Participatory prompting: a user-centric research method for eliciting AI assistance opportunities in knowledge workflows

Advait Sarkar
Microsoft Research
University of Cambridge
University College London
advait@microsoft.com

Ian Drosos
Microsoft Research
t-iandrosos@microsoft.com

Rob Deline
Microsoft Research
rob.deline@microsoft.com

Andrew D. Gordon
Microsoft Research
University of Edinburgh
adg@microsoft.com

Carina Negreanu
Microsoft Research
cnegreanu@microsoft.com

Sean Rintel
Microsoft Research
serintel@microsoft.com

Jack Williams
Microsoft Research
jack.williams@microsoft.com

Ben Zorn
Microsoft Research
ben.zorn@microsoft.com

Abstract

Generative AI, such as image generation models and large language models, stands to provide tremendous value to end-user programmers in creative and knowledge workflows. Current research methods struggle to engage end-users in a realistic conversation that balances the actually existing capabilities of generative AI with the open-ended nature of user workflows and the many opportunities for the application of this technology. In this work-in-progress paper, we introduce participatory prompting, a method for eliciting opportunities for generative AI in end-user workflows. The participatory prompting method combines a contextual inquiry and a researcher-mediated interaction with a generative model, which helps study participants interact with a generative model without having to develop prompting strategies of their own. We discuss the ongoing development of a study whose aim will be to identify end-user programming opportunities for generative AI in data analysis workflows.

1. Introduction and motivation

Generative AI presents many opportunities for assistance and automation for end-users and end-user programmers. Our research team is interested in exploring how Large Language Model (LLM) assistance can be used in data-driven sensemaking (Russell, Stefik, Pirolli, & Card, 1993; Pirolli & Card, 2005) in spreadsheets, to identify key areas of strength and weakness in LLM assistance and identify opportunities for LLM assistance that address specific parts of the overall workflow. End-user data analysis workflows are complex and range over many steps, including problem conceptualization, identifying relevant datasets, data cleaning and structuring, developing an analysis strategy, learning how to use relevant features, open-ended exploration, and presenting results.

The effect of generative AI on knowledge work has been described as a shift *"from material production to critical integration"* (Sarkar, 2023). Critical integration consists of *"deciding where in the workflow to use the productive power of AI, how to program it correctly [...], and how to process its output in order to incorporate it"*. Sarkar builds on the theory of double-loop learning in organizations (Argyris, 1977), observing that there is both an inner-loop aspect to applying AI in knowledge workflows (incorporating AI assistance in various steps of existing workflows) as well as an outer-loop aspect (reconfiguring knowledge workflows to take better advantage of AI, and developing new ones which are only possible with AI).

For example, in data-driven sensemaking, critical integration in the inner loop might consist of finding applications for AI in data visualization, or data cleaning. Critical integration in the outer loop might

consist of applying AI towards identifying a suitable analysis strategy or automating large portions of the sensemaking workflow (e.g., in the spirit of the "automatic statistician" (Steinruecken, Smith, Janz, Lloyd, & Ghahramani, 2019)) and developing new tools for human overseers focusing on auditing and quality control.

A key question for researchers and designers at this point in time is how to study the needs of users involved in such workflows. We are concerned with the first phase of the design "double diamond"; we first need to design the right thing, and only later can we attend to getting the design of the thing right (Buxton, 2010). As generative AI technology is new and continuously evolving, its use in society is limited and uneven. It may not be possible, for example, to simply observe participants working with generative AI, or interview them about their work practices with generative AI, if generative AI is not widely adopted within their workflows (which is the case for the vast majority of knowledge work at the time of writing). For example, code completion in code editors for professional software developers has been an early commercialization of generative AI, which gives researchers a wide pool of experienced users with mature behaviours to study (Sarkar et al., 2022). Other work has discussed code assistants for data analysis within computational notebooks (Mcnutt, Wang, Deline, & Drucker, 2023). Unfortunately, for our end-user scenario of interest (data analysis workflows in spreadsheets), this is not yet the case.

Furthermore, it is not ideal for researchers to develop high-fidelity experiences for generative AI as a way of testing its applicability to different workflow. It is time-consuming and expensive. Moreover, it is also limiting; out of the wide variety of potential interventions at the inner and outer loops, only a very small number can be feasibly explored using a functional prototype.

The traditional solution to this has been to use lower-fidelity methods such as Wizard-of-Oz (Gould, Conti, & Hovanyecz, 1983; Landauer, 1986), paper prototyping (Snyder, 2003) and champagne prototyping (Blackwell, Burnett, & Jones, 2004), which allow researchers to rapidly simulate a wide variety of user experiences with significantly lower engineering costs, while also enabling interaction with experiences that may be extremely challenging or impossible to build due to technical limitations. However, these methods have limitations as well; for a Wizard-of-Oz study to have direct implications for design, the Wizard protocol must correspond to the actually existing capabilities of the system(s) that are eventually built.

In particular, the mythologizing of AI's capabilities by the media, academia, and industry has led to a warped public conception of what AI can do and how it works (Sarkar, 2022). Thus Siddharth et al. (Siddarth et al., 2021) urge us to focus not on this collective mirage of what AI might be, but on "actually existing AI (AEAI)". There is a real risk that participant responses in low-fidelity studies will draw from their own biased and inflated expectations of AI capabilities to fill in the "gaps" left by the incomplete nature of the prototype. A poorly designed Wizard protocol, which allows too much improvisational deviation from a script, can exacerbate this. This problem is even greater in generic "need-finding" interviews where no prototypes are used.

There is thus a need for a research method that combines the advantages of low-fidelity methods such as Wizard-of-Oz, and rapidly exploring a wide range of potential interactions at both the inner and outer loops of a knowledge workflow, while still grounding conversations with participants in the capabilities of actually existing AI. In response, we have been developing a method called **participatory prompting**.

The participatory prompting method takes the form of a researcher-mediated interaction between a study participant and a working generative AI system. During the session, the researcher guides the participant through a workflow, seeking to test the potential applications for AI at each step. The researcher plays multiple roles in facilitating this interaction. Most importantly, they restructure user requests according to pre-identified prompting strategies, and help the user continue the interaction and recover from errors. The semi-structured interview is grounded in a specific real problem of interest to the user, drawing on principles of contextual inquiry (Raven & Flanders, 1996). The name of the method is inspired by participatory design (Spinuzzi, 2005), and we hope that in the spirit of participatory design, the method of participatory prompting contributes to the design of AI systems that empower and enfranchise users

with their involvement from the outset. The next section describes the method in detail.

2. The participatory prompting method

2.1. Materials required

Choice of system. The participatory prompting method uses a real, functional generative AI system as representative of the functionality of generative AI in general. It is therefore important to choose the system carefully and consider multiple alternatives for their suitability to the particular study.

We compared the following four systems for their suitability for use in our study: OpenAI playground, OpenAI ChatGPT, Google Bard, and Microsoft Bing Chat. We compared them by entering some example queries that a user might have into each system, and attempting to elicit guidance and multiple stages of the data analysis process, as we were intending to do during the study. We then discussed and evaluated the comparative quality of the responses and how a participant might react to each response, with a view to choosing the system which would help produce the most insightful interactions during the study.

It is worth noting that of the four systems we tested, the latter three (ChatGPT, Bard, and Bing Chat) are consumer-facing products: they are built upon one or more large language models and consist of UI elements and other modules and heuristics which come together to create a coherent experience for the non-expert consumer. They can be considered to be significantly "opinionated" in a number of ways. One obvious user-facing manifestation of this is in the so-called "guardrails" which kick in whenever the conversation topic approaches an area deemed inappropriate by the system designers (such as violent or sexual content). Another example of opinionation is the turn limits imposed by Bing Chat: at the time of writing, a conversation with Bing Chat cannot exceed 15 turns (after 15 turns, the conversation is erased, and a new conversation is started).

In contrast, the OpenAI playground is intended for developers to interactively test different models, and as such allows for the choice between multiple individual LLMs, and control over parameters such as temperature (most of our testing on the OpenAI playground was using the default temperature of `0.7` and the `text-davinci-003` model, a GPT-3 model which was the state of the art at the time of testing). While there are still some heuristics and guardrails in place, using the OpenAI playground is much closer to getting the "raw" output of a language model.

For many studies, using a highly opinionated experience may not be ideal; the heuristics and modules used in these systems are proprietary, and researcher control and visibility into parameters such as temperature is poor. For our purposes, however, this was not a dealbreaker.

For our study we have chosen to use Bing Chat for the primary reason that it is the only system (at the time of writing) that is designed to seek and include information from the Web as part of its responses. In our testing, we found that for many steps of the data analysis journey (ideating potential analysis paths, identifying relevant datasets, learning to use relevant features), the ability to report information from the Web resulted in much better and more actionable suggestions for users.

Due to the complexity of deploying these systems at scale, during our comparative evaluation we noticed many outages, where the system was overloaded and did not respond to queries and/or displayed an error message. For a user study to go smoothly, a system that is stable and consistent is key. While we did not quantify the outages we experienced, our informal assessments of a particular system's reliability and uptime did influence our final decision.

Prompt strategies. Identifying performant and consistent strategies for prompting LLMs is a well-documented challenge. In consumer-facing products, the user query is rarely sent directly to an LLM; instead it is processed and augmented with additional instructions and prompts that have been determined by the system developers.

Thus, when non-experts directly interact with a "raw" LLM (e.g., via tools such as OpenAI playground),

or with a generic chat application that is not tuned towards particular knowledge workflows, they may not be able to develop suitable prompting strategies to elicit good performance from the model. This is a key reason that our method involves researcher-mediated interaction, and why we do not simply study how end-users interact directly with the model.

A key strength of the participatory prompting method is that researchers familiar with the design of prompting strategies can prepare these ahead of time.

For our study, a group of researchers collaboratively experimented with different prompting strategies with Bing Chat over a period of several weeks, documenting screenshots of their interactions with Bing Chat and successful prompts in a shared document. A provisional list of prompting strategies which we developed through this process is given in Appendix B. For example, through this process we identified that Bing Chat:

- did not consistently use data sources from the web even if they were available, and we could bias it towards doing so by including a phrase in the prompt such as "use an online data source", "based on publicly available information", "with data from the web", and "use information from the web".
- did not consistently offer citations for sources, but could be biased to do so by including a phrase in the prompt such as "prove your sources are real" or "cite your sources".
- often provided multiple suggestions for types of data analysis the user could conduct, but the answers did not support an end-user's decision for what to do next. In this case we found that adding "justify your answer" or "justify your criterion" improved the actionability of the model's responses.
- is capable of rendering tables inline within the chat, which is very helpful for exploring ideas related to spreadsheet-based data analysis, but it does not consistently do so. We found that we could bias it towards generating tables by specifying "with an example", "make an example spreadsheet", or "make an example table".

Our method for identifying prompts is largely a pragmatic craft practice, based on trial-and-error and the intuitions of researchers. Due to the many sources of variability in LLM output, as well as variability between researchers' experience and the working examples they choose for testing different prompting strategies, our resulting prompts are subjective and difficult to reproduce. Another team, or the same team choosing different working examples, or a different model, may well have developed a different set of prompts, which will have significant downstream effects on the user study. Improving the consistency and systematicity of this step is a major challenge for user research with generative AI, as many libraries, toolkits, and even prompt marketplaces have been created to assist in this endeavour.

Demographics and generative AI experience. Participants will complete a standard demographics questionnaire which includes questions about spreadsheet experience, formula experience, and programming experience (Sarkar et al., 2020). In future participatory prompting studies, this can be replaced with another demographics questionnaire that gathers information relevant to those studies instead.

Based on the model of other questions in that questionnaire, we also developed a simple questionnaire item for assessing prior experience with generative AI, as follows: *"Which of the following BEST describes your experience with generative AI tools such as ChatGPT, DALL-E, Stable Diffusion, Midjourney, Google Bard, Bing Chat?"*

In response, participants choose from the following options:

1. Never heard of them
2. Heard of them but haven't tried any
3. Casually tried one or more
4. Occasionally use one or more
5. Regularly use one or more

As with other studies which use the aforementioned spreadsheet experience questionnaire, this item can be used in one of two ways: first, it can be used as part of the qualitative interpretation of participant interview data, to add context to their responses. Second, it can be used to group participants into rough categories of high and low prior experience (e.g., response levels 1-3 can be considered "low" experience and 4-5 can be considered "high" experience) for studying quantitative interactions between experience and any dependent variables gathered during the study (e.g., cognitive load (Hart & Staveland, 1988)).

Unlike spreadsheet experience or programming experience, the landscape of end-user experience with generative AI is shifting rapidly. The specific wording of this question and its response categories are thus likely to require periodic revision and updates.

2.2. Main interview activity

The main phase of the participatory prompting study takes the form of a semi-structured interview run concurrently with a researcher-mediated "conversation" between the participant and the model.

The interview consists of a number of "turns" consisting of 5 steps. (1) A turn begins by the participant expressing a query (e.g., asking for assistance, posing a question, asking for clarification). (2) Next, the researcher takes the user query, modifies and augments it according to the previously identified prompting strategies and sends it to the model. (3) The participant reads the model's response. (4) Next, the researcher asks the participant to reflect on the response. (5) Finally, the researcher guides the participant in continuing the conversation and choosing the next query.

We wish to explore the possibility for LLM assistance in the following scenarios:

- Problem conceptualization, decomposition, identifying parts of the problem that could be tackled in a spreadsheet
- Identifying relevant datasets
- Figuring out how to clean and structure data
- Developing an analytical strategy, involving applying multiple features in sequence
- Learning how to use relevant features
- Exploration of alternative analyses
- Presenting and communicating results

The problem chosen is ideally seeded by the participant's own problem domain. This can be elicited using a question such as: *"Can you share an example of a decision you had to make recently? The decision should be reasonably complex, requiring an evaluation of multiple criteria or sources."*

If elicited ahead of the study (e.g., in a pre-study communication, or as part of the initial demographics questionnaire), researchers could prepare a spreadsheet and problem that is familiar to the participant's own experience, or we can have the participant bring a shareable spreadsheet within their domain to work on. Alternatively, a suitable problem can be determined at the start of the interview. In practice we have found it is better to ask participants to think of such problems ahead of time, so that more time can be spent on the interactive portion of the interview.

The problems users bring can further be divided into the following types:

1. A well-established spreadsheet workflow where the user is already using spreadsheets.
2. An open-ended problem where the user has not tried to apply spreadsheets before.

From a research perspective, both types of seed problem have advantages, as they correspond respectively to the inner and outer loop of the double loop of AI assistance opportunities. In our study context, we are not interested in one or the other type in particular, or in comparing between the two, so we will not aim to control the distribution of types. However, future studies may be interested mainly in inner loop opportunities, or outer loop opportunities, or a direct comparison between them. In such cases care must be taken to ensure the seed problems used with participants are either predominantly of the type of concern, or roughly evenly distributed between the types to facilitate comparison.

With a suitable seed problem, we walk through the participant's problem step by step, entering their requests into the LLM system (using pre-identified prompt strategies) and relaying their response back to the user. We find that it is useful for participants themselves to also view the screen on which the LLM interaction is taking place, as the study can progress faster when participants can read the output directly themselves.

The user is then asked questions at each step such as:

- Is this useful? Why or why not?
- Are you confused or surprised? Why or why not?
- Does it contain anything that is factually incorrect or misleading?

To advance the conversation to the next turn, the experimenter may prompt the participant with a question such as

- Does this give you any further ideas?
- What would you like to know next, to continue your analysis?
- What information is missing?

Participants may also leverage suggested follow-up questions provided by the model as inspiration. However, since these suggestions may not adhere to experimenter prompting strategies, the suggestions may need intervention by the experimenter to align them.

2.2.1. Post-activity interview

After the turn-taking phase of the study, the participant is interviewed and asked to reflect on the experience. For instance, they could be asked how such a tool would fit into their workflow, what features they feel would improve the experience, and what were the strengths and weaknesses of the new modes of working enabled with generative AI.

This phase may additionally involve eliciting participants' responses to mock-ups of potential interface designs in a design probe. Importantly, because the participant has just had the experience of interacting with an actually existing AI system, they are more likely to have an accurate understanding of what a different interface design might actually achieve for their workflow in terms of usability. The participant's grounding in actual AI capabilities improves the validity of research insights over simply interviewing participants about their response to mock-ups.

If the participant reported that they had experience with generative AI, experimenters could elicit the participant to compare their previous workflows with what they experienced with participatory prompting. This might include the differences in solving similar or the same problems they saw in the study, or even understanding how the participant might change their own prompting strategies after the study.

The structure of and questions asked during the post-activity interview depends on the aims of the participatory prompting study. In our situation, we are interested in how generative AI can help non-expert end-users in data analysis workflows, particularly within spreadsheets, and so we selected our questions accordingly. Our full final script (incorporating revisions made after a pilot study detailed in Section 2.3) is given in Appendix A.

2.3. Pilot

The first version of this study protocol was piloted on a convenience sample of two participants who are familiar with spreadsheets and who use spreadsheets for their work. Each pilot took approximately 1 hour, as intended.

The pilots resulted in the following observations and adaptations:

It can be difficult for participants to settle on a suitable seed problem that is complex enough that it requires generative AI (as opposed to a traditional web search) to solve, but simple enough that the required context can be described to the system using a few sentences or a paragraph at most. We introduced more questions in the problem elicitation phase of the study that the experimenter could use

to help the participant (e.g., *"Can you share an example of a problem that required you to develop a new workflow?"*). However, our recommendation is that if possible, the participant should be asked to think of a suitable problem ahead of the scheduled study session, to maximize the time available to engage with the problem in the turn taking phase. We also noticed that participants were not familiar with some jargon in our questions (e.g., "data-driven decision-making") and we modified our questions to elaborate and clarify these terms.

We noticed that the participants were able to go through 5-6 turns in the time allotted. This may seem like a small number of turns, but it nonetheless produced a wide range of qualitative insights. The turn-taking phase can be elongated in future studies if this is felt to be necessary, study duration targets and participant fatigue notwithstanding.

The most time-consuming aspect of each turn is in the reflection step, where the participant is asked to reflect on the system's response, and the advancement step, where the researcher guides the participant to decide what to do next. This observation helped us decide on setting Bing Chat to "creative" mode for the study. Bing Chat has a single user-facing setting. The user can choose between creative mode (described by the Bing Chat UI as "original and imaginative"), balanced mode ("informative and friendly"), and precise mode ("concise and straightforward"). We initially used "precise" mode because we believed that it would be the least likely to hallucinate misinformation, and because generating short responses would allow the user to read through them more quickly and therefore enable more turns overall. Since the number of turns is largely dominated by the time spent on the reflection and advancement phases, the small time advantage gained in precise mode by having to read less text did not accumulate to allow an increased number of turns overall. Moreover, in practice we observed that "creative" mode was no more likely to generate hallucinations, and since it was far more verbose, often emitting several paragraphs in response, it improved participants' reflections (by giving them more to reflect on) and the ease with which a suitable next query was selected.

We noticed that if the model's response is completely generic or not useful, especially at an early stage of the conversation, our pilot participants were not motivated to continue the interaction. In response to this, we introduced a number of different advancement-oriented questions the researcher could use to help suggest a way forward, such as: *"What would you change in your query to make this more useful? Would you ask this a different way?"*.

We noticed participants' preconceived notions about the system's capabilities were heavily influenced by their prior experience with search engines, and initially thought to use short queries of the kind used with search engines. This is unsurprising given Bing Chat's positioning within a more traditional search interface. Previous studies have also noted that participants' use of generative AI systems is influenced by their experience with search engines (Liu et al., 2023; Sarkar et al., 2022). However, such short queries cannot adequately capture the user's context and intent. We introduced a guidance statement in the protocol whereby the experimenter explains that the generative AI system permits longer and more conversational interaction.

We also introduced a couple of strategies for the experimenter to gently suggest a way to continue the conversation, if the participant was having difficulties ideating a next step. These include the experimenter directly suggesting an action (e.g. *"Let's see what happens if we try <some query>"*) but also suggesting an action as a baseline to help the participant conceive a contrastive alternative (e.g., *"I propose to continue by <some query>, but what would you have done instead?"*). However, it is important that the experimenter's suggestions do not bias or significantly change the course of the interaction, and serve mainly to unblock the participant. Much as with regular interviewing, the ability to elicit rich responses from the participant without introducing bias depends on the skill of the interviewer. To help guard against such bias, we recommend in the protocol that these experimenter-led strategies should not be employed in consecutive turns (i.e., if the experimenter led the query in the previous turn, they should not do so in the current turn).

There were issues with understanding participant expectations of model output, where even after work-

ing with the experimenter to craft a prompt, the participant did not know they would need to specifically request images. This led to needing to re-prompt the model to obtain the desired result. To prevent needing to do multiple prompts to obtain desired output, which can take up study time, experimenters should inquire on the expected results, including data types, from the participant to close this gap. Understanding what types of data could be useful for the participant, and explicitly requesting them in the prompt, is a necessary strategy.

Participants noticed that content the model had previously shown in the output could be missing in successive outputs. When the motivation of the participant was to build upon previous results, they wanted to make sure the data was consistent throughout the conversation with Bing Chat. Therefore, prompts crafted for the study need to include or refer to previous output in an attempt to have the model consider this data for further prompts.

One participant was suspicious of the data the model produced as a column in a table and wanted to verify this data by going to the websites the model referenced. This is a third workflow separate from prompting and spreadsheeting that requires a tangent into navigating to the source and verifying the data. This is a realistic strategy for users of chat based LLMs, but it is removed from prompting interactions. While this is not explicitly part of the protocol, we will allow participants the freedom to explore and verify the results returned by the model if desired.

In the post-activity interview, we noticed participants speculating on multiple occasions that *"if it could do <some action>, that would be helpful"*. Since these types of questions can actually be put to the system to test whether it can do it, we amended the protocol to permit the researcher to spot-test such participant speculations and get feedback from the participant. We also introduced the following question to specifically elicit perceived barriers to sensemaking with AI assistance: *"What barriers or frustrations did you have with this experience that prevented you from exploring the question to your satisfaction?"*

Our full revised script after the pilot is given in Appendix A.

2.4. Effectiveness of protocol during pilot

While we do not claim these are usable findings due to currently running a pilot of n=2 and the protocol was adapted live during these pilot runs, we believe there is evidence that this protocol was effective at revealing valuable insights from participants. These include:

- After a few prompts, a participant noted they would start a spreadsheet to maintain the data they were receiving from the model, but thought it would be difficult to switch between the spreadsheet and further prompting. Getting the model to generate a table to help the participant visualize a future spreadsheet was helpful in this situation.
- A participant was unsure the model considered the entire context of the prompt it was given, even when this context was in the prompt, and felt there was no way to verify this with the model.
- Upon noticing a result of potentially summarized or hallucinated data was given by the model, a participant noted that if they could not trust the results and had to manually search to verify the data in the table. They said they felt it severely limited the benefits of Bing Chat.

We believe this protocol is an improved adaptation of the traditional Wizard-of-Oz approach for studies of generative AI, since participants interact with a real AI model, but the implementation costs were extremely low.

3. Discussion and limitations

The participatory prompting approach detailed in this paper raises the question of whether some activities carried out by the human experimenter could be supported with AI. One question that remains to be answered is what affordances human prompt strategies have over an AI that is focused in helping the participant best interact with the generative AI.

However, such a protocol might increase the turn time or number of turns taken as the participant has to interact with this new AI prompting assistant while also performing their sensemaking task. Human-

driven participatory prompting also allows the experimenter to ask user experience questions and inquire on participant motivations, which can provide valuable insights for researchers but may not be valuable for the actual prompting and might not be asked by an AI assistant.

One clear extension of this protocol is for the experimenter to also draw upon the library of existing AI plugins and recommend useful experiences that help the participant solve their problem. This could be directly compared to the effectiveness of the existing plugin experience where the model chooses which plugin to use, assuming the user has the installed plugins.

One limitation of this protocol is that because the experimenters will take a turn at helping craft prompts with the participant, results following this protocol may not give a clear understanding of where and when a user's unsupported prompting would have had issues. We attempt to address this limitation by having the participant reflect on how they would have modified a prompt (see Appendix A).

Similarly, because we are interested in how users might perform sensemaking in spreadsheets by leveraging AI, we have created an environment and crafted prompts that emphasize the use of spreadsheets and organized data. This might mean that the choice to move from prompting to spreadsheeting may not be as organic as it would if the user interacted with the model without experimenter assistance. There is a multitude of data analysis experiences that might also be useful for users (e.g., OpenAI's Code Interpreter, which performs data analysis tasks with Python code (OpenAI, 2023)). The freedom to choose from available interactions would provide useful insights for user needs for end-user data analysis and sensemaking tasks.

Some interactions were limited by the inability to re-trigger generation of a response with respect to a specific query within the conversation in Bing Chat (re-generation and editing a query is possible for instance, within ChatGPT and the OpenAI playground; this enables a kind of flexibility and fluidity that is akin to being able to independently edit and run different code cells out-of-order in a Jupyter notebook). For instance, if the participant changed their mind about a query, or if the system stopped generating text partway through a response, which worried one participant about continuing to prompt the model. In Bing Chat the only option is to submit a follow-up query within the same conversation, but which will include the undesired or incomplete previous queries and responses as part of the "context". The alternative is to start a fresh conversation and then laboriously "replay" the conversation, building up the same conversational state (or more likely, a *similar* state, since the system's responses are nondeterministic) through the same series of prompts until you arrive at the point in the conversation at which you wish to try a different query. Neither of these options is practical or predictable in a time-limited study.

4. Conclusion

In this work-in-progress paper we have presented the ongoing development of **participatory prompting**: a lightweight user research method for eliciting opportunities for AI assistance in knowledge workflows. It uses a real, functional generative AI system, thus improving on traditional Wizard-of-Oz or paper prototyping, where the user interaction can become unmoored from the technical reality of these systems. On the other hand, it allows researchers to use an "off-the-shelf" AI model with no additional engineering costs for fine-tuning, customization, or UI development, enabling rapid and broad-ranging testing of user experiences.

We reported a pilot study (n=2) in which we tested the participatory prompting protocol. The pilots have resulted in improvements to the protocol, changes to the script, and reflections on how to get the most insight out of a participatory prompting session. The pilots have validated the feasibility of the protocol as a method for understanding the user experience of generative AI in knowledge workflows. In future work, we are planning to run a full-scale participatory prompting study to elicit opportunities for AI assistance in the data analysis workflows of end-user programmers in spreadsheets.

5. References

Argyris, C. (1977). Double loop learning in organizations. *Harvard business review*, *55*(5), 115–125.

Blackwell, A. F., Burnett, M. M., & Jones, S. P. (2004). Champagne prototyping: A research technique for early evaluation of complex end-user programming systems. In *2004 ieee symposium on visual languages-human centric computing* (pp. 47–54).

Buxton, B. (2010). *Sketching user experiences: getting the design right and the right design*. Morgan kaufmann.

Gould, J. D., Conti, J., & Hovanyecz, T. (1983). Composing letters with a simulated listening typewriter. *Communications of the ACM*, *26*(4), 295–308.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139–183). Elsevier.

Landauer, T. K. (1986). Psychology as a mother of invention. *ACM SIGCHI Bulletin*, *18*(4), 333–335.

Liu, M. X., Sarkar, A., Negreanu, C., Zorn, B., Williams, J., Toronto, N., & Gordon, A. D. (2023). "What It Wants Me To Say": Bridging the Abstraction Gap Between End-User Programmers and Code-Generating Large Language Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–31).

Mcnutt, A. M., Wang, C., Deline, R. A., & Drucker, S. M. (2023). On the design of ai-powered code assistants for notebooks. In *Proceedings of the 2023 chi conference on human factors in computing systems.* New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/3544548.3580940` doi: 10.1145/3544548.3580940

OpenAI. (2023). *Chatgpt plugins: Code interpreter.* Retrieved from `https://openai.com/blog/chatgpt-plugins#code-interpreter`

Pirolli, P., & Card, S. (2005). The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis* (Vol. 5, pp. 2–4).

Raven, M. E., & Flanders, A. (1996). Using contextual inquiry to learn about your audiences. *ACM SIGDOC Asterisk Journal of Computer Documentation*, *20*(1), 1–13.

Russell, D. M., Stefik, M. J., Pirolli, P., & Card, S. K. (1993). The cost structure of sensemaking. In *Proceedings of the interact'93 and chi'93 conference on human factors in computing systems* (pp. 269–276).

Sarkar, A. (2022, March). Is explainable AI a race against model complexity? In *Workshop on Transparency and Explanations in Smart Systems (TeXSS), in conjunction with ACM Intelligent User Interfaces (IUI 2022)* (pp. 192–199). Retrieved from `http://ceur-ws.org/Vol-3124/paper22.pdf`

Sarkar, A. (2023). Exploring Perspectives on the Impact of Artificial Intelligence on the Creativity of Knowledge Work: Beyond Mechanised Plagiarism and Stochastic Parrots. In *2023 Symposium on Human-Computer Interaction for Work (CHIWORK 2023)* (pp. 1–11).

Sarkar, A., Borghouts, J. W., Iyer, A., Khullar, S., Canton, C., Hermans, F., ... Williams, J. (2020). Spreadsheet use and programming experience: An exploratory survey. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–9).

Sarkar, A., Gordon, A. D., Negreanu, C., Poelitz, C., Srinivasa Ragavan, S., & Zorn, B. (2022, September). What is it like to program with artificial intelligence? In *Proceedings of the 33rd Annual Conference of the Psychology of Programming Interest Group (PPIG 2022).*

Siddarth, D., Acemoglu, D., Allen, D., Crawford, K., Evans, J., Jordan, M., & Weyl, E. (2021). How AI fails us. *arXiv preprint arXiv:2201.04200.*

Snyder, C. (2003). *Paper prototyping: The fast and easy way to design and refine user interfaces*. Morgan Kaufmann.

Spinuzzi, C. (2005). The methodology of participatory design. *Technical communication*, *52*(2), 163–174.

Steinruecken, C., Smith, E., Janz, D., Lloyd, J., & Ghahramani, Z. (2019). The automatic statistician. *Automated machine learning: Methods, systems, challenges*, 161–173.

A. Study script

A.1. Materials/activities pre-interview

Ask to prepare spreadsheet / reflect on data-driven workflows.

A.2. [5 minutes] Opening

Introductions and pleasantries, consent form, demographics form.

A.3. [10 minutes] Discussion of current data decision practices.

- Can you briefly describe your role?
- Can you describe, with examples, what kinds of data-driven decision making you do as part of your role?
- Can you describe, with examples, what tools you use?
- Can you describe, with examples, how you approach an unfamiliar data-driven decision making problem? An unfamiliar problem where you had to make a decision based on some data. This could be tabular data, lists, or text, etc.
- Can you describe an unfamiliar data decision problem, potentially fictional, you may encounter in the future?

If this produces a satisfactory scenario, proceed to turn taking, else ask: Can you share an example of a problem that required you to develop a new workflow?

A.4. [30 minutes] Participatory prompting, turn taking

Per turn:

- Is this useful? Why or why not?
- Are you confused, surprised, or indirectly inspired?

To progress, choose one or more of:

- What would you like to know next? What else would you need to know to follow these suggestions?
- What would you change in your query to make this more useful? Would you ask this a different way?
- (Experimenter driven, at most once in a row) Let's see what happens if we try (X). Alternatively: I propose to continue by X, what would you have done (e.g., continued by Y, different task, abandon tool)?
- (If issue with result) I see that there is an issue here with (X), if we do this (new prompt) we can get that data back for you (if participant wants this, continue, else 1st question).
- (If participant is stuck, only thinking in terms of "classical" search engines) Imagine you're talking to a colleague, and bouncing ideas off them.

A.5. [15 minutes] Post activity interview

How would a tool like this fit or not fit into your workflow? If the participant says something like "If it could do X that would be helpful", try it out, get feedback from the participant, but keep time in mind.

1. What benefits does this hybrid spreadsheet-chat workflow provide over your existing workflow?
2. When you were surprised/inspired by X (from turn taking), what features/capabilities would be useful in exploring this inspiration further?
3. What features do you believe would increase the frequency and effectiveness of these inspiring results/moments (e.g., visualizations, videos, suggested prompts)?
4. How do you audit data/decisions now and how would that change with these AI-powered features?
5. How would your decision making workflow change with a tool like this?
6. What barriers or frustrations did you have with this experience that prevented you from exploring the question to your satisfaction?
7. What are the advantages or disadvantages of using a chat-based interface?

B. Pre-identified prompts

This section lists prompts that we have determined through trial and error for use during the study.

1. Problem conceptualization, decomposition, identifying parts of the problem that could be tackled in a spreadsheet
   (a) <Description of user problem>. Explain how to use a spreadsheet for this with an example.
   (b) Explain a different way to use a spreadsheet for this with an example.
   (c) I am trying to make a data-driven decision about <X>. Is this a good problem to use data tools such as spreadsheets to solve? Explain why or why not. What sub-problems or related problems are good candidates for spreadsheet solutions? Justify your answer.
2. Identifying relevant datasets
   (a) What data is relevant to this problem. List sources.
   (b) Use an online data source to add a useful column to this table. Prove your sources are real.
   (c) Add a column to the table containing a score representing <X>. Invent a criterion for this score based on publicly available information. Justify your criterion.
   (d) Add more rows and columns to the table based on information you consider relevant to the decision of <X>.
   (e) Add columns to the table with data from the web such as <X>. Cite your sources.
   (f) Use information from the web to populate the spreadsheet with more accurate figures. Cite your sources.
   (g) Make an example spreadsheet according to your suggestions above. Use information from the web to populate the spreadsheet with accurate information. Cite your sources.
3. Figuring out how to clean and structure data
   (a) Explain how to put this data in a spreadsheet with an example.
4. Developing an analytical strategy, involving application of multiple features in multiple steps
   (a) Explain how to <user problem> in Excel with steps.
   (b) Explain another way to <user problem> in Excel with steps.
   (c) It is not possible to <suggestion>. Explain an alternative method with steps.
   (d) What spreadsheet features can I use, such as charts, formulas, conditional formatting, pivot tables, etc. Show examples.
   (e) <Model suggestion> Explain how to do this with an example.
5. Learning how to use relevant features
   (a) Explain how to use <feature> to solve this problem in Excel with an example.
6. Exploration of alternative analyses
7. Presenting and communicating results

Others (non-categorised)

- Make a spreadsheet example
- <Model mistake> is not correct. Provide an alternative answer and prove that your answer is correct.
- Sometimes Bing Chat will refuse to make a spreadsheet. Try asking for a ‘table’ instead. Or ask repeatedly.