Visualizing Tree-of-Analysis: Facilitating Conversational Visual Analytics for Novices

Feiyuan Qu
Laboratory of Art and Archaeology Image, Zhejiang University
Hangzhou, Zhejiang, China
feiyuan_qu@zju.edu.cn

Tan Tang*
Laboratory of Art and Archaeology Image, Zhejiang University
Hangzhou, Zhejiang, China
tangtan@zju.edu.cn

Zeyang Fu
College of Computer Science and Technology, Zhejiang University
Hangzhou, Zhejiang, China
zeyangfu@zju.edu.cn

Yan Chen
State Key Lab of CAD&CG
Zhejiang University
Hangzhou, Zhejiang, China
clarence_cy@zju.edu.cn

Hanze Jia
State Key Lab of CAD&CG
Zhejiang University
Hangzhou, Zhejiang, China
hzjia@zju.edu.cn

Junming Gao
Laboratory of Art and Archaeology Image, Zhejiang University
Hangzhou, Zhejiang, China
junminggao@zju.edu.cn

Songela Nurdawuliet
Laboratory of Art and Archaeology Image, Zhejiang University
Hangzhou, Zhejiang, China
songelanur@zju.edu.cn

Yingcai Wu
State Key Lab of CAD&CG
Zhejiang University
Hangzhou, Zhejiang, China
ycwu@zju.edu.cn

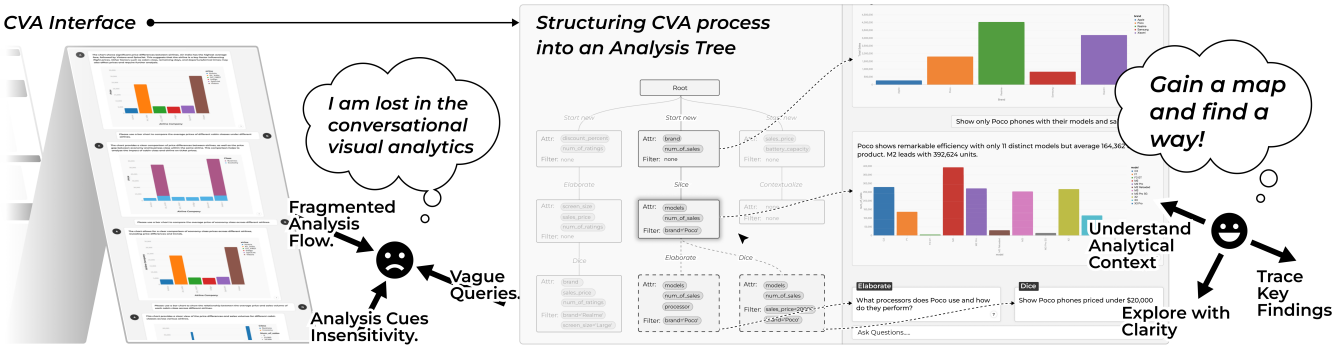

Figure 1: Facilitating conversational visual analytics with Tree-of-Analysis (ToA for short). Compared to traditional CVA interface where novices feeling lost in a fragmented analysis, our method, namely ToA, structures the process into an interactive analysis tree, guiding novices step by step. Our study reported that novices with ToA could better understand their analytical context, trace key findings, and explore alternative analysis branches with clarity.

Abstract

Conversational visual analytics (CVA) make data exploration accessible to novices but often leave users disoriented during multi-turn conversations. Previous approaches provide data-centric recommendations, but fail to help users regain orientations. To bridge this gap, we conducted a formative study ($N = 12$) revealing that novices are insensitive to analytical cues and rely on vague queries, leading to disorientation and task failures. In contrast, experts are sensitive to two types of analytical cues and use seven types of queries to organize workflows. Based on these findings, we propose ToA, a novel approach that structures the CVA process as an interactive analysis tree. Moreover, we visualize this tree, with AI outputs as nodes (containing two cue types) and user queries as edges (categorized by seven query types), to provide novices with an overview of their analysis journey. We evaluated ToA through user studies ($N = 12$) and expert interviews ($N = 3$). The results suggest that ToA eliminates task failure and increases per-turn insights (+58.3%), despite longer per-turn thinking time (+17.7%). Expert interviews further confirm its potential to democratize visual analytics.

*Corresponding Author.

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3791690

CCS Concepts

• **Human-centered computing → Natural language interfaces**; *Graphical user interfaces*; *Visual analytics.*

Keywords

Conversational Visual Analytics, Novices, Human-AI Interaction

ACM Reference Format:

Feiyuan Qu, Tan Tang, Zeyang Fu, Yan Chen, Hanze Jia, Junming Gao, Songela Nurdawuliet, and Yingcai Wu. 2026. Visualizing Tree-of-Analysis: Facilitating Conversational Visual Analytics for Novices. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3772318.3791690

1 Introduction

The emergence of large language models (LLMs) has made conversational visual analytics (CVA) an appealing choice for novices to conduct visual-driven data exploration [56, 59, 75, 76, 78]. Unlike traditional visual analytics (VA) systems that require users to master complex tool operations [48], CVA systems enable intuitive data exploration through natural language conversation.

While CVA reduces barriers to visual analytics, it also introduces new difficulties to novices. Unlike traditional visual analytics tools (e.g., Tableau) that guide users through graphic user interface, CVA's natural language interface lacks discoverability, leaving novices struggling with what to do next [68]. Additionally, the linear conversational interface complicates maintaining analytical context across multi-turn interactions [85]. Together, these characteristics of CVA make novices disoriented in their analysis process, struggling to understand where they are, where they have been, and what areas remain unexplored. To help novices regain orientation in CVA, existing work draws from two complementary research areas. In the visual analytics domain, researchers focus on data-centric recommendations that suggest insights based on statistical patterns [70, 87, 88, 96]. In the natural language processing domain, recent work leverages conversational history to generate contextually relevant follow-up questions [24, 52]. However, both approaches offer isolated suggestions rather than a global perspective on the analysis journey. As a result, novices still struggle to track their analysis progress and identify promising directions for further investigation.

To better serve novices in CVA, it is essential to provide a global perspective that continually situates their current position within the broader context. Inspired by the effectiveness of tree structures in providing hierarchical overviews for complex tasks [26, 73, 94, 95], we propose a real-time algorithm to represent the CVA process as an analysis tree. By visualizing this tree, we enable novices to gain an overview and easily identify next steps based on analysis context. Constructing such trees presents the following challenges:

C1: Vast User Intent Search Space. The flexibility of natural language creates an infinite space of user intents. Building a well-structured tree requires distilling this diversity into a predictable set of query types.

C2: Complex Multimodal Context. CVA context contains analytical information distributed across text, data, and visualizations. Constructing the tree requires extracting and organizing key analytical elements scattered across the multimodal context.

In this study, we first conducted a formative study ($N = 12$) comparing experts' and novices' behaviors in CVA. Our goal is to understand why novices become disoriented and identify expert strategies for effective analysis navigation. Through observations and interviews, we discovered that novices are insensitive to analytical cues in AI output and rely heavily on vague queries, leading to fragmented analysis flow that results in task failures and disorientation. In contrast, experts systematically identify 2 types of analytical cues from AI output and use 7 distinct query types to organize a tree-like analysis flow. Based on these findings, we address **C1** by formalizing user queries into 7 distinct types. This classification enables us to define a well-structured analysis tree, where user queries serve as edges (categorized by the 7 query types) connecting nodes defined by AI outputs (containing 2 analytical cue types). To address **C2**, we developed an LLM-based tree construction algorithm. Guided by our formative study, this algorithm extracts analytical cues from multimodal AI output and appropriately positions them within the analysis tree. Based on the tree, novices can receive step-by-step guidance while preserving a global overview of the analysis. Finally, we conducted a user study ($N = 12$) demonstrating that ToA significantly improves CVA efficiency. The results show ToA increases per-turn insights by 58.3% despite 17.7% longer per-turn thinking time. ToA also eliminates task failures and encourages users to engage in a deeper and more reflective exploration. Expert interviews ($N = 3$) further validate its potential to democratize visual analytics. Our main contributions are as follows:

- We present a formative study ($N = 12$) to explain why novices become disoriented in CVA and distill the general CVA workflow that includes two types of analytical cues and seven types of queries.
- We propose a real-time tree construction algorithm for representing CVA process as an analysis tree, which incorporates the practical guidelines derived from our formative study.
- We conducted an in-lab study ($N = 12$) comparing ToA with existing CVA tools and found that ToA effectively alleviates disorientation for novices by providing a clear overview that encourages more proactive exploration.

2 Related Work

The following sections describe related work about conversational visual analytics systems (Sec. 2.1), direction maintenance in visual analytics (Sec. 2.2) and conversation systems (Sec. 2.3), and the application of tree and graph visualization in complex tasks (Sec. 2.4). Existing approaches cannot help users maintain holistic analysis perspective, motivating our analysis tree visualization.

2.1 Conversational Visual Analytics

Traditional VA systems require users to master complex interfaces to explore data effectively [43, 48]. To reduce this barrier, researchers integrate natural language interfaces (NLIs) into VA systems, allowing users to communicate through natural language commands [22, 28, 36, 66, 67, 70, 71, 74, 96]. However, even with NLI

integration, the fundamental complexity of VA systems remains a barrier for novices.

With the advent of LLMs, CVA has emerged as a more appealing alternative for novices to conduct visual data exploration [25, 56, 58, 59, 75, 78]. CVA systems eliminate the complex interfaces of traditional VA systems, enabling users to explore data solely through natural conversation. Recent research has further enhanced CVA through multimodal interaction, preserving conversational intuitiveness while improving operational precision. Existing enhancement methods can be divided into two approaches: pre-interaction methods [17, 55, 80] and post-interaction methods [42, 89]. Pre-interaction approaches improve user input clarity through multimodal capabilities. For example, DirectGPT and VizTA allow users to drag visual elements into dialogue boxes [55, 80], while Interchat enables users to link chart elements to prompts [17]. Post-interaction approaches enable users to easily validate and refine LLM-generated output. Stepwise and Phasewise allow users to verify LLM-generated results and make corrections [42], while WaitGPT provides on-the-fly visualization of generated code for immediate revision [89].

However, while CVA reduces barriers to visual analytics, it also amplifies novices' tendency to become disoriented during their analysis journey. Unlike traditional VA tools (e.g., Tableau) that guide users through graphic user interface, CVA's natural language interface lacks discoverability, leaving novices struggling with what to do next [68]. Additionally, the linear conversational interface complicates analysis context management, causing users to lose their analysis direction during multi-turn conversations [85]. Our research addresses these difficulties by structuring CVA conversations into analysis trees that provide comprehensive support.

2.2 Maintaining Direction in Visual Analytics

Visual analytics is a complex iterative process where users easily lose track of their analysis direction during exploration [61]. To help users maintain direction in this complex process, recent work mainly focuses on data-centric recommendations that suggest insights based on statistical characteristics [49].

One category of work attempts to maintain users' analysis direction by automatically detecting statistical patterns and generating insight-rich visualizations from entire datasets [21, 37, 44, 81, 93]. These systems employ diverse technical approaches: AdaVis uses knowledge graphs with attention mechanisms to recommend adaptive visualizations that guide users' focus [93], while LLM4Vis leverages ChatGPT with demonstration examples to generate appropriate charts that suggest analysis paths [81]. Extending this approach, another category of work incorporates user interaction to provide more targeted guidance, while still relying primarily on statistical characteristics [70, 87, 88, 96]. Voyager2 enables users to specify partial data interests and automatically completes visualizations to guide their next steps [88]. Snowy helps users maintain analysis direction by suggesting deictic queries based on user-selected marks on a chart [70]. LEVA supports focused exploration by generating insights for user-selected chart regions, helping users deepen their analysis in specific areas [96].

However, these data-centric approaches provide isolated recommendations without offering users a global perspective of their analysis journey. Without this perspective, novices struggle to connect individual steps into a coherent logic, leading to repetitive or aimless exploration. To address this limitation, we construct and visualize an analysis tree that provides novices with a navigable view of their analysis journey.

2.3 Maintaining Direction in Conversation

In the natural language processing domain, follow-up question generation (FQG) has emerged as an effective technique for maintaining direction in complex conversational interactions. It aims to generate subsequent inquiries that help users maintain conversational flow, deepen dialogue, explore new perspectives, or seek more precise information [65, 82].

FQG has been applied across various domains, such as medical consultations [29, 51, 84, 86], education [30, 52], and e-commerce [24], demonstrating their effectiveness in maintaining the direction of conversation. Recently, FQG has evolved from rule-based methods [57, 69], which lack diversity, to more sophisticated LLM-based approaches that can generate more diverse questions while maintaining contextual relevance. For example, Liu et al. fuse knowledge graphs with LLMs, using external knowledge to generate more relevant and in-depth questions that better maintain conversation direction [52]; Winston et al. convert domain-specific content into follow-up question examples for in-context learning to improve directional guidance [86]; Dong et al. develop a context-aware model that mimics user behavior to generate follow-up questions for better maintaining conversation direction [24].

However, while FQG produces contextually relevant questions, it cannot yet provide novices with a global view of their analysis journey. Additionally, they are designed for purely textual conversations rather than multimodal CVA environments. Our work addresses both limitations by constructing an analysis tree from multimodal conversations. This tree provides novices with a navigable view of their ananlysis journey and enables contextual guidance based on their current position.

2.4 Graph Visualizations for Complex Tasks

Complex tasks are inherently non-linear, involving backtracking, branching, and parallel exploration [61, 98]. To capture this complexity, researchers in visualization have adopted graph structures to manage analysis history. For example, VisTrails [16], VisFlow [90], and FlowSense [91] use graphs to help analysts track and revisit their analysis steps.

Recently, the rise of LLMs has popularized conversational user interfaces (CUIs) due to their simplicity. However, unlike the graph-based Interfaces mentioned above, CUIs face several issues when applied to complex tasks, such as difficulties with version control [47] and context management [85]. Researchers have attempted to address these limitations by integrating tree and graph visualization with CUIs, applying this approach across writing [46, 95], sensemaking [40, 73], design [18, 72], coding [94] and visual analytics [23, 26, 79, 85]. Specifically, VISAR uses trees to visualize writing outlines, supporting rapid draft prototyping with LLMs [95]; Graphologue and Sensescape use tree and graph structures to visualize textual LLM output, aiding comprehension of lengthy responses [40, 73]; CoExploreDS and Luminate use graphs to visualize

design alternatives, supporting design ideations with LLMs [18, 72]; and NeuroSync uses graphs to visualize LLM reasoning steps, supporting human-LLM alignment in coding tasks [94].

Most relevant to our work, InsightLens and Jupybara use graphs to visualize relationships among discovered insights in LLM-assisted VA tasks [79, 85]. However, they focus on organizing ananlysis results rather than the exploration process. Visualizing the process is critical for novices, as they easily lose track of their previous steps and current status during multi-turn conversations. Ding et al. and Flowco allow users to explicitly construct graphs, either by manually adding nodes or by commanding LLMs, to support data analysis [23, 26]. However, this explicit construction imposes a significant cognitive burden on novices, who may lack the expertise to proactively structure their analysis. To better serve novices, we represent the CVA process as a tree rather than a complex graph and construct it automatically in real time. This approach provides a clear global overview and reduces visual clutter, helping novices maintain direction without the overhead of manual management.

3 Formative Study

The following sections describe our formative study [1]. We begin by describing our methodology and participant demographics (Sec. 3.1). Next, we investigate why novices experience disorientation in CVA and how experts sustain effective analytical navigation (Sec. 3.2). Finally, we distill a set of design goals that informed the development of ToA (Sec. 3.3).

3.1 Study Design

3.1.1 Participants. We recruited 12 participants with varying data analysis experience from different disciplines (5 male, 7 female; age: $M = 27.5$, $SD = 3.3$; see Table 1 in supplemental materials). All participants had experience with data analysis in their work and used ChatGPT regularly (5+ days per week). Based on the data literacy self-efficacy scale (detailed in Sec. 3.1.3), we classified 6 participants as experts (E1-E6) and 6 as novices (N1-N6).

3.1.2 Apparatus and Tasks. We selected ChatGPT (GPT-4o) with advanced data analysis plugin [59] as our experimental platform, representing the state-of-the-art publicly accessible CVA system at the time of our study. We used a job listings dataset (approximately 30,000 records, 7 dimensions) as our experimental dataset. This dataset was selected because it requires no specialized domain knowledge while providing moderate analytical complexity. Additionally, its clean structure minimizes LLM processing errors, ensuring that observed user struggles stemmed from capability issues rather than system limitations. The task required participants to explore job characteristics and recommend a position with data-driven explanations, mirroring the open-ended yet goal-directed nature of visual analytics [11]. Two experts (3+ years VA experience) completed the task in 20 minutes in a pilot study, establishing a performance baseline. The pilot also confirmed the stability of LLM performance on this dataset.

3.1.3 Procedure. Before the study, all participants reviewed and signed an informed consent form which explains the study purpose and procedures, potential risks and benefits, data handling practices, and participants' right to withdraw at any time. All participants followed the designated study protocol and each of them received $10 compensation for their participation. The study consisted of three parts lasting 55 minutes in total:

Introduction and Tutorial (15 minutes): Participants first completed a demographic questionnaire covering age, gender, job, and self-rated expertise in visual analytics. We assessed participants' visual analytics literacy using the data literacy self-efficacy scale [45]. This scale contains 10 subscales. We selected two subscales: data analysis and data visualization. These subscales were selected because they directly relate to the core skills required in our CVA tasks. Following the questionnaire, all participants received a study briefing and completed a hands-on tutorial to familiarize themselves with ChatGPT's advanced data analysis plugin.

Task Completion (25 minutes): Participants performed the CVA task by analyzing job listing dataset and making a job recommendation. After participants provided their final job recommendations, we conducted brief post-task verification, asking them to provide data-driven explanations for their choices. Participants were classified as task failures under three conditions: taking too long compared to experts' 20-minute baseline (e.g., 2 to 3 times longer than experts), being unable to provide explanations, or losing confidence and giving up. Failed participants proceeded directly to the interview phase. We collected screen recordings, observational notes, and queries sent to ChatGPT during task completion for subsequent analysis.

Interview (15 minutes): After completing the tasks, we interviewed the participants with the following questions:

(1) *How did you decide what to ask ChatGPT in each turn?*
(2) *What were the most challenging moments and why?*
(3) *How would you describe your overall analysis workflow?*

3.1.4 Data Encoding. To analyze behavioral differences between experts and novices, we coded participants' queries. We adopted *conversational transitions model* [77] as our initial framework, as it provides well-defined categories based on visualization modifications. Unlike taxonomies classifying queries by data operations [35], this model focuses on visualization changes, better reflecting how users interact with charts in CVA. Two co-authors coded all queries independently. During the process, we found the initial framework insufficient to capture the diverse behaviors in CVA. Consequently, we inductively expanded the taxonomy by developing new categories and subdividing original ones into more fine-grained types. Ultimately, we identified eight query types from 205 queries (see Table 1). Four were adapted from the initial model (*elaborate, dice, reshape, start new*), while four emerged inductively (*slice, clarify, contextualize, fumble*). Excluding *fumble*, the remaining seven types represent clear analytical intents: five involve direct visualization manipulation (*start new*, *elaborate*, *dice*, *slice*, *reshape*), and two reflect broader conversational behaviors (*contextualize*, *clarify*).

3.2 Study Results

Fig. 2 depicts how task success varies with participant expertise and the proportion of vague queries. The figure reveals two distinct groups. Experts (bottom-right, Fig. 2b) had low vague query ratios

[1]This study was approved by the Ethics Committee of the College of Biomedical Engineering & Instrument Science, Zhejiang University (Approval No. 2025-26).

Table 1: Eight query types identified in our formative study. Four derive from conversational transitions model [77]; four emerged inductively. Unlike Hong and Crisan's taxonomy [35], which classifies queries based on how they transform data, our taxonomy classifies queries based on how they transform charts.

Query Type	Definition	Example from User Queries
Start New (27.8%; 57) Derived from *start new*	Creating a new chart that departs from the current chart's analytical thread	*"Now let's look at geographic distribution of jobs"*
Fumble (19.5%; 40) Our induction	Making vague or unclear questions without specific analytical goals	*"Show me something interesting"*
Elaborate (13.2%; 27) Derived from *elaborate*	Augmenting a chart by introducing additional data attributes	*"Add the education level to the (existing) salary distribution chart"*
Dice (11.7%; 24) Derived from *adjust*	Adding filter conditions to a chart to restrict the displayed data subset	*"Show only jobs with salaries above $100,000"*
Slice (9.3%; 19) Our induction	Selecting specific attribute values from a chart as filters for detailed analysis	*"Show me the city distribution for the 'Software Engineer' positions (on this chart)"*
Contextualize (7.3%; 15) Our induction	Incorporating external domain knowledge to interpret insights of a chart	*"Now I have 5 years of experience and a master's degree, which companies would you recommend"*
Reshape (5.9%; 12) Derived from *adjust*	Modifying the visual encodings or transformations of a chart while retaining data attributes and filters	*"Change this pie chart to a bar chart"*
Clarify (5.4%; 11) Our induction	Requesting explanations of the data or visual elements displayed within a chart	*"How many company types does each job title have? Based on the pie chart"*

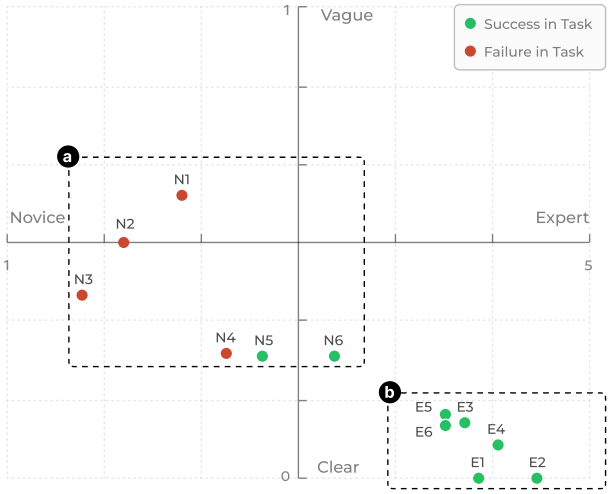

Figure 2: Study results reveal two user groups based on data literacy self-efficacy (horizontal axis: average score on self-efficacy scale) and vague query proportion (vertical axis: *fumble* queries / total queries): (a) novices; (b) experts.

and all succeeded. Novices (top-left, Fig. 2a) had higher vague query ratios and most failed.

3.2.1 Three Fundamental Stages in CVA. To better analyze the root causes of performance differences between experts and novices, we first characterized the CVA process into three fundamental stages based on our empirical observations:

Result Interpretation: Upon receiving AI output, participants examine the charts and text to comprehend the results. Crucially, they identify specific visual elements (e.g., outliers in line charts or prominent bars in bar charts) that indicate promising directions for further investigation. Adopting terminology from follow-up question generation research [30], we term these cues **analytical triggers**, as they stimulate the user's intent to explore further.

Query Formulation: Based on identified triggers, participants develop analytical intentions for further exploration, and then construct queries by translating these intentions into natural language.

Context Management: As conversations progress, participants track prior findings, understand relationships between different analytical paths, and make exploration decisions based on accumulated knowledge, often requiring them to scroll through conversation history to review previous content.

While result interpretation and query formulation occur iteratively per turn, context management persists throughout the entire analytical session. Each stage is critical for success, yet users exhibit significant performance differences. We next analyze how experts and novices perform at each stage.

3.2.2 Novice Weaknesses. Most novices (4/6) failed to complete the CVA task: one spent over 40 minutes without finishing, one provided inadequate explanations, and two abandoned the task (see Fig. 2a). They relied heavily on *fumble* (33.0%) and *start new* (30.8%) queries, indicating frequent confusion and lack of analytical continuity. In contrast, they used fewer *elaborate* (7.7%), *dice* (7.7%),

Table 2: Two categories of analytical triggers identified from formative study, showing how experts systematically recognize exploration opportunities in AI output.

Trigger Type	Definition	Observed Subtype
Presentation Trigger	Visual elements that indicate presentation issues, prompting restructuring requests.	*Occlusion*: overlapping or obstructed visual elements impeding data reading; *Overload*: excessive visual elements exceeding perceptual capacity
Observational Trigger	Visual elements that reveal interesting data patterns or values, prompting further investigation.	*Majority*, *Extreme*, *Outlier*, *Difference*, *Trend*, *Turning point*, *Correlation*, following established fact taxonomy [50, 83]

slice (5.5%), and *reshape* (4.4%) queries. *contextualize* (6.6%) and *clarify* (4.4%) queries accounted for only a small fraction of the interactions. Through interviews and observations, we identified core weaknesses novices face across each stage:

W1: Trigger Insensitivity in Result Interpretation. Novices lack sensitivity to analytical triggers, frequently overlooking them in AI output. For example, N4 noticed San Francisco had the highest average salary but was not triggered to pursue this outlier further, saying *"I just thought there wasn't much to explore."* This trigger insensitivity causes novices to miss exploration opportunities and explains why they frequently used start new (30.8%) to abandon current analysis threads.

W2: Vague Expression in Query Formulation. Even after recognizing analytical triggers, novices struggle to develop clear analytical intentions and translate them into precise queries. For example, N3, after discovering salary differences and wanting to explore further, asked *"Can you show me something interesting about these salary differences?"* rather than *"What's the salary range for software engineer across different cities?"* These vague queries lead to poor analysis quality and incorrect output, explaining both frequent *fumble* queries (33.0%) and task abandonment by two participants who lost confidence in ChatGPT.

W3: Fragmented Analysis Flow Impairs Context Management. Novices exhibit a fragmented analysis flow, frequently switching between analytical directions without establishing logical connections. For instance, N1 repeatedly switched between salary, education, experience, and city analyses without exploring their relationships. He later admitted: *"I remember analyzing lots of charts, but can't find logical connections. It's hard to find why I recommend this (job)."* This fragmentation explains both why one participant spent over 40 minutes without finishing the task and why another provided inadequate explanations.

3.2.3 Expert Strengths. All experts (6/6) successfully completed CVA tasks (see Fig. 2b). Compared to novices, they used fewer *start new* (25.4%) and *fumble* (8.8%) queries, while using more *elaborate* (17.5%), *dice* (14.9%), *slice* (12.3%), and *reshape* (7.0%) queries. Usage of *contextualize* (7.9%) and *clarify* (6.1%) showed no major differences. Experts demonstrated strengths across each stage:

S1: Trigger Sensitivity in Result Interpretation. Experts are highly sensitive to two types of triggers (see Table 2 for completed trigger taxonomy we summaries from interview and observation). For example, when E2 observed densely clustered data points in a scatter plot, this *occlusion* immediately caught his attention. P8 explained: *"When I see these overlapping features, I know that for better readability, it's best to switch to a heatmap."* When E1 observed salary patterns across different company types, this *difference* immediately caught his attention. E1 explained: *"Technology companies seem to pay significantly more than other company types, so I want to explore technology vs. traditional industry salary differences."* This trigger sensitivity enables experts to conduct deeper analysis and explains why they used more follow-up queries (*elaborate, dice, slice, reshape*: 51. 7% combined) than novices.

S2: Purposeful Expression in Query Formulation. After recognizing analytical triggers, experts develop clear analytical intentions and precisely translate them into seven types of queries. E5's queries exemplify this: *"Break down salary by company type (elaborate)"*, *"Focus on technology companies only (slice)"*, *"Show software engineer positions in technology companies across different cities (dice)"*. This precision improves the quality of AI output.

S3: Tree-like Analysis Flow Enhances Context Management. Experts implicitly organize the linear conversation into a tree-like workflow. For example, E6 first analyzed the relationship between salary and experience, discovering large salary ranges for senior positions. He then drilled down into this finding, systematically exploring the experience requirements and education backgrounds specifically for these high-salary senior roles. Only after exhausting this analytical thread did he backtrack to explore a different dimension: how salaries vary across cities. This tree-like workflow enables experts to maintain a clear awareness of their exploration boundaries and rapidly locate key insights within the conversation history, consistent with strategies observed in traditional VA systems [11, 87, 88].

3.3 Design Goals

Informed by the weaknesses of the novices and the strengths of the experts, we derive the following design goals (DGs) to facilitate CVA for novices (see Table 3).

DG1 Extract Analytical Triggers to Guide Further Exploration (derived from W1 and S1). To address novices' trigger insensitivity, we automatically detect and classify the two types of analytical triggers (presentation, observational) that experts systematically identify from AI output. By explicitly surfacing these hidden cues, we reveal exploration opportunities that novices miss and guide them toward promising analytical directions.

DG2 Categorize User Queries to Guide Query Formulation (derived from W2 and S2). To address novices' vague query construction, we classify user queries into the seven distinct types identified in our formative study. By systematically

Table 3: This table summarizes key contrasts observed in our formative study across three stages, highlighting novices weaknesses (W1-W3) and expert strengths (S1-S3) that inform our design goals (DG1-DG3).

CVA Stage	Novice Weakness	Expert Strength	Design Goal
Result Interpretation	**W1: Trigger Insensitivity.** Overlook analytical triggers in AI output	**S1: Trigger Sensitivity.** Identify 2 types of analytical triggers from AI output	**DG1: Extract Triggers.** Extract analytical triggers from AI output to guide further exploration
Query Formulation	**W2: Vague Queries.** Resort to *fumble* when unable to formulate specific questions	**S2: Purposeful Queries.** Use 7 distinct query types for targeted exploration	**DG2: Categorize Queries.** Systematically categorize user queries to guide targeted formulation
Context Management	**W3: Fragmented Analysis Flows.** Jump between different directions without logical connections	**S3: Tree-like Analysis Flows.** Implicitly adopt a tree-like workflow to maintain analysis context	**DG3: Construct Trees.** Construct and visualize analysis trees to enable navigation and prevent fragmented analysis flow

categorizing user queries, we can provide specific query suggestions and help novices formulate targeted questions instead of vague fumble queries.

DG3 Construct and Visualize an Analysis Tree for Navigation (derived from W3 and S3). To address novices' fragmented analysis flow, we restructure and visualize their analysis into an analysis tree that mirrors experts' tree-like workflows. By organizing the CVA process into such a tree with AI outputs as nodes and user queries as edges, we provide navigation support that prevents user disorientation.

4 Algorithm Design

The following sections describe the algorithm design. We first define the data structure for the analysis tree that captures CVA context (Sec. 4.1, **DG3**). Then we present our method for automatically constructing the tree and identifying question triggers from conversation content (Sec. 4.2, **DG1**). Finally, we demonstrate how we generate contextual-aware recommendations by leveraging the analysis tree and identified triggers (Sec. 4.3, **DG2**).

4.1 Analysis Tree Definition

We define an analysis tree (see Fig. 3) as the data structure that encodes the complete CVA context. Given a tabular dataset D and a sequence of n query-answer pairs $\{(Q_1, A_1), (Q_2, A_2), ..., (Q_n, A_n)\}$ in a CVA session, where each Q_i represents the user's natural language query over D and A_i represents the AI output including text and chart, an analysis tree T is constructed as follows:

$$T := \{r, N, E\}$$

where $N = \{answer_1, answer_2, ..., answer_n\}$ contains n nodes corresponding to $\{A_1, A_2, ..., A_n\}$; $E = \{query_1, query_2, ..., query_n\}$ contains n edges corresponding to $\{Q_1, Q_2, ..., Q_n\}$, with each $query_i$ transforming $answer_j$ to $answer_i$, where $j < i$; and $r = answer_0$ represents the root node (initial CVA state).

4.1.1 Node. We define each node as a 3-tuple representing the AI output A_i, where "?" indicates optional elements:

$$answer_i := \{chart_i?, answerText_i, triggers_i\}$$

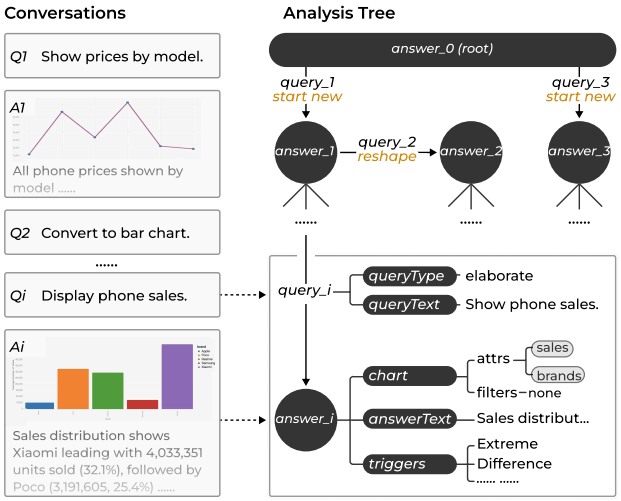

Figure 3: Structuring linear CVA conversations (left) into analysis tree (right). AI outputs (A_i) become nodes $(answer_i)$ containing charts, triggers, and text. User queries (Q_i) become edges $(query_i)$ that connect nodes and capture query types and original text.

$chart_i$ is a 2-tuple representing the chart in AI output A_i:

$$chart_i := \{attrs_i, filters_i?\}$$

where $attrs_i$ is data attributes from dataset D displayed in the chart, $filter_i$ is a filter applied to dataset D. We use only these two components to represent charts because $\{attrs_i, filters_i\}$ precisely captures what data users are currently viewing [49], which is essential for understanding how users move between analysis steps. For example, if a bar chart shows job titles and salaries, $attrs_i$ could be job title and salary columns, $filter_i$ could be a selected experience range or specific company types, $chart_i$ is null for text-only output.

$triggers_i$ is an array representing analytical triggers detected in the AI output A_i, where each element is defined as:

$$trigger := \{triggerType, triggerContent\}$$

with $triggerType$ represents two trigger categories identified from our formative study: *presentation, observational*, and $triggerContent$ containing a textual description of each trigger.

4.1.2 Edge. We define each edge as a 4-tuple representing user query Q_i:

$$query_i := \{answer_j, answer_i, queryType_i, queryText_i\}$$

where $answer_j$ represents the source node that is transformed to $answer_i$ (the target node) through user query Q_i, with $j < i$, $queryType_i$ represents seven candidate query types identified from our formative study: six types *(elaborate, dice, reshape, slice, clarify, contextualize)* that connect to any previous node, and one special type *(start new)* that exclusively connects to the root node r, and $queryText_i$ contains the original text of user query Q_i.

4.2 Tree Construction

Given an existing tree T and a new query-answer pair (Q_i, A_i), the algorithm determines the appropriate parent node while extracting analytical triggers (see Alg. 1 in supplemental materials). We decompose the tree construction process into two stages: node generation (see Fig. 4a) and insertion (see Fig. 4b).

4.2.1 Node Generation. Given new query-answer pair (Q_i, A_i), we construct node $answer_i$ as follows. First, we extract $answerText_i$ from the AI output A_i. Then, if A_i contains a chart, we extract $attrs_i$ and $filter_i$ from the chart specifications. Next, we identify triggers in A_i using multi-modal question answering capability of LLMs [38]. We provide LLMs with our trigger definitions and few-shot examples [20]. The LLM performs multiple choice selection to determine which triggers are present in the chart, then generates corresponding textual descriptions for the identified triggers. We then store these triggers in $triggers_i$. Finally, we construct the node as: $answer_i := \{chart_i?, answerText_i, triggers_i\}$.

4.2.2 Node Insertion. Given a new node $answer_i$, we find its parent node using the connection rules from Table 4. For each existing node $answer_j$ in the tree, we test whether it satisfies any of the six connection rules. If multiple nodes satisfy the rules, we select the most recent one and connect $answer_i$ to it with the corresponding query type. If no existing node satisfies any rule, we connect $answer_i$ to the root with query type *start new*. We complete the process by creating an edge $query_i := \{answer_j, answer_i, queryType, Q_i\}$ and updating the tree: $T' = T \cup \{answer_i, query_i\}$.

4.3 Query Recommendation

Given an analysis tree T, a tabular dataset D, and the current node $answer_i$, the algorithm generates a set of three recommended edges $\{edge_{rec1}, edge_{rec2}, edge_{rec3}\}$ that help novices identify promising next steps (see Alg. 2 in supplemental materials). We decompose the query recommendation process into two stages: trigger selection (see Fig. 4c) and query generation (see Fig. 4d).

4.3.1 Trigger Selection. Given the current node's triggers $triggers_i$ and the analysis context of ancestor nodes, this stage outputs a candidate trigger set $\{trigger_{rec1}, trigger_{rec2}, trigger_{rec3}\}$. We first collect analysis context by extracting data attributes and filters from all ancestor nodes along the path from root to current node. We then prompt the LLM with the exploration context and a multiple-choice question over $triggers_i$, asking it to select the three triggers most likely to lead to new discoveries. Finally, we identify three recommended triggers: $triggers_{rec} := \{trigger_{rec1}, trigger_{rec2}, trigger_{rec3}\}$.

4.3.2 Query Generation. Given a selected trigger $trigger_{rec}$, the current node $answer_i$, and tabular dataset D, this stage outputs a recommended edge $edge_{rec} := \{answer_i, answer_{rec}, queryType_{rec}, queryText_{rec}\}$. First, we provide the LLM with $trigger_{rec}$ and dataset D, asking it to select the most appropriate query type from six identified types (excluding *start new*) and generate the corresponding target node $answer_{rec} = \{chart_{rec}?, \emptyset, \emptyset\}$, where $chart_{rec}$ contains the predicted data attributes and filters. Second, the LLM generates $queryText_{rec}$ based on the selected query type and the predicted chart content. Finally, we construct the recommended edge: $edge_{rec} := \{answer_i, answer_{rec}, queryType_{rec}, queryText_{rec}\}$.

5 Interface Design

To achieve **DG3**, we designed an interface that visualizes the analysis tree to provide structured navigation support. It consists of two main parts: an interactive tree panel for managing the CVA context (Sec. 5.1) and an enhanced conversation panel integrating tree-based query recommendations (Sec. 5.2) We then describe our implementation details (Sec. 5.3).

5.1 Tree-Based Navigation Panel

The tree navigation panel (Fig. 5a) addresses exploration fragmentation by visualizing the CVA process as an interactive analysis tree. Nodes, rendered as rectangles, represent AI outputs ($answer_i$) and display the corresponding data attributes ($attr_i$) and filters ($filters_i$). Edges connect these nodes to represent user queries, labeled with types (e.g., *"Start new", "Elaborate"*) to depict analytical transitions. The toolbar (Fig. 5c) supports two primary interactions: the filter tool, which allows users to select multiple nodes to isolate specific conversation threads, and the selection tool, which enables direct navigation to any previous analysis step. Additionally, hovering over a node (Fig. 5b) reveals a floating window with available triggers ($triggers_i$). Clicking the "REC." button next to a trigger generates targeted query recommendations, which simultaneously appear as interactive cards in the conversation panel and dotted candidate nodes in the analysis tree.

5.2 Enhanced Conversation with Recommendations

The enhanced conversation panel (Fig. 5d) augments standard CVA interfaces by integrating tree-based query recommendations. When the LLM returns a response with charts and textual descriptions, the system also provides three recommended follow-up queries. Additionally, users can manually generate targeted recommendations by hovering over nodes in the tree panel and clicking the "REC." button next to specific triggers. These recommendations appear as interactive cards (Fig. 5e) and are simultaneously rendered as dotted nodes ($answer_{rec}$) and edges ($edge_{rec}$) in the tree panel (Fig. 5f) to indicate future exploration directions. Each card features a question mark icon that reveals the reasoning behind the recommendation.

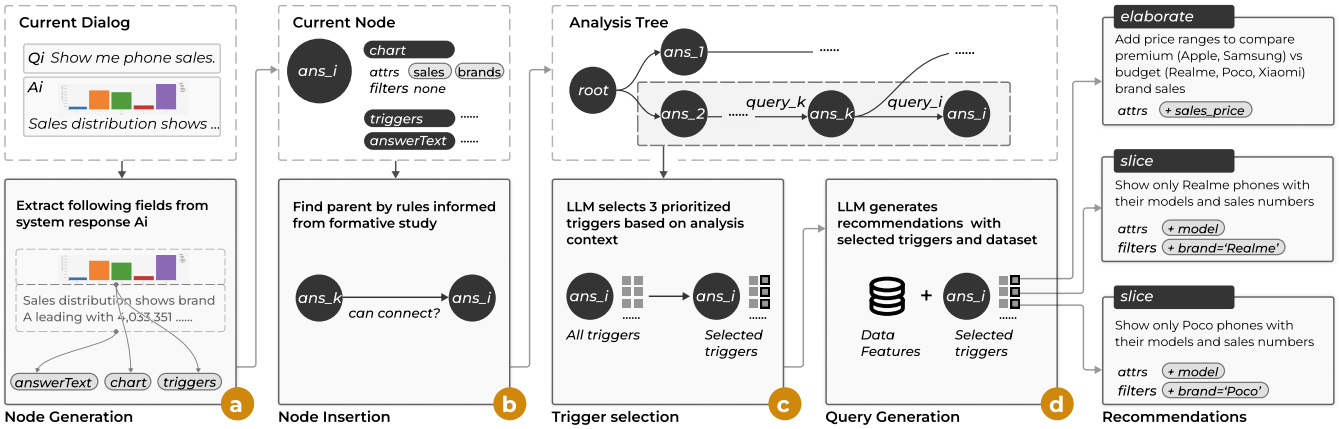

Figure 4: Overview of the analysis tree construction and query recommendation process. When a new QA pair arrives, (a) our algorithm first extracts information from the QA pair and generates a new node, (b) then inserts this node into the existing tree, completing the tree construction. For query recommendations, (c) the algorithm first selects three promising triggers from the current node, using ancestor nodes as context, (d) then combines these triggers with the dataset to generate six types of recommendations (the figure shows *elaborate*, *slice*, and *slice* as examples).

Table 4: Connection rules for inserting new nodes into a analysis tree. When adding a new node, we test each existing node to see if it satisfies any of the six connection rules. Each rule corresponds to a specific query type. The first four rules compare data attributes and filters using rule-based logic, while the last two assess text-chart relationships using LLMs' multiple choice capability. When both *slice* and *dice* conditions are satisfied simultaneously, *slice* takes precedence. If no rule is satisfied, the node connects to the root as a new branch.

Query Type	Rule
Elaborate	$child.chart \neq null$ AND $parent.attrs \subset child.attrs$ AND $parent.filter = child.filter$
Dice	$child.chart \neq null$ AND $parent.attrs = child.attrs$ AND $parent.filter \subset child.filter$
Reshape	$child.chart \neq null$ AND $parent.attrs = child.attrs$ AND $parent.filter = child.filter$
Slice	$child.chart \neq null$ AND $child.filter$ derived from $parent.attrs$
Clarify	$child.chart = null$ AND $child.answerText$ explains $parent.chart$
Contextualize	$child.chart = null$ AND $child.answerText$ connects $parent.chart$ to domain knowledge
Start New	No existing node satisfies any rule above $\rightarrow$ connect $child$ to r

This transparency promotes interpretability, helping novices understand expert exploration strategies and gradually cultivate their own analytical thinking.

5.3 Implementation

We implemented ToA using a modern web technology stack with a decoupled frontend-backend architecture. The frontend employs Vue3 [8] with TypeScript and Vite. For visualization, we use Vega-Lite [7] for statistical charts and G6 [4] for interactive conversation trees. The backend consists of a Django-based RESTful API service [1]. We selected Google Gemini 2.5 Flash (June 2025) [32] for its rapid reasoning and leading performance across most chart understanding tasks [39]. We employed LangChain [5] to orchestrate structured LLM interactions through carefully designed prompt engineering, enabling chained processing for chart generation, operation type classification, and recommendation tasks. ToA leverages pandas [6] for data processing and analysis.

6 Usage Scenario

Raj, a new Flipkart seller with no data analysis background, wants to understand mobile phone market trends before deciding his inventory strategy but doesn't know where to start. He uploads the mobile phone sales dataset into ToA, relieved to see a simple interface rather than complex statistical tools he's struggled with before. The tree panel shows a root node, ready to guide his exploration.

Unsure how to phrase analytical questions, Raj types a simple prompt: "Show me phone sales." ToA understands his intent and generates a bar chart showing sales distribution across brands. As the output appears, a new node branches from the root, connected by an edge labeled "Start new." Below the output, three recommendation cards immediately appear. Raj is grateful not having to formulate follow-up questions himself. One recommendation catches his eye: "Add price to compare premium (Apple, Samsung) vs budget (Realme, Poco, Xiaomi) brand sales *(Elaborate)*". Curious about the reasons, Raj clicks the question mark icon on the card.

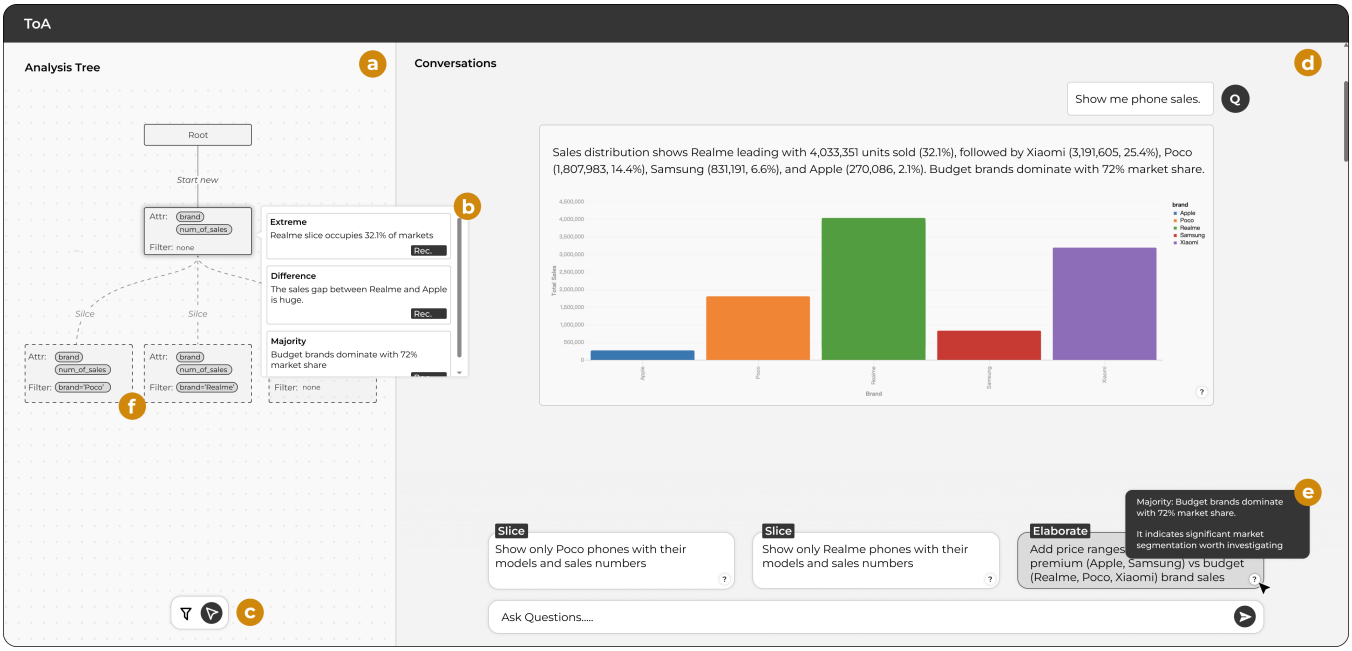

Figure 5: User interface of ToA: (a) Tree navigation panel for exploration overview; (d) Conversation panel with contextual recommendations. Upon query submission, new outputs appear in the conversation with recommendation cards (e), which are simultaneously previewed as dotted nodes (f) in the tree.

A tooltip explains: "*Majority*: Budget brands dominate with 72% market share. It indicates significant market segmentation worth investigating." This explanation helps Raj understand why this direction warrants investigation (see Fig. 5e).

As Raj continues exploring through recommendations, accepting one after another, the tree grows with multiple branches. After several steps, he examines a scatter plot showing phone prices versus battery capacity. The recommendation cards below this chart do not interest him. Wanting to see what other opportunities this node offers, he hovers over this node to expand all triggers. Among several triggers, he spots an intriguing *Correlation*: "Battery capacity shows negative correlation with price" Raj clicks "REC." next to this trigger, and a new recommendation immediately appears: "Explain why premium phones have smaller batteries *(Contextualize)*." He accepts this recommendation, and ToA generates a text explanation about how premium phones prioritize design thinness over battery capacity (see Fig. 6a).

Raj continues using the selection tool to jump between nodes and the "REC." button to generate further recommendations. The tree rapidly expands with branches spreading in multiple directions, over seven nodes now populate the tree. Raj feels overwhelmed by the fragmented information scattered across different analytical threads. Staring at the tree, Raj clicks through various nodes reviewing his previous explorations. One bar chart suddenly catches his attention. It shows sales-to-price ratios across brands, and Xiaomi stands out with a remarkably high value. Despite being positioned in the low-price segment, Xiaomi achieves strong sales performance far exceeding other brands. This counterintuitive finding could significantly impact his inventory strategy. However, his Xiaomi-related discoveries are now buried within the lengthy conversational history alongside analyses of other brands and features. Wanting to focus specifically on Xiaomi without interference from other analytical threads, he activates the filter tool and selects the three Xiaomi-related nodes (see Fig. 6b1 b2 b3). The conversation reorganizes to show only these three Xiaomi analyses in sequence, helping him see exactly what aspects of Xiaomi he has already. With this focused view, Raj enters the query: "Analyze Xiaomi's processor choices by model." ToA processes the query with the three filtered nodes as context. This generates a grouped bar chart revealing that the top-selling Xiaomi models predominantly use Qualcomm processors. Building on this insight, Raj continues using the filter to explore multiple related dimensions of Xiaomi. Through this focused exploration, he ultimately discovers that Xiaomi's best-selling models combine Qualcomm processors with 3GB RAM, priced around $8,500 (see Fig. 6b).

By the session's end, Raj can visually see his entire exploration journey through the analysis tree. With this global perspective, he identified his most valuable finding: Xiaomi models with Qualcomm processors and 3GB RAM at $8,500 represent the optimal inventory choice for his business. Raj successfully decides to focus on these specific configurations, achieving actionable insights despite never writing complex queries or understanding statistical concepts.

7 Algorithm Evaluation

ToA's effectiveness depends on two critical tasks: trigger detection for identifying analytical opportunities, and tree construction for maintaining analysis context. Trigger detection requires LLMs to comprehend and extract data features from a chart. Previous studies

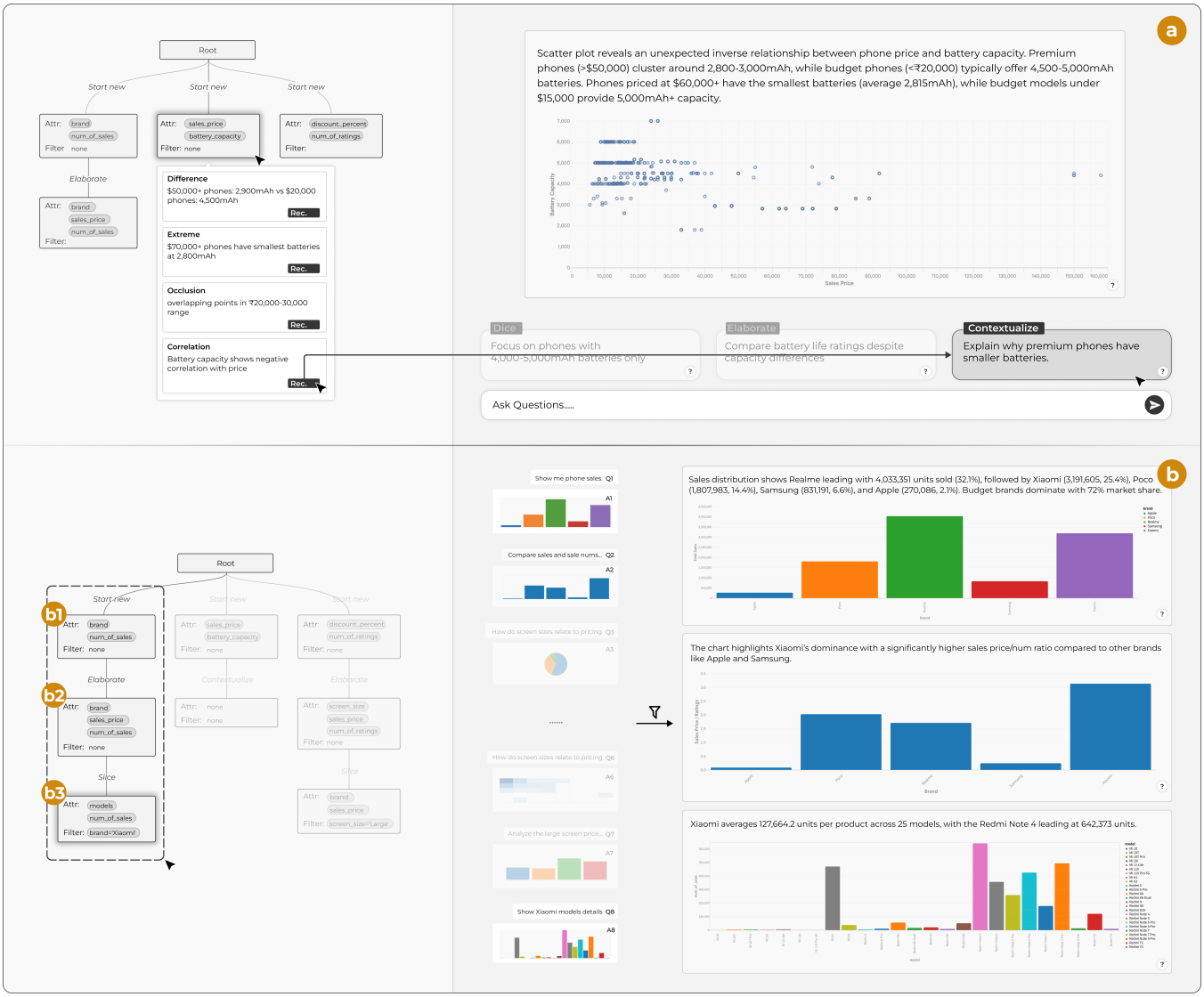

Figure 6: Raj's usage scenario: (a) Clicking "REC" next to a trigger generates a targeted recommendation. (b) Selecting multiple nodes filters the conversation panel to show only those analyses, helping Raj focus on related findings.

demonstrate that LLMs outperform human baselines on such tasks (e.g., 50.17/53 vs. 28.82/53 on VLAT [20]). Given that LLM performance is already well-established, re-evaluating it offers limited theoretical contribution. Therefore, our evaluation focuses on tree construction. Tree construction uses six connection rules (Table 4) to determine parent nodes. The first four rules (*elaborate*, *dice*, *reshape*, *slice*) compare data attributes and filters deterministically, achieving theoretical error-free performance. The challenge lies in the last two rules, *clarify* and *contextualize*, which require aligning a text description with a specific parent chart. This becomes particularly difficult in CVA, where users iteratively refine analyses and generate sequences of highly similar charts (e.g., bar charts differing only in filters). The LLM[2] must distinguish among similar candidates based on subtle semantic cues. Existing datasets (e.g., [9]) contain diverse but distinct charts and cannot support this task. We therefore constructed a CVA-specific dataset (Sec. 7.1) and evaluated the last two rules (Sec. 7.2).

7.1 Dataset and Ground Truth

We extracted complete analysis logs from the two participants who generated the most charts (19 and 20 respectively), providing a stress test for high-load scenarios. The dataset comprises 39 charts: 25 bar charts, 4 line charts, 2 scatter plots, 3 boxplots, 3 heatmaps, and 2 pie charts. To establish ground truth, two authors with rich research experiences (> 2 years) in VA wrote text descriptions based on Lundgard et al.'s visualization semantic model, a description taxonomy published in IEEE TVCG [54]. Each author handled one

[2]We evaluated Google Gemini 2.5 Flash, which is used in our system implementation.

Table 5: Tree construction evaluation results. Task 1 (parent identification) achieved 71.37% accuracy across four semantic levels. Task 2 (type classification) achieved 99.10% accuracy. System overall achieved 70.73% accuracy.

Task 1: Parent Identification	**L1: Elemental**	**L2: Statistical**	**L3: Perceptual**	**L4: Contextual**	**Overall Acc.**
Matching Accuracy	84.62% (99/117)	71.79% (84/117)	64.96% (76/117)	64.10% (75/117)	**71.37% (334/468)**
Task 2: Type Classification	**Clarify**		**Contextualize**		**Overall Acc.**
Metric Results (P / R)	99.23% (258/260) / 99.61% (258/259)		98.64% (73/74) / 97.33% (73/75)		**99.10% (331/334)**
System Overall	End-to-End Accuracy: **70.73%** (331/468)				

participant's data and cross-reviewed the other's work to reach a consensus. This model defines four description levels: L1 describing elemental features like chart type, L2 describing statistical patterns like outliers, L3 describing perceptual observations like trends, and L4 describing contextual domain knowledge. Accordingly, the first three levels (L1–L3) map to *clarify* as they describe intrinsic visual facts, while the last level (L4) maps to *contextualize* as it describes external knowledge. Following this model, two authors created four levels of descriptions for each chart (156 text-chart pairs total).

7.2 Tasks and Results

We evaluated tree construction using 156 text-chart pairs. To simulate real-world CVA scenarios, we restricted the search space to charts generated prior to the current step, with candidates presented in a randomized order. For example, when processing participant A's text for the 5th chart, the search space was limited to A's first five charts. Evaluation comprised two sequential tasks: (1) parent identification: matching the text to the target chart; (2) type classification: classifying the connection type (*clarify* or *contextualize*). For robustness, we repeated the experiment three times. We report the matching accuracy for Task 1, and the classification precision, recall and accuracy for Task 2.

Table 5 shows tree construction performance. Parent identification achieved 71.37% accuracy overall. Accuracy decreased from L1 to L4 descriptions, indicating that matching is more reliable for elemental descriptions but becomes challenging for contextual descriptions. Qualitative analysis of 134 failures revealed three error types. Context similarity interference (58 cases, 42.34%) was the primary and CVA-specific errors. In these instances, the model incorrectly selected a high-similarity distractor, struggling to resolve the visual ambiguity between fine-grained chart variants. In contrast, the remaining failures occurred when the model failed to identify any match, attributed to fundamental chart misinterpretation (39 cases, 29.10%) or text misinterpretation (37 cases, 27.61%) where obvious features or descriptions were not recognized. Type classification achieved 99.10% accuracy (two *contextualize* misclassified as *clarify*; one *clarify* misclassified as *contextualize*). This indicates the primary challenge is identifying the correct parent node rather than distinguishing between *clarify* and *contextualize*. Although overall accuracy was 70.73%, real-world performance should be higher. In practice, the algorithm prioritizes recent charts rather than searching randomly, which aligns with users' tendency to explore based on latest results and reduces interference from similar historical charts. Future improvements could enhance accuracy through fine-tuning the model on CVA-specific datasets to better capture the nuances of distinguishing similar charts in iterative analysis contexts [60].

8 User Study

The following sections describe our user study [3]. To evaluate how effectively our analysis tree prevents novices from getting lost during CVA, we conducted a user study. We assessed this effectiveness through three key indicators: analysis performance improvements, user perception changes, and behavioral pattern shifts during CVA. The following sections describe our experimental setup and methodology (Sec. 8.1), present both objective performance results and subjective perception findings (Sec. 8.2)

8.1 Study Design

8.1.1 Participants. We recruited 12 novices (6 male, 6 female; age: $M = 23.1$, $SD = 2.7$, see Table 2 in supplemental materials) from diverse backgrounds, such as cultural heritage restoration and industrial design. Our screening criteria included: (1) average scores below 3.0 on the data analysis and visualization subscales of the data literacy self-efficacy scale [45]; (2) no background in data science or statistics; (3) using conversational AI 5+ days per week; (4) no participation in our formative study.

8.1.2 Baseline and Apparatus. We created the baseline system by removing the analysis tree from ToA (see Fig. 7). Without analysis tree, the baseline generates recommendations based on conversation history and dataset features. Meanwhile, the baseline shares the same appearance and AI capability as ToA, ensuring that observed differences can be attributed solely to the analysis tree. All participants completed tasks in a standardized environment. They used identically configured desktop computers with two monitors.

8.1.3 Tasks. We designed two CVA tasks using distinct datasets: Task A utilized a flight price dataset [2], while Task B employed the mobile phone market dataset described in Sec. 6 [3]. Each task comprised two stages:

Exploration: Participants conducted open-ended exploration using the system to discover valuable insights. Task A focused on key factors affecting flight pricing, while Task B examined consumer purchasing behaviors. Participants had access to all system functionalities during this phase.

Summarization: Participants organized insights from their exploration and documented findings in a structured report with

[3]This study was approved by the Ethics Committee of the College of Biomedical Engineering & Instrument Science, Zhejiang University (Approval No. 2025-26).

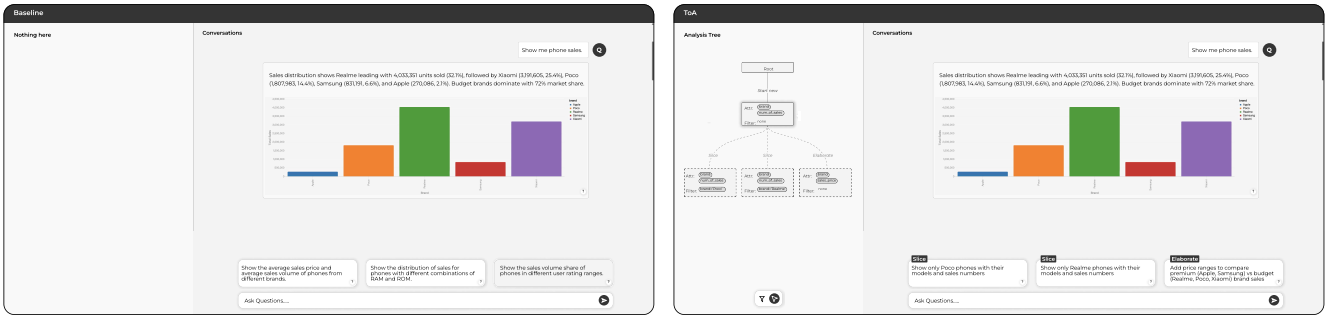

Figure 7: A comparison between the baseline system (left) and ToA (right). Both systems support conversational visual analytics and follow-up recommendations. The baseline differs only by removing the analysis tree.

data-driven evidence. Following established categorization frameworks in visual analysis [34, 53], participants were informed that insights can be observations, hypotheses, and generalizations. Participants could use all system functions but submitting new queries during this phase.

This two-stage task simulates common real-world VA scenarios, where users typically conduct in-depth data analysis first, and subsequently report findings to colleagues for further discussion [92].

8.1.4 Procedure. We employed a counterbalanced within-subjects design, with 12 participants randomly assigned to two groups (6 participants each). Group I used the baseline system for Task A, followed by ToA for Task B; Group II used ToA for Task A, followed by the baseline system for Task B. Before the study began, all participants reviewed and signed an informed consent form and were informed that they could withdraw at any time. The experimental procedure was as follows: participants first received a system introduction and demonstration (5 minutes), then completed the exploration phase (25 minutes) and summarization phase (10 minutes), followed by questionnaires (5 minutes). After a 10-minute break, they repeated the same sequence for the second analysis task. Finally, participants engaged in a semi-structured interview (15 minutes). Researchers conducted one-on-one observations, taking detailed behavioral notes and providing necessary technical support. All task processes were captured through screen recording, interview content was audio recorded, and observational notes were documented. Each participant received $15 compensation for their participation.

8.1.5 Measures. We assessed task performance using objective metrics: task outcomes, total conversation turns, and per-turn thinking time (interval between system response and next query). We measured user perceptions using two instruments: the NASA-TLX for workload, and a custom 7-point Likert scale for system usefulness (Q1–Q5) and logic alignment (Q6; ToA only). We also analyzed screen recordings to examine behavioral changes in exploration and navigation. Semi-structured interviews complemented these quantitative measures, offering qualitative insights into participants' ratings and tree usage patterns.

8.1.6 Data Analysis. We employed insight-based evaluations to analyze participants' task outcomes [63, 64]. To process multiple insights reported in a single statement, two researchers independently segmented all task outcomes into individual insight units and classified them according to the three predefined types, with disagreements resolved through discussion. We then calculated total insight counts for each participant. For continuous measures, we used paired t-tests to analyze differences between the two systems. For discrete measures, we employed Wilcoxon signed-rank tests. Finally, thematic analysis [14] was applied to interview transcripts and observational notes to further uncover insights into user behavior and interaction patterns.

8.2 Study Results

Both systems demonstrated improvement compared to our formative study. This improvement may be attributed to recommendation systems enhancing the discoverability of CVA, enabling users to identify potential next steps more easily. This finding aligns with previous research [22, 68]. However, ToA showed further advantages: all 12 participants successfully completed both tasks when using ToA, while one participant (P6) abandoned the baseline system task midway. During the semi-structured interview, P6 explained: *"After exploring price patterns, departure time trends, and arrival times for about 15 turns, when I asked follow-up questions, the system seemed confused and failed to correctly understand my intent. The incorrect output further amplified my confusion, so I lost confidence in the system and abandoned the task."* In contrast, P6 successfully completed the task using ToA. The filter tool of ToA allowed him to selectively focus on specific nodes for deeper analysis, maintaining a clear analysis context. This highlights the importance of filter tool in maintaining analytical focus.

8.2.1 Task Completion. Fig. 8 shows the performance differences between baseline system and ToA. Participants using ToA generated more insights overall, averaging 16.750 insights per person ($SD = 6.166$) compared to 14.333 insights ($SD = 8.520$) with baseline system. More importantly, ToA significantly improved per-turn insight productivity: participants generated 1.167 insights per turn ($SD = 0.353$) versus 0.737 insights per turn ($SD = 0.404$) with baseline system ($p = 0.002, t = 4.056, d = 1.171$). This represents a 58.3% improvement. However, this improved productivity came with a trade-off in interaction speed. Participants spent significantly longer per turn when using ToA, averaging 71.311 seconds ($SD = 22.243$)

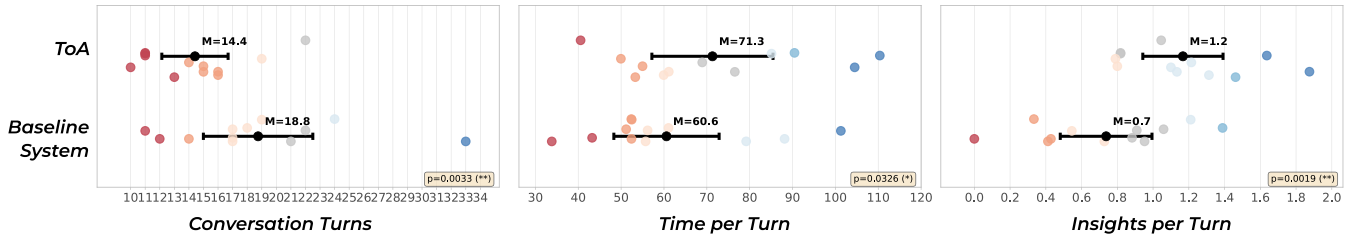

Figure 8: A comparison of task completion efficiency between baseline system and ToA over three utility metrics. Each scatter point in the graph represents one individual record from one participant. The black dots and black lines indicate the mean and 95% confidence interval for each metric.

compared to 60.587 seconds ($SD = 19.329$) with baseline system ($p = 0.033, t = 2.444, d = 0.706$). This represents a 17.7% increase in per-turn thinking time. Consequently, participants completed significantly fewer conversation turns with ToA, averaging 14.417 turns ($SD = 3.579$) compared to 18.750 turns ($SD = 5.910$) with baseline system ($p = 0.003, Z = -2.938, r = -0.600$). Notably, one participant (P11) engaged in 33 turns with the baseline versus 22 with ToA, yet generated fewer insights per turn with the baseline (0.73 vs 1.05). This suggests that the additional turns in the baseline system reflected exploration inefficiency rather than deeper analysis. Behavioral observations and Interviews revealed distinct interaction patterns that may explain these performance differences, which we discuss in Sec. 8.2.3.

8.2.2 User Perception. NASA-TLX results (see Fig. 9) revealed significant workload improvements with ToA compared to baseline system. Participants reported significantly lower mental demand ($p = 0.021, Z = -2.291, r = -0.468$), effort ($p = 0.039, Z = -2.057, r = -0.420$), and improved performance ($p = 0.021, Z = -2.242, r = -0.458$) when using ToA. We found no significant effects on physical demand, frustration, and temporal demand. However, visual analysis revealed that ToA showed lower physical demand ($M = 5.917$ vs. $M = 7.167$) and frustration ($M = 6.500$ vs. $M = 9.333$) compared to baseline system, while temporal demand remained identical ($M = 10.083$ vs. $M = 10.083$).

The usefulness scale (see Fig. 10) revealed significant improvements in gaining orientation and building mental maps. ToA achieved significantly higher ratings for overview tracking (Q1: $p = 0.021, Z = -2.308, r = -0.471$), interested insights tracing (Q2: $p = 0.005, Z = -2.810, r = -0.574$), interstep relation grasping (Q3: $p = 0.017, Z = -2.382, r = -0.486$), and alternative perspectives spotting (Q4: $p = 0.007, Z = -2.716, r = -0.554$). ToA also showed improvement in directions identifying (Q5: $p = 0.092, Z = -1.683, r = -0.344$), though this difference was not statistically significant. Regarding tree-user alignment, most participants (9/12) rated the analysis tree above the neutral point (Q6: $M = 5.167, SD = 1.115$), indicating general alignment with their analysis processes. Interviews revealed several perceived misalignments, which we discuss in Sec. 8.2.4.

8.2.3 User Behavior. Analysis of screen recordings, observational notes, and interview typescripts revealed distinct behavioral patterns when participants used ToA.

ToA Promotes More Proactive Analysis during Exploration. Interviews revealed how ToA transformed participants' interactions with system recommendations. Without the tree, many participants (6/12) described a "click-first, reflect-later" approach to avoid mental effort. P12 explained, *"The suggestions looked fine, but I couldn't tell how they related to my current analysis. I kept clicking anyway just to move forward."* This strategy often increased confusion instead of aiding analysis. P12 added, *"It sometimes made me more confused."* With the tree, however, only one participant (1/12) continued this click-first pattern. We believe this behavior changed for two reasons. First, ToA enabled participants to see how recommendations connected to their current steps, making it easier to evaluate relevance before acting. As P7 noted, *"I could see the path I was on and where the recommendation would take me, so I only clicked if it was actually related."* Second, ToA enabled the system to provide more relevant suggestions based on the user's current context. P5 mentioned, *"the recommendations seemed to match my current focus more"*. Together, these factors encouraged users to make more proactive decisions, which may explain the increased thinking time.

ToA Encourages Retrospective Browsing during Exploration. Observations revealed how ToA changed participants' engagement with analysis history. Without the tree, many participants (8/12) never reviewed previous steps because scrolling through the linear conversation history was cumbersome. As P4 noted, *"I sometimes forgot what I had asked before, and it was hard to review this. I ended up not thinking about it anymore."* With the tree, this pattern shifted. Only a few participants (2/12) did not review their analysis history, while others engaged with previous analysis steps during exploration. P4 explained, *"With the tree, I could quickly see what I had explored before without having to scroll through everything. It made me more willing to look for things I might have missed earlier."* Notably, a subset of participants (3/12) transitioned to using the analysis tree as their primary workspace instead of the conversation interface. Rather than manually inputting queries or scrolling through conversation history, these users continuously navigated to specific nodes to examine analytical triggers. They then leveraged these triggers to generate follow-up questions. As P9 explained, *"I was always worried I might miss something, so I kept searching through the tree to make sure I hadn't overlooked any important insights."* This approach encouraged users to actively review previous steps, which may explain both the longer thinking time and higher insight productivity.

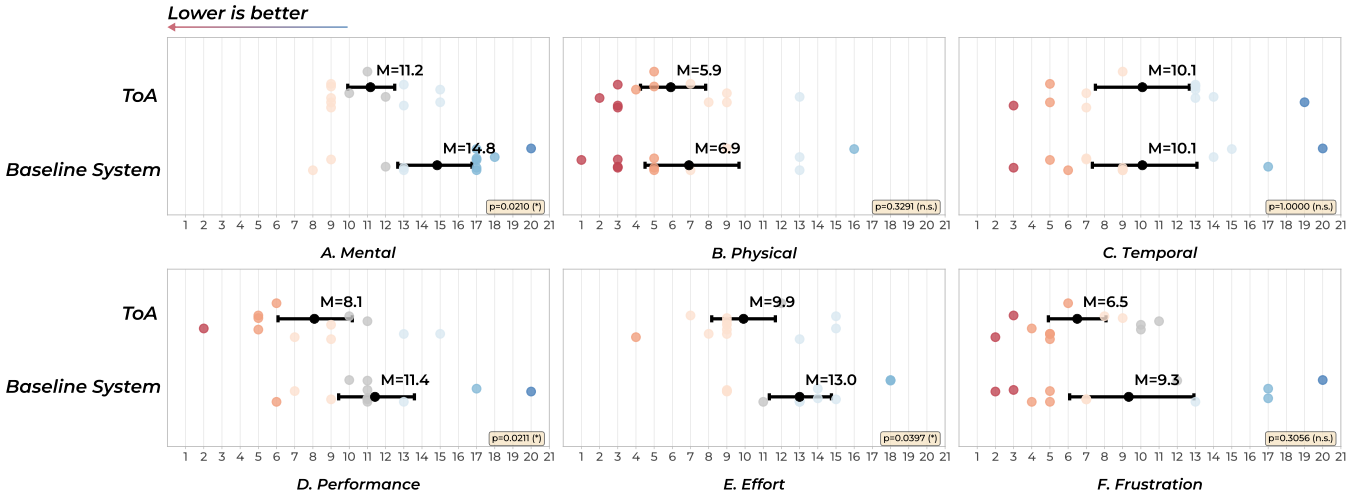

Figure 9: A comparison of perceived workload between ToA and baseline system over NASA-TLX metrics. Each scatter point in the graph represents one individual record from one participant. We map the values with a divergent color scheme from red (1-very low) to blue (21-very high). The black dot and the black line indicate the mean and 95% confidence interval of each metric. Lower scores indicate better outcomes for all dimensions.

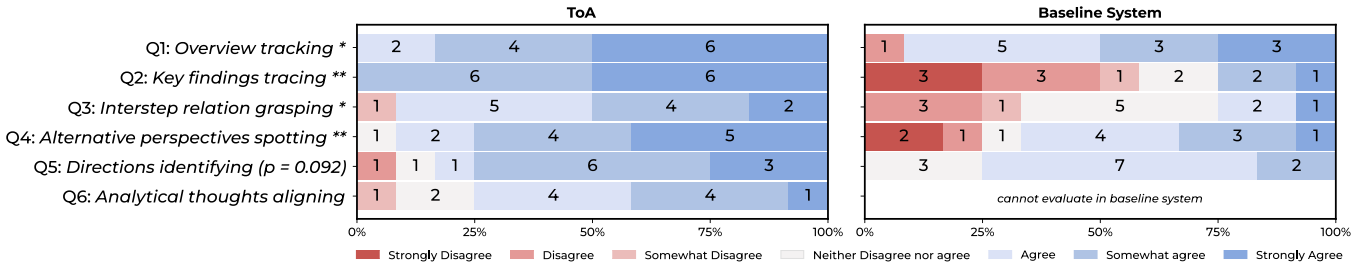

Figure 10: User usefulness ratings on the baseline system and ToA, with numbers indicating participant counts for each rating level. Q6 evaluates the alignment between the tree structure and users' analysis processes (ToA condition only).

ToA Facilitates Comparative Analysis during Summarization. Observations revealed how ToA supported comparative analysis during the summarization phase. Without the tree, some participants (5/12) followed a linear summarization approach, reviewing their analysis sequentially from beginning to end. As P11 who asked 33 questions noted, *"I had asked many questions before, and I originally wanted to look back at the previous charts (during summarization), but I didn't know exactly where to find them so I gave up."* With the tree, only one participant (1/12) continued this linear browsing; the rest used the tree's quick navigation to compare findings across different steps. As P11 explained, *"When I saw Vistara's high prices, I recalled something about their market volume but wasn't sure. I used the tree to confirm and found that Vistara had the lowest presence, explaining their premium strategy. Without the tree, I probably wouldn't have reached this conclusion."* This approach encouraged users to perform cross-reference comparisons, which may explain the higher per-turn insight productivity.

8.2.4 Interview Feedback. Semi-structure interviews provided valuable user suggestions for further improvement.

Participants Perceived Misalignments between Analysis Logic and ToA. Some participants (4/12) found that nodes connected through *reshape* queries introduced unnecessary complexity. Since these queries modify visual encodings without altering the underlying data, participants questioned whether they warranted separate tree branches. As P3 explained, *"Since the right side already shows the time sequence, the left tree doesn't need to represent timing again. When I see reshaped connections, it feels more like the tree is showing a sequence rather than conveying logic. It would be better to simply update the original node."*

A few participants (3/12) reported that some nodes should have multiple parents. For example, P10 noted, *"I had analyzed phone brands earlier to see which were most popular, and separately looked at battery sizes to understand power specifications. When I wanted to explore how battery capacity varies across different brands, I felt like this analysis should connect to both those previous nodes since I was combining brands and battery data together. But the tree only showed it connecting to my most recent query, making it look like a simple follow-up instead of a synthesis of two different analysis threads."*

One participant (P8) noted that ToA could not effectively represent his associative thinking. P8 felt that departure times, arrival times, and flight duration were conceptually connected, but *"the system treated them as separate analyses of three distinct data dimensions, failing to recognize the conceptual link between them."*

Participants Preferred Editable ToA. Half of the participants (6/12) felt that allowing manual tree modifications would be a valuable improvement. They wanted to manually adjust node positions, modify connections, or remove unnecessary nodes to ensure ToA accurately reflected their analysis. P10 suggested, *"It would be great if I could drag nodes around or change the connections when the tree doesn't match how I'm thinking about my analysis. I wish I could manually reorganize nodes or remove dead-end explorations that didn't lead anywhere useful."* This highlights a design trade off: while automatic generation minimizes user effort, it increases the risk of misalignment between ToA and users' actual logic.

9 Expert Interview

To further evaluate ToA's potential for democratizing visual analytics, we conducted semi-structured interviews with three experts from visualization and HCI.[4] **E7** is a senior visualization researcher who has published extensively in information visualization and visual analytics for over 8 years. He evaluated its suitability for representing CVA processes. **E8** is a UX researcher at a university with 4 years of experience designing intuitive and user-friendly interfaces. He assessed the readability of the visual encodings and the learnability of the interactions within the analysis tree. **E9** is a visualization educator who teaches introductory visualization courses at a university. He evaluated the pedagogical value and practical applicability of the analysis tree for learning data visualization skills. Each interview included a 10-min system demonstration followed by a 20-min focused discussion. Each expert signed an informed consent form before participation and received $25 compensation.

E7 primarily focused on ToA's expressiveness in capturing CVA processes. He appreciated that *"the tree metaphor naturally captures the branching nature of data exploration"* and found the representation of AI outputs as nodes and user queries as edges to be *"intuitive for understanding analysis flow."* However, he raised concerns about visual scalability, noting that *"complex analysis sessions might result in overly dense trees that could overwhelm users rather than provide clarity."* He suggested implementing interactive mechanisms such as filtering and pruning to manage tree complexity. In our user study, the largest tree contained 22 nodes. During semi-structured interviews, no participants spontaneously mentioned difficulties in reading or navigating the tree structure, suggesting that complexity remained manageable within our 25-minute sessions. However, we acknowledge E7's concerns that overlong trees may introduce readability challenges in extended analyses. We discuss potential solutions in Sec. 10.3.

E8 focused on the user experience for novice analysts. He commented that *"the tree structure is intuitive for novices to understand, and the interactions are easy to learn. Users can quickly grasp how to navigate their analysis history."* He suggested enhancing the tree nodes with visual previews of the charts to provide richer contextual information. We believe that word scale visualizations may be a potential approach that can be integrated to enhance the intuitiveness of the tree [12, 31, 99]. Additionally, he noted that visual encodings should remain consistent across thumbnail previews. For example, if sales data is represented by blue bars in one node, it should also be represented by blue bars in other nodes, ensuring users can quickly understand what each visual element means without confusion. We believe this is indeed a promising direction for future exploration.

E9 evaluated the educational value of ToA. He was particularly enthusiastic about the pedagogical potential, stating that *"the tree makes the analysis process visible and teachable in a way that traditional CVA systems don't."* He noted that the categorization of queries could help novices *"learn different ways to approach data exploration."* However, he pointed out that *"for users who simply want to use the system without investing time in learning, there are still some comprehension barriers."* He also identified a promising educational application: highlighting what changes between each step in the analysis tree. This would show learners how to adjust chart specifications to deepen their data exploration.

10 Discussion

The following sections discuss the design implications of ToA for VA systems, acknowledge limitations of our approach, and outline promising future research directions.

10.1 Design Implications

Democratizing Visual Analytics with Mental Maps. While CVA promises accessibility for novices, our findings reveal that users frequently become disoriented and experience task failure during exploration. ToA addresses this by providing a visual navigation map that organizes conversational history into a clear hierarchical structure, effectively externalizing the cognitive map that experts implicitly form. This approach draws inspiration from the psychological concept of mental maps [10, 62]. By explicitly displaying analytical triggers and exploration branches, the tree enables novices to establish global awareness: understanding where they are, where they have been, and where they can go next. This promotes active analytical decision-making, empowering users to move from reactive responses to proactive exploration planning. Future VA systems could integrate similar navigation mechanisms to better support novices.

Nonlinear Expressiveness versus Linear Intuitiveness. Conversational interfaces are inherently linear, conflicting with the nonlinear nature of analytical thinking [61]. However, fully representing nonlinear analytical processes would create overwhelming complexity [40, 73, 91]. This creates a fundamental design trade-off. ToA resolves this design tension by providing a structured overview of the analytical journey. Users can leverage the intuitive benefits of natural language conversation while performing nonlinear navigation and exploration. Future CVA systems should not make binary choices between linear and nonlinear, but rather support both interaction modalities simultaneously. The key lies in finding an appropriate balance between intuitiveness and expressiveness.

[4]This study was approved by the Ethics Committee of the College of Biomedical Engineering & Instrument Science, Zhejiang University (Approval No. 2025-26).

10.2 Limitation

Complexity of User Analysis Logic. ToA simplifies the inherent complexity of human reasoning. Human analysis often involves complex, network-like associations between multiple interconnected hypotheses, rather than a strictly hierarchical structure [18]. However, attempting to explicitly visualize every cross-branch dependency or non-linear association risks creating visual clutter that could overwhelm novices, who already struggle with disorientation. Therefore, we explicitly prioritized structural clarity over expressive completeness. By simplifying the non-linear thought process into a navigable tree, we aim to provide novices with a clear mental map, even though this implies losing some nuanced relationships between distant analysis threads.

Boundaries of Rule-Based Heuristics. Our rule-based tree construction method comprises six connection rules (see Table 4). The first four rules (*elaborate*, *dice*, *reshape*, *slice*) deterministically connect chart nodes by comparing data attributes and filters. The last two rules (*clarify*, *contextualize*) require LLMs to judge semantic relationships between text descriptions and charts. While this approach performs well in most cases, each rule type has its limitations. For the first four rules, our algorithm assumes users typically continue from their most recent analysis step. the algorithm works well when this assumption holds. However, when users continue from earlier analysis steps, the system may incorrectly connect new nodes to the most recent matching node rather than the intended parent node. For the last two rules, the core challenge is context similarity interference. In CVA, users frequently generate highly similar charts, such as bar charts differing only in filters. When judging which chart a text description corresponds to, LLMs struggle to distinguish between these similar candidates using textual cues alone. This was confirmed in our technical evaluation: context similarity interference accounted for 42.34% of all errors. Following participants' suggestions, future work could introduce human-AI collaboration that enables users to manually correct connections.

Generalizability of Evaluation. Our evaluation involved 15 participants (12 novices and 3 experts). While this sample size is relatively small, it is sufficient for several reasons. First, this scale reaches saturation for qualitative data [13, 33]. Guest et al. found that new themes emerge infrequently after analyzing 12 interviews [33]. Second, our sample size aligns with CHI standards [15] and established CVA research. For example, Xie et al. used 12 participants to evaluate code visualization in CVA [89], while Chen et al. employed 10 participants to study multimodal interaction [17]. Third, 12 participants provide appropriate statistical precision, as gains diminish beyond this threshold [41]. Crucially, the large effect size in task performance (per-turn insights, $d > 1.0$) confirms the statistical robustness of our results. However, our evaluation still has limitations. Participants may not represent all novice demographics across cultures, ages, or technical backgrounds. We evaluated only two datasets over short-term sessions, leaving questions about long-term behavior patterns. Future work should include larger, more diverse samples, varied datasets, and extended usage studies.

10.3 Future Works

Managing Complexity of ToA. While no participants spontaneously mentioned difficulties in reading the analysis trees in our study, **E7** raised concerns about potential readability challenges as trees grow larger in extended analyses. As the number of nodes increases, deeply nested branches and numerous nodes may reduce visual clarity and increase navigation difficulty. Following E7's advice, future work could introduce interactive pruning into ToA. Additionally, layout optimizations such as integrating fisheye and focus+context visualization techniques [19, 27] represent another promising direction, as they optimize display density without altering the underlying tree structure.

ToA for Storytelling and Presentation. Data storytelling represents a promising application area for ToA. The tree structure naturally documents the analytical journey, preserving every critical step from initial exploration to final insights. Unlike existing tools that can only generate stories from users' linear exploration sequences [50, 97], ToA captures the branching and iterative nature of analytical reasoning. This provides rich material for automated story generation. Future work could enhance ToA by developing algorithms that identify valuable exploration paths and extract key narrative nodes and logical flows. This would help analysts transform complex exploration histories into coherent storylines highlighting critical discoveries.

11 Conclusion

This study proposes a novel analysis approach, namely ToA, to democratic conversational visual analytic for novices. We first conducted a formative study to compare experts' and novices' behaviors, revealing that experts relied on a general CVA workflow while novices suffered from cue insensitivity and relied on vague questions. To ease difficulty, we propose ToA that organizes CVA conversations into an analysis tree where AI outputs serve as nodes containing analytical cues and categorized queries function as edges. Moreover, we develop an LLM-based tree construction algorithm that processes multi-modal content to recommend next-step action for novice users. Finally, we validate the effectiveness of ToA through an in-lab user study ($N = 12$) and an expert interview ($N = 3$). The results show that ToA completely eliminated task failure and increased per-turn insights by 58. 3%, despite longer per-turn thinking time. In the future, we plan to expand ToA to support more analytical tasks.

Acknowledgments

This work was supported by NSFC (U22A2032, 62421003, 62302440). The work was also supported by the Fundamental Research Funds for the Central Universities. We thank Yanhong Wu, Junxiu Tang, Chengye Huang, Hanrui Xu, and Ke Ren for their constructive suggestions and assistance with the revision. We also thank the anonymous reviewers for their valuable comments.

References

[1] 2025. Django - The web framework for perfectionists with deadlines. Website. https://www.djangoproject.com/.
[2] 2025. Flight Price Prediction. Kaggle Dataset. https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction.

[3] 2025. Flipkart Mobile Dataset. Kaggle Dataset. https://www.kaggle.com/datasets/shubhambathwal/flipkart-mobile-dataset.

[4] 2025. G6 Graph Visualization Engine | AntV. Website. https://g6.antv.antgroup.com/.

[5] 2025. LangChain. Website. https://www.langchain.com/.

[6] 2025. pandas - Python Data Analysis Library. Website. https://pandas.pydata.org/.

[7] 2025. Vega: A Visualization Grammar. Website. https://vega.github.io/.

[8] 2025. Vue.js - The Progressive JavaScript Framework. Website. https://vuejs.org/.

[9] Mubashara Akhtar, Oana Cocarascu, and Elena Simperl. 2023. Reading and Reasoning over Chart Images for Evidence-Based Automated Fact-Checking. In *Proceedings of the Association for Computational Linguistics: EACL 2023*. 399–414. doi:10.18653/v1/2023.findings-eacl.30

[10] Daniel Archambault and Helen C. Purchase. 2013. Mental Map Preservation Helps User Orientation in Dynamic Graphs. In *Proceedings of International Symposium on Graph Drawing*. 475–486. doi:10.1007/978-3-642-36763-2_42

[11] Leilani Battle and Jeffrey Heer. 2019. Characterizing Exploratory Visual Analysis: A Literature Review and Evaluation of Analytic Provenance in Tableau. *Computer Graphics Forum* 38, 3 (2019), 145–159. doi:10.1111/cgf.13678

[12] Fabian Beck and Daniel Weiskopf. 2017. Word-Sized Graphics for Scientific Texts. *IEEE Transactions on Visualization and Computer Graphics* 23, 6 (2017), 1576–1587. doi:10.1109/TVCG.2017.2674958

[13] Clive Roland Boddy. 2016. Sample size for qualitative research. *Qualitative Market Research: An International Journal* 19, 4 (2016), 426–432. doi:10.1108/QMR-06-2016-0053

[14] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. doi:10.1191/1478088706qp063oa

[15] Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *Proceedings of ACM Conference on Human Factors in Computing Systems*. 981–992. doi:10.1145/2858036.2858498

[16] Steven P. Callahan, Juliana Freire, Emanuele Santos, Carlos E. Scheidegger, Cláudio T. Silva, and Huy T. Vo. 2006. VisTrails: Visualization Meets Data Management. In *Proceedings of ACM International Conference on Management of Data*. 745–747. doi:10.1145/1142473.1142574

[17] Juntong Chen, Jiang Wu, Jiajing Guo, Vikram Mohanty, Xueming Li, Jorge Piazentin Ono, Wenbin He, Liu Ren, and Dongyu Liu. 2025. InterChat: Enhancing Generative Visual Analytics using Multimodal Interactions. *Computer Graphics Forum* 44, 3 (2025), e70112. doi:10.1111/cgf.70112

[18] Pei Chen, Jiayi Yao, Zhuoyi Cheng, Yichen Cai, Jiayang Li, Weitao You, and Lingyun Sun. 2025. CoExploreDS: Framing and Advancing Collaborative Design Space Exploration Between Human and AI. In *Proceedings of ACM Conference on Human Factors in Computing Systems*. 1–20. doi:10.1145/3706598.3713869

[19] Andy Cockburn, Amy Karlson, and Benjamin B. Bederson. 2009. A Review of Overview+Detail, Zooming, and Focus+Context Interfaces. *Comput. Surveys* 41, 1 (2009), 2:1–2:31. doi:10.1145/1456650.1456652

[20] Amit Kumar Das, Mohammad Tarun, and Klaus Mueller. 2025. Charts-of-Thought: Enhancing LLM Visualization Literacy through Structured Data Extraction. *IEEE Transactions on Visualization and Computer Graphics* (2025), 1–11. doi:10.1109/TVCG.2025.3634813

[21] Çağatay Demiralp, Peter J. Haas, Srinivasan Parthasarathy, and Tejaswini Pedapati. 2017. Foresight: Recommending Visual Insights. *Proceedings of the VLDB Endowment* 10, 12 (2017), 1614–1617. doi:10.14778/3137765.3137813

[22] Kedar Dhamdhere, Kevin S. McCurley, Ralfi Nahmias, Mukund Sundararajan, and Qiqi Yan. 2017. Analyza: Exploring Data with Conversation. In *Proceedings of ACM International Conference on Intelligent User Interfaces*. 493–504. doi:10.1145/3025171.3025227

[23] Zijian Ding, Michelle Brachman, Joel Chan, and Werner Geyer. 2025. "The Diagram is like Guardrails": Structuring GenAI-Assisted Hypotheses Exploration with an Interactive Shared Representation. In *Proceedings of ACM Conference on Creativity and Cognition*. 606–625. doi:10.1145/3698061.3726935

[24] Shujing Dong, Yuan Ling, Shunyan Luo, Shuyi Wang, Yarong Feng, Zongyi (Joe) Liu, Hongfei Li, Ayush Goyal, and Bruce Ferry. 2024. Context-Aware and User Intent-Aware Follow-Up Question Generation (CA-UIA-QG): Mimicking User Behavior in Multi-Turn Setting. In *Proceedings of IEEE International Conference on Big Data (BigData '24)*. 4086–4094. doi:10.1109/BigData62323.2024.10825543

[25] Ethan Fast, Binbin Chen, Julia Mendelsohn, Jonathan Bassen, and Michael S. Bernstein. 2018. Iris: A Conversational Agent for Complex Tasks. In *Proceedings of ACM Conference on Human Factors in Computing Systems*. 1–12. doi:10.1145/3173574.3174047

[26] Stephen N. Freund, Brooke Simon, Emery D. Berger, and Eunice Jun. 2025. Flowco: Mixed-Initiative Authoring of Reliable End-to-End Data Analyses via Dataflow Graphs and LLMs. In *Proceedings of Annual ACM Symposium on User Interface Software and Technology*. 1–20. doi:10.1145/3746059.3747636

[27] George W. Furnas. 1986. Generalized Fisheye Views. In *Proceedings of ACM Conference on Human Factors in Computing Systems*. 16–23. doi:10.1145/22627.22342

[28] Tong Gao, Mira Dontcheva, Eytan Adar, Zhicheng Liu, and Karrie G. Karahalios. 2015. DataTone: Managing Ambiguity in Natural Language Interfaces for Data Visualization. In *Proceedings of Annual ACM Symposium on User Interface Software and Technology*. 489–500. doi:10.1145/2807442.2807478

[29] Joseph Gatto, Parker Seegmiller, Timothy Burdick, Inas S. Khayal, Sarah DeLozier, and Sarah M. Preum. 2025. Follow-up Question Generation For Enhanced Patient-Provider Conversations. In *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL '25)*. Bangkok, Thailand, 22693–22714. doi:10.18653/v1/2025.acl-long.1226

[30] Yubin Ge, Ziang Xiao, Jana Diesner, Heng Ji, Karrie Karahalios, and Hari Sundaram. 2023. What should I Ask: A Knowledge-driven Approach for Follow-up Questions Generation in Conversational Surveys. In *Proceedings of Pacific Asia Conference on Language, Information and Computation*. 113–124. https://aclanthology.org/2023.paclic-1.12/

[31] Pascal Goffin, Wesley Willett, Jean-Daniel Fekete, and Petra Isenberg. 2014. Exploring the Placement and Design of Word-Scale Visualizations. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 2291–2300. doi:10.1109/TVCG.2014.2346435

[32] Google. 2025. Gemini Developer API | Gemma open models. https://ai.google.dev/

[33] Greg Guest, Arwen Bunce, and Laura Johnson. 2006. How Many Interviews Are Enough? An Experiment with Data Saturation and Variability. *Field Methods* 18, 1 (2006), 59–82. doi:10.1177/1525822X05279903

[34] Hua Guo, Steven R. Gomez, Caroline Ziemkiewicz, and David H. Laidlaw. 2016. A Case Study Using Visualization Interaction Logs and Insight Metrics to Understand How Analysts Arrive at Insights. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016), 51–60. doi:10.1109/TVCG.2015.2467613

[35] Matt-Heun Hong and Anamaria Crisan. 2025. Data has Entered the Chat: How Data Workers Conduct Exploratory Visual Analytic Conversations with GenAI Agents. *ACM Transactions on Interactive Intelligent Systems* 16, 2 (2025), 1–16. doi:10.1145/3744750

[36] Enamul Hoque, Vidya Setlur, Melanie Tory, and Isaac Dykeman. 2018. Applying Pragmatics Principles for Interaction with Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2018), 309–318. doi:10.1109/TVCG.2017.2744684

[37] Kevin Hu, Michiel A. Bakker, Stephen Li, Tim Kraska, and César Hidalgo. 2019. VizML: A Machine Learning Approach to Visualization Recommendation. In *Proceedings of ACM Conference on Human Factors in Computing Systems*. 1–12. doi:10.1145/3290605.3300358

[38] Kung-Hsiang Huang, Hou Pong Chan, Yi R. Fung, Haoyi Qiu, Mingyang Zhou, Shafiq Joty, Shih-Fu Chang, and Heng Ji. 2025. From Pixels to Insights: A Survey on Automatic Chart Understanding in the Era of Large Foundation Models. *IEEE Transactions on Knowledge and Data Engineering* 37, 5 (2025), 2550–2568. doi:10.1109/TKDE.2024.3513320

[39] Mohammed Saidul Islam, Raian Rahman, Ahmed Masry, Md Tahmid Rahman Laskar, Mir Tafseer Nayeem, and Enamul Hoque. 2024. Are Large Vision Language Models up to the Challenge of Chart Comprehension and Reasoning. In *Proceedings of the Association for Computational Linguistics: EMNLP 2024*. 3334–3368. doi:10.18653/v1/2024.findings-emnlp.191

[40] Peiling Jiang, Jude Rayan, Steven P. Dow, and Haijun Xia. 2023. Graphologue: Exploring Large Language Model Responses with Interactive Diagrams. In *Proceedings of Annual ACM Symposium on User Interface Software and Technology*. 1–20. doi:10.1145/3586183.3606737

[41] Steven A. Julious. 2005. Sample Size of 12 per Group Rule of Thumb for a Pilot Study. *Pharmaceutical Statistics* 4, 4 (2005), 287–291. doi:10.1002/pst.185

[42] Majeed Kazemitabaar, Jack Williams, Ian Drosos, Tovi Grossman, Austin Zachary Henley, Carina Negreanu, and Advait Sarkar. 2024. Improving Steering and Verification in AI-Assisted Data Analysis with Interactive Task Decomposition. In *Proceedings of Annual ACM Symposium on User Interface Software and Technology*. 1–19. doi:10.1145/3654777.3676345

[43] Daniel A. Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. 2008. Visual Analytics: Definition, Process, and Challenges. In *Information Visualization*, Andreas Kerren, John T. Stasko, Jean-Daniel Fekete, and Chris North (Eds.). Lecture Notes in Computer Science, Vol. 4950. 154–175. doi:10.1007/978-3-540-70956-5_7

[44] Alicia Key, Bill Howe, Daniel Perry, and Cecilia Aragon. 2012. VizDeck: self-organizing dashboards for visual analytics. In *Proceedings of ACM International Conference on Management of Data*. 681–684. doi:10.1145/2213836.2213931

[45] Jeonghyun Kim, Lingzi Hong, and Sarah Evans. 2024. Toward Measuring Data Literacy for Higher Education: Developing and Validating a Data Literacy Self-Efficacy Scale. *Journal of the Association for Information Science and Technology* 75, 8 (2024), 916–931. doi:10.1002/asi.24934

[46] Jeongyeon Kim, Sangho Suh, Lydia B. Chilton, and Haijun Xia. 2023. Metaphorian: Leveraging Large Language Models to Support Extended Metaphor Creation for Science Writing. In *Proceedings of ACM Designing Interactive Systems Conference*. 115–135. doi:10.1145/3563657.3595996

[47] Tae Soo Kim, Yoonjoo Lee, Minsuk Chang, and Juho Kim. 2023. Cells, Generators, and Lenses: Design Framework for Object-Oriented Interaction with Large Language Models. In *Proceedings of Annual ACM Symposium on User Interface Software and Technology*. 1–18. doi:10.1145/3586183.3606833

[48] Bongshin Lee, Petra Isenberg, Nathalie Henry Riche, and Sheelagh Carpendale. 2012. Beyond Mouse and Keyboard: Expanding Design Considerations for Information Visualization Interactions. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2689–2698. doi:10.1109/TVCG.2012.204

[49] Doris Jung-Lin Lee, Vidya Setlur, Melanie Tory, Karrie Karahalios, and Aditya Parameswaran. 2022. Deconstructing Categorization in Visualization Recommendation: A Taxonomy and Comparative Study. *IEEE Transactions on Visualization and Computer Graphics* 28, 12 (2022), 4225–4239. doi:10.1109/TVCG.2021.3085751

[50] Haotian Li, Lu Ying, Haidong Zhang, Yingcai Wu, Huamin Qu, and Yun Wang. 2023. Notable: On-the-fly Assistant for Data Storytelling in Computational Notebooks. In *Proceedings of ACM Conference on Human Factors in Computing Systems*. 1–16. doi:10.1145/3544548.3580965

[51] Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. 2024. MEDIQ: Question-Asking LLMs and a Benchmark for Reliable Interactive Clinical Reasoning. In *Proceedings of International Conference on Neural Information Processing Systems (NeurIPS '24)*. 31. https://openreview.net/pdf?id=W4pIBQ7bAI

[52] Jianyu Liu, Yi Huang, Sheng Bi, Junlan Feng, and Guilin Qi. 2025. From Superficial to Deep: Integrating External Knowledge for Follow-up Question Generation Using Knowledge Graph and LLM. In *Proceedings of International Conference on Computational Linguistics*. 828–840. https://aclanthology.org/2025.coling-main.55/

[53] Zhicheng Liu and Jeffrey Heer. 2014. The Effects of Interactive Latency on Exploratory Visual Analysis. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 2122–2131. doi:10.1109/TVCG.2014.2346452

[54] Alan Lundgard and Arvind Satyanarayan. 2022. Accessible Visualization via Natural Language Descriptions: A Four-Level Model of Semantic Content. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2022), 1073–1083. doi:10.1109/TVCG.2021.3114770

[55] Damien Masson, Sylvain Malacria, Géry Casiez, and Daniel Vogel. 2024. DirectGPT: A Direct Manipulation Interface to Interact with Large Language Models. In *Proceedings of ACM Conference on Human Factors in Computing Systems*. 1–16. doi:10.1145/3613904.3642462

[56] Microsoft. 2024. Power BI AI Features for All Data Analysts. Website. https://techcommunity.microsoft.com/t5/educator-developer-blog/power-bi-ai-features-for-all-data-analysts/ba-p/3835447

[57] Kyo-Joong Oh, Ho-Jin Choi, Gahgene Gweon, Jeong Heo, and Pum-Mo Ryu. 2015. Paraphrase Generation Based on Lexical Knowledge and Features for a Natural Language Question Answering System. In *Proceedings of the International Conference on Big Data and Smart Computing (BIGCOMP '15)*. Jeju, Korea, 35–38. doi:10.1109/35021BIGCOMP.2015.7072846

[58] OpenAI. 2024. ChatGPT. Website. https://openai.com/chatgpt/overview/.

[59] OpenAI. 2024. Improvements to data analysis in ChatGPT. Website. https://openai.com/index/improvements-to-data-analysis-in-chatgpt/.

[60] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training Language Models to Follow Instructions with Human Feedback. In *Proceedings of International Conference on Neural Information Processing Systems*. 27730–27744. https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html

[61] Peter Pirolli and Stuart Card. 2005. The Sensemaking Process and Leverage Points for Analyst Technology as Identified Through Cognitive Task Analysis. In *Proceedings of International Conference on Intelligence Analysis*, Vol. 5. 2–4.

[62] Helen C. Purchase, Eve Hoggan, and Carsten Görg. 2007. How Important Is the "Mental Map"? – An Empirical Investigation of a Dynamic Graph Layout Algorithm. In *Proceedings of International Symposium on Graph Drawing*. 184–195. doi:10.1007/978-3-540-70904-6_19

[63] Purvi Saraiya, Chris North, and Karen Duca. 2005. An Insight-Based Methodology for Evaluating Bioinformatics Visualizations. *IEEE Transactions on Visualization and Computer Graphics* 11, 4 (2005), 443–456. doi:10.1109/TVCG.2005.53

[64] Purvi Saraiya, Chris North, Vy Lam, and Karen A. Duca. 2006. An Insight-Based Longitudinal Study of Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics* 12, 6 (2006), 1511–1522. doi:10.1109/TVCG.2006.85

[65] Thomas Scialom and Jacopo Staiano. 2020. Ask to Learn: A Study on Curiosity-driven Question Generation. In *Proceedings of International Conference on Computational Linguistics*. 2224–2235. https://aclanthology.org/2020.coling-main.202/

[66] Vidya Setlur, Sarah E. Battersby, Melanie Tory, Rich Gossweiler, and Angel X. Chang. 2016. Eviza: A Natural Language Interface for Visual Analysis. In *Proceedings of Annual ACM Symposium on User Interface Software and Technology*. 365–377. doi:10.1145/2984511.2984588

[67] Vidya Setlur, Enamul Hoque, Dae Hyun Kim, and Angel X. Chang. 2020. Sneak Pique: Exploring Autocompletion as a Data Discovery Scaffold for Supporting Visual Analysis. In *Proceedings of Annual ACM Symposium on User Interface Software and Technology*. 966–978. doi:10.1145/3379337.3415813

[68] Leixian Shen, Enya Shen, Yuyu Luo, Xiaocong Yang, Xuming Hu, Xiongshuai Zhang, Zhiwei Tai, and Jianmin Wang. 2023. Towards Natural Language Interfaces for Data Visualization: A Survey. *IEEE Transactions on Visualization and Computer Graphics* 29, 6 (2023), 3121–3144. doi:10.1109/TVCG.2022.3148007

[69] Sarvesh Soni and Kirk Roberts. 2019. A Paraphrase Generation System for EHR Question Answering. In *Proceedings of BioNLP Workshop and Shared Task*. 20–29. https://aclanthology.org/W19-5003/

[70] Arjun Srinivasan and Vidya Setlur. 2021. Snowy: Recommending Utterances for Conversational Visual Analysis. In *Proceedings of Annual ACM Symposium on User Interface Software and Technology (UIST '21)*. 864–880. doi:10.1145/3472749.3474792

[71] Arjun Srinivasan and John Stasko. 2018. Orko: Facilitating Multimodal Interaction for Visual Exploration and Analysis of Networks. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2018), 511–521. doi:10.1109/TVCG.2017.2745219

[72] Sangho Suh, Meng Chen, Bryan Min, Toby Jia-Jun Li, and Haijun Xia. 2024. Luminate: Structured Generation and Exploration of Design Space with Large Language Models for Human-AI Co-Creation. In *Proceedings of ACM Conference on Human Factors in Computing Systems*. 1–26. doi:10.1145/3613904.3642400

[73] Sangho Suh, Bryan Min, Srishti Palani, and Haijun Xia. 2023. Sensecape: Enabling Multilevel Exploration and Sensemaking with Large Language Models. In *Proceedings of Annual ACM Symposium on User Interface Software and Technology*. 1–18. doi:10.1145/3586183.3606756

[74] Yiwen Sun, Jason Leigh, Andrew Johnson, and Sangyoon Lee. 2010. Articulate: A Semi-automated Model for Translating Natural Language Queries into Meaningful Visualizations. In *Proceedings of Smart Graphics*. 184–195. doi:10.1007/978-3-642-13544-6_18

[75] Tableau. 2024. Artificial Intelligence | Tableau. Website. https://www.tableau.com/products/artificial-intelligence.

[76] Tan Tang, Yanhong Wu, Junming Gao, Kejia Ruan, Yanjie Zhang, Shuainan Ye, Yingcai Wu, and Xiaojiao Chen. 2024. ArtEyer: Enriching GPT-based agents with contextual data visualizations for fine art authentication. *Visual Informatics* 8, 4 (2024), 48–59. doi:10.1016/j.visinf.2024.11.001

[77] Melanie Tory and Vidya Setlur. 2019. Do what I mean, not what I say! Design considerations for supporting intent and context in analytical conversation. In *Proceedings of IEEE Conference on Visual Analytics Science and Technology*. 93–103. doi:10.1109/VAST47406.2019.8986918

[78] VizGPT. 2024. VizGPT - Turn your data into charts with AI. Website. https://vizgpt.ai/.

[79] Huichen Will Wang, Larry Birnbaum, and Vidya Setlur. 2025. Jupybara: Operationalizing a Design Space for Actionable Data Analysis and Storytelling with LLMs. In *Proceedings of ACM Conference on Human Factors in Computing Systems (CHI '25)*. New York, NY, USA. doi:10.1145/3706598.3713913

[80] Liangwei Wang, Zhan Wang, Shishi Xiao, Le Liu, Fugee Tsung, and Wei Zeng. 2025. VizTA: Enhancing Comprehension of Distributional Visualization with Visual-Lexical Fused Conversational Interface. *Computer Graphics Forum* 44, 3 (2025), e70110. doi:10.1111/cgf.70110

[81] Lei Wang, Songheng Zhang, Yun Wang, Ee-Peng Lim, and Yong Wang. 2023. LLM4Vis: Explainable Visualization Recommendation using ChatGPT. In *Proceedings of Conference on Empirical Methods in Natural Language Processing: Industry Track*. 675–692. doi:10.18653/v1/2023.emnlp-industry.64

[82] Yansen Wang, Chenyi Liu, Minlie Huang, and Liqiang Nie. 2018. Learning to Ask Questions in Open-domain Conversational Systems with Typed Decoders. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*. 2193–2203. doi:10.18653/v1/P18-1204

[83] Yun Wang, Zhida Sun, Haidong Zhang, Weiwei Cui, Ke Xu, Xiaojuan Ma, and Dongmei Zhang. 2020. DataShot: Automatic Generation of Fact Sheets from Tabular Data. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2020), 895–905. doi:10.1109/TVCG.2019.2934398

[84] Ziyu Wang, Hao Li, Di Huang, Hye-Sung Kim, Chae-Won Shin, and Amir M. Rahmani. 2025. HealthQ: Unveiling Questioning Capabilities of LLM Chains in Healthcare Conversations. *Smart Health* 35 (2025), 100570. doi:10.1016/j.smhl.2025.100570

[85] Luoxuan Weng, Xingbo Wang, Junyu Lu, Yingchaojie Feng, Yihan Liu, Haozhe Feng, Danqing Huang, and Wei Chen. 2025. InsightLens: Augmenting LLM-Powered Data Analysis with Interactive Insight Management and Navigation. *IEEE Transactions on Visualization and Computer Graphics* 31, 6 (2025), 3719–3732. doi:10.1109/TVCG.2025.3567131

[86] Caleb Winston, Cleah Winston, Claris Winston, and Chloe Winston. 2024. Medical Question-Generation for Pre-Consultation with LLM In-Context Learning. In *GenAI for Health: Potential, Trust and Policy Compliance, Workshop at NeurIPS 2024*. https://neurips.cc/virtual/2024/106861

[87] Kanit Wongsuphasawat, Dominik Moritz, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer. 2016. Voyager: Exploratory Analysis via Faceted Browsing of Visualization Recommendations. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016), 649–658. doi:10.1109/TVCG.2015.2467191

[88] Kanit Wongsuphasawat, Zening Qu, Dominik Moritz, Riley Chang, Felix Ouk, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer. 2017. Voyager 2: Augmenting Visual Analysis with Partial View Specifications. In *Proceedings*

of ACM Conference on Human Factors in Computing Systems. 2648–2659. doi:10.1145/3025453.3025768

[89] Liwenhan Xie, Chengbo Zheng, Haijun Xia, Huamin Qu, and Chen Zhu-Tian. 2024. WaitGPT: Monitoring and Steering Conversational LLM Agent in Data Analysis with On-the-Fly Code Visualization. In *Proceedings of Annual ACM Symposium on User Interface Software and Technology*. 1–14. doi:10.1145/3654777.3676374

[90] Bowen Yu and Cláudio T. Silva. 2017. VisFlow – Web-based Visualization Framework for Tabular Data with a Subset Flow Model. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (2017), 251–260. doi:10.1109/TVCG.2016.2598497

[91] Bowen Yu and Cláudio T. Silva. 2020. FlowSense: A Natural Language Interface for Visual Data Exploration within a Dataflow System. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2020), 1–11. doi:10.1109/TVCG.2019.2934668

[92] Emanuel Zgraggen, Zheguang Zhao, Robert Zeleznik, and Tim Kraska. 2018. Investigating the Effect of the Multiple Comparisons Problem in Visual Analysis. In *Proceedings of ACM Conference on Human Factors in Computing Systems*. 1–12. doi:10.1145/3173574.3174053

[93] Songheng Zhang, Haotian Li, Huamin Qu, and Yong Wang. 2024. AdaVis: Adaptive and Explainable Visualization Recommendation for Tabular Data. *IEEE Transactions on Visualization and Computer Graphics* 30, 9 (2024), 5923–5938. doi:10.1109/TVCG.2023.3316469

[94] Wenshuo Zhang, Leixian Shen, Shuchang Xu, Jindu Wang, Jian Zhao, Huamin Qu, and Linping Yuan. 2025. NeuroSync: Intent-Aware Code-Based Problem Solving via Direct LLM Understanding Modification. In *Proceedings of Annual ACM Symposium on User Interface Software and Technology*. 1–19. doi:10.1145/3746059.3747668

[95] Zheng Zhang, Jie Gao, Ranjodh Singh Dhaliwal, and Toby Jia-Jun Li. 2023. VISAR: A Human-AI Argumentative Writing Assistant with Visual Programming and Rapid Draft Prototyping. In *Proceedings of Annual ACM Symposium on User Interface Software and Technology*. 1–30. doi:10.1145/3586183.3606800

[96] Yuheng Zhao, Yixing Zhang, Yu Zhang, Xinyi Zhao, Junjie Wang, Zekai Shao, Cagatay Turkay, and Siming Chen. 2025. LEVA: Using Large Language Models to Enhance Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics* 31, 3 (2025), 1830–1847. doi:10.1109/TVCG.2024.3368060

[97] Chengbo Zheng, Dakuo Wang, April Yi Wang, and Xiaojuan Ma. 2022. Telling Stories from Computational Notebooks: AI-Assisted Presentation Slides Creation for Presenting Data Science Work. In *Proceedings of ACM Conference on Human Factors in Computing Systems*. 1–20. doi:10.1145/3491102.3517615

[98] Jiayi Zhou, Renzhong Li, Junxiu Tang, Tan Tang, Haotian Li, Weiwei Cui, and Yingcai Wu. 2024. Understanding Nonlinear Collaboration between Human and AI Agents: A Co-design Framework for Creative Design. In *Proceedings of ACM Conference on Human Factors in Computing Systems*. 1–16. doi:10.1145/3613904.3642812

[99] Ruishi Zou, Yinqi Tang, Jingzhu Chen, Siyu Lu, Yan Lu, Yingfan Yang, and Chen Ye. 2025. GistVis: Automatic Generation of Word-scale Visualizations from Data-rich Documents. In *Proceedings of ACM Conference on Human Factors in Computing Systems*. 1–18. doi:10.1145/3706598.3713881