Supplementary Material for Uncovering ChatGPT's Capabilities in Recommender Systems

SUNHAO DAI[*], Gaoling School of Artificial Intelligence, Renmin University of China
NINGLU SHAO[*], Gaoling School of Artificial Intelligence, Renmin University of China
HAIYUAN ZHAO[*], School of Information, Renmin University of China
WEIJIE YU, School of Information Technology and Management, University of International Business and Economics
ZIHUA SI, Gaoling School of Artificial Intelligence, Renmin University of China
CHEN XU, Gaoling School of Artificial Intelligence, Renmin University of China
ZHONGXIANG SUN, Gaoling School of Artificial Intelligence, Renmin University of China
XIAO ZHANG, Gaoling School of Artificial Intelligence, Renmin University of China
JUN XU[†], Gaoling School of Artificial Intelligence, Renmin University of China

ACM Reference Format:
Sunhao Dai, Ninglu Shao, Haiyuan Zhao, Weijie Yu, Zihua Si, Chen Xu, Zhongxiang Sun, Xiao Zhang, and Jun Xu. 2023. Supplementary Material for 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, 5 pages. https://doi.org/10.1145/3604915.3610646

1 RELATED WORK

Large Language Models Pioneering studies [4, 22] demonstrated that LLMs can perform a diverse range of tasks without requiring gradient updates, solely based on textual instructions or a few examples. This has drawn significant attention towards improving the capabilities of LLMs. Previous studies [15] have investigated the performance limits of pre-trained language models (PLMs) by training larger models, as they have noted that augmenting the model or data size typically enhances the model's ability on downstream tasks, such as Megatron-turing NLG [27] with 530B parameters, Gopher [23] with 280B parameters, Ernie 3.0 Titan [31] with 260B parameters, BLOOM [25] with 175B parameters, and PaLM [7] with 540B parameters. These LLMs have exhibited exceptional performance on challenging tasks, showcasing new abilities that were not apparent in smaller pre-trained language models (PLMs). For a more comprehensive overview of LLMs, we would recommend referring to [36].

Existing Evaluation of ChatGPT As ChatGPT continues to gain worldwide popularity, more studies are focusing on evaluating it since it is perhaps one of the strongest LLMs to date. Bang et al. [2] propose to quantitatively evaluate ChatGPT from a multitask, multilingual, and multimodal perspective by analyzing their performance on 8 common NLP tasks. Their study reveals that while ChatGPT performs well on most tasks but may also have limitations and biases in

[*] Equal Contribution.
[†] Jun Xu is the corresponding author. Work partially done at Engineering Research Center of Next-Generation Intelligent Search and Recommendation, Ministry of Education.

reasoning, hallucination, and interactivity. Qin et al. [21] empirically show that ChatGPT possesses some zero-shot capabilities as a generalist model on 7 NLP tasks and conclude that ChatGPT performs poorly in solving specific tasks such as sequence tagging. Other researchers also do evaluations and case studies on ChatGPT's robustness [5, 30], ethics [13, 37] and its applications in education [10, 17, 19], medicine [1, 3, 24], recommender[8, 16], search[29] and law [6]. To the best of our knowledge, there has been no comprehensive evaluation of probing the ChatGPT's capabilities in recommender systems from different ranking perspectives. Therefore, we conducted an exhaustive evaluation of ChatGPT (as well as other GPT-3.5 series LLMs) on four recommendation domain benchmarks to fill this research gap.

Language Models for Recommendation The remarkable success of pre-trained LMs in NLP community has motivated researchers in recommender systems to explore their potential in recommendation tasks. Existing works can be categorized into two types: (i) utilizing LMs training strategies to reformulate and model recommendation tasks, such as BERT4Rec (*masked language modeling*) [28], UnisRec (*pre-train and finetune*)[12], P5 (*pre-train and prompting*) [9] and (ii) using LMs to obtain better representations of users, items and context based on textual information [14, 32, 34]. More recently, some researchers have explored leveraging off-the-shelf pre-trained LMs as recommender systems by reformulating the recommendation tasks with prompts as multi-token cloze tasks [20, 26, 35]. In this paper, we aim to conduct a preliminary evaluation of ChatGPT's potential and limitations in recommender systems.

2 EXPERIMENTAL DETAILS AND MORE EXPERIMENTAL RESULTS

2.1 Dataset

To better probe the different capabilities of ChatGPT and GPT-3.5s (text-davinci-002 and text-davinci-003) on personalized recommendation, we conducted evaluations on datasets from four different domains.

Movie: We use the widely-adopted MovieLens-1M[1] dataset that contains 1M user ratings for movies.

Book: We use the "Books" subset of Amazon Reviews[2] dataset that contains 1.8M user ratings for books.

Music: We use the "CDs & Vinyl" subset of Amazon Reviews[2] to conduct experiments on the music domain.

News: We use the MIND-small[3] dataset as the benchmark for news domain.

Following the common practices [11, 18, 33], for the Movie, Book, and Music datasets, we treat ratings above 3 as positive feedbacks (labeled as 1) and otherwise as negative feedbacks (labeled as 0). For the News dataset, we used the original binary feedback labels. For more details about the processing of datasets, please refer to the link[4].

2.2 Performance of Zero-shot Prompt

A natural question on the off-the-shelf LLMs for recommendation is whether LLMs can work without examples (i.e., $M = 0$). However, with the original 0-shot learning approach, we found that more than 50% of cases were invalid and difficult to evaluate in practice. Fortunately, OpenAI provides an API[2] that allows us to control the logits bias of output tokens. We were able to upweight the logit bias of the indexes of answers, which improved the compliance rate. For instance, for pair-wise ranking, we increased the probabilities of both outputs 'A' and 'B' by the same magnitude, while ensuring that their relative order remains unchanged. Under this setting, we conduct the zero-shot example experiments and the results are shown in Table 1. These findings demonstrate the potential of LLMs as recommendation systems, as

[1] https://grouplens.org/datasets/movielens/1m/

[2] http://jmcauley.ucsd.edu/data/amazon/

[3] https://msnews.github.io/

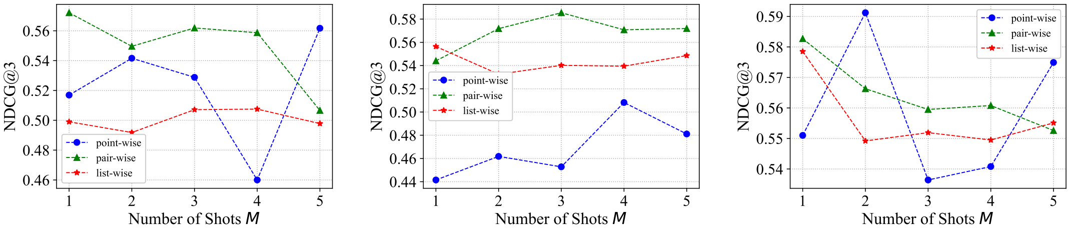
[4] https://github.com/rainym00d/LLM4RS/tree/main/data

[1] For zero-shot setting, valid outputs are derived by manipulating the 'logit_bias' and 'top_p' parameters of the API. However, it should be noted that due to a limitation in gpt-3.5-turbo's control over the 'top_p' parameter, there are currently no available experimental results in this regard.

[2] https://platform.openai.com/docs/api-reference/completions/create#completions/create-logit_bias

Table 1. 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		pair-wise		list-wise	
				zero-shot	few-shot	zero-shot	few-shot	zero-shot	few-shot
text-davinci-002	NDCG@3	0.4264	0.4761	0.5168	0.5416	0.5253	**0.5728**	0.4544	0.4990
	MRR@3	0.3667	0.4103	0.4519	0.4824	0.4643	**0.5071**	0.3950	0.4363
text-davinci-003	NDCG@3	0.4264	0.4761	0.4674	0.4618	0.5249	**0.5441**	0.5062	0.5564
	MRR@3	0.3667	0.4103	0.4092	0.3998	0.4633	**0.4763**	0.4450	0.4950
gpt-3.5-turbo (ChatGPT)	NDCG@3	0.4264	0.4761	0.5413	**0.5912**	0.5833	0.5827	N/A[1]	0.5785
	MRR@3	0.3667	0.4103	0.4742	**0.5260**	0.5243	0.5162		0.5167



(a) text-davinci-002

(b) text-davinci-003

(c) gpt-3.5-turbo (ChatGPT)

Fig. 1. Impact of the number of shots prompts in LLMs on Movie dataset.

they perform better than a random policy and a popularity-based policy under the zero-shot setting. Furthermore, as expected, LLMs under few-shot settings outperform those under zero-shot settings in most cases, demonstrating the effectiveness of few-shot prompts in-context learning.

2.3 Performance Under Different Shots Examples

Previous studies in natural language processing (NLP) have emphasized that the number of examples M is very important for in-context learning. To assess the impact of M in LLMs for recommendation, we conducted experiments on Movie dataset by varying M from 1 to 5. Figure 1 illustrates the performances of different M in terms of the $NDCG@3$ of ChatGPT and GPT3.5s. Surprisingly, we observe that the best results did not always correspond to the maximum number of examples. One possible explanation is that while more example shots can provide more context and information for LLMs to understand the recommendation task, they may also introduce more noise, causing the models to learn unhelpful patterns. Therefore, the optimal number of prompt shots may depend on the specific LLM, task, and dataset.

REFERENCES

[1] Fares Antaki, Samir Touma, Daniel Milad, Jonathan El-Khoury, and Renaud Duval. 2023. Evaluating the performance of chatgpt in ophthalmology: An analysis of its successes and shortcomings. *medRxiv* (2023), 2023–01.

[2] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023* (2023).

[3] James RA Benoit. 2023. ChatGPT for Clinical Vignette Generation, Revision, and Evaluation. *medRxiv* (2023), 2023–02.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[5] Xuanting Chen, Junjie Ye, Can Zu, Nuo Xu, Rui Zheng, Minlong Peng, Jie Zhou, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. How Robust is GPT-3.5 to Predecessors? A Comprehensive Study on Language Understanding Tasks. *ArXiv* abs/2303.00293 (2023).

[6] Jonathan H Choi, Kristin E Hickman, Amy Monahan, and Daniel Schwarcz. 2023. Chatgpt goes to law school. *Available at SSRN* (2023).

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

[8] Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. 2023. Chat-REC: Towards Interactive and Explainable LLMs-Augmented Recommender System. *ArXiv* abs/2303.14524 (2023).

[9] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2022. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5). In *Proceedings of the 16th ACM Conference on Recommender Systems*. 299–315.

[10] Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. *arXiv preprint arXiv:2301.07597* (2023).

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

[12] Yupeng Hou, Shanlei Mu, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. 2022. Towards Universal Sequence Representation Learning for Recommender Systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 585–593.

[13] Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech. *Companion Proceedings of the ACM Web Conference 2023* (2023).

[14] Chanwoo Jeong, Sion Jang, Eunjeong Park, and Sungchul Choi. 2020. A context-aware citation recommendation model with BERT and graph convolutional networks. *Scientometrics* 124 (2020), 1907–1922.

[15] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).

[16] Junling Liu, Chao Liu, Renjie Lv, Kang Zhou, and Yan Zhang. 2023. Is ChatGPT a Good Recommender? A Preliminary Study. *arXiv preprint arXiv:2304.10149* (2023).

[17] Muneer M Alshater. 2022. Exploring the role of artificial intelligence in enhancing academic performance: A case study of ChatGPT. *Available at SSRN* (2022).

[18] Kelong Mao, Jieming Zhu, Jinpeng Wang, Quanyu Dai, Zhenhua Dong, Xi Xiao, and Xiuqiang He. 2021. SimpleX: A simple and strong baseline for collaborative filtering. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 1243–1252.

[19] Desnes Nunes, Ricardo Primi, Ramon Pires, Roberto de Alencar Lotufo, and Rodrigo Nogueira. 2023. Evaluating GPT-3.5 and GPT-4 Models on Brazilian University Admission Exams. *ArXiv* abs/2303.17003 (2023).

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

[21] Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a General-Purpose Natural Language Processing Task Solver? *arXiv preprint arXiv:2302.06476* (2023).

[22] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. [n. d.]. Language Models are Unsupervised Multitask Learners. ([n. d.]).

[23] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446* (2021).

[24] Arya S Rao, John Kim, Meghana Kamineni, Michael Pang, Winston Lie, and Marc Succi. 2023. Evaluating ChatGPT as an adjunct for radiologic decision-making. *medRxiv* (2023), 2023–02.

[25] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100* (2022).

[26] Damien Sileo, Wout Vossen, and Robbe Raymaekers. 2022. Zero-Shot Recommendation as Language Modeling. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part II*. Springer, 223–230.

[27] Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990* (2022).

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

[29] Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agent. *ArXiv* abs/2304.09542 (2023).

[30] Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, et al. 2023. On the Robustness of ChatGPT: An Adversarial and Out-of-distribution Perspective. *arXiv preprint arXiv:2302.12095* (2023).

[31] Shuohuan Wang, Yu Sun, Yang Xiang, Zhihua Wu, Siyu Ding, Weibao Gong, Shikun Feng, Junyuan Shang, Yanbin Zhao, Chao Pang, Jiaxiang Liu, Xuyi Chen, Yuxiang Lu, Weixin Liu, Xi Wang, Yangfan Bai, Qiuliang Chen, Li Zhao, Shiyong Li, Peng Sun, Dianhai Yu, Yanjun Ma, Hao Tian, Hua Wu, Tian Wu, Wei Zeng, Ge Li, Wen Gao, and Haifeng Wang. 2021. ERNIE 3.0 Titan: Exploring Larger-scale Knowledge Enhanced Pre-training for

Language Understanding and Generation. arXiv:2112.12731 [cs.CL]

[32] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2021. Empowering news recommendation with pre-trained language models. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1652–1656.

[33] Chen Xu, Jun Xu, Xu Chen, Zhenghua Dong, and Ji-Rong Wen. 2022. Dually Enhanced Propensity Score Estimation in Sequential Recommendation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 2260–2269.

[34] Qi Zhang, Jingjie Li, Qinglin Jia, Chuyuan Wang, Jieming Zhu, Zhaowei Wang, and Xiuqiang He. 2021. UNBERT: User-News Matching BERT for News Recommendation.. In *IJCAI*. 3356–3362.

[35] Yuhui Zhang, Hao Ding, Zeren Shui, Yifei Ma, James Zou, Anoop Deoras, and Hao Wang. 2021. Language models as recommender systems: Evaluations and limitations. (2021).

[36] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A Survey of Large Language Models. arXiv:2303.18223 [cs.CL]

[37] Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Exploring ai ethics of chatgpt: A diagnostic analysis. *arXiv preprint arXiv:2301.12867* (2023).