

Data Analysis in the Era of Generative AI

Jeevana Priya Inala Chenglong Wang Steven Drucker Gonzalo Ramos Victor Dibia Nathalie Riche
 Dave Brown Dan Marshall Jianfeng Gao

Microsoft Research

Abstract

This paper explores the potential of AI-powered tools to reshape data analysis, focusing on design considerations and challenges. We explore how the emergence of large language and multimodal models offers new opportunities to enhance various stages of data analysis workflow by translating high-level user intentions into executable code, charts, and insights. We then examine human-centered design principles that facilitate intuitive interactions, build user trust, and streamline the AI-assisted analysis workflow across multiple apps. Finally, we discuss the research challenges that impede the development of these AI-based systems such as enhancing model capabilities, evaluating and benchmarking, and understanding end-user needs.

1. Introduction

Data-driven decisions and insights are crucial across a wide range of industries and for everyday end-users. Businesses rely on data to make strategic decisions, healthcare providers track trends to improve patient outcomes, and journalists craft data-based news stories that inform the public. On a more personal level, individuals can use data to manage their finances, monitor their health, optimize their schedules, and even for shopping and making travel plans. However, the cost of current data analysis is high, meaning that only a select group of experts—data analysts—have the tools and skills to ask data-related questions, analyze information, and generate reports/insights based on these insights. For the rest of the population, this process is largely inaccessible; they must accept the insights provided by others, without the ability to ask their own questions or conduct data analysis for personal tasks.

Democratizing data analysis has the potential to revolutionize how we operate. Imagine a world where anyone, not

just trained analysts, could effortlessly analyze data to make informed decisions. A small business owner could independently explore market trends, a patient could analyze their own health data to manage a chronic condition, or a traveler could optimize their trip based on past travel data. This shift would empower individuals and organizations, leading to more informed decisions.

The Challenge of Data Analysis: A Complex and Iterative Process. Deriving insights from data is a complex process. Data analysts need to iterate among task formulation, data collection, exploratory analysis, and creating visual representations to discover and validate data insights before drawing conclusions, and they further need to document the analysis as reports for sharing (Figure 1). To achieve these diverse analysis steps, the analysts need not only conceptual knowledge (e.g., data sense-making, domain, statistics, visualization design knowledge) to make right analysis decisions, they also need tool expertise and programming skills to take actions. Furthermore, because the analysis process is rarely linear and different steps are segregated across various tools (e.g., once the analysts discover missing data based on their current visualizations, they may need to backtrack and revisit data collection or cleaning steps), analysts face considerable overhead switching between tools and managing branching analysis steps.

To address these challenges, many interactive and automated tools and solutions have been developed, both to enhance analysts' ability to understand and explore data and to close the gulf of execution. These include interactive visualization platforms like Tableau and Power BI, automated data preparation platforms like Alteryx or Trifecta, and programming environments like Jupyter Notebooks where data analysis can be integrated with data cleaning and visualization. However, these systems often have to trade-off between the flexibility (and the expressiveness) of tasks that can be achieved with the system and the ease of learning and specifying the tasks. For example, compare Tableau, which provides user-friendly interface and drag-and-drop chart creation, with Python's Matplotlib library, which offers extensive customization and fine control over chart details but requires coding expertise.

. Correspondence to: Jeevana Inala <jinala@microsoft.com>.

The Generative AI (GenAI) revolution. The emergence of large language and multimodal models presents new opportunities to develop tools that are both expressive and user-friendly. These AI models possess the ability to perform human-like reasoning and transform high-level user specifications into low-level executable steps (Bubeck et al., 2023). This reduces the user’s need to learn new tools and languages while assisting with a variety of tasks. Foundation models like ChatGPT, GPT-4 (Achiam et al., 2023), Claude (Anthropic, 2024), Phi-3 (Abdin et al., 2024) have advanced proficiency in language understanding, conversation ability, and planning; they encode knowledge across various fields such as medicine and finance, along with code generation capabilities. Multimodal models like GPT-4o and Phi-3-Vision (Abdin et al., 2024) further augment these abilities by integrating language and vision modalities, enhancing reasoning across diverse input forms. With these abilities, we are in the era where AI-powered tools can empower novice users to perform data analysis tasks previously out of reach and even significantly boost the productivity of experienced data analysts.

Uniqueness of AI-powered tools for data analysis. AI-powered tools are being developed for many domains, including video and image generation, gaming, medicine, and software engineering. So, why does the data analysis domain require special treatment? Unlike generation tasks, data analysis involves planning and exploration over multiple steps. It also requires working with multiple modalities—natural language, structured data, code, images, and charts. Moreover, the task is fairly complicated that even non-AI solutions are spread across multiple apps and tools (for e.g. a user might clean data in Excel, generate charts in PowerBI, and then create a report in PowerPoint). These complexities create interesting challenges for existing Generative AI models and systems, highlighting the need for better models and algorithms. Moreover, solving these tasks may require multiple AI tools and require novel techniques to avoid fragmented and disconnected experience for users.

Data analysis is inherently iterative. Users must be involved throughout the process because the full specification of the task is usually unknown at the beginning. As users see initial results, they may want to follow up with new questions or adjust their goals. This iterative and multimodal nature of data analysis means that current AI tools, which often rely on natural language interfaces and static modalities such as images and file upload, may not be well-suited for this domain. The data analysis domain inspires invention of new interfaces and experiences that allow humans to collaborate with AI tools more effectively.

Moreover, data analysis is a sensitive domain. Errors can have serious consequences, especially in fields like healthcare or finance. Imagine an LLM that incorrectly identifies

a trend in financial data, leading to poor investment decisions, or one that misinterprets medical data, resulting in incorrect diagnoses. This underscores the importance of accuracy and reliability in AI systems, as well as the need for infrastructure that allows end-users to verify and debug the outputs from these AI tools.

Unlocking the potential of GenAI in data analysis. The goal of this paper is to explore how we can unlock the potential of generative AI tools in the data analysis domain. We aim to provide a view into parts of the vast design space of AI-powered data analysis tools that AI and HCI practitioners can use to organize existing AI systems/tools as well as identify areas requiring further attention. Our focus centers on three key themes.

- **GenAI opportunities in data analysis.** We explore how GenAI systems can support each individual stage of the data analysis process. While the utilization of LLMs for code generation in tasks like data cleaning, transformation, and visualization is well-known, we delve into additional avenues such as aiding users in finding required data, ensuring analyses maintain statistical rigor, assisting in hypothesis exploration and refinement, and personalized report generation (see section 4).
- **Human-driven design considerations.** The design of these AI-powered tools and their user interfaces can have a huge impact on user experiences. For instance, some interfaces facilitate a more intuitive approach for users to specify intents such as selecting the color of a chart using a color picker widget vs. using natural language description. Similarly, distinguishing between seemingly correct but analytically different charts poses a challenge for users, especially without additional interface showing provenance analysis. Concretely, we discuss various design considerations for reducing users’ intent specification efforts, enhancing users’ ability to understand and verify system outputs, and streamlining the analysis workflow to reduce iteration overhead across multiple steps and tools (see section 5).
- **Research challenges ahead of us.** Finally, we conclude the paper by discussing the research challenges that must be addressed to implement AI-powered data analysis systems. These challenges include enhancing existing models’ capabilities, addressing the scarcity of training and evaluation data, ensuring system reliability and stability of the data analysis process, and conducting user research to align system designs to meet users’ cognitive abilities and practical needs (see Figure 6).

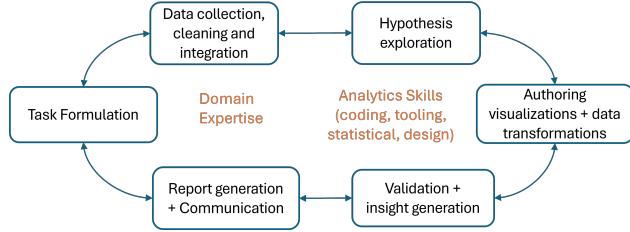


Figure 1: Overview of steps involved in deriving insights from data, highlighting the diverse skills required for these steps. We explicitly exclude any steps related to building/using models in this paper.

2. Background

2.1. The data analysis process

Figure 1 and Table 2 show some of the common steps involved in deriving insights from data. This process has multiple steps and often involves iterating back and forth among these steps. The process usually starts with task formulation, where the problem, question, or decision to be addressed is identified (e.g. What's the trend of renewable energy adoption over the years for different countries?). This step also involves operationalizing the task into sub-tasks (such as figuring out global trends over time, ranking of countries, and compare trends among top countries). Then comes the task of collecting relevant data to answer the above questions. This might involve actions like querying databases, instrumenting applications, web scraping, data cleaning, and integration from multiple sources. The sequence of these steps can vary; analysts may already possess data and seek insights or may identify data gaps or data cleaning needs later in the analysis.

Following data collection, the exploration phase begins, where analysts delve into the dataset to gain familiarity and identify potential patterns or relationships. This involves forming further sub-tasks such as generating descriptive statistics (e.g., mean, median, and count), visualizing the relationship between several variables, trends, and correlations, and generating hypotheses to explore. Then, they transform data or generate visualization artifacts to help support their hypotheses. Once artifacts (transformed data and charts) are generated, one has to understand them and validate them to ensure the reliability and accuracy of their findings. This may involve statistical analysis, domain knowledge-based validation, or comparison with external sources. The insights gathered from these artifacts are then used to refine the hypotheses accordingly.

Finally, analysts communicate their findings and generate the final results. This may involve creating reports, dashboards, or presentations to effectively communicate insights to stakeholders. Analysts may iterate on their communica-

tion based on the feedback received, further refining their presentation of the data and insights.

The data analysis process is inherently complex and multi-faceted, characterized by its iterative nature and the diverse set of skills required at each stage. Each step demands a distinct set of domain knowledge, statistical expertise, coding proficiency, and familiarity with various tools. Moreover, the fragmented nature of the process requires analysts to continuously juggle between different tools and methodologies to derive meaningful insights.

Please note that the steps outlined above are not exhaustive. They do not cover areas such as model building, automated machine learning, collaborative authoring, or accessibility. While these topics are crucial to data science and analysis, they require specialized attention beyond the scope of this discussion. Instead, we focus on the iterative and exploratory aspects of deriving insights from data and how generative AI can influence this process.

2.2. LLMs, LMMs, and agents

The research on LLMs, LMMs, and the AI systems that are built using these models has advanced drastically in recent years. On the base level, we have several foundation models starting with language models (ChatGPT, GPT4 (Achiam et al., 2023), Claude (Anthropic, 2024), and Llama (Touvron et al., 2023)) and then recently multimodal models (GPT-4o (OpenAI, 2024), LLava (Liu et al., 2024a), and Phi-3-Vision (Abdin et al., 2024)). These models are well-known for their ability to generate and understand code and have been shown to have good performance in generating algorithms and solutions to interview/competition problems (Li et al., 2022), write real-world software with GitHub copilot (Nguyen & Nadi, 2022), and various data transformations and visualization authoring tasks (Dibia, 2023; Narayan et al., 2022). These models are also instruction-tuned to understand and converse with people in a chat-like environment (Ouyang et al., 2022).

Using foundation models as a basis, researchers have developed AI agents or AI assistants tailored to different domains. These agents act as interfaces around foundation models, embedding any application-specific knowledge through system prompts or leveraging external tools like code execution engines, search engines, or calculators to respond to user queries (Lu et al., 2024). For example, a code interpreter (OpenAI, 2023) is an AI assistant that can generate code, execute it in a sandbox environment, and use the execution results to produce further text or code.

On top of all of these, there are also multi-agent AI systems that have a series of agents, each supposed to be specialized in a particular subtask, with all agents acting together to collaborate and respond to a user query (Wu et al., 2023b;

Hong et al., 2023). For example, one such multi-agent system is a generator and a evaluator-agent combination that is used to enhance the reliability of AI systems (Du et al., 2023).

To establish terminology, in the rest of the paper we use “AI systems” to refer to systems powered by the latest breakthroughs in LLMs, LMMs and Generative AI, rather than traditional symbolic AI systems. In some places, we also use LLMs and LMMs explicitly based on the context.

3. Case study: User experiences with AI-tools for visualization authoring

Data visualizations are prevalent in different data analysis stages: data analysts create visualizations to assess data quality issues, explore relations between data fields, understand data trends and statistical attributes, and communicate their insights from data to the audience. Although modern visualization tools and libraries have greatly eased the requirements of user expertise and effort, the authoring process remains challenging since the analysts not only need to learn to use visualization tools to implement charts, they also need to transform data into the right format for their chart design.

With the emergence of LLMs, AI-powered visualization authoring tools have been developed to reduce the authoring barrier (Dibia, 2023; Wang et al., 2023a; Maddigan & Susnjak, 2023; Vaithilingam et al., 2024). Here, we use an example visualization task to compare user experiences with (1) traditional programming tools, (2) direct conversational interaction with LLMs via a chat-based interface, and (3) LLM-powered interactive data-analysis tools, highlighting various human-driven design principles which will be discussed later in this paper. We center our comparison around the user experience for (1) specifying the initial intent, (2) consuming the AI system’s output, (3) editing, (4) iterating, (5) verifying AI system’s output, and (6) the workflow in the larger data analysis context.

3.1. Example task and the traditional experience

Table 1 shows a sample dataset that shows the CO₂ and the electricity generated from different sources for 20 countries between the years 2000 and 2020. Let us assume the user is at a stage in the data analysis pipeline where they want to investigate the trend of percentage of electricity from renewable energy sources over time for the five countries with the highest CO₂ emissions.

The traditional (non-AI) approach to this task involves two steps: (a) data transformation, where the user creates new columns (e.g., “percentage of renewable electricity”) and performs operations like grouping, summing, and filtering. More complex tasks may require pivoting, unpivoting, and

Year	Entity	CO ₂ emissions (kt)	Electricity from fossil fuels (TWh)	Electricity from nuclear (TWh)	Electricity from renewables (TWh)
2000	Australia	339450	181.05	0	17.11
2000	Brazil	313670	28.87	4.94	308.77
2000	Canada	514220	155.56	69.16	363.7
2000	China	3346530	1113.3	16.74	225.56
2000	France	373120	50.61	415.16	67.83
2000	Germany	830280	367.22	169.61	35.47
2000	India	937860	475.35	15.77	80.27
...
2020	United States	null	2431.9	789.88	821.4

Table 1: Example dataset: CO₂ and electricity data of 20 countries between 2000 and 2020. Target visualization: Renewable energy percentage over the years for the top 5 CO₂ emission countries. Dataset obtained from (Wang et al., 2024a).

other complex data transformations; and (b) chart authoring, where the user determines the chart type and encodings to appropriately represent the data trends. Data transformations can be performed using various methods, such as formulas in Excel or Python libraries like Pandas within a Jupyter notebook environment. Similarly, charts can be created with libraries such as Seaborn and Matplotlib or in interactive environments like PowerBI and Tableau. While this traditional approach offers full control for users, it presents a steep learning curve for beginner data analysts and non-programming end-users.

3.2. Conversational LLM interface (ChatGPT with GPT-4o and CodeInterpreter)

LLMs have made the above process easier for a wide range of users with different experience levels. For this running example, we use ChatGPT’s interface with GPT-4o model as the conversational LLM together with CodeInterpreter assistance capability. In this conversational experience, the user loads the dataset and expresses their intent as a natural language command, as shown in Figure 2(a). ChatGPT processes the request and performs several steps such as loading and understanding the dataset, figuring out subtasks, writing code for the subtasks, and executing the code to generate the final chart. ChatGPT shows its steps as code snippets, natural language descriptions, and renders the final chart for the user (Figure 2(b)). For this task, ChatGPT requires multiple invocations of the underlying model for the different sub-tasks, leading to long wait times for users. Moreover, repeated model invocations increase the likelihood of failure, and the system can sometimes get stuck on one of the steps, making it difficult to recover.

In conversational LLM interface, users can edit and iterate on their visualizations through follow-up instructions in a conversational manner (see Figure 3 (a)). The chat interface also allows users to backtrack to a previous conversation and edit from there. However, there is no easy way for

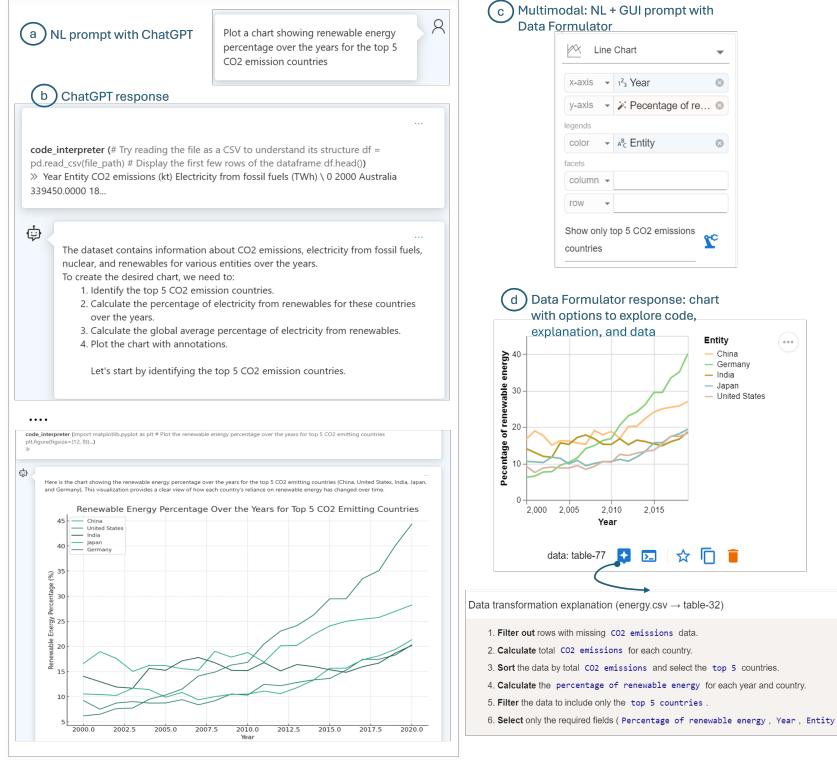


Figure 2: User experience for providing inputs and output formats with ChatGPT vs Data Formulator (Wang et al., 2023a)

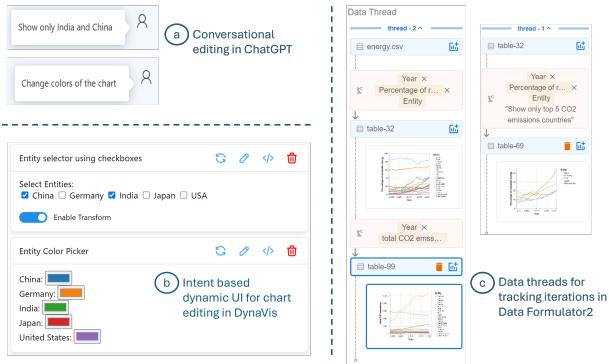


Figure 3: Contrasting conversational editing and iterations in (a) ChatGPT with (b) intent based dynamic UIs for chart editing in DynaVis (Vaithilingam et al., 2024) and (c) data threads in Data Formulator2 (Wang et al., 2024a).

users to track the different iterations. The ChatGPT system might also lose some context (such as an already performed data transformation) while backtracking, leading to the need for re-computation. Since the non-deterministic nature of the underlying model can produce varying outcomes upon re-computation, this again increases the chances of failure and leads to unreliable user experiences.

The latest ChatGPT interfaces and similar systems like Claude (Anthropic, 2024) support multimodal prompts, including natural language, images, and other documents, yet the intention remains primarily mediated through natural language, which can limit expressiveness. For example, some intents, like the exact color preferences in RGB values, are difficult to specify in natural language but easy to select from a color picker. Other commands can be imprecise, such as “put that there” in a complex figure, where users could more accurately specify the context by directly selecting the region. Additionally, users might want to repeat certain actions multiple times with slight variations to see their impact (for e.g. try out different colors before choosing one), but with conversational LMMs, they have to invoke the model for each change, which can again lead to different outcomes each time.

The conversational framework also allows users to ask for further clarification or explanation of how the model per-

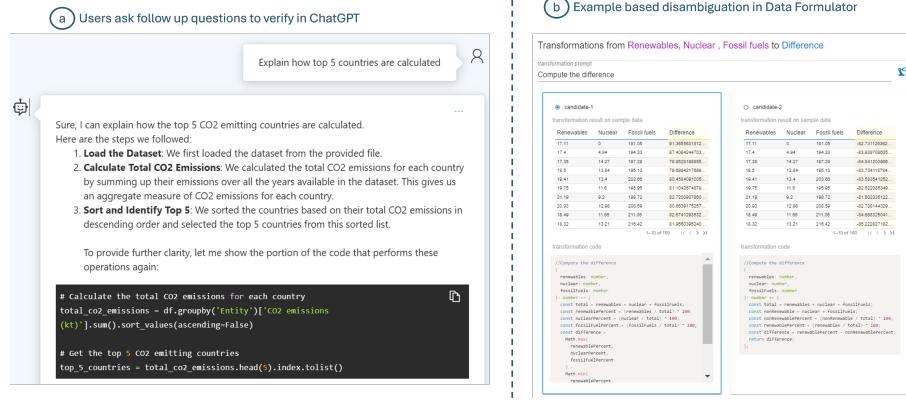


Figure 4: User experience for trust and verification of AI outputs in ChatGPT vs Data Formulator.

formed a certain computation in its analysis ([Figure 4 \(a\)](#)). The system responds back with the explanation as well as code snippets to help users understand and make any changes if needed, allowing users to trust and validate the model’s response.

Other examples of conversational AI interfaces include the copilots or AI assistants for individual tools such as Microsoft’s Office Copilot. However, these copilots function as separate entities. Consequently, if a user needs to perform a task using multiple apps (e.g., using Excel to clean the data and PowerBI to create charts and dashboards), they must manually switch and transport the context between these tools and their respective copilots ([Figure 5 \(a\)](#)), which leads to an inefficient workflow.

3.3. LLM-powered interactive data analysis tools

Now, we describe a different user experience for the above case study using a few selected interactive LLM-powered data analysis tools. Some systems and experiences rely on the capabilities of LLMs to support data analysis, while departing from the types of chat-like experience described above. First, for specifying the user’s intent, Data Formulator (Wang et al., 2023a) has devised a multi-modal UI that allows users to specify the intent using a combination of field-encodings and natural language (NL) instructions (Figure 2(c)). Based on the user’s multi-modal inputs, an LLM model generates the necessary data transformations to create a new table. Using this table and the user’s specified chart encodings, a Vega-Lite specification can then be straightforwardly generated to produce the chart. The chart is then rendered along with options for users to read/edit code, read code explanations, and view the newly transformed data table (Figure 2 (d)). This multi-modal UI allows users to have more control on the type/orientation/high-level details of the chart they want to create but also offload low-level tasks, such as performing the necessary transformations needed for a new concept, to the LLM. From the multi-modal in-

put, the model also has more information to ground the user's instruction for better code generation. For instance, the system can generate a Vega-Lite specification without requiring a separate model invocation. Moreover, this AI system is tailored for data visualization tasks, incorporating components for data processing and summarization, along with domain-specific prompts and instructions for LLMs. These features enhance reliability and reduce latency in the AI-driven data visualization authoring process compared to the experience with the ChatGPT interface.

DynaVis (Vaithilingam et al., 2024) is another interactive AI tool that provides a multi-modal UI for users to specify their chart editing intents. Based on the high-level intent specified in natural language, such as “I want to change the colors of this chart,” the system dynamically generates widgets (e.g., color pickers) on the fly, allowing users to try out several similar options (see Figure 3(b)). The system only calls the LLM when initially generating the widget, not for every change requested by the widget, enabling users to experiment with many variants and receive instant feedback for their edits. These intent-based widgets help users specify intents that are difficult to express in natural language and give users a sense of control and trust by allowing them to easily explore various chart editing options.

Data Formulator2 (Wang et al., 2024a) extends Data Formulator (Wang et al., 2023a) by supporting iterative exploration as a first-class interaction. It allows users to pick any existing visualization and express their modification intent again through the multi-modal UI from Figure 2 (c). The system’s agents reuse previously generated code and data to generate new results, thus giving a consistent experience to users. The system organizes the user’s interaction history as data threads (see Figure 3 (c)) to help the user manage the analysis session. Data threads enable users to easily locate existing plots for refinement, branch out from previous steps to explore alternatives, or backtrack to correct mistakes.

As a form of trust and verification mechanism for users, Data Formulator creates multiple chart/transformation candidates based on the user’s intent. It allows users to inspect these candidates via their code and a few examples (see Figure 4(b)), enabling users to disambiguate or refine their specifications.

Finally, to streamline the data analysis workflow, Data Formulator combines both data transformations and visualization authoring into a single tool, eliminating the need to switch between separate tools for each step. Further streamlining the workflow are multi-app systems such as UFO (Zhang et al., 2024), which is an innovative UI-focused agent tailored for applications on Windows OS using GPT-Vision (Figure 5(b)). UFO employs models to observe and analyze GUI and control information, allowing seamless navigation and operation within and across applications, simplifying complex tasks.

3.4. Remarks

In our comparative analysis of user experiences with various AI tools for visualization authoring, we observe that the design of these tools significantly impacts users’ ability to create the desired visualizations. These design considerations are critical not only for visualization authoring tools but also for AI tools across the different stages of data analysis. In section 5, we elaborate on these design principles and explore various strategies—beyond those examined in this case study—to alleviate user specification burdens, enhance trust and verification strategies, and facilitate workflows across multiple stages and tools.

Before delving into these design principles, section 4 will investigate additional AI opportunities within the broader data analysis landscape, extending beyond visualization authoring.

4. Opportunities for AI systems in the data analysis domain

In this section, we explore how AI systems can help users navigate the complexities of data analysis workflows. Each stage of the analysis process presents unique challenges, from complex reasoning tasks and iterative processes to the need for diverse skills in coding, domain knowledge, and scientific methodologies across multiple tools. To address these challenges, AI systems offer numerous opportunities to streamline and enhance these workflows.

The goal of this section is to establish some concrete problem definitions for a few potential AI-driven workflows in the data-analysis domain (i.e. establishing user’s requirements and expectations from the AI system). While there is an aspiration to automate the entire analysis process, it is imperative to dissect the process into its sub-tasks to identify

areas where AI systems can make a significant impact and where they may fall short. This way of solving a complex problem by breaking into multiple sub-tasks also helps users evaluate the intermediate outputs of the AI systems and to intervene and iterate when necessary, thus helping to ensure the reliability and effectiveness of the overall process.

Once we establish the opportunity and its impact on users, we do a literature review of studies that are related or useful for that particular opportunity. This is not an exhaustive set of opportunities or related works, but rather an preliminary overview aimed at establishing a foundation on which future researchers can build upon.

4.1. Closing the skills gap: Empowering users in data analysis

A data analysis task can require diverse skills including domain knowledge, data science knowledge, coding, as well as knowledge of and how to use tools. The need to possess this diverse skill set often acts as a barrier for non-expert individuals, hindering their ability to engage in effective data analysis and fully realize its benefits. LLMs bring new opportunities to mitigate this barrier by supplementing the existing expertise of users with complementary skills.

Low code/no code experiences. During a data analysis process, once users identify a specific next step, such as visualizing temporal trends, they still face the additional challenge of writing code or using specialized tools to accomplish this objective. For non-programmers, LLMs can alleviate this burden by aiding users in seamlessly transitioning from intent to automatically generating the required code.

LLMs have shown great proficiency in generating code snippets for data analysis from natural language descriptions, especially in languages such as Python. There are several AI systems for taking a user’s specification in natural language or other forms and generating code for various sub-stages such as data cleaning, transformations, querying databases, and visualization authoring.

LIDA (Dibia, 2023) is an open-source tool that enables goal-based data visualization generation by generating, refining, filtering, and executing code to produce visualizations. It also includes an infographer module that creates data-faithful, stylized graphics using image generation models. The system supports multiple Python libraries, including Matplotlib, Seaborn, and Altair, for visualization generation. ChatGPT’s code interpreter API (Cheng et al., 2023) lets users upload their dataset and pose queries in natural language (e.g., analyze the trend of renewable energy over time). Then the model generates a code snippet for this task and executes it to generate a plot. Chat2Vis (Maddigan & Susnjak, 2023) and ChartGPT (Tian et al., 2024) are some

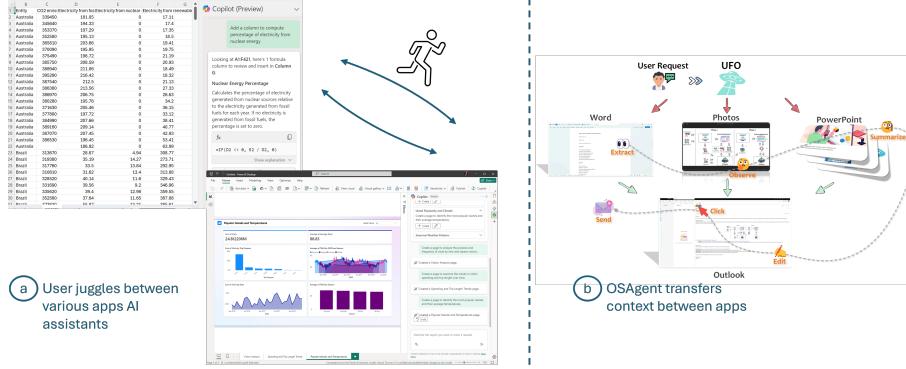


Figure 5: User experience for multi-app workflows with (a) existing AI assistants for individual apps vs (b) an OS-Agent (Zhang et al., 2024) that transfers context between apps. Figure obtained from (Zhang et al., 2024).

Step	Description	GenAI opportunities
Task formulation	Identify the problem to be solved, the question to be answered, or the decision to be made.	Helping people go from fuzzy specifications to concrete, measurable tasks; Adding domain knowledge to task formulation; Finding similar tasks and analyses.
Data collection, cleaning, and integration	Collect relevant structured data and prepare it for analysis.	Finding appropriate data from existing sources; Extracting data from unstructured documents or media; Executing extraction code from relevant databases/web; Automated and semi-automated cleaning or anomaly detection and integrating data from multiple sources.
Hypothesis exploration	Generate hypotheses about data to understand the data.	Domain knowledge based exploratory analysis; appropriate statistical tests based on task and data.
Execution and authoring	Create visualizations, tables, or other forms of structured data representation to support the hypothesis.	Circumventing deep procedural knowledge required for use of libraries and tools (e.g., Pandas, PowerData, etc.).
Validation and insight generation	Read charts and data to validate hypothesis, generation, and assumptions; gather insights for downstream analysis.	Automatic evaluation and repair of analysis; validate analysis and assumptions with domain knowledge; multi-modal analysis to generate insights from data and charts, interpretable results.
Report generation and communication	Make decisions/choose actions and communicate results to stakeholders.	Interactive decision support, automatic recommendations based on analyses; customized representations for different presentations and audiences; dashboard generation; generating creative charts and infographics.
Domain and analytical skills	Need coding, domain knowledge, statistics, and tool skills for different stages of the analysis pipeline.	Low/no code experiences; Offer domain knowledge support and statistical expertise; tool copilots

Table 2: Various analysis steps along with potential opportunities for assistance that GenAI can provide.

of the other LLM-based tools that generate code snippets for data analysis and visualization tasks based on natural language input. DataFormulator (Wang et al., 2023a), which was introduced in subsection 3.3, allows users to specify their visualization intent through drag and drop in addition to natural language and uses these multi-modal contexts to automatically generate the code for visualization. Some of these tools also allow users who may be slightly more proficient in coding skills to inspect the generated code and make any changes if needed.

There are also systems to enable processing of datasets (Sharma et al., 2023), performing data cleaning (Singh et al., 2023; Narayan et al., 2022) and transformations (Joshi et al., 2023) by incorporating domain knowledge in natural language form. Another well-established area of research, known as semantic parsing, involves converting natural language to SQL queries (Li et al., 2024; Floratou et al., 2024) for extracting data from databases.

However, the effectiveness of these systems may differ, highlighting the need to improve reliability and the need to establish evaluation benchmarks and metrics. Additionally, these tools can benefit further from the design considerations we present later, such as enabling natural forms of interactions and fostering user trust in the process.

Statistical assistance. While AI systems for coding components of the data analysis pipeline have received attention, the exploration of these models for other skills, such as statistical proficiency, remains relatively limited.

Here are a few examples where LLMs can offer statistical assistance: LLMs can help users select appropriate statistical tests based on tasks and data, such as t tests to compare the means of two groups. They can also help users avoid common pitfalls (Zuur et al., 2010) such as selection bias, misinterpretation of p-values, and overfitting by providing guidance on issues like multiple comparisons and confounding variables.

AI systems that expose LLM’s statistical reasoning capabilities for the above cases and trigger them at relevant stages during the analysis process hold significant promise. Some relevant prior work in this area comes from the pre-LLM era. Systems such as (Jun et al., 2019; 2022) are steps towards simplifying statistical analysis by allowing users to provide domain knowledge and automatically extracting statistical constraints.

Domain knowledge support. LLMs contain vast amount of knowledge about various domains from their pre-training datasets which can enable us to automate domain-specific contextual understanding of the data and the task (Wadden et al., 2022; Liang et al., 2023a). For example, LLMs can

help users understand the various columns in a dataset, help generate semantically rich features for downstream analysis tasks (Hollmann et al., 2024), and can help users interpret the results (such as finding anomalies (Lengerich et al., 2023)) of the analysis within the domain context.

Tool copilots. Recently, numerous applications feature their own AI copilots or AI assistant systems to facilitate a natural language and chat-based interface for their tools. Examples include BingChat, Office365 copilot, PowerBI copilot, Einstein copilot for Tableau and so on. These copilots notably lower the entry barrier for users engaging in data analysis tasks with their tools. In some instances, these copilots also act as an intelligent tutor that guides users to familiarize themselves with complex software interfaces, offering step-by-step instructions and clarifications on tool functionality.

Currently, these tools are application specific. Each app designs their own API/domain-specific language and uses an LLM/LMM (such as GPT4, Gemini, Llava) to trigger appropriate APIs based on the user’s chat context. The language model’s ability to follow instructions and APIs is a crucial component here. To reduce the burden on copilot developers, AI systems that can automatically understand and utilize existing tools and APIs offer a promising solution, extending AI assistance to a broader range of applications. In subsection 5.3, we also discuss design considerations for enabling these copilot frameworks to communicate with each other to provide a more streamlined experience for users.

In summary, LLMs/LMMs exhibit proficiency in coding, statistical analysis, domain knowledge, and tool usage, opening up opportunities to not only enhance users’ data analysis abilities but also lower the skill barrier of entry to do so.

4.2. Potential AI workflows for the different stages of data analysis process

Task formulation. One of the first steps of the analysis process is to formulate high-level goals for the analysis task and refine them into specific actionable tasks. Often, people face a cold-start problem, where they are unsure of what kind of analysis to perform or what insights they want to gain from their dataset. In such cases, AI systems can assist by leveraging domain-knowledge to understand the dataset and formulate meaningful questions. LIDA’s goal explorer module (Dibia, 2023) addresses this problem by generating high-level goals grounded in data and an optional persona. Each hypothesis is generated as a triplet - goal (e.g., Understand trends in CO_2) a visualization that addresses the hypothesis (e.g., Line chart of CO_2 per year) and a rationale on why the goal is informative.

LLMs can also assist in translating high-level user require-

ments (e.g., evaluating a company’s performance) into specific and actionable tasks (e.g., assessing revenue, employee satisfaction, etc.) again by leveraging the domain knowledge. Some AI systems to refine fuzzy specifications include InsightPilot (Ma et al., 2023) takes a vague user specification and uses an LLM to issue concrete analysis actions that are then given to a dedicated insight engine to explore data and generate insights. NL4DV Toolkit (Narechania et al., 2020) is another tool that generates analytic specifications for data visualization from natural language queries by breaking down an input query into attributes, tasks, and visualizations.

In addition to assisting with goal formulation and refining user tasks, AI systems can further enhance the analysis process by drawing inspiration from existing examples. People sometimes get inspiration from existing charts and data analysis (Bako et al., 2022). Similarly, AI systems can also assist by identifying similar datasets and questions, and their data analysis tasks and use them to guide the user into formulating their own task for their dataset. This draws some similarities to Retrieval Augmented Generation (RAG) based approaches that are commonly used to overcome a model’s limited prompt window size by selecting relevant sources from a vast body of external knowledge (Lewis et al., 2020; Izacard et al., 2022; Peng et al., 2023; Lu et al., 2024) and also few-shot learning approaches with examples chosen based on a similarity metric (Poesia et al., 2022; Li et al., 2023b; Liu et al., 2021).

Hypothesis exploration. Another avenue for AI systems to assist is by reducing the risk of biased data analysis through the support for multiverse analysis, wherein various analytical approaches are systematically examined. This helps users evaluate the strength of conclusions and improves transparency in decision-making by presenting all potential outcomes. Boba (Liu et al., 2020) is an early pre-LLM system to simplify multiverse analysis in data science by letting analysts write shared code with local variations, generating multiple analysis paths.

Automated hypothesis exploration, while highly beneficial, poses significant challenges due to the need to mitigate exponential blow-up of hypothesis space. Hence, it is an interesting research opportunity for AI systems to carefully combine domain knowledge and iterative exploration and at the same time avoid LLM’s own inherent biases.

Execution + Authoring. Given a concrete goal, the next step is to generate the code necessary for any data transformations and for configuring the charts for the visualization. There are numerous works on automating this process with LLMs and we refer to (Dibia, 2023; Wang et al., 2023a; Maddigan & Susnjak, 2023; Shen et al., 2022) for a more thorough exploration of these works.

Validation, insight generation, and refinement. Once data analysis artifacts (charts and new data representations) are created, it is now time to inspect them carefully. This inspection serves two primary purposes. The first is to validate the analysis and assumptions against general expectations derived from domain knowledge. The second is to extract insights that can inform downstream iterative analysis and decision-making processes. Vision language models such as GPT-V and Phi-3-Vision, have shown a lot of potential in analyzing images, charts, and figures. This gives us the opportunity to use these models to reason about the visualizations and generate insights from them.

LLMs and LMMs have been used to evaluate visualizations (or their representations). LIDA (Dibia, 2023) uses an LLM to examine the generated code for a visualization and judge it based on a set of predefined criteria. ChartQA (Masry et al., 2022), PlotQA (Methani et al., 2020) datasets are established to test LMM performance in answering charts-based questions. There are several models (Han et al., 2023; Masry et al., 2023) and approaches (Obeid & Hoque, 2020; Huang et al., 2024b) in the literature for using LLMs and LMMs to understand and gather insights from charts. LLMs have also been instrumental in consolidating insights from various analysis steps to construct a cohesive and comprehensive narrative (Zhao et al., 2021; Lin et al., 2023; Weng et al., 2024).

Finally, these chart understanding capabilities of vision models can be used to automatically iterate and refine hypotheses. For example, for regression analysis, the AI system can fit various curves to the scatter plot of the data, analyze the goodness of fit for each, and try other types of curves based on the initial analysis. This also draws similarity to LLM-based self-repairing for code (Olausson et al., 2023; Le et al., 2023) works where LLMs reason about their own generated code and refine them further.

It is important to note that the user might still play an instrumental role in validation of the AI assistant’s output. Gu et al., (Gu et al., 2024b) found that providing a combination of a high-level explanation of the steps that the AI is taking, the actual code, as well as intermediate data artifacts is important to help resolve ambiguous statements, AI misunderstandings of the data, conceptual errors, or simply erroneous AI generated code. We discuss more on user-verification strategies in subsection 5.2.

Data discovery. In many data analysis workflows that leverage AI, it is often assumed that users have access to all the necessary data. However, ongoing data analysis may reveal the need for additional data to enhance the depth and accuracy of the analysis. For example, an analysis focusing on county-based statistics may require population data to perform proper averaging. Alternatively, some users

may prefer a streamlined experience akin to a search engine, where they can pose direct questions like "which car should I buy?" instead of searching for the required dataset themselves. AI systems present an opportunity to help users locate the necessary datasets in such situations (Fernandez et al., 2023).

Data discovery in a data lake is a well-known problem. LLMs extend beyond traditional keyword/tag-based search methods, offering natural language query-based search in data lakes. By learning to embed popular datasets in a semantically meaningful form, LLMs facilitate dataset retrieval based on user queries. For example, Starmie (Fan et al., 2022) uses pre-trained language models like BERT for semantic-aware dataset discovery from data lakes, focusing on table union search.

Furthermore, AI systems are capable of interacting with various web APIs, such as flight APIs and web search APIs, to dynamically extract data (Liang et al., 2023b). For example, WebGPT (Nakano et al., 2021) can browse the Web to answer long-form questions. There are other AI systems that extract real-world knowledge using a search engine to answer questions (Lu et al., 2024; Schick et al., 2024; Peng et al., 2023). This opens up possibilities for LLMs to gather data from diverse websites based on user queries, organize the collected information into structured tables, and utilize it for reasoning and analysis. Although there have been previous non-AI based web automation tools (Barman et al., 2016; Inala & Singh, 2017), LLM-based approaches offer enhanced capabilities in data extraction and knowledge synthesis.

AI systems can also help with data cleaning and integrating multiple datasets from different sources (Narayan et al., 2022). This can include handling data formatting issues, resolving inconsistencies, and merging datasets to create a unified dataset for analysis.

Another significant opportunity lies in extracting structured data from non-tabular data formats like customer reviews, meeting notes, and video transcripts. LLMs can address this by analyzing the text data to extract structured data such as sentiments and trends for downstream analysis. There are several recent works on using LLMs to convert unstructured data to structured data (Vijayan, 2023), particularly in the medical domain (Goel et al., 2023; Bisercic et al., 2023). Another example is HeyMarvin (HeyMarvin, 2024), a qualitative data analysis platform, that exemplifies this approach by (semi) automating tasks such as transcription and thematic coding of user interviews, facilitating deeper insights into customer needs and preferences. However, there is still room to further improve this process by integrating it into the broader task context, enabling the extraction of necessary features based on the specific requirements of the data analysis task.

Personalized reports, sharing, and creativity. One of the final steps of the data analysis workflow is to create dashboards, reports, and presentations to communicate the results and insights to stakeholders. Current tools like PowerBI, Tableau, and Powerpoint require significant proficiency and development effort. Some ways AI systems can assist here are by facilitating the generation of dashboards and what-if analyses with minimal developer involvement (Pandey et al., 2022; Wu et al., 2023a; Sultanum et al., 2021). AI systems also bring the opportunity to transform static elements, such as papers and documents, into interactive versions (Dragicevic et al., 2019; Heer et al., 2023). Another example is Microsoft Office copilots that answer questions on the side about the document in an interactive manner.

Furthermore, AI systems, particularly LLMs, can tailor the content and format of reports to suit various audiences and devices by dynamically adjusting tone, detail level, and emphasis. Another aspect in which LLMs and other types of GenAI models such as DALL-E and Stable Diffusion can help is by creating artistic and aesthetically pleasing charts and infographics (Dibia, 2023; Schetinger et al., 2023). By interpreting the underlying patterns and insights in a dataset, GenAI models can craft visually representative charts that not only communicate information effectively but also evoke a sense of creativity and design. For example, for a chart regarding global warming trends, we can create an infographic with a melting ice metaphor to convey the urgency and severity of climate change.

In general, creating interactive and personalized artifacts for communication is challenging, as it involves linking various forms of media such as static text, images, videos, and animations, as well as employing diverse interaction techniques such as details-on-demand, drag-and-drop functionality, scroll-based interactions, hover effects, responsive design, and gamification to enrich communication (Hohman et al., 2020). The vast space of potential interactive designs presents an interesting problem for AI-based systems, which could generate these artifacts by understanding user preferences and intentions and executing them accordingly.

5. Human-driven design considerations for AI-based data analysis systems

In the previous section, we explored how AI systems can impact various workflows within the data analysis pipeline. However, the user experience of these AI-powered workflows varies significantly based on several design considerations. As Hutchins et al., (Hutchins et al., 1985) highlighted, two main challenges users must overcome when interacting with technology are (a) Execution: taking action to accomplish a particular goal, and (b) Evaluation: understanding the state of the system. While GenAI alters the notions of

execution (for e.g., specifying intentions instead of writing visualization code) and evaluation (for e.g., understanding AI-generated code in addition to reading charts), the gulfs of execution and evaluation still exist. Therefore, it is crucial to design AI systems with users in mind, aiming to reduce these new gulfs. Figure 6 presents an overview of potential human-driven design considerations, as well as outlines the research challenges that need to be addressed to implement these designs, which will be discussed in more detail in section 6.

5.1. Enhancing user-AI interaction: For natural intent communication

Relying solely on natural language based interfaces can pose challenges for users of AI-based data analysis systems, as expressing complex intents can become lengthy or difficult due to limited expressiveness. For instance, altering the legend placement in a chart requires users to acquire the literacy to talk about charts (in this case, learn the word “legend”) so that an AI agent would unambiguously understand. Similarly, for complex tasks, users might find it necessary/beneficial to interact with intermediate artifacts, explore variations, and integrate them gradually, rather than specifying everything at once. Below we discuss some potential ways AI systems can enable natural modes¹ for users to communicate their intent.

Multi-modal inputs. Beyond chat and natural language interaction, there are other modalities like mouse-based operations, pen and touch, GUI based interactions, audio, and gestures that could provide a more powerful and easier way for users to provide their intents.

For example, Data Formulator (Wang et al., 2023a) is an AI system that supports multi-modal inputs. Data Formulator combines graphical widgets and natural language inputs to let users specify their intent more clearly with less overhead compared to using just natural language interface (Figure 2(c)). Data Formulator provides a shelf-configuration user interface (Grammel et al., 2013) for the user to specify visual encodings, and the user only needs to provide short descriptions to augment UI inputs to explain chart semantics. In Data Formulator, the shelf is pre-populated based on the current columns in the dataset, but one could also imagine auto-generating additional concepts on the fly based on the task and context and let users choose the right concepts rather than fully specifying them in natural language. DirectGPT (Masson et al., 2024) is another system that allows users to specify parts of the intent by clicking, highlighting, and manipulating its canvas. This type of di-

¹by natural modes, we mean straightforward (possibly personalized) modes that do not require users to learn new communication skills

rect manipulation has been previously explored for image generation (Dang et al., 2022). InternGPT (Liu et al., 2023d) combines chatbots like ChatGPT with non-verbal instructions like pointing movements that enable users to directly manipulate images or videos on the screen.

Multimodality can also surface as demonstrations from the users. Pure demonstration-based inputs have been explored before (Barman et al., 2016), but combining demonstrations with natural language can be complex. For example, ONYX (Ruoff et al., 2023) is an intelligent agent that interactively learns to interpret new natural language utterances by combining natural language programming and programming-by-demonstration. Sketching or inking is another way for users to communicate their intent. Especially for designing charts, users often have a clear idea of the chart design they want, which can be challenging to specify using natural language alone (Lee et al., 2013).

Another form of user specification can be in the form of input-output examples. For instance, programming by example is a very common strategy to specify data transformations like pivoting (Gulwani, 2016). However, traditional non-AI approaches cannot handle noise (or manual errors) in the provided examples, which LLMs can potentially alleviate.

Audio input can be viewed as another representation of chat, but it is more powerful when combined by demonstrations or other user manipulations by associating specific utterances to what the user was doing at that point of time.

Supporting multimodal inputs requires careful system design. For example, we need to be able to automatically map user’s interaction on a chart/UI visually to the appropriate underlying component (You et al., 2024; Liu et al., 2018). We might even require entities (like charts) to be modular and interactive to enable actions like drag-and-drop. Dealing with continuous inputs like audio and real-time screen activity is also challenging for current models. However, recent model innovations such as GPT-4o (OpenAI, 2024), which support audio, video, and text input forms, are promising. Some models such as InternGPT (Liu et al., 2023d) also finetune large vision-language models for high-quality multi-modal dialogue.

Multi-step interactions. Human-AI interaction is often a multi-step process. Initial user specifications can be ambiguous, requiring adjustments based on outputs, or the tasks themselves might be multi-step, driven by feedback from intermediate results, such as in data exploration. Thus, enabling users to interact and iterate during the analysis process is crucial.

Some linear forms of interactions include chat based interaction, that aggregates context from previous conversa-

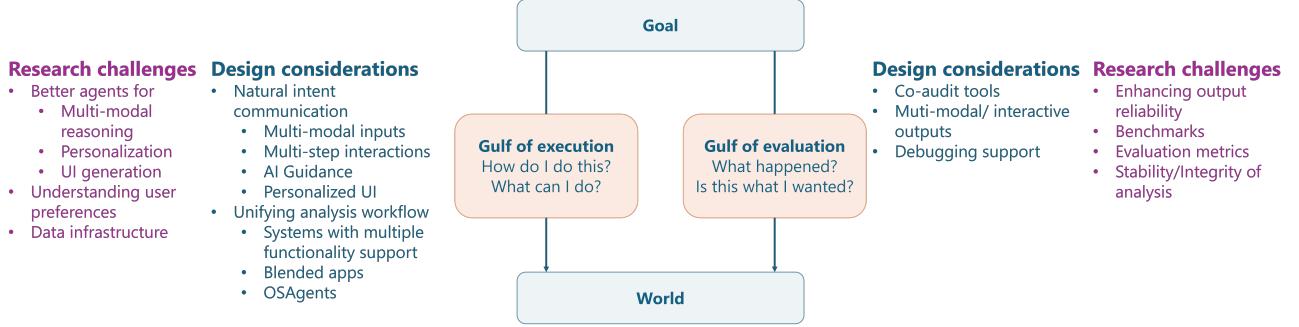


Figure 6: Human-driven design principles and research challenges based on the gulf of execution and gulf of evaluation principles from (Hutchins et al., 1985).

tions (Wang et al., 2023c), and computation notebook style, that enables users to inspect intermediate outputs. However, sometimes non-linear forms are needed, such as generating multiple outputs (charts, designs, etc.) in the first step and allowing users to choose or combine them to proceed further. Users might also need to backtrack and fork again. Threads and tree like representations can be used for non-linear interactions (Wang et al., 2024a; Hong & Crisan, 2023). For example, Data Formulator2 (Wang et al., 2024a) organizes the user’s interaction history as data threads to help the user manage analysis sessions (Figure 3(c)). With data threads, the user can easily locate an existing plot to refine, fork a new branch from a previous step to explore alternatives, or backtrack to a previous state to correct mistakes. AI Threads (Hong & Crisan, 2023) is another system that facilitates multi-threaded conversations, enhancing a chatbot’s contextual focus and improving conversational coherence. Sensescape (Suh et al., 2023) provides multi-level abstraction and seamlessly switches between foraging and sensemaking modes. Graphologue (Jiang et al., 2023), an interactive system that converts text-based responses from LLMs into graphical diagrams to facilitate followup information-seeking and question-answering tasks.

Tracking an AI agent’s state and users’ specifications along a particular path in the interaction tree is crucial for drafting the appropriate context for the models. Prompt summarization (such as (Wang et al., 2023b)) is a potential technique to make sure that each node in the tree is stateless and appropriately captures the full context that led to that particular node. Along with this, it is also necessary to have an appropriate user interface to enable users to interact in a potentially non-linear fashion with the AI agents.

AI guidance vs user control. Another important aspect of user-AI interaction design is whether users start with a blank slate, warm start with AI-driven recommendations, or use a mixed-initiative approach where both the AI system and the user take the initiative as appropriate. Starting with a blank state empowers users with maximum flexibility and

control over their tasks, whereas AI-driven recommendations offer proactive guidance and suggestions, especially when users are unsure of what to do next or unsure of what the AI system can do for them. For example, Lida (Dibia, 2023) exemplifies this approach by presenting users with a list of potential goals upon loading a dataset as well as recommendations of relevant visualizations once an initial visualization has been created. Similarly, various copilots and tool assistants offer prompt suggestions to guide user interactions. Some AI assistant systems can also prepopulate entire documents or slides based on notes or other user documents, providing users with a starting point that they can edit and refine through follow-up interactions.

GenAI models are increasingly gaining the ability to analyze datasets, user history, and the behaviors of other users to predict and recommend next steps, or even entire workflows. This shifts user interaction from building components from scratch to refining AI-generated suggestions. For example, after a model generates a slide deck, users might probe the system to understand/modify the data that was used to create the charts in the slide. This shift underscores the need for AI systems to support probing mechanisms, offering interfaces that allow users to ask follow-up questions, intervene and modify specific parts, and provide clear explanations of the AI’s processes and decisions.

Personalized and dynamic UI. The language model’s generative capabilities present an opportunity to generate customized UIs on the fly. DynaVis (Vaithilingam et al., 2024) is one such system that showcases LLM’s capacity to dynamically generate UI widgets for the subtask of visualization editing. As the user describes an editing task in natural language, such as “change the legend position to left”, DynaVis performs the edit and synthesizes a persistent widget that the user can interact with to make further modifications (to try other legend positions and find the one that best fits their needs) (Figure 3(b)). Stylette (Kim et al., 2022) is another system that allows users to specify design goals in natural language and uses an LLM to infer relevant

CSS properties. PromptInfuser (Petridis et al., 2023) explored using prompting in creating functional LLM-based UI prototypes. LLMs have also been used to generate feedback on UI designs (Duan et al., 2023). Expanding on the above techniques to empower LLMs in generating personalized user interfaces for data analysis tools holds significant promise for enhancing user experience with tools in the future.

5.2. Facilitating trust and verification: Enhancing model output reliability for users

GenAI models can produce unintended outputs either because of model hallucinations or the ambiguity of the user input. This makes it essential for users to verify the accuracy of the models' outputs. However, the verification process should not impose a cognitive load comparable to or greater than that of a person performing the task by themselves. For example, verifying data transformations may require more than visual inspection, requiring code review or editing. Hence, it is important for AI system designers to also facilitate easier ways to allow users to verify the outputs. This aligns with the argument that GenAI systems impose significant metacognitive demands on users, as highlighted in (Tankelevitch et al., 2024), where the authors propose integrating metacognitive support strategies into GenAI systems to reduce these demands, particularly by enhancing explainability and customizability.

Co-Audit tools. There are several ways in which users can verify the quality of a model's output, such as gauging the high-level correctness from charts and transformed data, inspecting corner cases in data, understanding code, and conducting provenance analysis for selected data points (Gu et al., 2024b). AI agents can be designed to support these types of quality control, as highlighted by the emerging concept of co-audit tools (Gordon et al., 2023), which aim to assist users in checking AI-generated content.

One of the co-audit approaches is to provide code explanations either in natural language (Wang et al., 2023a; Cui et al., 2022) or ground the output code in editable step-by-step natural language utterances (Liu et al., 2023a). Another verification approach is to generate multiple possible charts/data for the same user instruction and show any differences to the user to disambiguate (Mündler et al., 2023; Lahiri et al., 2022) (Figure 4(b)). Inspection tools, such as ColDeco (Ferdowsi et al., 2023), provide spreadsheet interfaces to inspect and verify calculated columns without requiring users to view the underlying code. Provenance analysis is another critical aspect that tracks how a particular output (for e.g. a single data point) is obtained through a series of transformations from the input. (Buneman et al., 2001) provides one of the early characterization of data provenance especially for data obtained through

SQL queries. There are several recent systems such as XNLI (Feng et al., 2023), which shows the detailed visual transformation process to the user along with a suite of interactive widgets to support error adjustments and a mechanism for the user to provide hints for revision in the data visualization domain. Datamations (Pu et al., 2021) is another tool that shows animated explanation of the various steps that led to a particular plot or table.

Multi-modal and interactive outputs. Additionally, the format of the output has a significant influence on the user's understanding and verification ability. For example, a long text describing the analysis results and insights can be difficult for users to parse, and lengthy code can be challenging for users to understand. But a multimodal document interlaced with text and images will be more readable. Similarly, interactive charts and what-if analyses allow users to quickly experiment with different parameters, helping them verify outcomes more effectively than by reading through extensive code. Dynamic user interfaces customized for the user and the task in the style of DynaVis (Vaithilingam et al., 2024) can also help by providing UI and widgets on the fly to help users interactively probe and refine AI generated outputs.

Debugging support for users. AI systems, which often consist of multiple models, agents, and components like retrieval-augmented generation (RAG) or self-repair mechanisms, can be challenging for end users to debug or understand. Users typically see only the final outcome without insight into the internal workings of the system, making it difficult to grasp why the AI system is behaving in a certain way, producing specific outputs, or failing to complete tasks.

To improve trust, one potential strategy is to design these systems to provide real-time visibility into the actions of these agents, along with support for checkpointing and interruptibility. This would allow users to stop the process at any given point, provide feedback, and then resume. Recent work, such as AutoGen Studio (Dibia et al., 2024) which supports building of AI systems by orchestrating multiple agents to collaborate, addresses this challenge by providing a no-code interface tool for authoring, debugging, and deploying multi-agent workflows. In particular, AutoGen Studio's debugging view visualizes the messages exchanged between agents using charts and displays key metrics such as costs (tokens used by an LLM, dollar costs, number of agent turns, and duration), tool usage, and tool execution status (e.g., failed code execution).

5.3. Unified analysis experience: Streamlining data tools and workflows

While most current data analysis tools focus on automating one or a few steps of the data analysis pipeline, upcoming

innovations in AI models and complex multi-agent AI systems have the potential to change the entire end-to-end data analysis stack. In the current workflow, users must navigate multiple tools like Excel, PowerBI, Jupyter Notebooks, and SQL servers because each serves a unique purpose and offers interfaces tailored to different aspects of data analysis. Even with GenAI techniques integrated into these tools, significant overhead remains in switching contexts, transferring data, and carrying over intent from one application to another. Therefore, another key design consideration for AI-powered data analysis tools is finding ways to streamline this process for users.

One tool multiple capabilities. Unifying multiple capabilities into a single tool is one way to streamline the analysis process by allowing users to work in a single environment, eliminating the need to re-explain domain contexts and provide feedback across multiple platforms. It also enables the unified AI system to seamlessly debug and reason across the various steps of the process. For example, if a visualization appears inaccurate due to improper data cleaning, the unified AI system can identify and address issues, without user intervention to identify and debug the problem.

Some of the AI-powered tools we have mentioned so far try to combine more than one data analysis step into a single tool. For example, Data Formulator (Wang et al., 2023a) unifies data transformation with visualization authoring, which improves the user experience over traditional visualization tools (e.g., PowerBI, Tableau) that need users to perform a separate step to transform their data to tidy formats as needed by the respective tools. LIDA (Dibia, 2023) is another tool that combines data summarization, goal exploration, chart authoring, and infographics generation into one tool.

However, it may not always be feasible to design a single tool that supports all stages of data analysis and accommodates the diverse preferences of users, which motivates alternate approaches, such as blended apps and OS Agents, which are discussed below.

Multi-agent systems. A key advancement that enables the integration of multiple capabilities within a single tool is the development of multi-agent systems (Wu et al., 2023b; Hong et al., 2023). These systems allow us to create AI systems composed of multiple agents, each specialized in a particular capability, and configure the agents to interact with one another to plan and collaborate on complex tasks. They have shown promise in producing complex AI tools in various domains such as scientific reasoning (Lu et al., 2024), software engineering (Hong et al., 2023; Qian et al., 2023), and embodied agents (Gong et al., 2023). Closest to the data analysis domain is the software engineering domain, where systems such as ChatDev (Qian et al., 2023)

are modeled as a system of AI “software engineers” such as programmers, code reviewers, designers, and testers who collaborate to facilitate a seamless software engineering workflow. They show that this kind of architecture in conjunction with high quality models is effective in identifying and alleviating potential software bugs, and also at reducing hallucinations of the LLMs.

Blended apps. Modern applications such as Microsoft Loop and Notion are starting to create all-in-one workbenches, where one brings multiple tools/data that are needed for the task into a single interface. In the context of building AI tools for the data analysis domain, a similar approach would be to design AI tools that can be contextualized in the overall analysis process. Individual AI systems for one part of the analysis workflow should be able to talk to AI systems for the other parts, share their contexts, and collectively solve/debug tasks. Compared to a single AI system handling multiple capabilities, this approach allows developers to build AI systems for different capabilities separately, each with its own customized user interface and verification strategies, while still providing users with a unified experience. Additionally, the user experience can be further improved by AI systems talking to and extracting context from existing non-AI systems. For example, going back to the Data Formulator tool, even though it helps users automate certain data transformations like table pivoting, it does not allow users to perform manual cleaning of the data; however, users can easily manually edit data in tools like Microsoft Excel. Therefore, a blended app experience that transitions automatically and ports any necessary context from one tool (AI or non-AI based) to another can enhance the user experience drastically.

OS Agents. Another notable way to reduce tool overhead is by using “OS Agents” or a “Self-driving OS” that automatically controls and interacts with multiple applications on the desktop. For example, UFO (Zhang et al., 2024) is a GPT vision-based agent that observes and analyzes the GUI and control information on Windows applications and performs tasks across multiple applications based on natural language commands from the user (Figure 5(b)). For applications that cannot be controlled via their APIs or co-pilots, this way of direct UI manipulation with AI agents is a promising means to enable a unified analysis experience for end-users.

6. Challenges in developing AI powered data analysis systems

Now that we have explored various design considerations to enhance human-AI experiences in data analysis tools, several research challenges remain in developing AI-powered systems that effectively incorporate these design principles.

As outlined in Figure 6, these include developing new models to support multimodality, planning, and personalization; understanding user preferences; creating data infrastructure to facilitate AI-driven suggestions; enhancing the reliability of existing AI models; and establishing robust system benchmarking and evaluation metrics. While some of these challenges are common across any LLM-based interactive system (Kaddour et al., 2023), this section particularly contextualizes the problems to the data analysis domain.

6.1. Ensuring reliability and trust

AI models are known to suffer from several issues such as hallucinations (Tonmoy et al., 2024), sensitivity to prompts (Sclar et al., 2023), failure to follow instructions (Liu et al., 2024b), lack of acknowledgment of uncertainty (Key et al., 2022), and biases (Gallegos et al., 2023). This is true even for the data analysis domain (Liu et al., 2023a; Singha et al., 2023). For e.g. (Singha et al., 2023) shows how LLMs’ performance is susceptible to table representation in the prompts. Thus, given the potential consequences of bad results, it is crucial to improve the reliability of AI-based data analysis systems.

Improving code correctness of LLMs. The first step is to ensure that LLM’s output satisfies the specification, such as using the correct API and being syntax error free. Grounding techniques, such as RAG (Retrieval-Augmented Generation) and GPT-4 Assistants API, enable LLMs to ground their outputs in external knowledge such as API documentation and relevant examples, as well as use tools such as calculator and custom functions to reduce errors in the output (Zhang et al., 2023a; Patil et al., 2023; Peng et al., 2023). For code generation parts of the data analysis pipeline, we can increase the accuracy of the models by verifying the code against a suite of input-output examples (which are either provided by the user (Wang et al., 2023a; Wen et al., 2024) or automatically generated but verified by the user (Lahiri et al., 2022)). It has also been shown that self-repair (Gero et al., 2023; Olausson et al., 2023) and self-rank (Inala et al., 2022) mechanisms that let models evaluate their own outputs, and even repair them, have been very beneficial in improving the accuracy of the code generated by the models.

Handling failure cases. Sometimes, despite the above efforts, models might still produce undesirable outcomes (such as a code that crashes). Some strategies for handling such failures include providing fallback options or having the model intelligently request additional information from the user. Consider generating UIs on the fly using LLMs, where the robustness of the output is crucial to maintain a seamless user experience. In addition to having mechanisms for error identification and correction, incorporating fallback

options such as static UIs or dynamically composing from static UI elements can mitigate disruptions caused by the errors.

Ensuring Stability and Integrity of the Analysis. Many factors, such as ambiguous user intents, faulty data, model biases, and hallucinations, can cause an AI system to produce unintended data analysis outputs. In the analysis process, LLMs may be used for hypothesis generation, making assumptions about the dataset or domain, or generating statistical reasoning. Therefore, it is essential to ensure that these assumptions and hypotheses are not flawed due to the above factors. Therefore, AI system developers should assess the stability and integrity of the AI system’s outputs before presenting them to users.

To identify hallucinations and ambiguous user intents, one approach involves sampling multiple outputs from the LLMs and examining their agreement to detect inconsistencies. The predictability, computability, and stability (PCS) framework (Yu, 2020) offers another structured method for evaluating the trustworthiness of results by ensuring that they are predictable, stable, and aligned with real-world contexts. A key aspect of this framework is demonstrating that the analysis outcome remains consistent when slight perturbations are applied to inputs, assumptions, and hypotheses at various stages of the data analysis pipeline.

Another threat to AI-driven analysis is the possibility of overlooking crucial considerations during hypothesis exploration, even if the analysis is stable. For example, when selecting a store with strong sales performance, the analysis might focus solely on total profit but miss important factors such as the store’s profitability relative to its size or operational costs, which could lead to misleading conclusions. One potential approach to identify such failures is to actively probe the model to consider alternative analyses (for e.g., for the above scenario, the AI system can actively probe the models to output an analysis that makes a different store than what was predicted before to appear as the strong performer). After generating multiple possible analyses, the system can reason over them to determine which makes the most sense given the user’s intents, the task, and the domain.

Overall, ensuring the integrity of the entire AI-driven analysis workflow is significantly more challenging than ensuring the reliability of individual AI steps, such as code generation for charts. This requires developing new frameworks, such as the PCS framework, and incorporating active probing, mutating, and testing strategies as integral components within the AI-data analysis system.

6.2. System benchmarking and evaluation metrics

Desired evaluation benchmarks. To enable researchers to prototype and evaluate data analysis models and systems

effectively, we need a comprehensive benchmark suite that covers a broad range of data sources and tasks within the data analysis domain. This suite should include both low-level tasks for each step of the data analysis process and high-level tasks that span multiple steps. Other domains, such as AI for Code, already have rich and rapidly growing benchmark sets (e.g., HumanEval (Chen et al., 2021), CodeContests (Li et al., 2022), and cross file code completion tasks (Ding et al., 2024)) that helped evaluate new techniques, such as self-repair (Olausson et al., 2023) and test-driven generation (Lahiri et al., 2022). The goal is to create a similar resource in the data analysis domain.

While the data analysis domain already has several benchmark collections, they often focus on specific steps of the data analysis pipeline. For example, existing benchmarks address visualization authoring (Luo et al., 2021; Srinivasan et al., 2021), data transformations (DS1000 (Lai et al., 2023), Juice (Agashe et al., 2019), CoCoNote (Huang et al., 2024a)), machine learning (MLAgentBench (Huang et al., 2024c)), and statistical reasoning (Blade (Gu et al., 2024c), QRData (Liu et al., 2024c)). DABench (Wang et al., 2024c), another benchmark, tests LLMs' abilities to write and execute complex code for tasks such as outlier detection and correlation analysis, which are essential for hypothesis evaluation. These benchmarks are very valuable, but they are spread across multiple benchmark suites, each with different evaluation metrics and procedures devised by individual research groups. Moreover, they often use data sources and tasks of varying difficulty levels, making it challenging to compare AI systems across different suites. Additionally, most of these benchmarks focus on the code generation and the code execution aspect of the data analysis pipeline and have tasks with unambiguous intents and single word, easily evaluated answers, which limits their ability to test all aspects of AI systems in the data analysis domain.

Moreover, benchmarks are often created to highlight specific model capabilities or techniques, typically tailored to the strengths of existing models. This approach can result in narrow benchmarks that do not fully represent the complexities found in real-world data analysis. Therefore, more research is needed to develop benchmarks that reflect the diverse challenges encountered by data analysts, such as tasks requiring multiple iterations of analysis, gathering additional data based on previous results, fixing data issues before creating visualizations, and exploring hypotheses. These domain-oriented benchmarks can also reveal the limitations of current models in areas such as multi-modal reasoning, multi-step planning, and exploration versus generation capabilities of models, driving AI researchers to push the boundaries of their models and techniques in new directions. Therefore, it would be useful to compile a diverse set of data tables and sources into a centralized location, along with potential high-level questions or goals (e.g., decision-making,

report creation) and low-level concrete tasks for each step of the data analysis process. A taxonomy of tasks, categorized by complexity for both humans and models (e.g., tasks requiring multimodal reasoning, planning, or exploration), would help researchers assess test coverage, understand the strengths and limitations of AI systems, and identify areas for improvement.

There are other interesting kinds of benchmarks that are still lacking in the community. These include benchmark tasks that involve potential human intervention, where the full intent isn't known at the outset and evolves based on what the model generates in earlier steps. We also need benchmarks that require models to engage in open-ended exploration and iterative reasoning, similar to human scientific reasoning. Additionally, it would be beneficial to develop session-based tasks that capture a user's activity during an entire session or across multiple sessions. These tasks could involve scenarios that would demand multi-modal and dynamic UIs, such as when users need to specify something difficult to articulate in natural language, repeat certain actions, or analyze multiple datasets in a single session. Such benchmarks would also help evaluate an AI system's ability to personalize to user preferences over time, track context based on past actions, and reuse contexts from previous sessions.

Evaluation metric challenges. Benchmarks should be accompanied by robust evaluation metrics. While it is valuable to evaluate a system in real usage with users in their own workflows with user studies and telemetry data, this process is often expensive and hence, are only used after promising results are observed in controlled offline environment. These Offline metrics are useful for fast prototyping because they do not require real time user interaction.

For general purpose coding tasks, there are some established ways to measure accuracy of the output by adding a suite of test cases (input and expected output) for each task and executing the generated code on these test cases to check the outputs match the expected output in the test case. However, for the data analysis domain, the generated code executes to produce complex artifacts such as charts or UIs which are hard to compare because they are subjective. Moreover the design principle of supporting multi-modal forms of communication (i.e., natural language, gestures, etc.) and interactive and iterative human-AI collaboration, makes offline evaluation even more challenging. To address these challenges, there are some preliminary techniques for using a model such as GPT4, GPT-Vision to simulate a human in the loop to both judge and evaluate multi-modal artifacts (Liu et al., 2023b) and to provide intermediate inputs and interact with GUIs by generating web inputs (Liu et al., 2023c).

Another key consideration in designing evaluation metrics

is the ability to assess partial correctness of AI systems. For example, an AI system might successfully explore multiple hypotheses but make a small mistake in one, leading to cascading effects on the final output. Additionally, evaluation metrics should account for multiple orthogonal dimensions of performance. For example, an AI system might excel at task breakdown, general code generation but struggle using particular visualization APIs or creating intuitive designs, or vice versa.

Behavioral evaluation. In addition to assessing the correctness of AI-generated outputs for data analysis tasks—i.e., whether the output matches the expected result—it is equally important to evaluate whether the AI system’s output fosters user trust. For instance, when asked which car is the best to buy, a system that provides an answer along with explanations of why the car is better across multiple dimensions, and details how it performed calculations as part of the analysis, is more trustworthy than a system that simply gives an answer, even if the answer is the same. Similarly, it’s important to evaluate not only whether the model can, on average, produce the correct answer (such as the $\text{pass}@k$ metric, which assesses if the model provides a correct answer within k attempts), but also measure worst-case metrics such as how frequently the model hallucinates, as wrong answers have potentially bad consequences. Lastly, it is crucial to include tasks and evaluation metrics that explicitly perform adversarial testing of the AI system. (for e.g. variations of prompts that can trigger a misleading but convincing analyses), similar to the adversarial robustness experiments conducted for NLP tasks in (Raina et al., 2024; Kumar et al., 2023).

6.3. Need more advances in models/agents

We need further advances in models and agents to fully realize multi-modal, iterative, and trustworthy AI systems for the data analysis domain.

Inference cost. Firstly, while LLMs like GPT-4 offer powerful capabilities, they come with high inference costs. Conversely, smaller models such as Llama and Phi-3 may not consistently produce reliable outputs. Hence, there is a need for advancements in smaller language models to strike a balance between efficiency and accuracy.

Training data. One of the challenges of using LLMs for data analysis is that many of the latest models are not specifically trained on data-related or data-analysis-specific tasks. While there may be adequate data coverage for programming languages such as Python or JavaScript, there is a lack of training data for other languages used in data analysis, such as R or VBA scripts. Additionally, although LLMs are potentially trained on web content and UI designs, they lack

sufficient training data on UI interactions, such as capturing what happens when a user clicks a specific button in applications like Excel. This kind of data is crucial for LLMs to generate personalized UIs to help user better interact with the AI system. Another challenge arises from the difficulty of finding training data that maps high-level user intents to low-level data analysis tasks, since the underlying human intent (such as why a particular analysis is performed) is often not apparent in the final artifacts—such as Jupyter notebook code—which are scraped from the web for training models.

Finetuning; RLAIF; Multi-agent finetuning. Finetuning the AI models for specific data-analysis capabilities on a smaller dataset is one possible solution to overcome the lack of training data problem. Avenues for finetuning in this domain include finetuning pre-trained LLMs to understand the semantics of the tables/data (Li et al., 2023a; Zhang et al., 2023b; Dong et al., 2022), finetuning to learn to generate domain-specific code/actions for which there is not much training data, and finetuning multi-agent systems over the entire data-analysis workflow. A significant challenge in fine-tuning is the scarcity of ground truth data. For multi-agent workflows, for instance, there is no supervised data to guide the intermediate actions of each agent. Similarly, in newer code domains or user-interaction-based tasks, it is difficult to gather annotated data manually. Hence, automated ways of gathering ground truth data are useful. Here, we can leverage the same techniques that are needed for automatic offline-evaluation of AI-data analysis systems; utilizing current AI models to generate multiple candidate outputs and then employing offline evaluation to label these candidates. This technique is called RLAIF (Lee et al., 2023) as it involves using AI feedback, such as simulated human agents, to enhance models (especially to improve smaller models using feedback from larger models).

Personalization and continual learning Personalizing AI system’s outputs is crucial for enhancing user experience. As users are likely to engage with AI-based data analysis systems across multiple sessions, AI agents need to learn from previous interactions to understand user preferences. Advancements in agents and AI models are needed to enable memory storage (Wang et al., 2024b), learning from past actions and human feedback, and adapting to user preferences over time (for example, the amount of code they would like to see, colors they would like in their plots, their reporting style, etc.). Additionally, AI agents should be able to leverage data from other users, offer tailored recommendations, and predict users’ next steps similar to current recommendation systems used by streaming platforms and e-commerce websites.

Lastly, AI systems that can evolve and adapt to new tasks over time is an exciting area of research. For instance, if an

AI system is repeatedly used to perform actions on an Excel document via the OfficeJS API, it could potentially learn to improve its use of the existing API. Moreover, it could also learn to create new APIs and better abstractions that are optimized for the most common user queries. For example, let's say the original API lacks direct support for certain tasks, for which the AI agent may need to write longer, more complex code, increasing the chances of failure. In this case, can the AI system through its self-evolution process create a more specialized, direct API to make the generation process more reliable?

Multi-modal reasoning. The data analysis tasks require AI models to reason across multi-modalities such as code, text, speech, gestures, images, and table modalities. As multi-modal models such as GPT-4o, Llava, Phi-3-Vision evolve, we can expect improvements in AI-assisted data analysis tools. However, current models face several challenges: for example, LMMs are often trained on natural images (e.g., scenery, human faces) and struggle to understand images such as charts (Wu et al., 2024). Understanding gestures is another challenge, for instance, drawing an arrow between two columns in a table might indicate that the user wants to add a new column at that location. Additionally, while current models are not yet perfect in any single modality (e.g., interpreting a chart from the image modality), they can improve performance by combining reasoning across multiple modalities, such as analyzing a chart, the underlying data, and the output of code executions simultaneously.

Planning and exploration. Data analysis is not a linear process; it involves planning, reasoning, and exploration, requiring AI agents to be skilled at hierarchical planning, iterative problem solving, and flexible adaptation to new information and unexpected challenges. LLM-based planning has been explored in other contexts, such as embodied agents (Xi et al., 2023). However, evaluations on international planning competition problems have yielded mixed results (Valmeekam et al., 2023), showing that LLMs perform better when collaborating with humans rather than functioning autonomously. This highlights the need for further advancements in both LLM capabilities and human-AI collaboration for effective planning for data-analysis tasks.

6.4. Understanding user preferences and abilities

A key component to building interactive AI systems is understanding user preferences and abilities to tailor an experience that is smooth and intuitive for the users. This can be in the form of formative studies to inform and shape the design process such as (Gu et al., 2024b; McNutt et al., 2023) or summative studies to assess the performance of a system after development such as (Wang et al., 2023a;

Vaithilingam et al., 2024). Understanding user preferences can not only guide system design but can also potentially be integrated into the AI system itself, allowing it to deduce user preferences based on user's actions, history, and the task as the session progresses.

Existing user studies have provided valuable insights into user preferences in AI-driven data analysis. On user interaction preferences, the study from Data Formulator2 has found that the GUI + NL approach is more effective than chat-based AI assistants in communicating and constraining intent. Another study on the DynaVis tool (Vaithilingam et al., 2024) revealed that participants preferred AI-generated widgets that they can use to perform actions rather than having the AI perform actions directly. However, they also noted that frequently changing UIs could increase cognitive load. On preferences on AI suggestions and outputs, (Gu et al., 2024a) found that users' varying levels of statistical expertise influenced their reactions to AI-generated suggestions, with some finding them helpful and well-matched to their expertise, while others found them overly basic or requiring significant effort to understand due to mismatches with their background. In another study on how users verify AI-assisted data analysis (Gu et al., 2024b), participants needed both procedure-oriented artifacts (e.g., natural language explanations, code, and code comments) for understanding analysis steps and data artifacts (e.g., intermediate/output data tables and summary visualizations) for confirming low level details. Finally, on users' preference on understanding and leveraging AI capabilities, the study performed by (McNutt et al., 2023) has found that users want to understand and control the context given to the LLM model to ensure relevant and accurate outputs and users like linter-like assistants that highlight inappropriate usage of AI systems for users.

Despite these insights, there are still gaps in our understanding of user preferences. For example, how can we deduce contextual preferences, such as the most effective modalities or the desired level of AI autonomy, based on user history, past actions, and evolving tasks?. There's also a gap in understanding how to personalize UI elements for individual users without overwhelming them with frequent changes. Furthermore, research is needed on dynamic multi-application environments and AI-coordinated complex multi-step workflows, focusing, for example, on user intervention interfaces, user's ability to understand/manage AI-context across multiple apps, UI designs that span different applications, and trust and verification with output artifacts that might span multiple apps.

6.5. Data Infrastructure

Another key challenge is ensuring the availability of high-quality data tables in various domains that AI systems can

use to provide initiative analysis suggestions for users. This infrastructure might involve search engine style infrastructure such as crawling, indexing, and ranking data tables from the internet so that the AI system can easily extract the data tables necessary for user queries. We also need infrastructure to enable domain experts to create data table APIs, facilitate real-time updates of the data tables using online data, and manage enterprise and proprietary data tables effectively. Additionally, mechanisms for evaluating and ranking data tables for quality, potentially through crowdsourcing, would further enhance the reliability and relevance of available data resources. Another aspect to consider is data privacy, security, and having mechanisms to anonymize and aggregate data, which is a significant research area in itself, but beyond the scope of this work.

7. Conclusion

This paper explores the vast potential that generative AI-powered tools have in unlocking actionable insights from data. We first outline a set of tasks that AI systems can assist with, providing related works for some tasks and inspiration for others that are ripe for further exploration. We then discuss human-driven design considerations aimed at optimizing user interactions, workflows, and enhancing trust and reliability. Most of these design considerations can in fact be applied more broadly to enhance the design and functionality of AI systems in other domains. Finally, our work highlights research challenges such as improving the robustness of AI models in data analysis scenarios and developing benchmarks and evaluation metrics. We also note several model advancements that are needed such as improved continual learning, multi-modal reasoning, and planning and that we can leverage multi-step, multi-modal interactive data analysis scenarios as benchmarks to drive these model innovations. In addition, we emphasize the need for comprehensive user studies to discern preferences and optimize AI-driven tools for data analysis users. Our goal is that these efforts collectively will help bridge the gap between complex data and actionable insights for a wide-range of users.

References

- Abdin, M., Jacobs, S. A., Awan, A. A., Aneja, J., Awadallah, A., Awadalla, H., Bach, N., Bahree, A., Bakhtiari, A., Behl, H., et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.
- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Agashe, R., Iyer, S., and Zettlemoyer, L. Juice: A large scale distantly supervised dataset for open domain context-based code generation. *arXiv preprint arXiv:1910.02216*, 2019.
- Anthropic, A. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1, 2024.
- Bako, H. K., Liu, X., Battle, L., and Liu, Z. Understanding how designers find and use data visualization examples. *IEEE Transactions on Visualization and Computer Graphics*, 29(1): 1048–1058, 2022.
- Barman, S., Chasins, S., Bodik, R., and Gulwani, S. Ringer: web automation by demonstration. In *Proceedings of the 2016 ACM SIGPLAN international conference on object-oriented programming, systems, languages, and applications*, pp. 748–764, 2016.
- Bisercic, A., Nikolic, M., van der Schaar, M., Delibasic, B., Lio, P., and Petrovic, A. Interpretable medical diagnostics with structured data extraction by large language models. *arXiv preprint arXiv:2306.05052*, 2023.
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., and Zhang, Y. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.
- Buneman, P., Khanna, S., and Wang-Chiew, T. Why and where: A characterization of data provenance. In *Database Theory—ICDT 2001: 8th International Conference London, UK, January 4–6, 2001 Proceedings* 8, pp. 316–330. Springer, 2001.
- Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. D. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.
- Cheng, L., Li, X., and Bing, L. Is gpt-4 a good data analyst? *arXiv preprint arXiv:2305.15038*, 2023.
- Cui, H., Wang, C., Huang, J., Inala, J. P., Mytkowicz, T., Wang, B., Gao, J., and Duan, N. Codeexp: Explanatory code document generation. *arXiv preprint arXiv:2211.15395*, 2022.
- Dang, H., Mecke, L., and Buschek, D. Ganslider: How users control generative models for images using multiple sliders with and without feedforward information. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–15, 2022.
- Dibia, V. Lida: A tool for automatic generation of grammar-agnostic visualizations and infographics using large language models. *arXiv preprint arXiv:2303.02927*, 2023.
- Dibia, V., Chen, J., Bansal, G., Syed, S., Fourney, A., Zhu, E., Wang, C., and Amershi, S. Autogen studio: A no-code developer tool for building and debugging multi-agent systems. *arXiv preprint arXiv:submit/5782427*, January 2024.
- Ding, Y., Wang, Z., Ahmad, W., Ding, H., Tan, M., Jain, N., Ramanathan, M. K., Nallapati, R., Bhatia, P., Roth, D., et al. Cross-codeeval: A diverse and multilingual benchmark for cross-file code completion. *Advances in Neural Information Processing Systems*, 36, 2024.
- Dong, H., Cheng, Z., He, X., Zhou, M., Zhou, A., Zhou, F., Liu, A., Han, S., and Zhang, D. Table pre-training: A survey on model architectures, pre-training objectives, and downstream tasks. *arXiv preprint arXiv:2201.09745*, 2022.

- Dragicevic, P., Jansen, Y., Sarma, A., Kay, M., and Chevalier, F. Increasing the transparency of research papers with explorable multiverse analyses. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19*, pp. 1–15, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300295. URL <https://doi.org/10.1145/3290605.3300295>.
- Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.
- Duan, P., Warner, J., and Hartmann, B. Towards generating ui design feedback with llms. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–3, 2023.
- Fan, G., Wang, J., Li, Y., Zhang, D., and Miller, R. Semantics-aware dataset discovery from data lakes with contextualized column-based representation learning. *arXiv preprint arXiv:2210.01922*, 2022.
- Feng, Y., Wang, X., Pan, B., Wong, K. K., Ren, Y., Liu, S., Yan, Z., Ma, Y., Qu, H., and Chen, W. Xnli: Explaining and diagnosing nli-based visual data analysis. *IEEE Transactions on Visualization and Computer Graphics*, 2023.
- Ferdowsi, K., Williams, J., Drosos, I., Gordon, A. D., Negreanu, C., Polikarpova, N., Sarkar, A., and Zorn, B. Coldeco: An end user spreadsheet inspection tool for ai-generated code. In *2023 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pp. 82–91. IEEE, 2023.
- Fernandez, R. C., Elmore, A. J., Franklin, M. J., Krishnan, S., and Tan, C. How large language models will disrupt data management. *Proceedings of the VLDB Endowment*, 16(11):3302–3309, 2023.
- Floratou, A., Psallidas, F., Zhao, F., Deep, S., Hagleither, G., Tan, W., Cahoon, J., Alotaibi, R., Henkel, J., Singla, A., Grootel, A. V., Chow, B., Deng, K., Lin, K., Campos, M., Emani, K. V., Pandit, V., Shnayder, V., Wang, W., and Curino, C. Nl2sql is a solved problem... not! In *Conference on Innovative Data Systems Research*, 2024. URL <https://api.semanticscholar.org/CorpusID:266729311>.
- Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., and Ahmed, N. K. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*, 2023.
- Gero, Z., Singh, C., Cheng, H., Naumann, T., Galley, M., Gao, J., and Poon, H. Self-verification improves few-shot clinical information extraction. *arXiv preprint arXiv:2306.00024*, 2023.
- Goel, A., Gueta, A., Gilon, O., Liu, C., Erell, S., Nguyen, L. H., Hao, X., Jaber, B., Reddy, S., Kartha, R., Steiner, J., Laish, I., and Feder, A. Llms accelerate annotation for medical information extraction. In Hegselmann, S., Parziale, A., Shanmugam, D., Tang, S., Asiedu, M. N., Chang, S., Hartvigsen, T., and Singh, H. (eds.), *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pp. 82–100. PMLR, 10 Dec 2023. URL <https://proceedings.mlr.press/v225/goel23a.html>.
- Gong, R., Huang, Q., Ma, X., Vo, H., Durante, Z., Noda, Y., Zheng, Z., Zhu, S.-C., Terzopoulos, D., Fei-Fei, L., and Gao, J. Mindagent: Emergent gaming interaction, 2023.
- Gordon, A. D., Negreanu, C., Cambronero, J., Chakravarthy, R., Drosos, I., Fang, H., Mitra, B., Richardson, H., Sarkar, A., Simmons, S., et al. Co-audit: tools to help humans double-check ai-generated content. *arXiv preprint arXiv:2310.01297*, 2023.
- Grammel, L., Bennett, C., Tory, M., and Storey, M.-A. D. A survey of visualization construction user interfaces. In *EuroVis (Short Papers)*, pp. 019–023, 2013.
- Gu, K., Grunde-McLaughlin, M., McNutt, A., Heer, J., and Althoff, T. How do data analysts respond to ai assistance? a wizard-of-oz study. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–22, 2024a.
- Gu, K., Shang, R., Althoff, T., Wang, C., and Drucker, S. M. How do analysts understand and verify ai-assisted data analyses? In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–22, 2024b.
- Gu, K., Shang, R., Jiang, R., Kuang, K., Lin, R.-J., Lyu, D., Mao, Y., Pan, Y., Wu, T., Yu, J., et al. Blade: Benchmarking language model agents for data-driven science. *arXiv preprint arXiv:2408.09667*, 2024c.
- Gulwani, S. Programming by examples: Applications, algorithms, and ambiguity resolution. In *Automated Reasoning: 8th International Joint Conference, IJCAR 2016, Coimbra, Portugal, June 27–July 2, 2016, Proceedings 8*, pp. 9–14. Springer, 2016.
- Han, Y., Zhang, C., Chen, X., Yang, X., Wang, Z., Yu, G., Fu, B., and Zhang, H. Chartllama: A multimodal llm for chart understanding and generation. *arXiv preprint arXiv:2311.16483*, 2023.
- Heer, J., Conlen, M., Devireddy, V., Nguyen, T., and Horowitz, J. Living papers: A language toolkit for augmented scholarly communication. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–13, 2023.
- HeyMarvin. Heymarvin, 2024. URL <https://heymarvin.com/>. Accessed: 2024-07-03.
- Hohman, F., Conlen, M., Heer, J., and Chau, D. H. P. Communicating with interactive articles. *Distill*, 5(9):e28, 2020.
- Hollmann, N., Müller, S., and Hutter, F. Large language models for automated data science: Introducing caafe for context-aware automated feature engineering. *Advances in Neural Information Processing Systems*, 36, 2024.
- Hong, M.-H. and Crisan, A. Conversational ai threads for visualizing multidimensional datasets. *arXiv preprint arXiv:2311.05590*, 2023.
- Hong, S., Zheng, X., Chen, J., Cheng, Y., Wang, J., Zhang, C., Wang, Z., Yau, S. K. S., Lin, Z., Zhou, L., et al. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 2023.
- Huang, J., Guo, D., Wang, C., Gu, J., Lu, S., Inala, J. P., Yan, C., Gao, J., Duan, N., and Lyu, M. R. Contextualized data-wrangling code generation in computational notebooks. *arXiv preprint arXiv:2409.13551*, 2024a.

- Huang, K.-H., Chan, H. P., Fung, Y. R., Qiu, H., Zhou, M., Joty, S., Chang, S.-F., and Ji, H. From pixels to insights: A survey on automatic chart understanding in the era of large foundation models. *arXiv preprint arXiv:2403.12027*, 2024b.
- Huang, Q., Vora, J., Liang, P., and Leskovec, J. Mlagentbench: Evaluating language agents on machine learning experimentation. In *Forty-first International Conference on Machine Learning*, 2024c.
- Hutchins, E. L., Hollan, J. D., and Norman, D. A. Direct manipulation interfaces. *Human-computer interaction*, 1(4):311–338, 1985.
- Inala, J. P. and Singh, R. Webrelate: integrating web data with spreadsheets using examples. *Proceedings of the ACM on Programming Languages*, 2(POPL):1–28, 2017.
- Inala, J. P., Wang, C., Yang, M., Codas, A., Encarnación, M., Lahiri, S., Musuvathi, M., and Gao, J. Fault-aware neural code rankers. *Advances in Neural Information Processing Systems*, 35:13419–13432, 2022.
- Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., Dwivedi-Yu, J., Joulin, A., Riedel, S., and Grave, E. Atlas: Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*, 2022.
- Jiang, P., Rayan, J., Dow, S. P., and Xia, H. Graphologue: Exploring large language model responses with interactive diagrams. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–20, 2023.
- Joshi, H., Ebenezer, A., Cambronero, J., Gulwani, S., Kanade, A., Le, V., Radićek, I., and Verbruggen, G. Flame: A small language model for spreadsheet formulas. *arXiv preprint arXiv:2301.13779*, 2023.
- Jun, E., Daum, M., Roesch, J., Chasins, S., Berger, E., Just, R., and Reinecke, K. Tea: A high-level language and runtime system for automating statistical analysis. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, pp. 591–603, 2019.
- Jun, E., Seo, A., Heer, J., and Just, R. Tisane: Authoring statistical models via formal reasoning from conceptual and data relationships. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2022.
- Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., and McHardy, R. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*, 2023.
- Key, D., Li, W.-D., and Ellis, K. Toward trustworthy neural program synthesis. *arXiv preprint arXiv:2210.00848*, 2022.
- Kim, T. S., Choi, D., Choi, Y., and Kim, J. Stylette: Styling the web with natural language. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–17, 2022.
- Kumar, A., Agarwal, C., Srinivas, S., Feizi, S., and Lakkaraju, H. Certifying llm safety against adversarial prompting. *arXiv preprint arXiv:2309.02705*, 2023.
- Lahiri, S. K., Naik, A., Sakkas, G., Choudhury, P., von Veh, C., Musuvathi, M., Inala, J. P., Wang, C., and Gao, J. Interactive code generation via test-driven user-intent formalization. *arXiv preprint arXiv:2208.05950*, 2022.
- Lai, Y., Li, C., Wang, Y., Zhang, T., Zhong, R., Zettlemoyer, L., Yih, W.-t., Fried, D., Wang, S., and Yu, T. Ds-1000: A natural and reliable benchmark for data science code generation. In *International Conference on Machine Learning*, pp. 18319–18345. PMLR, 2023.
- Le, H., Chen, H., Saha, A., Gokul, A., Sahoo, D., and Joty, S. Codechain: Towards modular code generation through chain of self-revisions with representative sub-modules. *arXiv preprint arXiv:2310.08992*, 2023.
- Lee, B., Kazi, R. H., and Smith, G. Sketchstory: Telling more engaging stories with data through freeform sketching. *IEEE transactions on visualization and computer graphics*, 19(12):2416–2425, 2013.
- Lee, H., Phatale, S., Mansoor, H., Lu, K., Mesnard, T., Bishop, C., Carbune, V., and Rastogi, A. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.
- Lengerich, B. J., Bordt, S., Nori, H., Nunnally, M. E., Aphinyanaphongs, Y., Kellis, M., and Caruana, R. Llms understand glass-box models, discover surprises, and suggest repairs. *arXiv preprint arXiv:2308.01157*, 2023.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.
- Li, J., Hui, B., Qu, G., Yang, J., Li, B., Li, B., Wang, B., Qin, B., Geng, R., Huo, N., et al. Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls. *Advances in Neural Information Processing Systems*, 36, 2024.
- Li, P., He, Y., Yashar, D., Cui, W., Ge, S., Zhang, H., Fainman, D. R., Zhang, D., and Chaudhuri, S. Table-gpt: Table-tuned gpt for diverse table tasks. *arXiv preprint arXiv:2310.09263*, 2023a.
- Li, X., Lv, K., Yan, H., Lin, T., Zhu, W., Ni, Y., Xie, G., Wang, X., and Qiu, X. Unified demonstration retriever for in-context learning. *arXiv preprint arXiv:2305.04320*, 2023b.
- Li, Y., Choi, D., Chung, J., Kushman, N., Schrittwieser, J., Leblond, R., Eccles, T., Keeling, J., Gimeno, F., Dal Lago, A., et al. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097, 2022.
- Liang, J. T., Badea, C., Bird, C., DeLine, R., Ford, D., Forsgren, N., and Zimmermann, T. Can gpt-4 replicate empirical software engineering research?, 2023a.
- Liang, Y., Wu, C., Song, T., Wu, W., Xia, Y., Liu, Y., Ou, Y., Lu, S., Ji, L., Mao, S., et al. Taskmatrix. ai: Completing tasks by connecting foundation models with millions of apis. *arXiv preprint arXiv:2303.16434*, 2023b.
- Lin, Y., Li, H., Yang, L., Wu, A., and Qu, H. Inksight: Leveraging sketch interaction for documenting chart findings in computational notebooks. *IEEE Transactions on Visualization and Computer Graphics*, 2023.
- Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024a.

- Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., and Chen, W. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*, 2021.
- Liu, M. X., Sarkar, A., Negreanu, C., Zorn, B., Williams, J., Toronto, N., and Gordon, A. D. “what it wants me to say”: Bridging the abstraction gap between end-user programmers and code-generating large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–31, 2023a.
- Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024b.
- Liu, X., Wu, Z., Wu, X., Lu, P., Chang, K.-W., and Feng, Y. Are llms capable of data-based statistical and causal reasoning? benchmarking advanced quantitative reasoning with data. *arXiv preprint arXiv:2402.17644*, 2024c.
- Liu, Y., Kale, A., Althoff, T., and Heer, J. Boba: Authoring and visualizing multiverse analyses. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1753–1763, 2020.
- Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. Gpteval: NLg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023b.
- Liu, Z., Thompson, J., Wilson, A., Dontcheva, M., Delorey, J., Grigg, S., Kerr, B., and Stasko, J. Data illustrator: Augmenting vector design tools with lazy data binding for expressive visualization authoring. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2018.
- Liu, Z., Chen, C., Wang, J., Chen, M., Wu, B., Che, X., Wang, D., and Wang, Q. Chatting with gpt-3 for zero-shot human-like mobile automated gui testing. *arXiv preprint arXiv:2305.09434*, 2023c.
- Liu, Z., He, Y., Wang, W., Wang, W., Wang, Y., Chen, S., Zhang, Q., Yang, Y., Li, Q., Yu, J., et al. Internchat: Solving vision-centric tasks by interacting with chatbots beyond language. *arXiv preprint arXiv:2305.05662*, 2023d.
- Lu, P., Peng, B., Cheng, H., Galley, M., Chang, K.-W., Wu, Y. N., Zhu, S.-C., and Gao, J. Chameleon: Plug-and-play compositional reasoning with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.
- Luo, Y., Tang, J., and Li, G. nvbench: A large-scale synthesized dataset for cross-domain natural language to visualization task. *arXiv preprint arXiv:2112.12926*, 2021.
- Ma, P., Ding, R., Wang, S., Han, S., and Zhang, D. Insightpilot: An llm-empowered automated data exploration system. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 346–352, 2023.
- Maddigan, P. and Susnjak, T. Chat2vis: Generating data visualisations via natural language using chatgpt, codex and gpt-3 large language models. *Ieee Access*, 2023.
- Masry, A., Long, D. X., Tan, J. Q., Joty, S., and Hoque, E. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.
- Masry, A., Kavehzadeh, P., Do, X. L., Hoque, E., and Joty, S. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning. *arXiv preprint arXiv:2305.14761*, 2023.
- Masson, D., Malacria, S., Casiez, G., and Vogel, D. Directgpt: A direct manipulation interface to interact with large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2024.
- McNutt, A. M., Wang, C., Deline, R. A., and Drucker, S. M. On the design of ai-powered code assistants for notebooks. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2023.
- Methani, N., Ganguly, P., Khapra, M. M., and Kumar, P. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1527–1536, 2020.
- Mündler, N., He, J., Jenko, S., and Vechev, M. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *arXiv preprint arXiv:2305.15852*, 2023.
- Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.
- Narayan, A., Chami, I., Orr, L., Arora, S., and Ré, C. Can foundation models wrangle your data? *arXiv preprint arXiv:2205.09911*, 2022.
- Narechania, A., Srinivasan, A., and Stasko, J. Nl4dv: A toolkit for generating analytic specifications for data visualization from natural language queries. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):369–379, 2020.
- Nguyen, N. and Nadi, S. An empirical evaluation of github copilot’s code suggestions. In *Proceedings of the 19th International Conference on Mining Software Repositories*, pp. 1–5, 2022.
- Obeid, J. and Hoque, E. Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model. *arXiv preprint arXiv:2010.09142*, 2020.
- Olausson, T. X., Inala, J. P., Wang, C., Gao, J., and Solar-Lezama, A. Is self-repair a silver bullet for code generation? In *The Twelfth International Conference on Learning Representations*, 2023.
- OpenAI. Code interpreter, 2023. URL <https://platform.openai.com/docs/assistants/tools/code-interpreter>. Accessed: 2024-07-03.
- OpenAI. Gpt-4o, 2024. URL <https://openai.com/index/hello-gpt-4o/>. Accessed: 2024-07-03.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- Pandey, A., Srinivasan, A., and Setlur, V. Medley: Intent-based recommendations to support dashboard composition. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):1135–1145, 2022.

- Patil, S. G., Zhang, T., Wang, X., and Gonzalez, J. E. Gorilla: Large language model connected with massive apis. *arXiv preprint arXiv:2305.15334*, 2023.
- Peng, B., Galley, M., He, P., Cheng, H., Xie, Y., Hu, Y., Huang, Q., Liden, L., Yu, Z., Chen, W., et al. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*, 2023.
- Petridis, S., Terry, M., and Cai, C. J. Promptinfuser: Bringing user interface mock-ups to life with large language models. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI EA '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394222. doi: 10.1145/3544549.3585628. URL <https://doi.org/10.1145/3544549.3585628>.
- Poesia, G., Polozov, O., Le, V., Tiwari, A., Soares, G., Meek, C., and Gulwani, S. Syncromesh: Reliable code generation from pre-trained language models. *arXiv preprint arXiv:2201.11227*, 2022.
- Pu, X., Kross, S., Hofman, J. M., and Goldstein, D. G. Datamations: Animated explanations of data analysis pipelines. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2021.
- Qian, C., Cong, X., Yang, C., Chen, W., Su, Y., Xu, J., Liu, Z., and Sun, M. Communicative agents for software development. *arXiv preprint arXiv:2307.07924*, 2023.
- Raina, V., Liusie, A., and Gales, M. Is llm-as-a-judge robust? investigating universal adversarial attacks on zero-shot llm assessment. *arXiv preprint arXiv:2402.14016*, 2024.
- Ruoff, M., Myers, B. A., and Maedche, A. Onyx: Assisting users in teaching natural language interfaces through multimodal interactive task learning. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2023.
- Schettering, V., Di Bartolomeo, S., El-Assady, M., McNutt, A., Miller, M., Passos, J. P. A., and Adams, J. L. Doom or deliciousness: Challenges and opportunities for visualization in the age of generative models. In *Computer Graphics Forum*, volume 42, pp. 423–435. Wiley Online Library, 2023.
- Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., and Scialom, T. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36, 2024.
- Sclar, M., Choi, Y., Tsvetkov, Y., and Suhr, A. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*, 2023.
- Sharma, A., Li, X., Guan, H., Sun, G., Zhang, L., Wang, L., Wu, K., Cao, L., Zhu, E., Sim, A., et al. Automatic data transformation using large language model-an experimental study on building energy data. In *2023 IEEE International Conference on Big Data (BigData)*, pp. 1824–1834. IEEE, 2023.
- Shen, L., Shen, E., Luo, Y., Yang, X., Hu, X., Zhang, X., Tai, Z., and Wang, J. Towards natural language interfaces for data visualization: A survey. *IEEE transactions on visualization and computer graphics*, 29(6):3121–3144, 2022.
- Singh, M., Cambronero, J., Gulwani, S., Le, V., Negreanu, C., and Verbruggen, G. Datavinci: Learning syntactic and semantic string repairs. *arXiv preprint arXiv:2308.10922*, 2023.
- Singha, A., Cambronero, J., Gulwani, S., Le, V., and Parnin, C. Tabular representation, noisy operators, and impacts on table structure understanding tasks in llms. *arXiv preprint arXiv:2310.10358*, 2023.
- Srinivasan, A., Nyapathy, N., Lee, B., Drucker, S. M., and Stasko, J. Collecting and characterizing natural language utterances for specifying data visualizations. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–10, 2021.
- Suh, S., Min, B., Palani, S., and Xia, H. Sensecape: Enabling multilevel exploration and sensemaking with large language models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–18, 2023.
- Sultanum, N., Chevalier, F., Bylinskii, Z., and Liu, Z. Leveraging text-chart links to support authoring of data-driven articles with vizflow. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–17, 2021.
- Tankelevitch, L., Kewenig, V., Simkute, A., Scott, A. E., Sarkar, A., Sellen, A., and Rintel, S. The metacognitive demands and opportunities of generative ai. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–24, 2024.
- Tian, Y., Cui, W., Deng, D., Yi, X., Yang, Y., Zhang, H., and Wu, Y. Chartgpt: Leveraging llms to generate charts from abstract natural language. *IEEE Transactions on Visualization and Computer Graphics*, 2024.
- Tonmoy, S., Zaman, S., Jain, V., Rani, A., Rawte, V., Chadha, A., and Das, A. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 2024.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- Vaithilingam, P., Glassman, E. L., Inala, J. P., and Wang, C. Dynavis: Dynamically synthesized ui widgets for visualization editing. *arXiv preprint arXiv:2401.10880*, 2024.
- Valmeekam, K., Marquez, M., Sreedharan, S., and Kambhampati, S. On the planning abilities of large language models-a critical investigation. *Advances in Neural Information Processing Systems*, 36:75993–76005, 2023.
- Vijayan, A. A prompt engineering approach for structured data extraction from unstructured text using conversational llms. In *2023 6th International Conference on Algorithms, Computing and Artificial Intelligence*, pp. 183–189, 2023.
- Wadden, D., Lo, K., Wang, L. L., Cohan, A., Beltagy, I., and Hajishirzi, H. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V. (eds.), *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 61–76, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.6. URL <https://aclanthology.org/2022.findings-naacl.6>.

- Wang, C., Thompson, J., and Lee, B. Data formulator: Ai-powered concept-driven visualization authoring. *IEEE Transactions on Visualization and Computer Graphics*, 2023a.
- Wang, C., Lee, B., Drucker, S., Marshall, D., and Gao, J. Data formulator 2: Iteratively creating rich visualizations with ai, 2024a. URL <https://arxiv.org/abs/2408.16119>.
- Wang, Q., Ding, L., Cao, Y., Tian, Z., Wang, S., Tao, D., and Guo, L. Recursively summarizing enables long-term dialogue memory in large language models. *arXiv preprint arXiv:2308.15022*, 2023b.
- Wang, W., Dong, L., Cheng, H., Liu, X., Yan, X., Gao, J., and Wei, F. Augmenting language models with long-term memory. *Advances in Neural Information Processing Systems*, 36, 2024b.
- Wang, W., Ni, W., Han, T., Bai, L., Duan, B., and Ren, K. Dabench: A benchmark dataset for data-driven weather data assimilation. *arXiv preprint arXiv:2408.11438*, 2024c.
- Wang, Z., Li, J., Li, G., and Jin, Z. Chatcoder: Chat-based refine requirement improves llms' code generation. *arXiv preprint arXiv:2311.00272*, 2023c.
- Wen, Y., Yin, P., Shi, K., Michalewski, H., Chaudhuri, S., and Polozov, A. Grounding data science code generation with input-output specifications. *arXiv preprint arXiv:2402.08073*, 2024.
- Weng, L., Wang, X., Lu, J., Feng, Y., Liu, Y., and Chen, W. Insightlens: Discovering and exploring insights from conversational contexts in large-language-model-powered data analysis. *arXiv preprint arXiv:2404.01644*, 2024.
- Wu, G., Guo, S., Hoffswell, J., Chan, G. Y.-Y., Rossi, R. A., and Koh, E. Socrates: Data story generation via adaptive machine-guided elicitation of user feedback. *IEEE Transactions on Visualization and Computer Graphics*, 2023a.
- Wu, Q., Bansal, G., Zhang, J., Wu, Y., Zhang, S., Zhu, E., Li, B., Jiang, L., Zhang, X., and Wang, C. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*, 2023b.
- Wu, Y., Yan, L., Luo, Y., Wang, Y., and Tang, N. Evaluating task-based effectiveness of mllms on charts. *arXiv preprint arXiv:2405.07001*, 2024.
- Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., et al. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*, 2023.
- You, K., Zhang, H., Schoop, E., Weers, F., Swearngin, A., Nichols, J., Yang, Y., and Gan, Z. Ferret-ui: Grounded mobile ui understanding with multimodal llms. *arXiv preprint arXiv:2404.05719*, 2024.
- Yu, B. Veridical data science. In *Proceedings of the 13th international conference on web search and data mining*, pp. 4–5, 2020.
- Zhang, C., Li, L., He, S., Zhang, X., Qiao, B., Qin, S., Ma, M., Kang, Y., Lin, Q., Rajmohan, S., et al. Ufo: A ui-focused agent for windows os interaction. *arXiv preprint arXiv:2402.07939*, 2024.
- Zhang, K., Li, G., Li, J., Li, Z., and Jin, Z. Toolcoder: Teach code generation models to use apis with search tools. *arXiv preprint arXiv:2305.04032*, 2023a.
- Zhang, T., Yue, X., Li, Y., and Sun, H. Tablellama: Towards open large generalist models for tables. *arXiv preprint arXiv:2311.09206*, 2023b.
- Zhao, J., Xu, S., Chandrasegaran, S., Bryan, C., Du, F., Mishra, A., Qian, X., Li, Y., and Ma, K.-L. Chartstory: Automated partitioning, layout, and captioning of charts into comic-style narratives. *IEEE transactions on visualization and computer graphics*, 29(2):1384–1399, 2021.
- Zuur, A. F., Ieno, E. N., and Elphick, C. S. A protocol for data exploration to avoid common statistical problems. *Methods in ecology and evolution*, 1(1):3–14, 2010.