A Comparative Evaluation of Natural Language and Dashboard Visualization Interfaces for Data Monitoring Tasks

Melika Besharati Amirkandeh*
Dalhousie University
Halifax, Canada
melika.besharati76@dal.ca

Derek Reilly
Dalhousie University
Halifax, Canada
reilly@cs.dal.ca

Abstract

Natural Language Interfaces (NLIs) are emerging as an alternative to or supplement for dashboard visualization interfaces, allowing users to formulate queries using conversational input rather than structured commands, and have responses provided as visualizations tailored to their query. However, despite growing interest in both natural language and dashboard-based visualization systems, there is limited empirical evidence from controlled user studies that directly compare their usability, cognitive workload, and efficiency on realistic analytical tasks in applied, real-world monitoring contexts. In a controlled study (N=24) we compare a chatbot-driven visualization interface and a visualization dashboard for a set of realistic analytics tasks in a specific applied domain (monitoring fish farm data). While we did not find statistically significant differences in System Usability Scale (SUS) scores or task accuracy scores between interface conditions, NASA TLX scores show significant higher temporal demand and higher mental demand when using the chatbot vs. the dashboard, and the chatbot yielded significantly higher task times overall. Participant feedback indicated complementary strengths, however: overall, the chatbot was praised for simplicity and adaptability, and the dashboard for precision and clarity. We consider ways of integrating natural language and traditional interfaces to enhance data exploration and decision-making.

CCS Concepts

• **Human-centered computing → Human computer interaction (HCI)**; *Interactive systems and tools*; *Visualization.*

Keywords

Data Visualization, Chatbots, Natural Language Interface, Dashboard, Large Language Models, Human-Computer Interaction

ACM Reference Format:
Melika Besharati Amirkandeh and Derek Reilly. 2025. A Comparative Evaluation of Natural Language and Dashboard Visualization Interfaces for Data Monitoring Tasks. In *Proceedings of Graphics Interface 2026 (GI 2026).* ACM, New York, NY, USA, 12 pages. https://doi.org/XXXXXXX.XXXXXXX

GI 2026, Kitchener-Waterloo, Canada

ACM ISBN 978-1-4503-XXXX-X/25/06
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Data visualization plays a critical role in fields that require real-time monitoring of data, including finance, environmental monitoring, and healthcare. Professionals rely on visualization tools to detect anomalies, track trends, and make timely, data-driven decisions from large-scale temporal datasets. Graphical user interface (GUI) dashboards have been the standard, offering structured workflows, predefined widgets, and precise control over visualization parameters, allowing users to filter, customize, and refine data representations. GUI dashboards support visualization techniques that are well-suited for trend detection and comparative analysis, such as line charts, heat maps, and moving averages. Domain-specific dashboards combine multiple coordinated views tailored to the workflows, analytical goals, and expertise of domain specialists, making them effective tools for in-depth analysis of complex temporal datasets. However, while dashboards are effective for targeted analysis, they can be complex to navigate, often require significant domain knowledge to fully comprehend, and may increase cognitive load for users without dashboard interface expertise [8, 15].

With advances in artificial intelligence, natural language interfaces (NLIs) have emerged as an alternative, allowing users to query data conversationally. Supported by large language models (LLMs), these systems can interpret everyday language and generate relevant visualizations [38, 45]. NLIs lower the technical barrier to data exploration but also face challenges such as query ambiguity, reduced parameter-level control, and limited transparency in AI-generated outputs [28, 29, 37].

Although many studies evaluate NLIs or dashboards separately, few compare them directly. Prior systems such as Eviza [37] and Chat2VIS [29] demonstrate that natural language interfaces can effectively translate user queries into visualizations, but they do not provide direct usability comparisons with dashboard-based interfaces. Empirical studies by Hearst et al. [15] and Setlur and Tory [38] offer early evidence that conversational interfaces support flexible and expressive data exploration, while also introducing higher interaction effort compared to structured visual controls. However, few studies have directly compared the effort required to iteratively refine natural language queries versus adjusting visual parameters to tune visualizations, particularly in real-time monitoring contexts. In such settings, where users must rapidly adjust views under time pressure, differences in interaction cost can meaningfully affect efficiency, cognitive load, and decision-making. Addressing this gap therefore requires evaluating performance, cognitive workload, and user feedback across both modalities.

Prior work suggests that natural language interfaces can lower barriers to data access by enabling flexible, conversational querying, while dashboards provide precise control and efficient parameter

adjustment through structured visual interactions. However, comparative studies have focused on exploratory or static analysis tasks rather than continuous, time-sensitive monitoring [37, 39, 44]. Real-time monitoring is distinct in that users must rapidly interpret evolving data streams and iteratively refine visualizations under time pressure, making interaction cost and cognitive load critical factors. It is important to compare NLIs and dashboards in this context to understand how different interaction mechanisms support timely and accurate decision-making. In fish farming and other domains, users involved in day-to-day monitoring activities typically lack extensive training with analytical tools, making differences in usability and cognitive effort between interaction modalities more pronounced.

To conduct a naturalistic comparative evaluation involving real-time monitoring, we focused on a single applied domain and validated study tasks with domain experts to reflect real-world scenarios. This research uses live fish farming data as a representative real-world use case. Conducted in collaboration with an interdisciplinary team and industry partners, the work addresses challenges in managing complex aquaculture environments where parameters such as dissolved oxygen, temperature, and salinity must be tracked in real-time to ensure stock health and operational efficiency. For this case study, we focus on users who are not data visualization experts or data scientists, because in many applied monitoring settings, the primary users of these tools are practitioners with limited or no formal training in data analysis.

The aquaculture domain was selected because it offers real-world, real-time monitoring data that remains accessible enough for our study. Our tasks focus on understandable, one-dimensional environmental variables such as temperature, salinity, and dissolved oxygen—variables that can be readily modelled as time series data. This means that specialized domain knowledge is not required to understand the data at the level required to effectively perform the tasks. This permits us to recruit non-experts for evaluation and facilitates comparisons between interfaces. Recruiting domain novices reduces the influence of domain knowledge on how the tools are used and how visualizations are interpreted. At the same time, unlike artificial or toy datasets, the fish-farm context provides a realistic and grounded setting. This contextual richness makes the tasks more relatable for participants and allows them to offer more thoughtful and experience-driven feedback. A limitation of this choice is that we do not capture how domain experts can leverage their knowledge to detect patterns or interpret ambiguous signals in data. We therefore complement our findings with reflections from domain experts on the perceived benefits and drawbacks of each tool.

This study contributes by directly comparing an NLI-based chatbot and a GUI dashboard for a realistic task set involving the monitoring and analysis of fish farm data. We conduct a controlled comparative evaluation with 24 domain novices, all with limited to no experience using data visualization dashboards or natural language visualization tools.

We focus on usability, efficiency, and cognitive load, asking three research questions: 1. How do natural-language and dashboard interfaces compare in supporting the task performance of domain novices? 2. How do domain novices perceive the usability and cognitive effort associated with each interface? 3. What practical strengths and weaknesses do domain novices report when using these tools?

2 Related Work

Natural Language Interfaces (NLIs) for data visualization have evolved rapidly, moving from early rule-based systems to neural network approaches and, more recently, large language model (LLM)-powered tools. This section reviews the main methodological approaches, evaluation strategies, usability considerations, hybrid interaction designs, and relevant work on dashboards and monitoring analytics.

2.1 Natural Language to Visualization Methods

Early systems such as DataTone and Eviza used manually defined rules and grammar-based parsing to translate user queries into visualizations [11, 37]. These systems handled ambiguity through interactive disambiguation but required structured phrasing and manual rule updates. FlowSense extended this approach to dataflow visualizations using semantic parsing [54]. Neural network-based systems like NL4DV and LIDA introduced more flexibility by using learned models for query interpretation [10, 34], while Data2Vis used sequence-to-sequence learning for chart generation [9]. LLM-based systems such as Chat2VIS, Visistant, and EvaLLM expanded expressiveness and multilingual capabilities [22, 29, 33], while recent extensions like NL4DV-LLM integrated analytic specification prompts for better explainability [36]. Surveys highlight persistent challenges, including ambiguity resolution, multi-turn interactions, and balancing automation with user control [19, 39, 55].

2.2 System Evaluation Methods

Evaluation typically focuses on two aspects: model performance and usability. Model performance is measured through accuracy, error rates, and semantic correctness [22, 34], sometimes enhanced with prompt engineering [7] and automated feedback loops [23]. Comparative studies of LLMs for visualization show GPT-4 outperforming others in code correctness but still struggling with complex analytical queries [20]. Usability evaluations employ metrics such as completion time, error count, and subjective ratings (SUS, NASA-TLX) [11, 29], along with open-ended tasks to reduce phrasing bias [43].

2.3 Usability Considerations of Chat Interfaces

NLI usability depends on user expertise, interaction style, and transparency. Beginners often benefit more from guided interactions and conversational refinement [3, 4, 6], while experts use NLIs for rapid exploration. Explainability features, such as rule-based explanations or natural language justifications, improve trust and understanding [21, 24, 46]. Iterative and dialog-based designs [13, 16, 32, 38] support multi-turn refinement and help manage ambiguity. These findings suggest tailoring NLIs to diverse user needs and balancing automation with user control, especially in real-time monitoring contexts.

2.4 Hybrid Interaction Models

The classic HCI debate between direct manipulation and interface agents frames an important trade-off between user control and automated assistance, which is central to hybrid visualization systems [40]. Hybrid systems combine NLIs with GUI controls to leverage the flexibility of language and the precision of direct manipulation. In such systems, natural language can also operate over the current dashboard context, allowing users to issue situated commands such as modifying an existing chart's encoding, colours, or filters rather than generating a new visualization from scratch. DynaVis generates UI widgets from language input for iterative refinement [49], while multimodal dashboards integrate voice and touch input for improved monitoring efficiency [8]. Snowy and Voyager guide exploration with proactive query and chart recommendations [42, 51]. Articulate+ interprets background conversation to create visualizations [47], and Stigall et al. advocate for chatbots as analytical partners that retain context [44]. Voice-based systems like VOICE and ChartMimic further extend conversational modalities into hands-free and multimodal scenarios [18, 53].

2.5 Dashboards for Real-Time Monitoring

Dashboards remain the standard tool for real-time monitoring in domains such as environmental sensing, finance, and industrial operations. Tools like Tableau [48], Power BI [31], and Grafana [12] offer structured layouts, interactive controls, and persistent feedback, supporting quick anomaly detection and targeted analysis. Research on dashboard design patterns [1] and domain-specific decision support systems [14] highlights their adaptability and effectiveness. Studies such as Plume [26] and MEDLEY [35] integrate storytelling and recommendation into dashboards, while MultiVision automates component generation from user data [52]. In this study, dashboards serve as the baseline for evaluating NLI-based systems.

2.6 Analytics for Monitoring and Time-Series Visualizations

Monitoring often involves time-series data, requiring quick detection of trends, anomalies, and seasonal patterns. Visualization techniques such as line charts, heatmaps, and moving averages are common in dashboards and increasingly supported in NLIs. Prior work includes anomaly detection in streaming data [5], neural forecasting for real-time decision support [27], and interactive filtering for multi-attribute streams [50]. In environmental and agricultural domains, custom monitoring dashboards have been developed to track water quality, crop growth, and aquaculture conditions [8, 14]. These studies provide methodological foundations for our evaluation of dashboard and NLI-based monitoring tools.

3 Methodology

3.1 Interfaces

The experiment involved two interactive systems designed for real-time monitoring data visualization: a GUI-based dashboard and a chatbot-based Natural Language Interface (NLI). Both interfaces used the same underlying dataset to ensure consistency in task comparison.

The GUI dashboard (RealFish Pro) is a web-based application with dropdown menus, checkboxes, and chart configuration panels, allowing users to interact with visual elements directly [17]. Figure 1 shows the dashboard interface that was used in this study. Users can filter data, select parameters (e.g., time range, variable), and apply visualization options such as line charts or moving averages.

The RealFishPro dashboard includes many industry-standard features, including line-chart panels, date-range selectors, and tabular summaries—components common across platforms like Tableau and Power BI. RealFishPro is a fully featured tool used today in the aquaculture industry, providing a valuable opportunity to evaluate interaction with a real-world system, using real data. Core UI elements found in other systems were relied on for our tasks rather than domain-specific, proprietary, or advanced features.

The NLI chatbot is a conversational interface, as shown in Figure 2, where users enter natural language queries (e.g., "Show me the salinity trend over the past month") [34]. The system interprets the queries and converts them into Vega-Lite commands, generating relevant visualizations from the dataset. The underlying large language model used in this implementation was Gemini 2.5 Pro; however, since system response times were negligible and the study does not aim to compare model performance, this choice is not expected to affect the overall findings. It was designed to support flexible, quick interaction without requiring users to understand visualization tools or terminology.

Among available NLI systems, NL4DV-LLM was chosen because it is a state-of-the-art tool, open-source, and easy to integrate with custom datasets [36]. NL4DV-LLM offers a transparent and modular design, supports multi-attribute queries, and is well-documented, making it a practical and research-aligned choice for our experimental setup. Our work provides an evaluation of the tool from an HCI and usability perspective, which is lacking in prior research [34]. Data from the fish farming dataset was extracted in CSV format and integrated into NL4DV-LLM before each trial to ensure that study participants worked with live data.

3.2 Task Design and Assignment

To evaluate the usability of the dashboard and the natural language interface (NLI), we developed two sets of 10 tasks, each representing realistic data monitoring and analysis activities in a fish farm context. Each participant interacted with both interfaces but was randomly assigned to one of the two task sets per interface to minimize bias from repetition and memorization. The task set was validated with domain experts to ensure that it reflected realistic analytical and monitoring practices while remaining comprehensible and achievable for domain novices.

Tasks were structured to reflect a range of demands. They varied along several dimensions:

- **Parameter scope**: single-variable tasks versus multi-variable comparison tasks.
- **Focus type**: trend analysis, threshold-based filtering, and extreme value identification (e.g., maximum or minimum).
- **Spatial dimension**: some tasks required location- or depth-specific reasoning.

Below, we provide representative examples illustrating how one could use both the dashboard and the NLI interface to complete

Figure 1: GUI-based dashboard interface for real-time fish farm monitoring [17]

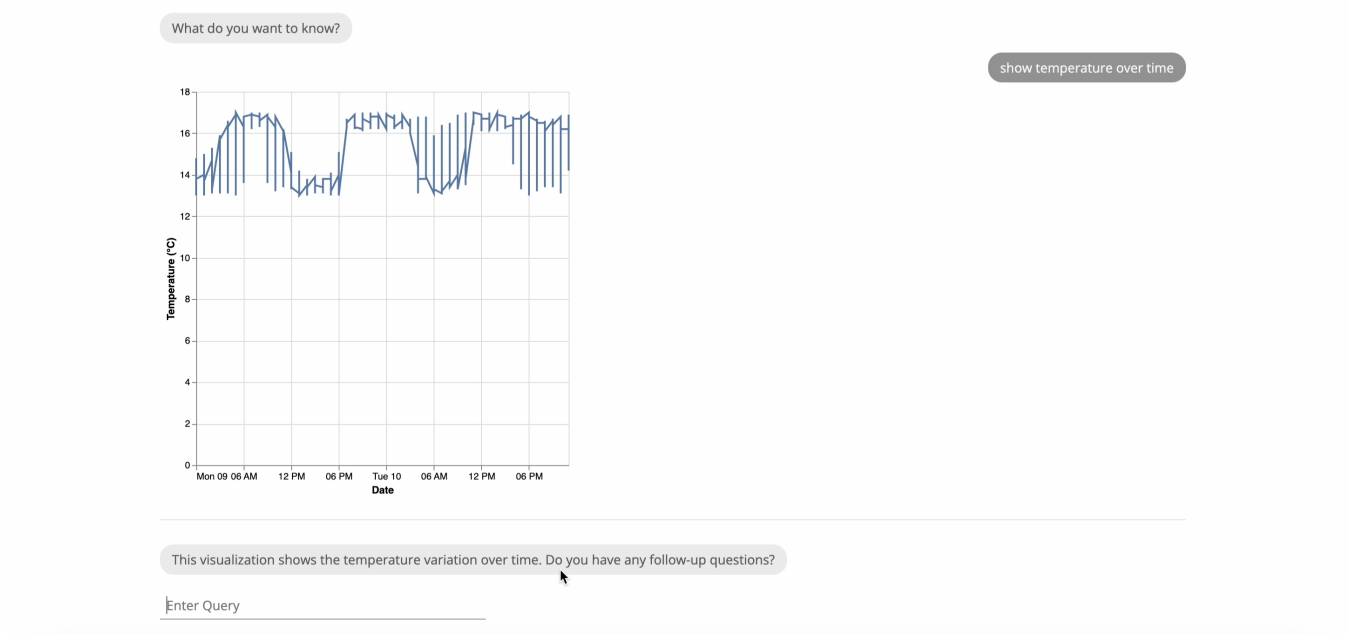

Figure 2: Chatbot-based Natural Language Interface [34]

various monitoring and analysis tasks. The solution approaches model those taken by someone with tool expertise.

3.2.1 Single-Parameter: Trend or Threshold. **Example:** *Does the Temperature (°C) decrease throughout Jan 7?*

Dashboard expert approach: The user filters the time range to "Jan 7" and selects "Temperature (°C)" from the parameter list to visualize changes over the day.
NLI expert approach: The user types "Show the temperature trend

for Jan 7." If the response is unclear, they may follow up with "Add a trendline."

3.2.2 Multi-Parameter: Relationship Analysis. **Example:** *When Dissolved Oxygen (%) increases, does Salinity decrease on Jan 7?*
Dashboard expert approach: The user overlays both "Dissolved Oxygen (%)" and "Salinity (PSU)" on the same line chart and checks if increases in oxygen align with decreases in salinity.
NLI expert approach: The user asks "Plot Dissolved Oxygen (%) and Salinity for Jan 7 on the same chart" and may follow up with "Highlight regions where DO increases and Salinity decreases."

3.2.3 Extreme Values: Maximum or Minimum Detection. **Example:** *What is the minimum Salinity recorded on Jan 7, and where does it occur?*
Dashboard expert approach: The user filters for Jan 7, inspects the line chart, and hovers to find the minimum value and corresponding location in the tooltip.
NLI expert approach: The user types "What is the lowest Salinity (PSU) recorded on Jan 7 and at which location?"

3.3 Study Design

The study employed a within-subject design, where each participant used both the GUI dashboard and the NLI chatbot. The order of interface exposure was randomized to minimize potential bias and learning effects. Participants completed the same number and types of tasks using both interfaces (similar tasks in each task set typically differed in terms of the dates, sensors, or attributes involved), allowing for a direct comparison of performance between interaction methods. Task set assignment was balanced across interface conditions.

3.4 Participants

A total of 24 participants (12 female, 12 male) were recruited for the study, a sample size commonly used in controlled within-subject HCI studies to ensure sufficient statistical power while maintaining feasibility. Participants were aged between 21 and 38 years ($M = 27.9$, $SD = 4.6$). All participants were students from varying academic majors with at least basic computer literacy and some familiarity with data visualization tools. Participants had no prior experience with NL4DV interfaces or fish farm data.

3.5 Tasks and Scenarios

The study employed two sets of ten tasks each. Each task was based on real-world data exploration scenarios. Tasks required participants to extract insights from real-time monitoring data, such as identifying recent trends and anomalies, and conducting comparative or temporal analyses. The tasks were designed to have different levels of complexity, ranging from simple fact retrieval (e.g., identifying peak values) to more complex analytical queries (e.g., comparing multivariate trends over time). The task design was informed by established evaluation strategies from prior work, particularly the task-framing methods outlined by Srinivasan and Stasko [43], by defining tasks around common analytical intents such as lookup, comparison, and trend identification, and varying task complexity to enable fair comparison across interfaces. Task sets were directly comparable: each task in one set had a counterpart in the other. In the presentation of results, tasks are numbered T1-T10; these numbers refer to the "counterpart" pair of tasks in each set. The assignment of task set to interface condition was counterbalanced.

3.5.1 Task Validation. The experimental tasks were designed in collaboration with a Fish Tracking Application Scientist at Innovasea (an expert user of the RealFishPro dashboard) to ensure that they reflected typical monitoring workflows, including trend detection, correlation checks, and value lookups. In a validation session, the same expert completed all tasks using the chatbot interface to confirm that the task sets and chatbot interaction design aligned with expectations.

3.6 Procedure

After giving informed consent, participants engaged in two interface conditions (Dashboard, Chatbot, counterbalanced order). Within each condition, participants completed three phases: video tutorial and practice session, task execution, and post-condition questionnaires. After both conditions, a semi-structured interview was conducted.

In the first phase of each interface condition, participants watched a tutorial video specific to that interface to familiarize themselves with its functionalities. Following the video tutorial, they engaged in a short practice session to interact with the interface before attempting the experiment tasks. This ensured that all participants had sufficient understanding and comfort with both interaction methods before data collection commenced. To mitigate learning effects on experiment task performance, every participant also completed a five-minute guided tour and two warm-up tasks using the interface before phase two began. In phase two, participants completed a set of 10 tasks using the interface. Task order was randomized, and a 180-second time limit was imposed for each task. In phase three, we administered the System Usability Scale (SUS) to measure perceived usability, the NASA Task Load Index (NASA-TLX) to evaluate cognitive workload, and a custom questionnaire (see Table 1).

3.7 Evaluation Metrics

The effectiveness of the GUI dashboard and NLI chatbot was assessed using a combination of quantitative and qualitative metrics.

Task completion time was recorded to evaluate the efficiency of each interface in facilitating analytical workflows. System response times for both interfaces (dashboard rendering and NLI query execution) were consistently fast and considered negligible relative to user interaction time; therefore, latency was not included as a separate evaluation factor. In addition, task completion rate was recorded based on the proportion of participants who completed each task within the defined 180-second time limit. Task correctness was evaluated by verifying the accuracy of user responses against the dataset they used.

Cognitive load was assessed using the NASA-TLX questionnaire. Usability was assessed using the System Usability Scale (SUS) questionnaire.

Paired-samples t-tests were used for continuous outcome measures when the assumptions of the test were appropriate. Specifically, we verified that observations were paired within participants,

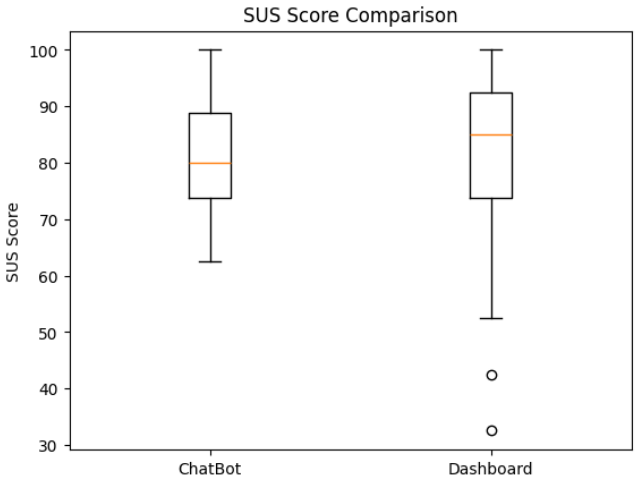

Figure 3: Boxplot of SUS scores for the Chatbot and Dashboard interfaces. Boxes show the interquartile range, horizontal lines indicate medians, whiskers show the non-outlier range, and points indicate outliers. Interface differences were assessed using a paired-samples t-test; no significance marker is shown because the difference was not statistically significant.

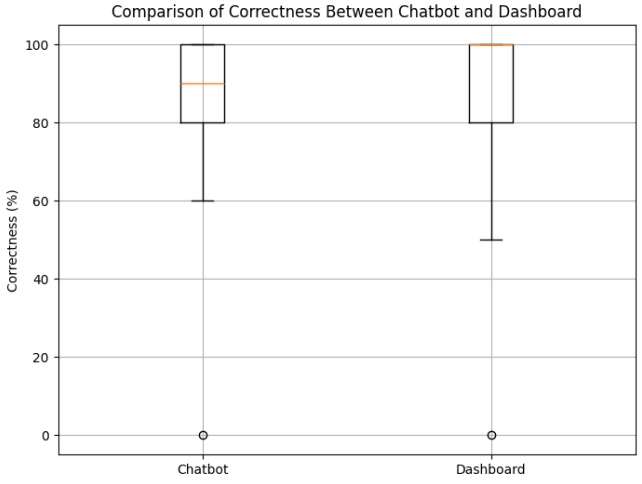

Figure 4: Boxplot of task correctness scores for the Chatbot and Dashboard interfaces. Scores represent the percentage of correctly completed tasks per participant. Boxes show the interquartile range, horizontal lines indicate medians, whiskers show the non-outlier range, and points indicate outliers. Interface differences were assessed using paired statistical comparisons; no significance marker is shown because significance results are reported in the text.

that the dependent measures were continuous, and that the within-participant differences were approximately normally distributed using Q-Q plots and Shapiro-Wilk tests. For ordinal questionnaire items, or where parametric assumptions were not appropriate, we used Wilcoxon signed-rank tests. McNemar tests were used for paired binary task-completion outcomes. Corrections for multiple comparisons were applied where appropriate.

4 Results

4.1 Usability Scores and Task Performance

To assess overall usability differences, the System Usability Scale (SUS) scores were compared between the chatbot and dashboard interfaces. A paired-samples t-test was conducted to compare System Usability Scale (SUS) scores between the chatbot and dashboard interfaces. The analysis did not provide evidence of a statistically significant difference in usability scores between the two interfaces, $t(23) = 0.18$, $p = 0.86$. SUS scores were treated as approximately normally distributed after checking the distribution of paired differences using Q-Q plots and Shapiro-Wilk tests, consistent with prior HCI studies reporting comparative usability results. The boxplot representation of SUS scores (Figure 3) illustrates a similar distribution, with overlapping interquartile ranges, though the chatbot interface exhibited slightly less variability. Additionally, two outliers in the dashboard scores suggest that some participants experienced difficulties. These results suggest that participants gave broadly similar usability ratings to both interfaces, while individual user experiences varied, individual user experiences varied, particularly for the dashboard, which had a broader range of responses.

4.2 Task Correctness and Efficiency

To evaluate task accuracy, correctness scores were analyzed for both interfaces. The boxplot comparison of correctness percentages (Figure 4) shows that users achieved similar accuracy when completing tasks using the chatbot and dashboard interfaces. A paired-samples t-test was conducted to compare correctness between the two conditions. The results indicated no significant difference in correctness, $t(23) = -0.2955$, $p = 0.7704$, suggesting that despite differences in interaction styles, users were able to obtain correct responses with comparable reliability in both systems.

For each interface, we consider the proportion of participants who completed each task within the 180-second time limit, based on the recorded completion logs and task completion times. Overall, participants were more likely to complete tasks within the time limit when using the dashboard interface. When pooled across all tasks, the dashboard yielded a higher completion rate (87.7%) than the chatbot (70.5%).

At the task level, completion rates were consistently higher for the dashboard on most tasks, with particularly large differences observed for T3 and T6 (both 100.0% for the dashboard vs. 59.1% for the chatbot). Paired McNemar tests indicated statistically significant differences for T3 and T6 after correction for multiple comparisons ($p = .004$); no other significant differences were found.

4.3 Task Completion Time

Task completion times were analyzed for all fully completed tasks. A paired t-test comparing mean completion times across all tasks revealed a significant overall difference between the chatbot and dashboard interfaces, $t = 4.34$, $p = .0025$, $d_z = 1.45$, 95%, indicating

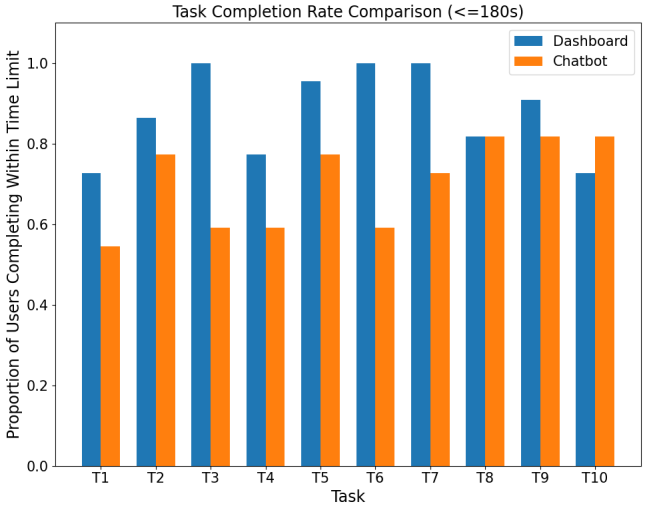

Figure 5: Task completion rates for the Chatbot and Dashboard interfaces across the ten study tasks. Bars show the proportion of participants completing each task within the 180-second time limit. Task-level differences were assessed using McNemar tests for paired binary completion outcomes; no significance markers are shown because corrected significance results are reported in the text.

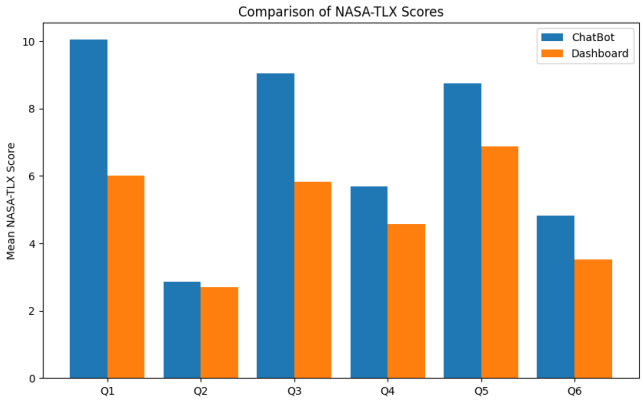

Figure 6: Mean NASA-TLX workload scores for the Chatbot and Dashboard interfaces across six workload dimensions: mental demand, physical demand, temporal demand, perceived performance, effort, and frustration. Bars show mean scores for each dimension. Interface differences were assessed using paired comparisons with correction for multiple comparisons; no significance markers are shown because corrected significance results are reported in the text.

that participants generally completed tasks faster using the dashboard. To identify which specific tasks contributed the most to this effect, separate paired t-tests were conducted for each task (T1–T10) with Bonferroni corrections applied to control for multiple comparisons ($\alpha_{\text{Bonf}} = .005$). After correction, three tasks (T3, T6, T7) showed significantly lower completion times, while other tasks did not reach statistical significance.

4.4 Cognitive Load (NASA-TLX) Analysis

Cognitive load was assessed using the NASA Task Load Index (NASA-TLX), comparing the perceived effort required when using each interface. Paired t-tests were conducted across the six NASA-TLX dimensions with Bonferroni correction applied ($\alpha_{\text{Bonf}} = .0083$). The chatbot interface yielded higher temporal demand ($t = 2.96$, $p = .0072$) than the dashboard, in line with the task time results. While there were higher NASA-TLX scores for the chatbot across multiple dimensions, as illustrated in Figure 6, no other differences were significant after correction.

4.5 Post-condition Questionnaires

After each interface condition, participants answered a custom set of ordinal questions to capture their perceptions and preferences. Each participant responded to the same set of questions for both interfaces, allowing for direct comparison. The questions are shown in Table 1.

The distribution of ordinal responses across seven paired questions reveals notable contrasts between the Chatbot and RealFish-Pro (Dashboard) interfaces (see Figure 7). In general, the Dashboard interface received higher proportions of favourable ratings (scores

Table 1: Custom Questionnaire for Chatbot and Dashboard Interfaces

Q1	How clear were the visualizations in the interface?
Q2	Did the visualizations in the interface accurately represent the data?
Q3	How helpful were the interface visualizations in understanding the data?
Q4	Did the interface visualizations give enough detail for your analysis?
Q5	How visually appealing were the interface visualizations?
Q6	How satisfied were you with the interface for visualizing data?
Q7	How likely are you to use the interface again for data visualization?

4-5). For instance, Q1 (Dashboard) shows a greater concentration in the top two categories (91%) compared to Q1 Chatbot (61%), indicating that more participants found Dashboard visualizations clearer. In contrast, the Chatbot interface displayed a wider spread across mid to lower scores, suggesting a more mixed perception.

Paired Wilcoxon signed-rank tests were conducted for each question. After Holm correction for multiple comparisons, statistically significant differences favouring the Dashboard were observed for Q1 (clear visualizations), Q2 (accurate data representation) and Q4 (sufficiently detailed) (all $p < .05$), while the differences for the remaining items did not reach statistical significance.

4.6 Thematic Analysis

To conduct the thematic analysis, we collected user responses from three sources: open-ended questions for both interfaces, interview

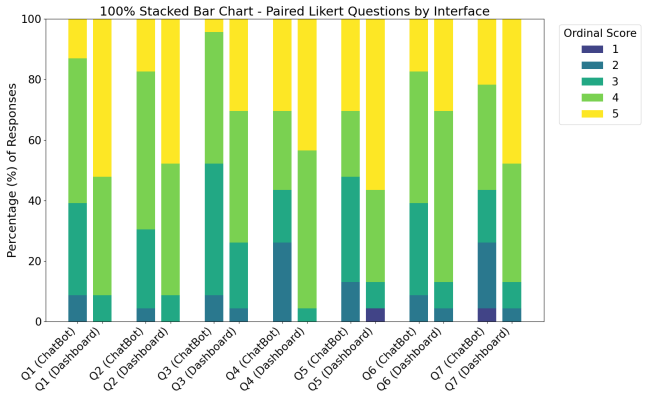

Figure 7: Distribution of responses to the custom paired Likert questionnaire for the Chatbot and Dashboard interfaces. Stacked bars show the percentage of responses for each ordinal score from 1 to 5, where higher scores indicate more favourable ratings. Interface differences were assessed using Wilcoxon signed-rank tests for paired ordinal responses; no significance markers are shown because significance results are reported in the text.

transcripts, and verbal feedback recorded during the study. We reviewed each comment to understand its content and identify which broad topic it related to. Comments that conveyed similar concerns or ideas were grouped together, and, through several iterations, we refined these groupings by analyzing their relationships. With each round of analysis, we updated and reorganized the structure, gradually developing a hierarchical tree that reflected thematic clusters and their internal logic. This process ensured a data-driven understanding of user feedback.

The iterations resulted in the following main categories and sub-categories:

A. Interactivity and Fine-Grained Control.
Participants described interactivity as two complementary strengths that are not yet balanced. The chatbot offers broad control—one participant noted that "all control of parameters and graphs [was] in my control" (P17)—but this freedom often stops at fine-grained actions such as zooming, adjusting time intervals, or changing axis settings. The dashboard, meanwhile, supports pinpoint detail through direct interaction with data points; as P15 described, they "just hovered over the line and found the highest salinity at specific time and depth easily," although fixed controls can limit certain adjustments and require additional filtering steps.

The subthemes show that users use dashboards when they need precise values and quick clicks, and they use the chat interface when they want to explore and compare ideas. Yet they still ask for easier controls, clearer graphics, and faster replies overall.

A.1 Retrieving Specific Details
Participants drew a clear line between precision and flexibility, generally preferring the dashboard when they needed exact numbers, and the chatbot for custom comparisons. For example, P6 noted that "When one is interested in specific values such as minimum or maximum, GUI is more effective." Several participants also reported

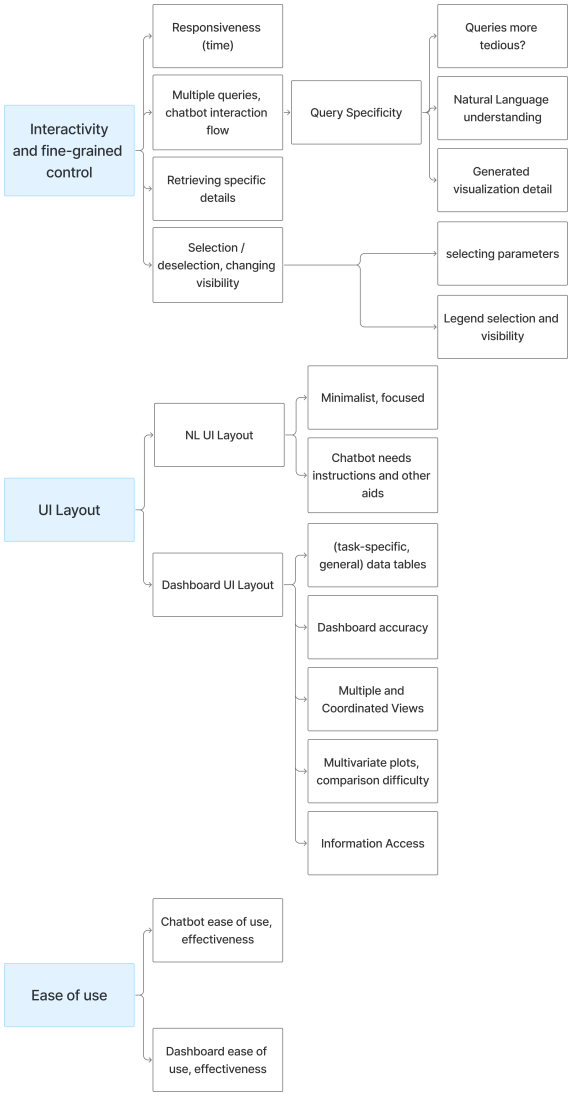

Figure 8: Bottom-up thematic map showing the hierarchical structure of codes and themes derived from user feedback. The analysis began from raw comments and iteratively identified groupings, subthemes, and their relationships, eventually forming high-level themes such as *Interactivity and Fine-Grained Control, UI Layout, Ease of Use.*

difficulty retrieving exact values from the chatbot and emphasized that successful outcomes often depended on prompt wording and refinement, with one participant noting that ambiguity in phrasing could lead to incorrect or incomplete results (P1).

A.2 Multiple Queries, Specificity, and Output Quality
Participants described the chatbot's conversational flow as flexible for refining visualizations in real-time, but often requiring multiple prompts, which made longer tasks tiring (P3: "have to ask many questions to reach your goal"). When queries were simple, responses

were generally accurate and quick, but precision-oriented tasks were perceived as more reliable in the dashboard. Participants also noted readability issues in generated charts, and several reported that repeatedly typing prompts was slower than using dashboard controls for routine tasks (P20).

A.3 Selection, Parameter, and Legend Controls

Participants valued the dashboard's ability to toggle lines, choose parameters, and use legends for filtering, but found these controls increasingly cumbersome with larger datasets. While hiding lines for comparison was praised—P19 noted that "the lines can be hidden for users to compare the specific data"—deselecting multiple items required too many clicks and lacked bulk actions such as "select all" or "reset." Participants suggested clearer grouping, persistent menus, and style adjustments to handle dense charts more efficiently.

A.4 Responsiveness (time)

Participants agreed that speed shapes their experience, noting that fast responses supported smooth interaction, while delays disrupted task flow. For example, P13 reported that it "took some 10–20 sec to load data," which broke interaction flow. Participants observed latency in loading or rendering for both tools and highlighted the overhead of typing prompts in the chatbot compared to clicking dashboard controls. They suggested reducing latency and providing clearer feedback during loading and error states.

B. UI Layout.

The dashboard looks polished, and the chatbot appears clean, yet crowded visuals and small layout issues still reduce clarity. Across subtopics, participants asked to reduce clutter, improve supporting tables, and include more in-context guidance.

B.1 Dashboard UI Layout

Participants praised the dashboard's polished appearance and single-screen view, but reported that crowded charts and dense multivariate plots sometimes obscured key details. For example, P6 noted that having "all visuals on one screen" supported exploration, yet others described the displays as "clustered" (P4) when many variables were shown. While participants valued the accuracy of the data and supporting tables, they requested better placement of summary tables, clearer labels, and improved filtering to reduce visual overload and support faster comparisons.

B.2 Natural Language UI Layout

Participants praised the chatbot's clean, minimalist design for lowering the entry barrier and keeping interactions straightforward. For example, P18 described it as "very easy to use. No complicated syntax," and P6 noted that the simple visuals "did not intimidate me." However, clarity sometimes broke down once results appeared, with participants reporting dense charts, confusing legends, and minor visual bugs. As a result, participants requested more in-context guidance, such as auto-complete suggestions, clarification questions for ambiguous input, and clearer feedback during errors or delays.

C. Ease of Use and Effectiveness (general comments).

C.1 Dashboard ease of use, effectiveness (general comments)

Participants judged the dashboard largely effective, highlighting familiar workflows, strong controls, robust functionality, and clear visuals. For example, P8 described it as "easy to understand on your own and not at all mentally demanding," and P6 emphasized its visual clarity, noting "Clarity … Pleasing to the eye." Participants also valued its capability and control; P20 reported, "I was able to answer the 10th question … which I couldn't do with the chatbot," underscoring the dashboard's effectiveness for more demanding tasks.

At the same time, participants noted areas for improvement, particularly around defaults and data loading. Some reported missing selected trends initially or experiencing friction during loading, suggesting that smoother loading behavior and smarter default settings could further reduce effort and improve efficiency.

C.2 Chatbot ease of use, effectiveness (general comments)

Participants largely agreed that the natural-language interface feels easy and helpful due to its user-friendly design, quick learning curve, and familiar chat-based interaction style. For example, P11 described it as "very intuitive," and P21 noted that it "saves so much time, doesn't take much to learn how to use it," highlighting the low barrier to entry and alignment with everyday interaction habits.

At the same time, some participants reported limitations in depth and polish. Several felt the interface was less equipped for complex or highly specific queries, with P14 noting that it "just needs more humanization" and was "still less equipped & less helpful overall," suggesting that ease of use does not always translate into sufficient analytical power.

D. Expert vs Novice Perception.

Participants expressed mixed views on how expertise influenced interaction with the two interfaces. Some perceived the natural-language tool as more welcoming to novices; for example, P1 noted that it "can help unquantified people work as data analysts and do EDA." By contrast, the dashboard was often seen as requiring prior technical knowledge, with P15 describing it as "made for more technical people" because users need to "understand every component."

However, participants also cautioned that the natural-language interface does not fully eliminate the need for expertise. P16 remarked that it "felt like not for beginners to read all the data," suggesting that both systems still assume some familiarity with data analysis concepts and would benefit from stronger guidance and onboarding.

E. Suggested Extra Features.

Participants suggested additional features to reduce friction once basic tasks felt easy. A fast search box was frequently mentioned to enable quicker parameter access; for example, P5 proposed "adding a searching field to find the needed parameters faster." Others highlighted hands-free interaction, with P14 suggesting a voice interface for situations where users cannot easily interact with the screen. Participants also requested more customization options, including additional chart types, to better tailor visual outputs to their needs.

4.7 Expert Feedback Session

We conducted a live detailed run-through of the study tasks using the NLI chatbot prototype with six Innovasea stakeholders. During this process, the experts provided task-specific feedback,

highlighting cases where queries were easy to complete and visualizations were clear, as well as situations where interaction limitations emerged, such as difficulties comparing closely related values, splitting data at the desired temporal granularity, or displaying multiple variables simultaneously.

The experts also suggested practical improvements, including clearer onboarding, the inclusion of example prompts, support for table-based output, and simplified interface layouts to reduce initial complexity. Collectively, this feedback reinforces many of the study findings. Finally, they noted that NLI integration into dashboard interfaces aligns with current industry trends, citing Yellowfin BI within the Farm360 platform as an example.

Overall, the expert feedback was not used as a separate evaluative dataset, but as contextual validation that the participant findings reflected practical monitoring concerns, particularly around onboarding, output clarity, comparison tasks, and the potential value of integrating NLI capabilities into existing dashboard workflows.

5 Discussion

This study examined how a Natural Language Interface (NLI) and a GUI-based dashboard support real-time monitoring tasks, combining quantitative measures with qualitative feedback from both end-users and domain experts. While the analysis did not show statistically significant differences in usability ratings or task accuracy, they did so through distinct interaction styles, each with corresponding strengths and weaknesses.

The dashboard's fixed controls, hover-to-inspect features, and persistent visual context allowed participants to complete tasks more quickly and with lower mental demand. These patterns mirror the efficiency benefits of direct manipulation reported by Setlur et al. [38] and demonstrated in systems such as DataTone [11] and LIDA [10]. By externalizing query structure and offering stable controls, the dashboard reduced the need for extensive cognitive planning, which is particularly important in time-critical monitoring.

In contrast, the chatbot supported open-ended exploration, enabling participants to combine attributes, pose follow-up questions, and change perspectives without navigating multiple panels. This flexibility reflects the conversational strengths described in Eviza [37], FlowSense [54], and Snowy [42]. However, such freedom can shift interaction costs to linguistic planning and iterative clarification, as cautioned by Srinivasan and Stasko [43]. In this study, vague or overly specific prompts sometimes produced incomplete or incorrect outputs, requiring rewording and multiple attempts, echoing similar challenges identified in Chat2VIS [29].

Quantitative measures reinforced these trade-offs. Robustness analysis indicates that the chatbot as implemented is not great for tasks that involve multiple comparisons. When incomplete tasks are capped at the maximum time limit, significant differences in tasks levels are observed for tasks involving multiple comparison. Although SUS scores did not differ significantly—similar to the results of Bhattacharya et al. [2]—NASA-TLX scores showed that the chatbot required greater mental and temporal effort. This suggests that even with domain-specific schema hints, NLIs may still struggle to match the efficiency of direct manipulation for rapid, precise queries. This pattern is consistent with previous work suggesting that beginners often benefit more from guided interactions and conversational refinement, whereas expert users tend to leverage NLIs for rapid exploratory analysis [3, 4, 6]. It is important to note that learnability and familiarity effects were not explicitly measured in this study; since the evaluation was conducted in a single-session setting, longer-term adaptation and skill development with either interface may influence performance but are outside the scope of this work.

Thematic analysis provided more context for these findings. The dashboard was valued for "trustworthy detail" (P6), "one-click comparisons" (P18), and "seeing everything at a glance" (P3), while the chatbot was described as a "quick teammate" for exploratory checks and ambiguous starting points (P20). Seven participants explicitly recommended a hybrid approach: use the chatbot to create or adjust a visualization, then refine it with dashboard controls (e.g., P9, P14, P21). This aligns with the design direction of hybrid systems like DynaVis [49] and multimodal workspaces described by León et al. [25].

Taken together, these qualitative findings indicate that participants perceived the dashboard and chatbot as complementary rather than competing tools for real-time monitoring. When pushed beyond their respective strengths, both modalities introduced friction: dashboards became cumbersome as the number of variables increased (P11), while chat-based interaction required repeated refinement when prompts were ambiguous or outputs lacked clarity (P1, P23), consistent with limitations noted in prior conversational visualization work [29, 43]. These observations reinforce the need for hybrid designs that combine conversational flexibility with structured visual controls to better support time-sensitive monitoring tasks [49].

The group validation session with six Innovasea stakeholders pointed to the industry relevance of these findings. NLI integration into dashboards aligns with industry trends, with SMEs citing their adoption of Yellowfin BI in the Farm360 platform. They emphasized that conversational tool adoption is still in early stages, making studies like this critical for understanding workflow fit and trust. Their practical suggestions for improving the NLI interface—sample prompts, better onboarding, and simplified layouts—closely matched participant feedback, underscoring the value of blending conversational freedom with structured controls.

5.1 Limitations

Several limitations should be considered when interpreting these results. The evaluation compared only one NLI (NL4DV) and one commercial dashboard (RealFishPro). As shown in Chen et al. [7], different implementations can yield varied usability outcomes. The study was conducted in a controlled lab setting with predefined tasks, ensuring comparability but not capturing long-term adaptation, trust development, or response to unplanned events [32]. While participants came from diverse analytical backgrounds, the absence of complete novices or frontline operational staff limits generalizability to all potential user groups [30].

5.2 Future Work

Future research should evaluate multiple NLIs, including LLM-based systems [41], and a variety of dashboard designs to assess

whether these findings generalize beyond aquaculture monitoring. Longitudinal field studies would help capture learning curves, trust calibration, and sustained usage patterns in operational settings. Hybrid designs that allow seamless switching between conversational and GUI modes—such as initiating a chart through natural language and refining it with direct manipulation—should be prototyped and tested, given strong user and expert support for this approach. Additionally, improvements to NLI capabilities, including better handling of precise queries, enhanced error recovery, and richer domain adaptation, could reduce cognitive load and make them more competitive with dashboards in time-sensitive environments.

6 Conclusion

This study compared a Natural Language Interface (NLI) and a GUI-based dashboard for real-time monitoring of time-series data, combining controlled usability testing with thematic analysis and expert validation. While the analysis did not provide evidence of statistically significant differences in task accuracy or usability ratings, the dashboard excelled in speed, precision, and lower cognitive demand, and the chatbot offered greater flexibility for exploratory and open-ended queries. These results, consistent with prior research, demonstrate that each modality addresses different user needs and that hybrid approaches could harness their complementary strengths. Expert feedback confirmed the operational relevance of this finding, highlighting industry interest in conversational tools but also the current challenges of adoption and integration. By situating these insights within real-world monitoring workflows, the study contributes evidence-based guidance for designing multimodal systems that balance conversational freedom with the efficiency of direct manipulation.

References

[1] Benjamin Bach, Euan Freeman, Alfie Abdul-Rahman, Cagatay Turkay, Saiful Khan, Yulei Fan, and Min Chen. 2022. Dashboard design patterns. *IEEE transactions on visualization and computer graphics* 29, 1 (2022), 342–352. doi:10.1109/TVCG.2022.3209448

[2] Abari Bhattacharya, Barbara Di Eugenio, Veronica Grosso, Andrew Johnson, Roderick Tabalba, Nurit Kirshenbaum, Jason Leigh, and Moira Zellner. 2024. A Conversational Assistant for Democratization of Data Visualization: A Comparative Study of Two Approaches of Interaction. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 17, 6 (2024), e11714. doi:10.1002/sam.11714

[3] Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. 2025. Generative AI at work. *The Quarterly Journal of Economics* (2025), qjae044. doi:10.1093/qje/qjae044

[4] Christopher Bull and Ahmed Kharrufa. 2023. Generative artificial intelligence assistants in software development education: A vision for integrating generative artificial intelligence into educational practice, not instinctively defending against it. *IEEE Software* 41, 2 (2023), 52–59. doi:10.1109/MS.2023.3300574

[5] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *ACM computing surveys (CSUR)* 41, 3 (2009), 1–58. doi:10.1145/1541880.1541882

[6] John Chen, Xi Lu, Yuzhou Du, Michael Rejtig, Ruth Bagley, Mike Horn, and Uri Wilensky. 2024. Learning agent-based modeling with LLM companions: Experiences of novices and experts using ChatGPT & NetLogo chat. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18. doi:10.1145/3613904.3642377

[7] Nan Chen, Yuge Zhang, Jiahang Xu, Kan Ren, and Yuqing Yang. 2024. Viseval: A benchmark for data visualization in the era of large language models. *IEEE Transactions on Visualization and Computer Graphics* (2024). doi:10.1109/TVCG.2024.3456320

[8] Imran Chowdhury, Abdul Moeid, Enamul Hoque, Muhammad Ashad Kabir, Md Sabir Hossain, and Mohammad Mainul Islam. 2020. Designing and evaluating multimodal interactions for facilitating visual analysis with dashboards. *Ieee Access* 9 (2020), 60–71. doi:10.1109/ACCESS.2020.3046623

[9] Victor Dibia and Çağatay Demiralp. 2019. Data2vis: Automatic generation of data visualizations using sequence-to-sequence recurrent neural networks. *IEEE computer graphics and applications* 39, 5 (2019), 33–46. doi:10.1109/MCG.2019.2924636

[10] Stan Franklin, Tamas Madl, Sidney D'mello, and Javier Snaider. 2013. LIDA: A systems-level architecture for cognition, emotion, and learning. *IEEE Transactions on Autonomous Mental Development* 6, 1 (2013), 19–41. doi:10.1109/TAMD.2013.2277589

[11] 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 the 28th annual acm symposium on user interface software & technology*. 489–500. doi:10.1145/2807442.2807478

[12] Grafana Labs. 2024. Grafana. https://grafana.com/. Version 10.0.

[13] Yi Guo, Danqing Shi, Mingjuan Guo, Yanqiu Wu, Nan Cao, and Qing Chen. 2024. Talk2Data: A natural language interface for exploratory visual analysis via question decomposition. *ACM Transactions on Interactive Intelligent Systems* 14, 2 (2024), 1–24. doi:10.1145/3643894

[14] Francisco Gutiérrez, Nyi Nyi Htun, Florian Schlenz, Aikaterini Kasimati, and Katrien Verbert. 2019. A review of visualisations in agricultural decision support systems: An HCI perspective. *Computers and Electronics in Agriculture* 163 (2019), 104844. doi:10.1016/j.compag.2019.05.053

[15] Marti Hearst and Melanie Tory. 2019. Would you like a chart with that? Incorporating visualizations into conversational interfaces. In *2019 IEEE Visualization Conference (VIS)*. IEEE, 1–5. doi:10.1109/VISUAL.2019.8933766

[16] Jieying Huang, Yang Xi, Junnan Hu, and Jun Tao. 2022. Flownl: Asking the flow data in natural languages. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (2022), 1200–1210. doi:10.1109/TVCG.2022.3209453

[17] Innovasea. 2025. *Innovasea: Aquaculture Technology and Fish Tracking Solutions*. https://www.innovasea.com/ Accessed: 2025-03-31.

[18] Donggang Jia, Alexandra Irger, Lonni Besancon, Ondrej Strnad, Deng Luo, Johanna Bjorklund, Anders Ynnerman, and Ivan Viola. 2023. Voice: Visual oracle for interaction, conversation, and explanation. *arXiv preprint arXiv:2304.04083* (2023). doi:10.1109/TVCG.2025.3579956

[19] Ecem Kavaz, Anna Puig, and Inmaculada Rodríguez. 2023. Chatbot-based natural language interfaces for data visualisation: A scoping review. *Applied Sciences* 13, 12 (2023), 7025. doi:10.3390/app13127025

[20] Saadiq Rauf Khan, Vinit Chandak, and Sougata Mukherjea. 2025. Evaluating LLMs for Visualization Tasks. *arXiv preprint arXiv:2506.10996* (2025). doi:10.48550/arXiv.2506.10996

[21] Dae Hyun Kim, Enamul Hoque, and Maneesh Agrawala. 2020. Answering questions about charts and generating visual explanations. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13. doi:10.1145/3313831.3376467

[22] Tae Soo Kim, Yoonjoo Lee, Jamin Shin, Young-Ho Kim, and Juho Kim. 2024. Evallm: Interactive evaluation of large language model prompts on user-defined criteria. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–21. doi:10.1145/3613904.3642216

[23] Woosung Koh, Jang Han Yoon, MinHyung Lee, Youngjin Song, Jaegwan Cho, Jaehyun Kang, Taehyeon Kim, Se-young Yun, Youngjae Yu, and Bongshin Lee. 2024. [illegible]: Scalable Auto-Feedback for LLM-based Chart Generation. *arXiv preprint arXiv:2410.18652* (2024). doi:10.18653/v1/2025.naacl-long.232

[24] Biagio La Rosa, Graziano Blasilli, Romain Bourqui, David Auber, Giuseppe Santucci, Roberto Capobianco, Enrico Bertini, Romain Giot, and Marco Angelini. 2023. State of the art of visual analytics for explainable deep learning. In *Computer Graphics Forum*, Vol. 42. Wiley Online Library, 319–355. doi:10.1111/cgf.14733

[25] Gabriela Molina León, Petra Isenberg, and Andreas Breiter. 2023. Talking to Data Visualizations: Opportunities and Challenges. *arXiv preprint arXiv:2309.09781* (2023). doi:10.48550/arXiv.2309.09781

[26] Maxim Lisnic, Vidya Setlur, and Nicole Sultanum. 2025. Plume: Scaffolding Text Composition in Dashboards. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–18. doi:10.1145/3706598.3713580

[27] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th annual meeting of the association for computational linguistics*. 4487–4496. doi:8653/v1/P19-1441

[28] Alan Lundgard and Arvind Satyanarayan. 2021. Accessible visualization via natural language descriptions: A four-level model of semantic content. *IEEE transactions on visualization and computer graphics* 28, 1 (2021), 1073–1083. doi:10.1109/TVCG.2021.3114770

[29] Paula Maddigan and Teo Susnjak. 2023. Chat2vis: Generating data visualizations via natural language using chatgpt, codex and gpt-3 large language models. *Ieee Access* 11 (2023), 45181–45193. doi:10.1109/ACCESS.2023.3274199

[30] Homeyra Mahmoudi, Silvana Camboim, and Maria Antonia Brovelli. 2023. Development of a voice virtual assistant for the geospatial data visualization application on the web. *ISPRS International Journal of Geo-Information* 12, 11 (2023), 441. doi:10.3390/ijgi12110441

[31] Microsoft Corporation. 2024. Power BI. https://powerbi.microsoft.com/. Version July 2024.

[32] Rishab Mitra, Arpit Narechania, Alex Endert, and John Stasko. 2022. Facilitating conversational interaction in natural language interfaces for visualization. In *2022 IEEE Visualization and Visual Analytics (VIS)*. IEEE, 6–10. doi:10.1109/VIS54862.2022.00010

[33] V Muthumanikandan and Santhosh Ram. 2024. Visistant: A Conversational Chatbot for Natural Language to Visualizations with Gemini Large Language Models. *IEEE Access* (2024). doi:10.1109/ACCESS.2024.3465541

[34] Arpit Narechania, Arjun Srinivasan, and John Stasko. 2020. NL4DV: A toolkit for generating analytic specifications for data visualization from natural language queries. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2020), 369–379. doi:10.1109/TVCG.2020.3030378

[35] Aditeya Pandey, Arjun Srinivasan, and Vidya Setlur. 2022. Medley: Intent-based recommendations to support dashboard composition. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (2022), 1135–1145. doi:10.1109/TVCG.2022.3209421

[36] Subham Sah, Rishab Mitra, Arpit Narechania, Alex Endert, John Stasko, and Wenwen Dou. 2024. Generating Analytic Specifications for Data Visualization from Natural Language Queries using Large Language Models. Presented at the NLVIZ Workshop, IEEE VIS 2024. arXiv:2408.13391 [cs.HC] doi:10.48550/arXiv.2408.13391

[37] 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 the 29th annual symposium on user interface software and technology*. 365–377. doi:10.1145/2984511.2984588

[38] Vidya Setlur and Melanie Tory. 2022. How do you converse with an analytical chatbot? revisiting gricean maxims for designing analytical conversational behavior. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–17. doi:10.1145/3491102.3501972

[39] Leixian Shen, Enya Shen, Yuyu Luo, Xiaocong Yang, Xuming Hu, Xiongshuai Zhang, Zhiwei Tai, and Jianmin Wang. 2022. Towards natural language interfaces for data visualization: A survey. *IEEE transactions on visualization and computer graphics* 29, 6 (2022), 3121–3144. doi:10.1109/TVCG.2022.3148007

[40] Ben Shneiderman and Pattie Maes. 1997. Direct Manipulation vs. Interface Agents. *interactions* 4, 6 (1997), 42–61. doi:10.1145/267505.267514

[41] Yuanfeng Song, Xuefang Zhao, and Raymond Chi-Wing Wong. 2024. Marrying dialogue systems with data visualization: Interactive data visualization generation from natural language conversations. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2733–2744. doi:10.1145/3637528.3671935

[42] Arjun Srinivasan and Vidya Setlur. 2021. Snowy: Recommending utterances for conversational visual analysis. In *The 34th annual ACM symposium on user interface software and technology*. 864–880. doi:10.1145/3472749.3474792

[43] Arjun Srinivasan and John Stasko. 2020. How to ask what to say?: Strategies for evaluating natural language interfaces for data visualization. *IEEE Computer Graphics and Applications* 40, 4 (2020), 96–103. doi:10.1109/MCG.2020.2986902

[44] Brodrick Stigall, Ryan Rossi, Jane Hoffswell, Xiang Chen, Shunan Guo, Fan Du, Eunyee Koh, and Kelly Caine. 2023. On Chatbots for Visual Exploratory Data Analysis. In *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 5924–5929. doi:10.1109/BigData59044.2023.10386335

[45] Andreas Stöckl. 2022. Natural language interface for data visualization with deep learning based language models. In *2022 26th International Conference Information Visualisation (IV)*. IEEE, 142–148. doi:10.1109/IV56949.2022.00031

[46] Md Kazi Abdul Halim Sumon, Md Hamjajul Ashmafee, Md Rafiqul Islam, and Abu Raihan Mostofa Kamal. 2022. Explainable NLQ-based visual interactive system: Challenges and objectives. In *Proceedings of the 2nd International Conference on Computing Advancements*. 420–425. doi:10.1145/3542954.3543014

[47] Roderick Tabalba, Nurit Kirshenbaum, Jason Leigh, Abari Bhatacharya, Andrew Johnson, Veronica Grosso, Barbara Di Eugenio, and Moira Zellner. 2022. Articulate+: An always-listening natural language interface for creating data visualizations. In *Proceedings of the 4th Conference on Conversational User Interfaces*. 1–6. doi:10.1145/3543829.3544534

[48] Tableau Software. 2024. Tableau. https://www.tableau.com/. Version 2024.1.

[49] Priyan Vaithilingam, Elena L Glassman, Jeevana Priya Inala, and Chenglong Wang. 2024. Dynavis: Dynamically synthesized ui widgets for visualization editing. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–17. doi:10.1145/3613904.3642639

[50] Chris Weaver. 2009. Cross-filtered views for multidimensional visual analysis. *IEEE Transactions on Visualization and Computer Graphics* 16, 2 (2009), 192–204. doi:10.1109/TVCG.2009.94

[51] Kanit Wongsuphasawat, Dominik Moritz, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer. 2015. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE transactions on visualization and computer graphics* 22, 1 (2015), 649–658. doi:10.1109/TVCG.2015.2467191

[52] Aoyu Wu, Yun Wang, Mengyu Zhou, Xinyi He, Haidong Zhang, Huamin Qu, and Dongmei Zhang. 2021. MultiVision: Designing analytical dashboards with deep learning based recommendation. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 162–172. doi:10.1109/TVCG.2021.3114826

[53] Cheng Yang, Chufan Shi, Yaxin Liu, Bo Shui, Junjie Wang, Mohan Jing, Linran Xu, Xinyu Zhu, Siheng Li, Yuxiang Zhang, et al. 2024. Chartmimic: Evaluating lmm's cross-modal reasoning capability via chart-to-code generation. *arXiv preprint arXiv:2406.09961* (2024).

[54] Bowen Yu and Cláudio T Silva. 2019. FlowSense: A natural language interface for visual data exploration within a dataflow system. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 1–11. doi:10.1109/TVCG.2019.2934668

[55] Weixu Zhang, Yifei Wang, Yuanfeng Song, Victor Junqiu Wei, Yuxing Tian, Yiyan Qi, Jonathan H Chan, Raymond Chi-Wing Wong, and Haiqin Yang. 2024. Natural language interfaces for tabular data querying and visualization: A survey. *IEEE Transactions on Knowledge and Data Engineering* (2024). doi:10.1109/TKDE.2024.3400824