Please quote as: Göldi, A. & Rietsche, R. (2024). Insert-expansions for Large Language Model Agents. Americas Conference on Information Systems (AMCIS), .

August 2024

Insert-expansions for Large Language Model Agents

Andreas Göldi
University of St.Gallen, andreas.goeldi@unisg.ch

Roman Rietsche
Institute for Digital Technology Management, roman.rietsche@bfh.ch

Insert-expansions for Large Language Model Agents

Completed Research Full Paper

Andreas Göldi
University of St.Gallen
andreas.goeldi@unsig.ch

Roman Rietsche
Bern University of Applied Sciences
roman.rietsche@bfh.ch

Abstract

AI agents use language generation for internal reasoning and interacting with users. Agents often employ tools beyond language generation, such as calculators or search, to further augment these capabilities. We focus on how such tools can give the agent too much external context, diverting it from the user's original intent. According to Conversation Analysis, human-human dialogue often uses "insert-expansion" - inserted utterances for clarification - to resolve ambiguities. Building on this, we introduce a "user-as-a-tool" approach, enabling the AI agent to solicit clarification from the user while still reasoning, thereby realigning it with the user's intent. Initial evidence shows that our approach has benefits for conversational recommendation systems. We present a novel interaction method and empirical findings that enhance the user's role in guiding agent reasoning. This research is especially relevant as AI agents become increasingly common, and holds significance for optimizing the human-chatbot interaction loop.

Keywords

LLM, Agent, Sequence Organisation.

Introduction

The emergence of AI agent platforms, such as those by OpenAI, has democratized the use of AI for tasks ranging from brainstorming to service management. Powered by Large Language Models (LLMs), these agents simulate human reasoning in iterative calls, by generating intermediate natural language steps. Here, external tools like search can gather additional information to help prompt effective final responses. However, this information gathering can introduce delays and make the agents' output deviate from user intentions, posing alignment challenges.

Addressing these complexities, this paper suggests implementing a technique known in human-to-human dialogue, namely insert-expansion (Schegloff 2007). Namely, humans often clarify intent before providing their final answer. Considering the user as yet another tool for the agent, the agent can gather information about their intent just as they gather other relevant information during the intermediate reasoning. This is meant to help the agent retain focus on user intent. To investigate this, we ask:

RQ: If Large Language Model-based AI agents use external tools to gather information before responding with a final answer, can insert-expansion prevent misalignment to user intent during this information-gathering?

This is particularly pertinent given the challenges AI faces in maintaining user intent as the central focus of interaction (Hornbæk and Oulasvirta 2017). By actively seeking user input in scenarios of ambiguity or when additional context is required, the AI agent's conversation becomes more dynamic and user-centered. This approach aligns with the principles of human-AI teaming and is crucial for tasks where human-in-the-loop systems offer significant benefits (De-Arteaga et al. 2020). The paper presents an approach for implementing this approach in tool-enabled AI agents. We call the approach 'user-as-a-tool', as the user informs about their intent like another agent tool would about other relevant information. We investigate the approach in a pilot and two empirical studies. We do not claim causal evidence, as these are not randomized experiments. Instead, we aim to demonstrate plausibility and stimulate future work. Our theoretical contribution lies in advancing the understanding of AI-agent interactions, particularly how incorporating human conver-

sational patterns, such as insert-expansions (Schegloff 2007), can realign AI reasoning with user intent. We also contribute to the discourse on the efficacy of human-in-the-loop systems in AI interactions, highlighting the balance between AI autonomy and user guidance (De-Arteaga et al. 2020).

In conclusion, this paper introduces a significant advancement in AI-agent interaction by incorporating a human-like conversational pattern into the AI reasoning process. This approach not only bridges the gap between current AI capabilities and the nuanced dynamics of human conversation but also offers a practical framework for enhancing the interaction quality between users and AI agents. By keeping the user's intent at the forefront of AI interactions, this method offers a promising direction for future HCI research and development, emphasizing the importance of user-centric design in the evolving landscape of AI technology.

Background

Recent studies in human-computer interaction (HCI) emphasize the application of human-to-human interaction principles in AI agent development. Key traits like mutual understanding and trust, essential in human interactions, are found to be equally valued in human-agent conversations, though trust in agents also depends on their observed performance (Clark et al. 2019; Yin et al. 2019). The concept of "repair" in human conversations, which deals with correcting misunderstandings or errors, has been extended to human-computer dialogues. This extension leads to the categorization of repairs into user- and bot-initiated types, demonstrating the adaptability of human interaction methodologies for enhancing human-agent communication (Ashktorab et al. 2019). There is a noted discrepancy between user expectations and the actual capabilities of conversational agents, highlighting the necessity for more human-like interactions. Studies show that users often approach prompt design in a manner similar to end-user programming and interactive machine learning, facing comparable challenges (Zamfirescu-Pereira et al. 2023).

Insert-Expansion

The organization of dialog in human interaction is a complex process that involves various elements to ensure smooth communication. A key aspect of this organization, as described by Schegloff 2007, is the use of "insert-expansions" in talk-in-interaction. These expansions are essential components in natural dialogs, allowing conversations to extend beyond basic adjacency pairs. An adjacency pair is a structure where an initial utterance prompts one of four types of responses: preferred (e.g., acceptance), dispreferred (e.g., rejection), temporizing (e.g., maybe), or blocking (e.g., counter or change of subject). Insert-expansions can occur as pre-extensions to draw attention or gauge interest, or as post-first insert expansions to recover from misunderstandings and clarify intent. This structure is not only present in natural spoken dialog but also extends to text-based communication, highlighting the importance of replicating these patterns in digital interactions for more natural and effective communication, especially in the context of human-chatbot interactions. In our view, integrating 'insert-expansion' techniques in Large Language Model Agents is crucial for enhancing their dialog management. It has been discussed in the HCI literature (Fast et al. 2018), and recently been applied to inspire interaction patterns in code debugging (Chopra et al. 2024). By adopting insert-expansions, agents can more accurately interpret and respond to user inputs, making their assistance more effective and user-centric.

Language Models Can Mimic Both Speech and Reasoning

Human language serves both for communication and thinking (Kompa 2023). Thanks to advances like deep learning (LeCun et al. 2015), it's now easier to handle language with computers. These models use layered artificial neurons to map, say, text to labels (Hornik et al. 1989). With modern architecture (Vaswani et al. 2017), a focus on text generation (Radford et al. 2018), and bigger models (Brown et al. n.d.), we are now getting human-level results in language tasks. These models not only meet human expectations (Christiano et al. 2017), but they are also good at non-language tasks like taking tests (Bubeck et al. 2023). Both everyday users and researchers are finding value in these advances (Liu et al. 2023).

Large Language Model Agents Integrate Reasoning and Tools Problem-Solving

These advancements owe a lot to models that tackle tasks in a step-by-step manner. The concept of breaking down tasks is not new; it dates back to Cartesius (Dika 2023). By chaining simple tasks together, language models can solve complex tasks more effectively (Wu et al. 2022). This allows us to consult a language model multiple times until we get a final answer. Humans often think through writing (Applebee 1984), and language models simulate this by generating text step-by-step. Given the allowance for intermediate steps, it has become viable to integrate external tools like search engines or calculators into the process (Parisi et al. 2022). This makes language models more than just simple predictors: they become augmented language models (Mialon et al. 2023) or Large Language Model based agents Park et al. 2023, thus mimicking key human traits like reasoning and tool-use (Oakley 1957).

Large Language Model Agents Can Get Off Track If Stuck Too Long in Reasoning Loops

Large Language Model Agents aim to give answers, but their multi-step reasoning can go off course and stray from the user's original question. Each step is based on the last, so if they don't find an answer quickly, they can get increasingly off track. This phenomenon has now been observed by researchers (Yao et al. 2023) and even government agencies (Government Office for Science 2023).

In Human-to-Human Dialogue, This is Solved via Insert-Expansions

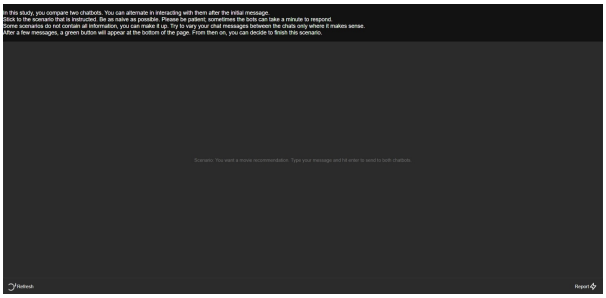
People often use insert-expansion in dialogues for better decision-making (Schegloff 2007). For instance, a buyer might ask a seller about the price, adding a step to the conversation for more informed choices. This pattern is common across languages (Kendrick et al. 2020), text chats (Tudini 2015), and is under continuous study (Meredith 2019). It's useful for realigning discussions to the main topic and shared goals.

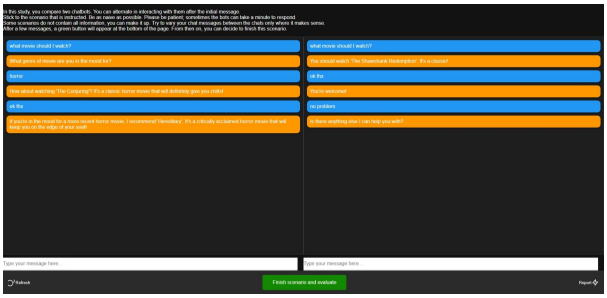
Methodology

The purpose of our research was to compare conversational agents, both with and without insert-expansion capability. Lacking previous studies on this effect, we had participants evaluate both types simultaneously to avoid repeated interactions. We employed brief scenarios to control usage variability and simplify the study design, enabling smaller sample sizes without the need for balancing. The study's design yields correlational results, suitable for preliminary guidance in this emerging field, rather than causal conclusions.

English-speaking participants from prolific.com were recruited for three studies: a pilot with 10 participants (average age 31.50, 60% female), Study 1 with 71 participants (average age 31.45, 27% female), and Study 2 with 36 participants (average age 32.22, 31% female, 1 other), excluding two for inconsistencies. Sample sizes were based on power estimations for detecting medium to large effects. The study's focus was a chat interface with two conversational agents, comparing their performance and discussing results from the pilot and both studies sequentially. In this study, we compared two augmented language model conversational agents: a standard agent ("standard bot") and an enabled agent with user-as-a-tool features ("enabled bot"), both utilizing OpenAI's gpt-3.5-turbo-0301 and external tools like Wikipedia queries, simple mathematics, and an embedding-based PDF reader. The setup involved users interacting with both bots in a chat interface, as shown in Figure 1a, where they initiated conversations and compared the bots' responses. Participants engaged with each bot for a set number of messages before evaluating their performance, as depicted in Figure 1c (Schegloff 2007).

For the comparison, we used bipolar rating scales, with the placement (left, right) of the two bots as anchors (See Figure 1c). We deemed this appropriate as the placement constitutes a definite reference without biasing results by naming the bots. We visually indicated the middle of the scale with a small vertical line; however, we did not label intermediary steps so as to not jeopardize equidistance Höhne et al. 2021. In Study 1, we allowed no-preference, in Study 2, we used a forced-choice procedure by eliminating the midpoint. The former approach does not presume a preference, and may thus be more appropriate as a measure of spontaneous preference; the latter forces participants to decide, and thus evokes cognition about preferences, which may lead to preference formation even if none was spontaneously available.





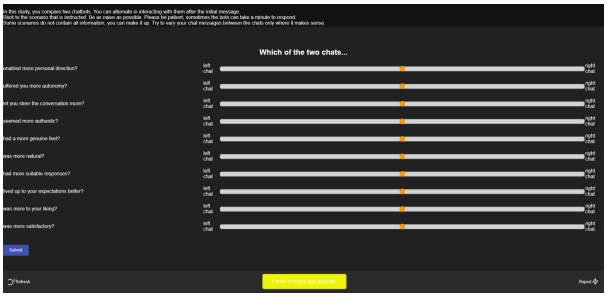


Figure 1a: Before seeing the two separate chat interfaces, participants sent one query to both chatbots.
Figure 1b: Example chat (scenario 1 is depicted). On the left side, insert expansion is enabled, the right side is a standard chat. This was held constant throughout the study.
Figure 1c: Bipolar rating scales after each scenario (Study 1: 7 points; Study 2: 6 points).

Pilot

The Bot Usability Scale (15-item version) was used to assess both standard and enabled bots Borsci et al. 2022. Results indicated minimal differences in usability scores (mean = 3.59, SD = 0.68 for standard bots; mean = 3.63, SD = 0.25 for enabled bots). In addition to usability testing, participants were asked for open-ended feedback, covering their perceptions of the study, interface, and comparative bot evaluations.

Study 1

The continuation of our research involved addressing issues identified in the pilot study, involving 71 participants. We introduced 7-level bipolar rating scales for a direct comparison of user experiences and also assessed potential moderators like the 15-item Big Five Inventory-2 Short (BFI-2-XS) (Soto and John 2017) and assessed the need for cognitive closure (NFC-15) (Roets and van Hiel 2011). The focus was on participants' perceived control, the naturalness of interactions, fulfillment of intent, and overall satisfaction (see Table with reliabilities of items at https://osf.io/gvjq5/?view_only=8f9ce8b6f3fa402e995dd31b2305583e). After a short demographic questionnaire, n=71 participants were given 3 scenarios (see Figure 2), rating each before turning to the next. Having completed all scenarios and evaluations, feedback was elicited and interindividual variables were assessed.

The distribution of control ratings was normal; of naturalness normal with noticeable but statistically acceptable kurtosis (-0.88); of intent-effectiveness seemingly bimodal, with a population at the extreme end of the enabled bot, however, with the main distribution at an offset towards the standard bot; and satisfaction almost uniform, or potentially multimodal as well, meaning normality assumptions were not fulfilled here.

The interindividual variables all approximated normal distributions, with expected ranges. Notably, openness, agreeableness, and conscientiousness were relatively high (out of 5) with 3.76 (SD=.79) 3.75 (SD=.87) 3.78 (SD=.71), cutting off the extreme positive tail. In terms of significance, only control in scenario 1 differed between the two bots (the difference in satisfaction for scenario 1 was marginal with p=0.05985, as tested with a Wilcoxon Signed Rank test). Control in scenario 1 is normally distributed; and a difference was observable in a one-sample, one-tailed t-test (t=-1.8494, df=70, p=0.03431). In the other situations, control was not significantly different from 0, although the tendency persists. This is because the effect size seems to be small (for this t-test, Cohen's d is d=.22). In the other two scenarios, no significant differences were observed. Overall preference (treating all scales as one factor) is marginally not significant with V=827,

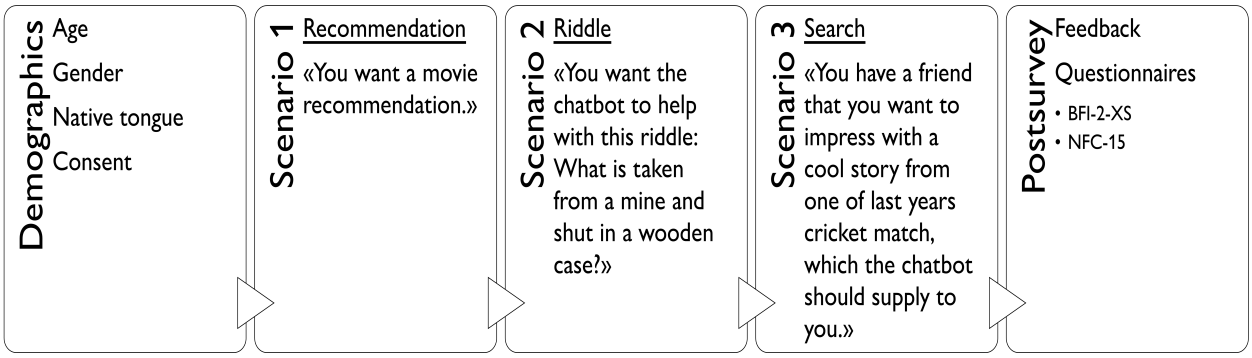


Figure 2. Study 1 procedure.

p=0.05463 in a Wilcoxon Signed Rank test. There is visible clustering in the scenarios for the direct comparison measures. For scenario 2, where the bots were supposed to help with solving a riddle, openness seemed to be important across ratings. However, most associations do not exceed τ=.2. To quantify the feedback, we repeated the procedure described for the pilot. Feedback on the study was 69.01% positive (neutral opinions tend to be rated as negative as well, e.g. ”I have no opinion about the study.”), on the interface only 52.11% (negatives include no feedback at all, neutral statements such as ”It was fast and responsive, just feel like the loading is too big and the lettering also”, but also some on bugs like ”It was fine, although sometimes my prompt would trigger a loading animation that the bot would never reply to, so I had to prompt again, which left the loading anim on the screen for one of the bots but not the other, not a big deal.” or ”It was frusting when it could not listen or answer all my question”). 46.48% of participants preferred the enabled bot (16.90% preferred standard; 2.82% neutral, and the rest unclear). While there was a tendency to see the enabled bot [i.e, the left bot] as more personable, opinions differed, for example, one participant stated ”The chatbot on the left had a more personal touch to it. The one on the right felt more like a robot and more focused on facts.” but another ”The one on the left was robot-like and the one on the right was very human”. This goes for overall preference too: ”Not sure but I felt the first one was more helpful.” versus ”I think the right one was more intelligent, I mean, it gave better responses”. Other participants gave feedback that nicely puts the two options into contrast, such as ”Both are good, but I think the left one is more cautious, and the one on the right answers faster.”, ”The chat on the left was a lot quicker and asked more questions in terms of my desired answer. The chat on the right was more strict to the point.”. So, in summary, adding insert-expansion capability does not automatically make conversational agents superior, especially as the standard bot is already trained to fulfill social expectations. Even if this imitation is only on the surface, it is apparently enough for many users.

In summary,we found that adding insert-expansions to conversational agents varies in impact across scenarios. Internal Consistency Coefficients (ICCs) were generally low in intent-effectiveness and satisfaction, but control and naturalness were more consistent. The only significant difference was in control for a recommendation scenario (Scenario 1), possibly due to diverse user preferences. These preferences, influenced by individual characteristics or specific conversation outcomes, might explain the bimodal distributions in intent-effectiveness and satisfaction. Satisfaction could even show a trimodal pattern in larger samples, reflecting different preferences for the enabled or standard bots. This variation is linked to the efficacy of insert-expansions in conversational repair and expectation management (Schegloff 2007).

Study 2

As the necessity of adding insert-expansions depends on scenarios, we added a new scenario for a follow-up study. We chose: ”You want to work on the most important sustainable development goal in the 2022 UN report but do not know which it is. Type your message and hit enter to send to both chatbots.” In this scenario, both bots had an information retrieval tool and access to the United Nations annual report of 2022

(which is after the training data of the underlying model ended). They were prompted to ask for a grounded recommendation, as Study 1 indicated recommendation as a relevant domain to introduce insert-expansion capabilities. We reduced the number of insert-expansion prompts for the user-as-a-tool tool in the enabled bot, only retaining the scope_response tool.

We retained all measures from Study 1. However, we exchanged the 7-level bipolar rating with midpoint for a 6-level forced-choice bipolar rating. In this, the midpoint was still visually indicated; however, participants had to decide whether they would rather indicate preference for the enabled or standard bot. We used the forced-choice approach because in Study 1, many ratings tended towards the midpoint, meaning that only some participants spontaneously formed a preference; however, formed preferences is what we were interested in, and therefore, we elicited preference formation via forced choice. The procedure remained the same as in Study 1, with the exception that only one scenario was given (for more details, see Figure 3). Both study and interface feedback was positive with 66.67% each. 36.11% preferred the enabled version, 27.77% standard, while 13.88% were neutral, and the rest of the preferences unclear. Qualitatively, the feedback largely conformed with the one to Study 1. The smaller sample size leads to less clean normal distributions, rendering non-parametric tests safer. We can still observe potential bimodality in intent-effectiveness and satisfaction, specifically regarding clumping at the extreme end of preference for the enabled bot, which is out of the normal distribution. Extreme preference for the standard bot may have been reduced by the choice of scenario. In this study, we observed significant differences from null preference in control and naturalness, both towards the enabled bot (using Wilcoxon Signed Rank tests, for control: V=217, p=0.03427; for naturalness: V=208, p=0.02484). For intent-effectiveness and satisfaction, effects were non-significant (V=271, p=0.3266 and V=299, p=0.2976, respectively), with a tendency towards the standard bot. Overall preference (treating all scales as one factor) is, as in Study 1, not significantly different from null preference (V=216, p=0.08296).

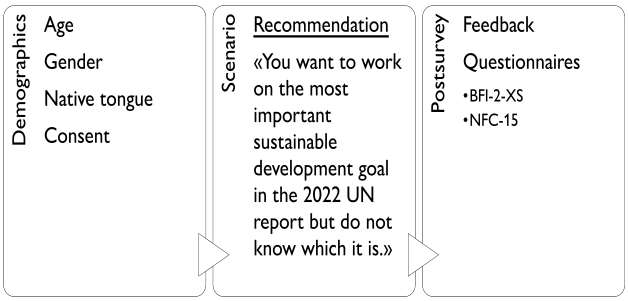


Figure 3. Study 2 procedure.

Study 2 confirmed a preference for insert-expansions in tool-enabled conversational agents, specifically in recommendation contexts. It also found a significant difference in naturalness, expanding on the marginal difference observed in Study 1. Despite this, heterogeneous user preferences were still evident. The mediators examined showed only small effect sizes, echoing the findings of Study 1. Incorporating interindividual variables into complex models of preference was not effective, as the minor associations needed larger sample sizes to yield meaningful results.

General Discussion

The emerging technology of Large Language Model agents is about to automate many tasks; however, they tend to get stuck in their own reasoning or action loops Yao et al. 2023; new methods are to address this problem while reclaiming user control (Yang et al. 2020). We propose a ”user-as-a-tool” method, inspired by ”insert-expansion” conversational patterns (Schegloff 2007), allowing agents to request more context and stay aligned with user goals, the key interaction metric (Hornbæk and Oulasvirta 2017). Our pilot and two empirical studies tested this approach using direct comparison.

Our research indicates that the effectiveness of insert-expansions in user-agent interactions varies by con-

Table 1. Two example reasoning paths that have access to a pdf reading tool and the 2022 UN report on sustainable development goals: https://www.un.org/annualreport/files/2022/09/SG-Annual-Report-2022_E_1.pdf. The example illustrates how 'user-as-a-tool' can help to reign in tool-enabled language models, and provide more useful answers. Note that a user only sees the final answer (and the query "What is your field of work or area of interest?" in the user-as-a-tool version). The terms in bold are generated by the agents themselves.

Standard	User-as-a-tool
Entering new chain... I should check the UN annual report for 2022 to find the most important sustainable development goal. **Action**: UN info **Action Input**: Search for the UN annual report for 2022 **Observation**: I don't know the answer. **Thought**: I should try searching for the specific sustainable development goals mentioned in the report. **Action**: UN info **Action Input**: Search for sustainable development goals in the UN annual report for 2022 **Observation**: The annual report for 2022 mentions the sustainable development goals several times. It highlights the support provided to 162 countries and territories to rescue the goals, the delivery of $17 billion in operational activities for development, the facilitation of $95 million in trade investment deals through support to South-South agreements, the protection of 138 million workers from work-related deaths, injuries, and disease, the prevention of 77 million tons of CO2 emissions through clean energy initiatives, the support provided to 183 million children with access to remote learning, and the convening of a conference on sustainable transport as a means to achieve the sustainable development goals. **Thought**: Based on the information provided in the UN annual report for 2022, it is difficult to determine the most important sustainable development goal. However, some of the goals mentioned in the report include rescuing the goals, delivering operational activities for development, supporting South-South agreements, protecting workers, preventing CO2 emissions, providing access to remote learning, and promoting sustainable transport. It would be helpful to further analyze the report to identify the specific goals that are highlighted as the most important. **Final Answer**: The most important sustainable development goal in the UN annual report for 2022 is not specified. Further analysis of the report is needed to determine the most important goal. *Finished chain.*	*Entering new chain...* I need to find the UN annual report for 2022 to determine the most important sustainable development goal. **Action**: UN info **Action Input**: UN annual report 2022 **Observation**: I don't know the answer to the question as it is not provided in the given context. **Thought**: I need to use the scope_response tool to gather more information about the person's specific interests or field of work. **Action**: scope_response First pair part of insert expansion (bot query): **Action Input**: What is your field of work or area of interest? Second pair part of insert expansion (user input): **Observation**: Finance **Thought**: I can now provide a tailored answer based on the person's interest in finance. **Final Answer**: The most important sustainable development goal in the UN annual report 2022 that you could work on is Goal 8: Decent Work and Economic Growth. This goal focuses on promoting sustained, inclusive, and sustainable economic growth, full and productive employment, and decent work for all. As someone interested in finance, you could contribute to this goal by supporting initiatives that promote job creation, entrepreneurship, and financial inclusion. *Finished chain.*

text. Effective in recommendation settings, these expansions enhance user control and conversational naturalness but are less impactful in riddle-solving and conversational search, likely due to the need for extended interactions or suboptimal implementation. The variability in achieving user satisfaction suggests a requirement for agent designs that are either personalized or context-adaptive. The "user-as-a-tool" approach should only be deployed when it significantly outperforms standard agents, considering the additional cognitive load and time involved. Modern LLMs trained with human feedback may mimic insert-expansion without any agent additional implementation. However, they do not use insert-expansion during reasoning, as agents enabled with the user-as-a-tool approach do. This means that standard agents may return what really should be an insert-expansion as their final answer, thereby forgetting intermediate steps, and having to redo them. Our approach not only emulates the surface aspects of natural dialogue but also replicates its deeper mechanisms, embedding reasoning and anticipating further information needs, similar to human interactions (Schegloff 2007). This method may reduce the need for training agents on surface dialogue traits, potentially making interactive system development more efficient.

Deep learning employs biological exemplars, specifically the architecture and function of the brain, to construct complex models capable of sophisticated abstractions (LeCun et al. 2015). In contrast, instruction-tuning (Wei et al. 2021) and products like ChatGPT (Liu et al. 2023) utilize social exemplars, particularly social expectations, to generate accurate and context-aware responses. The next significant category of exemplars for artificial intelligence will likely be cognitive, concentrating on patterns of thought to improve machine reasoning. Crucially, these categories of exemplars are not isolated; they are interconnected. For example, Reinforcement Learning (RL) applies neural exemplars to implement social expectations effectively. Analogously, the "user-as-a-tool" approach we propose enhances agent reasoning by leveraging dialog structures as social exemplars. By demonstrating this analogy, we hope to stimulate research in how to improve imitation of cognition in AI agents by implementing social exemplars, such as sequence organization in dialogue.

Limitations and Future Research

While our study shows high reliability in its primary outcome measures, it falls short in assessing their convergent and discriminant validity; future research should aim to establish these forms of validity for a more nuanced interpretation of the results. We relied on the Direct Comparison methodology in one-sample cases, limiting our findings to associative rather than causal interpretations; alternative empirical designs could provide a more comprehensive understanding and even establish causality. We can also not rule out order-effects, since we did not randomize whether the chats were displayed to the right or left, and we did not randomize the order of scenarios. Another limitation is the observed bimodal distributions in intent-effectiveness and user satisfaction, which we did not fully explore. It points to a minority of participants benefiting immensely from the "user-as-a-tool" approach, while the majority does not, pointing towards either interindividual differences or, more likely, the conclusion that the "user-as-a-tool" approach should only be used if otherwise, problems are introduced into the dialogue with reasoning agents. Subsequent research should more systematically investigate the root causes behind this bimodality to better identify which users or contexts would most benefit from insert-expansion capabilities. Additionally, although we examined certain potential mediators like the Big 5 personality traits and the need for closure, we could not pinpoint the specific factors responsible for the observed bimodal distributions; future studies might explore other variables such as quality of answers and conversational breakdowns as factors affecting user satisfaction. Finally, our work did not explore the impact of adding mechanisms for additional sequence organization features like pre- or post-insertions; future research could expand upon these aspects to offer a richer understanding of how different features can improve user interaction with AI agents. Overall, while this study has limitations, it lays the groundwork for future research, providing foundational insights into the role of insert-expansion and the "user-as-a-tool" approach in conversational agents, and thereby contributing to the broader field of human-AI interaction.

Conclusion

In this study, we have introduced and empirically explored a "user-as-a-tool" approach in the context of tool-enabled conversational agents. By implementing a mechanism for the agent to gather information from the

user without leaving its reasoning loop, this approach implements the natural dialogue strategy of "insert-expansion", thereby improving reasoning performance. The approach excels particularly in recommendation scenarios, serving as a pivotal advancement for creating effective, human-centered interactive systems. By elevating the user's role in guiding AI reasoning, our findings contribute a robust strategy to the emerging discourse on optimizing AI-human interaction loops.

References

Applebee, A. N. (1984). "Writing and Reasoning," *Review of Educational Research* (54:4), pp. 577–596.

De-Arteaga, M., Fogliato, R., and Chouldechova, A. (2020). "A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems,* R. Bernhaupt, F. '. Mueller, D. Verweij, J. Andres, J. McGrenere, A. Cockburn, I. Avellino, A. Goguey, P. Bjørn, S. Zhao, B. P. Samson, and R. Kocielnik (eds.). New York, NY, USA: ACM, pp. 1–12.

Ashktorab, Z., Jain, M., Liao, Q. V., and Weisz, J. D. (2019). "Resilient Chatbots," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems,* S. Brewster, G. Fitzpatrick, A. Cox, and V. Kostakos (eds.). New York, NY, USA: ACM, pp. 1–12.

Borsci, S., Malizia, A., Schmettow, M., van der Velde, F., Tariverdiyeva, G., Balaji, D., and Chamberlain, A. (2022). "The Chatbot Usability Scale: the Design and Pilot of a Usability Scale for Interaction with AI-Based Conversational Agents," *Personal and Ubiquitous Computing* (26:1), pp. 95–119.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (n.d.). *Language Models are Few-Shot Learners*.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., and Zhang, Y. (2023). *Sparks of Artificial General Intelligence: Early experiments with GPT-4*.

Chopra, B., Bajpai, Y., Biyani, P., Soares, G., Radhakrishna, A., Parnin, C., and Gulwani, S. (2024). *Exploring Interaction Patterns for Debugging: Enhancing Conversational Capabilities of AI-assistants*.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. (2017). "Deep reinforcement learning from human preferences," *Advances in neural information processing systems* (30).

Clark, L., Pantidi, N., Cooney, O., Doyle, P., Garaialde, D., Edwards, J., Spillane, B., Gilmartin, E., Murad, C., Munteanu, C., Wade, V., and Cowan, B. R. (2019). "What Makes a Good Conversation?," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems,* S. Brewster, G. Fitzpatrick, A. Cox, and V. Kostakos (eds.). New York, NY, USA: ACM, pp. 1–12.

Dika, T. R. (2023). "Descartes' Method," in *The Stanford Encyclopedia of Philosophy,* Edward N. Zalta and Uri Nodelman (eds.). Metaphysics Research Lab, Stanford University.

Fast, E., Chen, B., Mendelsohn, J., Bassen, J., and Bernstein, M. S. (2018). "Iris," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems,* R. Mandryk, M. Hancock, M. Perry, and A. Cox (eds.). New York, NY, USA: ACM, pp. 1–12.

Government Office for Science (2023). *Frontier AI: Capabilities Risks and Scenarios*.

Höhne, J. K., Krebs, D., and Kühnel, S.-M. (2021). "Measurement properties of completely and end labeled unipolar and bipolar scales in Likert-type questions on income (in)equality," *Social science research* (97), p. 102544.

Hornbæk, K. and Oulasvirta, A. (2017). "What Is Interaction?," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems,* G. Mark, S. Fussell, C. Lampe, m. schraefel m.c., J. P. Hourcade, C. Appert, and D. Wigdor (eds.). New York, NY, USA: ACM, pp. 5040–5052.

Hornik, K., Stinchcombe, M., and White, H. (1989). "Multilayer feedforward networks are universal approximators," *Neural Networks* (2:5), pp. 359–366.

Kendrick, K. H., Brown, P., Dingemanse, M., Floyd, S., Gipper, S., Hayano, K., Hoey, E., Hoymann, G., Manrique, E., Rossi, G., and Levinson, S. C. (2020). "Sequence organization: A universal infrastructure for social action," *Journal of Pragmatics* (168), pp. 119–138.

Kompa, N. A. (2023). "Inner Speech and 'Pure' Thought – Do we Think in Language?," *Review of Philosophy and Psychology* ().

LeCun, Y., Bengio, Y., and Hinton, G. (2015). "Deep learning," *Nature* (521:7553), pp. 436–444.

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., Zhu, D., Li, X., Qiang, N., Shen, D., Liu, T., and Ge, B. (2023). *Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models*.

Meredith, J. (2019). "Conversation Analysis and Online Interaction," *Research on Language and Social Interaction* (52:3), pp. 241–256.

Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., Rozière, B., Schick, T., Dwivedi-Yu, J., Celikyilmaz, A., Grave, E., LeCun, Y., and Scialom, T. (2023). *Augmented Language Models: a Survey*.

Oakley, K. (1957). "Tools Makyth Man," *Antiquity* (31:124), pp. 199–209.

Parisi, A., Zhao, Y., and Fiedel, N. (2022). *TALM: Tool Augmented Language Models*.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. (2023). *Generative Agents: Interactive Simulacra of Human Behavior*.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). *Improving language understanding by generative pre-training*.

Roets, A. and van Hiel, A. (2011). "Item selection and validation of a brief, 15-item version of the Need for Closure Scale," *Personality and Individual Differences* (50:1), pp. 90–94.

Schegloff, E. A. (2007). *Sequence organization in interaction: A Primer in Conversation Analysis I,* A primer in conversation analysis : v. 1. Cambridge: Cambridge University Press.

Soto, C. J. and John, O. P. (2017). "Short and extra-short forms of the Big Five Inventory–2: The BFI-2-S and BFI-2-XS," *Journal of Research in Personality* (68), pp. 69–81.

Tudini, V. (2015). "Extending Prior Posts in Dyadic Online Text Chat," *Discourse Processes* (52:8), pp. 642–669.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, ., and Polosukhin, I. (2017). "Attention is All you Need," in *Advances in Neural Information Processing Systems,* G. I., V. L. U., B. S., W. H., F. R., V. S., and G. R. (eds.). Vol. 30. Curran Associates, Inc.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. (2021). *Finetuned Language Models Are Zero-Shot Learners*.

Wu, T., Terry, M., and Cai, C. J. (2022). "AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts," in *CHI Conference on Human Factors in Computing Systems,* S. Barbosa, C. Lampe, C. Appert, D. A. Shamma, S. Drucker, J. Williamson, and K. Yatani (eds.). New York, NY, USA: ACM, pp. 1–22.

Yang, Q., Steinfeld, A., Rosé, C., and Zimmerman, J. (2020). "Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems,* R. Bernhaupt, F. '. Mueller, D. Verweij, J. Andres, J. McGrenere, A. Cockburn, I. Avellino, A. Goguey, P. Bjørn, S. Zhao, B. P. Samson, and R. Kocielnik (eds.). New York, NY, USA: ACM, pp. 1–13.

Yao, W., Heinecke, S., Niebles, J. C., Liu, Z., Feng, Y., Le Xue, Murthy, R., Chen, Z., Zhang, J., Arpit, D., Xu, R., Mui, P., Wang, H., Xiong, C., and Savarese, S. (2023). *Retroformer: Retrospective Large Language Agents with Policy Gradient Optimization*.

Yin, M., Wortman Vaughan, J., and Wallach, H. (2019). "Understanding the Effect of Accuracy on Trust in Machine Learning Models," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems,* S. Brewster, G. Fitzpatrick, A. Cox, and V. Kostakos (eds.). New York, NY, USA: ACM, pp. 1–12.

Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., and Yang, Q. (2023). "Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems,* A. Schmidt, K. Väänänen, T. Goyal, P. O. Kristensson, A. Peters, S. Mueller, J. R. Williamson, and M. L. Wilson (eds.). New York, NY, USA: ACM, pp. 1–21.