“It’s like a rubber duck that talks back”: Understanding Generative AI-Assisted Data Analysis Workflows through a Participatory Prompting Study

Ian Drosos*
t-iandrosos@microsoft.com
Microsoft Research
Cambridge, UK

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

Xiaotong (Tone) Xu†
xt@ucsd.edu
University of California San Diego
La Jolla, USA

Carina Negreanu
cnegreanu@microsoft.com
Microsoft Research
Cambridge, UK

Sean Rintel
serintel@microsoft.com
Microsoft Research
Cambridge, UK

Lev Tankelevitch
lev.tankelevitch@microsoft.com
Microsoft Research
Cambridge, UK

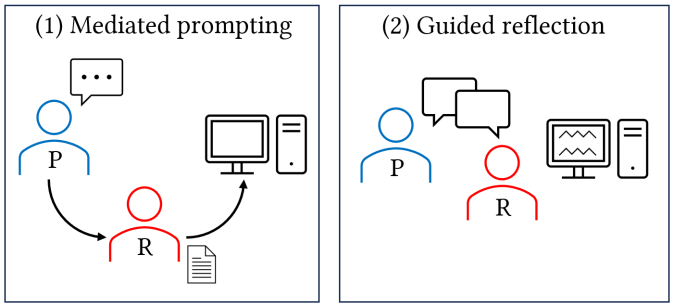

Figure 1: The turn-taking phase of the participatory prompting method. (1) Mediated prompting: the participant (P, blue) expresses their intent. The researcher (R, red) formulates a prompt based on this intent and a set of pre-prepared prompting strategies, and enters the prompt into the system. (2) The participant reflects on the result, guided by the researcher, and forms their next intent, after which the study returns to step (1) for the next turn.

ABSTRACT

Generative AI tools can help users with many tasks. One such task is data analysis, which is notoriously challenging for non-expert end-users due to its expertise requirements, and where AI holds much potential, such as finding relevant data sources, proposing analysis strategies, and writing analysis code. To understand how data analysis workflows can be assisted or impaired by generative AI, we conducted a study (n=15) using Bing Chat via participatory prompting. Participatory prompting is a recently developed methodology in which users and researchers reflect together on tasks through co-engagement with generative AI. In this paper we demonstrate the value of the participatory prompting method. We found that generative AI benefits the information foraging and sensemaking loops of data analysis in specific ways, but also introduces its own barriers and challenges, arising from the difficulties of query formulation, specifying context, and verifying results.

CCS CONCEPTS

• **Human-centered computing** → *HCI theory, concepts and models*; **Natural language interfaces**; **Participatory design**; • **Computing methodologies** → Natural language processing; Neural networks; Machine learning; • **Social and professional topics** → *User characteristics.*

ACM Reference Format:
Ian Drosos, Advait Sarkar, Xiaotong (Tone) Xu, Carina Negreanu, Sean Rintel, and Lev Tankelevitch. 2024. “It’s like a rubber duck that talks back”: Understanding Generative AI-Assisted Data Analysis Workflows through a Participatory Prompting Study. In *Proceedings of the 3rd Annual Meeting of the Symposium on Human-Computer Interaction for Work (CHIWORK ’24), June 25–27, 2024, Newcastle upon Tyne, United Kingdom.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3663384.3663389

*Equal contribution.
†The author was affiliated with Microsoft Research when this research was conducted.

CHIWORK ’24, June 25–27, 2024, Newcastle upon Tyne, United Kingdom

ACM ISBN 979-8-4007-1017-9/24/06
https://doi.org/10.1145/3663384.3663389

1 INTRODUCTION

End-user tools based on generative deep learning, i.e., "generative AI" (defined in Section 2.2) can substantially improve the ability of users to analyse and make sense of data, particularly those without formal expertise or training in data analysis. Data analysis workflows are notoriously tedious, challenging, error-prone, and have high expertise requirements. Generative AI significantly advances the state of the art in facilitating the authoring and debugging of data analysis scripts, reuse of analysis workflows, comprehension of analysis scripts, learning, and exploration [58]. The potential change in user behaviour has been described as the *generative shift* [58]. The generative shift posits three axes of change: intensification (more sophisticated automation will be applied to existing workflows), extensification (more workflows will be automated), and acceleration (workflows which were previously costly will be applied in more contexts, as they become cheaper due to their automation).

An important user scenario for the generative shift is in *end-user data-driven sensemaking*, that is, conducting analyses (often open-ended, ill-defined, and exploratory) within the context of some data (detailed in Section 2.1). Classic examples of end-user data-driven sensemaking include personal and corporate budgeting, financial modelling in spreadsheets, and quantified self [39] activities. Less conspicuous examples include travel planning, or choosing a restaurant to visit or film to watch. These involve a mixture of qualitative and quantitative information, and of subjective and "objective" criteria; to choose a film, one might consider one's personal preferences and mood, the preferences of any companions, one's reactions to the trailer, critical reviews and ratings, film duration, genre, director, cast, and so on.

As previously noted, generative AI has many applications in data-driven sensemaking. It can suggest relevant datasets or analysis procedures, write data transformation and analysis scripts or spreadsheet formulae, help debug or repurpose existing scripts, suggest subjective criteria for evaluating different options, teach the user how to apply an unfamiliar statistical procedure or tool, or even act as a critic or sounding board, to help the user decompose and refine an ill-defined problem. Faced with such a breadth of applications, the key question facing system designers is therefore one of scope: *where are the greatest opportunities and challenges for improving the end-user experience of data-driven sensemaking with generative AI?*

Our study is the first to apply the participatory prompting protocol by Sarkar et al. [63] to explore the opportunities and challenges of generative AI for end-user sensemaking with data. Participatory promoting is a researcher-mediated interaction between the participant and a broad, open-ended AI system, such as OpenAI ChatGPT or Microsoft Bing Chat. The latter are "broad" in the sense that they are designed to support assistance in a wide range of workflows. By virtue of being researcher-mediated, participant experiences can be grounded in actual AI capabilities, scoped down by the researcher to a particular domain (in our case, data-driven sensemaking). We further discuss the value of participatory prompting in the description of our method (Section 3).

Our study found that generative AI supports data analysis workflows in the information foraging loop by streamlining information gathering, and the sensemaking loop by helping users generate hypotheses and develop strategies to test them (Section 4.2). However, we also found challenges to effective use of generative AI in data sensemaking workflows. These included forming effective queries, giving context to the AI, long or vague responses causing information overload, and frustrations with the verification of generated results (Section 4.3). These results provide a range of implications for design, such as assisting users build detailed prompts that contain the context needed by AI to be effective, helping users verify AI responses, and better integration with feature-rich application workflows (detailed in Section 5.2).

As well as the domain-specific results, in this paper we also reflect on the value of the participatory prompting method for developing insights via mediated interaction that might otherwise remain unidentified. We discuss how it might expand to other fields of interest (Section 5.4), but also note some of its limitations in practice. These limitations include striking a balance in experimenter intervention to prevent over-influencing participant workflows, and potential inconsistencies between how researchers create and apply prompt strategies, which may reduce the reproducibility of results (detailed in Section 5.5).

2 BACKGROUND

To clarify our guiding question, in this section we explain the concepts of sensemaking (Section 2.1), generative AI (Section 2.2), and end-user programming (Section 2.3), and summarise previous work on intelligent assistance for data analysis (Section 2.4).

2.1 Sensemaking

We adopt Pirolli and Card's concept of sensemaking [47], which shares roots with Weick's [29, 30] organizational sense-making, but is focused on data analysis rather than social psychology. Sensemaking is the process by which individuals gather information, represent it schematically for interpretation, and develop insights into its meaning to create useful knowledge products. Sensemaking involves two iterative processes: (1) information foraging [46] and (2) hypothesis development and testing (the latter by itself is also called the "sensemaking loop").

The sensemaking framework is heavily influential and has been applied to understand data analyst workflows in multiple scenarios, such as navigating large datasets [51], and understanding unfamiliar data visualisations [35]. Notably, the latter study suggested that novices struggle to construct correct initial mental models ("frames") to inform exploration, tending to persist with incorrect frames. To support sensemaking, the authors suggest that system designers should consider strategies like scaffolded introduction of visualizations or targeted annotation to aid formation of valid initial mental models.

A recent study explored how novice data analysts make sense of computational notebooks [8]. They developed an interface called Porpoise that groups code cells and adds structured labels to support these tasks (thus implementing the scaffolding and targeted annotation suggested by previous work). A counterbalanced user study with 24 practitioners found Porpoise facilitated comprehension and supported the building of mental models compared to default notebooks.

2.2 Definition of generative AI

The term "generative AI" is extremely broad and encompasses many types of systems [58]. The term can variously refer to core algorithms (e.g., the transformer architecture), specific instantiated models (e.g., GPT-4), or fully productized systems consisting of an ensemble of models plus additional components (e.g., ChatGPT).

To provide clarity around this term, Sarkar [58] defines generative AI as *"an end-user tool, applied to programming, whose technical implementation includes a generative model based on deep learning"*. The term "end-user tool" refers to tools that end-users directly interact with, not the underlying algorithms or models. The tool may consist of an ensemble of models, heuristics, engineered prompts, and interfaces. The definition is restricted to generative models based on contemporary deep learning techniques. Finally, the definition is restricted to the programming domain. Examples that fit this original definition include code completion tools leveraging large language models such as GitHub Copilot, and naturalistic language programming in spreadsheets using such models.

In this paper we adopt the "end-user tool" and "technical implementation [...] based on deep learning" aspects of the definition, but rather than programming, our domain of interest is sensemaking with data. Thus, we define generative AI as *"an end-user tool, applied to sensemaking with data, whose technical implementation includes a generative model based on deep learning"*.

2.3 End-user programming

End user programming refers to programming primarily for personal use rather than public use, with the goal of supporting one's work or hobbies rather than developing commercial software. While end user programmers prioritize external goals over software quality, they face many software engineering challenges such as requirements elicitation, design, testing, debugging, and code reuse. Ko et al. provide a survey of the field [32].

Much end-user programming research has focused on spreadsheets. Many techniques help with authoring spreadsheets, ranging from templating systems [13] to programming by example [21]. Testing methods like WYSIWYT (What You See Is What You Test) integrate white box testing into spreadsheet use [50]. Debugging tools analyse formula dependencies or suggest fixes [14, 83]. Other work focuses on developing higher level abstractions to facilitate reuse within spreadsheets, such as lambdas [67], sheet-defined functions [28, 41], and grid-based reuse [26]. Previous research has variously explored how spreadsheets are comprehended [74], learned and adopted [64], or structured [7]. Sensemaking theory has also been applied to end-user programming, for example, to explain and scaffold end-user debugging strategies [18, 22].

While many studies have investigated the potential of AI assistance for data analysis (which will be detailed in Section 2.4), a relatively smaller number have focused on the impact of generative AI more broadly on the activities of programming and end-user programming. Notably, no prior studies have investigated how generative AI tools can impact the data-driven sensemaking workflows of end-user programmers.

In a study exploring the emerging paradigm of artificial intelligence-assisted programming [65], the authors observed shifts in the workflows of programmers, away from directly writing code and toward identifying suitable opportunities for AI aid, forming mental models of when AI support benefits workflows, and evaluating AI-generated output. The challenge for programmers transforms from writing code to activities such as judiciously *"breaking down prompts at the 'correct' level of detail,"* seen as an emerging core programmer competency. Other challenges involve constantly gauging whether any given scenario warrants AI involvement and debugging model outputs post-generation. Working with AI demands qualitatively different skill sets from programmers than previous workflows. More broadly, the theory of *"critical integration"* [56], i.e., the effortful and conscious evaluation, repair, and integration of AI output into a partially automated workflow, appears to be representative of how AI integration affects knowledge work.

An open question in end-user programming research is: to what extent people will still need to write code directly, if generative AI can do this for them from natural language prompts [58]? As generative models advance, the author argues, they may facilitate a significant expansion in the scope and scale of end-user programming activities. However, this "generative shift" also raises questions about the continued relevance of traditional programming languages as an interface. In confronting these questions, the author proposes the focus of end-user programming research should transition from improving formal system usage to new questions around how to design for control and explanation, while mediating user intent through natural language.

2.4 Intelligent assistance for data analysis

AI assistance for data analysis has long been studied under the paradigm of "Intelligent Discovery Assistants" (IDAs). Serban et al. provide an overview of IDAs [71], which predate generative AI technologies and instead rely on AI planning and expert system techniques. Previous research has also considered the end-user activity of *interactive analytical modelling*, i.e., building machine learning models as part of data analysis [61, 66, 68], and developed design principles for designing tools for non-experts [53].

More recently, AI assistance has been studied in connection with exploratory data analysis and computational notebooks. Gu et al. [20] investigate how data analysts from diverse technical backgrounds verify analyses generated by artificial intelligence (AI) systems, finding that analysts shift between procedure-oriented and data-oriented workflows. McNutt et al. [43] conducted an interview study exploring the design space of AI code assistance in notebooks. Among other observations, analysts varied in their preferences in terms of the context provided to the AI system (full context or user-specified), and how assistance should be integrated into the workflow (e.g., in inline cells, in a sidebar, via pop-ups etc.). Chen et al. present WHATSNEXT, an interactive notebook environment that aims to facilitate exploratory data analysis with guidance and a low-code approach [9]. The tool augments standard notebooks with insight-based recommendations for follow-up analysis questions or actions. Li et al. present EDAssistant, an interactive system that facilitates exploratory data analysis (EDA) in Jupyter notebooks through in-situ code search and recommendation [37]. Wang et al.

investigate how professional data scientists interact with a data science automation tool called AutoDS to complete an analysis task [78]. They observed that data scientists expressed more confidence in their manually-created models than models from AutoDS, even though AutoDS models performed better.

A particularly relevant study is Gu et al. [19], who explored analysts' responses to AI assistance that supports *planning* of analyses. They first identified categories of suggestions that such a system could provide, including about data wrangling, conceptual model formulation, operationalisation of constructs, results interpretation, and others. In their Wizard-of-Oz setup, participants interacted with a JupyterLab notebook and received proactive analysis suggestions from a human wizard interacting with a LLM behind the scenes (the wizard was able to observe the notebook for context). Participants' generally valued planning assistance in the form of suggestions, but found them cognitively effortful to consider. Suggestions were helpful when accompanied by commented code, provided at an appropriate time in analysts' workflows, and when matching the analysts' statistical background, domain knowledge, and own analysis plan. However, in some cases, analysts became distracted by the suggestions or over-relied on them.

Researchers have also explored AI assistance from a sensemaking perspective, albeit theoretically, and not yet with empirical evidence from users. Wenskovitch et al. conceptualize how human-machine teams could facilitate AI-driven data sensemaking [82]. The authors propose four roles that humans may assume in such teams: Explorer, Investigator, Teacher, and Judge. Similarly, Dorton and Hall propose a "collaborative" human-AI framework for sensemaking in intelligence analysis [12], notwithstanding critiques of the term "human-AI collaboration" and the collaboration metaphor for human-AI interaction more generally [55].

In summary of the previous work:

- Sensemaking theory gives us a framework for understanding the process of analysing datasets, particularly with open-ended or ill-defined questions. It decomposes the process into a set of interdependent loops of activity, and exposes opportunities for tool design. Sensemaking theory has been applied widely to visual analytics, intelligence analysis, and aspects of software development and end-user programming. However, the broader process of data analysis by non-expert end-users has not been studied with a sensemaking perspective.
- End-user programming research addresses the needs and challenges faced by people, typically non-programmers, writing programs for their own use. A particularly important site of end-user programming activity pertinent to data analysis is the spreadsheet. Numerous studies have elaborated the challenges that spreadsheet users face in learning and comprehending spreadsheets, and writing and debugging formulas. Sensemaking theory has been applied to study some aspects of end-user programming, but the potential impact of generative AI on the broader end-user activity of data analysis has not yet been studied.
- Intelligent assistance for data analysis has been explored in a number of ways, such as suggesting analysis paths and automatic experimentation. Many augmentations of computational notebooks, a common site for exploratory data analysis, have been proposed. Sensemaking theory has been considered in the context of AI assistance for data analysis, but prior explorations have been theoretical. Moreover, the efforts in this space have largely been directed towards expert data analysts.

Crucially, what is missing from previous literature is an understanding of the potential opportunities and challenges with applying generative AI to data-driven sensemaking workflows conducted by non-expert end-user programmers. This is the gap we aimed to fill. This research objective is incredibly broad; we cannot claim to have answered it definitively. However, our study has significantly advanced our understanding of the issue over previous work, and thrown light on new phenomena arising from the confluence of generative AI and end-user data analysis.

3 METHOD

3.1 Participatory prompting

At this stage in generative AI's development, exploratory research questions can be difficult to interrogate in ways that provide sufficient balance of ecological validity with both system access and researcher control. While generative AI systems with low usage barriers are available off-the-shelf, they can be difficult to focus on the task at hand without blockages, hallucinations, or other non-task-related issues that derail engagement. Alternatives are limited prototypes, mock-ups, or design fictions that can be too far removed from the actual capabilities of the technology, and lead to participant responses being based on an imagined caricature of AI conditioned by media narratives.

The participatory prompting method, first proposed by Sarkar et al. [63], aims to bridge this gap. Participatory prompting is a user-centric research method for eliciting AI assistance opportunities. The method combines principles of contextual inquiry and participatory design [69], in which researchers mediate participant interactions with a real generative AI system.

In a participatory prompting study, researchers first identify a domain problem and the relevant form of generative AI system. They experiment with different prompting formulations to elicit targeted responses, and then recruit participants who bring self-selected scenarios within the domain, and potentially also resources to be used. Researchers then conduct sessions in which participants work through their scenarios in multi-step turns (illustrated in Figure 1).

A key advantage of participatory prompting over low-fidelity prototyping and Wizard-of-Oz methods is that it grounds studies in "actually existing AI" [73] capabilities rather than simulations or speculative design probes. A benefit in comparison to experiments with fully functional prototypes is that it can leverage off-the-shelf AI systems with minimal engineering costs, and flexibly explore different use cases during a study, for which a functional prototype might be too constrained.

Participatory prompting studies also have an advantage over some forms of purely observational studies, because by virtue of being researcher-mediated, participatory prompting can account for discrepancies in participants' *a priori* prompting strategies, enabling participants to be appropriately challenged while not fixating on

practical problems in generative AI usage that are not relevant to the research questions.

Participatory prompting may involve various kinds of researcher mediation. The format used in this study is that of the researcher-as-relay. In this form, a participant poses an open-ended query to the researcher. The researcher reformulates the query using prompting strategies, and sends this prompt to the model. The participant reviews, reflects on, and builds upon the model's response to determine their next query, guided by the researcher. The 'dialogue' with the system and the participants' reflections, together with optional quantitative measures of interactions such as response satisfaction, can then be analysed. Other formats could include researcher-as-guide, where the participant directly interacts with the AI system but discusses their thought processes with the researcher.

The interaction between the participant and the researcher creates valuable opportunities to elicit participant reasoning. First, the researcher can probe participant reasoning turn by turn (or sets of turns, as appropriate), to capture sequential expectations and responses. Second, when the researcher is involved in the translation of participant queries into prompts, participants may see and comment on the researcher's prompting strategies as reference point in comparison or contrast to what the participant might have done without guidance. While in some research methods this could be seen as influence or bias, in the participatory design context, this collaborative engagement on solving the problem of prompting reveals the differences and similarities between users' and technologists' assumptions, methods, and success criteria, and hence where either social or technical interventions or features are needed.

3.2 Preparation

The first step of the participatory prompting method is to choose a suitable functional generative AI system as a representative of AI capabilities more broadly. This involved careful evaluation of the possible alternatives. Four candidates were considered: OpenAI Playground, OpenAI ChatGPT, Google Bard, and Microsoft Bing Chat. We tested the systems by eliciting multi-stage guidance for data analysis through example queries, examining the quality and potential reception of each system's responses in a manner similar to a cognitive walkthrough [36].

We noted how particular design decisions in each system shaped and imposed limitations upon discourse. For instance, ChatGPT, Bing Chat, and Bard, as consumer products, incorporate "guardrails" against content considered inappropriate by the system developers, e.g., violent or sexual content. At the time of our study, Bing Chat restricted conversational exchanges to fifteen turns. In contrast, the OpenAI Playground allows more unrestrained exploration, and options for model and parameter selection. For our purposes, such constraints did not definitively preclude any options. However, the proprietary and opaque nature of commercial systems does restrict controllability, and this may render them unsuitable for some investigations.

At the time of our study, Bing Chat had the unique ability to source knowledge from the Web within replies. In a data-driven sensemaking activity, this can enhance suggestions at each problem-solving phase, such as by identifying relevant open datasets, and retrieving tutorials and recommendations for tool features (such as spreadsheet formulae). We found that information from the Web significantly improved the breadth and utility of the AI responses. This outweighed other limitations, and we therefore chose Bing Chat for our study.

The next step is preparing prompting strategies for the study. The challenge of developing reliable and effective prompting strategies to optimize large language models' performance has been comprehensively documented [84]. Users, particularly non-experts directly engaging with AI systems, struggle to devise suitable prompts to elicit high-quality responses. To overcome this limitation, the participatory prompting protocol involves the mediation of an expert researcher with knowledge and practice of prompt design, to help users formulate suitable prompts. Besides this, the mediation also helps users rapidly iterate on queries, can help users focus on the relevant aspects of the interaction and avoid distraction from incidental elements of the user interface that are not relevant to the research questions, and eliminate variations in typing speed, as the researcher relays user queries to the system, rather than the user interacting with it directly.

For our study, three researchers individually experimented with developing prompting strategies for Bing Chat across four weeks, using a range of real data-driven sensemaking tasks drawn from their own personal or professional experience, including quantitative analysis of a poem text, choosing a bar to visit with colleagues, developing a spreadsheet for evaluating World of Warcraft game strategies, selecting a car for purchase, and choosing a plot type for a statistical report. These interaction logs and screenshots, successful and unsuccessful prompting strategies, error recovery methods, and other observations were catalogued in a shared repository.

Through this process, we identified that despite having the capability to do so, Bing Chat did not consistently use information from the Web, render tabular data visually as a table, or attribute its sources. We developed prompting strategies through which this behaviour could be reliably induced when needed. It often provided multiple options without further support to the user for choosing between them; we developed prompts (e.g., "use information from the Web", "cite your sources", and "show result in a table") to induce more such support when needed.

At the end of the experimentation period, the researchers convened to negotiate and codify a list of prompting strategies and how they would be applied in different situations that might arise during the study. Despite having access to a thus carefully designed "bank" of prompting strategies, we found that in practice a lot of ad-hoc and in-situ adjustment was needed (discussed in Section 5.5).

3.3 Pilot

We conducted a pilot study with a convenience sample of 2 regular spreadsheet users. The pilots revealed that it can be difficult for participants to choose a suitable seed problem that is complex enough to require generative AI assistance but simple enough to describe concisely. To address this, more guiding questions were added to help participants during the problem elicitation phase. We also recommended that participants prepare a problem in advance of the study if possible. Terms such as "data-driven decision-making" were unclear to participants and had to be clarified.

We found that 5-6 turns could be completed in the allotted time (45–60 minutes), eliciting detailed qualitative insights despite the small number of turns. The turn-taking phase could be extended if needed. Reflecting on responses and choosing a next step was the most time-consuming and insightful aspect of each turn. This led to us changing the Bing Chat system mode from "precise" to "creative" (the exact nature of these modes is proprietary, but the salient aspect is that the latter is more verbose and the responses typically carry more information), to give more to reflect on and help guide next queries.

If early responses were generic or unhelpful, participants lost motivation. To counter this, advancement questions were added to the protocol to suggest ways forward, like rephrasing queries. Participants also tended to use short queries typical of web searches, which were more likely to result in generic responses; we added guidance to explain that longer, conversational queries were more effective.

We included steps for experimenters to more deeply understand participant expectations, including desired output types, to avoid multiple incremental prompts, which while useful to study, could slow down the progress of the task and thus impair the study of more complex interactions with the AI. Prompts also needed to refer to previous outputs to maintain consistency in the system responses; we updated the protocol to include this. While not initially part of the protocol, we noted that it was useful for participants to explore and verify outputs online, thus navigating temporarily away from the chat session. Finally, we revised the protocol so that participant speculations about helpful system capabilities could be immediately tested, and barriers to sensemaking were specifically elicited.

3.4 Participatory prompting sessions

We conducted a study with a fresh sample of spreadsheet users (N=15, 5 women, 0 non-binary, 10 men). Participants were recruited partly via email from a database of spreadsheet users who signalled interest to take part in research studies, and partly through a recruitment consultancy firm specialising in user research with African participants. Participation was voluntary, and all participants were free to withdraw from the study at any time without penalty and without having to cite a reason. All recruited participants were compensated with a USD $50 (or local currency equivalent) gift voucher for an online retailer. Participants read and signed a consent form detailing the study format, data collection, and risks. The study method and data collection protocols were reviewed and approved by our institution's ethics review board.

Participants provided demographic information (Table 1) relating to their experience with spreadsheets, programming, and generative AI via a survey. We directly use the spreadsheet and programming experience [62], and generative AI experience [63] survey items, and corresponding integer coding scheme, from previous work. Participants varied in spreadsheet usage (1 beginner, 7 experienced and basic usage, 7 experienced and advanced usage) and generative AI usage (3 never used, 1 casually use, 6 occasionally use, 5 regularly use), as well as programming experience (7 never programmed, 3 novices, 3 moderately experienced, 2 experts). Participants resided in various locations (7 in Africa, 3 in Europe, and 5 in North America).

The study sessions were conducted remotely using a Microsoft Teams video call, with the researcher handling the interaction with Bing Chat which was screen-shared with remote control to the participant, so that they could view and explore the results.

Experimenters first elicited an example problem from the participant before entering the turn-taking phase as previously outlined. At each turn the participant was asked to read the response from Bing Chat and reflect aloud on the usefulness of the response and if anything was surprising, inspiring, or confusing. The experimenter would then ask the participant if they wanted to follow up with Bing Chat on the response, ask another question relating to the original problem, or pivot to a new problem they were interested in, thus proceeding to the next turn on the basis of the participant's response.

We tried to stay neutral, passive, and open-ended in terms of affecting the topic that the participant wanted to work on (and how to follow up in each turn). However, many participants found it hard to imagine what they would want to do with Bing Chat, and then how to follow up its responses. As such, we had to be active in eliciting their thought process and moving the study forward, by suggesting options to follow up and drawing their attention to certain aspects of the output.

The degree to which the researcher needed to intervene to reformulate the participants' query into a prompt varied depending on the context. At one extreme, the intervention was extremely minimal: a participant would dictate a query for the researcher to type verbatim. At the other extreme, when the participant found it challenging to articulate their need concisely, the researcher proceeded by writing a candidate query and asked the participant to confirm or disconfirm it, e.g. "does this prompt capture what you wanted to ask the system to do?" In between these two extremes, we would express their query directly, but suggest the addition of context (what columns existed in their spreadsheet), or append a prompt from our list of strategies (e.g. "output a table", or "cite your sources"). The level of researcher intervention and the impact of query reformulation on our findings is a complex issue and we address the trade-offs in detail in Sections 5.4 and 5.5.

Participants worked with experimenters through several turns between Bing Chat prompt and response until the task was achieved, or the allotted time was reached. Finally, participants gave further feedback on their experience through semi-structured interviews.

3.5 Analysis method

We transcribed the audio recordings of participant think-alouds and interviews. One researcher initially organised participant quotes through affinity mapping [3] into four broad categories: remarks about interaction with AI, remarks about workflows, remarks about barriers encountered, and remarks about specific features. The organisation was negotiated with a second researcher. This categorisation was not the final analysis, but a data management step to facilitate the final analysis.

The final analysis was a directed, negotiated coding between two researchers, with the aim of discovering emergent themes. We coded remarks relevant to the question of how AI assistance can support data sensemaking workflows according to the main categories of activities identified by sensemaking theory. We report

Table 1: Participant profession (self-reported description). Spreadsheet experience ((1) Little or no experience; (2) Some experience, but I'm still a beginner; (3) A lot of experience, but my use is basic; (4) A lot of experience, and I use some advanced features; (5) A lot of experience with many advanced features). Programming experience ((1) I have never programmed; (2) I have learnt a little bit but never used it; (3) I know enough to use it for small infrequent tasks; (4) I am moderately experienced and write programs regularly; (5) I am highly experienced). Generative AI experience ((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).

P No.	Profession	Spreadsheet Experience	Programming Experience	Generative AI Experience
1	PhD candidate (Anthropology)	3	2	2
2	Consultant	3	2	4
3	Statistician	4	2	4
4	Account Officer	3	2	2
5	Student (Geophysics)	4	4	5
6	Software Engineer	2	4	4
7	Data Analyst	4	2	5
8	Software Engineer	4	4	5
9	Student (Informatics)	3	2	5
10	PhD student (History)	3	2	2
11	PhD candidate (Computer Science)	4	5	5
12	PhD student (Anthropology)	3	3	4
13	Computer Scientist	4	5	3
14	Student (Philosophy)	4	3	4
15	Student (Nursing Science)	3	3	4

our findings organised by these frameworks in Section 4.2. We coded remarks relevant to the question of how AI assistance can create barriers for data sensemaking workflows according to the *iterative goal satisfaction* framework, described in Section 4.3, which also reports the results accordingly.

Our final analysis relied on the application of prior theoretical frameworks to supply the basis of code organisation, and thus differs from the more commonly applied inductive approach [4]. We were not developing a reusable coding scheme and quantitative measures of inter-rater reliability are inappropriate here. Instead, in accordance with qualitative coding best practices, the two researchers iteratively discussed their interpretation of the findings and negotiated each disagreement until it was resolved [42]. Our analysis focused on identifying and characterising themes, rather than on quantifying the prevalence of each in our sample. As such, it is not helpful to be concerned with the precise participant counts associated with each identified theme, although this may be inferred from the list of participant IDs mentioned under each theme.

4 RESULTS

4.1 Overview of tasks

Recall that we did not design study tasks *a priori* but rather developed them in a participatory manner with each participant at the start of the study using task elicitation questions. This resulted in a set of unique but highly ecologically valid tasks that were directly relevant to each participant. Participants explored a variety of data sensemaking tasks during the study, most related to their professional work, but also some personal workflows such as job searching or scheduling a pub crawl. The full list of participant tasks elicited is given in Table 2.

Each participant's task involved key sensemaking activities when seeking assistance with different aspects of data analysis. As part of the information foraging loop, participants often began by describing their data and its format (e.g., row and column descriptions), and their overall analysis goal. Some participants even began by requesting that Bing Chat generate or find example data (this tendency is corroborated by previous studies of analysts, which have found that analysis often begins in the absence of data [40]). For example, P10 requested that Bing Chat provide a list of potential career paths they could follow based on their skills and experience as a History PhD candidate. When P10 found a career path that was interesting to them (archivist), they continued by requesting the requirements for that career and example open positions that they could apply to. These interactions represented data filtering and searching within the information foraging loop.

As part of the sensemaking loop, most participants asked Bing Chat for help with formulating potential research questions (*hypothesis generation*) or strategies for analysing their data (P1-5, 8, 9, 11-15), code or formulas for a specific analysis (P1, 2, 5, 8), or step-by-step instructions for applying Excel features such as filtering and visualizations (P4, 7) (*hypothesis testing*). For example, P9 first asked Bing Chat about data analysis strategies using Excel for data they collected in a survey. P9 then iterated with Bing Chat to generate potential research hypotheses and analysis plans for testing them.

4.1.1 Example turn-taking sessions. An illustrative example of a complete turn-taking session (P1's) is described as follows. P1 wished to analyse a dataset about "cooperative behaviour in literature" they had collected. P1's first mediated query told Bing Chat they had a spreadsheet with data, where the *"rows are the data*

Table 2: Overview of the tasks developed in collaboration with participants. Columns (left to right): participant ID, brief description of task, first prompt issued for that task to Bing Chat, and count of turns taken over the course of the task.

P	Description of task	First prompt	Turns
1	Performing a literary analysis with spreadsheets	I have a spreadsheet with data. In rows are the data for "tales" and in columns are the data for "cooperative behaviour". Each cell contains an example of a cooperative behaviour in a certain tale, e.g., "brother saved brother". I need a way to code each cell according to different categories. Explain how to use a spreadsheet for this with an example.	3
2	Categorizing age data in a spreadsheet	I have data about people with a column for their age. I need to regroup the people with age between 18-35 into categories Gen X, Gen Y, Gen Z. Explain how to use a spreadsheet for this with an example.	2
3	Analysis on how discrimination impacts workplace performance	I am trying to determine the extent to discrimination affects employee performance in my company. What data is required? What online data sources may be relevant? Explain how to solve this in Excel with an example.	4
4	Creating a reusable expense tracker in spreadsheet	I am making a form in Excel where people have to categorize their business expenses. For each expense the user has to choose a category. Instead of typing out the category I want them to be able to filter on the cell. Explain how to do this in Excel with an example.	4
5	Data analysis exploration of ridesharing data	I have a dataset of rides taken on a bike sharing service. For each ride we know the rider type (casual or annual member), start and end points time and location, and bike type (classical or electric). I am trying to understand what differentiates the usage of the casual and annual members. Suggest a data analysis strategy for solving this problem using Excel and R.	4
6	Apartment hunting organization	I am looking for an apartment. I have a spreadsheet with the data about various apartments and the following column headers: Name, Google Rating, Location, Distance from Msft, 2Br Price, # of Baths, Sq. ft., Move-in Date, Notes. Explain a few ways I can analyse this data in a spreadsheet to help me make my decision.	6
7	Strategies to share data analyses	I am a data analyst who does his analysis in Microsoft Excel. It is challenging to share the findings from my analysis because Excel files stored on my computer are not live, in the sense that I am the only one who can have access to it at any one point in time. Suggest a few solutions to this problem.	3
8	Writing Javascript code	Write me javascript code that adds 2 days on top of current today and has these conditions: - When it's Monday/Tuesday/Wednesday/Thursday, then the day would be Wednesday/Thursday/Friday/Saturday - When it's Friday, then the day would be Monday - When it's Saturday/Sunday, the day would be Tuesday	6
9	Requesting analysis strategies for an HRI survey	I am a researcher studying Human-Robot Interaction. I have data from a survey in which 100s of respondents were asked to rate their perception of two robot voices on a Likert scale from -3 to +3. I am interested in the differences between voices. Suggest a few analytical strategies for solving this problem. Suggest how to implement the strategies in Excel.	5
10	Potential career paths based on background	I am a history PhD candidate with strong research, writing, and education skills. I would like to know what potential career paths I could follow based on my skills when I graduate.	7
11	Comparing object detection models	I'm interested in the performance of object detection models. Recommend ways to compare these models. Focus on accuracy, speed in ms, and size of the model.	3
12	Finding strategies for literary analysis	I am interested in narrative structure. Recommend appropriate frameworks for analyzing narrative text within short stories. Cite your sources.	5
13	Performing data analysis on student grading data	I want to disaggregate categorical and numerical grading data by gender as part of my research study. I want to calculate the difference between the median grade and each gender. How do I do this in a spreadsheet? After this is done, I would like to export the data to latex and create a visualization of the data, can you give me the steps to perform these tasks?	5
14	Discovering data about lifestyle impact on diabetic patients	Are type-2 diabetic patients compliant with lifestyle modifications? Cite your sources please.	3
15	Understanding what factors impact working for corporations	Does university curriculum prepare people for work in corporate organizations after graduation? I want to know the answer and all the potential factors that contribute to this. Please cite your references.	7

for 'tales' and columns contain the data for 'cooperative behaviour' of a certain tale (e.g., 'brother saved brother'). The query indicated that they *"need a way to code each cell according to different categories, explain how to use a spreadsheet for this with an example."* Bing Chat's response confirmed its understanding, suggested options like using Excel VBA, thematic analysis, Google Sheets, and SPSS, and gave an example table and showed how thematic analysis might be applied to complete the task.

P1's mediated follow-up response was to quote the fourth suggestion (use SPSS) and ask how this could be done in R with another example. Bing Chat's response explained how to use R in R Studio, and provided R code to complete the task and visualize the output. Each section of R code also contained a natural language description of the code.

P1's final mediated query asked for the same code but with their specific categories in mind by asking *"show me how to do it when the categories follow Hamilton's categories of biological cooperation"*. P1 was satisfied with the response, and this ended the turn-taking session for this task.

Another example, in brief: P6 was currently apartment hunting, so their first turn involved asking the Bing Chat to recommend ways to sort apartments based on the data they had previously collected. Subsequent turns involved recommending alternative apartments based on their criteria (by searching the Web). Finally, the participant requested Bing Chat to draft a letter to landlords to request extra information that Bing Chat recommended P6 collect, since it was missing from the spreadsheet they made.

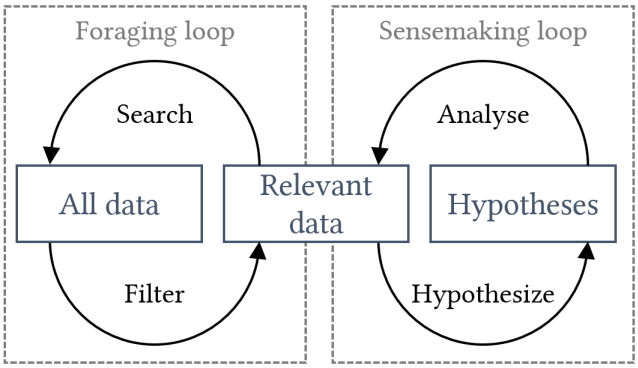

Figure 2: A simplified version of the data analyst's process, adapted from Pirolli and Card [47]. The process consists of (1) a foraging loop, in which the analyst transitions back and forth between a large set of data and a smaller set of interest through searching and filtering, and (2) a sensemaking loop, in which the analyst transitions back and forth between relevant data and hypotheses through hypothesis generation and testing of the hypotheses (analysis).

4.2 Data sensemaking workflow support

Participants saw generative AI as a versatile tool that enabled various stages of data sensemaking. P11 saw generative AI as useful for *"any part of a workflow"*, from *"starting a new project"* to *"preparing PowerPoint slides"* for presenting the project. Several participants thought generative AI supported their workflow by making their *"work easier"* (P2, 4) by streamlining the search for *"the desired result"* (P4), adding new perspectives on how to analyse their data (P3), and *"scaffolding"* the solution to a task to *"speed up the process"* of working with data (P1). P8, a business owner, believed generative AI would *"save time"* and *"greatly decrease cost"* for many of the tasks they needed to perform.

The data analyst's work process as characterised by Pirolli and Card [47] consists of two loops: an information foraging loop whose purpose is to identify a smaller set of relevant data out of a larger set, and a sensemaking loop whose purpose is to generate and test hypotheses. A high-level overview of this process is summarised in Figure 2. This overview is helpful for further delineating aspects of our participants' experiences, presented in the following sections.

4.2.1 Generative AI in the information foraging loop. Several participants compared generative AI to traditional search. P9 thought that generative AI workflows improved upon the information overload caused by traditional search workflows since *"search engines will give you multiple results, and it's very messy, but this [Bing Chat] directly gives one thing to do."* P14 also enjoyed that generative AI output was specific to their question, while search results would require *"converting the result"* to your specific task.

However, participants also cited concerns about the ability to apply generative AI tools to information foraging. P10 said their PhD research was *"super-duper niche"* and frequently required them to *"travel to archives all over the world"* to find data and thought generative AI would be unable to assist them for these types of tasks, because unlike textual data from the web, heterogenous archival data may not have uniform and easily accessible indices, and might be highly unstructured, mixed-media, only partially digitised, and therefore difficult or impossible for generative AI to operate over.

4.2.2 Generative AI in the sensemaking loop. Participants noted the opportunity to be assisted both in generating hypotheses and in identifying strategies to test them.

Hypothesis generation. P4 thought generative AI was useful to *"have another perspective, like conversing with another person to see how their perspective is different from yours"* which *"could be inspiring"* for their own workflow by *"giving a whole new way to do a task."* Others believed generative AI was useful for brainstorming (P10) or getting unstuck by using the AI to give alternatives or options to explore (P7). P6 appreciated how Bing Chat's response considered aspects of the problem that they *"didn't really get a chance to think about... so, it's good that Bing Chat was able to cover that as well."* P13 said they were *"directly inspired"* by Bing Chat, as it allowed them to *"move further in the research analysis"* by introducing methods to do their task *"in a different way"* than they had planned.

However, some participants were sceptical of using generative AI for their creative process (P1) or forming research questions (P9), and instead saw its primary application as being for specific data analysis tasks. This could be due to concerns about personal agency in the analysis process; for instance, P9 thought that even when generative AI generated useful text, it would still *"miss your own style of writing"*.

Hypothesis testing. Participants noted how Bing Chat helped them avoid *"spending ages try to figure out code"*(P1) and *"insightful"* when offered analysis techniques they *"had never thought of"* (P3). Participants also liked when Bing Chat provided *"step-by-step process on how to get a chart in Excel"* that gave *"headway on how to get the desired results"* (P4), and *"some kind of direction"* (P5). P5 thought generative AI enabled this understanding by both *"streamlining your thought process [...] with step-by-step instructions"* and giving *"inspiration on how you can analyse data."* P9 similarly valued the *"step-by-step [instructions] on what to do"* and *"other possible strategies"*.

P12 likened generative AI to "rubber duck" debugging, an informal technique from software engineering where in order to fix a bug, the programmer explains their problem aloud to an inanimate object (archetypically, a rubber duck, hence the name) – the idea is that verbalising the problem can often trigger the understanding and insight needed to fix the bug. P12 stated, *"it's like a rubber duck that actually talks back and is useful."*. This analogy highlights how, even if the AI system does not introduce new information, it may facilitate problem-solving and sensemaking by providing a channel for the reification and refinement of the user's thought process.

An additional benefit to the sensemaking loop was the ability to *learn* new skills as part of the analysis process, which enriches the space of hypotheses it is possible to generate and test. These can be fairly straightforward technical skills, such as learning particular features of spreadsheet software. P7 had *"a good learning experience"* in using an unfamiliar formula. P4 similarly *"initially thought you could only create bar charts with a pivot table"*, but learnt from a Bing Chat suggestion that they *"could just select the particular cell to create and insert the bar chart."*

There is also the potential for learning broader skills. P5 saw Bing Chat's recommendations of unfamiliar functions and statistical packages as a potential *"learning direction on how to go about carrying out descriptive statistics and visualizations to assist with that task."* P10 saw generative AI as a potential learning surface that assists in critical thinking, because when P10 asked for a biography of Thomas Jefferson, the response did not initially raise the problematic issue of Jefferson's slave ownership, which P10 expected. P10 reflected that generative AI could be used to explore *"what kind of questions we can ask and what kind of information is being omitted"*. This finding aligns with the constructivist theory of learning in interactive machine learning systems, which holds that users construct mental models of their task through iterative exposure to AI model responses [52].

4.3 Barriers to sensemaking with generative AI

Rather than thematising barriers according to the analyst process, we found that it is more helpful to consider them in terms of a workflow we term *iterative goal satisfaction*. Broadly, this is the process by which a user satisfies a series of goals with AI assistance.

The iterative goal satisfaction workflow is presented in Figure 3. The user moves through different phases: goal formulation, query formulation, and response inspection. There is an outer "goal iteration" loop as the user attempts to achieve a high-level goal, and an inner "prompt-response-audit" loop as the user attempts to achieve specific steps towards that goal. The elements of this workflow are as follows:

- *Goal formulation*: the user reflects on their goals, needs, intents, and research questions, and identifies a need for assistance where AI could be applied.
- *Query formulation*: the participant composes the information, context, and data that the AI might need to address a goal (in our study, the query is relayed to the mediating researcher who then further shapes it into a prompt). Query formulation can proceed directly from goal formulation, or it may be in the context of iterating on a previously identified goal, as a result of having inspected a previous response (described next).
- *Response inspection*: the participant checks for readability and relevance to the goal. If the output is readable and relevant, the participant reads with the aim of deeper comprehension, checking quality and correctness. If the response failed any checks, participants would either reformulate their query to attempt to elicit a better response, or change their overall goal. The sequence of query formulation and re-formulation in response to deficiencies identified by inspecting the output maps directly to the *prompt-response-audit* cycle described by Gordon et al. [17].
- *Response acceptance*: when the AI response satisfies their goal, participants might exit the goal iteration workflow entirely (e.g., to apply the results by copying a formula into a spreadsheet, or add code to their IDE), or develop a new goal. We thus observed two situations in which participants could develop entirely new goals: either as a result of having their previous goal satisfied, or a "pivot" as a result of inspecting a response and reflecting upon it. Consequently, we broaden Gordon et al.'s prompt-response-audit loop by showing that there are two distinct reasons for exiting it, and that it is itself part of a larger goal iteration loop.

With this picture of the iterative goal satisfaction workflow, we are in a better position to understand the barriers to effective sensemaking with generative AI encountered by our participants. Broadly, these fell into three categories: barriers to *query formulation*, barriers to the *utilisation of responses*, and barriers to *verification and trust*. We detail each of these in turn.

4.3.1 Barriers to query formulation. Participants faced difficulties in understanding, gathering, and expressing their request. These are difficulties they experienced in their own articulation of their needs.

Detailed expression of intent. Part of the challenge was in fully articulating their need. Participants had trouble *"wording it in the right way that the AI understands [...] writing [what is in your head] down is the hard part."* (P1) and giving *"a very explicit explanation in the prompt that is detailed"* (P13), though P1 noted that Bing Chat could generate helpful responses for *"convoluted"* questions (i.e., prompts worded in a noticeably vague or unnatural manner). P5 similarly was frustrated by their inability to *"really define the problem because there are a lot of components, a lot of things to factor in before clearly defining the problem."*; it was challenging to *"be as detailed as possible when you are putting information [into a prompt, but], you can't just be lazy about it and get the most useful answer [...] you have to feed [Bing Chat] with as much detail as possible."* Such difficulties led to P9 asking *"where should I learn this kind of stuff when I'm chatting with Bing Chat"*.

Barriers to query formulation resulted in, but also stemmed from, inadequate output from the AI, with P12 stating *"it is frustrating to figure out what is it that is being miscommunicated."* P8 pointed out that *"generative AI can't read your mind, so you just have to formulate your question 'correctly'"*, and they would *"be annoyed at myself for not writing the prompt correctly"* rather than blame the system for an inadequate output. Other participants similarly attributed this issue to their having *"communicated 'wrongly' at first"* (P4). P2 observed that the prompts that the experimenter wrote were *"very different"* than their own in that they were more specific and *"direct"*. P2 described their current prompting methods as *"too general"* in comparison, and having difficulty understanding *"where to start from"* when interacting with generative AI.

Participants developed strategies to manage the challenge of detailed expression. Participants used follow-up prompts to *"ask it specifically to focus"* on a specific part of their data (P3) or on a *"specific list of categories"* (P4). P1 thought the solution was *"just asking the right questions"*, which meant being *"clear and real specific in the details"*, though this was challenging and left them *"a bit confused."* P13's received a response localised to a different country, so they realised they should *"be even more specific"* about their location. P5 decomposed their queries to *"streamline them to focus on things I actually need and not just suggest the entire data analysis strategy."* P12 thought they would improve their prompts by practising through *"having to use it over and over again."* Others developed more ad-hoc techniques, such as avoiding acronyms (e.g., 'MSFT' instead of 'Microsoft') (P6), to reduce the likelihood of miscommunication.

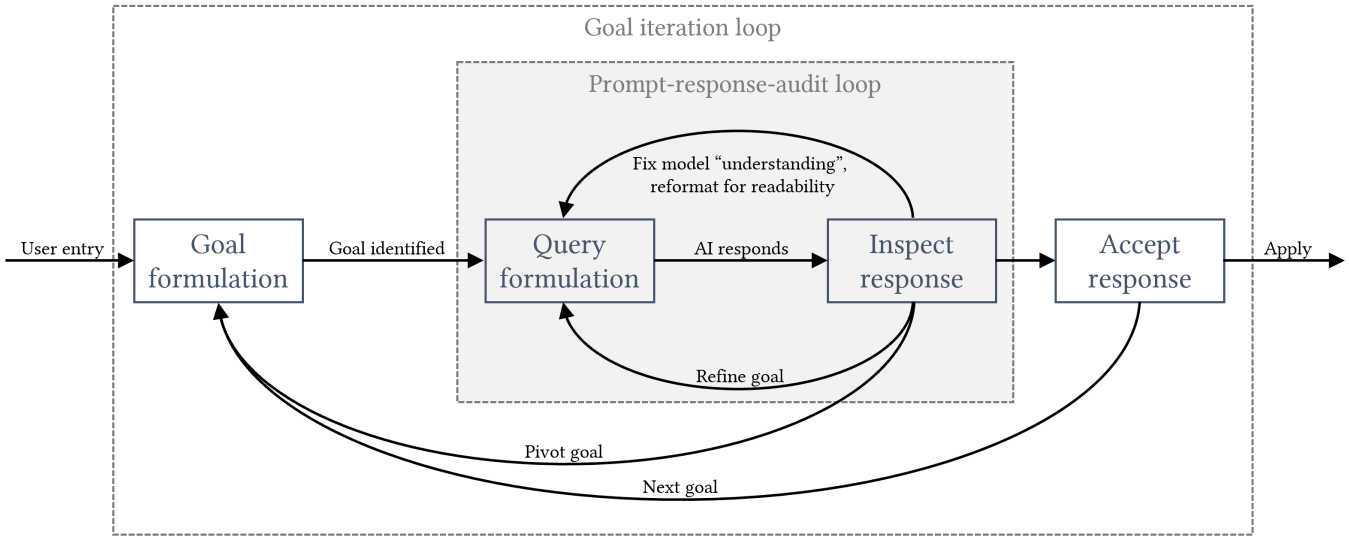

Figure 3: The workflow of iterative goal satisfaction with generative AI.

Determining and expressing context. Participants were also challenged by the need to determine what contextual information was relevant to fulfilling or interpreting their request, and then articulating it. For example, after being recommended 'thematic analysis' as a way to analyse their data (which was not applicable to the kind of data they had), P1 noted that giving context (in this case, information that would enable the system to rule out thematic analysis as a plausible method) to generative AI was important for making sure AI suggestions *"actually work"* for the task and data. Participants drew a comparison to their experience of human-human collaboration. P12 found giving this context to generative AI was more difficult than giving it to a human co-worker, as they usually framed questions with what they had attempted previously and what went wrong before asking *"what should I change?"* when asking a co-worker for help. P12 felt that their co-workers are *"more familiar with examples"* that they would provide as context to their problem, and worried this context seemed more difficult to convey to the system.

Researcher mediation of prompts occasionally impacted participant awareness of these barriers. For example, researchers asked participants for needs around the data format of the response or gathered extra context about the problem being solved, which revealed to participants the specific prompting strategies we applied. Researcher-mediated queries served as a reference point for participants to compare their own experiences in forming queries. While some aspects of effective prompting could be handled by the mediating influence of the researcher and thus "smoothed over" from the participant's perspective, as the examples above show, even with such guidance, participants are challenged by the activity of expressing their intent.

4.3.2 Barriers to utilisation of responses. Participants faced barriers to being able to effectively use the responses, such as an overwhelming volume of information; poorly or incorrectly formatted results; output that while not strictly *incorrect* was nonetheless incomplete or inadequate in some other qualitative manner; and responses which were not easily intelligible because they referred to unfamiliar concepts.

Volume of information in the response. Bing Chat's responses were often lengthy, likely due to our choice of using Bing Chat's "creative mode" which is designed to be more verbose. This required the participant to read several paragraphs of text. P2 experienced information overload with the text results from Bing Chat, which they originally expected to be returned as a table or spreadsheet. P3 complained about the level of technical detail in one of the responses, finding it *"not easily understandable for someone who is being introduced or does not have much experience in statistics"*. This points to the need for tailoring responses according to user expertise. When given different options to complete a task, P13 thought it was useful, but also *"excessive information"*.

Excessive length also applied to generated code. P8 received *"additional unnecessary code"* based on what they asked for, but nonetheless believed the result to be correct. In follow up, P8 asked Bing Chat multiple times to *"make it [the code] shorter"*, until Bing Chat successfully reduced a 15 line function into 3 lines.

Participants' preferences regarding a suitable default length and contents for generative AI output varied (P3, 8, 10-12). For example, P3 preferred a specific order of generative AI output: first the answer, then an explanation of that answer, and finally an example of how to implement it in Excel. P8 shared a preference for seeing examples and expected Bing Chat responses go beyond *"just some sort of summary"* by producing examples that apply Bing Chat's recommendations (e.g., showing how the A/B testing model Bing Chat generated might apply to a video advertising campaign for a company). P5 considered extra or irrelevant results from generative AI as harmful when under *"tight time constraints"*, as they *"would not want to spend time on things that are unneeded to complete the task."* P10 wondered about balancing *"how much versus how little information"* that generative AI puts into a response, and how

they could control this amount of information produced to their preferences. P12 expressed appreciation for responses that were *"a good balance"* of information *"between bullet points and short paragraphs"*, and *"not just a two sentence answer that doesn't give any information."*

Goal-satisfaction of the response. Participants could face barriers in progressing with their task if the results only repeated what they already knew and did not add any further information, or if the results were incomplete, or incorrect, or too broad.

Some participants were suggested solutions they already knew about, but which could be useful for novices *"unaware of these methods"* (P7) or *"starting from scratch"* (P11). P7 requested *"three more suggestions"* to elicit more unfamiliar solutions. Occasionally, the model would fail to interpret very basic and clear instructions correctly. For example, P12 was surprised that the system incorrectly applied a literary analysis framework to one story ("The Glowing Coal") when specifically asked to apply the framework to a different story ("ATU 333, Little Red Riding Hood"). P12 wondered if the data needed to complete the task was not available to Bing Chat.

Participants also received incomplete responses from Bing Chat (P11, 13-15). P11 said they needed Bing Chat to provide justification for its choices. P13 and P15 both had replies that were useful, but incomplete since it failed to address every part of their question. E.g., one result was *"not able to achieve the task"*, since it missed out the step to *"convert a column"* (P13).

Other participants noted that some responses were not applicable to their specific preference, but could nonetheless be helpful in other situations (P1, 3). P14 considered a response to be *"just an introduction"* to the topic and not applicable to their task. P4 wanted *"the data to be shown in a different form."* Similarly, P9 asked for a data visualization which Bing Chat provided, but P9 instead preferred a bar chart instead as it was *"much more useful than a pie chart."*

Moreover, model "misinterpretations" could also function as a sort of tolerance for imprecise or incorrect querying: P11 was surprised that Bing Chat ignored part of a prompt and gave what was more likely correct when P11 tried to modify a table of object detection models produced by Bing Chat by asking it to *"add a column of the platforms (e.g., iOS, Android, Raspberry Pi) supported by each model"*. Bing Chat instead added a column with values like *"CPU, GPU, DSP, EdgeTPU"*, which P11 realized was actually what they wanted to see in the table. P11 thought that had Bing Chat provided what was originally asked for it would have been incorrect, and instead preferred that Bing Chat intervene and recommend *"corrected information"* like it had.

Formatting of the response. Another issue that participants faced was getting responses in a useful format. For example, P11 attempted to compare popular object detection models and their characteristics so they might choose the best one, but the initial reply was a bulleted list of several models and their characteristics, which made it difficult to compare between models. P11 requested Bing Chat to produce a table that specifically compared *"accuracy, speed, and size"* and link to the code repository of each model. After inspecting the resultant table, P11 iterated to add columns for additional model properties. While P11 could have potentially created a detailed prompt to get a satisfactory answer with a completed table in a single step, P11 preferred to iterate and make incremental progress.

Similarly, when textual results were reformatted into a table P10 thought the results were *"perfect"* since the original outputs were *"very text heavy"*, but did not originally ask for a table. Thus, P10 placed the blame on Bing Chat's vague answers on the vagueness of the question they asked.

Intelligibility of the response. Finally, participants faced difficulty comprehending responses which referred to unfamiliar concepts (in a scenario where the participant was not expecting to encounter an unfamiliar concept). For example, when Bing Chat replied to a question with R functions that P5 did not know about, P5 requested an explanation of the functions and their relevance to the problem being analysed. In another example, Bing Chat recommended"Pivot Tables" to P9, which they were unfamiliar with, but P9 said they would *"just ask [Bing Chat] how to use pivot tables and for examples"* to learn more about unfamiliar concepts that generative AI recommends.

4.3.3 Barriers to verification and trust. Another category of barriers was associated with the work required to assess the reliability and validity of generative AI's output, both in specific instances of AI output but also in terms of developing a mental model for the system's strengths and weaknesses in different tasks, and an overall conception of trust in the system.

Verification strategies. Participants developed strategies for detecting and addressing incorrect output. To understand non-working code, P1 thought they would leverage traditional resources *"like textbooks"* that seemed *"slightly more professional"* than Bing Chat, or ask co-workers for help.

A common validation strategy was to follow the inline references (P10, 12, 14). Bing Chat provides references to the URLs from which it derives its responses using footnote-style superscripts (Figure 4). During the study, P14 followed a reference link, then described a previous experience with ChatGPT where it could not present similar reference links which P14 wanted to save in EndNote. P9 also compared Bing Chat to ChatGPT, finding the citation feature *"much better and more reliable"*.

Citations were seen as a fairness mechanism that *"gives credit where credit is due"* (P10). However, P12 found that checking citations *"becomes a process of verifying all the information it's giving you, and it might have just been quicker to find the sources yourself."* P6 said that they will *"have to verify"* each source and *"use those sources to further search"*. When performing data analysis, P2 said they need to *"validate that the data is from the right source"*, including the timestamps and recency of the data.

Source quality mattered. P7 preferred sites they *"already trust"*, rather than unfamiliar ones. P9 and P12 manually inspected the sources cited by Bing Chat for quality and relevance, which increased their trust of the output. P9 checked if a reference was *"a scholarly article or just a website"*, preferring *"trustable research"* publications, and inspected the publication date to ensure recency. P10 liked the citations, but if they were missing, they said they would just use traditional web search to verify the result themselves.

Some participants considered the seriousness of the task when deciding how much to trust and verify the response. For one task,

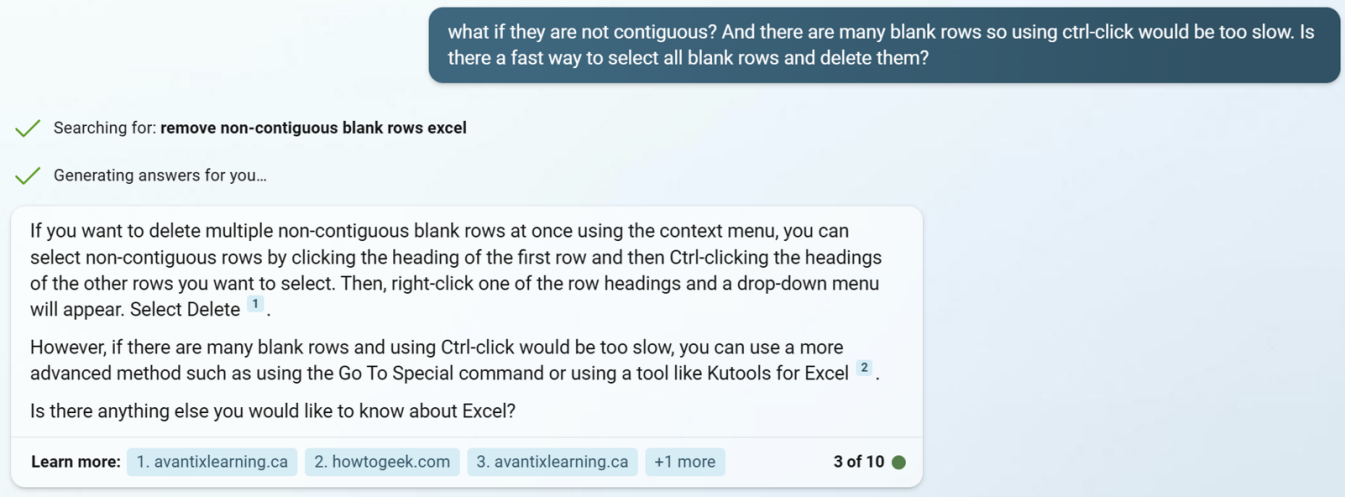

Figure 4: Bing Chat referencing UI (public design at the time of the study). The user message is in the dark blue bubble, top right. Bing Chat's response is in the white bubble, bottom left. Footnote-style superscripts indicate a supporting URL. The URLs are listed along the bottom of the chat response. A user can follow the superscripts or the links to read the web sources.

P10 said they *"trust the results, because this is such a low stakes query."* P12 said they would trust output if it *"sounded right"* to them, unless they *"really needed it to be right."*

Verification might also involve testing and applying AI output in a different tool (P1-3, 9-11). P2 would take generated Excel formulas and *"test it directly"* on their data, but P1 noted that this might be challenging without first *"cleaning up the data."* P3 also tested a generated formula *"as an example"*, and then edited it to fit their needs. P4 said when they were presented with step-by-step instructions, they would *"try it, and if it's not working out, do further research"* by searching online, asking colleagues for help, or watching video tutorials on YouTube. P9 had a similar approach to generated SPSS code: they first ran the code on test data to *"see if it makes sense"*, before applying it to their dataset. P13 stated that *"the only way to know the code is correct is to put it into an IDE."* P3 worried about errors, which decreased their trust of generative AI, saying that they *"don't rely on it [generative AI]"* and always rigorously verified any generated formulas.

Hallucinations. Hallucination, defined as "generated content that is nonsensical or unfaithful to the provided source content" [25], limited generative AI's usefulness for data analysis for our participants, because it was difficult to detect (especially when the hallucination is about a domain in which the user is not an expert) and time-consuming, requiring careful and detailed attention to every part of the output.

For example, P9 stated that they would not use it for literature reviews because of this risk. P6 felt a burden of *"always having to double-check and read every line"* of the response. P11 said that while their *"personal strategy is to verify everything"*, it was time-consuming and *"not always possible or feasible"* to do so. Moreover, P5 described difficulty in verifying generative AI output for domains they did not *"have a strong understanding in"*.

When Bing Chat started hallucinating data for P6's task, P6 said they started to *"understand when you should use [generative AI] and when you shouldn't"*. P6 subsequently formed a belief that Bing Chat was not able to index copyrighted media like books, and stated that the system ought to *"say 'I cannot access this book or its chapters' rather than continuing to make things up"*. P11 had an experience of *"generating a function in the code that looked very authentic, but didn't exist."* To mitigate the impact of such hallucinations, P11 aimed to *"always verify all information"*, but noted that users who *"blindly trust these AI tools can easily be misguided"* by hallucinations. P9 suggested that *"more specific prompts to focus on a specific topic"* might address hallucinations.

4.4 Explicit feature speculations

Participants on some occasions explicitly speculated about features that would help them with sensemaking. In the traditions of HCI research deriving from sociology and cognitive psychology, study participants are not conventionally involved in the direct design of products, and as such, explicit feature requests and speculations are treated as potential evidence of a deeper underlying need which may or may not be best satisfied by implementing the feature requested. On the other hand, since we are invoking the participatory design tradition [69], we are explicitly interested in participant's design speculations and consider them as first-class design contributions, at face value. We report these feature speculations in this section.

Application integration. Some participants saw a need for integration with the data applications they already used (P1, 2, 7, 11-13). For P13 to *"be comfortable in the analysis flow of using generative AI, it would be integrated in whatever system being used on the side, and not taking up the whole screen."* P7 thought that if they could *"do it all in just Excel"*, the generative AI to have access to charts and data within the spreadsheet, reducing the effort of *"going between different tabs"*. Further, P11 said that their analyses frequently ended up in slideshow presentations, so they wanted the generative AI to

leverage features of one app (Excel) and place them into the final app (PowerPoint). Similarly, P6 wanted to go from chat to spreadsheet by having Bing Chat create a spreadsheet for them, avoiding a *"very manual"* process of creating spreadsheets by iterating on *"what categories should be included and filling out information"* (P6).

P12 believed that app-specific generative AI would *"save a lot of time spent procrastinating"* such as *"going down Wikipedia rabbit holes"* (e.g., exploring various related topics that are not critical to solving the task at hand). However, P1 enjoyed the broad possibilities of a general generative AI chat and worried that when leveraging generative AI within an application, the AI might limit their suggestions to operate within that application, even if a better solution might exist in another application. Instead, P1 thought that both in-app generative AI and a general generative AI would be useful, where when the in-app generative AI failed to accomplish the task, the general generative-AI could act as *"the big boss who is like 'alright, we'll sort this out'"*.

Context. We previously noted that providing context was a challenge (part of the larger set of barriers to query formulation, Section 4.3). Several participants offered suggestions for sharing context more easily. P1 desired a way to easily include topics and keywords of interest. P4 wanted to give negative examples, to showing what *"does not exactly fit into what is wanted"* to tune the responses to be *"more specific"* to the goal. P1 wished to upload their entire dataset and *"have it [the AI] go"*. P5 wanted to upload *"particular columns"* of their dataset as context for questions like *"what can I do with this particular column"* rather than getting *"generalized responses"*.

P11 described the need for chat histories they could revisit and reuse *"after months"* away, to *"pick up where we stopped last time and continue from there without redoing everything I did before"*. P7 wanted to go further and share chat histories with others, which might help collaborators understand the provenance of some analysis activity.

Formatting and modality. Participants saw a need for better intelligence and flexibility in output formatting. For example, P10 desired the data they received from Bing Chat to be in a table, which Bing Chat was able to provide after re-prompting. P10 then thought the data was *"organized nicely and not overwhelming"* and could be exported easily to other applications. P11 ran into a similar issue while comparing two paragraphs, noting that placing the data in a table and comparing the columns would be *"more useful"* than reading each in sequential text.

Participants also described how generative AI might go beyond text and into other modalities (P4-11, 13). Several participants saw videos, imagery, interactive maps, and other visualizations as improvements over purely textual output (P4, 10, 11, 13) depending on the problem being solved (P11). Video tutorials could provide *"further clarity to see the step-by-step process"* (P4). P13 described example visualizations provided by Bing Chat as inspirational examples for how they could themselves visualize their data. However, P13 worried that visualizations could also be distracting and take user attention away from the text.

On the input side, P8 also wanted to provide images and videos as part of a prompt to provide context to Bing Chat, instead of just providing text. P11 suggested that voice interaction would feel *"more natural, like talking to a human assistant."*

Anthropomorphisation and social cues. Some participants reflected positively on Bing Chat's ability to use emojis and seem *"friendly"* (P10, 13). P10 noted it was a *"almost human reaction"* and said it was *"nice to feel like you're talking to some sort of person or feel kind of happy [...] like texting a friend"*. However, P9 thought this style of reply *"felt strange"* and was *"confusing"* for them in the context of doing work with Bing Chat, since they felt like they had to make conversation with the chatbot rather than just getting answers from it.

5 DISCUSSION

5.1 Connections with related work

How generative AI conversations compare to search workflows. Participants in our study compared generative AI to traditional search workflows, finding that the linear, summarised and aggregated nature of Bing Chat responses required less effort in comparison to manually viewing multiple search results and developing a mental summary oneself (Section 4.2). The consumer-facing positioning of the Bing Chat interface is as a complement to the more traditional Bing search engine, so to some extent this comparison is a natural one to draw, but other studies have also noted the comparison to search engines even in interfaces without such associations. For instance, studies of language model assistance in programming through code completion tools such as GitHub Copilot also find that participants cite a reduced effort in comparison to manual web search as a benefit of these tools [65], though there are also drawbacks: due to the limited scope of sources and generation formats, language model interfaces generally offer a less media-rich experience, with fewer opportunities for learning and tangential exploration, and with fewer cues about the provenance of the results. A related observation from our participants is that search results for data analysis workflows require further work in order to adapt to the task at hand, whereas generative AI can often perform part or all of the adaptation needed. This benefit has also been observed in previous studies [65], and it is an important benefit given that many end-user data sensemaking workflows involve such search and adaptive reuse of resources on the Web (i.e., "transmogrification" [33]).

Generative AI and creativity in data sensemaking. Participants generally valued the creative potential of Bing Chat for ideation and the generation of alternative perspectives, though some participants stated a preference for first ideating and forming research questions privately (i.e., without generative AI assistance) and only using generative AI for specific data analysis tasks (Section 4.2). At least one participant was concerned about the preservation of personal voice and style when using AI-generated text. This mix of optimism and caution has been reflected in multiple other fields, such as programming, creative writing, and visual art [56], where similarly, some aspects of creativity can be usefully attributed to the AI system, and AI can be viewed as a potential source and enhancer of creativity, but there are still important roles for humans to play, as curators, as editors, as critics, and as integrators.

Generative AI and common ground. A key set of challenges faced by our participants revolved around understanding and providing the context needed by Bing Chat to address their request (Section 4.3). Participants explicitly drew a comparison to interacting with human colleagues, where interactions were simplified to due to the vastly greater degree of shared implicit context, some deriving from the shared domain of work, others from the broader shared experience of culture and language. A concept from linguistics that captures this is the notion of *common ground* [75], the set of contextual presuppositions held by interlocutors that allows any speech acts to be performed and interpreted at all, without devolving into an infinite regression of "but what does *that* mean?". Human users and generative AI do share a certain amount of common ground (deriving from the fact that generative AI behaviour is stochastic replay of real human behaviour [55]), but the quality of this common ground in our study was perceived as both alien and inferior to that shared between human collaborators. This aligns with the conclusions of Gu et al. [19], who suggest that AI assistance should be grounded in an understanding of users' current analysis plan, statistical and domain background, and overall goals; likewise, users should understand the goals of the AI assistance (e.g., to help with analysis execution, high-level planning etc.). Researchers have thus proposed to investigate how design might facilitate the notation and sharing of such contextual information without burdening the user [19, 54], but to our knowledge there are no compelling solutions, and this is one of the trickier open challenges for interaction design of generative AI.

Folk theories and external influences. When confronted with a response that did not fit their needs or expectations, participants usually proceeded by developing a hypothesis about why the model had responded in the way it had, and adapting their next prompt accordingly, including specific strategies such as using full names of entities rather than abbreviations (e.g., "Microsoft" and not "MSFT"), despite not necessarily having evidence that such hypotheses were correct, or that such strategies would be effective (Section 4.3). This echoes findings from other studies such as Liu, Sarkar et al. [38], who found that participants drew from a wide range of linguistic influences, from web search to programming languages, to inform their hypotheses about how to prompt the AI system effectively. Due to the stochastic nature of generative AI, these hypotheses and consequent prompt refinements can very well produce an improved result, affirming the participants' mental model. Over time, this may result in the development of folk theories [27] about prompting and behaviour of generative AI that may not necessarily be reliable.

Anthropomorphism of Generative AI in data sensemaking. Bing Chat is mildly anthropomorphised and frequently introduces emoji into its responses. Some participants noted this as a benefit as it improved the collegiality of the interaction, while others felt that it introduced an unwarranted expectation of politeness, verbosity, and conversationality on the part of the user (Section 4.3). This is also reflected in other studies of anthropomorphism in AI, which have found that introduction of human-like features can help users be more forgiving of a system that makes errors [24], and improves its perceived likeability, but can be counterproductive for a system with high performance and focus on task completion [11]. It is unclear from our findings whether there is a single correct approach for data sensemaking, which includes a blend of activities, some of which the system may be able to perform with high accuracy, and some not. More likely, the suitability of anthropomorphising features such as emoji appears to be dependent on the context and individual preferences.

Iteration and incremental progress. We noted that participants iterated with Bing Chat to incrementally build up an optimal response (Section 4.3), by issuing a series of prompts to slightly refine the previous response, as opposed to building up a single detailed prompt to satisfy all the requirements. This tendency to favour incremental progress has been noted in multiple previous studies of end-user interaction with AI in spreadsheets (e.g., building up a complex result through a series of intermediate columns [38], or incrementally training a machine learning model through an "edit, learn, guess" loop [66]). This preference for incremental interaction is similar to the motivation for direct manipulation interfaces and their property of being "rapid, incremental, and reversible" [72], and might be the result of the same cognitive factors that underlie the success of the direct manipulation paradigm. However, more research is needed to understand whether this is the case, and if so why, since it would appear to contradict the well-documented tendency of end-user programmers to favour the shortest path to their goal.

The burden of verification. Participants found that manually verifying sources was burdensome, and in some cases the work of verifying a response might be greater than the work required to conduct a web search manually (Section 4.3). The increased burden for users to check content has been observed in several studies (e.g., [56, 65, 76]). One approach to resolving this is "co-audit", where AI tools themselves can help to check AI-generated content [17]. What co-audit tools might look like in the context of the diverse range of data sensemaking workflows is an open research question.

Expertise and over-reliance. Recall that participants varied in spreadsheet usage (1 beginner, 7 experienced and basic usage, 7 experienced and advanced usage) and generative AI usage (3 never used, 1 casually use, 6 occasionally use, 5 regularly use), as well as programming experience (7 never programmed, 3 novices, 3 moderately experienced, 2 experts). In making data analysis more accessible to a wider range of non-experts through generative AI, over-reliance may become an unintended consequence (a review of the literature on AI over-reliance is given by Passi and Vorvoreanu [45]). We observed multiple phenomena during our study that could contribute to over-reliance, such as AI-generated output referring to concepts unfamiliar to end-users, and verification fatigue. While mitigating over-reliance was not within the scope of our study, multiple approaches have been explored such as explanations [77], cognitive forcing functions [5], and encouraging critical thinking [59, 60], to create appropriate reliance [34], which is important to consider in future work.

Metacognitive demands of generative AI. Several of the issues that participants encountered align with what has been termed the 'metacognitive demands' of generative AI [76]. These are usability issues that reflect a need for users to have a degree of self-awareness,

task decomposition, and well-adjusted confidence in their own abilities when working with generative AI systems. For example, participants struggled to formulate prompts because it was difficult to verbalise what was in their mind and break down their overall goal into sub-goals for the AI system to address—i.e., difficulties with self-awareness and task decomposition, as described and observed in other studies [2, 10, 23, 76, 84]. Moreover, some participants found it difficult to disentangle their prompting ability from the AI system performance when certain interactions went wrong, suggesting a challenge with calibrating one's self-confidence in working with the system, as also observed in prior studies [65, 84]. Participants' comments touched upon the role of self-confidence in verifying outputs, particularly for domains in which they have little expertise, as also observed in previous work [48, 81]. In some cases, this was magnified by the volume of information in generated responses.

Conversely, some participants implicitly noted how the AI system provided them with metacognitive support, as outlined in Tankelevitch et al. [76]. For example, participants commented how the system helped them think in a "step by step" manner, reflecting support with task decomposition. They also noted how the alternatives suggested by the system acted as inspiration when they were stuck, suggesting benefits to their metacognitive flexibility, analogously to that observed in Gmeiner et al. [16], which used human guides to support users co-creating with generative AI.

These observations suggest that there are opportunities to design systems which explicitly provide metacognitive support to users as they approach a task, formulate prompts, and evaluate system outputs.

5.2 Implications for design

Interaction design can support generative-AI assisted data sensemaking workflows (Section 4.2) by addressing barriers discovered in our study (Section 4.3).

For query formulation: Participants had challenges in conveying their goals and context to generative AI. These led to irrelevant, unhelpful, or partially helpful responses that required iteration to improve. This might be addressed by a design that helps a user build more detailed prompts, e.g., proactive questions that the system provides for the user to respond to (i.e., a form of metacognitive support [76]). Ambiguous or missing context could be detected and flagged before producing a response to avoid low quality responses. Output formats relevant to the user's request could be recommended as prompt addenda. For example, if the user asks how to perform a specific data analysis workflow, "step-by-step instructions" could be suggested. This could help users improve and calibrate their confidence in their prompting ability. Designers could also explore restricted vocabularies and grammars (as opposed to unrestricted natural language queries) [44], or techniques such as grounded abstraction matching [38] to help users develop clearer mental models of effective querying styles.

For response inspection: Participants also spoke about a need to verify generative AI responses for correctness, quality, and hallucinations. To do this, they inspected references provided by Bing Chat or testing code and formula suggestions. However, user expertise plays a major role in detecting incorrect output (a similar role for user expertise was observed in Gu et al. [19]). Further, verification was effortful and time-consuming. Therefore, users need verification assistance, e.g., through co-auditing features [17]. The system might share strategies with the user for identifying high quality references, or assist with specifying which types of references are suitable for supporting a particular response. To assist users in verifying AI-generated code or formulas, the system might generate and run tests to help detect failure cases. This would speed up iteration on coding tasks. In some cases, users may not be appropriately calibrated relative to their own expertise, potentially leading to over-reliance (e.g., as in Gu et al. [19]). Thus, as suggested in Tankelevitch et al. [76], there is scope for systems to prompt users to consider their own expertise and whether additional verification assistance might be helpful.

For goal formulation: Participants in our study used generative AI to help them think about their data by having Bing Chat provide potential research questions or alternative analysis strategies. However, it can go further by helping users critically think about their data-driven decisions. For example, when a user asks for AI assistance to recommend a data analysis task, the system could accompany its recommended approach with a critique of that approach outlining its potential limitations. This might prevent overreliance on the initial recommendation. The system could identify when a user's data might not be able to answer the questions they are asking, and recommend data collection strategies that would enable them to do so. To this end, Gu et al. [19] suggest that, alongside an 'analysis execution' mode, AI assistance can enter a *"'think' mode for specific planning suggestions, a 'reflection' mode for connecting decisions and highlighting potential missed steps, and an 'exploration' mode for higher-level planning suggestions"*. A step further would be to help users realise that they may not yet have a clear problem or hypothesis in mind. For example, systems can surface self-evaluation notices that encourage users to reflect on their broader aims and help them in clarifying and scoping them into concrete goals [16, 76].

For streamlining workflows: Previous research has noted the challenges of cross-application workflows, particularly when using feature-rich software termed *praxisware* [57]. Participants described a desire to integrate generative AI within the feature-rich applications they already use, rather than a separate experience which requires context switching between generative AI and application. This integration could help provide much of the context that our participants had trouble elaborating, as the application state already contains much of the context relevant to the task. It may also address issues with responses containing unfamiliar concepts, features, and programming languages.

However, some participants were wary of this type of integration and saw it as potentially limiting the recommendations that generative AI could provide. For example, a question asked within R studio would produce methods and code suited for doing data analysis in R, but there might be more effective strategies in other applications (e.g., Excel) that might not be provided. This limitation could be circumvented if application-specific AI systems were able to delegate queries to other applications when appropriate.

5.3 Implications for AI research

So far we have discussed design opportunities to improve the user experience of generative AI-assisted data analysis. This section discusses current technical developments that could positively impact the underlying issues, describe remaining gaps, and hypothesize why some issues might be addressed with foreseeable advances in technology.

In the user journey, writing the first prompt is a significant step and our study shows that there are several issues that make query formulation difficult. Several approaches have been investigated, such as improving user prompts automatically [6, 49] (including commercial solutions[1]), methods to better select prompt templates [1], prompt banks[2] and prompt documentation[3]. A less explored avenue relates to tuning prompts such that the output is not only correct, but aligned with the users' goals. There are secondary goals when users pose a question such as learning or brainstorming (as identified in our study) and more research is needed on supporting users to write prompts that produce outputs aligned with personal goals.

In our study, users also observed the importance and challenge of providing context. As Large Language Model (LLM) providers are continuously expanding model prompt windows (over 100,000 tokens in some cases), one might imagine that just by automatically ingesting more aspects of the user's work (e.g., the content of files on the user's filesystem, messages to collaborators, etc.) and passively relaying these to the model, we might be able to solve the context problem. Alas, several studies have shown that models struggle to identify the relevant portions in large prompts, and methods such as RAG (retrieval augmented generation) have been proposed [15]. The problem worsens when the context is not inherently textual; for example, when the task needs structured knowledge via (complex) tables or knowledge bases. Despite much research effort, current evaluation still shows a significant performance gap [80].

Users identified that generative AI can provide useful and diverse responses: new datasets, complex logic, general knowledge, and inspirational ideas. Unfortunately reliability is an issue and hallucinations, or even worse inconsistent hallucinations (similar or same prompts sometimes resolve successfully, other times produce incorrect outputs), are a significant problem. Researchers have explored how to improve detection [79], and counteract hallucinations by grounding in verified sources [70]. No current approach can guarantee that the results generated by an LLM are correct, and research is moving towards building tools and agents that can support users to validate outputs [17]. This work is still at an early phase, but can draw from large bodies of related research such as verification, scientific reviews, and design critiques. An interesting technical challenge is to develop an approach that lets us predict whether a generation is *likely* to be correct. Because LLMs are typically optimised for next-token generation, this might require significant architectural changes. Nonetheless, this would open the door to better feedback integration in LLM generations.

[1] e.g., https://www.junia.ai/tools/prompt-generator
[2] e.g., https://github.com/f/awesome-chatgpt-prompts
[3] e.g., https://platform.openai.com/docs/guides/prompt-engineering/prompt-engineering

5.4 Expanding the Participatory Prompting method to other fields of interest

Our research approach takes its name and inspiration from the participatory design tradition [69]. That being said, the domain of data sensemaking to which we have applied it has very specific requirements that may not generalize to all use cases of the highly-flexible technology of generative AI. We believe that the method can be extended to other domains, and here make five suggestions for fundamental aspects of the method that researchers should consider.

Level of researcher intervention: The nature of what participatory design will find depends on the interrelationship of the maturity of the technology being investigated and the level of domain expertise of the participants, but, crucially, mediated by the nature and level of activity and intervention of the researchers. Researcher mediation is a necessary part of participatory design because the approach is fundamentally about *helping* end-users find agency in a context of uncertainty around technology design. Researchers may be able to take the role of a passive conduit when the participatory design process is needed to enable access to a technology that is otherwise out of reach of end-users. However, when the technology or its application are very new or involve high levels of uncertainty, researchers may need to be an active helper for participants to articulate, enact, and reflect on their own needs. This is particularly valuable in the context of generative AI, where researcher involvement enables richer, in-the-moment collection of participant data at the level of individual prompts, rather than post-hoc recollections obtained after task completion. In Sections 3.4 and 4.1.1 we describe the nature of our participatory prompting sessions, and how we tried to stay passive – and in some cases could – but often had to be more active. The more active researchers are, the more potential there is for introducing biases, but this needs to be balanced against getting reasonable results when participant uncertainty is high, and also balanced against the spirit of enabling end-user agency that is central to participatory design. As such, it is important to plan, document, and account for how active researchers need to be and actually were, so that the results can be calibrated against others in the future.

End-user agency and ascribing agency to generative AI: Related to the first point and researcher intervention, is that in participatory prompting, end-user agency must be more than an issue of 'just' giving users a voice in the design process. The *joint agency* of people and systems in participatory prompting needs to be carefully planned for, documented, and accounted for when the technology *itself* is generative. That is, while researchers guide participants to see how generative technology opens up pathways for tasks hitherto difficult or impossible for them, researchers also need to guide participants on unpacking their agency in the process and track where participants ascribe agency to the technology (as our participants sometimes did in discussing how Bing Chat was part of the sensemaking loop in Section 4.2.2, and anthropomorphising Bing Chat in Section 4.4).

Ecological validity: Ecological validity is the extent to which a study mimics a real situation and its findings can be generalized outside the research setting [31]. While this is an issue in all research, it has a scale of relevance to participatory design depending

on the domain of interest and nature of the technology. In participatory prompting studies, beyond researcher intervention mentioned above, two key aspects affecting ecological validity are: the use of participants' own materials as resources for the generative AI system, and, relatedly, the persistence of both resources and generative AI outputs across time and across technology surfaces (as noted by participants at the beginning of Section 4.4). To get meaningful results, researchers will need to decide in advance how they will represent to participants the nature of ecological validity of the participatory prompting exercise and its use and persistence of resources.

Users in groups: Related to the third point, our study focused on one human using one generative AI system, such that the researcher was a facilitator of an individual participant's work. However, future participatory prompting studies will likely need to extend to participants acting in groups, and potentially even a hierarchy of groups (e.g. a team, the group the team belongs to, and the organisation that comprises the groups). This will entail decisions around whether participatory prompting will require exploration of each individual in a group having their own personal generative AI experience that they use in parallel to contribute to a wholly human group experience, or the group having one shared generative AI system that all can see and access serially or even some combination of both. While such group action is quite common in traditional participatory design studies, such group action maybe outside current capabilities of generative AI systems (especially group action across time and technology surfaces), necessitating some combination of real and speculative usage (or increased design fiction or Wizard-of-Oz engagement). It may also require one or more complex meta-prompts for the generative AI system so that it can (appear to) act on behalf of groups or even whole organisations. These prompts will need to be carefully designed not to misrepresent both what is feasible and what is desirable in such situations.

Domain of interest and expected outcomes: Our study focused on sensemaking from data, which will naturally only account for a proportion of the possible workflows for the flexible technology of generative AI. The method can clearly be extended to paradigms outside data sensemaking, such as artistic creativity, idea synthesis, personal reflection on goals, Socratic dialogue, educational testing and explanation, therapeutic discussion, team project planning, and more. While some of these (e.g. education) have empirically factual outcomes that users and researchers alike could agree on, others will have outcomes more related to personal satisfaction (e.g. therapeutic discussion) or shared satisfaction (e.g. the results of a creative output), potentially some combination of both factual and satisfaction outcomes, and personal and shared outcomes (e.g. the output of a team project plan). When extending the method, then, participants and researchers need to be clear about how the nature of the domain of interest is related to the nature of preferred and expected outcomes. This is especially important given the generative AI issues around non-determinative outcomes and the potential for hallucinations, as participants voiced concerns about in Section 4.3.3. Such issues will be more relevant to some domains more than others. For example, verification of sources will be crucial in some domains (e.g. information analysis), others may have no sources to be verified (e.g. creative expression), the source for some will be the participant themselves (e.g. articulating and synthesising rough ideas into a coherent draft), and the 'sources' for others will be the stochastic patterns of apparent human behaviour output by the models, to then be treated as satisfactory or not by participants (e.g. role-based prompting, such as asking a generative AI system to act as a travel agent or car mechanic when giving advice, planning etc.).

5.5 Limitations

There were limitations inherent to the Bing Chat interface which limited the kinds of behaviours we could explore. For example, some chat interfaces allow queries to be edited and re-submitted, but Bing Chat does not. If a participant wished to revise an earlier query, the best option was simply to submit the revised query as a new message, but the result might then be contaminated by the results from the previous version of the query due to the manner in which the context from the entire conversation is used in Bing Chat's responses. Nor was starting an entirely new conversation a good option, as participants often wished to continue and build on a successful conversation when revising a query. Moreover, there are features supported by other tools (e.g., ChatGPT supports plugins with varied functionality; Anthropic's Claude supports uploading and querying large documents) which we could not study. Thus, the choice of any particular tool will influence the scope of interactions which can be studied.

The set of prompting strategies was developed by trial and error, guided by the experience and subjective judgments made by a particular set of researchers. There will be differences between how different groups approach the process of developing prompting strategies, and thus this aspect of the participatory prompting process is not easily reproducible. Making this process more consistent is an important avenue for research.

As part of our protocol, each participant developed their own unique and personalised sensemaking task (Table 2). The themes emerging from a single participant engaging with a particular task may not generalise to other participants engaging with that same task. However, for our study this was an acceptable trade-off for three reasons. First, having a wider variety of tasks improves our coverage and generalisability of insights for *data-driven sensemaking* as a broad activity, which is more important than establishing generalisability for particular tasks. Second, having personal tasks developed by participants achieves ecological validity to a level that is very difficult to achieve using a synthetic suite of uniform tasks. Third, as previously mentioned, another aim of this study was to evaluate participatory prompting as a method, which more holistically and rigorously achieved using a diverse range of ecologically valid tasks.

As noted in Section 4.3, when encountering the researcher's mediation and pre-prepared prompting strategies, participants reflected on their own lack of awareness and perceived deficiencies in prompting strategies. Many participants described their own unmediated prompting strategies as "too general" and reported difficulty understanding "where to start from." To some extent this validates the utility of the participatory prompting protocol; by mediating participant requests and reformulating them according

to effective prompting strategies, the protocol bypasses many potential sources of frustration and shallow experiential dead-ends that might derail a 1-hour interactive study and compromise the ability to study meaningful tasks. On the other hand, this reduces the external validity of these experiences, since participants will not have access to expert mediation during real work. The amount of mediation is therefore a balance that needs to be carefully struck, to avoid over-influencing the participants' workflow; enough intervention to enable interesting and meaningful interaction but not so much that the interaction is completely different to the kind that the participant might have had on their own.

6 CONCLUSION

We studied how generative AI might affect the workflow of open-ended data analysis, i.e., sensemaking with data. We conducted participatory prompting sessions, in which participants worked with a researcher experienced in prompting strategies, to explore a data analysis problem of interest with the Bing Chat generative AI. Participants were asked to think aloud and reflect on the output at each turn of the conversation. The transcripts of the conversations with Bing Chat and the think-aloud data were thematically analysed.

We found that generative AI was useful in both the information foraging loop (by reducing the manual effort required to search for relevant information) and in the sensemaking loop (by helping ideate hypotheses, and proposing strategies to test them). On the other hand, participants faced barriers to query formulation (such as expressing their intent in detail, and determining what context needed to be shared with the system); in the utility of the responses (such as being overwhelmed by the amount of information, the response failing to meet their needs, or being unable to understand unfamiliar concepts in the response); and to verification and trust (such as the manual effort of looking for supporting information, and detailed checking for hallucinations).

The findings have design implications regarding balancing generative AI as a standalone application versus integration with other applications, helping users understand and provide context, managing the format and modality of responses, and metacognitive support. Besides viewing these as interaction design opportunities, we also highlight opportunities for technical research in machine learning to address some of these challenges. Further, we find that our data complements and extends our understanding of phenomena observed in previous research, such as the relationship of generative AI to search, creativity, common ground, folk theories, and metacognition. Finally, we reflect on the participatory prompting method as a research technique for eliciting opportunities and challenges for generative AI in knowledge workflows, consider its limitations, and how it might be applied to other domains.

ACKNOWLEDGMENTS

We thank our participants for their time, and our reviewers for their helpful feedback.

REFERENCES

[1] Lisa P. Argyle, E. Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, Taylor Sorensen, and David Wingate. 2022. An Information-theoretic Approach to Prompt Engineering Without Ground Truth Labels. *Political Analysis* 31 (2022), 337 – 351. https://api.semanticscholar.org/CorpusID:252280474

[2] Shraddha Barke, Michael B James, and Nadia Polikarpova. 2023. Grounded copilot: How programmers interact with code-generating models. *Proceedings of the ACM on Programming Languages* 7, OOPSLA1 (2023), 85–111.

[3] Hugh Beyer and Karen Holtzblatt. 1997. *Contextual Design: Defining Customer-Centered Systems.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[4] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[5] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[6] Tingfeng Cao, Chengyu Wang, Bingyan Liu, Ziheng Wu, Jinhui Zhu, and Jun Huang. 2023. BeautifulPrompt: Towards Automatic Prompt Engineering for Text-to-Image Synthesis. In *Conference on Empirical Methods in Natural Language Processing*. https://api.semanticscholar.org/CorpusID:265150243

[7] George Chalhoub and Advait Sarkar. 2022. "It's Freedom to Put Things Where My Mind Wants": Understanding and Improving the User Experience of Structuring Data in Spreadsheets. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 585, 24 pages. https://doi.org/10.1145/3491102.3501833

[8] Souti Chattopadhyay, Zixuan Feng, Emily Arteaga, Audrey Au, Gonzalo Ramos, Titus Barik, and Anita Sarma. 2023. Make It Make Sense! Understanding and Facilitating Sensemaking in Computational Notebooks. *arXiv preprint arXiv:2312.11431* (2023).

[9] Chen Chen, Jane Hoffswell, Shunan Guo, Ryan Rossi, Yeuk-Yin Chan, Fan Du, Eunyee Koh, and Zhicheng Liu. 2023. WHATSNEXT: Guidance-enriched Exploratory Data Analysis with Interactive, Low-Code Notebooks. In *2023 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, 209–214.

[10] Hai Dang, Sven Goller, Florian Lehmann, and Daniel Buschek. 2023. Choice over control: How users write with large language models using diegetic and non-diegetic prompting. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[11] Ewart J De Visser, Samuel S Monfort, Ryan McKendrick, Melissa AB Smith, Patrick E McKnight, Frank Krueger, and Raja Parasuraman. 2016. Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied* 22, 3 (2016), 331.

[12] Stephen L Dorton and Robert A Hall. 2021. Collaborative human-AI sensemaking for intelligence analysis. In *International conference on human-computer interaction*. Springer, 185–201.

[13] Gregor Engels and Martin Erwig. 2005. ClassSheets: automatic generation of spreadsheet applications from object-oriented specifications. In *Proceedings of the 20th IEEE/ACM international Conference on Automated software engineering*. 124–133.

[14] Kasra Ferdowsi, Jack Williams, Ian Drosos, Andrew D. Gordon, Carina Negreanu, Nadia Polikarpova, Advait Sarkar, and Benjamin Zorn. 2023. COLDECO: An End User Spreadsheet Inspection Tool for AI-Generated Code. In *2023 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. 82–91. https://doi.org/10.1109/VL-HCC57772.2023.00017

[15] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-Augmented Generation for Large Language Models: A Survey. *ArXiv* abs/2312.10997 (2023). https://api.semanticscholar.org/CorpusID:266359151

[16] Frederic Gmeiner, Humphrey Yang, Lining Yao, Kenneth Holstein, and Nikolas Martelaro. 2023. Exploring Challenges and Opportunities to Support Designers in Learning to Co-create with AI-based Manufacturing Design Tools. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.

[17] Andrew D Gordon, Carina Negreanu, José Cambronero, Rasika Chakravarthy, Ian Drosos, Hao Fang, Bhaskar Mitra, Hannah Richardson, Advait Sarkar, Stephanie Simmons, et al. 2023. Co-audit: tools to help humans double-check AI-generated content. *arXiv preprint arXiv:2310.01297* (2023).

[18] Valentina Grigoreanu, Margaret Burnett, Susan Wiedenbeck, Jill Cao, Kyle Rector, and Irwin Kwan. 2012. End-user debugging strategies: A sensemaking perspective. *ACM Transactions on Computer-Human Interaction (TOCHI)* 19, 1 (2012), 1–28.

[19] Ken Gu, Madeleine Grunde-McLaughlin, Andrew M McNutt, Jeffrey Heer, and Tim Althoff. 2023. How Do Data Analysts Respond to AI Assistance? A Wizard-of-Oz Study. *arXiv preprint arXiv:2309.10108* (2023).

[20] Ken Gu, Ruoxi Shang, Tim Althoff, Chenglong Wang, and Steven M. Drucker. 2023. How Do Analysts Understand and Verify AI-Assisted Data Analyses? arXiv:2309.10947 [cs.HC]

[21] Sumit Gulwani. 2011. Automating string processing in spreadsheets using input-output examples. *ACM Sigplan Notices* 46, 1 (2011), 317–330.

[22] Amber Horvath, Brad Myers, Andrew Macvean, and Imtiaz Rahman. 2022. Using Annotations for Sensemaking About Code. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–16.

[23] Dhanya Jayagopal, Justin Lubin, and Sarah E Chasins. 2022. Exploring the learnability of program synthesizers by novice programmers. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–15.

[24] Theodore Jensen and Mohammad Maifi Hasan Khan. 2022. I'm Only Human: The Effects of Trust Dampening by Anthropomorphic Agents. In *International Conference on Human-Computer Interaction*. Springer, 285–306.

[25] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 12, Article 248 (mar 2023), 38 pages. https://doi.org/10.1145/3571730

[26] Nima Joharizadeh, Advait Sarkar, Andrew D. Gordon, and Jack Williams. 2020. Gridlets: Reusing Spreadsheet Grids. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3334480.3382806

[27] Philip N Johnson-Laird and Keith Oatley. 1998. Basic emotions, rationality, and folk theory. In *Consciousness and Emotion in Cognitive Science*. Routledge, 289–311.

[28] Simon Peyton Jones, Alan Blackwell, and Margaret Burnett. 2003. A user-centred approach to functions in Excel. In *Proceedings of the eighth ACM SIGPLAN international conference on Functional programming*. 165–176.

[29] Karl E. Weick. 1969. *The Social Psychology of Organizing*. Addison Wesley, Reading, MA.

[30] Karl E. Weick. 1995. *Sensemaking in Organizations*. SAGE Publications, Thousand Oaks, CA.

[31] Suzanne Kieffer. 2017. ECOVAL: Ecological Validity of Cues and Representative Design in User Experience Evaluations. *AIS Transactions on Human-Computer Interaction* 9, 2 (June 2017), 149–172. https://aisel.aisnet.org/thci/vol9/iss2/4

[32] Amy J Ko, Robin Abraham, Laura Beckwith, Alan Blackwell, Margaret Burnett, Martin Erwig, Chris Scaffidi, Joseph Lawrance, Henry Lieberman, Brad Myers, et al. 2011. The state of the art in end-user software engineering. *ACM Computing Surveys (CSUR)* 43, 3 (2011), 1–44.

[33] Sam Lau, Sruti Srinivasa Srinivasa Ragavan, Ken Milne, Titus Barik, and Advait Sarkar. 2021. TweakIt: Supporting End-User Programmers Who Transmogrify Code. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 311, 12 pages. https://doi.org/10.1145/3411764.3445265

[34] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[35] Sukwon Lee, Sung-Hee Kim, Ya-Hsin Hung, Heidi Lam, Youn-ah Kang, and Ji Soo Yi. 2015. How do people make sense of unfamiliar visualizations?: A grounded model of novice's information visualization sensemaking. *IEEE transactions on visualization and computer graphics* 22, 1 (2015), 499–508.

[36] Clayton Lewis and Cathleen Wharton. 1997. Cognitive walkthroughs. In *Handbook of human-computer interaction*. Elsevier, 717–732.

[37] Xingjun Li, Yizhi Zhang, Justin Leung, Chengnian Sun, and Jian Zhao. 2023. EDAssistant: Supporting Exploratory Data Analysis in Computational Notebooks with In Situ Code Search and Recommendation. *ACM Trans. Interact. Intell. Syst.* 13, 1, Article 1 (mar 2023), 27 pages. https://doi.org/10.1145/3545995

[38] Michael Xieyang Liu, Advait Sarkar, Carina Negreanu, Benjamin Zorn, Jack Williams, Neil Toronto, and Andrew D. Gordon. 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* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 598, 31 pages. https://doi.org/10.1145/3544548.3580817

[39] Deborah Lupton. 2016. *The quantified self*. John Wiley & Sons.

[40] Mariana Mărăşoiu, Alan F Blackwell, Advait Sarkar, and Martin Spott. 2016. Clarifying hypotheses by sketching data. In *Proceedings of the Eurographics/IEEE VGTC Conference on Visualization: Short Papers*. 125–129.

[41] Matt Mccutchen, Judith Borghouts, Andrew D Gordon, Simon Peyton Jones, and Advait Sarkar. 2020. Elastic sheet-defined functions: Generalising spreadsheet functions to variable-size input arrays. *Journal of Functional Programming* 30 (2020), e26.

[42] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–23.

[43] Andrew M McNutt, Chenglong Wang, Robert A Deline, and Steven M Drucker. 2023. On the design of ai-powered code assistants for notebooks. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[44] Jesse Mu and Advait Sarkar. 2019. Do We Need Natural Language? Exploring Restricted Language Interfaces for Complex Domains. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312975

[45] Samir Passi and Mihaela Vorvoreanu. 2022. Overreliance on AI: literature review. *Microsoft Research* (2022).

[46] Peter Pirolli and Stuart Card. 1999. Information foraging. *Psychological review* 106, 4 (1999), 643.

[47] Peter Pirolli and Stuart Card. 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. McLean, VA, USA, 2–4.

[48] James Prather, Brent N Reeves, Paul Denny, Brett A Becker, Juho Leinonen, Andrew Luxton-Reilly, Garrett Powell, James Finnie-Ansley, and Eddie Antonio Santos. 2023. " It's Weird That it Knows What I Want": Usability and Interactions with Copilot for Novice Programmers. *arXiv preprint arXiv:2304.02491* (2023).

[49] Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic Prompt Optimization with "Gradient Descent" and Beam Search. In *Conference on Empirical Methods in Natural Language Processing*. https://api.semanticscholar.org/CorpusID:258546785

[50] Gregg Rothermel, Lixin Li, Christopher DuPuis, and Margaret Burnett. 1998. What you see is what you test: A methodology for testing form-based visual programs. In *Proceedings of the 20th international conference on Software engineering*. IEEE, 198–207.

[51] Daniel M Russell, Mark J Stefik, Peter Pirolli, and Stuart K Card. 1993. The cost structure of sensemaking. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*. 269–276.

[52] Advait Sarkar. 2016. Constructivist Design for Interactive Machine Learning. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI EA '16)*. Association for Computing Machinery, New York, NY, USA, 1467–1475. https://doi.org/10.1145/2851581.2892547

[53] Advait Sarkar. 2016. *Interactive analytical modelling*. Technical Report UCAM-CL-TR-920. University of Cambridge, Computer Laboratory. https://doi.org/10.48456/tr-920

[54] Advait Sarkar. 2022. 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) (CEUR Workshop Proceedings, 3124)*. 192–199. http://ceur-ws.org/Vol-3124/paper22.pdf

[55] Advait Sarkar. 2023. Enough With "Human-AI Collaboration". In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 415, 8 pages. https://doi.org/10.1145/3544549.3582735

[56] Advait Sarkar. 2023. Exploring Perspectives on the Impact of Artificial Intelligence on the Creativity of Knowledge Work: Beyond Mechanised Plagiarism and Stochastic Parrots. In *Proceedings of the 2nd Annual Meeting of the Symposium on Human-Computer Interaction for Work* (Oldenburg, Germany) *(CHIWORK '23)*. Association for Computing Machinery, New York, NY, USA, Article 13, 17 pages. https://doi.org/10.1145/3596671.3597650

[57] Advait Sarkar. 2023. Should Computers Be Easy To Use? Questioning the Doctrine of Simplicity in User Interface Design. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 419, 10 pages. https://doi.org/10.1145/3544549.3582741

[58] Advait Sarkar. 2023. Will Code Remain a Relevant User Interface for End-User Programming with Generative AI Models?. In *Proceedings of the 2023 ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software* (Cascais, Portugal) *(Onward! 2023)*. Association for Computing Machinery, New York, NY, USA, 153–167. https://doi.org/10.1145/3622758.3622882

[59] Advait Sarkar. 2024. AI Should Challenge, Not Obey. *Communications of the ACM (in press)* (2024).

[60] Advait Sarkar. 2024. Large Language Models Cannot Explain Themselves. In *ACM CHI 2024 Workshop on Human-Centered Explainable AI (HCXAI)*.

[61] Advait Sarkar, Alan F Blackwell, Mateia Jamnik, and Martin Spott. 2014. Teach and try: A simple interaction technique for exploratory data modelling by end users. In *2014 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. 53–56. https://doi.org/10.1109/VLHCC.2014.6883022

[62] Advait Sarkar, Judith W. Borghouts, Anusha Iyer, Sneha Khullar, Christian Canton, Felienne Hermans, Andrew D. Gordon, and Jack Williams. 2020. Spreadsheet Use and Programming Experience: An Exploratory Survey. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3334480.3382807

[63] Advait Sarkar, Ian Drosos, Rob Deline, Andrew D. Gordon, Carina Negreanu, Sean Rintel, Jack Williams, and Ben Zorn. 2023. Participatory prompting: a user-centric research method for eliciting AI assistance opportunities in knowledge workflows. In *Proceedings of the 34th Annual Conference of the Psychology of Programming Interest Group (PPIG 2023)*.

[64] Advait Sarkar and Andrew D. Gordon. 2018. How do people learn to use spreadsheets? (Work in progress). In *Proceedings of the 29th Annual Conference of the Psychology of Programming Interest Group (PPIG 2018)*. 28–35.

[65] Advait Sarkar, Andrew D. Gordon, Carina Negreanu, Christian Poelitz, Sruti Srinivasa Ragavan, and Ben Zorn. 2022. 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)*.

[66] Advait Sarkar, Mateja Jamnik, Alan F. Blackwell, and Martin Spott. 2015. Interactive visual machine learning in spreadsheets. In *2015 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. 159–163. https://doi.org/10.1109/VLHCC.2015.7357211

[67] Advait Sarkar, Sruti Srinivasa Ragavan, Jack Williams, and Andrew D. Gordon. 2022. End-user encounters with lambda abstraction in spreadsheets: Apollo's bow or Achilles' heel?. In *2022 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. 1–11. https://doi.org/10.1109/VL/HCC53370.2022.9833131

[68] Advait Sarkar, Martin Spott, Alan F. Blackwell, and Mateja Jamnik. 2016. Visual discovery and model-driven explanation of time series patterns. In *2016 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. 78–86. https://doi.org/10.1109/VLHCC.2016.7739668

[69] Douglas Schuler and Aki Namioka. 1993. *Participatory design: Principles and practices*. CRC Press.

[70] Sina J. Semnani, Violet Z. Yao, He Zhang, and Monica S. Lam. 2023. WikiChat: Stopping the Hallucination of Large Language Model Chatbots by Few-Shot Grounding on Wikipedia. In *Conference on Empirical Methods in Natural Language Processing*. https://api.semanticscholar.org/CorpusID:258841157

[71] Floarea Serban, Joaquin Vanschoren, Jörg-Uwe Kietz, and Abraham Bernstein. 2013. A survey of intelligent assistants for data analysis. *ACM Computing Surveys (CSUR)* 45, 3 (2013), 1–35.

[72] Ben Shneiderman. 1983. Direct manipulation: A step beyond programming languages. *Computer* 16, 08 (1983), 57–69.

[73] Divya Siddarth, Daron Acemoglu, Danielle Allen, Kate Crawford, James Evans, Michael Jordan, and E Weyl. 2021. How AI fails us. *arXiv preprint arXiv:2201.04200* (2021).

[74] Sruti Srinivasa Ragavan, Advait Sarkar, and Andrew D Gordon. 2021. Spreadsheet Comprehension: Guesswork, Giving Up and Going Back to the Author. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 181, 21 pages. https://doi.org/10.1145/3411764.3445634

[75] Robert Stalnaker. 2002. Common ground. *Linguistics and philosophy* 25, 5/6 (2002), 701–721.

[76] Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. 2023. The Metacognitive Demands and Opportunities of Generative AI. *arXiv preprint arXiv:2312.10893* (2023).

[77] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S Bernstein, and Ranjay Krishna. 2023. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–38.

[78] Dakuo Wang, Josh Andres, Justin D. Weisz, Erick Oduor, and Casey Dugan. 2021. AutoDS: Towards Human-Centered Automation of Data Science. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 79, 12 pages. https://doi.org/10.1145/3411764.3445526

[79] Weixuan Wang, Barry Haddow, Alexandra Birch, and Wei Peng. 2023. Assessing the Reliability of Large Language Model Knowledge. *ArXiv* abs/2310.09820 (2023). https://api.semanticscholar.org/CorpusID:264146357

[80] Zilong Wang, Hao Zhang, Chun-Liang Li, Julian Martin Eisenschlos, Vincent Perot, Zifeng Wang, Lesly Miculicich, Yasuhisa Fujii, Jingbo Shang, Chen-Yu Lee, and Tomas Pfister. 2024. Chain-of-Table: Evolving Tables in the Reasoning Chain for Table Understanding. *ArXiv* abs/2401.04398 (2024). https://api.semanticscholar.org/CorpusID:266899992

[81] Justin D Weisz, Michael Muller, Stephanie Houde, John Richards, Steven I Ross, Fernando Martinez, Mayank Agarwal, and Kartik Talamadupula. 2021. Perfection not required? Human-AI partnerships in code translation. In *26th International Conference on Intelligent User Interfaces*. 402–412.

[82] John Wenskovitch, Corey Fallon, Kate Miller, and Aritra Dasgupta. 2021. Beyond visual analytics: Human-machine teaming for ai-driven data sensemaking. In *2021 IEEE Workshop on TRust and EXpertise in Visual Analytics (TREX)*. IEEE, 40–44.

[83] Jack Williams, Carina Negreanu, Andrew D. Gordon, and Advait Sarkar. 2020. Understanding and Inferring Units in Spreadsheets. In *2020 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. 1–9. https://doi.org/10.1109/VL/HCC50065.2020.9127254

[84] JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 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*. 1–21.