For model parameters in the recommender f_{rec}, including user and item embeddings, we adopt SGD for optimization with a learning rate of 10^{-3} and weight decay of 10^{-4}. For the remaining parameters, we rely on Adam optimization with a learning rate of 5×10^{-4} and weight decay of 10^2. Since two optimizers may converge with different speeds, we make Adam backpropagate gradients every other epoch, while SGD updates across all 10 epochs. We set the batch size to 256. For model parameters, the sizes of word embedding, KG embedding, and user/item embedding are 200, 100, 100, respectively, and the latent vector dimensionality is $d = 100$. The multi-head attention size in the Transformer encoder is set to 4.

5 EXPERIMENTS

We extensively evaluate the proposed benchmark method over the HOOPS benchmark data. First of all, the model should be able to accurately conduct the next attribute prediction within the graph, to demonstrate the capability of pruning off irrelevant candidates within the HitL graph reasoning paradigm. Moreover, we expect the proposed HitL conversational recommendation to not only facilitate offering accurate recommendations but also to properly select the next questions to ask with user feedback, which correspond to the recommendation task and the next question prediction tasks, respectively. For each of these tasks, we compare our model against several state-of-the-art baselines.

Experimental Settings. Recall that our HOOPS dataset includes Cellphones & Accessories, Grocery & Gourmet, Toys & Games, and Automotive. Each provides a unique KG and a set of conversations, implying that results are not necessarily comparable across different domains. We split the conversations into training (60%), validation (20%), test (20%) portions. For each user–item pair, we take one conversation with a maximum utterance length of 50 and a maximum conversation length of 10, applying zero-padding if the number of utterances is less than 10. There are 10 question candidates to predict, out of which only one is the correct ground truth choice. The same setup also applies for next-hop entity prediction. For recommendation, we sampled 100 items with which the user has not interacted as negative candidates. Our goal is to retrieve 1 correct labeled item out of a pool of 100 candidates, 1 question out of 10 question candidates, as well as 1 entity out of 10.

Baselines. For recommendation task, we consider Bayesian personalized ranking **BPR** [29], collaborative knowledge base embedding **CKE** [43], **RippleNet** [33], and the knowledge graph attention network **KGAT** [35] as baselines. For next-question prediction, we compare popular response ranking methods, including the deep matching network **DMN** [40], deep attention matching network **DAM** [48], and multi-hop selector network **MSN** [48]. The baselines above each either yield recommendations or address the next question prediction task. However, none of them is able to accommodate both tasks. Hence, we implement the following modified baselines targeted at jointly conducting both tasks. **KBRD** [7]: This is a conversational recommender system that originally couples recommendation with dialogue generation. We applied Transformers [32] with a decoder designed for our response selection downstream task. **OpenDialKG** [24]: The DialKG Walker model is able to conduct conversational reasoning. The original version supports predicting a KG entity via an attention-based graph path decoder. We modified the model by encoding the target question with an LSTM, which enables next question prediction.

Next Attribute Prediction. We study the performance of descriptive attribute prediction to justify whether the HitL graph reasoning is able to correctly predict the next attribute entity. Since the KG incorporates meta-information of both users and items, predicting the most relevant entities manifests a proper user participation that enables pruning off irrelevant candidates. The results in Table 2 indicate that our baseline approach obtains the best results compared to all prior baselines. Seq2Seq and LSTM are typical methods designed for sequential prediction, but they are unable to perform well with the aid of graph structures. Moon et al. [24] deployed a graph decoder by walking over knowledge graphs. However, without considering the hybrid user behavior in the modeling, it remains less convincing in terms of the transparency.

Next Question Prediction. In our benchmark dataset, we assume users may occasionally struggle to provide useful requests to the agent, since they initially may not be entirely aware of their preferences. Thus, learning to ask the right question given the past conversation context reveals whether the model successfully predicts user preferences. The benchmark results are shown in Table 2. In our HitL graph reasoning for conversational recommendation scenario, next question prediction closely resembles response ranking. The OpenDialKG and KBRD baselines exploit KGs in order to leverage sentence, dialogue, and KG structural features. Our proposed benchmark method not only takes advantage of the extracted coarse-to-fine entities within the KG, but also models the user feedback within the conversational turns. This enables it to outperform other baselines in most of the evaluation results.

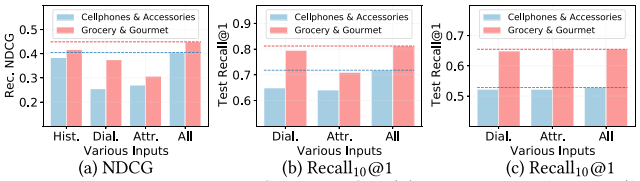
Recommendation. We adopt standard metrics to evaluate the recommendations of each user in the testset, including Normalized Discounted Cumulative Gain (**NDCG**), **Recall**, and Mean Average Precision (**MAP**). The top-10 recommendation results of different models are given in Table 2. The benchmark method is able to outperform other approaches, as it draws on human feedback and HitL graph reasoning to enhance the recommendation quality.

Ablation Study. We show the influence of different modules taking care of corresponding inputs on the three sub-tasks to demonstrate the effectiveness of our designed framework. As shown in Figure 3(a), we first consider the recommendation performance with each input separately with abbreviations Hist. = User History, Dial. = dialogue, and Attr. = descriptive attributes. While keeping all other parameters unchanged, we observe that each input contributes substantially to the performance, but retaining only one of them leads to a performance drop. This suggests that each ingredient of our HitL approach is complementary rather than redundant. The model is almost equal to user-based collaborative filtering when the input is solely user behavior, which takes the dominant role for personalized recommendation. In contrast, although the dialogue provides more semantics than pure attributes, it is worth noting that the conversational utterances may also introduce noise in the input. Therefore, there is a slight recommendation performance gap between dialogue-alone and attribute-alone as input.

Furthermore, we also evaluate how the various inputs contribute to the next question prediction and next attribute prediction subtasks in Figures 3(b) and (c). We find that user-readable dialogue is

Tasks	Benchmarks	Cellphones & Accessories			Grocery & Gourmet			Toys & Games			Automotive		
Next Attribute	Metrics	MAP	R_{10}@1	R_{10}@3	MAP	R_{10}@1	R_{10}@3	MAP	R_{10}@1	R_{10}@3	MAP	R_{10}@1	R_{10}@3
	Seq2Seq	0.612	0.430	0.738	0.707	0.544	0.845	0.593	0.431	0.674	0.701	0.547	0.817
	LSTM	0.642	0.465	0.772	0.726	0.569	0.859	0.566	0.408	0.626	0.659	0.499	0.768
	OpenDialKG	0.643	0.467	0.774	0.707	0.555	0.822	0.656	0.501	0.754	0.706	0.557	**0.838**
	HOOPS (Ours)	**0.688**	**0.528**	**0.810**	**0.789**	**0.655**	**0.917**	**0.705**	**0.561**	**0.806**	**0.712**	**0.564**	0.825
Next Question	Metrics	MAP	R_{10}@1	R_{10}@3	MAP	R_{10}@1	R_{10}@3	MAP	R_{10}@1	R_{10}@3	MAP	R_{10}@1	R_{10}@3
	DMN [40]	0.475	0.269	0.564	0.502	0.304	0.587	0.456	0.253	0.518	0.469	0.267	0.553
	DAM [48]	0.514	0.373	0.590	0.581	0.394	0.635	0.579	0.388	0.546	0.552	0.387	0.608
	MSN [42]	0.608	0.428	0.740	0.678	0.503	0.749	0.630	0.455	0.732	0.645	0.473	0.713
	OpenDialKG [24]	0.699	0.654	0.678	0.729	0.676	0.724	0.579	0.499	0.561	0.710	0.640	0.726
	KBRD [7]	0.669	0.498	0.771	0.768	0.626	**0.896**	0.688	0.559	**0.760**	0.711	0.552	0.809
	HOOPS (ours)	**0.781**	**0.718**	**0.788**	**0.854**	**0.812**	0.859	**0.693**	**0.562**	0.746	**0.850**	**0.805**	**0.858**
Recommend	Metrics	NDCG	Recall	MAP	NDCG	Recall	MAP	NDCG	Recall	MAP	NDCG	Recall	MAP
	BPR [29]	0.349	0.540	0.336	0.331	0.521	0.360	0.305	0.498	0.335	0.307	0.487	0.312
	CKE [43]	0.360	0.543	0.303	0.411	0.598	0.353	0.435	0.636	0.372	0.385	0.570	0.327
	RippleNet [33]	0.326	0.476	0.279	0.366	0.534	0.314	0.420	0.612	0.361	–	–	–
	HeteroEmbed [1]	0.388	0.583	0.327	0.439	0.637	0.377	0.467	0.654	0.409	0.395	0.598	0.335
	KGAT [35]	0.399	0.593	0.338	0.424	0.622	0.363	0.443	0.637	0.386	0.387	0.581	0.326
	KBRD [7]	0.253	0.424	0.201	0.293	0.475	0.237	0.210	0.366	0.162	0.249	0.409	0.200
	HOOPS (ours)	**0.405**	**0.611**	**0.341**	**0.449**	**0.650**	**0.386**	**0.477**	**0.668**	**0.418**	**0.403**	**0.605**	**0.341**

Table 2: Performance of selected baselines and our benchmark methods on four proposed sub-datasets. The best results are highlighted in bold and the second best results are underlined.



(a) NDCG (b) Recall$_{10}$@1 (c) Recall$_{10}$@1

Figure 3: Comparison of inputs for (a) recommendation, (b) next question prediction, and (c) next attribute prediction.

more useful than merely considering the attributes for the question prediction task. Interestingly, there is a small performance gap for next attribute prediction. This is also because the utterance incorporates the descriptive attribute information, while attribute-alone loses semantic information content. Thus, utterance-alone is better than attribute-alone on question prediction, but fairly similar on the attribute prediction sub-task.

6 RELATED WORK

There has been significant research in human-centered AI. Much of it has focused on societal goals rather than individual human needs and interests [16, 17, 21]. Recently, some progress has been made in the HCI field towards invoking ML to augment interactive and intelligent systems [2, 41]. In this regard, the notion of Human-in-the-Loop (HitL) AI has been proposed. We propose a concrete HiTL graph reasoning framework for conversational recommendation. At the same time, the integration of knowledge graphs [20] has enabled CRS models to make recommendations grounded in knowledge-driven reasoning [11, 18, 24, 37, 38]. For example, Lei et al. [18] propose an RL-based mechanism based on an interactive path reasoning algorithm. However, the lack of human-readable fluent utterances is replaced by crawling the attribute words from raw review contexts, which is less practical in real-world scenarios. In Chen et al. [7], item-related knowledge bases with entity-linked text leads to better performance than either of them alone in dialogue generation and recommendation. Comparing to these methods, we provide an open dataset for conversational recommendation that supports the HiTL graph reasoning paradigm and integrates knowledge graphs so that prominent knowledge with semantics can be used to consider user-involved feedback and provide transparent recommendations. Except for conversational recommendation, the dataset may also be used for conversational search [3], conversational QA [28] and Explainable Recommendation [8, 9, 19, 44, 46]. Since reasoning on graphs naturally provides transparency of the decision making process, it helps to provide explanations for users over the recommended items [1, 11, 38, 39].

7 CONCLUSION

Our work in this paper is the first exploration of human-in-the-loop (HitL) learning for recommendation. Specifically, we define a new HitL graph reasoning paradigm with the three properties of hybrid integration, coarse-to-fine resolution, and a transparent decision-making process. We instantiate the paradigm for the conversational recommendation problem, where the system can leverage interactive user feedback to shrink the large search space during the multi-step reasoning process. Accordingly, we construct a new dataset called HOOPS including a graph that structurally integrates diverse user behavior and item-related information, as well as a multi-round conversation corpus that simulates user–agent interaction. We also provide a benchmark model to approach the HitL graph reasoning for recommendation with reported performance in three tasks on the constructed dataset. We hope it opens up avenues for further research on more realistic applications for Human-in-the-Loop learning. All data and code are freely available under a CC-BY-SA license.[1]

ACKNOWLEDGEMENT

This work was supported in part by NSF IIS-1910154 and IIS-2007907. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

[1] https://github.com/zuohuif/HOOPS

REFERENCES

[1] Q. Ai, V. Azizi, X. Chen, and Y. Zhang. 2018. Learning heterogeneous knowledge base embeddings for explainable recommendation. *Algorithms* (2018).

[2] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. Bennett, K. Quinn, J. Teevan, R. Kikin-Gil, and E. Horvitz. 2019. Guidelines for Human-AI Interaction. *CHI* (2019).

[3] Keping Bi, Qingyao Ai, Yongfeng Zhang, and W Bruce Croft. 2019. Conversational product search based on negative feedback. In *CIKM*.

[4] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *NeurIPS*.

[5] P. Budzianowski, T. Wen, B. Tseng, I. Casanueva, S. Ultes, O. Ramadan, and M. Gasic. 2018. MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *EMNLP*.

[6] Hanxiong Chen, Shaoyun Shi, Yunqi Li, and Yongfeng Zhang. 2021. Neural Collaborative Reasoning. *WWW* (2021).

[7] Q. Chen, J. Lin, Y. Zhang, M. Ding, Y. Cen, H. Yang, and J. Tang. 2019. Towards Knowledge-Based Recommender Dialog System. *ArXiv 1908.05391* (2019).

[8] Xu Chen, Hanxiong Chen, Hongteng Xu, Yongfeng Zhang, Yixin Cao, Zheng Qin, and Hongyuan Zha. 2019. Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In *SIGIR*.

[9] Xu Chen, Yongfeng Zhang, and Zheng Qin. 2019. Dynamic explainable recommendation based on neural attentive models. In *AAAI*.

[10] M. Eric, L. Krishnan, F. Charette, and C. D. Manning. 2017. Key-Value Retrieval Networks for Task-Oriented Dialogue. In *SIGDIAL*.

[11] Zuohui Fu, Yikun Xian, Ruoyuan Gao, Jieyu Zhao, Qiaoying Huang, Yingqiang Ge, Shuyuan Xu, Shijie Geng, Chirag Shah, Yongfeng Zhang, et al. 2020. Fairness-aware explainable recommendation over knowledge graphs. In *SIGIR*. 69–78.

[12] Zuohui Fu, Yikun Xian, Yongfeng Zhang, and Yi Zhang. 2020. Tutorial on Conversational Recommendation Systems. In *RecSys*.

[13] Zuohui Fu, Yikun Xian, Yongfeng Zhang, and Yi Zhang. 2021. IUI 2021 Tutorial on Conversational Recommendation Systems. In *IUI*.

[14] Zuohui Fu, Yikun Xian, Yongfeng Zhang, and Yi Zhang. 2021. WSDM 2021 Tutorial on Conversational Recommendation Systems. In *WSDM*.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.

[16] A. Jaimes, D. G. Perez, N. Sebe, and T. Huang. 2007. Guest Editors' Introduction: Human-Centered Computing–Toward a Human Revolution. *Computer* (2007).

[17] R. Kling and S. L. Star. 1998. Human centered systems in the perspective of organizational and social informatics. *SIGCAS Comput. Soc.* 28 (1998), 22–29.

[18] Wenqiang Lei, Gangyi Zhang, Xiangnan He, Yisong Miao, Xiang Wang, Liang Chen, and Tat-Seng Chua. 2020. Interactive Path Reasoning on Graph for Conversational Recommendation. In *KDD*.

[19] Lei Li, Yongfeng Zhang, and Li Chen. 2020. Generate neural template explanations for recommendation. In *CIKM*.

[20] Zelong Li, Jianchao Ji, Zuohui Fu, Yingqiang Ge, Shuyuan Xu, Chong Chen, and Yongfeng Zhang. 2021. Efficient Non-Sampling Knowledge Graph Embedding. *WWW* (2021).

[21] Zhenyu Liao, Yikun Xian, Xiao Yang, Qinpei Zhao, Chenxi Zhang, and Jiangfeng Li. 2018. TSCSet: A crowdsourced time-sync comment dataset for exploration of user experience improvement. In *IUI*. 641–652.

[22] Michael E. J. Masson. 1983. Conceptual processing of text during skimming and rapid sequential reading. *Memory & Cognition* 11 (1983), 262–274.

[23] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NeurIPS*.

[24] Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *ACL*.

[25] Mark Newman. 2010. Networks: An Introduction.

[26] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In *EMNLP*.

[27] Namyong Park, Andrey Kan, Xin Dong, Tong Ke Zhao, and Christos Faloutsos. 2019. Estimating Node Importance in Knowledge Graphs Using Graph Neural Networks. *KDD* (2019).

[28] Chen Qu, Liu Yang, Minghui Qiu, Yongfeng Zhang, Cen Chen, W Bruce Croft, and Mohit Iyyer. 2019. Attentive history selection for conversational question answering. In *CIKM*.

[29] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *UAI*.

[30] Shaoyun Shi, Hanxiong Chen, Weizhi Ma, Jiaxin Mao, Min Zhang, and Yongfeng Zhang. 2020. Neural Logic Reasoning. In *CIKM*. 1365–1374.

[31] Yueming Sun and Yi Zhang. 2018. Conversational Recommender System. In *SIGIR '18*.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention is All you Need. In *NIPS*.

[33] Hongwei Wang, Fuzheng Zhang, Jialin Wang, Miao Zhao, Wenjie Li, Xing Xie, and Minyi Guo. 2018. Ripplenet: Propagating user preferences on the knowledge graph for recommender systems. In *CIKM*. ACM, 417–426.

[34] Hongwei Wang, Miao Zhao, Xing Xie, Wenjie Li, and Minyi Guo. 2019. Knowledge Graph Convolutional Networks for Recommender Systems. In *WWW '19*.

[35] Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. 2019. KGAT: Knowledge Graph Attention Network for Recommendation. In *KDD*.

[36] Sam Wiseman, Stuart M. Shieber, and Alexander M. Rush. 2018. Learning Neural Templates for Text Generation. In *EMNLP*.

[37] Yikun Xian, Zuohui Fu, Qiaoying Huang, Shan Muthukrishnan, and Yongfeng Zhang. 2020. Neural-Symbolic Reasoning over Knowledge Graph for Multi-Stage Explainable Recommendation. *arXiv preprint arXiv:2007.13207* (2020).

[38] Yikun Xian, Zuohui Fu, S. Muthukrishnan, Gerard de Melo, and Yongfeng Zhang. 2019. Reinforcement Knowledge Graph Reasoning for Explainable Recommendation. *SIGIR* (2019).

[39] Yikun Xian, Zuohui Fu, Handong Zhao, Yingqiang Ge, Xu Chen, Qiaoying Huang, Shijie Geng, Zhou Qin, Gerard De Melo, Shan Muthukrishnan, et al. 2020. CAFE: Coarse-to-fine neural symbolic reasoning for explainable recommendation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 1645–1654.

[40] L. Yang, M. Qiu, C. Qu, J. Guo, Y. Zhang, W. B. Croft, J. Huang, and H. Chen. 2018. Response Ranking with Deep Matching Networks and External Knowledge in Information-seeking Conversation Systems. *SIGIR* (2018).

[41] Qian Yang, Nikola Banovic, and John Zimmerman. 2018. Mapping Machine Learning Advances from HCI Research to Reveal Starting Places for Design Innovation. *CHI* (2018).

[42] Chunyuan Yuan, Wen jie Zhou, MingMing Li, Shangwen Lv, Fuqing Zhu, Jizhong Han, and Songlin Hu. 2019. Multi-hop Selector Network for Multi-turn Response Selection in Retrieval-based Chatbots. In *EMNLP/IJCNLP*.

[43] Fuzheng Zhang, Nicholas Jing Yuan, Defu Lian, Xing Xie, and Wei-Ying Ma. 2016. Collaborative Knowledge Base Embedding for Recommender Systems. In *KDD*.

[44] Yongfeng Zhang and Xu Chen. 2018. Explainable Recommendation: A Survey and New Perspectives. *Found. Trends Inf. Retr.* 14 (2018), 1–101.

[45] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W Bruce Croft. 2018. Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 177–186.

[46] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 83–92.

[47] Wayne Xin Zhao, Gaole He, Kunlin Yang, Hongjian Dou, Jin Huang, Siqi Ouyang, and Ji-Rong Wen. 2019. KB4Rec: A Data Set for Linking Knowledge Bases with Recommender Systems. *Data Intelligence* 1, 2 (2019), 121–136.

[48] Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018. Multi-Turn Response Selection for Chatbots with Deep Attention Matching Network. In *ACL*.

[49] Yaxin Zhu, Yikun Xian, Zuohui Fu, Gerard de Melo, and Yongfeng Zhang. 2021. Faithfully Explainable Recommendation via Neural Logic Reasoning. *NAACL* (2021).