# Data Therapist: Eliciting Domain Knowledge from Subject Matter Experts Using Large Language Models

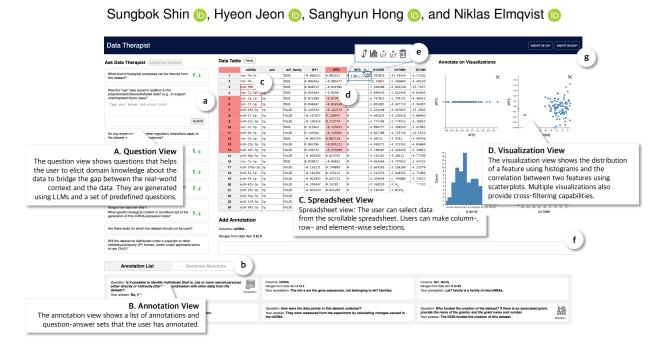


Fig. 1: **Data Therapist user interface.** Data Therapist helps to elicit domain knowledge from experts about a domain-specific dataset by asking questions yielding answers as annotations for the dataset. These annotations bridge the gap between the real-world context and the data. The user has three ways to add annotations. The first is to add annotations about the overall dataset. The second is to add data-specific annotations with the help of spreadsheets (C) and visualizations (D). Finally, the user can answer questions generated by Data Therapist in the question view (A). The annotation view (B) shows the list of annotations as well as questions and answers.

Abstract—Effective data visualization requires not only technical proficiency but also a deep understanding of the domain-specific context in which data exists. This context often includes tacit knowledge about data provenance, quality, and intended use, which is rarely explicit in the dataset itself. We present the Data Therapist, a web-based tool that helps domain experts externalize this implicit knowledge through a mixed-initiative process combining iterative Q&A with interactive annotation. Powered by a large language model, the system analyzes user-supplied datasets, prompts users with targeted questions, and allows annotation at varying levels of granularity. The resulting structured knowledge base can inform both human and automated visualization design. We evaluated the tool in a qualitative study involving expert pairs from Molecular Biology, Accounting, Political Science, and Usable Security. The study revealed recurring patterns in how experts reason about their data and highlights areas where Al support can improve visualization design. All supplemental materials are available at https://osf.io/e4fxu.

Index Terms—Domain knowledge, mixed-initiative interaction, LLMs, annotation, visualization recommendation.

#### 1 Introduction

- Sungbok Shin is with Aviz at Inria and Université Paris-Saclay, Saclay, France. E-mail: sungbok.shin@inria.fr.
- Hyeon Jeon is with Seoul National University, Seoul, South Korea. E-mail: hj@hcil.snu.ac.kr.
- Sanghyun Hong is with Oregon State University, Corvallis, OR, USA. E-mail: sanghyun.hong@oregonstate.edu.
- Niklas Elmqvist is with Aarhus University, Aarhus, Denmark. E-mail: elm@cs.au.dk.

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

"Tell me about your data." This question, or some variant thereof, is one that most data visualization experts will reach for whenever they start a new project in a new domain. As visualization researchers, our business is data first and foremost, but the data we visualize is invariably someone else's. And so, as part of this exploratory process to gather necessary information about a new domain—a process that has variously been referred to as design study methodology [35], usercentered design [12,49], or participatory design [30] depending on your background and approach to visualization—we tend to find ourselves engaging with the owners of this data in interviews that are not unlike what a psychotherapist will engage in with a new patient. Regardless of domain, the questions we ask will be more or less similar: the source of the data, how often it is updated, its quality, the stakeholders responsible for it, and so on. Nevertheless, understanding the data model of a domain, as well as its metadata and tacit knowledge, is a vital activity in designing a visualization that addresses user needs.

In this paper, we present the DATA THERAPIST (Figure 1), an LLMdriven web-based system designed to elicit domain knowledge from experts through a mixed-initiative approach [14, 17] based on iterative question-and-answers as well as user annotation. Starting with a dataset uploaded by the domain expert, the Data Therapist will analyze the data and ask a mix of questions based on the seven genres of metadata [11] as well as its growing body of knowledge about the data and the domain. Users can choose between answering questions, annotating the overall dataset, or adding data-specific annotations using their domain knowledge. Data-specific annotations can either be added to columns, rows, or individual cells in an interactive spreadsheet, or to interactive visualizations that are automatically created as the user selects data. All answers are automatically added and organized into a structured knowledge base that can be exported at the end of an elicitation session and shared with a visualization designer—or even used as the basis for an automated visualization recommendation system.

We have validated the Data Therapist through a web-based implementation that uses a server-side Large Language Model (GPT-4 in our case) to formulate questions, structure elicited data, and provide contextual memory for guiding further inquiries into the problem domain. Furthermore, we use the Data Therapist in a qualitative user study involving three sets of two domain experts drawn from three distinct problem domains: computer security, molecular biology and genomics, and accounting. In a two-phase evaluation protocol, we first ask the first domain expert in a pair to use Data Therapist to elicit domain knowledge in a 45-minute session using the tool. In the second phase, the second expert in the pair uses a view-only version of the Data Therapist to browse the resulting knowledge base and grade its coverage, quality, and insightfulness with regards to Gebru et al.'s genres of metadata [11]. Our qualitative findings show strong evidence that each knowledge base about the three different datasets score highly on all counts.

To summarize, we claim the following contributions in this paper: (1) a structured mixed-initiative interaction approach to question-answering and user-driven annotation for eliciting domain knowledge about a dataset; (2) the DATA THERAPIST LLM-driven system for implementing the above domain elicitation process in a practical web-based tool; and (3) results from a qualitative user study spanning 4 pairs of groups (8 expert participants) from Accounting, Molecular Biology, Political Science and Usable Security demonstrating the utility of the approach and our implementation. Information about our prompts and detailed results of our study is shown in our OSF link.<sup>1</sup>

#### 2 BACKGROUND

Below we review the relevant background for this research, including domain knowledge elicitation, metadata practices, the use of LLMs for knowledge management, and existing work aiding the visualization design process. For each topic, we review our proposed contributions.

# 2.1 Knowledge Elicitation in Visualization Design

Effective visualization design requires both technical visualization expertise and deep domain knowledge. Several theoretical models describe how data transforms into visualizations [20, 31, 35, 41], but these models often assume ideal knowledge transfer between domain experts and visualization designers.

In practice, professional visualization design typically involves a division of roles—domain experts who provide data and define goals, and visualization designers who create the visual representations [45]. This division exists because domain experts generally lack design expertise, time, and sometimes motivation to create effective visualizations themselves [13, 42]. However, it introduces significant communication gaps that can impede the design process.

These gaps manifest in several ways. Bigelow et al. [2] observed that "the designers' inferences about data behavior were often inaccurate" and "their inferences about the data structure are also separated from the actual data structure." Similarly, Walny et al. [42] noted that "for

designers, anticipating all of the implications of scale and interaction for a given dataset remains challenging." These findings highlight a fundamental issue: visualization designers lack crucial contextual knowledge that domain experts possess about their data.

This knowledge gap includes understanding data provenance, quality issues, implicit assumptions, and domain-specific interpretations. Lin et al. [21] conceptualized this tacit knowledge as "data hunches"—insights about mismatches between reality and data representation. These hunches acknowledge that data is always an imperfect representation of reality. Importantly, Lin et al. [22] argue that capturing these hunches at the data level, before visualization design begins, significantly affects the quality of the final visualization.

We think that the systematic approaches for eliciting and documenting this tacit domain knowledge can benefit the Visualization community. While interview techniques [35, 49] and collaborative design sessions [30] are common, they often rely on unstructured conversations that may miss critical domain insights. Furthermore, these approaches typically generate unstructured documentation that is difficult to systematically incorporate into visualization design decisions.

Our contribution. The Data Therapist addresses these challenges by providing a structured approach to eliciting and documenting domain experts' tacit knowledge about their data. Systematically capturing data hunches through guided questioning and annotation allows our system to bridge the gap between reality and data representation. This structured knowledge externalization helps visualization designers better understand the context of the data, enabling more accurate and effective visualization outcomes while reducing miscommunication between domain experts and visualization designers.

#### 2.2 Metadata and Documentation for Datasets

As datasets grow in complexity and importance, systematic approaches to documenting their context, provenance, and limitations have become increasingly vital. Gebru et al. [11] introduced the concept of "datasheets for datasets," drawing inspiration from the electronics industry where components are accompanied by standardized documentation. Their framework proposes documenting seven genres of metadata: motivation, composition, collection process, preprocessing, uses, distribution, and maintenance. This comprehensive approach aims to increase transparency and accountability in data-driven systems.

The visualization community has also explored how metadata can enhance data visualization practices. Burns et al. [4,5] investigate the risks and benefits of disclosing metadata in visualizations, finding that appropriate metadata can significantly impact viewer trust and interpretation. Their studies demonstrate that metadata disclosure influences how audiences perceive data reliability, especially when visualizations are used in decision-making contexts.

Beyond documentation frameworks, researchers have explored how metadata can be explicitly incorporated into visualization systems. Srinivasan et al. [38] introduce "data facts"—concise, automatically generated statements about statistical features or contextual information about a dataset. These data facts augment visualizations to facilitate interpretation and communication, addressing challenges that arise when viewers lack domain knowledge or analytical expertise.

The visualization of metadata itself has also emerged as a research focus. Makhija et al. [26] developed Loch Prospector, a system for visualizing metadata for data lakes. Their approach helps users navigate large open data repositories through metadata representations, demonstrating how metadata visualization can aid understanding.

Despite these advances, current approaches to metadata documentation remain primarily manual and often disconnected from the visualization design process. Documentation frameworks like datasheets [11] provide valuable structure but offer limited guidance on eliciting tacit knowledge from domain experts. Furthermore, while annotation systems exist for documenting datasets, they typically focus on explicit characteristics rather than the implicit domain knowledge that affects interpretation. Another challenge is that current metadata documentation approaches generally occur separately from visualization design, creating a disconnect between data context and visual representation.

This separation can lead to visualizations that fail to account for important contextual factors known to domain experts but not explicitly recorded in the dataset or its documentation.

Our contribution. The Data Therapist addresses these limitations by providing an interactive, mixed-initiative system for eliciting and documenting dataset metadata. By structuring our elicitation process around the seven genres of metadata proposed by Gebru et al. [11], we ensure comprehensive coverage of dataset context while focusing on domain-specific tacit knowledge. Our approach bridges the gap between manual documentation and visualization design by creating a structured knowledge base that can directly inform design decisions. Furthermore, by making the process interactive and guided by an LLM, we lower the barrier to thorough documentation while ensuring domain experts can contribute their specialized knowledge efficiently.

## 2.3 Language Models as Knowledge Bases

A language model is a statistical model, trained on a large corpus of text data, that works by the *next token prediction* objective. A token is typically a word or subword, depending on how the model is designed. Given a sequence of tokens, the model computes the probability distribution over possible next tokens and chooses the most likely one.

Recent work has demonstrated that the *scale* is the key factor in enabling them to accomplish complex tasks effectively [3, 10, 44]. These "large" language models (LLMs) typically contain billions of parameters [1, 3, 7, 33, 40], but the scale also makes their training computationally intensive and resource-demanding. As a result, training is almost exclusively done by large-budget organizations such as OpenAI, Google and Meta, with the resulting *pre-trained* models offered either as online services (e.g., OpenAI's ChatGPT or Google's Gemini) or through open-source releases (e.g., Meta's LLaMA models).

Because LLMs are trained on massive corpora of curated text from diverse sources, there have been studies investigated their use as knowledge bases. Heinzerling and Inui [15] examined the capabilities of LLMs in this role, analyzing their entity representations, storage capacity, and question-answering abilities. Their work demonstrated that LLMs can store and retrieve factual knowledge, albeit with limitations in consistency and reliability compared to traditional knowledge bases. This lack of consistency and reliability primarily arises during interactions with users—specifically when generating questions. To improve factual accuracy and reduce inconsistencies, our system leverages prompts designed to guide the reasoning process of LLMs.

Our contribution. The Data Therapist leverages LLMs not merely as knowledge repositories but as active participants in the knowledge elicitation process. Our system uses an LLM to generate contextually relevant questions, interpret free-text responses, and organize elicited knowledge into a structured format aligned with established metadata frameworks. This approach addresses the lack of domain knowledge by creating a structured conversation that helps domain experts externalize their tacit understanding. While our system does not directly address the lack of visualization expertise, it creates a bridge between domain knowledge and visualization design by providing a structured knowledge base that visualization systems can leverage. By combining the conversational capabilities of LLMs with structured annotation tools, we create a more accessible and comprehensive approach than either manual documentation or purely automated approaches alone.

## 2.4 Augmenting the Visualization Design Process

Automated visualization recommendation has evolved significantly since Mackinlay's pioneering work [24,25]. Current recommendation systems use diverse inputs, such as data properties [19,46,47], perceptual principles [46], expert feedback [23], and formalized design knowledge [29]. AI-driven approaches [9,18,39,50] further automate this process by incorporating both data properties and usage context.

Design feedback mechanisms offer another approach to augmentation. Systems employing a linting metaphor provide automated critiques of visualization code or output [6, 16, 28], highlighting potential issues or suggesting improvements. Other feedback-driven augmentation approaches incorporate perceptual signals such as visual salience and color usage to inform design decisions [36, 37]. Choi et al. [8] take a different approach with VisLab, employing crowdsourcing to gather diverse user feedback on visualizations.

Authoring tools focus on enhancing expressivity and flexibility during creation. Systems such as Charagraph [27] and ChartSpark [48] enable more sophisticated visualization creation with support for customization and narrative intent. Data Formulator [43] takes a conceptual approach by binding textual concepts to visualizations, creating a bridge between natural language and visual representation.

Recent work has begun to address issues of trust and transparency in visualization design by incorporating data context. Burns et al. [4,5] examine how embedding metadata within visualizations influences user trust and understanding. This highlights a growing recognition that effective visualization requires not just technical design knowledge but also a deeper understanding of data provenance and limitations. Nevertheless, current visualization recommendation and authoring tools—automated or otherwise—do not focus on eliciting domain knowledge from users even if such knowledge is critical for the design process [35].

Our contribution. The Data Therapist addresses this gap by providing a structured framework for capturing and communicating domain experts' implicit knowledge. Unlike existing augmentation approaches that focus primarily on the technical aspects of visualization design, our system targets the upstream problem of knowledge elicitation and documentation. This approach complements existing recommendation and feedback systems by providing them with the crucial contextual information needed to make appropriate design decisions.

## 3 THE DATA THERAPIST

We introduce the Data Therapist, a mixed-initiative [14, 17] annotation system for eliciting metadata and tacit knowledge about datasets. Drawing inspiration from human therapists, the Data Therapist engages domain experts in a structured dialogue about their data, combining directed questioning with freeform annotation capabilities. This dual approach allows experts to externalize both explicit and implicit knowledge throughout the annotation process. Below, we detail the design considerations, conceptual model, and method of the Data Therapist.

## 3.1 Design Considerations

The Data Therapist's design is guided by several key design considerations (DCs) that shape its functionality and user experience:

- **DC1 Bridge reality and representation.** The system must elicit annotations that minimize the gap between reality and its data representation. Questions should prompt domain experts to explain context at a level that allows non-experts to understand the dataset's limitations, assumptions, and real-world significance.
- DC2 Support multi-granular annotation. Users should be able to flexibly select and annotate data at multiple levels of granularity entire columns or rows, individual cells, or a combination depending on where domain knowledge applies.
- **DC3 Enable analysis.** The system must incorporate analytical capabilities that allow experts to identify and explain patterns and anomalies in their data, facilitating the externalization of insights that might otherwise remain implicit.
- DC4 Ensure quality. The tool should provide mechanisms to verify annotation quality, identify contradictions between annotations, detect incomplete responses, and prompt for clarification or additional detail when necessary.
- **DC5 Provide oversight.** Users need clear visibility into their annotation progress, including what aspects of the data have been addressed and what remains unexplained, enabling systematic and comprehensive knowledge capture.
- **DC6** Facilitate comprehensive coverage. The system should guide users toward complete domain knowledge capture by systematically addressing all relevant metadata categories. This includes

proactively identifying knowledge gaps and steering the elicitation process toward unexplored aspects of the dataset.

## 3.2 Conceptual Model

The overarching goal of our work is to bridge the gap between data and real-world context [22] by eliciting knowledge from domain experts (DC1). Here the Data Therapist plays the role of a non-expert—it has access to the data but lacks background knowledge about its problem domain. To address the design considerations, the approach provides two complementary modes seamlessly combined in a mixed-initiative interaction [14,17] model:

- A guided question-and-answer dialogue that probes for specific aspects of data context, and
- 2. A **direct annotation interface** where experts can attach observations to the data at various levels of data granularity.

By interweaving these modes, the Data Therapist helps users progressively build a comprehensive knowledge base that captures both formal metadata and tacit domain insights. The purpose is not to shift reliance onto the system by having experts merely answer its questions, but rather to enhance the annotation process by asking questions that complement aspects the expert may not have initially considered.

## 3.3 Addressing the Design Considerations

The Data Therapist translates our design considerations into practical functionality through a structured mixed-initiative annotation workflow. When a domain expert begins with an unannotated dataset, the system activates complementary pathways for knowledge capture:

For dataset-wide context (DC1, DC6): Experts can add global annotations without selecting specific data points. These annotations typically cover provenance, collection methodology, and intended use cases as the foundational reality behind the datasets.

For data-specific insights (DC2): The system enables multi-granular selection, allowing experts to highlight individual cells, entire columns or rows, or arbitrary regions to annotate. This flexibility accommodates different types of domain knowledge that may apply at varying levels of specificity.

Throughout this process, the Data Therapist actively participates by:

- Generating targeted questions based on observed data patterns and metadata frameworks (addressing DC3, DC6);
- Cross-referencing new annotations against existing ones to flag potential inconsistencies (addressing DC4);
- Tracking annotation coverage across the dataset and metadata categories (addressing DC5); and
- Suggesting areas for further exploration based on detected knowledge gaps (addressing DC6).

This bidirectional interaction creates a knowledge elicitation loop where expert-initiated annotations inform system-generated questions, which in turn can prompt deeper expert insights. The resulting knowledge base captures both explicit metadata and tacit domain knowledge in a structured format that can inform visualization design decisions.

## 4 IMPLEMENTATION: DATA THERAPIST

Here we delineate the Data Therapist system. First, we present an overview of how the system is structured. Then, we describe how the interface is structured. Third, we detail how we structured the prompts and our system's implementation details.

# 4.1 System Overview

The Data Therapist system adopts a server-client architecture to streamline data annotation and analysis. An overview of the system is shown in Figure 2. The client provides the annotation and question answering interface whereas the server stores annotations and generates questions. To begin with, we present our user interface in § 4.2. Then, we explain our three annotation options (§ 4.3, § 4.4, and § 4.5). Afterwads, we

detail our question generation mechanism in § 4.6 (DC1, DC4) and our prompting methodology in § 4.7.

Client frontend. The client frontend interface provides various functionalities to users to help elicit their domain knowledge about the dataset. First, it allows users to add annotations to the dataset and to help the process, it offers questions to guide towards enriching the annotation process. It also offers multiple methods for selecting data instances within the dataset, such as spreadsheet-style views and feature-distribution visualizations (DC2, DC3) (see § 4.2). Users can directly make annotations. The user can also choose to make annotations by answering the questions that Data Therapist generates.

The system helps users track their progress by displaying a list of annotations, along with a thematic summary of the information they contain (DC5). Once a sufficient number of annotations have been generated, users can either export the full list of annotations or generate a dataset report based on the annotation-driven analysis.

Finally, to export the annotations (Figure 1g), the user can click the button 'export annotations' to receive a JSON file of all the annotations.

Server backend. The primary goal of the server is to manage user annotations and generate questions that elicit user responses about the dataset. Each time a user uploads an annotation, it is stored on the server. In addition, the server performs background operations to update the system—such as adding and prioritizing new questions, or generating summaries. These updates are processed on the server before being reflected in the client interface. We detail the prompting methodology in § 4.7. Finally, to support ongoing work, the server continuously saves the most recent version of the annotated dataset.

## 4.2 User Interface

Fig. 1 shows an overview of the Data Therapist interface. The view is dominated by a scrollable spreedsheet view of the current dataset. This interface is shown as soon as the user uploads a dataset into the tool.

The **basic Data Therapist mixed-initiative workflow** involves the user continuously choosing between three options:

- 1. Annotate the overall dataset (GENERAL ANNOTATION);
- 2. Annotate a specific region of the dataset (DATA-SPECIFIC ANNOTATION); or
- 3. Answer a Data Therapist-initiated question about the dataset (QUESTION-ANSWERING).

Below we detail these three mechanisms (Figure 2): (1) general annotation, (2) data-specific annotation, and (3) question-answering.

## 4.3 General Annotation

Data Therapist allows the user to add general annotations to the dataset by entering text into the text box (see Figure 1f) below the main spreadsheet view. The implication is that these annotations apply to the full dataset and **not** to specific regions (cells, columns, or rows). There is a wide range of information that can be annotated about the dataset itself. For example, one can specify where the dataset originated, why it is important, its intended purpose, potential applications, and its broader significance. Each time an annotation is uploaded to the interface, the annotation is stored in the server, and is shown in the list of annotations in the annotation view (see Figure 1B) (DC5).

## 4.4 Data-Specific Annotation

Annotating individual data entries allows the user to provide finegrained context or observations about specific portions of the dataset. There are several methods to create this type of annotations (DC2).

First, they can directly select from the spreadsheet (Figure 1C). The spreadsheet enables selecting multiple instances by row, column, or individuals cells, as well as using a rectangular rubberband selection. It also allows reordering the dataset as well as changing the sorting order.

Second, users can choose instances from visualizations automatically created by selecting columns. The visualizations helps users quickly



Fig. 2: **The Data Therapist workflow.** The Data Therapist is composed of the user interface and a backend system. The user (a domain expert) provides both general and data-specific annotations through the user interface, which are then stored in our system. In turn, the backend stores these entries and generates additional questions relevant to the newly added or existing annotations, presenting them to the user via the interface.

understand and read patterns in the dataset that would be difficult to see in the table view (DC3). This helps the user not only to understand the data, but also easily select the instances that they want to annotate. Our implementation currently supports creating histograms and scatterplots. By right-clicking on a column header in the spreadsheet, the user can choose from several icons. Selecting the first icon sorts the column's values according to its feature. Selecting the second icon displays a histogram depicting the column's distribution next to the spreadsheet (see Figure 1e). The next two icons allow the user to designate two axes for a scatterplot. When multiple visualizations are shown in the interface, these visualizations work as a cross-filtering functions, where only corresponding instances in the visualizations are selected. The user can add annotations to the selected location using the textbox in Figure 1f.

In addition, to help users locate the relevant data, hovering over a data-specific annotation in the annotation view (Figure 1B) highlights the corresponding instance(s) in the spreadsheet (Figure 1c).

# 4.5 Question-Answering

Finally, the last approach is to elicit domain knowledge from the user by having them answer the questions automatically generated by Data Therapist in the *question view* (Figure 1A). Question and answering works in the following manner (question generation is is explained in § 4.6):

- Data Therapist initializes the question view list with 10 questions relevant to that dataset.
- 2. The user can choose to answer any question in the list by clicking on it, opening up a textbox where they can write their response.
- 3. If a question is irrelevant, the user can remove it from the list.
- 4. The user may refill the list of questions in the question view by clicking "generate new questions."

To support the question-answering process, Data Therapist provides two key features: (1) verification of user responses, and (2) automated thematic summarization of predefined questions. After the user submits an answer, the system verifies whether the response adequately addresses the question. If the answer passes this check, the question disappears from the list after a brief delay. Otherwise, the system displays a message explaining why the answer was rejected, allowing the user to revise and resubmit.

Specifically, the system evaluates two criteria: (1) whether the answer faithfully addresses the question, and (2) whether it conflicts with any existing annotations. This process is powered by ChatGPT, as detailed in § 4.7. We illustrate how a response is validated in Figure 3. Also, once the user has answered enough predefined questions (from when the user answered 2 questions per theme), Data Therapist automatically generates a thematic overview of those responses. This summary clarifies which metadata fields have been completed and which remain unfilled. Finally, to facilitate sharing the accumulated information, the



#### (a) Initial Question:

In what ways have the International Monetary Fund's considerations affected Argentina's position in the dataset over the years?



#### (b) Initial Answer:

The International Monetary Fund generally forces the government to facilitate massive layoffs.



#### (c) Feedback:

The answer mentions the impact of the International Monetary Fund's (IMF) actions on layoffs and collective fury, but it does not directly address how these considerations have affected Argentina's position in the dataset over the years.



#### (d) Revised Answer:

The International Monetary Fund (IMF) has influenced Argentina's position in the dataset over the years by enforcing austerity measures, such as public sector layoffs and spending cuts, as conditions for financial assistance.

Fig. 3: **How Data Therapist validates responses.** Given an initial question (a), the expert provides an answer (b). The Data Therapist system then evaluates it, either validating or rejecting it with feedback (c). If rejected, the expert can submit a revised answer until it is validated (d).

"export annotations" feature (Figure 1g) enables users to download a .JSON file containing all annotations.

## 4.6 Generating Questions

Here we describe the mechanism behind how the questions are generated and presented to the user.

Generating questions. In the backend, as soon as the dataset is uploaded, the LLM generates 30 questions that aim to elicit domain knowledge and information from the user. We maintain all questions in a single list and use a custom prioritizing function to select those that best align with our priorities whenever a request is made. When a dataset is first created, we generate 30 questions for two reasons. The first is to provide an initial set of questions as a starting point for the annotation process. The second is to prevent running out of questions in cases where no annotations are made but the user asks for more questions, since new questions are only generated upon user annotation. Each time the user leaves an annotation, five additional questions related to that annotation are generated. If there are fewer than ten LLM-generated questions left in the list, we immediately add ten more to maintain a sufficient number of available questions. Below we describe the prioritizing function used to select the questions.

Prioritizing questions. The prioritization of questions are partially motivated by the priorizing methodology conducted by Park et al. [34]. As shown in Equation 1, the score of a question is based on three factors, originality, recency, and importance. We let each metric be

integers between 1 and 5 in Likert scale. We add scores from these three factors, and select from the highest-scoring questions.

$$Score_q = Originality_q + Recency_q + Importance_q. \tag{1}$$

The originality of a question, Originality $_q$ , measures how distinct a given question is relative to existing annotations. We calculate it by first computing a similarity score based on the arithmetic average between the question and previously used annotation questions, then scale this score from 1 to 5 (rounded to integers). The recency, Recency $_q$ , tracks how recent the question is. Whenever a new annotation is created, the question starts with a recency score of 5. Each time the user adds an annotation, this score decreases by 1, but never below 1. Finally, importance, Importance $_q$ , indicates how critical the question is to our primary goal of understanding the dataset. We instruct an LLM to evaluate (1) the question's ability to clarify the dataset and (2) its potential to introduce novel insights.

In addition, when prioritizing predefined questions, we prioritize those from the theme with the most unanswered items. This approach ensures an even update of our thematic summary. If two questions share the same priority level, one is selected at random.

Filling in questions. The initial question view is populated with 10 questions. Among them, half of them are from the list of predefined questions by Gebru et al. [11], and the other half are selected from the list of LLM-generated questions. We recommend predefined questions for two main reasons: first, these datasets were originally created to describe themselves; second, we seek to further enrich their metadata. When filling in questions, the interface displays up to 10 questions at a time. A question can be removed in one of two ways: either by answering it or by manually deleting it. The "generate new questions" button becomes active once the interface has at most four remaining questions. At that point, half of the empty slots are populated with predefined questions, and the other half are populated with LLMgenerated questions. If the number of empty slots is odd, one additional LLM-generated question is added. For example, if the user presses the "generate new questions" button when only two questions remain, there are eight empty slots to fill, resulting in four predefined questions and four LLM-generated questions. However, if just one question remains, there are nine empty slots to fill, which will be split into four predefined questions and five LLM-generated questions. Additionally, once all predefined questions are exhausted, the system provides only LLM-generated questions.

# 4.7 Prompt Engineering

The Data Therapist takes a human-in-the-loop approach, where human knowledge is added to the system, and the system uses it to create questions generated from LLMs, and again, the questions help users elicit more detailed explanations and knowledge. We structured this procedure by employing the following 5 strategies. To this end, we refer to both prior work and instructions provided by ChatGPT [32].

- Describe the role and task precisely to ensure responses faithfully align with our intent.
- Structure responses clearly and concisely to maintain precision and avoid verbose outputs.
- Guide a specific response template (e.g., using delimiters for outputs) — to foster consistency and streamline parsing.
- Detail context and provide all relevant information to anchor the system's output in accurate domain knowledge.
- Use global and local variables in the prompt to maintain consistent references across global (e.g., specifying the role of Data Therapist) and local (e.g., tasks) specifications.

Below we detail the prompt templates used in the Data Therapist. We first show how we specify the role of Data Therapist in the interaction loop. Next, we present templates used for generating and validating questions. We lastly describe templates for smaller tasks, such as

summarizing texts and assessing the importance of questions. The full prompt details are provided in the OSF submission <sup>2</sup>.

Conceptualizing Data Therapist. In defining the role of a "data therapist," we sought to describe it in detail, clarifying its objectives, the problems it addresses, and the responsibilities we expect it to fulfill. This prompt serves as a global variable for each task prompt. Defining this as [Role] $_{DT}$  is as follows:

[Role]<sub>DT</sub>: The role of the Data Therapist is to elicit knowledge about the dataset from users by asking appropriate questions that, by answering, could help them understand their own data. The questions should aim to bridge the gap between reality (the situation, environment, and background surrounding the dataset) and the dataset itself. While the data is present, it does not, on its own, explain the background or issues. Ideally, annotations should be made so that even someone unfamiliar with the dataset can understand it by reading those annotations. Although the main goal is to help extract annotations by asking questions, the Data Therapist can also assist in other tasks related to annotation, such as validating questions.

Generating questions. We generate two types of questions.

The first type focuses on the dataset itself, while the second type is derived from the annotations. In generating T1, we leverage ChatGPT's built-in RAG and in-context learning (ICL) functionalities, while for follow-up questions in T2 we embed the user's previous, relevant responses into each subsequent query to the LLM. Below, we present the prompt templates used for generating each type of question.

We define  $\mathbf{T}_{T1}$  as the template for generating the first type of question. Let  $[Role]_{DT}$ , and  $[Output\text{-}format]_{T1}$  be textual strings, respectively, that prompts the role of the Data Therapist, the task of generating this question type, and the specific output format (for  $\mathbf{T}_{T1}$ , we present a json-style example of an output). {Dataset} is the dataset of interest in a textual form. We define  $[Task]_{T1}$  as follows:

[Task]<sub>T1</sub>: Read the dataset. Generate 30 key questions you would ask the experts of the dataset that would help people who do not know about the data understand the dataset. The generated questions should (1) help clarify the dataset from a non-expert's perspective and (2) bridge the gap between the explanations and the dataset itself.

Then, summing them all up, we define  $T_{T1}$  as:

$$\mathbf{T}_{T1} = [\text{Role}]_{DT} + \{\text{Dataset}\} + [\text{Task}]_{T1} + [\text{Output-format}]_{T1}.$$

 $\mathbf{T}_{T2}$  refers to the template for generating the second type of question. We define  $\{\text{Questions}\}_{all}$  and  $\{\text{Questions}\}_{rec}$  as, respectively, the full set of questions used for annotation and the most recently annotated question (if available).  $\{\text{Annotations}\}_{all}$  and  $\{\text{Annotations}\}_{rec}$  are the complete set of annotations and the most recent one, respectively. Let  $[\text{Output-format}]_{T2}$  and  $[\text{Task}]_{T2}$  be a JSON-style output format for this question type and the textual description of the corresponding task. We define  $\mathbf{T}_{T2}$  as follows:

```
\begin{split} \mathbf{T}_{T2} &= [\text{Role}]_{DT} + \\ & \{ \text{Dataset} \} + \{ \text{Questions} \}_{all} + \{ \text{Annotations} \}_{all} + \\ & \{ \text{Questions} \}_{rec} + \{ \text{Annotations} \}_{rec} + [\text{Task}]_{T2} + \\ & [\text{Output-format}]_{T2}. \end{split}
```

Validating answers. There are two types of validations for the answers checking (1) if the answer correctly addresses the question, and (2) for contradictions across questions and answers. The template for the first type of validation ensures that the answer aligns with the intent of question. For this, we define  $[Task]_{T1}$  as:

<sup>2</sup>https://osf.io/e4fxu/?view\_only= 0ac12e1c4b8e4b99a5739935aa0ca97e

Table 1: **Study demographics.** A total of 8 domain experts from 4 different expertise (Molecular Biology, Accounting, Political Science, and Usable Security) participated in our user study. The participants were all above the age 18. For each domain, we tried our best to bring experts with similar years of experience in the field. If this was not feasible, we ensured that each expert had at least three years of professional experience in their field.

Topic	ID	Gender	Age	Education	Job Title	Experience	Assistance from
Molecular Biology	M1	Male	34	Ph.D.	Assistant Professor in Molecular Biology	10 yrs	Data Therapist
	M2	Female	26	Ph.D.	Ph.D. Student in Molecular Biology	6 yrs	Human interviewer
Accounting	A1 A2	Male Male	33 34	B.S. B.S.	Certified Public Accountant Certified Public Accountant	7 yrs 6 yrs	Data Therapist Human interviewer
Political Science	P1	Male	37	Ph.D.	Assistant Professor in Political Science	8 yrs	Data Therapist
	P2	Male	36	Ph.D.	Research Professor in Political Science	8 yrs	Human interviewer
Usable Security	S1	Male	34	Ph.D.	Ph.D. Student in Usable Security	2.5 yrs	Data Therapist
	S2	Female	30	B.S.	M.S. Student in Usable Security	2.5 yrs	Human interviewer

 $[{
m Task}]_{V1}$ : Now, look at the answer and check if the answer makes sense, based on the question. If the answer makes no sense at all, then provide feedback to guide users to answer the question.

Then the  $T_{V1}$  becomes:

$$\begin{split} \mathbf{T}_{V1} &= [\text{Role}]_{DT} + [\text{Question}]_{test} + [\text{Answer}]_{test} + \\ &[\text{Task}]_{T1} + [\text{Output-format}]_{V1}. \end{split}$$

The template for the second type of validation  $(T_{V2})$  is constructed in a similar manner, comparing existing annotations with the most recently updated one. Additionally, we implement auxiliary text processing functions, such as text summarization. All prompt templates and further explanations, are provided in our supplementary materials.

## 4.8 Implementation

We implement Data Therapist as a client-server web application. The front-end of the system is implemented using JavaScript and React framework, where it communicates with the Python Flask server using axios. The backend server was hosted on Amazon Web Services (AWS). As the LLM, we use the API provided by OpenAI. In particular, we use o1 for initial question generation to leverage its reasoning capability, and GPT-40 for all other LLM calls.

# 5 USER STUDY

We are interested in understanding how a fully automated system such as Data Therapist compares to a human visualization expert when eliciting domain knowledge from subject matter experts. To be able to make this comparison, we conducted a user study to collect domain knowledge using both Data Therapist as well as a human assistant from two different experts working on the same topic. In this setup, the domain knowledge collected by the human interviewer is regarded as a gold standard and our goal is to see how the Data Therapist compares.

The study was organized into two sessions, each around 2 hours, comprising three parts in total. Our experiment was approved by the national institutional review board of South Korea. Below we provide details of the user study.

## 5.1 Participants

We recruited a pair of experts for four different domains, yielding a total of 8 participants. We screened participants to have at least an undergraduate degree, at least two years of research or practical experience in the field, fluency in reading, writing, and speaking English, and, most importantly, comparable career experience for both experts in a given domain. By applying these criteria, we selected pairs of participants in four domains: Molecular Biology, Accounting, Political Science, and Usable Security. Each participant received a \$20 gift card as compensation upon completing the study. The demographics of the participants are shown in Table 1.

## 5.2 Task

We compared the effectiveness of the Data Therapist system with that of a human interviewer. For each pair of domain experts, our experiment asked one expert ( HI) to collaborate with a human interviewer (the study administrator), while the other expert ( TDT) exclusively used the Data Therapist system. TDT participants were free to add annotations at their discretion and could also decide whether or not to seek help or respond to questions generated by Data Therapist. Each participant was given approximately 45 minutes to annotate the dataset.

Prior to the experiment, we asked each pair of domain experts to together select a type of data, while familiar to both, would be difficult for non-experts to understand. We asked for a small, manageable example tabular dataset (not more than 10,000 rows and 20 columns). After a brief email exchange, they chose the following datasets.

- Molecular Biology: A dataset that aims to understand the impact of arm selection in analyzing gene mutations.
- Accounting: A cash flow Statement for a resort company.
- Political Science: The Political Terror Scale dataset.<sup>3</sup>
- Usable Security: Computer Network Traffic dataset.<sup>4</sup>

#### 5.3 Procedure

We conducted the study entirely online through video teleconferencing in individual sessions. The study consisted of two sessions with three phases: (1) preparation and instruction, (2) annotation, and (3) domain knowledge assessment. Participants completed Phases 1 and 2 in the first session and Phase 3 in the second session.

Phase 1: Preparation and Instruction. The instructor first collected each participant's information via the consent form. After starting each session, the instructor outlined the research problem and objectives, clarifying expectations for each participant. For DT participants, we explained all of Data Therapist's features and ensured that they could use all the tools effectively. For HI participants, we presented the dataset in a Google Spreadsheet and asked the participant to enter annotations directly there. We also requested that any annotations generated from working with the human assistant include '[H]' at the end. This phase took approximately 15 to 20 minutes.

Phase 2: Annotation. All participants were given 45 minutes to annotate and describe the dataset. DT participants began by creating annotations independently. If they encountered difficulties or saw a question they wished to answer, they did so and reviewed any answered questions. They were also able to generate visualizations, especially for large datasets (e.g., in usable security) to identify patterns.

Liparticipants started by adding annotations they believed were important. We asked them to explicitly write responses and annotations using their keyboard rather than verbally—which might be more natural and similar to a typical domain expert interview—to ensure parity with the DT participants. If they became stuck or needed help, or if no activity occurred for over 20 seconds, the human interviewer provided feedback on missing or insufficient explanations. Participants could

<sup>3</sup>https://www.politicalterrorscale.org/Data/Download.html <sup>4</sup>https://www.kaggle.com/datasets/crawford/ computer-network-traffic/data

Table 2: **Annotation patterns.** Below, we present annotation statistics from the four domain groups that participated in our study: Accounting, Molecular Biology (Mole. Biology), Political Science (Political Sci.), and Usable Security. The second row shows the number of annotations generated by participants who collaborated with a human interviewer, including the total number of annotations (Total annotations) and a breakdown of those created with and without interviewer assistance (Anno. w/ human inter., Anno. w/o human inter.). The third row shows annotations from participants who used Data Therapist, including the total number, those created without the system (Anno. w/o Data Therapist), and those generated in response to LLM-generated questions (Ans. from q. by LLMs) or predefined questions (Ans. from q. by predef.). The bars and percentages in parentheses indicate each category's proportion relative to the total annotations per participant.

	Metric	Accounting	Molecular Biology	Political Science	Usable Security	Total
w/ human interviewer	Total annotations Anno. w/o human inter. Anno. w/ human inter.	24 12 (50%) 12 (50%)	21 12 (57%) 9 (43%)	33 10 (30%) 23 (70%)	15 10 (67%) 5 (33%)	93 44 (47%) 49 (53%)
w/ Data Therapist	Total annotations Anno. w/o DT Ans. from q. by LLMs Ans from q. by predef.	24 7 (29%) 15 (63%) 2 (8%)	27 5 (19%) 12 (44%) 10 (37%)	34 19 (56%) 10 (29%) 5 (15%)	24 9 (38%) 10 (42%) 5 (21%)	109 40 (37%) 47 (43%) 22 (20%)

Table 3: Informativeness and difficulty of annotations. We ask the domain experts in Accounting (Acct.), Molecular Biology (Mol. Bio.), Political Science (Pol. Sci.), and Usable Security (Usab. Sec.) to evaluate their own annotations as well as those from the other expert with different experiment conditions. We let them assess from two perspectives: informativeness (the degree of which the annotation helps explain the data), and difficulty (the degree to which the annotation is difficult). The detailed evaluation criteria are described in § 6.2.

		Informativeness				Difficulty					
		Acct.	Mol. Bio.	Pol. Sci.	Usab. Sec.	Mean	Acct.	Mol. Bio.	Pol. Sci.	Usab. Sec.	Mean
w/ human	Anno. w/o human inter.	3.59	3.55	3.20	3.55	3.47	2.75	2.90	2.18	1.90	2.43
interviewer	Anno. w/ human inter.	3.63	3.70	3.82	3.80	3.74	2.59	2.63	2.25	1.90	2.34
w/ Data Therapist	Anno. w/o DT	3.36	4.00	2.77	4.31	3.61	2.07	3.50	2.42	1.78	2.57
	Ans. from q. by LLMs	2.54	3.42	2.75	3.72	3.11	2.10	2.96	2.15	1.65	2.22
	Ans. from q. by predef.	1.00	2.56	1.90	3.40	2.22	1.00	3.00	1.80	1.60	1.85

accept this feedback and update their annotations or ignore it if they deemed it unnecessary. In both conditions, participants were notified when there were 10 minutes and 5 minutes remaining.

Phase 3: Design knowledge assessment. We conducted the second session at least a day after the first session ended. Here participants evaluated both their own annotations and those made by others, whether through Data Therapist or human assistance, using a 1 to 5 scale. We asked participants to assess based on two criteria: the annotation's **informativeness** in describing the dataset and its **difficulty**. The information about ratings are explained in § 6.2. Afterwards, we asked participants to compare annotations created from ChatGPT-based questions against those inspired by the human assistant. We concluded the session with an open-ended discussion, allowing participants to freely share their impressions and insights gained from this study.

## 5.4 Apparatus

The interview and experiment were conducted using the participants' own computers. To make the environmental setting consistent and convenient, we recommended the participants to use Google Chrome, and zoom out to 70% in a  $1920 \times 1080$  resolution display.

## 5.5 Data Collection

We collected the following data. First, we gathered participant demographics via an online form. With their permission, we recorded and filmed each participant's experimental session. We also compiled information on the questions they used and the annotations they added to the dataset. In addition, during the second session, we gathered the scores provided by participants in their evaluations.

#### 6 RESULTS

We report on the results from the qualitative experiment. We first describe the basic annotation patterns from participants (§ 6.1). We then present the evaluations of annotations from domain experts (§ 6.2). Afterwards, we also assess the coverage of the annotations (§ 6.3). We discuss the takeaways in the next section (see § 7.1).

#### 6.1 Usage Patterns

On average, DT participants produced four more annotations per group than HI participants, suggesting that the absence of direct human contact can slightly increase annotation output. From both DT and HI participants, we observed no consistent pattern in how much they relied on assistance when generating annotations. HI participants typically produced around ten annotations without human interviewers with little variance, whereas the numbers varied more when making annotations with human interviewers, ranging from 33% to 70% of all annotations. DT participants on average created 37% of their annotations without help from the tool, but this figure fluctuated considerably per participant. We think that this highlights the importance of flexible support systems in annotation tools. Furthermore, we observed that participants working without any assistance, whether DT or HI, tended to focus primarily on specific data, with few general annotations.

In three out of four cases among DT participants, annotations prompted by question-answering interactions exceeded the number of direct annotations, whether general or data-specific. Among the DT annotations, approximately 60% of the questions were produced by the LLM, while the remaining 40% were drawn from the predefined questions by Gebru et al. [11]. Although the system initially suggested an equal mix of LLM-generated and predefined questions, participants often removed the predefined ones. Indeed, most participants noted that "the predefined questions were not particularly relevant to what they were trying to do."

# 6.2 Annotation Evaluation

At the end of the experiment, we let the participants evaluate the difficulty and informativeness of annotations. Table 3 shows the results. The goal of this metric is twofold: (1) to understand the level of difficulty perceived by humans, and (2) to examine how experts and LLMs differ in their perception of the "lay audience." We define "level of difficulty" as the degree of expertise required—based on domain experts' judgment—to comprehend the information in each annotation. We explicitly outlined the scoring to participants as follows: a score of 1 refers to content that a lay audience can easily understand, a score of

Table 4: **Coverage of annotations.** We present a brief summary of the themes covered by different annotations from 4 groups of participants. We also analyze annotations by shared themes, and themes exclusive in human-assisted and Data Therapist-assisted participants. We present the number of annotations and answers (# anno.) that correspond to the theme, as well as the contents (Contents) of the themes. Below the number of annotations, we present the number of annotations assisted by human interviewers in HI, presented as H and by LLMs and predefined questions [11] in DT, presented as LLM and Pre. 6LLM means that 6 questions are assisted by LLMs.

	Accounting		Molecular Biology		Poli	tical Science	Usable Security	
	# anno.	Contents	# anno.	Contents	# anno.	Contents	# anno.	Contents
Shared themes	28 (8LLM, 6H)	Understanding cash flow mechanism	8 (2LLM, 2H)	Understanding key variables	13 (4LLM, 2H)	Understanding key variables	21 (6LLM, 2H)	Anomaly detection in traffic
Themes on participants w/ human assist. only	11 (2H)	Understanding key features	22 (10H)	Clarification of difficult definitions	25 (9H)	Understanding key variables	11 (3H)	Details on specific data instances
Themes on participants w/ DT only.	2 (2Pre)	Metadata explanation	19 (10LLM, 9Pre)	High-level data analysis	33 (8LLM)	Real-world data facts about dataset	7 (2LLM,5Pre)	High-level data analysis

3 refers to content that a lay audience would not understand but any domain expert would, and a score of 5 refers to content so specialized that only a select subset of domain experts can grasp it.

We find that the difficulty of most annotations without human assistance lie between 2 and 3. This implies that the annotations contain not only information specific to domain experts but also elements that could be considered common knowledge, though still relevant to the topic. Another noteworthy point is that on average, human-assisted annotations from HI and LLM-assisted annotations from DT tend to be slightly lower (4% and 13% less, respectively) in difficulty than those produced solely by human annotators. This suggests that experts may sometimes overestimate the knowledge level of a lay audience.

We also evaluated each annotation based on its informativeness to assess the value of the information it provides. To quantify this, we defined a difficulty metric ranging from 1 to 5. A score of 1 indicates information that is commonly known even to a lay audience. A score of 3 represents information that is generally unknown to laypeople but well-understood by domain experts. A score of 5 denotes highly specialized information that is known only to specific subgroups of domain experts. In both annotations involving a human assistant and those involving Data Therapist, human-provided annotations remained fairly consistent on average, scoring around 3.4 to 3.6. In annotations that included human assistance in **LHI**, we observed an approximately 8% increase in informativeness. Conversely, annotations created with LLM assistance in BDT were about 14% less informative than those generated solely by humans. We want to note that in comparison with scores on difficulty, the scores in informativeness exhibit larger variance in the scores. We believe this is because assessing the difficulty level of an annotation is a relatively objective task, reflected by an average standard deviation of 0.43 across all difficulty scores. On the other hand, evaluating informativeness is more subjective, as indicated by a higher standard deviation of 0.92. Experts may hold differing opinions on what constitutes 'informativeness' in the context of a given data point, leading to greater variability.

Finally, the predefined questions from Gebru et al. [11] showed lower informativeness scores and lower difficulty levels compared to questions generated through human or LLM-based assistance.

## 6.3 Coverage of Contents

Table 4 summarizes the results of the coverage conducted by participants who used Data Therapist and who received human assistance while conducting the task. We classify each participant group's annotations and question-answer sets into three categories: (1) shared themes, (2) themes exclusive to HI participants, and (3) themes exclusive to DT participants. Also, we summarize the key content of the annotations. Furthermore, we also provide information about how many are questions assisted by humans, LLMs, and predefined ones.

From the data, we see that shared themes appear across annotations in all four groups. One notable common point is understanding the mechanism captured in the dataset. This theme is especially evident in Accounting and Usable Security. In Accounting, 14 out of 28 annotations belonging to the shared theme came from participants assisted by both LLMs and human experts, concentrating on the mechanism behind cash flow. In Usable Security, 8 out of 21 annotations in the shared

theme are focused on detecting anomalies in network traffic. This indicates that participants consistently prioritize grasping the mechanism behind the dataset. Another common point emerges around identifying key variables within the dataset. In Molecular Biology, 4 out of 8 annotations belong to shared theme, and in Political Science, 6 out of 13 annotations are those assisted by both LLMs and human experts that contributed to understanding key variables. These two examples indicate that Data Therapist is capable of eliciting knowledge that humans think is important and can help augment their explanations.

However, we observe differences in tendencies between humanassisted and Data Therapist-assisted annotations. This can be observed by analyzing the themes exclusively in an DT or HI participants (see Table 4). For example, human-based annotations generally focused on clarifying unfamiliar features or definitions. In contrast, participants who used Data Therapist to augment their dataset tended to emphasize data-related details (e.g., who is responsible for maintaining the dataset?) and pose higher-level questions (e.g., what insights can be drawn from...?) about the dataset.

#### 7 DISCUSSION

Below we summarize the takeaways (§ 7.1) and limitations (§ 7.2) for Data Therapist.

# 7.1 When LLMs Ask Like Humans, and When They Don't

In our study, we compare the annotations generated through human assistance with those produced by LLMs via question-based prompts. One of the objectives in the experiment was to uncover both the aspects in which LLMs successfully emulate human-like behavior and the areas where they still fall short. To this end, we assess these differences.

We find that LLMs can generate questions in a manner similar to humans, particularly when it comes to understanding the underlying mechanisms within a dataset. For example, LLMs effectively support the interpretation of cash flow structures in accounting data and help identify anomalies in network traffic data within the domain of usable security. One possible explanation for this strength is that LLMs may have previously encountered similar datasets used by our participants. In fact, several participants made similar comments. A2, a certified accountant, remarked, "Sometimes ChatGPT asks are as though it knows about Accounting to some degree."

However, LLMs face challenges in accurately gauging the knowledge level of a lay audience. We can infer this from cases where sometimes LLMs overlook key variables that users may not understand. In domains such as Molecular Biology or Political Science, human assistants tend to probe further when encountering variables that remain unclear to the user. In contrast, LLM-generated questions focus more on asking questions that leads to high-level data analysis or inquire further about real-world datasets related to the given data, rather than addressing foundational gaps in user understanding.

While Data Therapist is clearly no replacement for a human interviewer, we suggest that Data Therapist be used as a simulacrum [34] to support the initial stages of data collection and screening to bridge the gap between real-world context and the data. In this role, it can serve as a starting point for conducting broad, general surveys, enabling human experts to concentrate more effectively on domain-specific details.

# 7.2 Limitations

Due to the constraints of our experimental design, participants were given a fixed duration of 45 minutes to generate annotations. However, in real-world visualization design settings, it is often unrealistic to expect that one can fully understand all labels and contextual information related to a dataset within a limited timeframe. As such, our results do not capture the potential effects of extended annotation time. Exploring how performance or interaction patterns evolve with longer engagement remains an avenue for future work.

#### 8 Conclusion

We presented Data Therapist, a structured mixed-initiative interaction approach to question-answering and user-driven annotation for eliciting domain knowledge about a dataset. In doing so, we developed the Data Therapist implementation as a practical web-based tool. We conducted a user study involving four groups, each consisting of two domain experts with similar expertise. Our findings indicate that while LLM-generated questions are effective in helping users understand the complex mechanisms embedded in the dataset, accurately estimating the knowledge level of a lay audience remains an interesting avenue for future work.

## REFERENCES

- [1] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen, E. Chu, J. H. Clark, L. E. Shafey, Y. Huang, K. Meier-Hellstern, G. Mishra, E. Moreira, M. Omernick, K. Robinson, S. Ruder, Y. Tay, K. Xiao, Y. Xu, and Y. Z. et al. PaLM 2 technical report. *CoRR*, abs/2305.10403, 93 pages, 2023. doi: 10.48550/arXiv.2305.10403
- [2] A. Bigelow, S. Drucker, D. Fisher, and M. Meyer. Reflections on how designers design with data. In *Proceedings of the ACM Conference on Advanced Visual Interfaces*, p. 17–24. ACM, New York, NY, USA, 2014. doi: 10.1145/2598153.2598175
- [3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901. Curran Associates, Inc., Red Hook, NY, USA, 2020. 3
- [4] A. Burns, C. Lee, T. On, C. Xiong, E. Peck, and N. Mahyar. From invisible to visible: Impacts of metadata in communicative data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 30(7):3427–3443, 2024. doi: 10.1109/TVCG.2022.3231716 2, 3
- [5] A. Burns, T. On, C. Lee, R. Shapiro, C. Xiong, and N. Mahyar. Making the invisible visible: Risks and benefits of disclosing metadata in visualization. In *Proceedings of the IEEE Workshop on Visualization for Social Good*, pp. 11–15, 2021. doi: 10.1109/VIS4Good54225.2021.00008 2, 3
- [6] Q. Chen, F. Sun, X. Xu, Z. Chen, J. Wang, and N. Cao. VizLinter: A linter and fixer framework for data visualization. *IEEE Transactions* on Visualization and Computer Graphics, 28(1):206–216, 2022. doi: 10. 1109/TVCG.2021.3114804
- [7] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing. Vicuna: An open-source chatbot impressing GPT-4 with 90%\* ChatGPT quality, March 2023. 3
- [8] J. Choi, C. Oh, Y.-S. Kim, and N. W. Kim. VisLab: Enabling visualization designers to gather empirically informed design feedback. In *Proceedings* of the ACM Conference on Human Factors in Computing Systems, pp. 813:1–813:18. ACM, New York, NY, USA, 2023. doi: 10.1145/3544548. 3581132.3
- [9] V. Dibia. LIDA: A tool for automatic generation of grammar-agnostic visualizations and infographics using large language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 113–126. Association for Computational Linguistics, Toronto, Canada, July 2023. doi: 10.18653/v1/2023.acl-demo.11 3
- [10] D. Ganguli and et al. Predictability and surprise in large generative models. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, pp. 1747–1764. ACM, New York, NY, USA, 2022. doi: 10. 1145/3531146.3533229
- [11] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. doi: 10.1145/3458723 2, 3, 6, 8, 9

- [12] S. Goodwin, J. Dykes, S. Jones, I. Dillingham, G. Dove, A. Duffy, A. Kachkaev, A. Slingsby, and J. Wood. Creative user-centered visualization design for energy analysts and modelers. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2516–2525, 2013. doi: 10. 1109/TVCG.2013.145
- [13] L. Grammel, M. Tory, and M.-A. Storey. How information visualization novices construct visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):943–952, 2010. doi: 10.1109/TVCG.2010.164
- [14] M. A. Hearst, J. F. Allen, E. Horvitz, and C. I. Guinn. Trends & controversies: Mixed-initiative interaction. *IEEE Intell. Syst.*, 14(5):14–24, 1999. doi: 10.1109/5254.796083 2, 3, 4
- [15] B. Heinzerling and K. Inui. Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries. In Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics, pp. 1772–1791. Association for Computational Linguistics, Apr. 2021. doi: 10.18653/v1/2021.eacl-main.153 3
- [16] A. K. Hopkins, M. Correll, and A. Satyanarayan. VisuaLint: Sketchy in situ annotations of chart construction errors. *Computer Graphics Forum*, 39(3):219–228, 2020. doi: 10.1111/cgf.13975 3
- [17] E. Horvitz. Principles of mixed-initiative user interfaces. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pp. 159–166, 1999. doi: 10.1145/302979.303030 2, 3, 4
- [18] K. Hu, M. A. Bakker, S. Li, T. Kraska, and C. Hidalgo. VizML: A machine learning approach to visualization recommendation. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, p. 128:1–128:12. ACM, New York, NY, USA, 2019. doi: 10.1145/3290605. 3300358
- [19] A. Key, B. Howe, D. Perry, and C. Aragon. VizDeck: Self-organizing dashboards for visual analytics. In *Proceedings of the ACM Conference* on *Management of Data*, p. 681–684. ACM, New York, NY, USA, 2012. doi: 10.1145/2213836.2213931 3
- [20] G. Kindlmann and C. Scheidegger. An algebraic process for visualization design. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2181–2190, 2014. doi: 10.1109/TVCG.2014.2346325
- [21] H. Lin, D. Akbaba, M. Meyer, and A. Lex. Data hunches: Incorporating personal knowledge into visualizations. *IEEE Transactions on Visualiza*tion and Computer Graphics, 29(1):504–514, 2023. doi: 10.1109/TVCG. 2022.3209451 2
- [22] H. Lin, M. Lisnic, D. Akbaba, M. Meyer, and A. Lex. Here's what you need to know about my data: Exploring expert knowledge's role in data analysis. 2023. 2, 4
- [23] Y. Luo, X. Qin, N. Tang, and G. Li. DeepEye: Towards automatic data visualization. In *Proceedings of the IEEE International Conference on Data Engineering*, pp. 101–112. IEEE, Piscataway, NJ, USA, 2018. doi: 10.1109/ICDE.2018.00019
- [24] J. Mackinlay. Automating the design of graphical presentations of relational information. ACM Transactions on Graphics, 5(2):110–141, 1986. doi: 10.1145/22949.22950 3
- [25] J. Mackinlay, P. Hanrahan, and C. Stolte. Show me: Automatic presentation for visual analysis. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1137–1144, 2007. doi: 10.1109/TVCG.2007.70594
- [26] N. Makhija, M. Jain, N. Tziavelis, L. D. Rocco, S. D. Bartolomeo, and C. Dunne. Loch Prospector: Metadata visualization for lakes of open data. In *Proceedings of the IEEE Visualization Conference*, pp. 126– 130. IEEE Computer Society, Los Alamitos, CA, USA, 2020. doi: 10. 1109/VIS47514.2020.00032
- [27] D. Masson, S. Malacria, G. Casiez, and D. Vogel. Charagraph: Interactive generation of charts for realtime annotation of data-rich paragraphs. In Proceedings of the ACM Conference on Human Factors in Computing Systems, pp. 146:1–146:18. ACM, New York, NY, USA, 2023. doi: 10. 1145/3544548.3581091
- [28] A. McNutt, G. Kindlmann, and M. Correll. Surfacing visualization mirages. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, p. 1–16. ACM, New York, NY, USA, 2020. doi: 10. 1145/3313831.3376420 3
- [29] D. Moritz, C. Wang, G. L. Nelson, H. Lin, A. M. Smith, B. Howe, and J. Heer. Formalizing visualization design knowledge as constraints: Actionable and extensible models in Draco. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):438–448, 2019. doi: 10.1109/TVCG. 2018.2865240 3
- [30] M. J. Muller and S. Kuhn. Participatory design. Commun. ACM, 36(6):24–28, 5 pages, June 1993. doi: 10.1145/153571.255960 1, 2

- [31] T. Munzner. A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):921–928, 2009. doi: 10.1109/TVCG.2009.111 2
- [32] OpenAI. ChatGPT API Documentation. https://platform.openai. com/docs/api-reference/introduction. Accessed: 2025-03-15. 6
- [33] OpenAI. GPT-4 technical report. CoRR, abs/2303.08774, 100 pages, 2023. doi: 10.48550/arXiv.2303.08774 3
- [34] J. S. Park, L. Popowski, C. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the ACM Symposium on User Interface Software and Technology*, article no. 74, 18 pages. Association for Computing Machinery, New York, NY, USA, 2022. doi: 10.1145/3526113.3545616 5, 0
- [35] M. Sedlmair, M. Meyer, and T. Munzner. Design study methodology: Reflections from the trenches and the stacks. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2431–2440, 2012. doi: 10. 1109/TVCG.2012.213 1, 2, 3
- [36] S. Shin, S. Chung, S. Hong, and N. Elmqvist. A Scanner Deeply: Predicting gaze heatmaps on visualizations using crowdsourced eye movement data. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):396–406, 2023. doi: 10.1109/TVCG.2022.3209472 3
- [37] S. Shin, S. Hong, and N. Elmqvist. Perceptual Pat: A virtual human visual system for iterative visualization design. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pp. 811:1–811:17. ACM, New York, NY, USA, 2023. doi: 10.1145/3544548.3580974
- [38] A. Srinivasan, S. M. Drucker, A. Endert, and J. Stasko. Augmenting visualizations with interactive data facts to facilitate interpretation and communication. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):672–681, 2019. doi: 10.1109/TVCG.2018.2865145
- [39] Y. Tian, W. Cui, D. Deng, X. Yi, Y. Yang, H. Zhang, and Y. Wu. ChartGPT: Leveraging LLMs to generate charts from abstract natural language. *IEEE Transactions on Visualization and Computer Graphics*, 31(3):1731–1745, 2025. doi: 10.1109/TVCG.2024.3368621
- [40] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. LLaMA: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 27 pages, 2023. doi: 10.48550/arXiv. 2302.13971
- [41] J. van Wijk. The value of visualization. In *Proceedings of the IEEE Visualization Conference*, pp. 79–86, Oct 2005. doi: 10.1109/VISUAL. 2005.1532781
- [42] J. Walny, C. Frisson, M. West, D. Kosminsky, S. Knudsen, S. Carpendale, and W. Willett. Data changes everything: Challenges and opportunities in data visualization design handoff. *IEEE Transactions on Visualization* and Computer Graphics, 26(1):12–22, 2020. doi: 10.1109/TVCG.2019. 2934538 2
- [43] C. Wang, J. Thompson, and B. Lee. Data Formulator: AI-powered conceptdriven visualization authoring. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):1128–1138, 2024. doi: 10.1109/TVCG.2023. 3326585.3
- [44] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus. Emergent abilities of large language models. *CoRR*, abs/2206.07682, 30 pages, 2022. doi: 10.48550/arXiv.2206.07682
- [45] Y. L. Wong, K. Madhavan, and N. Elmqvist. Towards characterizing domain experts as a user group. In *Proceedings of the IEEE Evaluation* and Beyond - Methodological Approaches for Visualization Workshop, pp. 1–10. IEEE Computer Society, Los Alamitos, CA, USA, 2018. doi: 10. 1109/BELIV.2018.8634026 2
- [46] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):649–658, 2016. doi: 10.1109/TVCG.2015.2467191 3
- [47] K. Wongsuphasawat, Z. Qu, D. Moritz, R. Chang, F. Ouk, A. Anand, J. D. Mackinlay, B. Howe, and J. Heer. Voyager 2: Augmenting visual analysis with partial view specifications. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pp. 2648–2659. ACM, New York, NY, USA, 2017. doi: 10.1145/3025453.3025768
- [48] S. Xiao, S. Huang, Y. Lin, Y. Ye, and W. Zeng. Let the chart spark: Embedding semantic context into chart with text-to-image generative model. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):284–294, 2024. doi: 10.1109/TVCG.2023.3326913

- [49] Y. Ye, F. Sauer, K.-L. Ma, K. Aditya, and J. Chen. A user-centered design study in scientific visualization targeting domain experts. *IEEE Transactions on Visualization and Computer Graphics*, 26(6):2192–2203, 2020. doi: 10.1109/TVCG.2020.2970525 1, 2
- [50] M. Zhou, Q. Li, X. He, Y. Li, Y. Liu, W. Ji, S. Han, Y. Chen, D. Jiang, and D. Zhang. Table2charts: Recommending charts by learning shared table representations. In *Proceedings of the ACM Conference on Knowledge Discovery & Data Mining*, p. 2389–2399. ACM, New York, NY, USA, 2021. doi: 10.1145/3447548.3467279