Uncovering ChatGPT's Capabilities in Recommender Systems

Jun Xu
Gaoling School of Artificial Intelligence
Renmin University of China

DataFunSummit # 2023

Uncovering ChatGPT's Capabilities in Recommender Systems

Sunhao Dai*
Gaoling School of Artificial Intelligence, Renmin University of China
Beijing, China
sunhaodai@ruc.edu.cn

Ninglu Shao*
Gaoling School of Artificial Intelligence, Renmin University of China
Beijing, China
ninglu_shao@ruc.edu.cn

Haiyuan Zhao*
School of Information, Renmin University of China
Beijing, China
haiyuanzhao@ruc.edu.cn

Weijie Yu
School of Information Technology and Management, University of International Business and Economics
Beijing, China
yuweijie23@gmail.com

Zihua Si
Gaoling School of Artificial Intelligence, Renmin University of China
Beijing, China
zihua_si@ruc.edu.cn

Chen Xu
Gaoling School of Artificial Intelligence, Renmin University of China
Beijing, China
xc_chen@ruc.edu.cn

Zhongxiang Sun
Gaoling School of Artificial Intelligence, Renmin University of China
Beijing, China
sunzhongxiang@ruc.edu.cn

Xiao Zhang
Gaoling School of Artificial Intelligence, Renmin University of China
Beijing, China
zhangx89@ruc.edu.cn

Jun Xu†
Gaoling School of Artificial Intelligence, Renmin University of China
Beijing, China
junxu@ruc.edu.cn

Sunhao Dai, Ninglu Shao, Haiyuan Zhao, Weijie Yu, Zihua Si, Chen Xu, Zhongxiang Sun, Xiao Zhang, and Jun Xu. 2023. Uncovering ChatGPT's Capabilities in Recommender Systems. In Seventeenth ACM Conference on Recommender Systems (RecSys '23), September 18–22, 2023, Singapore, Singapore. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3604915.3610646

PDF: https://arxiv.org/pdf/2305.02182.pdf
Github: https://github.com/rainymood/LLM4RS

Outline

- **LLMs for Recommendation**
- **Experimental Results**
- **Discussion**

Language Models for Recommendation

➢ Utilizing LMs training strategies

- BERT4Rec (masked language modeling)
- UniSRec (pre-train and fine-tune)
- P5 (pre-train and prompting)

……

➢ Using LMs to obtain better representations

- Better user representations
- Better item representations
- Better context representations

……

Key Task: Top-K Ranking of Items

- Pointwise ranking

$$s(i \mid u) = \Phi_{\text{point}}(\mathbf{x}_{u,i})$$

- Pairwise ranking

$$s(i_m > i_n \mid u) = \Phi_{\text{pair}}(\mathbf{x}_{u,i_m}, \mathbf{x}_{u,i_n})$$

- Listwise ranking

$$s(i_1 \mid u), s(i_2 \mid u), \cdots, s(i_k \mid u) = \Phi_{\text{list}}(\mathbf{x}_{u,i_1}, \mathbf{x}_{u,i_2}, \cdots, \mathbf{x}_{u,i_k})$$

LLMs for Recommendation: Overall Evaluation Procedure

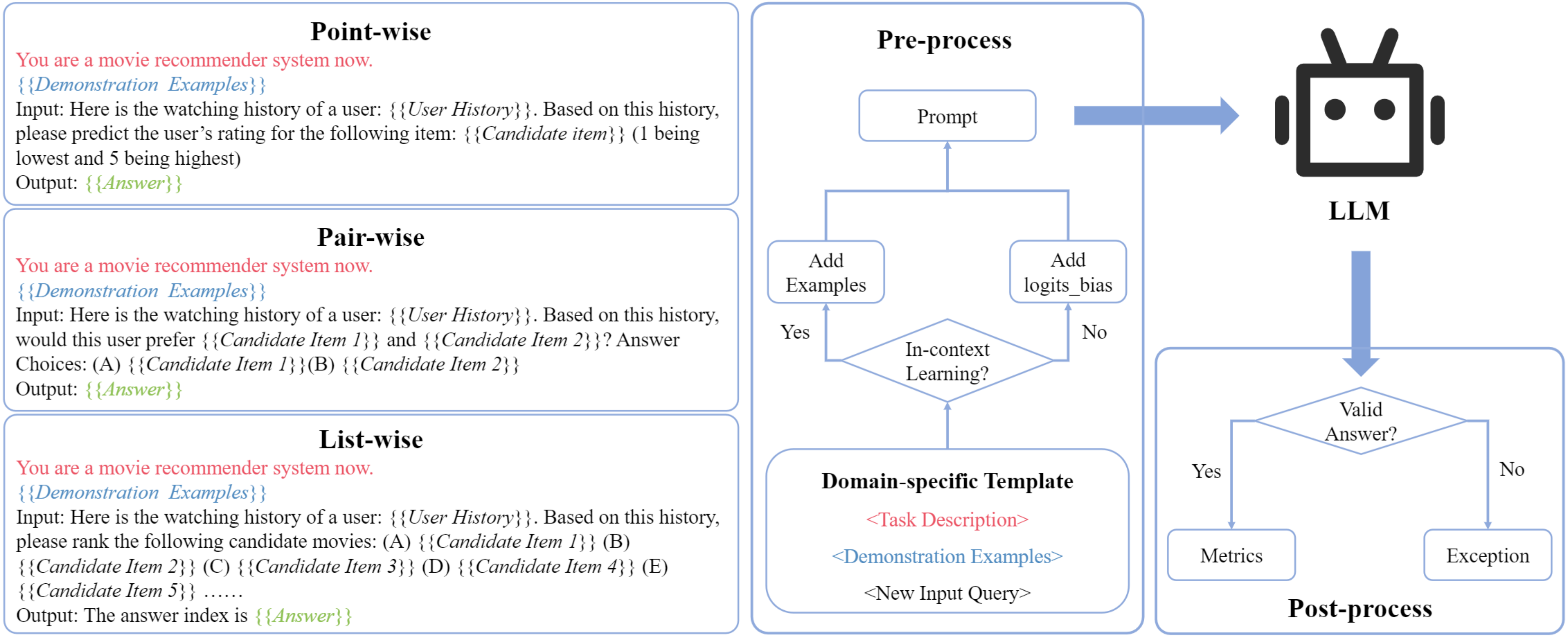


Fig. 1. The overall evaluation framework of LLMs for recommendation. The left part demonstrates examples of how prompts are constructed to elicit each of the three ranking capabilities. The right part outlines the process of employing LLMs to perform different ranking tasks and conduct evaluations.

The Prompt

- Task description I
- Demonstration examples $\mathcal{D} = \{f(\mathbf{h}_1, \mathbf{c}_1, \mathbf{y}_1), f(\mathbf{h}_2, \mathbf{c}_2, \mathbf{y}_2), \cdots, f(\mathbf{h}_N, \mathbf{c}_N, \mathbf{y}_N)\}$

  h: history interactions (can be any context features about each sample)
- New input query $f(\mathbf{h}', \mathbf{c}' \mid u)$

$$\mathbf{c} = \begin{cases} \{i\} & \text{for point-wise ranking capability;} \\ \{i_m, i_n\} & \text{for pair-wise ranking capability;} \\ \{i_1, i_2, \cdots, i_k\} & \text{for list-wise ranking capability;} \end{cases}$$

$$\hat{y}'_i = LLM_{\text{point}}\left(I, \mathcal{D}, f(\mathbf{h}', \mathbf{c}' \mid u)\right)$$

$$\hat{y}'_{i_m \succ i_n} = LLM_{\text{pair}}\left(I, \mathcal{D}, f(\mathbf{h}', \mathbf{c}' \mid u)\right)$$

$$\hat{y}'_{i_1}, \hat{y}'_{i_2}, \cdots, \hat{y}'_{i_k} = LLM_{\text{list}}\left(I, \mathcal{D}, f(\mathbf{h}', \mathbf{c}' \mid u)\right)$$

Prompt Examples

Pointwise

You are a movie recommender system now.

Task description

Input: Here is the watching history of a user: Gattaca, Armageddon, Big, Babes in Toyland, Gladiator. Based on this history, please predict the user's rating for the following item: Star Wars: Episode I - The Phantom Menace (1 being lowest and 5 being highest)

Output: 1.

Demonstration examples

Input: Here is the watching history of a user: The Virgin Suicides, Goya in Bordeaux, Snow Falling on Cedars, Gladiator, Hamlet. Based on this history, please predict the user's rating for the following item: The Last September (1 being lowest and 5 being highest)

Output: 5.

Input: Here is the watching history of a user: The Green Mile, Sling Blade, A River Runs Through It, Donnie Brasco, Sleepers. Based on this history, please predict the user's rating for the following item: American Beauty (1 being lowest and 5 being highest)

New input query

Output: 5.

Answer From LLM

Prompt Example

Pairwise

You are a movie recommender system now.

Task description

Input: Here is the watching history of a user: Gattaca, Armageddon, Big, Babes in Toyland, Gladiator. Based on this history, would this user prefer Donnie Brasco or Con Air? Answer Choices: **(A) Donnie Brasco** (B) Con Air

Output: The answer index is A.

Demonstration examples

Input: Here is the watching history of a user: The Adventures of Milo and Otis, The Pelican Brief, Dinosaur, Air Force One, The First Wives Club. Based on this history, would this user prefer George of the Jungle or Tom and Huck? Answer Choices: **(A) George of the Jungle** (B) Tom and Huck

New input query

Output: The answer index is A.

Answer From LLM

Prompt Example

高瓴人工智能学院
Gaoling School of Artificial Intelligence

Listwise

You are a movie recommender system now. Task description

Input: Here is the watching history of a user: Gattaca, Armageddon, Big, Babes in Toyland, Gladiator. Based on this history, please rank the following candidate movies: (A) Con Air (B) Mulan (C) Nikita **(D) Donnie Brasco** (E) Star Wars: Episode I - The Phantom Menace

Demonstration examples

Output: The answer index is D A E B C.

Input: Here is the watching history of a user: Rushmore, South Park: Bigger, Longer and Uncut, The Sixth Sense, Who Framed Roger Rabbit?, Trick. Based on this history, please rank the following candidate movies: **(A) Some Like It Hot** (B) Wild Wild West (C) Mystery Men (D) The Mosquito Coast (E) Boys Don't Cry

New input query

Output: The answer index is A C E D B.

Answer From LLM

Outline

高瓴人工智能学院
Gaoling School of Artificial Intelligence

- LLMs for Recommendation
- **Experimental Results**
- Discussion

Datasets and Baselines

- **Four datasets from different domains**
  - Movie: MovieLens-1M
  - Book: “Books” subset of Amazon
  - Music: “CDs & Vinyl” subset of Amazon
  - News: MIND-small
- **Baselines**
  - Random, Pop, MF, NCF
- **Metrics**
  - NDCG, MRR

Main Results

Table 1. Overall performance of different models on four datasets from different domains. Bold indicates the best result for each row and '_' indicates the best result for each LLM.

Domain	Metric	random	pop	text-davinci-002			text-davinci-003			gpt-3.5-turbo (ChatGPT)		
				point-wise	pair-wise	list-wise	point-wise	pair-wise	list-wise	point-wise	pair-wise	list-wise
Movie	Compliance Rate	-	-	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	99.98%	100.00%
	NDCG@1	0.2000	0.2240	0.3110	0.3203	0.2600	0.2259	0.2843	0.3260	**0.3342**	0.3230	0.3320
	NDCG@3	0.4262	0.4761	0.5416	0.5728	0.4990	0.4618	0.5441	0.5564	**0.5912**	0.5827	0.5785
	MRR@3	0.3667	0.4103	0.4824	0.5071	0.4363	0.3998	0.4763	0.4950	**0.5260**	0.5162	0.5167
Book	Compliance Rate	-	-	99.96%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	99.98%	99.80%
	NDCG@1	0.2000	0.2440	0.2420	0.2847	0.2000	0.2325	0.2887	0.2440	0.2823	0.3061	**0.3126**
	NDCG@3	0.4262	0.4999	0.4889	0.5298	0.4290	0.4585	0.5293	0.4597	0.5075	0.5350	**0.5395**
	MRR@3	0.3667	0.4340	0.4247	0.4646	0.3690	0.3993	0.4665	0.4040	0.4495	0.4774	**0.4800**
Music	Compliance Rate	-	-	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	99.96%	99.80%
	NDCG@1	0.2000	0.1780	0.2354	0.2397	0.2300	0.2377	0.2690	0.2540	0.2892	0.3077	**0.3086**
	NDCG@3	0.4262	0.4094	0.4623	0.4681	0.4277	0.4732	0.5072	0.4506	0.5201	0.5439	**0.5567**
	MRR@3	0.3667	0.3470	0.4030	0.4082	0.3750	0.4113	0.4448	0.4000	0.4605	0.4830	**0.4950**
News	Compliance Rate	-	-	99.80%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	99.60%
	NDCG@1	0.2000	**0.3080**	0.2183	0.2200	0.2920	0.2532	0.2630	0.2540	0.2591	0.2491	0.2892
	NDCG@3	0.4262	**0.5444**	0.4483	0.4550	0.5059	0.4880	0.4892	0.4742	0.4826	0.4991	0.5094
	MRR@3	0.3667	**0.4840**	0.3879	0.3936	0.4497	0.4271	0.4294	0.4173	0.4246	0.4354	0.4515

- LLMs performed much better than random/pop policies on almost all cases
- ChatGPT performed best across four domains

Main Results

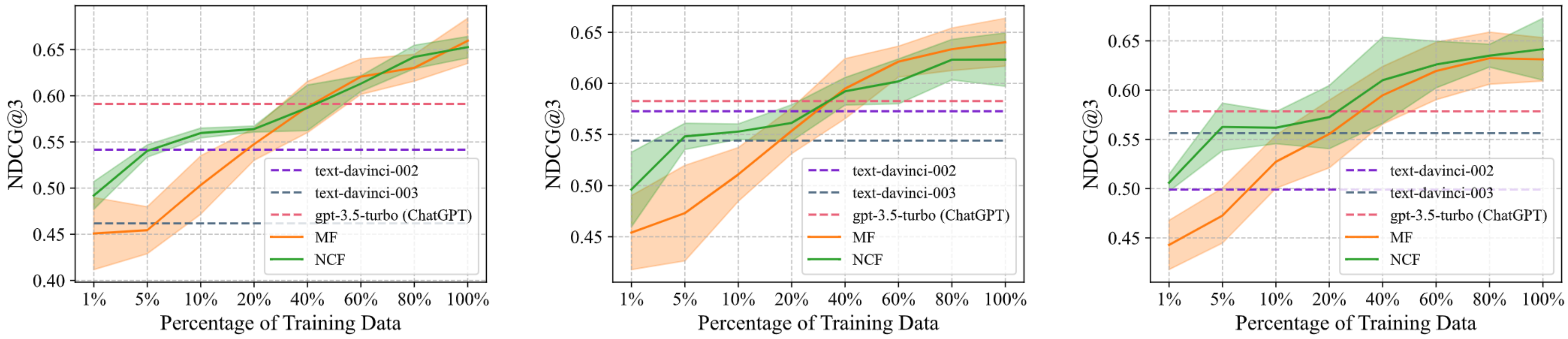
高瓴人工智能学院
Gaoling School of Artificial Intelligence

Table 2. Rank of different capabilities of different LLMs-based recommendation models on four datasets from different domains.

Domain	text-davinci-002	text-davinci-003	gpt-3.5-turbo (ChatGPT)
Movie	pair-wise $>$ point-wise $\gg$ list-wise	list-wise $\approx$ pair-wise $\gg$ point-wise	point-wise $>$ pair-wise $\approx$ list-wise
Book	pair-wise $\gg$ point-wise $\gg$ list-wise	pair-wise $\gg$ list-wise $\approx$ point-wise	list-wise $>$ pair-wise $\gg$ point-wise
Music	pair-wise $>$ point-wise $\gg$ list-wise	pair-wise $\gg$ point-wise $\gg$ list-wise	list-wise $>$ pair-wise $\gg$ point-wise
News	list-wise $\gg$ pair-wise $\approx$ point-wise	pair-wise $\approx$ point-wise $>$ list-wise	list-wise $>$ pair-wise $>$ point-wise

- text-davinci-002 and text-davinci-003 performed better with pair-wise ranking in most cases
- ChatGPT is better with list-wise ranking except on movie domain

LLMs v.s. Traditional Recommendation Models



(a) point-wise (b) pair-wise (c) list-wise

Fig. 4. Comparison with Matrix Factorization in terms of different percentages of training data on Movie dataset. The shaded area indicates the 95% confidence intervals of t-distribution.

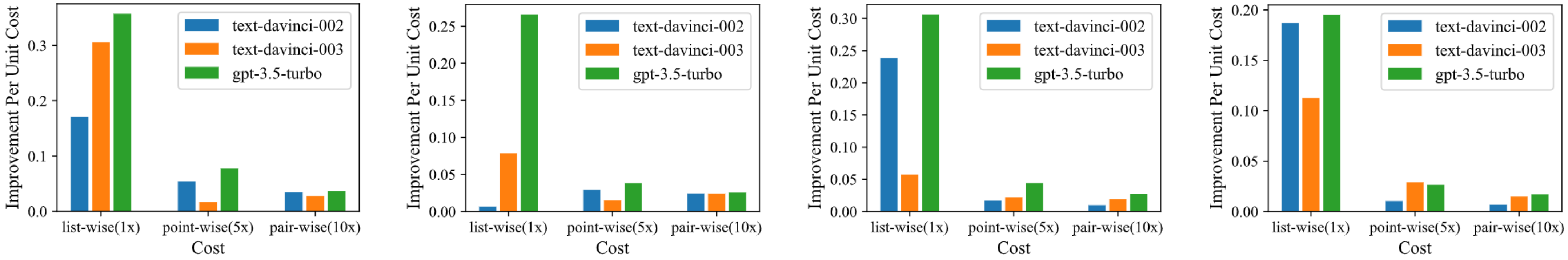
- Given enough user-item interactions, traditional models still performed better than LLMs
  - LLMs can be used to mitigate cold start

Costs for Improvements

$$\frac{\frac{V_{LLM}-V_{random}}{V_{random}}}{cost_{LLM}}$$

Cost:

- Pointwise: N
- Pairwise: N(N-1)/2
- Listwise: 1



(a) Movie (b) Book (c) Music (d) News

Fig. 2. Improvement of $NDCG@3$ per unit cost and five shots examples on four datasets. '1x 5x 10x' denote the cost of list-wise, point-wise, and pair-wise, respectively.

- ➢ **List-wise achieved the best improvements per unit cost**
- ➢ Point-wise and pair-wise are similar
  - ➢ Pair-wise achieved higher performances with the cost of more sessions

Zero-shot v.s. Few-shot

高瓴人工智能学院
Gaoling School of Artificial Intelligence

Table 3. Performance of different LLMs with zero-shot and few-shot examples on Movie dataset. Bold indicates the best result for each row and '_' indicates the best result for each wise of each LLM.

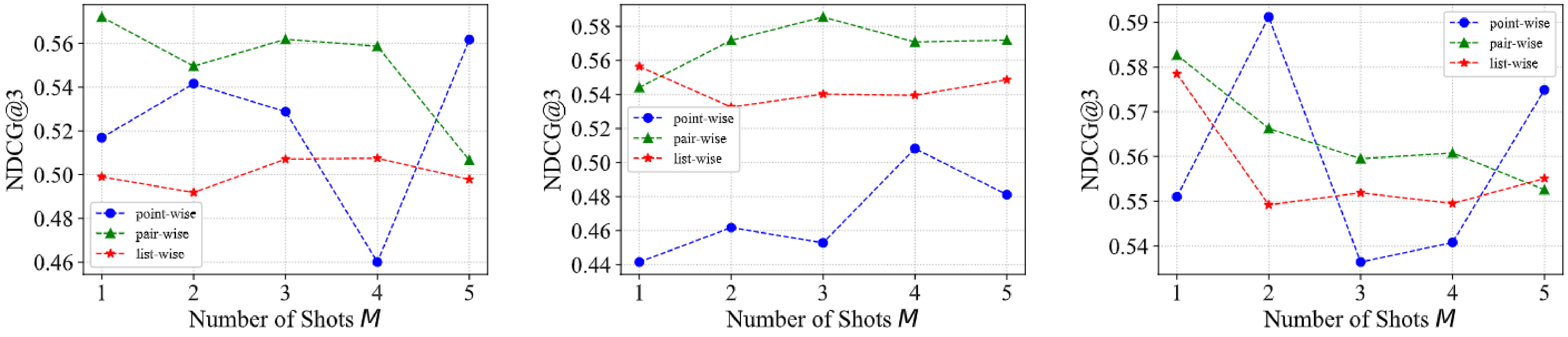
Model	Metric	random	pop	point-wise zero-shot	point-wise few-shot	pair-wise zero-shot	pair-wise few-shot	list-wise zero-shot	list-wise few-shot
text-davinci-002	NDCG@3	0.4264	0.4761	0.5168	<u>0.5416</u>	0.5253	**<u>0.5721</u>**	0.4544	<u>0.4990</u>
	MRR@3	0.3667	0.4103	0.4519	<u>0.4824</u>	0.4643	**<u>0.5066</u>**	0.3950	<u>0.4363</u>
text-davinci-003	NDCG@3	0.4264	0.4761	<u>0.4674</u>	0.4618	0.5249	**<u>0.5441</u>**	0.5062	<u>0.5564</u>
	MRR@3	0.3667	0.4103	<u>0.4092</u>	0.3998	0.4633	**<u>0.4763</u>**	0.4450	<u>0.4950</u>
gpt-3.5-turbo (ChatGPT)	NDCG@3	0.4264	0.4761	0.5413	**<u>0.5912</u>**	<u>0.5833</u>	0.5827	-	0.5785
	MRR@3	0.3667	0.4103	0.4742	**<u>0.5260</u>**	<u>0.5243</u>	0.5162	-	0.5167

- Zero-shot prompt outperformed the random and pop
  - ➔ capability of LLMs for zero-shot recommendation

Need at least one example to tell ChatGPT to output a ranking

- Few-shot performed much better than zero-shot in most cases
  - ➔ effective of few-shot prompts in-context learning

Analysis: Number of Prompt Shots



(a) text-davinci-002 (b) text-davinci-003 (c) gpt-3.5-turbo (ChatGPT)

Fig. 3. Impact of the number of shots prompts in LLMs on Movie dataset.

- More examples ≠ better results?
  - More examples → more context information, but **higher noise**

Analysis: Number of History Items

	HR@1			
History Length	1	3	5	10
text-davinci-003	0.325	0.314	0.332	**0.340**
gpt-3.5-turbo	**0.372**	0.362	0.368	0.355

- Given more historical interacted items ≠ better results?
  - More items → more context information, but **higher noise**
  - **Open question: how to retrieval historical items for recommendation**

Case Study: Ranking with Explanations

You are a movie recommender system now.

Input: Here is the watching history of a user: Sleepy Hollow, The Natural, Smoke, The Breakfast Club, Fearless. Based on this history, please rank the following candidate movies: **(A) Philadelphia** (B) FairyTale: A True Story (C) Four Weddings and a Funeral (D) Excess Baggage (E) Nick of Time

Output: The answer index is as follows: (A) Philadelphia (B) FairyTale: A True Story (C) Four Weddings and a Funeral (E) Nick of Time (D) Excess Baggage. Based on the user's watching history, they seem to enjoy a mix of genres including drama, sports, and coming-of-age films. Therefore, Philadelphia, a drama film about a lawyer fighting for justice for a man with AIDS, would be a good fit. FairyTale: A True Story, a family-friendly fantasy film, may not be the best match. Four Weddings and a Funeral, a romantic comedy, could be a good option as it has elements of drama and humor. Nick of Time, a thriller starring Johnny Depp, could also be a good fit based on the user's history of watching Sleepy Hollow. Excess Baggage, a crime comedy, may not be the best match.

Case Study: No Ranking, with Explanations

You are a movie recommender system now.

Input: Here is the watching history of a user: Blazing Saddles, The Outlaw Josey Wales, The Wild Bunch, The Good Bad and The Ugly, A Fistful of Dollars. Based on this history, please rank the following candidate movies: (A) Clueless (B) Shakespeare in Love **(C) True Grit** (D) Gone with the Wind (E) Hell in the Pacific

Output: The answer index is (C) True Grit. This movie fits well with the user's preference for westerns and action-packed films. The other options, such as Clueless and Shakespeare in Love, are not in the same genre and may not appeal to the user based on their watching history. Gone with the Wind is a classic, but it may not be as action-packed as the user's previous choices. Hell in the Pacific could be a good option, but True Grit is a more well-known and highly rated film in the same genre.

Case Study: Reject to Answer

高瓴人工智能学院
Gaoling School of Artificial Intelligence

You are a book recommender system now.

Input: Here is the reading history of a user: Dress Your Family in Corduroy and Denim, My Age of Anxiety: Fear, Hope, Dread, and the Search for Peace of Mind, Knuffle Bunny: A Cautionary Tale, Don't Let The Pigeon Drive The Bus!, Purplicious (Pinkalicious). Based on this history, would this user prefer Something Other Than God: How I Passionately Sought Happiness and Accidentally Found It or Moms Who Drink and Swear: True Tales of Loving My Kids While Losing My Mind? Answer Choices: **(A) Something Other Than God: How I Passionately Sought Happiness and Accidentally Found It** (B) Moms Who Drink and Swear: True Tales of Loving My Kids While Losing My Mind

Output: The answer index is A.

Input: Here is the reading history of a user: The Cellist of Sarajevo, After I'm Gone: A Novel, The Reason I Jump: The Inner Voice of a Thirteen-Year-Old Boy with Autism, The Serpent of Venice: A Novel, We Are All Completely Beside Ourselves: A Novel. Based on this history, would this user prefer The Secret Life of Bees or The Help? Answer Choices: (A) The Secret Life of Bees **(B) The Help**

Output: The answer index is N/A. It is difficult to determine the user's preference based on their reading history as neither book is similar to the books they have read.

Case Study: Reject to Answer

You are a music recommender system now.

Input: Here is the listening history of a user: Broken Bells, She & Him - Volume Two, April Uprising, Congratulations, You I Wind Land & Sea. Based on this history, would this user prefer This Is Happening or Thor?

Answer Choices: **(A) This Is Happening** (B) Thor

Output: The answer index is A.

Input: Here is the listening history of a user: Brian McKnight,I'll Be Home For Christmas,Razor & Tie,Holiday,Holiday Music,Soul/R&B,Christmas / Chanukkah />, Foreign Exchange, Rock Steady, Classique, Christy Nockels,Life Light Up,Sparrow Records,Pop,Religious / Contemp. Christian,Christian,Contemporary Christian Music,Christian - Contemporary,Christian - Modern Worship,Christian Music - Gospel (GS) />. Based on this history, would this user prefer Chinese Wall or VIP? Answer Choices: (A) Chinese Wall **(B) VIP**

Output: The answer index is N/A as neither option is a good fit based on the user's listening history. It would be better to recommend music that aligns with their preferred genres and artists.

- Reject to answer because they are two negative samples
  - Suggestion: evaluating LLMs for recommendation from more perspectives that **beyond rank**
  - LLMs can generate **recommendation reasons**.

Case Study: Incorrect Ranking with Explanations

You are a movie recommender system now.

Input: Here is the watching history of a user: Boys Don't Cry, Babe: Pig in the City, Babe, The Siege, King of the Hill. Based on this history, please rank the following candidate movies: (A) Jurassic Park (B) Apollo 13 (C) Braveheart (D) Evita **(E) The Poseidon Adventure**

Output: The answer index is as follows: (A) Jurassic Park, (B) Apollo 13, (C) Braveheart, (D) Evita, <u>(E) The Poseidon Adventure</u>. Based on the user's watching history, they seem to enjoy a mix of drama and action movies. Therefore, I would recommend (A) Jurassic Park and (B) Apollo 13 as they both have elements of action and drama. (C) Braveheart may also be a good choice as it is a historical drama with action scenes. (D) Evita and (E) The Poseidon Adventure may not be the best fit based on the user's watching history.

Outline

高瓴人工智能学院
Gaoling School of Artificial Intelligence

- LLMs for Recommendation
- Experimental Results
- **Discussion**

Summary

- Tested three types of prompts that can generate item rankings with LLMs
  - Pointwise, Pair-wise, and List-wise
  - ChatGPT are good at list-wise and pair-wise ranking, and list-wise is more economical
- Given enough user-item interactions, traditional models outperformed LLMs
  - LLMs can be used to mitigate the cold start problem
  - LLMs demonstrate the potential for interpretable recommendations (exhibits the capacity of comprehending item similarities)

Discussion

- LLMs v.s. traditional models (e.g., MF/NCF) for recommendation
  - LLMs: based on semantic information from the large scale text data (world knowledge)
  - Traditional: based on user-item interaction in (limited) training data (domain knowledge)
  - Are they complementary? (like LM4IR and PageRank in web search?)
  - How to better combine or align?
    - Embedding: user/item side info → LLMs → user/item text-embedding → combine/align with ID-embedding
    - LLMs as RecSys: natural language in and natural language out

- Ranking item with LLMs is not natural
  - New UI for recommendation
  - Do we really need natural language in ?

- Finetune is the way to go

Thank You!

Project Link： https://github.com/rainymood/LLM4RS