

Conversational Challenges in AI-Powered Data Science: Obstacles, Needs, and Design Opportunities

BHAVYA CHOPRA, Microsoft, India

ANANYA SINGHA, Microsoft, India

ANNA FARIHA, University of Utah, United States

SUMIT GULWANI, Microsoft, United States

CHRIS PARNIN, Microsoft, United States

ASHISH TIWARI, Microsoft, United States

AUSTIN Z. HENLEY, Microsoft, United States

Large Language Models (LLMs) are being increasingly employed in data science for tasks like data preprocessing and analytics. However, data scientists encounter substantial obstacles when conversing with LLM-powered chatbots and acting on their suggestions and answers. We conducted a mixed-methods study, including contextual observations, semi-structured interviews ($n=14$), and a survey ($n=114$), to identify these challenges. Our findings highlight key issues faced by data scientists, including contextual data retrieval, formulating prompts for complex tasks, adapting generated code to local environments, and refining prompts iteratively. Based on these insights, we propose actionable design recommendations, such as data brushing to support context selection, and inquisitive feedback loops to improve communications with AI-based assistants in data-science tools.

CCS Concepts: • Human-centered computing → HCI theory, concepts and models; Empirical studies in HCI; • Software and its engineering → Requirements analysis.

Additional Key Words and Phrases: Data Science, Generative Models, Large Language Models, ChatGPT, Computational Notebooks

1 INTRODUCTION

Data scientists have traditionally relied on resources like documentation, tutorials, online courses, colleagues, Q&A forums, and online communities to develop skills and solve tasks [25, 26, 38, 44]. These resources have been instrumental in helping them navigate the challenges of data acquisition, cleaning, wrangling, visualization, and presentation, especially for those with a non-programming background [23]. Although useful, it can be time consuming, tedious, and error-prone to rely on these resources for solving a data-science problem.

With the emergence of AI-powered chat assistants, data scientists now have access to potentially faster and more accessible resources through a chat interface. These AI-powered chatbots, like ChatGPT¹, enable users to ask questions

¹<https://chat.openai.com/>

Authors' addresses: Bhavya Chopra, Microsoft, India, t-bhchopra@microsoft.com; Ananya Singha, Microsoft, India, t-asintha@microsoft.com; Anna Fariha, University of Utah, United States, afariha@cs.utah.edu; Sumit Gulwani, Microsoft, United States, sumitg@microsoft.com; Chris Parnin, Microsoft, United States, chrisparnin@microsoft.com; Ashish Tiwari, Microsoft, United States, astiwar@microsoft.com; Austin Z. Henley, Microsoft, United States, austinhenley@microsoft.com.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2023 Association for Computing Machinery.

Manuscript submitted to ACM

in natural language, and get useful responses with little to no latency. For example, when prompted with “split my date column that is in MM/DD/YYYY format into three columns”, ChatGPT provided usable Python code:

```
df[['Month', 'Day', 'Year']] = df['Date'].str.split('/', expand=True)
```

ChatGPT can also explain how the code works, and supports follow-up questions or changes. Evaluations of these AI tools have demonstrated over 70% accuracy on programming benchmarks [7] and pass numerous tests designed for humans (e.g., the GRE, a college entrance exam, and the LSAT, a law school admission test) [34].

However, the effectiveness of these tools is dependent on data scientists successfully communicating their questions, context (overall problem, task at hand, datasets, etc.), assumptions, and domain knowledge to the AI assistant, and doing so through a back-and-forth conversation. Grice’s maxims of conversation² posits that a successful conversation involves content that has right amount of information, that is truthful and supported by evidence, that is relevant to the specific context, and that is presented clearly [14]. But conversations go awry—either side of the conversation may be making false assumptions, there may be ambiguities, and conversations may require numerous clarifications. All of these problems are possible with AI assistants; most notably, they are known to *hallucinate*, meaning they confidently respond with false statements. Furthermore, studies have found that users find conversations with AI assistants, such as ChatGPT, to have problematic communication styles [39] and responses provided to be repetitive and filled with outdated or irrelevant information [20, 39]—a far cry from Grice’s maxims. As such, for humans to communicate effectively with AI agents, humans must follow an iterative sense-making process that involves describing intent, building hypotheses, gathering evidence, and evaluating responses [5].

The barriers to express intents are exacerbated in the context of data-science tasks. First, data scientists work with a variety of artifacts, including raw datasets, code, computational notebooks, visualizations, documentation, and machine-learning pipelines. Second, datasets are often large and normalized (spilt across multiple tables), which may not be feasible to share (e.g., due to token limits of AI tools or due to data being spread across heterogeneous sources) or to summarize the relevant portions succinctly (e.g., specifying regions of the data). Third, real-world data is messy, and suffer from quality issues; they may not strictly adhere to a schema or homogeneous format. Fourth, data-science tasks often require domain expertise and numerous assumptions (e.g., negative values in a column represent an error, 0 means missing value) and it may be laborious to ensure that the AI assistant is aware of such domain knowledge and assumptions. Furthermore, even when a data scientist is able to articulate their question to the AI assistant, these same problems can cause difficulties to understand and validate the AI assistant’s response.

In this work, we aim to understand the fundamental problems of a [human] data scientist communicating to an AI chat agent. In particular, we address the following research questions:

- *RQ1*: How do data scientists interact with ChatGPT to complete data-science tasks?
- *RQ2*: What challenges and unmet needs do data scientists face when interacting with ChatGPT?
- *RQ3*: How well do these challenges and needs generalize to broader community of data scientists?

To answer these questions, we conducted two mixed-method need-finding studies (Section 3). For the first study, we observe 14 professional data scientists as they perform four diverse tasks that are common in data-science pipelines. The participants were presented with the browser-based chat assistant, ChatGPT, to help them solve the tasks whenever needed. We performed inductive thematic analysis to identify needs and challenges by triangulating findings from open coding of semi-structured interview transcripts, observations from field notes, and video annotations. For the second

²<https://effectiviology.com/principles-of-effective-communication/>

study, we conducted a confirmation survey with 114 professional data scientists to validate and generalize findings from the first study.

Our results (Sections 4 and 5) show that data scientists usually follow a typical pattern of articulating their question, collecting context, validating the response made by the AI assistant, and then applying the response to their task. Throughout this process, they face copious barriers while communicating with the AI assistant, such as spending considerable time writing their question and collecting relevant context. Then they face a variety of additional barriers while trying to make use of the AI assistant’s response, such as understanding the assumptions being made by the AI and modifying the provided code to work for their scenario. This arduous workflow resulted in participants applying strategies to overcome the barriers and still make use of the AI assistant. Based on our findings, we provide practical design recommendations that can improve user experiences while using LLM-powered assistants for their data-science workflows (Section 6).

2 BACKGROUND

How data scientists work. Data science is the broad discipline of using data to understand the underlying nature of a domain and provide insights [31]. Activities include data acquisition, cleaning, wrangling, visualization, sharing, and acting on insights [11, 25, 31]. The majority of time spent on data-science tasks often involves data cleaning and wrangling [15, 21]. Sutton et al. stated that handling expansive datasets often involves navigating a myriad of issues, culminating in a situation they term “death by a thousand wranglings” [41].

To support exploratory wrangling tasks, data scientists often employ *computational notebooks*, such as Jupyter notebooks³. These notebooks provide an interactive environment that allows for the seamless integration of text, code, and output, facilitating both the development and documentation of computational workflows in a single document. Unlike traditional code editors, notebooks enable a more narrative style of coding, where code cells and rich-text commentary can coexist and outputs can be immediately rendered inline. However, notebooks come with their own pain points: they take considerable effort to keep organized, hidden state information can lead to misconceptions, they can be difficult to deploy, and long-running tasks can be problematic [6, 24, 35].

Large language models and AI-powered chat assistants. AI-powered chat assistants are built on top of large language models (LLMs). LLMs are generative machine-learning models with billion parameters, and are trained on vast amount of data (text and images) to generate human-like responses and perform various language tasks. These models interact with users through a *prompt*, a natural-language query to which the LLM can respond to, also in natural language. There are restrictions on the *token limit*, or the amount of information that can be sent or generated from the LLM. Tools like ChatGPT add a user interface to enable a conversation, such that the model has *context* of the conversation history. However, these models do not have direct access to data sources, and only respond based on the data they have seen during training.

3 METHODOLOGY

We performed two studies to answer our research questions. First, we conducted task-based semi-structured interviews and observed 14 professional data scientists engaged in data-science tasks. Then, we distributed a general survey of our initial findings with 114 data scientists.

³<https://jupyter.org/>

3.1 Study 1: Using ChatGPT for Data Science Tasks

We used ChatGPT, a browser-based LLM-powered chat-based AI assistant, for the study. We avoided using any AI assistants integrated within computational notebooks to better understand fundamental challenges and patterns of usage. This deliberate choice to keep the AI assistant separate from the data science environment(s) of the participants allows us to give them complete control and transparency over the prompt-writing and code adaptation processes—while giving us an opportunity to understand the underlying process of assessing *which* context, and *how much* context is relevant for prompting. This choice also echoes one of the design recommendations proposed by Barke et al. [1] that developers would prefer control over the shared prompt context.

Participants. We recruited participants from a large technology company by randomly selecting employees with the job title “Data Scientist” mentioned in the company address book. We further screened participants based on having at least 1 year of experience with data science and prior experience with Python, computational notebooks, and ChatGPT (or any other LLM-based chat/code assistants) as a pre-requisite to their participation. Table 1 indicates the demographics collected for all participants, who reported a median of 5 years of experience in the domain of data science ($\mu = 5.4$, $sd = 2.8$). All participants have a background and educational degree in CS and Data Science related fields.

All sessions were audio and video recorded for transcription of responses and analysis of on-screen interactions. All participants signed a consent form prior to the study. Participation in the study was voluntary and participants were compensated with gift cards worth \$25 USD each.

Table 1. Overview of the participants, including the domain that they work in, their years of experience as a data scientist, and the programming languages they often use.

ID	Domain	Exp. (yrs)	Languages used
P1	Anomaly Detection	5	Python, SQL
P2	Big Data Processing	10	Python, SQL
P3	User Segmentation	4	Python
P4	NL Processing	1	Python
P5	Feature Engineering	4	SQL
P6	Statistical Analysis	2	Python
P7	Defect Detection	5	Python
P8	Finance & Forecasting	6	Python, SQL, Spark
P9	Business Analytics	2	Python
P10	Fraud Detection	5	Python, SQL
P11	User Segmentation	10	Python, SQL, Spark
P12	Cloud Computing	10	Python
P13	Failure Prediction	7	Python
P14	Fraud Detection	5	Python, SQL

3.1.1 Tasks. We used four essential tasks performed by data scientists in the lifetime of a typical project [31]: (T1 and T2) data wrangling and pre-processing, (T3) feature extraction and selection for training tasks, (T4) data visualization for insight-finding and reporting. The four tasks are described in Table 2.

We selected a sample of the *New York EMS emergency calls dataset*⁴, which contains rows corresponding to emergency incidents, and information about time, location, resources, and the Fire Department’s response to the emergency. The

⁴<https://www.kaggle.com/datasets/new-york-city/ny-ems-incident-dispatch-data>

Table 2. The four data science tasks used in our study and their descriptions.

#	Task	Description
T1	Datetime typecast	Change the datatype of column “INCIDENT_DATETIME” from string to datetime (e.g., “2020-07-31T23:59:51.000”).
T2	Split using delimiter	Split the (“INCIDENT_DESCRIPTION”) column using semicolon (“;”) as the delimiter. Rows contain information about the incident location – borough, incident dispatch area code, and zipcode. Few rows do not contain the zipcode (e.g. rows such as “Brooklyn; K7; 11211” contain all information, whereas rows like “Manhattan; M3” are missing the zipcode).
T3	Feature selection	Extract a new column titled “RESPONSE_TIME” using the difference between “FIRST_ON_SCENE_DATETIME” and “INCIDENT_DATETIME”; subsequently perform feature selection with the aim of predicting the “RESPONSE_TIME”.
T4	Heatmap plotting	Obtain a heatmap plot to analyze correlation between any columns of choice.

dataset was well-suited for supporting the tasks, as several columns had to be transformed before performing further analysis, and event data was amendable to feature and correlation analysis.

Study protocol. We conducted a one-hour long, task-based study with each of the 14 data scientists via video conferencing. The study begins with demographic questions about their role and data-science-related work, languages and tools, educational background, and prior experience with using LLMs. Next, participants were presented with 4 tasks, one at a time, to be solved using Python in any computational notebook of their choice, and the browser-based ChatGPT⁵. They were encouraged to make use of any preferred tools and resources (such as web search, notebooks, Excel, and so on), and think aloud while solving the tasks. The completion of any task was not a necessity, and participants were encouraged to use ChatGPT at least once for every task. Once the participants had explored a task sufficiently, they were asked semi-structured questions to assess if ChatGPT helped them solve the task satisfactorily, and if it could have helped them with the task any better.

Analysis. We transcribed the audio recordings for each session. Subsequently, we annotated videos from each session recording to timestamp and gathered: (1) prompts authored by participants, (2) corresponding responses provided by ChatGPT, and (3) a log of activities performed by the participants (including prompt writing, adapting generated code, and validating generated code). The first and second author analyzed the transcripts and annotations from video recordings to perform inductive thematic coding [36]. We first conducted discussions over a span of 2 weeks while collaboratively annotating data for one participant. Following this, we independently annotated 20% of the data (from three participants) with descriptive open codes. Next, we grouped similar codes and analyzed them to obtain axial codes, identifying usage patterns and pain-points faced by data scientists. We then computed the inter-rater agreement score to be 0.872, indicating ‘almost perfect’ agreement [27]. The first two authors then split the remaining participant interviews and annotated them independently, while meeting frequently to discuss and refine the codes and themes.

We observed theoretical saturation after thematic analysis of data from 9 participants. To further support the validity of our findings, we requested 5 study participants (P1, P3, P4, P8, P12) to respond to a member-check survey. The survey consisted of 18 Likert-scale questions, each question representing an open-code from our findings. The ratings indicated agreement with each finding.

⁵<https://chat.openai.com/> (Versions: March 14, 2023; March 23, 2023)

3.2 Study 2: Confirmatory Survey

To validate findings from Study 1, we conducted a survey with a broader population of 114 data scientists at a large technology company.

Survey protocol. The survey consisted of two demographic questions, asking respondents to: (D1) indicate years of experience with data science (a single choice question), and (D2) describe at least one scenario where they used any LLM for a data-science task (an open-field question). To ensure the quality of the survey responses, we used responses for D2 to filter out respondents who do not have any experience with solving data-science tasks using LLMs.

We used findings from the qualitative analysis of Study 1 responses to frame at least one question for each identified obstacle. The survey consisted of 8 ‘Agree’/‘Disagree’ statements, and all questions were optional to answer. Finally, to evaluate if our observations from Study 1 reached theoretical saturation, we asked respondents to indicate if they have faced any other difficulties in using LLMs for data science tasks that were not mentioned in the survey questions.

Recruitment and informed consent. The respondents were contacted randomly via e-mail invites based on their job title (“Data Scientist”) mentioned in the company address book. In our e-mail invites, we also requested participants to forward the survey to other data scientists who might be interested. We offered a drawing for one of two Amazon gift cards worth \$50 USD each for completing the survey.

Respondents. 114 data scientists from a large technology company responded to the survey. After inspecting whether the respondents have meaningful prior experiences (D2), we removed 16 respondents (14%) that had no experience with using LLMs for data science tasks. We report all findings using the screened responses. 76% of the respondents self-reported having more than 5 years of experience in data science related professional roles. We discuss survey results in Section 5, where we plotted the responses and inspected them manually to gather any other difficulties that were not mentioned in the survey questions.

4 OBSERVATIONS

We provide a summary of our participants interactions with ChatGPT and then report observed obstacles and strategies used by the participants. Table 4 summarizes the themes that emerged from our analysis of the data scientists’ actions and responses. We also report participants who identified with each obstacle and employed prompting strategies in Table 4.

4.1 Participants Interactions with ChatGPT

4.1.1 Typical workflow. Most participants started with loading the CSV files into a notebook of their choice, and constructing a pandas dataframe, while viewing the first few rows of the data. Sometimes they opened the CSV in Excel before creating a notebook, to glance over the data. Some participants used `pandas.DataFrame.describe()` to obtain a high-level summary of their data.

Next, they were sequentially presented with each of the tasks. Every participant worked on tasks T1, T2, and T3. However, 4 participants were unable to attempt T4 in the available time. For each task, participants generally began with phrasing their question, assessing and gathering context—such as description of the data, sample values—for prompt construction, understanding and taking action on ChatGPT’s response while writing additional prompts that would help them elicit better responses. Figure 1 presents a gallery of tasks involved in the study.

Task 1: Datetime typecast

C df

P4 Write a function in Python to convert a pandas column which is formatted as a string (e.g. "2020-07-31T23:59:51.000") to datetime format (day/month/year, hour, minute, second).

A Sure, here's an example Python function that you could use to convert a pandas column of strings to datetime format:

```
python
import pandas as pd

def convert_to_datetime(df, column_name):
    df[column_name] = pd.to_datetime(df[column_name])
    return df
```

You can use this function by passing in the pandas DataFrame and the name of the column containing the strings that you want to convert. For example, if your DataFrame is called 'my_df':

Task 2: Split using delimiter

C df['INCIDENT_DESC'].head() #region, code, zip

P6 format this into 3 columns in a data frame

A To format the data into a DataFrame with three columns, you can use the "pandas" library in Python. Here's an example code snippet that demonstrates how to do this:

```
python
import pandas as pd

# Define the data
data = [{"location": "BROOKLYN; K7; 112211"}, {"location": "QUEENS; Q3; 11419"}, {"location": "MANHATTAN; M4; 10025"}, {"location": "BROOKLYN; K7; 112211"}, {"location": "QUEENS; Q2; 11434"}]
```

Task 3: Feature selection

C i in df.columns if 'TIME' not in i.upper() and 'ID' not in i.upper()

P Intent to build a GBM using sklearn. The X variables are [INITIAL_CALL_TYPE, INITIAL_SEVERITY_LEVEL_CODE, FINAL_CALL_TYPE, FINAL_SEVERITY_LEVEL_CODE, HELD_INDICATOR, POLICEPRECINCT, CITYCOUNCILDISTRICT, COMMUNITYDISTRICT, COMMUNICHOOLDISTRICT, CONGRESSIONALDISTRICT, REOPEN_INDICATOR, SPECIAL_EVENT_INDICATOR, STANDBY_INDICATOR, TRANSFER_INDICATOR], target variable is RESPONSE_TIME which is timedelta.

A Sure, you can use the "GradientBoostingRegressor" class from sklearn to build a GBM model for predicting the "RESPONSE_TIME" target variable. Here is an example code:

```
python
import pandas as pd
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_absolute_error

# Load the dataset
df = pd.read_csv('data.csv')

# Convert the timedelta values to seconds
df['RESPONSE_TIME'] = df['RESPONSE_TIME'].dt.total_seconds()

# Select X and Y variables
X = df[INITIAL_CALL_TYPE, INITIAL_SEVERITY_LEVEL_CODE, FINAL_CALL_TYPE, FINAL_SEVERITY_LEVEL_CODE, HELD_INDICATOR, POLICEPRECINCT, CITYCOUNCILDISTRICT, COMMUNITYDISTRICT, COMMUNICHOOLDISTRICT]
```

Task 4: Heatmap plotting

P Can you show a correlation plot with seaborn?

A Sure, here's an example of how to create a correlation plot with Seaborn:

```
python
import pandas as pd
import seaborn as sns

# Load a sample dataset
data = pd.read_csv('data.csv')

# Create a correlation matrix
corr_matrix = data.corr()

# Create a Seaborn heatmap of the correlation matrix
sns.heatmap(corr_matrix, cmap='coolwarm', annot=True)

# Show the plot
plt.show()
```

Fig. 1. A gallery of tasks involved in Study 1. Images are taken from participant screen sharing recordings and are anonymized. Snippets marked (C) denote the local **Context** retrieved by participants to use in prompt-writing; (P) denotes the **Prompt** sent to ChatGPT; and (A) denotes the **Answer** provided by ChatGPT.

Participants faced several challenges in conversing with ChatGPT and using its responses for task completion. In particular, most data scientists struggled with T3 (feature selection), taking them significantly higher number of prompts and time to get a satisfactory response. Feature selection, being a broad task (as compared to the remaining study tasks), had participants thinking about decomposing it to plan for sub-tasks, and providing ChatGPT sufficient context to obtain concrete and actionable next steps.

4.1.2 Prompting behaviors. In total, we logged 111 prompts made by 14 participants. Table 3 presents an overview of the prompt distribution for each study task. We observed that participants spent 64% time on preparing prompts, 27% time on adapting the code returned by ChatGPT, and the remaining on validating the code, as seen in Figure 2 (Left). In raw time spent on these activities, on average, 302 seconds were spent on prompt writing, 57 seconds on adaptation and 24 seconds on validation. We also find that 37% of the total time was spent on writing the initial prompt, whereas

Table 3. Overview of the number of prompts queried per task, and average time spent in prompt-writing.

Task	# prompts	median	min	max	avg time spent (s)
T1 – Datetime typecast	22	1	1	3	55
T2 – Split using delimiter	19	1	1	2	59
T3 – Feature selection	47	3	0	6	159
T4 – Heatmap plotting	23	2	0	4	58

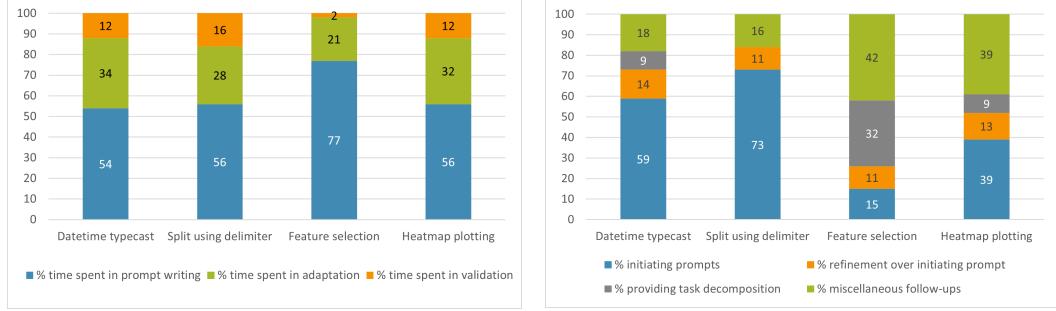


Fig. 2. Percentage of total time spent in each activity — prompt writing, code adaptation, and validation — for each task (Left). Stacked percentage distribution of initiating prompts, refinements, prompts for task decompositions, and follow-up prompts (Right).

making refinements to prompts and asking follow-up questions takes 28% of the total time. In particular, we observed that participants especially struggled to write an initial prompt for T3 (feature selection), a task with more subjectivity, when compared to relatively more objective tasks (T1, T2 and T4). The prompts also greatly varied in verbosity:

You are a data scientist, how can you do feature selection? (P13)

I have a dataset that has information about medical emergency services. I have a response time column which tells me historically how much time will it take to respond to a medical emergency incident. I have other columns some are categorical and some are continuous which explain more about the incident. My task is to predict the response time for a new incident and I am thinking about building a predictive model for that. What would be good features that can help the model predict response time? (P8)

Participants also made iterative refinements to their prompts. They sent across a plenitude of follow-up prompts for performing feature selection (T3)—re-emphasizing task goals, decomposing tasks, providing additional context and examples, and requesting for tweaks to code snippets. The need to send follow-up prompts was also exacerbated for plotting visualizations (T4), where participants would go back-and-forth with ChatGPT for minor tweaks—changes in plot size, color scheme, and font size. Figure 2 (Right) reflects on the distribution of follow-up and refined prompts for each task.

Overall, we observed that when participants interacted with ChatGPT as an assistant for providing recommendations for code, it often involved many steps for more involved tasks. Further, participants often struggled with providing information about their data and desired goals to ChatGPT and adapt its responses to solve their tasks.

		Tasks				Total
		Datetime typecast	Split using delimiter	Feature selection	Heatmap plotting	
Data Context	Sample cell	7	5	0	0	12
	Sample row(s)	0	3	7	1	11
	Data header	0	3	11	2	16
	Specific attribute name(s)	5	4	4	7	20
	Attribute format/summary	2	3	4	2	11
	Error message	1	0	2	2	5

Fig. 3. Frequency of data context shared by participants for prompt-construction for each task.

4.2 Obstacles when Communicating with ChatGPT

4.2.1 Sharing context is difficult. What information must be shared with ChatGPT? Since LLMs “cannot make sense of the raw data” (P2, P3, P7, P8, P10, P12), data scientists begin the process of prompt-writing by brainstorming what pieces of data, (such as filename, column names, column data types, sample rows, etc.) and which descriptive information (such as pattern of strings in a column, number of unique values, missing values, outlier information, minimum/maximum for numeric data, etc.) will help ChatGPT in solving the query satisfactorily. Participants reported that starting with curating a prompt is a “daunting” task (P9) because it requires decision-making on what data context needs to be included. P2 said, “*I doubt if this will be successful, I feel like we need a little bit more of numbers, but let's see if only using the column names we can get something.*”

How much information must be shared with ChatGPT? Participants often questioned *how much* information needs to be shared with ChatGPT to get a good response. Some participants made attempts to paste the entire CSV file into the prompt and encountered token limit errors (P6, P7, P13). Others also reflected on the importance of sharing only relevant instructions or snippets of the data as “*additional data seems to confuse ChatGPT into providing not-so-relevant results*” (P7, P9, P10). P9 said, “*The main obstacle I ran into here was, knowing how to prompt ChatGPT. Because I think at first I almost gave it too much information on how to perform a correlation analysis by describing it.*” After sharing 10 sample rows, P2 tried to estimate ChatGPT’s boundaries in being able to understand raw data beyond its semantics, as it provided a “*very basic response for the most relevant fields, by repeating the entire fields names and telling me their [semantic] meaning.*”

Challenges in fetching context. To construct the prompt, after assessing *what* context is relevant, participants had to go back and forth between ChatGPT’s browser window and data sources to fetch the context. They had to frequently write code snippets in their computational notebooks to gather the required context. For instance, consider a pandas DataFrame named df. Participants must write, execute, and copy the output of commands such as:

- `df.head()` to obtain a sample of rows and the data header
- `df.columns` or `df.dtypes` to obtain column names and their data types
- `df['INCIDENT_DATETIME'][0]` and `df['INCIDENT_DESCRIPTION'][0]` to obtain the first value in the respective columns to provide as few-shot examples, and so on.

We observed that P10 and P14 also implemented custom logic to extract information to share beyond what is typically shared in a prompt. For instance, P14 wanted to provide names of specific columns to ChatGPT and hence wrote the following code that computes column names that do not contain the sub-strings ‘TIME’ or ‘ID’:

```
>>> [i for i in df.columns if 'TIME' not in i.upper() or 'ID' not in i.upper()]
```

Table 4. Themes—obstacles and strategies when using ChatGPT for data-science tasks.

THEME	DESCRIPTION	REPRESENTATIVE EXAMPLES	PARTICIPANTS
Obstacles when Communicating with ChatGPT			
<i>Sharing context is difficult</i> (Section 4.2.1)	Gathering relevant information to prompt with (e.g., column names, datatypes, example datapoints) takes time.	“Should I add some rows for it to see? I am not sure if [it] will work...” “So this one is recommending scikit-learn which is not a package I typically use. I usually work in pyspark.” “It is all about the context which we provide. If we could refine it, it would do better.”	P1–P6, P8, P10, P12, P14 (10 of 14)
<i>ChatGPT opaquely makes assumptions</i> (Section 4.2.2)	Data scientists had to correct or adjust prompts in response to unanticipated assumptions made by ChatGPT about data.	“I will provide this example for the format” “Oh, it thinks the call type is int, let me say that it is categorical”	P1–P14 (14 of 14)
<i>Misaligned expectations</i> (Section 4.2.3)	ChatGPT often returned overly verbose answers or inappropriate code responses.	“I will skip reading any of this text, unless my code doesn’t work” “it didn’t bring up <code>to_datetime()</code> and suggested <code>strptime()</code> instead!”	P1–P10, P12– P14 (13 of 14)
Obstacles in Leveraging ChatGPT’s Responses			
<i>Generation of repeated code</i> (Section 4.3.1)	ChatGPT repeatedly generated code for imports, data ingestion and exploration, which had to be removed.	“It erased my data, I will quickly re-run the cells” “We already have pandas and the data, I’ll delete these lines [...] not required”	P1–P10, P12– P14 (13 of 14)
<i>Data and notebook management preferences</i> (Section 4.3.2)	Splitting code into cells and aligning it with data management preferences (such as creating temporary views, dropping parent columns, etc.).	“a temp df to see the new split columns” “I would not fill with ‘NULL’ [...] still need to understand these columns more”	P2–P9, P12–P14 (11 of 14)
<i>Code validation</i> (Section 4.3.3)	Data scientists often author additional code snippets to validate generated code.	“ChatGPT’s affirmative language is deceiving [...] it doesn’t really know my data” “let me print <code>df.dtypes</code> to be sure”	P1–P10, P12– P14 (13 of 14)
Strategies for Prompting and Alternate Resources			
<i>Techniques for prompt construction</i> (Section 4.4.1)	Data scientists use prompting techniques and create placeholders for streamlined context gathering from various sources	“lets ask why it selected these features, why not the others” “I will paste column names later, so that I can write the prompt first”	P1–P10, P12, P14 (11 of 14)
<i>Scaffolding with domain expertise</i> (Section 4.4.2)	Data scientists often employ techniques to guide ChatGPT with their domain knowledge to elicit better responses.	“I can be more specific and give it only the float columns” “I will exclude the time columns to get good features”	P5, P10, P14 (3 of 14)
<i>Choosing alternate resources over ChatGPT</i> (Section 4.4.3)	Making use of web-search; and reuse of previously authored notebooks/pipelines	“must write longer queries instead of 8-10 key words just because I can” “copy paste canned code [...] 80% of it would be repeating”	P3–P5, P7, P11, P12 (6 of 14)

Further, P5, P7 and P8 had to leave their notebook environment to open the CSV file in Excel. This was because the participants wanted to share values from a column, and the raw output from `df.head()` could not be selected freely

to be copied to the clipboard. Opening the file in Excel enabled them to select any range of cells and copy them with ease without writing additional code.

In addition to data context, participants also used snippets of their code with error messages to prompt ChatGPT for a fix (P6, P7, P12). Figure 3 shows the frequency of different types of data context shared with ChatGPT corresponding to each task.

After fetching the relevant data context, participants had to paste it into the prompt. Participants would often manually format the pasted information to make the prompt look “cleaner,” to ensure that ChatGPT can understand the context (P2, P3, P7). While solving T3 (feature selection), P2 wanted to fetch column names, and organize them such that each column name appears in a new line in the prompt. To this, he said, “*Can you copy and paste it for me please? By any chance would you have like all the columns of the dataset one name per line? In the meantime, I'll be typing [the prompt]... I am so lazy.*”

4.2.2 ChatGPT opaquely makes assumptions. Based on the context shared with ChatGPT, it made its own “mental model” of the data. Participants were often enthusiastic about the semantic capabilities of ChatGPT, and how it could infer domain knowledge about the data solely using the column names. P4 leveraged these capabilities and prompted ChatGPT to provide a data dictionary. Several participants (P5, P7, P10, P12–P14) were amazed when ChatGPT could understand the data domain sufficiently to convert the extracted ‘RESPONSE_TIME’ column from `timedelta` format to seconds—which was the most appropriate unit of measurement, since emergency medical services are likely to be dispatched quickly. However, despite the semantic abilities of the LLM, participants struggled with several scenarios, and had to re-assess ChatGPT’s understanding of the task and the data. We observed several instances across every task, where lack of adequate context in the prompt led ChatGPT to make assumptions about the data and its format.

Several participants (P1, P5, P6) received incorrect code generations for T1 (datetime typecast), where ChatGPT assumed the string to be of the format ‘%d%m%Y %H:%M:%S’ (e.g., 31-07-2020 23:59:51) instead of ‘%Y-%m-%dT%H:%M:%S.%f’ (e.g., 2020-07-31T23:59:51.000). This assumption led to generation of the following code snippet to typecast the string column to datetime format, resulting in a run-time error:

```
>>> pd.to_datetime(date_str, format='%d%m%Y %H:%M:%S')
Error: time data '2020-07-31T23:59:51.000' does not match the format '%d%m%Y %H:%M:%S' (match)
```

Upon providing ChatGPT with the error message, or an example string from the column, it re-generated the snippet without the `format` parameter, allowing successful execution, as pandas can now correctly infer the format of the input datetime string itself.

Next, as part of T2 (split using delimiter), participants had to work with the ‘INCIDENT_DESCRIPTION’ column, containing the borough, incident dispatch area code, and zipcode, separated by semicolons. However, few rows in this column did not contain the zipcode. This led to the column having two formats: (1) “<borough>; <incident_dispatch_area_code>; <zipcode>”, for example: “Brooklyn; K7; 11211”, and (2) “<borough>; <incident_dispatch_area_code>”, for example: “Manhattan; M3”. Participants generally did not discover this data-formatting inconsistency by themselves, since encountering such issues requires exploratory data analysis and on-ramping. As a result, ChatGPT often did not get visibility into this data quality issue through participant prompts (P1, P2, P5, P11), and it generated non-executable code for splitting the column under the assumption of having clean and consistent data, which contains all three splits in each row:

```
# split the INCIDENT_DESCRIPTION column by the ; separator
>>> df['Borough'] = df['INCIDENT_DESCRIPTION'].str.split(';').str[0]
```

```
>>> df['Precinct'] = df['INCIDENT_DESCRIPTION'].str.split(';').str[1]
>>> df['ZipCode'] = df['INCIDENT_DESCRIPTION'].str.split(';').str[2]
IndexError: list index out of range
```

On the other hand, some participants were able to receive correct code generations from ChatGPT owing to their relatively generic prompts. In such cases, where the generated code ran without errors, participants never discovered the underlying data quality issue—leading them to be unaware about missing values in the newly derived column for zipcode (P1, P3–P4, P7–P14). P4 expected similar behavior for other data smells and quality issues, such as “*typos and capitalization inconsistencies in categorical variable columns*,” and expressed the *need* to be able to quickly discover and convey such data quality issues to the LLM.

In tasks T3 (feature selection) and T4 (heatmap plotting), ChatGPT often made assumptions about the data type of columns solely based on the column names. That is, when participants did not share sample rows, or the column data types, ChatGPT used the column names to make a semi-informed guess about the data type. This led to run-time errors in T3 while trying to train a model (P5, P12), and visualizations that were not meaningful for any insight-finding in T4, such as scatter-plots for categorical and boolean variables (seen in Figure 4) (P2, P3, P10).

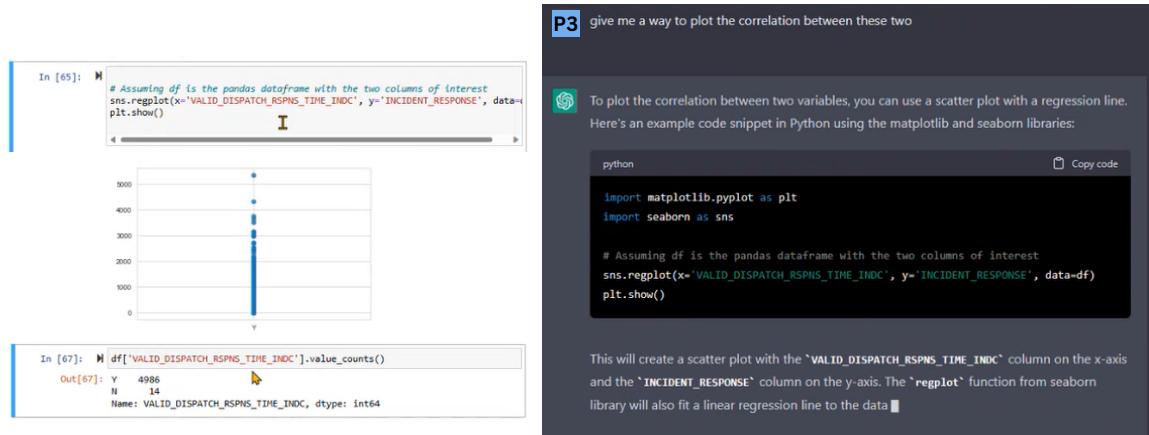


Fig. 4. ChatGPT suggests a highly skewed boolean variable for a scatter plot. P3 says that, “*Should I choose another column? [...] This being categorical and skewed isn’t useful for the visualization. So this is where I think [ChatGPT] only looked at the column names, and not the actual values in the data I shared.*”

4.2.3 Misaligned expectations. Responses generated by ChatGPT are highly temperamental to the phrasing of prompts. Participants form a mental model of the nature of responses they would get from ChatGPT, in terms of whether the response will contain code or not, and the libraries, packages and APIs it would use in code generation. However, participants report that the presence or absence of specific words in the prompt can lead to unanticipated responses from ChatGPT.

For instance, not mentioning “pandas” or “dataframe” in the initial prompt to ChatGPT led to code generations with APIs belonging to standard python libraries, instead of pandas APIs (P1, P3, P12). P12 reflected on the issue and realized that they “*didn’t talk about pandas, that’s why it didn’t bring up to_datetime() over here and suggested strftime()* instead!”

Convert datetime string like “2020-07-31T23:59:51.000” to datetime object in python (P12)

```
>>> from datetime import datetime
>>> datetime_str = '2020-07-31T23:59:51.000'
>>> datetime_obj = datetime.strptime(datetime_str, '%Y-%m-%dT%H:%M:%S.%f')
```

Similar challenges surfaced as participants solved T3 (feature selection). Owing to the subjectivity of T3, participants often mentioned that ChatGPT cannot be completely trusted for suggesting techniques and steps since it lacks domain expertise (P1–P5, P7, P8, P11). Participants observe that ChatGPT’s responses are almost never aligned with their domain knowledge, as it misses necessary data pre-processing steps, or provides technically incorrect responses. In one instance, as P5 went about prompting to train a model that predicts the ‘RESPONSE_TIME’, ChatGPT generates code to perform training without standardizing column values. P5 emphasized the importance of standardization and suggested careful examination of all generated code: *“One thing it has missed is standardizing data. It is very important for us to scale the data to check for outliers and understand the data distribution.”* In other instances, ChatGPT suggested selecting the features: ‘INCIDENT_DATETIME’ and ‘FIRST_ON_SCENE_DATETIME’, which were used to compute the ‘RESPONSE_TIME’ column. Participants expressed that this is technically incorrect since the target column was arithmetically derived using the features selected by ChatGPT (P2, P3, P5, P13, P14); and that the ‘FIRST_ON_SCENE_DATETIME’ information will not be available in a pragmatic setting—making it impossible to predict the ‘RESPONSE_TIME’ (P8).

Lastly, participants often found ChatGPT’s responses to be *“excessively verbose and unnecessary”*, especially for natural language (NL) code explanations. Participants wanted ChatGPT to omit explanations for APIs that they were already familiar with, and required NL explanations for only those unknown or uncommon APIs and parameters. Several participants held the long NL explanations for code generations as responsible for the prolonged inference time, that is, the time taken by ChatGPT to produce a complete response. P7 said, *“When it was generating the explanation content, I felt like it is time consuming. Just give me the code. I was being impatient.”*

We observed that all participants skip reading any NL explanations when the response included a code-snippet, and begin with copying the generated code to adapt it to their notebook. P4 said, *“I guess it takes a bit of time to finish running. It would just be easier for me if it just outputs the code, and then I can ask follow up question on what does n = 2 in the first line of the function do.”*

4.3 Obstacles in Leveraging ChatGPT’s Responses

We now present the challenges faced by participant in *using* the code generated by ChatGPT. Participants had to edit the generated code in certain ways to successfully adapt it to their notebook contexts. Since the chat instance with ChatGPT was completely detached from the notebook, unless participants explicitly provided variable and attribute names in their prompt, ChatGPT created dummy dataframes and attribute names and included them in the generated code. Thus, the adaptation process frequently involved copying the generated code and replacing the generated variable and attribute names with the actual identifiers. These trivial, yet additional, steps in the adaptation process can be fixed by augmenting the user prompt (i.e., scaffolding the user’s NL prompt with a prefix, and/or suffix to provide relevant context to the model) with the name of the dataframe, attribute names, and any other variables involved. However, we identify several other repeated obstacles across participants, which either require more involved prompt augmentation, post-processing of the generated code, or interface design interventions for easier adaptation. Figure 5 details the nature of code adaptation and validation activities and their frequencies across tasks, with managing import statements and removing additional code generated for data ingestion being the most frequent adaptations.

		Tasks				Total
		Datetime typecast	Split using delimiter	Feature selection	Heatmap plotting	
Code Adaptation	Fixed attribute names	5	6	4	1	16
	Managed imports	10	10	7	4	31
	Removed data ingestion	7	10	6	2	25
	Cell management	1	3	4	1	9
	Data management	2	5	0	1	8
	Apply to entire column	3	1	0	0	4
Validation	Test with a cell	3	1	0	0	4
	View DataFrame	4	6	2	0	12
	View derived attribute(s)	2	7	3	0	12
	View attribute summary	9	3	1	0	13
	View plot(s)	0	0	0	9	9

Fig. 5. Frequency of code adaptation and validation activities performed by participants for each task.

4.3.1 Generation of repeated code. A common adaptation the participants had to make to the generated code was removing lines that were already present in their notebook. This mostly included (1) repeated library import statements—especially `import pandas as pd`—which were generated with every code response generated by ChatGPT, (2) repeated data ingestion using the `pandas.read_csv()` API, or by hard-coding a sample of rows as a python dictionary while using the `pandas.DataFrame.from_dict()` API, and (3) performing exploratory data analysis using `pandas.DataFrame.describe()` and other APIs. Accidentally executing this code sometimes led the participants' dataframes to be over-written by the sample values generated by ChatGPT (Figure 6), causing them to re-run all notebook cells from the beginning (P1, P9).

```

df = pd.DataFrame({
    'INCIDETN_DESCP': ['Region1; value1; value2', 'Region2; value3; value4']
})

# split the INCIDETN_DESCP column into three columns
df[['Region', 'col2', 'col3']] = df['INCIDETN_DESCP'].str.split(';', expand=True)

# remove the original INCIDETN_DESCP column
df = df.drop('INCIDETN_DESCP', axis=1)
df

```

	Region	col2	col3
0	Region1	value1	value2
1	Region2	value3	value4

Fig. 6. P1 accidentally executes data ingestion code generated by ChatGPT, causing `df` to be over-written by a dummy dataset.

4.3.2 Data and notebook management preferences. We now discuss a few issues related to data and notebook management preferences:

Managing generated code in computational notebooks. We observe that data scientists have implicit preferences for code management, aligned with common practices in the use of computational notebooks, such as having all import statements in a single cell at the beginning of the notebook, splitting functions and data transformations into cells to view intermediate results, interleaving of code with markdown cells to segment the notebook, and so on.

ChatGPT often generated big chunks of code involving multiple data transformations as a single code-block in its responses. Participants often broke-down such code chunks into meaningful cells (P2, P3, P5, P7, P8, P14). This helped them in isolating lines of code into steps, and execute them sequentially to observe any execution errors. This also

enabled them to view the dataframe in between the data-transformation steps, while ensuring that the generated code matches their intent. Further, ChatGPT bundled library import statements with other data transformation/visualization code in the same block. This frequently required participants shifting the new import statements to the cells at the top of their notebook (P3, P5, P7, P12, P13).

Data scientists have also been studied to commonly author exploratory code; or use the fluent programming pattern, composing multiple transformations into a chain [4, 12]. When participants mentioned the phrase “write a function” in their prompts, it led to generation of functions for data transformation tasks, with excessive number of parameters and intermediate steps, contrary to the typical exploratory and fluent patterns. P3 and P7 refactored such code generations to obtain chain of transformations for the target column.

Data transformation and management. Participants often had preferences on how data transformation must be performed, such as—deciding to retain the parent column(s) after having derived newer column(s) (P3, P8, P9, P12), creating temporary dataframes for data transformation tasks (P12, P13), and deciding on appropriate techniques for data imputation (P6). Adaptations had to be made to the generated code to take these transformation preferences into account. For instance, when P6 prompted ChatGPT for T2 (split using delimiter), it generated an additional line of code to impute missing values in zipcode with “NULL”. P6 said, “*No, no, I would not have preferred filling them with ‘NULL’ because right now probably I still need to understand these columns more, so I may just remove null values afterwards,*” and subsequently removed the line of code.

4.3.3 Code validation. Participants emphasized on the importance of thoroughly validating every operation performed using LLM-generated code. P7 mentioned that validation is even more essential as “*ChatGPT’s confirmatory and affirmative language in responses—like ‘Definitely! Here’s the code you need’—is extremely deceiving because it doesn’t really know about my data.*”

Though code validation took the least amount of time when compared to other activities (prompt construction and code adaptation) (Figure 2), each task required participants to verify the functioning of the generated code in different ways. In some cases, ChatGPT self-generated the code snippet(s) required to validate the data transformation. In other cases, the participants wrote additional code snippets, or manually inspected the data to validate changes. For instance, to ensure that ‘INCIDENT_DATETIME’ has been typecast to pd.datetime (T1), participants used `df.dtypes` and `df.columns`. To check for newly derived columns after splitting ‘INCIDENT_DESCRIPTION’ (T2), participants printed and manually inspected specific columns. Some participants also checked for missing values after using `df.value_counts()`. On the other hand, for T4 (heatmap plotting), visual verification of the generated plot would be sufficient to validate the code generated by ChatGPT.

4.4 Strategies for Prompting and Alternate Resources

We now discuss strategies employed by data scientists to overcome communication and code-adaptation obstacles, along with scenarios where they prefer to leverage other resources (e.g., web-search, documentation, and previously written code).

4.4.1 Techniques for prompt construction. Alongside data context, participants also brainstormed prompting techniques that are relevant for ChatGPT to solve the problem. We observed frequent use of one-shot and few-shot prompting for T1 (datetime typecast) and T2 (split using delimiter) (P1–P4, P6, P8, P10, P12, P14); chain of thought prompting for T3 (feature selection), to elicit reasons behind *why*, and *on what basis* ChatGPT suggests features (P1, P3, P10); and making

ChatGPT assume the role of a data scientist to receive expert suggestions, by prefixing each prompt with “*You are a data scientist*” (P13).

To avoid the frequent context-switching between prompt-writing (in the ChatGPT browser-window) and data context gathering (from the notebook/Excel), we observe that participants focused on creating a prompt skeleton while embedding “placeholders” for data context that must be fetched from the notebook (discussed in section 4.2.1), and pasted into the prompt (P1, P2, P5, P7, P8, P10, P14). For instance, P2 created placeholders (enclosed in square brackets), to replace them afterwards with the required column name and sample value:

I have a dataframe with a column called [col name]. This column has data separated by ;, here is an example: [sample value]. Can you split this column? (P2)

This strategy was commonly used by participants to complete writing the natural-language query in one go, instead of their flow being disrupted by going back and forth between ChatGPT and their notebook to fetch relevant data context. Prompt drafts with placeholders also offered a low cost in switching out pieces of data context while assessing their importance for solving the problem at hand.

4.4.2 Scaffolding with domain expertise. Participants often expressed that ChatGPT lacks “*any actual understanding of the data*” (discussed in sections 4.2.2 and 4.2.3). To overcome these obstacles, we observe limited interactions where data scientists successfully guide the LLM with their domain expertise. P10 and P14 narrowed down ChatGPT’s output space by iterative refinement of the data context included in their prompts.

To obtain meaningful visualizations (T4), P10 refined their prompt to only include names of float columns.

I would like to plot out response time vs:

[List of ‘float’ attribute names]

Can you write python code that does this in seaborn? (P10)

Similarly, to prevent ChatGPT from suggesting timestamp-based features (T3), P14 provided only names of columns that do not contain the sub-strings ‘TIME’ or ‘ID’ in them.

I want to build a GBM using sklearn. The X variables are [List of attribute names not containing sub-strings ‘TIME’ or ‘ID’], target variable is “RESPONSE_TIME” which is timedelta (P14)

Surprisingly, we also observe an instance with P5, where ChatGPT asks a “clarifying question” in response to the user provided prompt:

*P5: Consider this dataset to be the dataframe which we are going to use for the prediction of the response time:
[Top 10 rows with data header]*

ChatGPT: What is response time in this dataset and how is it calculated?

P5: Response time is calculated as below: FIRST_ON_SCENE_DATETIME – INCIDENT_DATETIME

This interaction made P5 feel assured that ChatGPT “*is able to understand the prompt*”, and “*make sense of the data context*”. The clarifying question enabled P5 to provide context that they missed in their initial prompt, and allowed for complete intent expression instead of generating a hallucinated response.

4.4.3 Choosing alternate resources over ChatGPT. There were several factors that lead to a preference to use alternative resources over ChatGPT:

Data privacy concerns. Participants reflected on the importance of maintaining privacy for sensitive and personally identifiable information (PII). P2 was hesitant about using any confidential business data with ChatGPT and said, “Do Manuscript submitted to ACM

you want me to use normal ChatGPT? And we are not using that for any sensitive data, right, because I don't want to share any private data with ChatGPT.” Contemporary tools for data might inherently include data context in their prompt, to scaffold user queries. For this reason, participants expressed interest in controlling and reviewing any data context or information being passed on to the LLM. P3 expressed concerns over having their data become part of the model’s training and being presented to a different ChatGPT user as a result to their queries. Participants were comfortable with using enterprise ChatGPT only in accordance with company policy (P2, P7).

Re-using previously written code. Data scientists mentioned that they periodically receive new batches of data, for which they must re-run analyses and visualization reports, and re-train models (P3, P5, P6, P7, P11). In such cases, data scientists are already familiar with the structure of the data, and they reuse previously written code for data pre-processing and cleaning. P11 mentioned, “*We pretty much just copy and paste canned code for every project. A lot of variations can be created on the charts, but 80% of it would be repeating.*” The need to author new code for pre-processing or analytics in such scenarios is rare.

Using web search and documentation. Participants described that using LLMs can normally help them save time over making a web search, such as when “*asking for generic directions and steps, saving me the time to go through 4–5 blog posts on the web,*” and “*asking for code in my specific context, which needn't be the exact issue discussed on Stack Overflow*” (P4, P11). In these cases, they find the LLM’s responses to be highly contextualized to their data and previously applied steps, and “*less noisy*” compared to web-pages. On the other hand, P7 found it easy to make a web search, requiring only “*8–10 key words,*” whereas they must write “*longer queries and be more specific with ChatGPT just because it allows me to do that.*” P11 also mentioned that the use of LLMs will “*restrict me from learning new techniques or discovering a breadth of alternate approaches. Forums like Stack Overflow support these needs better.*”

5 SURVEY RESULTS

Our qualitative insights from the task-based study (Section 4) identify numerous obstacles that data scientists face in conversing with ChatGPT, and subsequently, in acting on the generated responses. The survey results help us understand how well do these obstacles generalize to a diverse set of data science tasks, and to a broader community of data scientists. The survey respondents report experience with using LLMs for varied data science tasks spanning across project timelines, including synthetic data generation, data wrangling, pre-processing, data labelling, exploratory data analysis, insight finding and summarization, hypothesis generation, training various models, outlier detection, and generating plots.

Figure 7 shows the distribution of responses for each ‘Agree’/‘Disagree’ question. Results indicate that data scientists find sharing of relevant data context to be important (86%), as well as tedious (59%), echoing findings from Section 4.2.1. 62% data scientists agreed that prompting ChatGPT for columns with mixed formats is challenging, reflecting the issues discussed in Section 4.2.2. These challenges lead to communication breakdowns between data scientists and ChatGPT—causing them to make repeated prompt refinements (92%) (Section 4.1.2). Respondents also indicate that LLMs do not have sufficient domain expertise in data science (60%) (Section 4.2.3), and that generated code requires several modifications before it can be used in their notebooks (87%) (Section 4.3). Lastly, code validation respondents have a split opinion on whether it is easy to verify code generated by ChatGPT (48%) (Section 4.3.3).

We manually inspected responses to look for any other challenges encountered by data scientists in conversing with LLMs for data science tasks. Respondents re-emphasized difficulties in “*forming well scoped prompts*” and deciding *what* context to include. They reflected on the need for repeated prompt refinements due to ChatGPT’s lack of domain

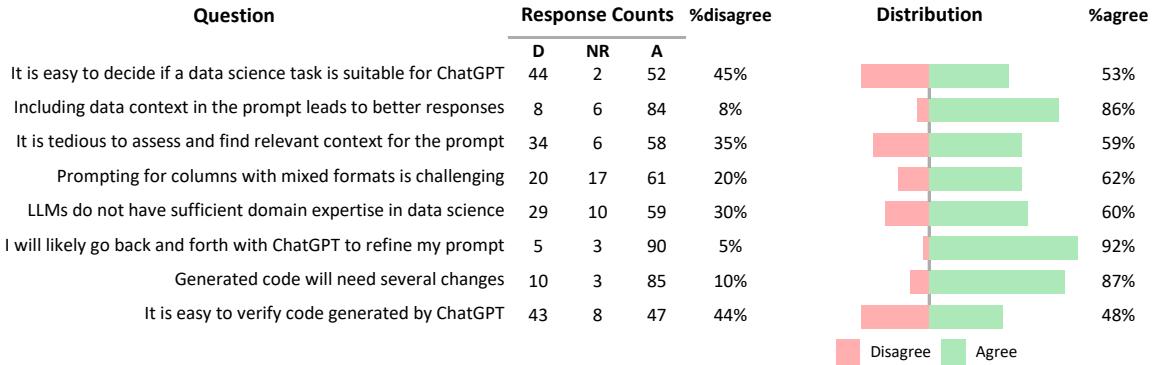


Fig. 7. Distribution of survey responses for ‘Agree’/‘Disagree’ questions. Response Counts: D (Disagree); NR (No Response); A (Agree).

expertise. A few responses brought up the “*instability*” in ChatGPT’s responses, wherein it generates drastically different answers despite sending the same prompt. We did not encounter any new obstacles in response to this question.

6 DESIGN RECOMMENDATIONS

In this section, we discuss the implications of our findings and identify the ways in which they could be adapted to various data-science environments and to broader contexts.

6.1 Recommendation I—Provide preemptive and fluid context when interacting with AI assistants

From our studies, we found that participants spent a significant amount of time constructing prompts (Section 4.1.2) and gathering and expressing their context (Section 4.2.1) to ChatGPT. Nearly 45% of the prompts included portions of data that they either manually entered, or wrote additional code snippets to obtain; however, participants did not enjoy being “*slowed-down*” in fetching this data context for every interaction. Thus, data scientists may need help in providing context about their environments, and we should support mechanisms to either preemptively provide implicit context, or provide fluid interactions that allow data scientists to easily refer to fragments of their work in conversations with AI assistants.

In data visualization, classic techniques such as *brushing* [3], enable a user to select and highlight specific data points, allowing the user to use those data points as a filter, or as a point of interest for further analysis. Similarly, providing mechanisms for supporting *data brushing* in data-science environments would allow users share their context with agents more directly. For example, in T2, participants had to split text into three columns. However, some rows had inconsistent formatting making the operation suggested by ChatGPT error-prone. With data brushing (Figure 8), participants could mark an example cell with a problematic entry and ask an AI assistant to help split the rows, while dealing with outliers such as the marked cell. Similarly, allowing inline code snippets that are expanded, or referring to entities within the data (@INCIDENT_RESPONSE), allow less tedious and more fluid expressions of context when prompt writing.

Despite these aids, some context, especially providing a summary and schema of the data would be tiresome. Furthermore, users had no clear understanding of what information should be provided, and even for researchers, this question is still being explored [8]. However, recall Grice’s maxim of relevance, which states that a successful conversation requires not too little or too much information be provided [14]. A simple intermediate measure would be

AI ASSISTANT	INCIDENT_DESC	# POLICEPRECINCT	# CITYCOUNCILDISTRICT
What is this dataset about?	BROOKLYN; K7; 11211	94	34
The EMS Incident dataset contains data that is generated by the EMS Computer Aided Dispatch System. The data spans from the time the incident is created in the system to the time the incident is closed in the system. This is a dataset hosted by the City of New York, and covers information about the incident and allocated resources.	QUEENS; Q3; 11419	106	28
	MANHATTAN; M4; 10025	24	7
	BROOKLYN; K7; 11221	81	41
	QUEENS; Q2; 11434	113	27
	BRONX; B2; 10456	44	16
	QUEENS; Q6; 11368	115	21
	QUEENS; Q1; 11692	101	31
	MANHATTAN; M8; 10030	32	9
	BRONX; B2; 10457	48	17
	BRONX; B1; 10456	40	17
	BRONX; B3; 10472	43	18
	BRONX; B1; 10456	42	16
	QUEENS; Q7	Missing value	Missing value
	MANHATTAN; M3	Missing value	Missing value
	BROOKLYN; K3; 11203	67	41
	QUEENS; Q5; 11435	107	24
	MANHATTAN; M4; 10025	24	7
	BROOKLYN; K5; 11221	79	36
	MANHATTAN; M8; 10030	22	9

Fig. 8. Design concept—Using data brushing to mark examples (selected in green), and inconsistent entries (selected in red).

to provide a preemptive summaries of data characteristics. This would help reduce the chances that ChatGPT makes assumptions about data format, which may lead to incorrect code or longer conversations (Section 4.2.2).

6.2 Recommendation II—Provide inquisitive feedback loops and validation-aware operations

For transactional tasks, such as T1 (datetime typecast), ChatGPT often provided satisfactory responses. Unfortunately, open-ended tasks, such as T3 (feature selection), required a more involved conversation with ChatGPT. Initial responses from ChatGPT were often “excessively long” and were not particularly helpful for the task, likely because it violates the fundamental HCI principle of progressive disclosure [32] and overwhelms the user with details. As a result, participants were uncertain how much value would be gained by continuing the conversation or what further information would help (Section 4.2.3). In other cases, participants had to make multiple clarifications and carry out editing operations across several iterations in their data-science environments. This task fragmentation often lead to errors, for example, because ChatGPT always included code to ingest the dataframe—regardless of the stage of conversation—some participants accidentally executed it, leading to their data being over-written (Section 4.3.1).

One mechanism to help participants have more productive conversations with AI assistants is through *inquisitive feedback loops*, where the system guides the user in expressing what they need by proactively asking the user a series of questions, rather than waiting for the user to know when they need help and how to get it [17]. Moreover, participants in our study could have been aided just by ChatGPT asking clarifying questions to resolve ambiguities instead of making assumptions. For example, one participant (P9) was able to get more useful advice from ChatGPT, after suggesting that they could reframe the problem of feature selection: “*I think I am more interested in finding the feature importance so I'll prompt it for something like a random forest*”. In this case, a user could be prompted with questions that helped the user think through which high-level strategy they wanted to use before giving all the details.

Tools for supporting *refactoring*, behavioral-preserving transformations, can be one approach for assisting data scientists who must make repeated edits of similar nature, or adapt generated code snippets to their local notebook contexts (Section 4.3). For instance, GHOSTFACTOR [13] is a light-weight program analysis tool with refactoring detection

algorithms that can identify incomplete changes to code. Simple approaches can be used for enabling automated post-processing and human-in-the-loop refactoring of generated code snippets based on observed nature of edits—deletion of repeated code, notebook and data management preferences, and validation.

6.3 Recommendation III—Provide transparency about shared context and domain expertise slots

Barke et al. [1] posit that lack of transparency about the shared context, and lack of control in refining the context selections can leave users in a confused state. Our participants echo a similar consideration, wherein they expressed the urge to “*get to the first response as fast as possible, but have the ability to look at what context was provided and edit it if I need to refine my prompt*” (P3, P4, P7, P12). Furthermore, participants also wanted to be able to provide more context that reflected their personal expertise and experience (Section 4.4.2).

Even with automated context management or prompt augmentation, it is important to consider how that context is ultimately shared. For example, by having sufficient transparency into the shared context can also enable data scientists to ensure that sensitive data does not leak into the prompt. Data-science environments interacting with AI should consider mechanisms for ensuring a two-way knowledge transfer between data scientists and AI. For example, an AI assistant chat interface could integrate a *context panel*, a dedicated, editable grid that mirrors the AI assistant’s and user’s assumptions about data and tasks.

Finally, users should be able to inspect and modify the context to provide their domain expertise or reflect their analysis style (e.g., always perform data scaling and impute missing values prior to training a model). One step in this direction is the recent addition to ChatGPT allowing for custom instructions⁶.

7 THREATS TO VALIDITY

All study participants had prior experience with using ChatGPT. However, most participants (9 of 14) had not used LLMs for solving data-science tasks. If they had been exposed to tasks of similar nature prior to the study session, their interactions with ChatGPT could have been different. We took this into consideration to ensure the quality of responses in the confirmatory survey, and asked all respondents to share at least one scenario where they used an LLM-powered tool for a data-science task. This enabled us to filter out respondents who did not have any prior experiences to provide informed responses.

We selected a diverse set of data-science tasks representative of data pre-processing, cleaning, feature selection, and plotting. However, the study duration limited the breadth of tasks that participants could have performed, such as data discovery, capture, and machine learning [31]. Nevertheless, our survey responses (Section 5), from a broader community of data scientists with diverse experiences, confirm the generalizability of our findings. Future work can involve empirical studies that measure the performance of LLMs on assorted data-science tasks with datasets that have not been “seen” by the LLMs during pre-training. We presented the data-science tasks to all participants through verbal descriptions of the objectives. This may have inevitably led to priming, wherein the participants’ formulation of prompts could have been influenced by our description of the task.

Browser-based version of ChatGPT version was updated as we conducted our task-based studies. P1–P9 used the March 14, 2023 version of ChatGPT, and P10–P14 used the March 23, 2023 version. This might have had a slight impact on study findings, though we did not observe any noticeable shifts in the nature of responses generated by ChatGPT for the study tasks. Lastly, we encouraged participants to think aloud as they solved study tasks, which could have

⁶<https://openai.com/blog/custom-instructions-for-chatgpt>

added to the time taken in performing different activities. Our video annotations take this into consideration to only include timestamps where participants are actively working with the data, or with ChatGPT.

8 RELATED WORK

AI-assisted tools for data cleaning suggestions—such as Trifacta⁷ [16, 22], Data Wrangler⁸, CoWrangler [9], AWS Glue DataBrew⁹, Salesforce Einstein Discovery¹⁰, AutoPandas [2], Auto-Suggest [45]—have existed, but LLMs bring forth new opportunities for enhancing these tools further. More recently, several commercial and open-source tools have emerged—DataChat AI¹¹, Anaconda Assistant¹², Databricks Assistant¹³, and Jupyter AI¹⁴—which enable data scientists to access AI-powered chat assistants within their notebook (such as Anaconda, Databricks, and Jupyter notebooks), and other alternate data-science environments (e.g., DataChat AI, SQL and file editors for Databricks, etc.). The semantic abilities of LLMs, coupled with a chat interface, provides users with the agency to converse about their data, ask follow-up questions, and receive highly contextualized responses.

McNutt et al. [30] studied, through an interview study, the design space of code assistants in computational notebooks, but their focus is on the design space as they do not focus on data-specific issues that data scientists face while communicating their needs to the AI assistants (e.g., communicating the data context or domain knowledge). A number of recent work conducted studies to understand the use of LLMs by software engineers and programmers [1, 19, 20, 28, 37, 42, 43]. However, these mostly focus on the general understanding of AI-powered code assistants and do not particularly focus on data-science tasks, where communicating the data at hand to the assistant is a big challenge.

There has been a recent line of work that try to make use of LLMs to automate data-science tasks and evaluate performance of LLMs on data-science tasks [18, 33, 46]. This is a promising, but orthogonal, area of research to ours as we focus on human-in-the-loop data science where AI-assistants and human data scientists work as a team. Similarly, there are work on designing and evaluating tools for end users [10, 29, 40], which is complementary to our work.

Most relevant to our work is the study by Liu et al. [29] on how users of a spreadsheet software articulate their question to an AI-powered tool and how a system they designed can better support this process. Our work is differentiated by (1) examining the fundamental interactions with a general AI tool like ChatGPT rather than a domain-specific AI tool embedded in a system, (2) we studied professional data scientists rather than non-programmers, and (3) they were specifically investigating the effect of co-editing a prompt by splitting it into natural language steps.

9 CONCLUSION AND FUTURE WORK

The recent rise of AI-powered chat assistants gives hope of greatly increasing the productivity of data scientists as they are engaged in tedious data-wrangling tasks. To understand the fundamental obstacles, needs, and design opportunities of data scientists using AI assistants, we conducted mixed-method studies involving a task-based study, interviews, and a survey with professional data scientists. We found that participants faced two key sets of barriers: when communicating with ChatGPT and then when adapting ChatGPT’s response to aid their specific situation. For example, participants spent considerable time collecting and formatting information to provide to ChatGPT, such as relevant regions of their

⁷<https://docs.trifacta.com/display/SS/Overview+of+Predictive+Transformation>

⁸<https://devblogs.microsoft.com/python/data-wrangler-release/>

⁹<https://docs.trifacta.com/display/SS/Overview+of+Predictive+Transformation>

¹⁰https://help.salesforce.com/s/articleView?id=sf.bi_edd_prep_terminology.htm

¹¹<https://datachat.ai/>

¹²<https://www.anaconda.com/blog/anaconda-assistant-launches-to-bring-instant-data-analysis-code-generation-and-insights-to-users>

¹³<https://www.databricks.com/blog/introducing-databricks-assistant>

¹⁴<https://jupyter-ai.readthedocs.io/>

dataset or descriptive statistics. Furthermore, ChatGPT often made assumptions unbeknownst to the participant, thus leading to bugs in the code and additional steps to remedy. When participants did get a useful response from ChatGPT, it required considerable effort to adapt the code to work with the rest of their code and to follow their style. We provide key design recommendations for tools based on our study, such as the need to provide inquisitive feedback loops and validate-aware operations. As a next step, we intend to operationalize these design guidelines into a working tool and assess its efficacy in alleviating the workflow challenges that data scientists encountered in our studies. In conclusion, our work aims to function as a significant catalyst in democratizing data science via AI-powered chat assistants.

REFERENCES

- [1] Shraddha Barke, Michael B. James, and Nadia Polikarpova. 2023. Grounded Copilot: How Programmers Interact with Code-Generating Models. *Proc. ACM Program. Lang.* 7, OOPSLA1, Article 78 (apr 2023), 27 pages. <https://doi.org/10.1145/3586030>
- [2] Rohan Bavishi, Caroline Lemieux, Roy Fox, Koushik Sen, and Ion Stoica. 2019. AutoPandas: Neural-Backed Generators for Program Synthesis. *Proc. ACM Program. Lang.* 3, OOPSLA, Article 168 (oct 2019), 27 pages. <https://doi.org/10.1145/3360594>
- [3] Richard A. Becker and William S. Cleveland. 1987. Brushing Scatterplots. *Technometrics* 29, 2 (1987), 127–142. <https://doi.org/10.1080/00401706.1987.10488204>
- [4] Mary Beth Kery and Brad A. Myers. 2017. Exploring exploratory programming. In *2017 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, New York, NY, USA, 25–29. <https://doi.org/10.1109/VLHCC.2017.8103446>
- [5] Ángel Alexander Cabrera, Marco Túlio Ribeiro, Bongshin Lee, Robert Deline, Adam Perer, and Steven M. Drucker. 2023. What Did My AI Learn? How Data Scientists Make Sense of Model Behavior. *ACM Trans. Comput.-Hum. Interact.* 30, 1, Article 1 (mar 2023), 27 pages. <https://doi.org/10.1145/3542921>
- [6] Souti Chattopadhyay, Ishita Prasad, Austin Z. Henley, Anita Sarma, and Titus Barik. 2020. What’s Wrong with Computational Notebooks? Pain Points, Needs, and Design Opportunities. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (*CHI ’20*). Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3313831.3376729>
- [7] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchin Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating Large Language Models Trained on Code. [arXiv:2107.03374 \[cs.LG\]](https://arxiv.org/abs/2107.03374)
- [8] Wenhui Chen. 2023. Large Language Models are few(1)-shot Table Reasoners. [arXiv:2210.06710 \[cs.CL\]](https://arxiv.org/abs/2210.06710)
- [9] Bhavya Chopra, Anna Fariha, Sumit Gulwani, Austin Z. Henley, Daniel Perelman, Mohammad Raza, Sherry Shi, Danny Simmons, and Ashish Tiwari. 2023. CoWrangler: Recommender System for Data-Wrangling Scripts. In *Companion of the 2023 International Conference on Management of Data* (Seattle, WA, USA) (*SIGMOD ’23*). Association for Computing Machinery, New York, NY, USA, 147–150. <https://doi.org/10.1145/3555041.3589722>
- [10] Kasra Ferdowsi, Jack Williams, Ian Drosos, Andy Gordon, Carina Negreanu, Nadia Polikarpova, Advait Sarkar, and Ben Zorn. 2023. ColDeco: An End User Spreadsheet Inspection Tool for AI-Generated Code. In *IEEE Symposium on Visual Languages and Human-Centric Computing*. IEEE, New York, NY, USA. <https://www.microsoft.com/en-us/research/publication/coldco-an-end-user-spreadsheet-inspection-tool-for-ai-generated-code/>
- [11] Danyel Fisher, Rob DeLine, Mary Czerwinski, and Steven Drucker. 2012. Interactions with Big Data Analytics. *Interactions* 19, 3 (may 2012), 50–59. <https://doi.org/10.1145/2168931.2168943>
- [12] Martin Fowler. 2005. Bliki: Fluentinterface. <https://www.martinfowler.com/bliki/FluentInterface.html>
- [13] Xi Ge and Emerson Murphy-Hill. 2014. Manual Refactoring Changes with Automated Refactoring Validation. In *Proceedings of the 36th International Conference on Software Engineering* (Hyderabad, India) (*ICSE 2014*). Association for Computing Machinery, New York, NY, USA, 1095–1105. <https://doi.org/10.1145/2568225.2568280>
- [14] Paul Grice. 1991. *Studies in the Way of Words*. Harvard University Press, Cambridge, Mass. [u.a.]
- [15] Philip J. Guo, Sean Kandel, Joseph M. Hellerstein, and Jeffrey Heer. 2011. Proactive Wrangling: Mixed-Initiative End-User Programming of Data Transformation Scripts. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology* (Santa Barbara, California, USA) (*UIST ’11*). Association for Computing Machinery, New York, NY, USA, 65–74. <https://doi.org/10.1145/2047196.2047205>
- [16] Philip J. Guo, Sean Kandel, Joseph M. Hellerstein, and Jeffrey Heer. 2011. Proactive Wrangling: Mixed-Initiative End-User Programming of Data Transformation Scripts. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology* (Santa Barbara, California, USA) (*UIST ’11*). Association for Computing Machinery, New York, NY, USA, 65–74. <https://doi.org/10.1145/2047196.2047205>
- [17] Austin Z. Henley, Julian Ball, Benjamin Klein, Aiden Rutter, and Dylan Lee. 2021. An Inquisitive Code Editor for Addressing Novice Programmers’ Misconceptions of Program Behavior. In *Proceedings of the 43rd International Conference on Software Engineering: Joint Track on Software Engineering Education and Training* (Virtual Event, Spain) (*ICSE-JSEET ’21*). IEEE Press, New York, NY, USA, 165–170. <https://doi.org/10.1109/ICSE-SEET52601>

- 2021.00026
- [18] Noah Hollmann, Samuel Müller, and Frank Hutter. 2023. LLMs for Semi-Automated Data Science: Introducing CAAFE for Context-Aware Automated Feature Engineering. arXiv:2305.03403 [cs.AI]
 - [19] 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* (Bend, OR, USA) (*UIST '22*). Association for Computing Machinery, New York, NY, USA, Article 64, 15 pages. <https://doi.org/10.1145/3526113.3545659>
 - [20] Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and Applications of Large Language Models. arXiv:2307.10169 [cs.CL]
 - [21] Sean Kandel, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. 2011. Wrangler: Interactive Visual Specification of Data Transformation Scripts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) (*CHI '11*). Association for Computing Machinery, New York, NY, USA, 3363–3372. <https://doi.org/10.1145/1978942.1979444>
 - [22] Sean Kandel, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. 2011. Wrangler: Interactive Visual Specification of Data Transformation Scripts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) (*CHI '11*). Association for Computing Machinery, New York, NY, USA, 3363–3372. <https://doi.org/10.1145/1978942.1979444>
 - [23] Sean Kandel, Andreas Paepcke, Joseph M Hellerstein, and Jeffrey Heer. 2012. Enterprise data analysis and visualization: An interview study. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2917–2926.
 - [24] Mary Beth Kery, Marissa Radensky, Mahima Arya, Bonnie E. John, and Brad A. Myers. 2018. The Story in the Notebook: Exploratory Data Science Using a Literate Programming Tool. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) (*CHI '18*). Association for Computing Machinery, New York, NY, USA, 1–11. <https://doi.org/10.1145/3173574.3173748>
 - [25] Miryung Kim, Thomas Zimmermann, Robert DeLine, and Andrew Begel. 2018. Data Scientists in Software Teams: State of the Art and Challenges. *IEEE Transactions on Software Engineering* 44, 11 (2018), 1024–1038. <https://doi.org/10.1109/TSE.2017.2754374>
 - [26] Sean Kross and Philip J. Guo. 2019. Practitioners Teaching Data Science in Industry and Academia: Expectations, Workflows, and Challenges. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) (*CHI '19*). Association for Computing Machinery, New York, NY, USA, 1–14. <https://doi.org/10.1145/3290605.3300493>
 - [27] J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1 (1977), 159. <https://doi.org/10.2307/2529310>
 - [28] Jenny T. Liang, Chenyang Yang, and Brad A. Myers. 2023. Understanding the Usability of AI Programming Assistants. arXiv:2303.17125 [cs.SE]
 - [29] 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>
 - [30] 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* (Hamburg, Germany) (*CHI '23*). Association for Computing Machinery, New York, NY, USA, Article 434, 16 pages. <https://doi.org/10.1145/3544548.3580940>
 - [31] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q. Vera Liao, Casey Dugan, and Thomas Erickson. 2019. How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) (*CHI '19*). Association for Computing Machinery, New York, NY, USA, 1–15. <https://doi.org/10.1145/3290605.3300356>
 - [32] Jakob Nielsen. 2006. Progressive disclosure. *nnngroup.com* (2006).
 - [33] David Noever and Forrest McKee. 2023. Numeracy from Literacy: Data Science as an Emergent Skill from Large Language Models. arXiv:2301.13382 [cs.CL]
 - [34] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]
 - [35] Adam Rule, Aurélien Tabard, and James D. Hollan. 2018. Exploration and Explanation in Computational Notebooks. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) (*CHI '18*). Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3173574.3173606>
 - [36] Johnny Saldaña. 2009. *The Coding Manual for Qualitative Researchers*. <http://ci.nii.ac.jp/ncid/BB20067005>
 - [37] 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)*.
 - [38] Nischal Shrestha, Titus Barik, and Chris Parnin. 2021. Remote, but Connected: How #TidyTuesday Provides an Online Community of Practice for Data Scientists. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 52 (apr 2021), 31 pages. <https://doi.org/10.1145/3449126>
 - [39] Marita Skjuve, Asbjørn Følstad, and Petter Bae Brandzaeg. 2023. The User Experience of ChatGPT: Findings from a Questionnaire Study of Early Users. In *Proceedings of the 5th International Conference on Conversational User Interfaces* (Eindhoven, Netherlands) (*CUI '23*). Association for Computing Machinery, New York, NY, USA, Article 2, 10 pages. <https://doi.org/10.1145/3571884.3597144>
 - [40] Sruti Srinivasa Ragavan, Zhitao Hou, Yun Wang, Andrew D Gordon, Haidong Zhang, and Dongmei Zhang. 2022. GridBook: Natural Language Formulas for the Spreadsheet Grid. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) (*IUI '22*). Association for Computing Machinery, New York, NY, USA, 345–368. <https://doi.org/10.1145/3490099.3511161>

- [41] Charles Sutton, Timothy Hobson, James Geddes, and Rich Caruana. 2018. Data Diff: Interpretable, Executable Summaries of Changes in Distributions for Data Wrangling. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (London, United Kingdom) (*KDD '18*). Association for Computing Machinery, New York, NY, USA, 2279–2288. <https://doi.org/10.1145/3219819.3220057>
- [42] Priyan Vaithilingam, Elena L. Glassman, Peter Groenwegen, Sumit Gulwani, Austin Z. Henley, Rohan Malpani, David Pugh, Arjun Radhakrishna, Gustavo Soares, Joey Wang, and Aaron Yim. 2023. Towards More Effective AI-Assisted Programming: A Systematic Design Exploration to Improve Visual Studio IntelliCode’s User Experience. (2023), 185–195. <https://doi.org/10.1109/ICSE-SEIP58684.2023.00022>
- [43] Priyan Vaithilingam, Tianyi Zhang, and Elena L. Glassman. 2022. Expectation vs. Experience: Evaluating the Usability of Code Generation Tools Powered by Large Language Models. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (*CHI EA '22*). Association for Computing Machinery, New York, NY, USA, Article 332, 7 pages. <https://doi.org/10.1145/3491101.3519665>
- [44] Bogdan Vasilescu, Alexander Serebrenik, Prem Devanbu, and Vladimir Filkov. 2014. How Social Q&A Sites Are Changing Knowledge Sharing in Open Source Software Communities. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Baltimore, Maryland, USA) (*CSCW '14*). Association for Computing Machinery, New York, NY, USA, 342–354. <https://doi.org/10.1145/2531602.2531659>
- [45] Cong Yan and Yeye He. 2020. Auto-Suggest: Learning-to-Recommend Data Preparation Steps Using Data Science Notebooks. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data* (Portland, OR, USA) (*SIGMOD '20*). Association for Computing Machinery, New York, NY, USA, 1539–1554. <https://doi.org/10.1145/3318464.3389738>
- [46] Pengcheng Yin, Wen-Ding Li, Kefan Xiao, Abhishek Rao, Yeming Wen, Kensen Shi, Joshua Howland, Paige Bailey, Michele Catasta, Henryk Michalewski, Alex Polozov, and Charles Sutton. 2022. Natural Language to Code Generation in Interactive Data Science Notebooks. arXiv:2212.09248 [cs.CL]