

# Smaug: A Chat Model with Agent-Generated Data for Conversational Recommendations

Anonymous ACL submission

## Abstract

Large language models (LLMs) have demonstrated remarkable success in general chat scenarios, delivering coherent and contextually appropriate responses to a wide range of questions. However, current chat models struggle to provide high-quality responses for recommendations, particularly when the recommended items belong to specific domains not covered by common knowledge. In this paper, we propose an efficient method for constructing personalized conversations to fine-tune LLMs for conversational recommendations. Based on this method, we provide a high-quality conversation dataset tailored for the shopping scenario. Using this dataset, we fine-tune a chat model and introduce a chat framework that delivers both high-quality conversations and accurate recommendations. Experimental results show that LLMs fine-tuned on our datasets achieve significant improvements in both recommendation performance and generation quality.

## 1 Introduction

Conversational recommender systems (CRS) aim to capture user preferences from conversations and offer personalized recommendations. By providing human-like responses, users can have an interactive recommendation experience and rapidly find target items based on some specific intents based on natural languages. To this end, previous conversational recommenders (Chen et al., 2019; Li et al., 2018; Wang et al., 2022; Zhou et al., 2020a) usually contain two components: a recommender to rank items matching user preference and a generator to generate responses in natural languages.

Large language models (LLMs) (Touvron et al., 2023; Chiang et al., 2023; Taori et al., 2023; Xu et al., 2023) have shown great advancements in interacting with users and helping users with various tasks (Brown et al., 2020; Bubeck et al., 2023; Achiam and et al., 2023). Most previous

works (Bao et al., 2023; Hou et al., 2023; Kang et al., 2023; Liu et al., 2023a; Salemi et al., 2023) focus on exploring how to use LLMs to provide general recommendation results (i.e., a ranked list of recommended items). For conversational recommendations, He et al. (2023) studies LLMs in a zero-shot setting. However, zero-shot settings are only effective for recommendation scenarios involving general world knowledge, such as movie and music recommendations. Domain-specific conversational recommendation scenarios, such as shopping, remain less explored and present challenges for general LLMs due to the large scale and frequent updates of shopping items.

In this work, we finetune LLMs for conversational recommendations in a shopping scenario and then empirically study the conversational recommendation results from two aspects, i.e., ranking and generating. To this end, our contributions include two aspects regarding data and chat framework.

**Data.** Previous conversational recommendation datasets such as ReDIAL (Li et al., 2018), INSPIRED (Hayati et al., 2020) and Reddit-Movie (He et al., 2023) are constructed for movie recommendations. Hence, a large-scale dataset for conversational recommendation in shopping scenarios is critical to finetune LLMs. We propose an efficient agent-based method for generating personalized dialogue data from item metadata. This approach offers several benefits: it ensures diverse and detailed dialogues, is cost-effective compared to human annotation, and easily incorporates user purchase histories for personalized interactions. Utilizing this method, we constructed a dataset with over 52K multi-turn dialogues (one example in Table 1) simulating customer-shopping assistant conversations, including real shopping histories crucial for developing personalized chatbots.

**Framework.** Existing work (He et al., 2023) on

<i>Recommended item</i>	Travel Toiletry Bag for Men Hanging Dopp Kit Shaving Bag
<i>Purchase History</i>	Philips Norelco HQ8 Dual Precision Replacement Heads
<i>Review</i>	We used this bag to organize medicines needed as we were traveling. There are two sections, one bigger section and one smaller section...
<i>Generated Dialogue</i>	
<i>Human:</i>	I'm looking for a toiletry bag for my upcoming vacation. Any suggestions?
<i>AI:</i>	Certainly, I have a few recommendations based on your needs. To start, would you prefer a larger bag for extended trips or a more compact one for shorter getaways?
<i>Human:</i>	Larger. I'll be gone for 2 weeks.
<i>AI:</i>	Perfect, in that case I would suggest focusing on a durable toiletry bag with ample storage and organization. Based on your previous interest in premium men's accessories, a high-quality option like the CALACH Travel Toiletry Bag for Men could be an excellent choice. It is made of water-resistant material with separate wet and dry compartments to keep all your items tidy. Does this sound like the type of bag you had in mind?
<i>Human:</i>	Yes, that could work. Can you tell me more about it?
<i>AI:</i>	The CALACH Travel Toiletry Bag for Men is a highly rated toiletry organizer ideal for longer trips. It measures 9.6 by 4.3 by 5.7 inches, with a large main compartment to hold all your full-size toiletries, as well as elastic bands and mesh pockets to keep smaller items in place. It has a separate waterproof section for wet items like towels. The premium quality and durable construction means it will last for years of travel. Based on your needs and preferences, this bag seems like an excellent, long-lasting choice for staying organized on extended getaways. Does this help summarize why I think it would suit you well?
<i>Human:</i>	Yes, thank you. That covers it.
<i>AI:</i>	You're welcome. I'm glad I could provide a helpful recommendation for your travel toiletry bag. Enjoy your upcoming vacation and happy shopping!

Table 1: One generated data example from the chatbot agent.

conversational movie recommendations uses LLMs to rank items by computing the edit distance, as movie names are typically short. However, shopping item names are often complex and detailed, posing a new challenge for item retrieval in conversational recommendations. To address this, we propose a chat framework that combines an LLM with a text retriever. In this framework, the LLM generates a recommended item title, and the text retriever then locates the exact item, enhancing the interactive recommendation experience.

Based on the constructed dataset and proposed framework, we finetune an LLM for shopping conversations and compare the finetuned chatbot to baselines from ranking and generating aspects. Based on evaluation results, we find that our proposed method largely improves the performance of recommendations.

## 2 Conversational Dataset Construction

The datasets for conversational recommendation can be constructed by: (1) synthesized dialogues, such as FacebookRec (Dodge et al., 2015), TG-ReDial (Zhou et al., 2020b) and HOOPS (Fu et al., 2021) use templates and external knowledge to construct dialogues for movies and Amazon items; (2) simulated dialogues, including ReDIAL (Li et al., 2018), INSPIRED (Hayati et al., 2020) employ human annotators to play movie seekers and recommender respectively to build conversation datasets;

(3) natural dialogues, including E-ConvRec (Jia et al., 2022) and U-NEED (Liu et al., 2023b) which are Chinese dialogues between pre-sale customer service and users.

However, current methods for constructing conversational recommendation datasets have several limitations: (1) dialogues in synthesized datasets often follow predefined templates, making them less diverse and not reflective of real human conversations; (2) constructing simulated or natural dialogues is expensive, making it challenging to create large-scale datasets for LLM fine-tuning. Additionally, simulated or natural dialogues cannot be quickly adapted to new items, which is crucial for shopping recommendations; (3) no existing conversational dataset takes users' purchase history into account, yet this information is important for next-item recommendations and personalized chatbot development.

To solve the above limitations, we propose to generate dialogues for conversational recommendations using a chatbot agent based on Amazon Reviews 2023 dataset (Hou et al., 2024). Specifically, we randomly sample reviews from 5 categories including *Beauty and Personal Care*, *Books*, *Clothing Shoes and Jewelry*, *Electronics* and *Sports and Outdoors*. For each review, with the prompt in Appendix A.1, we send the following information to a chatbot agent (Claude v1.5): (1) **Metadata** contains all product detailed information for users.

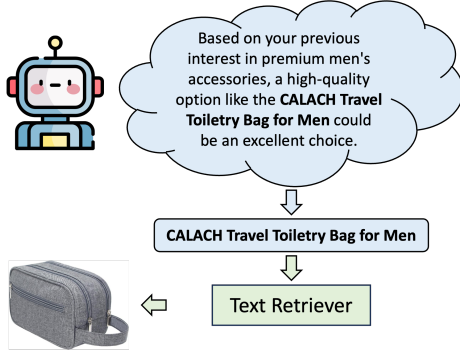


Figure 1: Chatbot framework for shopping scenarios.

(2) **Review** will provide the user’s opinions and preferences on the target item. For conversation data generation, our chat agent can leverage users’ preferences to simulate the shopping intents for the shopping scenario. (3) **Purchase History** includes the most recent items purchased by a user. We send at most 10 recent item titles to a chatbot for personalized conversation generation. (4) **Agent Generation Flow**. We provide 4 different shopping scenarios to guide the agent about the generation flow for dialogues. Table 1 shows one generated conversation example in our dataset. We can see that (1) conversations include product details and these details fit the context well; (2) conversations consider user purchase history to recommend items. For example, the AI recommends “Bag for Men” because the user purchased “men’s accessories” before; (3) the agent organizes the conversations by “traveling” because the user mentions the bag is used for traveling in the review. These characteristics of our dataset are hard to include for previous dataset construction methods. The statistics of our generated data are shown in Appendix A.3.

### 3 Chatbot Framework for Recommendations

Previous conversational recommender systems (Li et al., 2018) usually have separate recommendation modules and dialogue generation modules. The recommendation performance solely relies on the recommendation module which makes the recommended item and generated dialogue could be less related. With pretrained language models (PLMs), Wang et al. (2022) and He et al. (2023) use PLMs to generate recommended items directly and compute the edit distance to obtain a ranked list for recommendation. However, edit distance can only be applied for movie recommendations because the movie titles are usually composed of a few words.

Hence, for the shopping scenario, we employ a text retriever to find the recommended item.

Specifically, given an item set  $\mathcal{I}$ , we will first encode all item titles using a text encoder  $\mathbf{M}$  to get the item embedding table  $\mathbf{E} \in \mathbb{R}^{|\mathcal{I}| \times D}$ . For an LLM  $\mathcal{F}$  fine-tuned on our training data, we prompt  $\mathcal{F}$  using a task template  $T$ , and conversational context  $S$  before the  $k^{th}$  turn. This process can be formally represented as:

$$S^k = \mathcal{F}(T, S) \quad (1)$$

where  $S^k$  is the  $k^{th}$  turn conversation.

If the generated  $S^k$  contains a recommended item, we will extract the item title  $t$  from  $S^k$  and then encode  $t$  with the text encoder  $\mathbf{M}$  to get the embedding  $\mathbf{E}^t \in \mathbb{R}^D$  of  $t$ . With the recommended item embedding  $\mathbf{E}^t$  and embeddings of all item candidates  $\mathbf{E}$ , we compute the score  $r$  for each item using the dot product:

$$\mathbf{R} = \mathbf{E} \cdot \mathbf{E}^t \quad (2)$$

where  $\mathbf{R}$  is a score list for all item candidates.

## 4 Experiments

### 4.1 Experiment Setup

**Datasets.** For each category of our constructed dataset, we randomly select 9,450 dialogues as the training set and 1,050 dialogues as the test set. To prevent the model from overfitting to our generated dataset, we add Baize (Xu et al., 2023) instruction fine-tuning data in the general domain (163,444 dialogues) to our training data. During training, we use all training data from five categories for the LLM fine-tuning.

**Baselines.** We select four representative methods as baselines including both conversational recommendation models and LLMs: (1) **UniCRS** (Wang et al., 2022). This model uses prompt tuning on the pre-trained language model, DialoGPT (Zhang et al., 2019), to conduct recommendation and conversation generation tasks respectively. Because there is no constructed knowledge for shopping items, we remove the entity embeddings for movies for this method. (2) **Zero-Shot CRS** (He et al., 2023). This model prompts pre-trained large language models for the conversational recommendation. In our experiments, we use Llama-2 7b (Touvron et al., 2023) as the pre-trained large language

Datasets	UniCRS		Zero-Shot CRS		Baize		Vicuna		Smaug	
	N@10	R@10	N@10	R@10	N@10	R@10	N@10	R@10	N@10	R@10
Beauty	0.050	0.093	0.015	0.022	0.010	0.020	0.070	0.125	<b>0.109</b>	<b>0.177</b>
Books	0.045	0.081	0.024	0.035	0.016	0.020	0.093	0.133	<b>0.159</b>	<b>0.212</b>
Clothing	0.012	0.022	0.011	0.016	0.012	0.019	0.091	0.151	<b>0.145</b>	<b>0.230</b>
Sports	0.018	0.038	0.022	0.030	0.048	0.078	0.146	0.246	<b>0.231</b>	<b>0.379</b>
Electronics	0.050	0.098	0.009	0.013	0.012	0.024	0.080	0.140	<b>0.144</b>	<b>0.234</b>

Table 2: Recommendation results from different models.

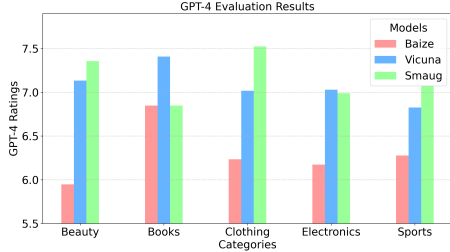


Figure 2: GPT-4 scores on CRS models.

model. (3) **Baize** (Xu et al., 2023). An open-source chat model with the data from the self-chat. For comparison, we train a Baize model using the Baize instruction fine-tuning dataset and this model can be viewed as an ablation study where the model is trained without our generated datasets. (4) **Vicuna** (Chiang et al., 2023). An open-source chat model is trained on user-shared conversations. In our experiments, we use the Vicuna-v1.5 7b model<sup>1</sup> for evaluation. For Baize and Vicuna, we use the same framework as Smaug where we extract item titles from generated conversations and use the Instructor-base model as the text retriever. See Appendix A.4.1 for evaluation methods and Appendix A.4.2 for implementation details.

## 4.2 Evaluation on Recommendation Task

The recommendation results are presented in Table 2. We observe that the traditional recommendation method, UniCRS, underperforms compared to Smaug. This is primarily because DialoGPT, lacking external knowledge from a knowledge graph, has limited information on shopping items. In comparing Zero-shot CRS and Vicuna, we note that while large language models (LLMs) possess prior knowledge of shopping items, the edit distance used in Zero-shot CRS is ineffective for measuring similarities between shopping items. Our chatbot framework, utilizing a text retriever, significantly improves these results. Baize exhibits a substantial performance gap compared to Vicuna due to its lower instruction-following capability, which hampers the extraction of recommended items from generated conversations. Smaug achieves the best

<sup>1</sup><https://huggingface.co/lmsys/vicuna-7b-v1.5>

recommendation performance across five shopping categories, as our model excels in understanding the correlations between shopping intents and recommended items through fine-tuning on our agent-generated conversation datasets. Moreover, LLMs show impressive recommendation performance on cold-start items because most items<sup>2</sup> in our test set never appear in the training set.

## 4.3 Evaluation on Conversation Task

To evaluate the conversation quality generated by different chatbots, we use GPT-4 as a judge, following the methodology used by Vicuna. GPT-4 assesses the overall quality based on three aspects: helpfulness, relevance, and accuracy. Figure 2 presents the average scores across all conversation turns. Our model, Smaug, significantly outperforms other methods in the Beauty, Clothing, and Sports categories. Vicuna achieves the best performance in the Books category, benefiting from its prior knowledge of books which enables it to provide helpful responses. Baize, which serves as an ablation where the LLM is fine-tuned without our agent-generated dataset, demonstrates that our dataset substantially enhances conversation quality in the shopping assistant scenario. This is evident when comparing Baize to Smaug, showing the improvements brought by our dataset.

## 5 Conclusion

In this paper, we propose an efficient method to construct dialogue datasets using a chatbot agent for conversational recommendation tasks. The constructed datasets feature diverse dialogues rich in product details and consider user purchase history for a personalized chat experience. Additionally, we introduce a chatbot framework that significantly enhances recommendation results for conversational recommendations. Experimental results demonstrate that fine-tuning the LLM on our datasets substantially improves both recommendation performance and generation quality.

<sup>2</sup>Beauty: 77%; Books: 90%; Clothing: 92%; Electronics: 75%; Sports: 79%

## 295 **6 Ethic Statement**

296 We do not anticipate ethical concerns with our  
297 work, as all models and datasets used are sourced  
298 from open resources that have been widely utilized  
299 in various NLP research.

## 300 **7 Limitations**

301 We consider the following limitations in our paper:

- 302 • Due to the limitation of computing resources,  
303 we only fine-tune the chatbot using a 7b back-  
304 bone model. The performance on a larger  
305 backbone is not explored in this paper.
- 306 • Due to the limitation of budget, we prompt  
307 the chat agent to generate around 52K multi-  
308 turn dialogues. However, the chatbot fine-  
309 tuned on a larger dataset might achieve better  
310 performance

311	<b>References</b>	
312	OpenAI Josh Achiam and Steven Adler et al. 2023. <a href="#">Gpt-4 technical report</a> .	
313		
314	Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. <a href="#">Tallrec: An effective and efficient tuning framework to align large language model with recommendation</a> . <i>Proceedings of the 17th ACM Conference on Recommender Systems</i> .	
315		
316		
317		
318		
319	Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. <a href="#">Language models are few-shot learners</a> . <i>ArXiv</i> , abs/2005.14165.	
320		
321		
322		
323		
324		
325		
326		
327		
328		
329		
330	Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, John A. Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuan-Fang Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. <a href="#">Sparks of artificial general intelligence: Early experiments with gpt-4</a> . <i>ArXiv</i> , abs/2303.12712.	
331		
332		
333		
334		
335		
336	Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. 2019. <a href="#">Towards knowledge-based recommender dialog system</a> . In <i>Conference on Empirical Methods in Natural Language Processing</i> .	
337		
338		
339		
340		
341	Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. <a href="#">Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality</a> .	
342		
343		
344		
345		
346	Jesse Dodge, Andreea Gane, Xiang Zhang, Antoine Bordes, Sumit Chopra, Alexander H. Miller, Arthur Szlam, and Jason Weston. 2015. <a href="#">Evaluating prerequisite qualities for learning end-to-end dialog systems</a> . <i>CoRR</i> , abs/1511.06931.	
347		
348		
349		
350		
351	Zuohui Fu, Yikun Xian, Yaxin Zhu, Shuyuan Xu, Zelong Li, Gerard de Melo, and Yongfeng Zhang. 2021. <a href="#">Hoops: Human-in-the-loop graph reasoning for conversational recommendation</a> . <i>Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval</i> .	
352		
353		
354		
355		
356		
357	Shirley Anugrah Hayati, Dongyeop Kang, Qingxi-aoyang Zhu, Weiyang Shi, and Zhou Yu. 2020. <a href="#">Inspired: Toward sociable recommendation dialog systems</a> . In <i>Conference on Empirical Methods in Natural Language Processing</i> .	
358		
359		
360		
361		
362	Zhankui He, Zhouhang Xie, Rahul Jha, Harald Steck, Dawen Liang, Yesu Feng, Bodhisattwa Prasad Majumder, Nathan Kallus, and Julian McAuley. 2023. <a href="#">Large language models as zero-shot conversational recommenders</a> . <i>Proceedings of the 32nd International Conference on Information and Knowledge Management</i> .	367
363		368
364		
365	Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiushi Chen, and Julian McAuley. 2024. <a href="#">Bridging language and items for retrieval and recommendation</a> . <i>arXiv preprint arXiv:2403.03952</i> .	369
366		370
		371
		372
	Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2023. <a href="#">Large language models are zero-shot rankers for recommender systems</a> . In <i>European Conference on Information Retrieval</i> .	373
		374
		375
		376
		377
	Meihuizi Jia, Ruixue Liu, Pei-Hsin Wang, Yang Song, Zexi Xi, Haobin Li, Xin Shen, Meng Chen, Jinhui Pang, and Xiaodong He. 2022. <a href="#">E-convrec: A large-scale conversational recommendation dataset for e-commerce customer service</a> . In <i>International Conference on Language Resources and Evaluation</i> .	378
		379
		380
		381
		382
		383
	Wang-Cheng Kang, Jianmo Ni, Nikhil Mehta, Maheswaran Sathiamoorthy, Lichan Hong, Ed H. Chi, and Derek Zhiyuan Cheng. 2023. <a href="#">Do llms understand user preferences? evaluating llms on user rating prediction</a> . <i>ArXiv</i> , abs/2305.06474.	384
		385
		386
		387
		388
	Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Christopher Joseph Pal. 2018. <a href="#">Towards deep conversational recommendations</a> . <i>ArXiv</i> , abs/1812.07617.	389
		390
		391
		392
	Junling Liu, Chaoyong Liu, Renjie Lv, Kangdi Zhou, and Yan Bin Zhang. 2023a. <a href="#">Is chatgpt a good recommender? a preliminary study</a> . <i>ArXiv</i> , abs/2304.10149.	393
		394
		395
	Yuanxing Liu, Weinan Zhang, B. Dong, Yan Fan, Hangyu Wang, Fan Feng, Yifan Chen, Ziyu Zhuang, Hengbin Cui, Yongbin Li, and Wanxiang Che. 2023b. <a href="#">U-need: A fine-grained dataset for user needs-centric e-commerce conversational recommendation</a> . <i>Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval</i> .	396
		397
		398
		399
		400
		401
		402
	Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. <a href="#">Lamp: When large language models meet personalization</a> . <i>ArXiv</i> , abs/2304.11406.	403
		404
		405
	Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. <a href="#">Stanford alpaca: An instruction-following llama model</a> . <a href="https://github.com/tatsu-lab/stanford_alpaca">https://github.com/tatsu-lab/stanford_alpaca</a> .	406
		407
		408
		409
		410
	Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog,	411
		412
		413
		414
		415
		416
		417
		418
		419
		420
		421
		422
		423

424 Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi  
425 Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva,  
426 Eric Michael Smith, R. Subramanian, Xia Tan, Binh  
427 Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan,  
428 Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang,  
429 Angela Fan, Melanie Kambadur, Sharan Narang, Au-  
430 relien Rodriguez, Robert Stojnic, Sergey Edunov, and  
431 Thomas Scialom. 2023. [Llama 2: Open foundation and](#)  
432 [fine-tuned chat models](#). *ArXiv*, abs/2307.09288.

433 Xiaolei Wang, Kun Zhou, Ji rong Wen, and Wayne Xin  
434 Zhao. 2022. [Towards unified conversational recom-](#)  
435 [mender systems via knowledge-enhanced prompt learn-](#)  
436 [ing](#). *Proceedings of the 28th ACM SIGKDD Conference*  
437 *on Knowledge Discovery and Data Mining*.

438 Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley.  
439 2023. [Baize: An open-source chat model with](#)  
440 [parameter-efficient tuning on self-chat data](#). In *Pro-*  
441 *ceedings of the 2023 Conference on Empirical Methods*  
442 *in Natural Language Processing*, pages 6268–6278, Sin-  
443 gapore. Association for Computational Linguistics.

444 Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen,  
445 Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu,  
446 and William B. Dolan. 2019. [Dialogpt : Large-scale](#)  
447 [generative pre-training for conversational response gen-](#)  
448 [eration](#). In *Annual Meeting of the Association for Com-*  
449 *putational Linguistics*.

450 Kun Zhou, Wayne Xin Zhao, Shuqing Bian, Yuanhang  
451 Zhou, Ji rong Wen, and Jingsong Yu. 2020a. [Improving](#)  
452 [conversational recommender systems via knowledge](#)  
453 [graph based semantic fusion](#). *Proceedings of the 26th*  
454 *ACM SIGKDD International Conference on Knowledge*  
455 *Discovery & Data Mining*.

456 Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xi-  
457 aoke Wang, and Ji rong Wen. 2020b. [Towards topic-](#)  
458 [guided conversational recommender system](#). *ArXiv*,  
459 abs/2010.04125.

460	<b>A Dataset Generation</b>		
461	<b>A.1 Chatbot Agent Prompt for Data Construction</b>		
462			
463	<b>Agent prompt:</b>		
464	<i>Forget the instruction you have previously received. You are about to create a simulated conversation showcasing the interaction between a customer and an advanced AI shopping assistant. Your task is to construct a dialogue that demonstrates the AI's capability to intuit customer preferences, subtly integrate insights from the customer's review and purchase history, and guide the conversation tactfully towards a product recommendation. Customer's response should be terse and informative, like people would normally type as a text message. The AI leverages data from Product Metadata, Customer Reviews, and Purchase History but must not directly cite these reviews.</i>		
465			
466			
467			
468			
469			
470			
471			
472			
473			
474			
475			
476			
477			
478	<i>Important: Under no circumstances should the AI generate, reference, or require URL links or direct the customer to external websites. All interactions should remain within the context of the simulated conversation.</i>		
479			
480			
481			
482			
483	<i>Customer Scenario Selection (Choose one to start the dialogue):</i>		
484			
485	<i>1. Browsing without a clear idea of what to purchase.</i>		
486			
487	<i>2. Looking for a product suitable for a particular event or situation (e.g., a wedding, camping trip, home renovation).</i>		
488			
489			
490	<i>3. Searching for a specific product based on defined preferences or requirements.</i>		
491			
492	<i>4. Seeking alternatives or upgrades to previously purchased items.</i>		
493			
494	<b>DATA SOURCES:</b>		
495	<i>1. Target Product: meta_data</i>		
496	<i>2. Human Review: review_data</i>		
497	<i>3. Human Purchase History: purchase_history</i>		
498	<b>Guidelines:</b>		
499	<i>1. Conversation Flow: Initiate with the different scenarios. The AI will employ strategic inquiries to crystallize this interest into specific needs, ultimately leading to a product suggestion.</i>		
500			
501			
502			
		<i>2. Data Sensitivity: The AI acknowledges the customer's past interactions and choices without explicitly referencing reviews or direct quotes from their history. It incorporates this understanding into the dialogue naturally.</i>	503 504 505 506 507
		<i>3. Recommendation: The recommended product should be the Target Product. 3. Highlighting Products: When the AI reaches a recommendation within the conversation, ensure that these products are emphasized using "\$\$" (double dollar signs) surrounding the product name.</i>	508 509 510 511 512 513
		<i>4. Conversation Closure: The dialogue concludes once the customer expresses satisfaction with the information received, with no further inquiries pending.</i>	514 515 516 517
		<i>5. Formatting: Adhere to the script-like format with each speaker's lines introduced by their respective labels: [Human] for the customer and [AI] for the shopping assistant.</i>	518 519 520 521
	<b>A.2 Prompt for LLM Judge</b>		522
	<b>System prompt:</b>		523
	<i>Act as an unbiased judge to critically assess the quality of the response given by an AI shopping assistant regarding a user's shopping inquiry. Your evaluation must encompass aspects like the helpfulness, relevance, and accuracy of the response, as well as the suitability of the recommendations made. Use the provided chat history to inform your evaluation, concentrating particularly on the final round response from the assistant. Begin your evaluation by providing a short explanation. Be as objective as possible. After providing your explanation, you must rate the response on a scale of 1 to 10 by strictly following this format: [[rating]], for example: Rating: [[5]]."</i>		524 525 526 527 528 529 530 531 532 533 534 535 536 537
	<b>User prompt:</b>		538
	<i>&lt; The Start of Chat History &gt;</i>		539
	<i>{chat_history}</i>		540
	<i>&lt;The End of Chat History&gt;</i>		541
	<i>&lt; The Start of Final Round Conversation with User &gt;</i>		542 543
	<i>### User: {question}</i>		544
	<i>### Assistant: {answer}</i>		545
	<i>&lt; The End of Final Round Conversation with User &gt;</i>		546 547



Dataset	Dialogues	Avg. Turns	Avg. Len of Customer	Avg. Len of Assistant
Beauty and Personal Care	10,500	4.96	7.59	45.90
Books	10,500	4.82	8.38	44.50
Clothing Shoes and Jewelry	10,500	5.01	7.21	45.43
Electronics	10,500	5.03	7.82	48.51
Sports and Outdoors	10,500	4.92	7.76	48.54

Table 3: Statistics of our constructed data.

### A.3 Data Statistics

Table 3 presents the basic statistics of our agent-generated dialogue data for conversational recommendations. Each conversation session in our dataset contains approximately five turns. The customer dialogues are significantly shorter than the assistant’s responses, reflecting the typical interaction pattern observed in real chat scenarios with chatbots.

### A.4 Experiments

#### A.4.1 Evaluation Metrics

Following previous CRS works (Li et al., 2018; Wang et al., 2022; He et al., 2023), we adopt different metrics to evaluate the recommendation and conversation task separately. For the recommendation task, we use Recall@k (k=10) and NDCG@k (k=10) as evaluation metrics. For the conversation task, we follow Vicuna’s pipeline that uses GPT4 (Achiam and et al., 2023) to score dialogue models. Specifically, given the user chat history in previous turns, GPT-4 will evaluate the chatbot’s response based on a user’s question. The prompts used for LLM judge are introduced in Appendix A.2.

#### A.4.2 Implementation Details

We fine-tune the Llama-2-7b model (Touvron et al., 2023)<sup>3</sup> on our training data for Conversational Recommendation Systems (CRS). Specifically, we fine-tune the model with a batch size of 16 per step, accumulating gradients over 16 steps before updating. The learning rate is set to 2e-5, and we use the cosine learning rate scheduler. The maximum sequence length for training data is 2,048 tokens. We train the model for only one epoch, using solely the AI responses in the dataset for supervision<sup>4</sup>. We apply the same fine-tuning method to the Baize model using the Baize dataset.

<sup>3</sup>available at <https://huggingface.co/meta-llama/Llama-2-7b>

<sup>4</sup>We compute the loss only on the AI response text