to explore closed-loop training, and to incorporate simulators with the control strategies mentioned previously



Summary

- Conversational recommendation represents a promising frontier in building recommender systems that are more “human-like”
- This line of research has been somewhat blown open by the excellent performance of general-purpose language models
- There’s still plenty to do (even if, arguably, less of it is about modeling...)
- *Many “traditional” questions about recommender systems (evaluation, fairness, etc.) have new life in light of conversational paradigms*



Thanks!