Enhancing Data Literacy On-demand: LLMs as Guides for Novices in Chart Interpretation

Kiroong Choe, Chaerin Lee, Soohyun Lee, Jiwon Song, Aeri Cho, Nam Wook Kim, Jinwook Seo

Abstract—With the growing complexity and volume of data, visualizations have become more intricate, often requiring advanced techniques to convey insights. These complex charts are prevalent in everyday life, and individuals who lack knowledge in data visualization may find them challenging to understand. This paper investigates using Large Language Models (LLMs) to help users with low data literacy understand complex visualizations. While previous studies focus on text interactions with users, we noticed that visual cues are also critical for interpreting charts. We introduce an LLM application that supports both text and visual interaction for guiding chart interpretation. Our study with 26 participants revealed that the in-situ support effectively assisted users in interpreting charts and enhanced learning by addressing specific chart-related questions and encouraging further exploration. Visual communication allowed participants to convey their interests straightforwardly, eliminating the need for textual descriptions. However, the LLM assistance led users to engage less with the system, resulting in fewer insights from the visualizations. This suggests that users, particularly those with lower data literacy and motivation, may have over-relied on the LLM agent. We discuss opportunities for deploying LLMs to enhance visualization literacy while emphasizing the need for a balanced approach.

Index Terms—Visualization literacy, large language model, visual communication

1 Introduction

A broader spectrum of people are encountering visualizations in daily contexts, such as in digital news media and social media platforms. While simple charts such as bar and line graphs are the most commonly utilized [1], more intricate visualizations are also reaching a wider audience [2]. For example, Bloomberg annually publishes dozens of data stories, showcasing visualizations like bubble sets, treemaps, and Sankey diagrams, often enhanced with composite visualizations and custom encodings for better storytelling (Fig. 1).

The majority of people still have limited proficiency in understanding complex visualizations beyond basic charts [3], [4]. When trying to make sense of complex visualizations online, these individuals are left with little choice but to rely on accompanying explanatory text. However, the text may have its limitations in fully grasping the visual encodings and in accurately discerning the insights as demonstrated by the visualization. Efforts to bridge the visualization literacy gap have, until now, mainly focused on formal learning environments [5] or instructional tools [6], [7]. These solutions, though, are indirect and distant, lacking the immediate and contextual support needed when a novice wants to understand charts in real-world situations.

The recent advancements in large language models (LLMs) such as ChatGPT [9] present new opportunities for visualization novices to engage with and learn from visualizations found in real-world contexts. Several tools have used LLMs to develop natural language interfaces for data visualizations [10], [11]. However, these tools primarily concentrate on articulating user queries to generate charts and analyze data, rather than aiding individuals in better understanding the charts' meanings. Given their extensive knowledge base, LLMs possess considerable potential as personalized tutors. Yet, their effect on enhancing a novice's comprehension of charts is only beginning to be explored.

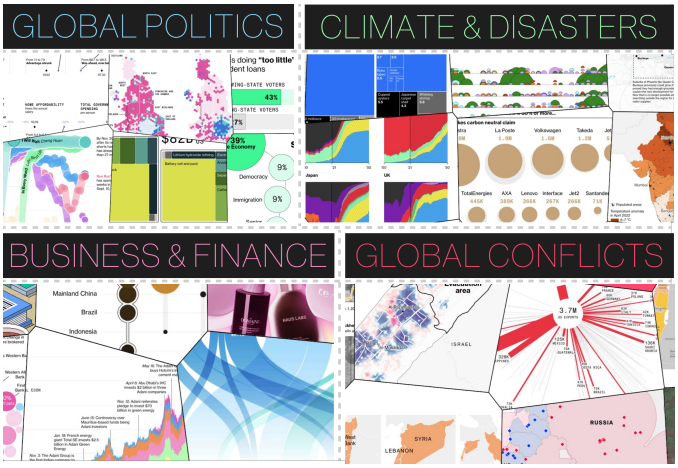


Fig. 1. In 2023, Bloomberg published visualization-based data storytelling articles on a wide range of public interest topics, such as health, politics, and the economy [8]. This figure is a collage of page contents, highlighting the visualization aspects.

The inherently visual nature of data representations makes visual cues crucial. However, most existing LLM applications focus predominantly on text-driven interactions. When interpreting charts, individuals often change their viewpoint and highlight specific visual elements, actively engaging with the visualization [12]. Active learning theory indicates that such cross-modal experiences enhance learning [13]. This suggests that the text-centric commu-

Kiroong Choe, Chaerin Lee, Soohyun Lee, Jiwon Song, Aeri Cho, and Jinwook Seo are with Seoul National University. e-mail: {krchoe, crlee, shlee, jwsong, archo}@hcil.snu.ac.kr, jseo@snu.ac.kr
Nam Wook Kim is with Boston College. e-mail: nam.wook.kim@bc.edu
(Corresponding author: Jinwook Seo.)

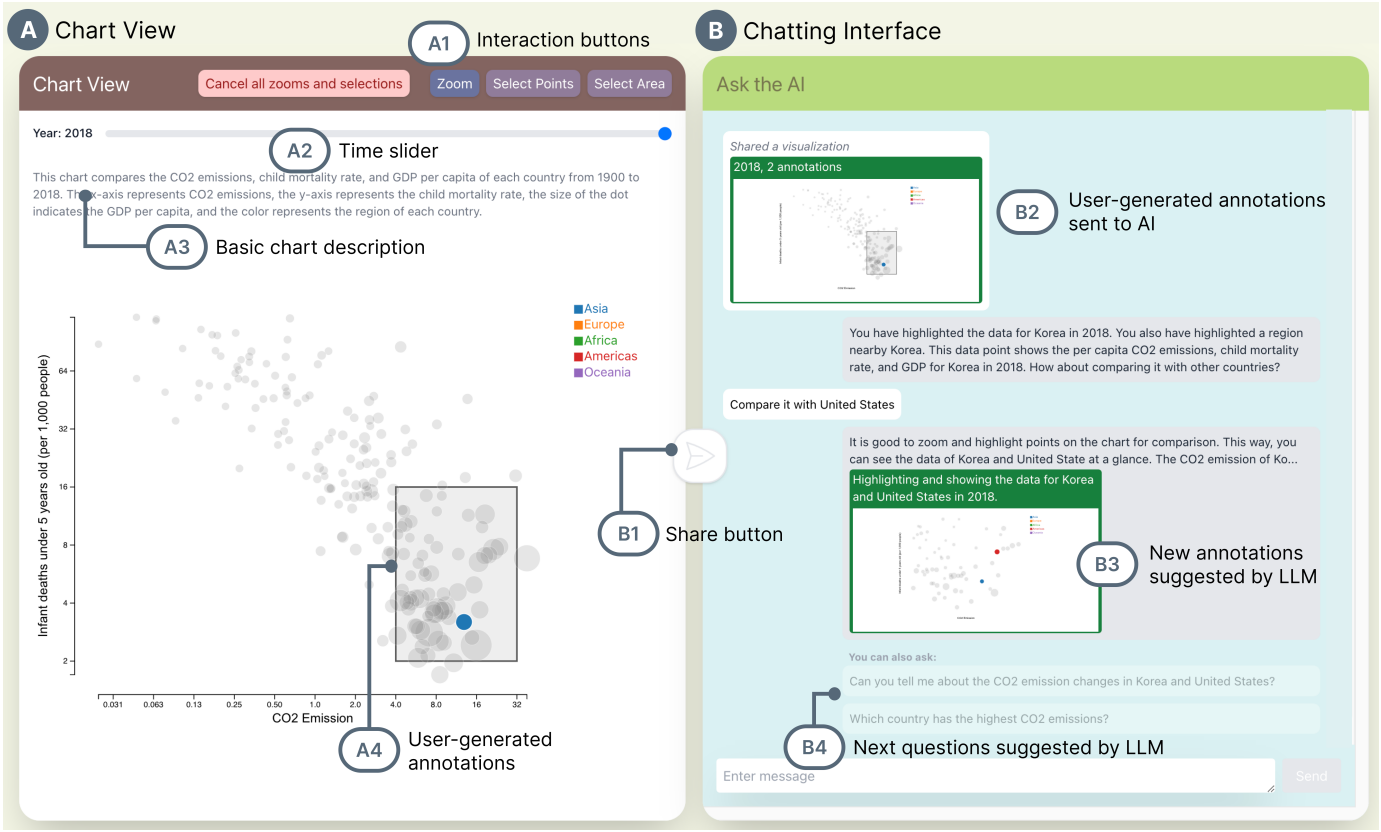


Fig. 2. The interface used in our experiment. The chart view (A) displays a chart alongside a basic caption (A3). Users can manipulate charts using interaction buttons (A1) and a time slider (A2), with their interactions represented as annotations (A4). In the chatting interface (B), the LLM provides assistance in chart comprehension. Users can ask a question or share a visualization (B1), which delivers the current chart annotations to the agent, displayed as an embedded snapshot (B2). The LLM agent may propose new annotations (B3) and suggest follow-up questions for further analysis (B4).

nication common to many LLM applications might pose challenges for novices engaging with visual data representations and ultimately degrade learning effectiveness. While cross-modal visual exploration has been proposed for data exploration [14], its impact on improving literacy in visualizations is still unexplored. This gap is particularly notable as we integrate AI into the learning domain, where issues like hallucinations [15] and ensuring user engagement [16], [17] are prevalent concerns.

We explored the design space of an LLM-based interface aimed at helping novices interpret and navigate unfamiliar visualizations. Here, we define a visualization novice as an individual who has not received formal education in visualization beyond basic statistical charts and seldom encounters visualizations in daily life. Our interface facilitates interaction through both textual questions and visual selections (such as zooming and selecting points and areas) on a chart. The LLM agent was also enabled to respond to questions with answers in both text and visual formats (Fig. 2). We conducted a within-subjects controlled experiment with 26 participants using this interface, where they explored three advanced chart types [3], [4]—scatterplots (with size and color encoding), treemaps, and parallel coordinate plots—which were mostly unfamiliar to them. For each chart exploration, participants received one of three types of assistance: all available text+vis LLM features, only text-based LLM features, or, as a control condition, a conventional web search engine without LLM support.

The results revealed a significant preference for the assistance provided by the LLM agent, largely due to its capability to handle chart-specific questions. Participants effectively utilized visual queries for various purposes and were influenced by the LLM agent's visual responses (Fig. 6). This visual communication was praised for reducing the effort needed to formulate questions, especially valuable when participants struggled to articulate their inquiries or found it burdensome to describe their analysis targets textually. Moreover, visual responses from the LLM enhanced the alignment between text-based discussions and visualizations, thereby fostering continued engagement with visualizations, and facilitated the visual tasks such as data point location and comparison.

However, we also noted the potential drawbacks of LLM-based situated support and visual communication aids. Despite the advanced features, participants derived fewer insights from chart exploration compared to traditional web search methods. Interestingly, participants believed they performed better with the LLM features, resulting in contrasting results.

Two distinct LLM usage patterns emerged from our observations. Participants who were less motivated and unfamiliar with data analysis, found the LLM's guidance immensely helpful in overcoming the confusion and uncertainty they faced during analysis. This group frequently

sought answers from the LLM for questions that could be easily answered by examining the chart, demonstrating a strong system dependence and reduced chart interaction. On the other hand, participants more experienced in data analysis utilized the LLM agent as a practical supplementary aid. They actively engaged with the charts manually and avoided unnecessary reliance on the LLM agent. These participants used the LLM for abstract, higher-level insights related to domain-specific and contextual knowledge beyond basic chart interpretation.

Our findings suggest that LLMs possess the potential to offer on-demand support for visualization novices encountering and seeking to interpret unfamiliar visualizations in daily contexts. Additionally, in such interactions with LLMs, visual communication can enhance textual communication by adding an extra layer of expressiveness. However, this comes with potential drawbacks; excessive reliance on LLMs could reduce the amount of engagement with visualizations and make the experience less fruitful. In our discussion, we align with the cognitive model of visualization novices [2] to understand these results. We also explore the potential of various annotation spaces to enhance visual communication and discuss how LLMs could expand their role beyond improving visualization literacy, such as by aiding in the creation and evaluation of visualizations.

2 Related Work

2.1 Visualization Novices and Education

Visualizations are becoming increasingly common in various public sectors such as healthcare and finance. While basic charts like pie and bar charts are the most frequently used [1], the public is also experiencing a growing exposure to advanced visualizations [2]. The majority of people still find complex charts such as treemaps and parallel coordinate plots (PCPs) unfamiliar and often struggle to interpret them compared to basic charts and graphs (e.g., pie charts, line graphs) [4], [3]. Lee et al. [2] investigated the challenges novices face when encountering unfamiliar visualizations. They emphasized the complexity of building accurate mental models for both understanding charts and interpreting data, a process of trial and error raising numerous questions at different stages.

Considering the difficulty of spontaneous engagement in such learning, efforts are being made to empower a broader demographic (e.g., those in remote, data-poor areas [18]) with the skills to read and comprehend visualizations. Research interests range from understanding basic visualizations [19], [20], to more advanced formats like treemaps [21], parallel coordinate plots [7], network visualizations [5], and interactive visualizations [22]. Despite the effort to establish clear objectives and research questions on visualization literacy education [23], traditional methods such as classrooms [24] and online courses [5] face limitations in cost and scalability.

Interactive learning tools have demonstrated their effectiveness in teaching visualizations [6], [21], presenting a potential solution for scalable and cost-effective education. However, these programs must be carefully tailored to both the chart type and the readers. This is because "visualization novices" can greatly vary across different groups [25]. For instance, while children may have a similar graphical perception to adults, their ability to decode visual representations is not as developed [26].

In this paper, we defined a visualization novice as an individual who has not received formal education in visualization beyond basic statistical charts and seldom encounters visualizations in daily life. Following this definition, we explored how adult novices learn advanced visualizations with the help of an LLM agent, seeking opportunities for developing on-demand literacy education programs for adults focused on unfamiliar visualizations.

2.2 LLMs in Visualization Education

Traditional courses present a higher barrier because individuals must actively seek out these courses and invest dedicated time and effort over a fixed period. Interactive learning tools are often tailored for very specific situations and chart types, which do not generalize well to the real-world visualizations people encounter in daily contexts.

Recently, large language models have demonstrated proficiency in a broad range of tasks. These tasks extend beyond typical natural language processing tasks such as summarization and translation [27], [28], to factual and reasoning tasks, as seen with OpenAI's ChatGPT [9]. The GPT-4 model has also reported notable performance in chart reasoning [29]. This highlights the potential of large language models to aid novices in understanding charts in everyday contexts. Yet, concerns such as hallucinations and confabulations [15] can emerge as we integrate LLMs into learning.

Research on human-AI collaboration, particularly in creative domains, has highlighted key principles for designing AI systems that effectively work with humans. Users prefer to lead and control their interactions with AI, choosing even lower-performing AI models (*e.g.,* [30], [31]) to maintain a sense of ownership [17], [16]. They find unexpected AI-generated results both inspiring [17], [32], [31], [33], [34] and useful as starting points [35], although these can be distracting for those with specific goals [17], [32], [34]. Expectations of AI systems vary, ranging from seeking help with mundane tasks [34], [36], desiring a partnership with a creative entity [32], [31], to using AI as a source of entertainment during tasks [34].

Our focus is on the application of LLMs in the context of learning visualizations, which may have different characteristics compared to creative domains. We investigated whether LLMs can be beneficial in this educational setting and identified any considerations or precautions that should be taken into account when designing educational systems incorporating LLMs.

2.3 Cross-modality in Understanding Visualization

Communicating on a visualization requires both visual(*e.g.,* pointing) and lexical(*e.g.,* describing) access. As there is an inherent gap between visual and lexical channels, significant research has aimed to bridge the gap. Efforts have been made to create visualizations from text [37], [38], [10], [11] or synchronize text and visuals through annotations [39]. Additionally, the accessibility domain has contributed to converting visualizations to text formats (*e.g.,* [40], [41]),

along with an understanding of various levels of description that ranges from basic chart encoding description to those of cognitive-level findings and domain-specific reflections [42].

Understanding and explaining visualizations involves a more complex, qualitative, and unpredictable process [43]. Such an iterative process involves switching between text and visual elements to explore the data [14], [44] and to communicate those findings [45], [39]. Yet, the specific interaction of text and visual communication in enhancing visualization literacy among novices remains unexplored.

We explored the distinct roles and effects of text and visual communication when learners aim to interact with an LLM with the goal of learning how to read visualizations and uncover inherent data insights. We investigated how text and visual modes interplay throughout the learning process.

3 Formative Study

Our objective is to probe the potential effectiveness of LLMs in supporting the understanding of advanced visualizations by novices. We conducted an iterative formative study with six visualization novices to make informed decisions regarding the specific objectives, tasks, and features of the experimental interface. Before participation, all participants were asked about their familiarity with visualizations. They all reported they had not encountered complex visualizations beyond basic bar and line charts and had not regularly engaged with any visualizations. In the study, participants were presented with unfamiliar visualizations and asked to interpret and articulate them. We additionally provided them with our prototype interface that enables communications with LLMs, allowing us to monitor their inquiries. The prototype interface was similar to our experiment interface in overall functionality, but it initially lacked several features, which were gradually added as the study progressed. For example, it utilized non-temporal datasets, leading to no timeline feature, and it did not provide LLM's suggested follow-up questions or visual responses.

Even though the visualizations were unfamiliar, participants formed an initial interpretation of the given charts without much struggle. However, their overall engagement with the chart was notably superficial. When given specific questions about charts to deepen their engagement, participants tended to focus solely on those questions, seeking answers with the least effort and overlooking the chart's details. The introduction of an LLM had a limited impact on this behavior. Some participants simply relayed the given questions to the LLM and accepted the answers without critique. One participant expressed a preference for obtaining direct answers from the LLM, bypassing personal interpretation of the charts.

Consequently, the primary aim of the main study shifted to promoting deep engagement with charts beyond just data interpretation. The focus was on encouraging participants to identify overarching patterns and relationships in the chart as well as integrating their insights and prior knowledge. To achieve this, participants were assigned tasks related to preparing an imaginary presentation. In this task, they were required to introduce the chart and highlight what they found interesting in it, thereby encouraging a more thorough and active exploration of the charts.

The majority of queries directed to the LLM were related to specific data details, such as questions about the maximum or minimum values in certain categories. Even with a prototype interface allowing for direct annotation on and reference to the chart, participants tended to convert their visual questions (*e.g., "Is the point in category A obscured by this cluster of points in category B?"*) into textual data queries that can indirectly answer the question (*e.g., "What is the minimum x and y value in the category A?"*. This observation underscored the need to enhance our LLM-based features to foster more visually oriented dialogue. As a result, both the user and the LLM agent gained the capability to make direct annotations on the chart, communicating their intentions in both textual and visual formats.

4 Interface Design for User Study

This section outlines how we engineered the user interface and LLM-based features for our experiment. We organize our description around three central components: (1) the chart view and associated interactions, (2) the chatting interface coupled with visual communication features, and (3) the engineering of prompts for the LLM, enabling it to serve as an assistant capable of both instructing in chart comprehension and handling chart annotations (Fig. 2).

4.1 Chart View and Interactions on Chart

The chart view displays a chart to the user with basic caption and interaction features. For chart interactions, we intentionally made the feature simple and straightforward to ensure a seamless user experience and focus on evaluating the LLM's efficiency in chart comprehension. There are three interaction buttons (Fig. 2-A1): "Zoom", "Select Points", and "Select Area." These tools allow users to zoom into smaller areas and highlight data points or rectangular regions through click-and-drag actions. The results of these user interactions are immediately reflected as annotations on the chart (Fig. 2-A4). For instance, selecting certain data points will cause other unselected points to fade, emphasizing the selected points.

Every chart selected for our study features a timeline slider (Fig. 2-A2), allowing users to select a specific year to visualize from a predefined range. In this way, we incorporated the temporal domain into the charts to broaden the exploration scope, thereby enhancing the potential for users to uncover personalized data insights.

4.2 Visual Communication with the LLM

The chatting interface implements an interface that resembles common messenger applications with message bubbles(Fig. 2-B). When a user sends a message, the LLM agent processes the request and offers an answer. Both the user and the LLM agent can communicate using visual annotations on the chart. Users can initiate this interaction either by sending a message or by clicking on the share button (Fig. 2-B1) even without an explicit prompt. The user's query is paired with current chart annotations, shown as a chart snapshot photo in the interface (Fig. 2-B2). The

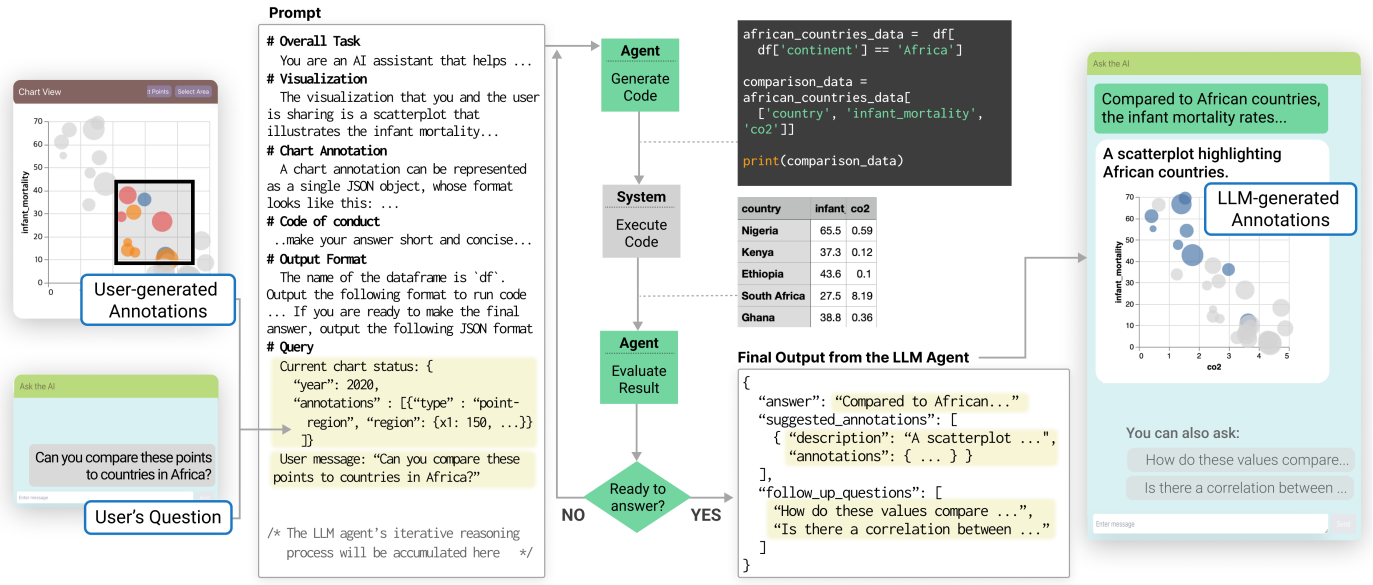


Fig. 3. The system pipeline for implementing visual communication with the LLM. The pipeline includes translating text and visual inputs into a JSON object and performing iterative code generation by the LLM agent. The final output consists of a text response and LLM-generated annotations

LLM agent can similarly respond with chart annotations in a format mirroring the user's, optionally embedding a chart snapshot photo in the response (Fig. 2-B3). If the user approves of the LLM's suggested chart annotation, they can click on the snapshot to apply it to the chart view on the left. Beyond its primary response, the LLM agent also recommends follow-up questions based on the prior messages(Fig. 2-B4), enhancing user interaction and engagement.

Fig. 3 illustrates our implementation of visual communication with the LLM. The user's text questions and visual annotations are translated into a JSON object following a predefined protocol and delivered to the LLM agent. To address the user's question, the LLM agent requires access to the chart data. The system enables this by making the LLM agent write code to query the database. The system then executes the code, observes the results, and provides these results back to the LLM agent. The iterative process of code generation, execution, and result evaluation continues until the LLM agent determines a final answer. The final output formats as a JSON object, containing the text answer and a set of recommended follow-up questions (each with text questions and visual annotations), which renders on the interface.

4.3 LLM Prompt Design

We formulated the prompt for the LLM based on the *"CSV agent"* template proposed by LangChain [46]. This approach implements the ReAct (Reasoning and Acting) framework [47] to enable the LLM agent to read a CSV data file using Python code. We enhanced and adapted the prompt to enable the agent to comprehend the chart, understand the annotations on it, memorize previous chat history, and respond with suggested questions and annotations in addition to an answer to the original question.

Our prompt consists of the following sections. The **overall task section** outlines the agent's role: assisting users in understanding unfamiliar charts. The **visualization section** details the chart type, encoding, axis, existence of a time slider, and other basic information like the presence of a tooltip, matching the level-1 description by prior work [42]. The **chart annotation section** explains the functionalities available to the user, such as setting the year, magnifying, and highlighting points and rectangular regions. It also details how agents receive this information in JSON format. The **code of conduct section** emphasizes the agent's need to produce concise and short answers, focusing on chart details first but also answering general questions based on common knowledge. The **output format section** defines an output structure that includes a textual answer, chart annotations in JSON, and a list of recommended questions. It also specifies an intermediate output format for use when the agent needs to read the raw chart database using Python code. The **chat history section** includes the last ten messages between the user and the agent. Lastly, the **query section** contains the actual user input to be answered.

In experimental conditions that do not permit visualization-mediated communication, the chart annotation section and any references to chart annotation in the output format section are omitted. We used the *"GPT-4-32k"* model for our experiment. Detailed information is provided in the supplementary material.

5 Experiment Design

5.1 Participants

We recruited participants through a local community platform accessible to residents of the local district, resulting in 26 participants (18 female, 8 male) with an average age of 28.0 ($\sigma = 6.6$), ranging from 22 to 54. Our recruitment survey featured questions on visualization familiarity, ensuring all participants selected were unfamiliar with visualizations. All 26 participants reported not having seen at least one of the three visualization types (*i.e.*, scatterplot, treemap, and

parallel coordinates plot); 15 (58%) had not seen any, and 6 (23%) had not seen two types. 16 participants (62%) had not encountered any visualizations in the last three months. 23 participants (88%) indicated they would depend entirely on external resources, such as web search results, to understand unfamiliar visualizations before attempting interpretation on their own.

5.2 Chart and Data

For the advanced visualization, which many participants might find unfamiliar, we chose three types: a scatterplot with size and color encoding, a treemap, and a parallel coordinate plot.

This selection was adapted from directly relevant prior work that utilized treemaps, parallel coordinates, and chord diagrams [48]. These three visualization types are not included in the K-12 curriculum, and their underlying data structures are complex. Another study involving 53 ordinary participants (i.e., museum visitors) also showed that these visualizations are less familiar and more challenging to interpret than basic charts and graphs, such as bar and line charts [4]. Building on the selection of these three visualizations, we made a single modification: replacing the chord diagram with a scatterplot. This decision was made to avoid the additional challenge of understanding network data, which extends beyond interpreting visual encoding.

When selecting specific datasets for the three chart types, we followed various criteria to ensure the suitability and efficacy of each chart type. The selected data should be straightforward and familiar, enabling effortless engagement even for individuals with limited experience in visualization. Additionally, the data should include a temporal domain in it, allowing users to explore several years using a time slider and build an integrated understanding across multiple charts.

Table 1 displays the final charts used in our experiment, along with example insights that participants could potentially derive from it. For the scatterplot, the chosen data is from the national economic indicators (GDP, fetal mortality rate, and CO2 emissions) from Gapminder, which highlights the significance of the correlation between several pairs of variables. This choice is informed by the data's compatibility with size and color coding, offering a clear insight into the relationships among the variables. The data for the treemap consists of the market capitalization and price changes of the top 200 stocks in the Korean national market, each associated with a specific sector. This data, inherently hierarchical and segmented, is ideal for a treemap, as it allows for meaningful comparisons between entities, with size and color coding reflecting distinct variables. The parallel coordinate plot, designed to represent multi-axial and inherently multivariate data, utilizes information on CO2 emission sources (e.g., from coals, from oils, from flaring, etc.) for each country.

5.3 Tasks and Conditions

The experiment adopted a within-subject design, wherein each participant underwent three distinct sessions. Each session was defined by a unique chart type—scatterplot, treemap, or parallel coordinate plot—and a specific form of assistance: web search, text-only LLM, or text+vis LLM. In the text+vis LLM condition, all features described in the Sec. 4 were all provided. In the text-only LLM condition, participants engaged exclusively in text-based communication for understanding the chart without the visualization-mediated communication features. For the web search condition, participants were not provided with LLM features and instead freely utilized their preferred search engine (e.g., Google). Participants divided the screen for both the chart and the web search window, ensuring constant, unhindered access to the search engine.

Among nine possible combinations from three chart types and three types of assistance, each participant randomly encountered three combinations in a balanced fashion, spanning all chart and assistance types. Utilizing an order-3 mutually orthogonal Latin square [49], we defined three sequences, each containing three combinations, to counterbalance the order effect of both conditions. Participants were randomly assigned to one of these three sequences.

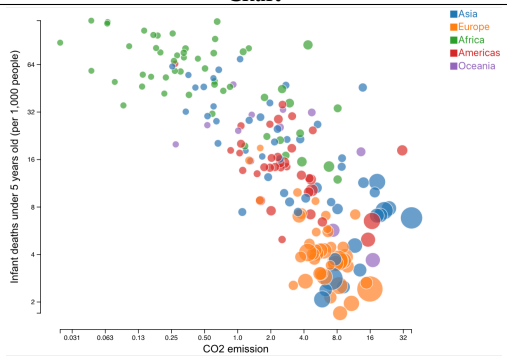
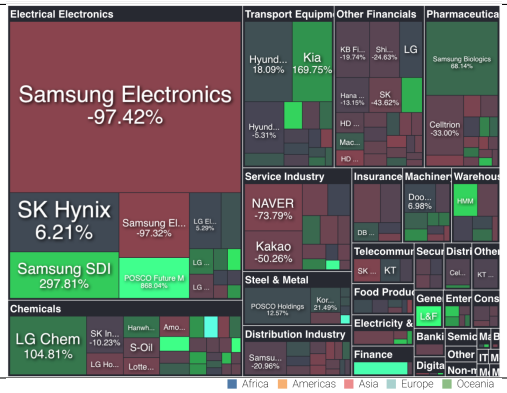
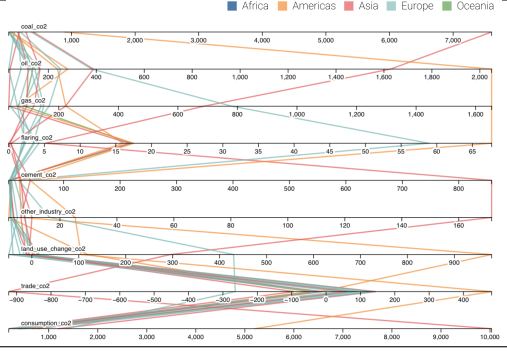
Fig. 4 illustrates the task design in our experiment. To avoid a problem-solving mindset—seeking answers to the given question with minimal effort and neglecting the chart's details, as discussed in Sec. 3—the task was designed to be a free exploration of the chart for the purpose of preparing for a team meeting. The following task prompt was given before every session to elicit active exploration and enhance motivation:

> You are in a video-creating company that uses charts to explain the economy. You have a team meeting on *(specific topic)*, and you must choose one chart on this topic and explain its contents and intriguing aspects to your team members. This chart was found on an online news media platform. Assume your team meeting is in 10 minutes, and use this time to explore the chart and prepare for the meeting.

We set an explicit 10-minute time limit to encourage participants to fully engage with the visualization within their assigned time, specifically aiming to deter them from exerting the minimum effort needed to complete the task and then disengaging. Additionally, the time limit helped keep the total duration of the experiment manageable to prevent participant fatigue.

During the **Exploration stage**, lasting up to 10 minutes, participants interacted with the chart and the corresponding assistance feature without any intervention from the experimenter. Following the exploration, participants entered the **Insight Articulation stage** where they articulated their understanding of the chart. In this stage, the interaction logs from the assistance feature were hidden, although the chart itself remained visible to the participants. In the **Interaction Articulation stage**, participants revisited and recounted their experiences with the assistance feature, examining their intentions and satisfaction from each transaction, with the interaction logs now made available. For instance, in scenarios where participants had chatted with the LLM, the chat history was hidden during the Insight Articulation stage, and then restored in the Interaction Articulation stage. They were instructed to recount from the beginning to the end of the chat, explaining their intention behind

TABLE 1
The scatterplot, treemap, and parallel coordinate plot used in our experiment. Each dataset is selected based on its relevance to generally familiar topics (economy, environment, and stock finances), and suitability to be represented by the corresponding chart type. Since each chart has a temporal domain, the chart shown is an example, representing one specific year. The table lists potential insights that can be gleaned from engaging with the charts.

Chart	Example insights to be found
	• Points symbolize countries, colored by continent, showing a negative correlation between CO2 emissions (x-axis) and fetal mortality rate (y-axis), questioning their direct relationship. • Observable regional variations include African countries generally at upper left (higher mortality, lower emissions) and European countries at lower right (lower mortality, higher emissions). Historical data indicates a significant shift for Asian countries towards lower mortality and higher emissions from the 1950s to the 2010s. • The size (GDP) hints at a confounding link. Higher GDP countries often have better medical infrastructure, lower fetal mortality rates, and higher CO2 emissions.
	• In 2013, many stock prices in green pointed to economic growth, contrasting with the red decline in 2018, worsening in 2023, potentially due to COVID-19, impacting the Korean economy. • Despite the downturn, the pharmaceutical sector displayed growth, highlighting its resilience or potential gain from the pandemic. • The service sector, especially IT, saw significant growth with Naver and Kakao emerging as major players by 2023.
	• China and the USA stand as major outliers with the highest CO2 emissions from coal and oil respectively. • The "flaring CO2" axis underscores Russia's substantial emissions, likely tied to significant natural gas production. • Most countries exhibit positive "exchange CO2," indicating energy importation, while Russia and China's negative values highlight their status as major global resource producers.

each question and assessing the usefulness of the answers received. In the text+vis LLM condition, when users had constructed chart annotations for a question, the intention behind such construction was also inquired. For the web search condition, participants recounted their search histories and explained the usefulness of their searches. Concluding each session, a **Debriefing stage** involved a semi-structured interview capturing participants' insights on the task difficulty, the utility of the assistance feature, and their overall experience. Participants also completed the NASA-TLX questionnaire [50] to assess cognitive load. The Insight Articulation, Interaction Articulation, and Debriefing stages held about 10 minutes in total.

After the completion of all three sessions, a **Final Interview** was conducted as a 20-minute semi-structured interview, wherein participants elaborated on how each assistance type uniquely contributed to their learning and compared between them. Finally, participants selected the most preferred assistance types based on their effectiveness in facilitating chart exploration.

5.4 Data Collection and Analysis

All data gathered from the experimental stages underwent both quantitative and qualitative analysis. For this purpose, the data were reorganized and subjected to a manual coding process as illustrated in Fig. 4.

5.4.1 *Interaction Log Coding*

Interaction logs from the Exploration stage of each session were compiled into an observation database. Each interaction between the user and the LLM agent is categorized as a transaction, which includes a user question, its origin (whether it was written by the user or was from the suggested questions of a previous message), the LLM agent's answer, and its suggested next questions. In the text+vis

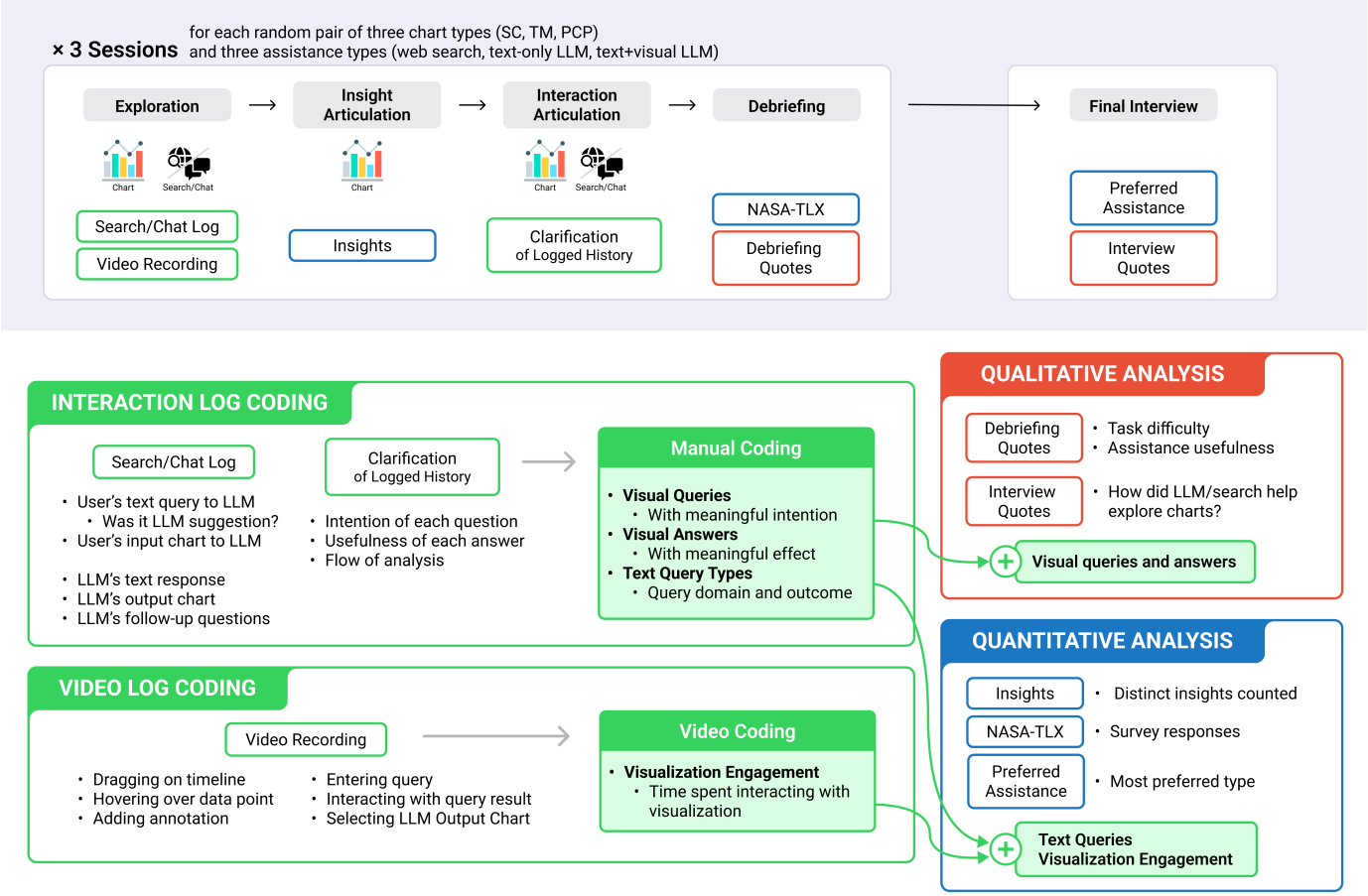


Fig. 4. The experiment followed a within-subject design. Each participant engaged in three sessions, each featuring a different chart and assistance type. Each session was divided into four stages, followed by a final interview after all sessions. Both quantitative and qualitative analyses were conducted on the collected data. Interaction logs were organized into a database and underwent a manual coding process. This process categorized user queries and LLM answers, which were also included in the analysis.

LLM condition, the transaction log also contains the user-input chart and LLM-output chart as snapshot images. The database is further augmented by aligning the user's interview quotes in the Interaction Articulation stage with each chat transaction. This offers in-depth insight into the intention behind each question, the perceived usefulness of each answer, and the overall analytical progression.

The combined observation database underwent a manual coding process. This process categorized each text and visual query from the user, and visual answer from LLM, into distinct semantic categories. Two independent coders initially explored the category through an open coding process, later discussing and finalizing the categories and definitions for each type. For instance, Table 2 outlines the classification of text queries based on their domain and outcome, which subsequently underwent quantitative analysis. Visual queries were classified based on the user's intended purpose, and the visual answers were categorized according to their impact on the user. Due to the relatively smaller number of identified visual queries and answers, these subsequently underwent qualitative analysis, which is detailed in the Result section.

TABLE 2
Categorization of text queries based on domain and outcome

Domain	
Narrow	Queries dealing exclusively with limited data points without any unspecified variables. (*e.g., "Tell me the 2023 CGV stock price"*)
Broad	Queries dealing with numerous data points with at least one unspecified variable. (*e.g., "Tell me the yearly stock prices of CGV"*)
Abstract	Queries encompassing the entire database or lacking specific domain indication. (*e.g., "Which is the country with the least economic development?"*)
Outcome	
Specific	Queries wherein the outcome format is predetermined and explicit. (*e.g., "What is the stock with the highest volatility in the service industry?"*)
Abstract	Queries where the outcome format remains ambiguous, potentially taking free-description forms. (*e.g., "How does this correlation differ by continent?"*)

5.4.2 Video Log Coding

To quantitatively measure the extent of participants' direct engagement with visualizations or assistance features, we conducted additional video coding. Initially, by coding the first five videos, we ascertained that user interactions could be distinctly categorized into engagements with the visualization on the left side of the interface and with the assistance feature on the right side. Interaction events such as clicking, dragging, hovering (to view tooltips), and typing

were used as criteria to segment each 10-minute exploration session into durations spent on these two types of activities. If a participant did not perform any action for more than 10 seconds, the duration thereafter was coded as time not attributed to either category until a new event occurred. Consequently, we quantified the total time spent interacting with either the visualization or the assistance feature during each 10-minute session. These quantifications were utilized for further quantitative analysis. The results of this video coding are provided in the supplementary material.

5.4.3 Quantitative analysis

In addition to the query type coded from the observation database, the quantitative analysis includes the number of insights, the NASA-TLX cognitive load scale, and the most favored assistance types. Insights are counted from the descriptions made in the Insight Articulation stage. While segmenting distinct insights from the textual description of the chart, we referred to the four-level model of semantic content [42], which defines multiple levels in the textual description of visualizations. This model categorizes insights from level 1 (elemental and encoded) to level 4 (contextual and domain-specific). Since we supplied the chart with captions explaining the data and basic encoding (classified as level 1 insight), we only counted levels 2 to 4 insights as a meaningful result of participants' chart exploration. The segmentation process was two-fold: insights were first categorized according to their levels, and then within each level, further distinctions identified the diverse insights present. Lastly, all numbers in levels 2 to 4 insights are totaled to obtain the final insight count, serving as an indicator of achievement for each session.

5.4.4 Qualitative analysis

In addition to examining the patterns and effects of both users' visual queries and LLMs' visual answers, the qualitative analysis involved reflexive thematic analysis of interview quotes from Debriefing sessions and the Final Interview.

6 Result

6.1 Situated Support through a Language Model

There was a significant difference in preference when asked for the type of assistance they found most beneficial ($\chi^2(2) = 10.23, p < .05$). Among 26 participants, only 3 preferred web search the most, 7 preferred text-only LLM, and a notable 16 preferred text+vis LLM.

Participants often found it less advantageous to search the web to understand the charts. As one participant, P3, noted, *"I first tried googling 'reading parallel coordinate plot,' but there weren't results that I could understand, so I tried to figure it out myself."* Facing the challenge of deciphering unfamiliar charts, the burden of formulating a search query added the complexity, as expressed by P9, *"I don't even know what this chart shows. [...] it is all unknown to me, and it says you should ask that unknown thing."* In situations where participants desired to inquire about specific, contextual questions from the chart, web search engines were found to be less relevant. For example, P25's web search on countries listed in the chart predominantly returned results related to Korea and East Asian countries, a bias introduced by the use of a Korean search engine. P14 wanted to understand the reason behind the apparent correlation between CO2 emissions and the child mortality rate shown in the chart. However, the search results primarily featured posts that focused on each attribute separately, without discussing the correlation between the two or exploring the general correlation in the topic.

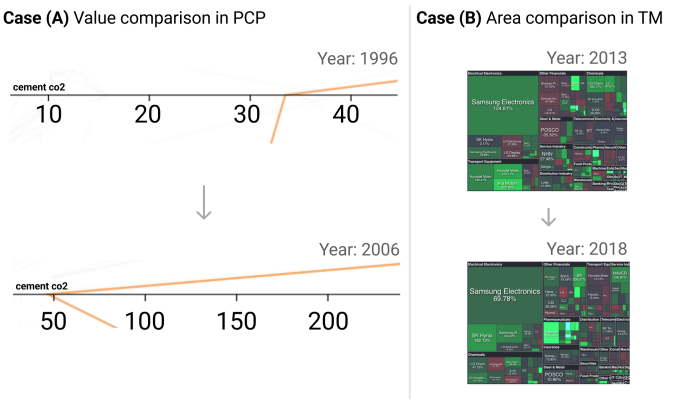


Fig. 5. Two exemplary cases from the experiment demonstrate the LLM agent's role in correcting users' misinterpretations of charts. Case (A): the agent clarified that a visual mark moving leftward actually depicted an increase in value due to a change in the axis domain. Case (B): the agent calculated and indicated that the non-red areas do not have significant differences, contrary to a participant's initial perceptions.

In contrast, LLM agents could answer specific, contextual questions, as shared by P13, *"We were looking at the same thing, so I could ask, 'why does it go to a negative value?'"*. LLM agents' support extended to various levels, offering background knowledge related to the chart's phenomenon (P14, P25) or providing summary statistics such as average, count, extrema, and correlations (P6, P20, P21). Many participants appreciated the LLM's ability to show only relevant results, compared to the diverse and often unrelated results from web searches. Additionally, the conversational nature of an LLM further improved the engagement. As P25 highlighted, *"Even though it's an AI, it's a conversation. I didn't go off on tangents, instead, I tackled one topic at a time. If I was searching by myself, I might not have gotten anywhere after bouncing around different topics."*

We observed notable instances where LLM's on-the-spot answers prevented participants from misinterpreting charts. P14 mistakenly thought a value was decreasing because the visual mark moved left. This error was corrected by the agent, highlighting that the value was actually increasing as the axis domain changed (Fig. 5-A). Another case involved P18, who tried to compare the area of *"non-red"* area between two treemaps (Fig. 5-B). Although he initially thought the two treemaps had significantly different areas in their non-red regions, the specific calculation provided by the LLM showed that the areas were not notably different, helping him to form a more accurate perception of the data presented.

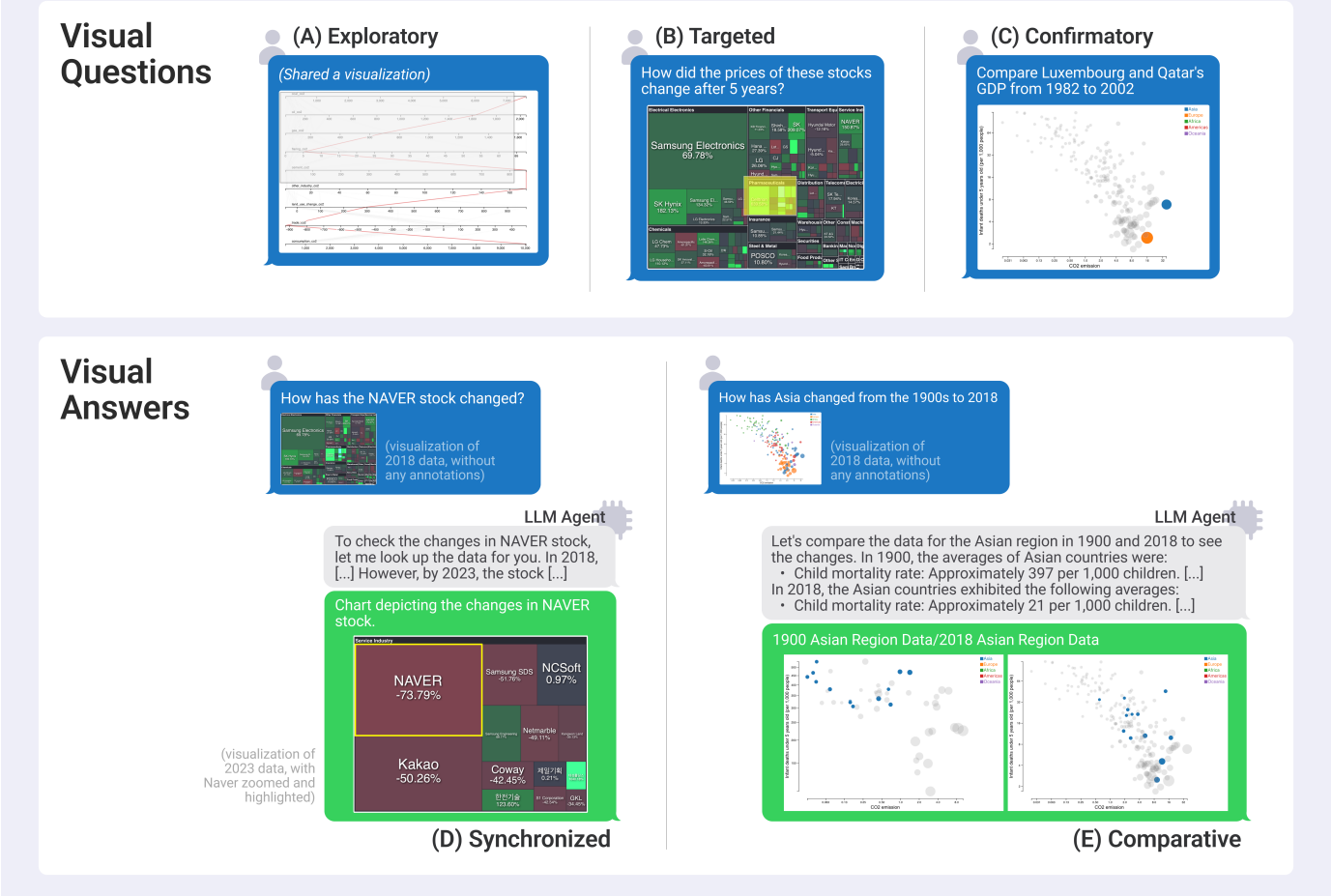


Fig. 6. We categorized participants' intentions behind visual queries into exploratory (A), targeted (B), and confirmatory (C), along with the role of the LLM agent's visual responses, namely synchronized (D) and comparative visual answers (E). Exploratory queries highlight vague areas of interest, targeted queries enable precise data point selection, and confirmatory queries reinforce textual questions with visual supplements. The agent's synchronized answers adjust a visualization to highlight discussed elements, while comparative answers facilitate visual comparison of different data points or timelines, enhancing user engagement and understanding of complex data visualizations.

6.2 Enhanced Interaction through Visual Communication

We analyzed interaction patterns in visualization-mediated communication, where the participants exchanged annotated visualizations with LLM agents alongside text communication. We focused on understanding user-side phenomena in visual communication, specifically users' motivations for making visual queries and the impact of visual responses on their analytical processes (Fig. 6).

6.2.1 Visual Queries

Participants' visual queries were categorized into three primary intentions: exploratory, targeted, and confirmatory.

Exploratory visual query is a generalized way to spotlight a specific area on the chart. This marks the user's exploratory intent to delve deeper into the highlighted area, albeit in a vague format since the user has not explicitly formulated what they want to investigate. Observations showed this could be a few manually selected points, points under certain categories, or regions marking clusters of data points. In certain cases, participants attempted to convey more meaning by sequentially highlighting different areas, expecting the LLM agent to grasp their higher intention. For instance, P10 and P13 sequentially shared charts of different years, anticipating the agent to present distinct summaries of changes and observations for each year. P21, highlighting two sectors in a treemap sequentially, anticipated the agent to provide insights about a noticed phenomenon where the two sectors appeared *"similar in size but reversed in the order,"* seeking an in-depth exploration and explanation of this occurrence. Many participants expressed the ease of conveying vague and abstract focuses to the agent through visual queries (P2, P7, P9, P14-P16, P19, P25). Participants highlighted the burden of formulating textual questions and emphasized that such inquiries are feasible only after understanding the chart (P3, P5, P7, P13-P16, P19, P21). P16 explained,

> *"I could just about manage to select areas or points and send them to the AI, then follow the recommended questions. But, having to come up with a specific text question? That was a hassle."*

Targeted visual query involves clearly marking certain points or areas on the chart that define the target of analysis, which is not specified in the textual question and can only be found in the visual selection. Targets range from data points, axes, half of the screen, or even the entire screen

(P22). Many participants found this enabled them to confine their analysis to their interests, proving particularly helpful when it was cumbersome to write data focus for numerous, small, and clustered items. P21 utilized the ease of selecting proximal points (Fig. 6-B), explaining,

> *"Even when I was just interested in one specific point, I could ask them to brief me about it along with other (visually) nearby data."*

Confirmatory visual query involves marking specific points or areas, overlapping with information already in the textual query. Despite the text already containing the analysis target, visual markings were still added, leading to a redundant representation. P22 remarked, *"Sending over a picture made me feel like the AI was right there with me, looking at it and getting what I was aiming for. It felt easy and comfortable."* Beyond clarification, participants reported a positive sense of relief and control because communicating with an LLM through visualizations gave them a feeling of being closely connected (P2, P9, P10, P22).

6.2.2 Visual Answers

In our analysis of visual responses, we aimed to understand how the added visualization influenced users' analytical process. We focused on the distinct impact of visual answers as opposed to text-only responses. For instance, while textual answers might introduce new data points for investigation, which are also highlighted in the visual answers, the contribution of visual answers might extend beyond merely presenting data points visually. We discovered that adding visual answers to textual responses enhanced user engagement with the visualization through two key functions: visual synchronization and visual comparison.

Synchronized visual answer refers to visual answers that adjust the visualization to mirror the ongoing discussion and improve the visibility of entities being discussed. This adjustment includes highlighting specific elements, zooming into pertinent areas to exclude irrelevant parts, and updating the timeline to correspond with the currently discussed year. Such patterns also occurred in composition; for instance, when a user inquired about the price change of a specific stock, *Naver*, the visual response provided a later timestamp visualization, zoomed into a *Service Industry* sector that *Naver* belonged to, and highlighted *Naver* (Fig. 6-D).

Visually synchronized responses not only made the dense text more readable and engaging but also reduced the barrier for users to interact with visualizations directly. Many participants appreciated the accompaniment of visual clarification, highlighting their benefits over textual responses alone. These benefits included clarifying the focus of the textual information (P4, P7), easing the burden of digesting unfamiliar and lengthy text (P24, P5), and facilitating the matching of text with visualizations (P3). Furthermore, visual synchronization enabled users to stay directly engaged with visualizations by making relevant visual information readily accessible without the need for additional user actions. P24 was able to verify visually thanks to the visual response, whereas P3 couldn't because there was no visual aid. P24 mentioned, *"One side of the graph was so tiny I couldn't even see the names. Without a picture, I might guess it's in that tiny bit, but I can't be sure. A picture let me check it myself."* Meanwhile, P3 shared, *"The AI told me which stock was the biggest, but I couldn't spot it. It didn't show me where, and I couldn't see it straight away, so I just didn't bother searching."*

Synchronized visual answers sometimes functioned to locate specific points within a visualization against the clutter of other points. This could be intentional, as in the case where P24 specifically requested a specific country to be highlighted in the visualization, or incidental, as P5 described, *"I completely missed that (the data point for) Ireland was hidden behind Luxembourg, but the AI's highlight made it visible to me."*.

Comparative visual answer involves placing two visualizations side by side (*i.e.,* juxtaposition) that include comparable entities. In our experimental setup, this was achieved through comparisons either between the main visualization and a suggested one, or between two suggested visualizations. These entities might represent different data points within a single visualization or the same points across different times. For instance, in response to inquiries about changes in a specific region's countries over time, the agent provided two visualizations highlighting the region in 1900 and 2018 (Fig. 6-E). This side-by-side arrangement initiated users to visually compare the entities. However, relying solely on visual comparison for deepening discussions had its limits, as highlighted by P16, who remarked, *"Oh, comparing is a neat idea [...] But it just highlighted, and that's all. Would've been nice to have some more explanation."*

6.3 Conflicting Effects of Chart Reading Guidance

We found a disparity between perceived performance and the actual insights gained from utilizing the LLM agent's situated supports and visual communication aids. Even though participants thought they performed better with the text+visual aids LLM, both the actual insights gained and the time spent engaging with visualizations were notably less compared to the web search condition. The types of queries made by participants further magnify this conflict, with variations in their approach to using LLM assistance.

Quantitative analysis on cognitive load, the number of insights, and the visualization engagement time showed a significant difference between web search and text+vis LLM condition (Fig. 7). Through the NASA-TLX 4th subscale (Performance), participants reported themselves to be more successful with text+vis LLM than with web search, showing statistical significance under the Friedman test with the Nemenyi post-hoc test ($Q = 7.68,\ p < .05$). Conversely, the actual number of insights reported was significantly higher in the web search condition than in text+vis LLM condition ($Q = 7.40,\ p < .05$). The Friedman test [51] and the Nemenyi post-hoc test [52] are non-parametric versions of the repeated measures ANOVA. These tests do not require additional assumptions and are more appropriate for ordinal Likert-scale data. Additionally, participants who most preferred web search ($n = 3$) and text-only LLM ($n = 7$) generated significantly more insights compared to those favoring text+vis LLM ($n = 16$) ($t(24) = 3.30,\ p < .05$). Furthermore, participants who most preferred text+vis LLM spent significantly less time engaging with visualizations than those

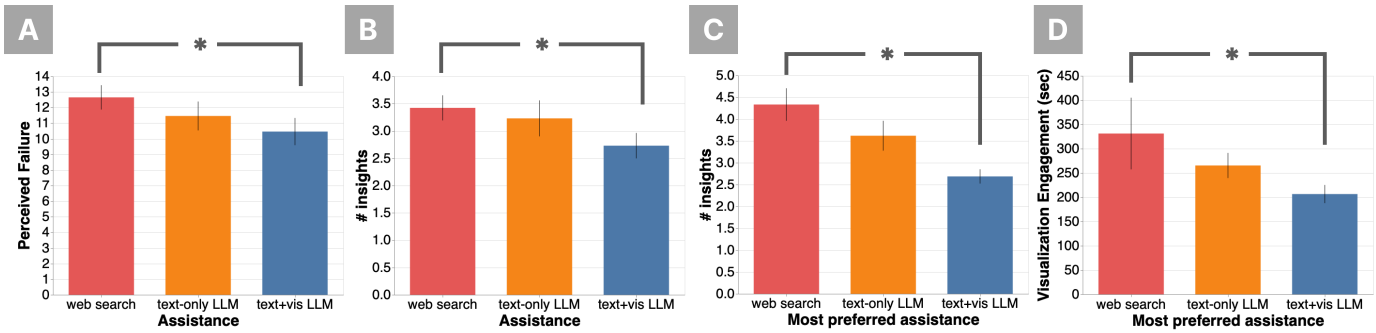


Fig. 7. Although participants believed they performed best with text+vis LLM compared to a web search engine through NASA-TLX 4th subscale (A), the actual number of insights gathered was lower (B). This pattern persisted among participants who expressed a preference for the text+vis LLM over the web search engine (C). Furthermore, participants who preferred the text+vis LLM spent significantly less time engaging with visualizations compared to those who preferred the web search engine (D).
$(* : p < .05)$

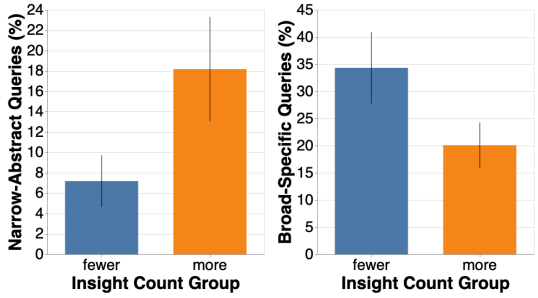


Fig. 8. Query types showed different trends between participants with insights below and above the median.

who preferred web search, under the one-way ANOVA with Tukey's HSD post-hoc test ($F(2, 75) = 3.66,\ p < .05$). While the statistical significance is mostly between text+vis LLM and web search condition, there is a trend showing a monotonic transition from web search condition to text-only LLM, and then to text+vis LLM (Fig. 7). These findings highlight a discrepancy between perceived performance and actual engagement and insight generation, indicating a potential adverse effect of LLM's situated support and the enhanced interaction of visualization-mediated communication.

To delve deeper into insight disparity, we divided participants into fewer-insight and more-insight group based on the median number of insights per participant. Analysis of their query types (Fig. 8) revealed the fewer-insight group predominantly used broad-domain specific-outcome queries (*e.g., "Which country releases the most CO2 from land use change?"*), while the more-insight group employed more narrow-domain abstract-outcome queries (*e.g., "Can you tell me about any issues related to CGV stock in 2023?"*). The proportion of broad-domain specific-outcome queries was higher in the fewer-insight group ($t(24) = 1.83, p = 0.08, d = 0.72$), although not meeting the traditional significance level but displaying a medium to large effect size. Conversely, the more-insight group used more narrow-domain abstract-outcome queries ($U(12) = 49.5, p = 0.07, r = 0.41$), again not significant but with a large effect size, highlighting a distinct query approach based on insight levels. Note that the latter comparison employed a non-parametric Mann-Whitney U test because it did not pass Shapiro-Wilk Test and Levene's Test, failing to meet the assumptions required for a t-test.

More-insight reporters who also preferred search or text-only LLM (P5, P11, P18, P21) engaged charts firsthand and used an LLM agent as a supplementary tool. Although they were unfamiliar with the visualization itself, their prior data analysis experience guided their chart analysis process, leading them to seek practical assistance to use their time more efficiently. They sought help in rearranging hard-to-read parts in charts and calculating statistics. LLM responses echoing identifiable chart content were deemed unhelpful (P5, P11, P18), and LLM's recommended follow-up questions and visual queries were seen as lacking informative value (P18). Their preference leaned towards familiar textual communication and web search rather than spending time formulating queries for LLM (P5, P21).

Fewer-insight reporters who also most preferred text+vis LLM (P3, P12, P22) appreciated guidance features for resolving their difficulties and uncertainty with charts. Reading charts posed a challenge, leading to reliance on the LLM agent which they believed *"comprehend the chart perfectly [. . .] and answer chart-related queries effectively"* (P22). They faced difficulties in formulating questions, and blamed themselves for insufficiently using the LLM agent due to this challenge. Recommended questions and visual queries provided a respite from such difficulties, making them *"feel more understood"* by the LLM.

7 Discussion

7.1 Language Models in Visualization Literacy

We investigated whether visual communication with the LLM can enhance literacy in complex visualization. Our results show that users could effectively learn to decode visualizations. They learned to interpret visualizations through direct questioning and avoid data misinterpretation with corrections from LLM agents. Yet, the concept of literacy spans from merely decoding underlying values to reading trends, connecting with outside knowledge, and personalizing information [4], [53], [19]. Conflicting results were observed in understanding bigger trends and personalizing insights. LLMs and visual communication led users to

believe they were more successful, while both the interaction time with visualizations and the number of user-reported insights decreased. This aligns with the disfluency effect [54], which states that easily learned knowledge is easily forgotten.

Upon closer examination, two contrasting interaction patterns emerged. Participants with more familiarity with data analysis leveraged the LLM to gain higher-level insights, actively interacting with the charts and avoiding trivial questions while discussing with LLM agents. Conversely, those less familiar or motivated preferred LLM interaction over engaging with the charts. These two groups also varied in their queries; the former group often asked about broader, abstract topics related to specific areas of interest, while the latter group preferred to ask the LLM to pinpoint specific results, instead of extracting information directly from the charts.

These results can be viewed in light of the concepts of visualization onboarding [55], [56], [57] and visualization guidance [58]. Although the distinction between these concepts is not always clear, onboarding focuses on teaching users how to read and interpret data visualizations, whereas guidance helps users engage with visualizations to achieve specific objectives (e.g., performing data analysis [56]). In this context, LLMs and visual communication effectively *onboarded* participants by enabling detailed questions about visualizations. As participants became familiar with visualizations and proceeded to analysis, the role of LLMs evolved into *guidance* for data analysis. For example, using LLMs for trivial questions about charts represents onboarding support, whereas detailed topic discussions with LLMs served as a guidance for finding deeper meaning in the data.

Guidance can be designed differently based on users' characteristics and goals. Objectives can range from encouraging engagement with visualizations in everyday scenarios from various perspectives, educating effective visualization techniques, to leading users towards more sophisticated analyses using visualizations as tools. One of the critical aspects in this topic is to distinguish independent learning from blind replication of guidance. We believe that such active learning occurred in our experiment. If the users had merely repeated the LLM responses without actual learning, they should have been able to report more insights with LLM than with web search because language models are more suitable for providing off-the-shelf insights compared to web search. The actual results were the opposite. Further research may explore whether learning can occur independently without guidance, and how LLM support for learning visualization differs from support for visualization consumption. In our experiment, users reported insights after exploring visualizations, but other evaluation schemes could also be considered. These might include identifying similar insights in different charts or employing literacy tests [19]. Additionally, preventing users from misinterpreting charts due to LLM hallucinations would be a critical consideration for fostering independent learning.

7.2 Comparison with NOVIS Model

Our motivation and task formation shares much with the NOVIS model [2], a cognitive model for novices encountering unfamiliar visualizations. Here we compare our findings with the model to understand how the incorporation of an LLM and visual communication altered the dynamics. According to the NOVIS model, novices encountering an unfamiliar visualization undergo a series of steps: they construct an interpretive frame, explore the visualization within this frame, and then either revise the frame based on their findings or struggle due to failure. Our finding suggests that the inclusion of an LLM affects most of these stages.

On one hand, during the construction and refinement of their interpretive frame, novices had the option to directly consult an LLM for specific inquiries about the chart. For example, they could highlight particular data points and elements they have yet to understand. The NOVIS model revealed that novices tend to adhere to their initially constructed frame, and when it comes to revising a misunderstood frame, they typically depended on familiar cues they could interpret. This context underscores the value of LLM's situated support in offering correct and timely feedback, addressing the cognitive challenges novices encounter.

On the other hand, during the exploration phase of the visualization, reliance on the LLM could detract from novices' direct interaction with the visualization. The NOVIS model breaks down the exploration process into three parts: retrieving information from the visualization, recalling domain and personal knowledge, and engaging in unrestricted exploration. In our study, some participants predominantly used the LLM for information retrieval and as a source of domain knowledge. While beneficial, this approach might have bypassed the inherent exploratory process that fosters direct engagement with the visualization, leading to reduced visual exploration. Indeed, we found that novices who most preferred the LLM's support spent significantly less time engaging with the visualization and gained significantly fewer insights from it.

In this context, visual communication with the LLM emerged as a pivotal mechanism that redirects users' attention back to the visualization. Visual responses, aligned with the users' textual discussions, served as constant prompts for users to consult the visualization. Additionally, the ability to pose visual questions allowed users to express abstract ideas directly onto the visualizations, facilitating a deeper engagement where the visualization remained central to their reasoning process. Both our study and the NOVIS model observed the challenges with verbalization; within the NOVIS model, participants struggled to articulate their thoughts upon encountering the visualization, tending to offer abstract impressions or focus on general visual traits. In our study, although participants encountered similar challenges in formulating textual questions for LLM, this obstacle was effectively overcome through intuitively highlighting their areas of focus using visual questioning.

7.3 Visual Communication with Language Models

We enabled visual communication with the LLM through visual annotation features, such as point highlighting, rectangular area marking, and zooming. These features supported users in delivering their ideas to the LLM agent; for users with a specific inquiry target in mind, annotations offered a means to select data points that were difficult or impossible

to identify using text alone. For those with a more abstract focus, yet to be articulated, annotations served as a visual language to express such focus. Participants developed their unique visual language schemes, expecting the agent to interpret these (e.g., using a rectangular area as a filter, or a sequence of annotations as a request for comparison). Whether driven by specific objectives or curiosity, annotations provided both new capabilities and affordances; they enabled expressions previously unavailable, and conversely, nudged users to explore visualizations with these capabilities.

Our findings call for further examination of annotations as a medium for communicating with visualizations. The scope of annotation capabilities extends beyond our prototype system [59], [60], covering diverse targets (*e.g.,* data marks, axes, labels), goals (*e.g.,* nuanced visualization, decoration, personalization), forms (e.g., arrows, background colors, custom icons), and computational relationships (e.g., marking average values, highlighting marks below a specific value, drawing trend lines). These variations can be personalized to serve as an individual's language and, conversely, can prompt different perspectives and cognitive engagements with visualizations. Understanding how visualization novices harness the expressive potential of annotations could inform future system development. Furthermore, annotation is also an active thought process that reflexively influences users' sense-making [61], [62]. During the active reading of visualization, individuals naturally engage with tools for pointing and drawing free-form annotations as part of their thought process [12]. Thus, incorporating free-drawing input into communication with LLMs could enhance the customization of on-demand visualization guides for everyday use. The recent advancements in LLMs, especially those with image modality capabilities [63], showcase the potential for achieving this integration.

7.4 Limitation and Future Work

7.4.1 Toward Complex, Composite Visualizations

We selected scatterplots, treemaps, and parallel coordinates as our target visualization for their relevance to everyday contexts and their balance of simplicity and informativeness. However, the complexity of a visualization may impact users' learning processes, calling for further examination in future work. There are many options to incorporate additional complexity to visualization. For instance, we could enhance scatterplots with trajectories to show changes over time, or increase the hierarchical depth in treemaps for more detailed data representation. Other advanced visualization types, such as network visualizations and flow diagrams, are also options. If we prioritize explorability over clear insights embedded in the visualizations, options can further extended to composite visualizations. Dashboard visualizations, like those consumed during COVID-19 [64], have been widespread within the general public. Visual analytic features on dashboards, such as brushing and linking, offer additional educational opportunities.

7.4.2 Toward Creating and Evaluating Visualization

Our study explored the impact of on-demand visualization support on enhancing visualization literacy. Yet, its application might also extend to visualization creation and evaluation. Many of our participants expressed a desire not only to analyze individual charts but also to create and compare multiple charts for deeper insights. Future research could explore whether this on-demand support effectively aids in creating and evaluating various visualizations, whether for exploratory data analysis or authoring personalized visualizations. Another opportunity can be enabling users to evaluate design issues of visualization. Enhancing visualization literacy could involve not just understanding but also critiquing and improving upon design shortcomings [65], which is a topic increasingly discussed by visualization practitioner on social media platforms [66]

7.4.3 Enriching Text-only Communication of Visualization

We observed that in the text-only condition, participants mostly referred to specific visual marks using their names or identifiers. It was possible to ask questions about visual aspects using text (e.g., *"the green square in the top right"*), but the frequent use of identifiers indicates that participants preferred to focus on the data aspect rather than the visual aspect. One hypothesis is that for novices, reading familiar text fragments is cognitively easier than describing visual elements. Another possibility is that users have low expectations of AI and thus aim to be as precise as possible. This tendency may vary with different visualization settings; for instance, a treemap without data labels might yield more visual-oriented questions. Future research could explore how users' primary focus shifts between textual, numerical, and visual aspects depending on specific contexts in text-based communication about visualizations with LLMs. Such research could inform the design of text-only visualization interactions with LLM, considering the limited access to visual communication methods.

7.4.4 Communicating Visualizations with LLM

In our research, we meticulously detailed every visible element of visualizations in text to effectively communicate with LLM. This method was adequate for our current scope, but extending it to a broader array of visualizations necessitates further exploration into how to ensure effective communication. This is particularly crucial for visualizations that rely heavily on human perception, such as those utilizing Gestalt principles, where simple textual descriptions may not suffice. We can leverage insights from research on generating captions for visualizations, such as from the accessibility domain, to bridge this gap.

Furthermore, we identified an opportunity to better align the LLM's understanding of annotation use with human practices. Our study did not specify preferences for types of annotations, leading to an observation where the LLM predominantly opted for zooming and highlighting points over rectangular selections, likely due to their directness compared to the more abstract nature of area annotations. This insight prompts us to consider teaching an LLM more about the human approach to annotations, potentially improving how it communicates with users through a more intuitive and human-like use of visual annotations.

7.4.5 Diversifying the Study Population

We defined novices as individuals who have not received formal education in visualization and do not frequently encounter visualization in their daily lives. We recruited users from a local community platform. The majority of our participants were in their 20s or 30s, and a significant portion of them were capable of or had at least experienced data analysis. As the learning patterns of visualization novices may vary based on other characteristics of the group [25], future research could expand the study to diverse age groups (e.g., individuals over 50, elementary school students) and consider controlling for data literacy as a new variable.

8 Conclusion

We explored the role of Large Language Models (LLMs) in supporting individuals who are unfamiliar with advanced visualizations. Our LLM-based interface, allowing both text and visual interaction on charts, was commended for its contextual support and enhanced guidance. Despite these benefits, we observed a potential unintended side effect of such guidance feature: reduced insight gained from and engagement with visualizations, especially among individuals with limited motivation and data literacy. The results suggest that LLMs hold both the potential to provide effective on-demand support for novices interpreting unfamiliar visualizations and the risk of leading to over-reliance. On one hand, we suggest that future research could more actively utilize language models to improve higher-level visualization literacy, such as the creation and evaluation of visualizations. On the other hand, to ensure users do not neglect the inherent strengths of visualization due to over-reliance on LLMs, Visual communication features could be further enhanced and diversified to cover a broader spectrum of visualization annotation spaces.

References

[1] L. Battle, P. Duan, Z. Miranda, D. Mukusheva, R. Chang, and M. Stonebraker, "Beagle: Automated extraction and interpretation of visualizations from the web," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–8.

[2] S. Lee, S.-H. Kim, Y.-H. Hung, H. Lam, Y.-A. Kang, and J. S. Yi, "How do People Make Sense of Unfamiliar Visualizations?: A Grounded Model of Novice's Information Visualization Sensemaking," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 499–508, Jan. 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7192668/

[3] B. C. Kwon and B. Lee, "A comparative evaluation on online learning approaches using parallel coordinate visualization," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016, pp. 993–997.

[4] K. Börner, A. Maltese, R. N. Balliet, and J. Heimlich, "Investigating aspects of data visualization literacy using 20 information visualizations and 273 science museum visitors," *Information Visualization*, vol. 15, no. 3, pp. 198–213, 2016.

[5] A. Zoss, A. Maltese, S. M. Uzzo, and K. Börner, "Network visualization literacy: Novel approaches to measurement and instruction," *Network Science In Education: Transformational Approaches in Teaching and Learning*, pp. 169–187, 2018.

[6] T. Kwon, M. Jeong, E.-S. Ko, and Y. Lee, "Captivate! Contextual Language Guidance for Parent–Child Interaction," in *CHI Conference on Human Factors in Computing Systems*. New Orleans LA USA: ACM, Apr. 2022, pp. 1–17. [Online]. Available: https://dl.acm.org/doi/10.1145/3491102.3501865

[7] E. E. Firat, A. Denisova, M. L. Wilson, and R. S. Laramee, "P-Lite: A study of parallel coordinate plot literacy," *Visual Informatics*, vol. 6, no. 3, pp. 81–99, Sep. 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2468502X22000377

[8] Bloomberg. (2024, Feb.) 2023: The year in graphics, data, maps, and visual stories. [Online]. Available: https://www.bloomberg.com/graphics/2023-in-graphics/

[9] OpenAI. (2023, Oct.) Chatgpt. [Online]. Available: https://chat.openai.com/

[10] E. Kavaz, A. Puig, and I. Rodríguez, "Chatbot-based natural language interfaces for data visualisation: A scoping review," *Applied Sciences*, vol. 13, no. 12, p. 7025, 2023.

[11] P. Maddigan and T. Susnjak, "Chat2vis: Generating data visualisations via natural language using chatgpt, codex and gpt-3 large language models," *IEEE Access*, 2023.

[12] J. Walny, S. Huron, C. Perin, T. Wun, R. Pusch, and S. Carpendale, "Active Reading of Visualizations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 770–780, Jan. 2018. [Online]. Available: http://ieeexplore.ieee.org/document/8017606/

[13] R. C. Schank, "Active learning through multimedia," *IEEE multimedia*, vol. 1, no. 1, pp. 69–78, 1994.

[14] L. Shen, E. Shen, Y. Luo, X. Yang, X. Hu, X. Zhang, Z. Tai, and J. Wang, "Towards natural language interfaces for data visualization: A survey," *IEEE transactions on visualization and computer graphics*, 2022.

[15] Y. Liu, Y. Yao, J.-F. Ton, X. Zhang, R. G. H. Cheng, Y. Klochkov, M. F. Taufiq, and H. Li, "Trustworthy llms: a survey and guideline for evaluating large language models' alignment," *arXiv preprint arXiv:2308.05374*, 2023.

[16] Q. Roy, F. Zhang, and D. Vogel, "Automation accuracy is good, but high controllability may be better," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–8.

[17] C. Oh, J. Song, J. Choi, S. Kim, S. Lee, and B. Suh, "I lead, you help but only with enough details: Understanding user experience of co-creation with artificial intelligence," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–13.

[18] E. M. Peck, S. E. Ayuso, and O. El-Etr, "Data is personal: Attitudes and perceptions of data visualization in rural pennsylvania," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–12.

[19] S. Lee, S.-H. Kim, and B. C. Kwon, "VLAT: Development of a Visualization Literacy Assessment Test," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 551–560, Jan. 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7539634/

[20] S. Pandey and A. Ottley, "Mini-vlat: A short and effective measure of visualization literacy," *arXiv preprint arXiv:2304.07905*, 2023.

[21] E. Firat, A. Denisova, and R. Laramee, "Treemap literacy: a classroom-based investigation," in *Eurographics Proceedings*, 2020.

[22] E. E. Firat, A. Joshi, and R. S. Laramee, "Interactive visualization literacy: The state-of-the-art," *Information Visualization*, vol. 21, no. 3, pp. 285–310, 2022.

[23] B. Bach, M. Keck, F. Rajabiyazdi, T. Losev, I. Meirelles, J. Dykes, R. S. Laramee, M. AlKadi, C. Stoiber, S. Huron *et al.*, "Challenges and opportunities in data visualization education: A call to action," *arXiv preprint arXiv:2308.07703*, 2023.

[24] B. Alper, N. H. Riche, F. Chevalier, J. Boy, and M. Sezgin, "Visualization literacy at elementary school," in *Proceedings of the 2017 CHI conference on human factors in computing systems*, 2017, pp. 5485–5497.

[25] A. Burns, C. Lee, R. Chawla, E. Peck, and N. Mahyar, "Who do we mean when we talk about visualization novices?" in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–16.

[26] L. Panavas, A. E. Worth, T. Crnovrsanin, T. Sathyamurthi, S. Cordes, M. A. Borkin, and C. Dunne, "Juvenile graphical perception: A comparison between children and adults," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–14.

[27] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[28] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary *et al.*, "Beyond english-centric multilingual machine translation." *J. Mach. Learn. Res.*, vol. 22, no. 107, pp. 1–48, 2021.
[29] OpenAI, "Gpt-4 technical report," 2023.
[30] P. Khadpe, R. Krishna, L. Fei-Fei, J. T. Hancock, and M. S. Bernstein, "Conceptual metaphors impact perceptions of human-ai collaboration," *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW2, pp. 1–26, 2020.
[31] K. I. Gero and L. B. Chilton, "Metaphoria: An algorithmic companion for metaphor creation," in *Proceedings of the 2019 CHI conference on human factors in computing systems*, 2019, pp. 1–12.
[32] Y. Lin, J. Guo, Y. Chen, C. Yao, and F. Ying, "It is your turn: collaborative ideation with a co-creative robot through sketch," in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–14.
[33] J. Koch, A. Lucero, L. Hegemann, and A. Oulasvirta, "May ai? design ideation with cooperative contextual bandits," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–12.
[34] M. Guzdial, N. Liao, J. Chen, S.-Y. Chen, S. Shah, V. Shah, J. Reno, G. Smith, and M. O. Riedl, "Friend, collaborator, student, manager: How design of an ai-driven game level editor affects creators," in *Proceedings of the 2019 CHI conference on human factors in computing systems*, 2019, pp. 1–13.
[35] J. D. Weisz, M. Muller, S. Houde, J. Richards, S. I. Ross, F. Martinez, M. Agarwal, and K. Talamadupula, "Perfection not required? human-ai partnerships in code translation," in *26th International Conference on Intelligent User Interfaces*, 2021, pp. 402–412.
[36] R. Louie, A. Coenen, C. Z. Huang, M. Terry, and C. J. Cai, "Novice-ai music co-creation via ai-steering tools for deep generative models," in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–13.
[37] 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 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–18.
[38] A. Narechania, A. Srinivasan, and J. Stasko, "Nl4dv: A toolkit for generating analytic specifications for data visualization from natural language queries," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 369–379, 2020.
[39] S. Latif, Z. Zhou, Y. Kim, F. Beck, and N. W. Kim, "Kori: Interactive synthesis of text and charts in data documents," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 1, pp. 184–194, 2021.
[40] N. W. Kim, S. C. Joyner, A. Riegelhuth, and Y. Kim, "Accessible visualization: Design space, opportunities, and challenges," in *Computer Graphics Forum*, vol. 40, no. 3. Wiley Online Library, 2021, pp. 173–188.
[41] C. Chen, R. Zhang, E. Koh, S. Kim, S. Cohen, and R. Rossi, "Figure captioning with relation maps for reasoning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 1537–1545.
[42] A. Lundgard and A. Satyanarayan, "Accessible visualization via natural language descriptions: A four-level model of semantic content," *IEEE transactions on visualization and computer graphics*, vol. 28, no. 1, pp. 1073–1083, 2021.
[43] C. North, "Toward measuring visualization insight," *IEEE computer graphics and applications*, vol. 26, no. 3, pp. 6–9, 2006.
[44] Y.-H. Kim, B. Lee, A. Srinivasan, and E. K. Choe, "Data@ hand: Fostering visual exploration of personal data on smartphones leveraging speech and touch interaction," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–17.
[45] S. Latif, S. Chen, and F. Beck, "A deeper understanding of visualization-text interplay in geographic data-driven stories," in *Computer Graphics Forum*, vol. 40, no. 3. Wiley Online Library, 2021, pp. 311–322.
[46] O. Topsakal and T. C. Akinci, "Creating large language model applications utilizing langchain: A primer on developing llm apps fast," in *International Conference on Applied Engineering and Natural Sciences*, vol. 1, no. 1, 2023, pp. 1050–1056.
[47] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," *arXiv preprint arXiv:2210.03629*, 2022.
[48] S. Lee, S.-H. Kim, Y.-H. Hung, H. Lam, Y.-a. Kang, and J. S. Yi, "How do people make sense of unfamiliar visualizations?: A grounded model of novice's information visualization sensemaking," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 1, pp. 499–508, 2015.
[49] H. B. Mann, "The construction of orthogonal latin squares," *The Annals of Mathematical Statistics*, vol. 13, no. 4, pp. 418–423, 1942.
[50] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Advances in psychology*. Elsevier, 1988, vol. 52, pp. 139–183.
[51] M. R. Sheldon, M. J. Fillyaw, and W. D. Thompson, "The use and interpretation of the friedman test in the analysis of ordinal-scale data in repeated measures designs," *Physiotherapy Research International*, vol. 1, no. 4, pp. 221–228, 1996.
[52] T. Pohlert, "The pairwise multiple comparison of mean ranks package (pmcmr)," *R package*, vol. 27, no. 2019, p. 9, 2014.
[53] J. Boy, R. A. Rensink, E. Bertini, and J.-D. Fekete, "A principled way of assessing visualization literacy," *IEEE transactions on visualization and computer graphics*, vol. 20, no. 12, pp. 1963–1972, 2014.
[54] T. Kühl and A. Eitel, "Effects of disfluency on cognitive and metacognitive processes and outcomes," *Metacognition and Learning*, vol. 11, pp. 1–13, 2016.
[55] C. Stoiber, F. Grassinger, M. Pohl, H. Stitz, M. Streit, and W. Aigner, "Visualization onboarding: Learning how to read and use visualizations."
[56] C. Stoiber, D. Ceneda, M. Wagner, V. Schetinger, T. Gschwandtner, M. Streit, S. Miksch, and W. Aigner, "Perspectives of visualization onboarding and guidance in va," *Visual Informatics*, vol. 6, no. 1, pp. 68–83, 2022.
[57] V. Dhanoa, C. Walchshofer, A. Hinterreiter, H. Stitz, E. Groeller, and M. Streit, "A process model for dashboard onboarding," in *Computer Graphics Forum*, vol. 41, no. 3. Wiley Online Library, 2022, pp. 501–513.
[58] D. Ceneda, T. Gschwandtner, T. May, S. Miksch, H.-J. Schulz, M. Streit, and C. Tominski, "Characterizing guidance in visual analytics," *IEEE transactions on visualization and computer graphics*, vol. 23, no. 1, pp. 111–120, 2016.
[59] D. Ren, M. Brehmer, B. Lee, T. Höllerer, and E. K. Choe, "Chartaccent: Annotation for data-driven storytelling," in *2017 IEEE Pacific Visualization Symposium (PacificVis)*. Ieee, 2017, pp. 230–239.
[60] E. K. Choe, B. Lee *et al.*, "Characterizing visualization insights from quantified selfers' personal data presentations," *IEEE computer graphics and applications*, vol. 35, no. 4, pp. 28–37, 2015.
[61] H. Romat, N. Henry Riche, K. Hinckley, B. Lee, C. Appert, E. Pietriga, and C. Collins, "Activeink: (th) inking with data," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–13.
[62] Y. B. Shrinivasan and J. J. Van Wijk, "Supporting the analytical reasoning process in information visualization," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2008, pp. 1237–1246.
[63] "Gpt-4v(ision) system card," 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:263218031
[64] (2024, Feb.) Who covid-19 dashboard. [Online]. Available: https://data.who.int/dashboards/covid19/cases
[65] L. W. Ge, Y. Cui, and M. Kay, "Calvi: Critical thinking assessment for literacy in visualizations," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–18.
[66] 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 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–18.

Kiroong Choe is a Ph.D. student at the Human-Computer Interaction Laboratory under the Department of Computer Science and Engineering, Seoul National University, Korea. His research interests include Human-AI Collaboration, Educational System Design, and Information Visualization. He is currently focused on designing AI systems for sensemaking in structured data such as academic literature, knowledge graphs, and data charts.

Chaerin Lee is a Ph.D. student at the Human-Computer Interaction Laboratory under the Department of Computer Science and Engineering, Seoul National University, Korea. Her research interests are in exploring sound and visual data in a multimodal manner. She is currently working on developing systems that utilize visualization and text to more effectively navigate through sound data.

Soohyun Lee is a Ph.D. student at the Human-Computer Interaction Laboratory under the Department of Computer Science and Engineering, Seoul National University, Korea. His research interests lie in data analysis, specifically focusing on creating effective visual analytic techniques based on dimensionality reduction. He is dedicated to developing methods that simplify complex data sets for easier interpretation and analysis.

Jiwon Song is a Ph.D. student at the Human-Computer Interaction Laboratory under the Department of Computer Science and Engineering, Seoul National University, Korea. Her research interests are in designing systems that help seniors manage their personal data more effectively. Currently, she is researching a system aimed at enabling seniors to better manage their blood pressure data, making it more accessible and understandable for them.

Aeri Cho is a Ph.D. student at the Human-Computer Interaction Laboratory under the Department of Computer Science and Engineering, Seoul National University, Korea. Her research interests are in data analysis, with a particular focus on introducing human-steerable interactions into dimensionality reduction using neural networks. She aims to create interfaces that allow users to interact more intuitively with data analysis processes, enhancing the accessibility and effectiveness of data interpretation.

Nam Wook Kim is an Assistant Professor of Computer Science at Boston College. His research vision is to lower barriers for everyone to understand and communicate complex data. He tackles this challenge by studying visualization within the broad context of human-computer interaction. His research investigates innovative approaches to interact with data, going beyond traditional expert systems and addressing the needs of a broader audience, including designers, journalists, and casual users.

Jinwook Seo is a professor in the Department of Computer Science and Engineering, Seoul National University, where he is also the Director of the Human-Computer Interaction Laboratory. His research interests include Human-Computer Interaction, Information Visualization, and Biomedical Informatics. He received his PhD in Computer Science from the University of Maryland at College Park in 2005.