ChatShop: Interactive Information Seeking with Language Agents

Sanxing Chen **Sam Wiseman** **Bhuwan Dhingra**
Duke University
sanxing.chen@duke.edu
{swiseman, bdhingra}@cs.duke.edu

Abstract

The desire and ability to seek new information strategically are fundamental to human learning but often overlooked in current language agent development. Using a web shopping task as an example, we show that it can be reformulated and solved as a retrieval task without a requirement of interactive information seeking. We then redesign the task to introduce a new role of shopper, serving as a realistically constrained communication channel. The agents in our proposed ChatShop task explore user preferences in open-ended conversation to make informed decisions. Our experiments demonstrate that the proposed task can effectively evaluate the agent's ability to explore and gradually accumulate information through multi-turn interaction. We also show that LLM-simulated shoppers serve as a good proxy to real human shoppers and discover similar error patterns of agents.

1 Introduction

Recent studies have explored Large Language Models (LLMs) as autonomous agents in general problem solving (Zhou et al., 2023; Liu et al., 2023; Xie et al., 2023). In their design, the component of information seeking is often against a static information source such as a knowledge graph or a pile of web documents. The unconstrained access to the information source reduces *interactivity*—the agent does not need to strategically seek new information from the user and its decision-making process is not informed by tracking accumulated information. To investigate this issue, we first examine Webshop (Yao et al., 2022), which evaluates an agent's ability to identify the correct product in a web shopping interface based on user instructions. For this task, we show that since the shopper provides full details of the target product in their instruction, a retrieval system can directly score the relevance of each product to locate the target.

In a realistic scenario, a shopper would start with partial information of the target product which would become clearer after seeing various options the agent might find. The key challenge with designing such a setup is that interactions between the agent and the shopper would require a human-in-the-loop environment, hindering scalable evaluation. Given the strong performance of recent LLM agents, we hypothesize that LLMs themselves would be capable of simulating humans in an interactive web shopping experience (Li et al., 2023b; Park et al., 2023). To test this hypothesis, we repurpose WebShop to propose ChatShop, in which the agent starts with an unspecific goal instruction—only the coarse type of product. The lack of specificity in the instruction creates a challenge of task ambiguity (Tamkin et al., 2023), which can only be resolved by effectively gathering information from the shopper and the website environment about products (Figure 1). The challenge is amplified by other inherent complexities

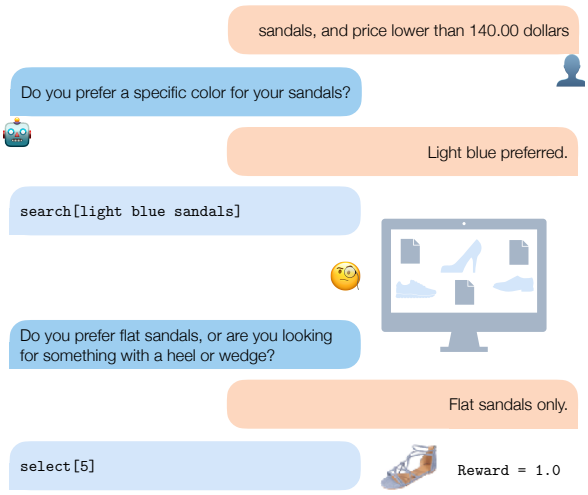

Figure 1: An illustration of the ChatShop task. In contrast to WebShop's original detailed instruction *a non-slip sandals for my wife that is blue in color, 5.5 size*, we provide only the product type (price) and require the agent to narrow down the search space by interacting with the shopper and product database.

such as searching the vast product space and tool usage.

We benchmark a range of agents with both GPT-3.5/4 and a Llama 2 variant as base models in environments where the role of the shopper is played by humans or LLMs. Experimental results verify the challenges introduced by the information need. We further evaluate how good an LLM at simulating the interaction with real human shoppers in a human study. The benchmarking results and the failure patterns show that the LLM simulated environment is as effective in recovering the gap between agents. We hope our work can drive the automatic evaluation of language agents towards more complex and meaningful interactions with (simulated) humans.

2 Related Work

Information Seeking Tasks Language agents' information-seeking ability has long been a focus of AI research, especially in the context of question answering and task-oriented dialogue (Bachman et al., 2016; Dhingra et al., 2017; Zamani et al., 2022; Zhou et al., 2023). In such tasks, the agent usually receives an information need from the user and accesses external knowledge sources to gather information, a task which can often be formulated as a single-turn retrieval problem. The constraints of such interaction are often artificial (Yuan et al., 2020). In contrast, the constraints in ChatShop task originate from a realistic situation of a human party in a web shopping scenario.

Human-AI Collaboration More recently, there has been a growing interest in studying human-AI collaboration via LLMs. MINT (Wang et al., 2023) benchmarks a range of LLM agents in leveraging human or AI-simulated feedback to improve multi-turn problem solving. Unlike ChatShop, this feedback can be viewed as a form of natural language supervision, which is beneficial but not required to solve the task. DialOp (Lin et al., 2023) focuses on the agent's ability of planning based on human preferences in a grounded dialogue setting. Compared to ChatShop, the tasks in DialOp has a narrower and synthetic search space. Li et al. (2023a) propose a learning framework for LLMs to elicit human preferences in tasks such as content recommendation, however, their tasks focus on exploration guided by the general world knowledge stored in the LLM weights internally, whereas in ChatShop, the exploration is grounded in an external real-world product space.

References

Philip Bachman, Alessandro Sordoni, and Adam Trischler. 2016. Towards information-seeking agents. *arXiv preprint arXiv:1612.02605*.

Bhuwan Dhingra, Lihong Li, Xiujun Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmed, and Li Deng. 2017. Towards end-to-end reinforcement learning of dialogue agents for information access. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–495, Vancouver, Canada. Association for Computational Linguistics.

Belinda Z. Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. 2023a. Eliciting human preferences with language models. *arXiv preprint arXiv: 2310.11589*.

Yuan Li, Yixuan Zhang, and Lichao Sun. 2023b. Metaagents: Simulating interactions of human behaviors for llm-based task-oriented coordination via collaborative generative agents. *arXiv preprint arXiv:2310.06500*.

Jessy Lin, Nicholas Tomlin, Jacob Andreas, and Jason Eisner. 2023. Decision-oriented dialogue for human-ai collaboration. *arXiv preprint arXiv: 2305.20076*.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. 2023. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv: 2308.03688*.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

Alex Tamkin, Kunal Handa, Avash Shrestha, and Noah D. Goodman. 2023. Task ambiguity in humans and language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. 2023. Mint: Evaluating llms in multi-turn interaction with tools and language feedback. *arXiv preprint arXiv: 2309.10691*.

Tianbao Xie, Fan Zhou, Zhoujun Cheng, Peng Shi, Luoxuan Weng, Yitao Liu, Toh Jing Hua, Junning Zhao, Qian Liu, Che Liu, Leo Z. Liu, Yiheng Xu, Hongjin Su, Dongchan Shin, Caiming Xiong, and Tao Yu. 2023. Openagents: An open platform for language agents in the wild. *arXiv preprint arXiv: 2310.10634*.

Shunyu Yao, Howard Chen, John Yang, and Karthik R Narasimhan. 2022. Webshop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems*.

Xingdi Yuan, Jie Fu, Marc-Alexandre Côté, Yi Tay, Chris Pal, and Adam Trischler. 2020. Interactive machine comprehension with information seeking agents. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2325–2338, Online. Association for Computational Linguistics.

Hamed Zamani, Johanne R Trippas, Jeff Dalton, and Filip Radlinski. 2022. Conversational information seeking. *arXiv preprint arXiv:2201.08808*.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2023. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv: 2307.13854*.