A Pragmatics-based Approach to Proactive Digital Assistants for Data Exploration

Roderick S Tabalba Jr.
University of Hawaii at Manoa
Laboratory for Advanced Visualization and Applications
Honolulu, Hawaii, USA
tabalbar@hawaii.edu

Christopher J Lee
University of Hawaii
Honolulu, Hawaii, USA
clee48@hawaii.edu

Giorgio Tran
University of Hawaii at Manoa
Laboratory for Advanced Visualizations and Applications
Honolulu, Hawaii, USA
ttran2@hawaii.edu

Nurit Kirshenbaum
University of Hawaii at Manoa
Information and Computer Science
Honolulu, Hawaii, USA
nuritk@hawaii.edu

Jason Leigh
University of Hawaii at Manoa
Information an Computer Science/Natural Science/LAVA
Honolulu, Hawaii, USA
leighj@hawaii.edu

Figure 1: ArticulatePro, a visualization application integrated in the Smart Amplified Group Environment (SAGE3) software [49]. Dyads of participants in the user study talked to a voice-activated digital assistant to create visualizations on the Hawaii Climate Data Portal (HCDP) [25–28].

Abstract

Recent advances in Natural Language Interfaces (NLIs) and Large Language Models (LLMs) have transformed the way we tackle NLP tasks, shifting the focus towards a more Pragmatics-based perspective. This shift enables more natural interactions between humans and voice assistants, which have historically been difficult to achieve. Pragmatics involves understanding how users often speak out of turn, interrupt one another, or provide relevant information without being explicitly asked (maxim of quantity). To explore this, we developed a digital assistant that continuously listens to conversations and proactively generates relevant visualizations during data exploration tasks. In a within-subject study, participants interacted with both proactive and non-proactive versions of a voice assistant while exploring the Hawaii Climate Data Portal (HCDP). Results suggest that interaction with the proactive

CUI '25, Waterloo, ON, Canada

ACM ISBN 979-8-4007-1527-3/25/07
https://doi.org/10.1145/3719160.3736632

assistant increased the total number of utterances and discoveries, facilitated quicker and more reliable insights, and led to greater usage of the system's chart capabilities. Our study highlights the potential of proactive AI in NLIs and identifies key challenges in its implementation, offering insights for future research.

CCS Concepts

• **Human-centered computing** → **Natural language interfaces**; **Empirical studies in visualization**; *Collaborative and social computing devices.*

Keywords

Proactive Digital Assistant, Data Exploration, Pragmatics, Natural Language Interfaces, NLI, Human Computer Interaction, HCI, Data Visualization, User Study, Comparative Analysis

ACM Reference Format:

Roderick S Tabalba Jr., Christopher J Lee, Giorgio Tran, Nurit Kirshenbaum, and Jason Leigh. 2025. A Pragmatics-based Approach to Proactive Digital Assistants for Data Exploration. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces (CUI '25), July 08–10, 2025, Waterloo, ON, Canada.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3719160.3736632

1 Introduction

Advances in Large Language Models (LLMs) like GPT, Llama, and BERT [2, 18, 53] have accelerated progress in NLP applications, despite ongoing ethical debates [1, 8, 31]. Li et al. synthesized the research in Natural Language Processing (NLP) and advocated for a more Pragmatic approach towards Natural Language Interfaces [24]. Pragmatics is a branch of linguistics that describes the nuances that occur in natural language conversations, highlighting context as a crucial factor towards understanding one another. This shift towards Pragmatics gives us the opportunity to reevaluate how we approach interacting with voice assistants to develop Pragmatics-focused Natural Language applications. A Pragmatic voice assistant could better model human communication, supporting roles like AI psychiatrists or tutors.

In human conversations, proactivity involves natural interactions such as interrupting, talking out of turn, and offering the right amount of information [6, 40]. These kinds of interactions are described by Pragmatics [10, 22]. However, current home assistants, such as Siri and Alexa, respond only to direct requests, and even advanced AI chat bots such as ChatGPT and Claude act only when explicitly instructed. This results in AI that mirrors the user's thoughts rather than prompting new ideas or challenging their thinking, which can limit the depth of conversations [56]. Previous research highlights the potential benefits and the growing demand for proactive AI [30, 54, 55]. This research aims to illuminate the benefits and challenges that may occur when users interact with a proactive voice assistant, motivated by the theory of Pragmatics.

To explore the concept of a proactive AI voice assistant, we focused on the context of data exploration. During the past decade, Natural Language Interfaces (NLIs) have made significant progress in assisting users with data exploration tasks [43]. These interfaces simplify the process of creating visualizations to enable users to understand data without programming skills or expertise in visualization [33, 48]. Data exploration serves as the testbed to evaluate the effectiveness of our approach towards proactive digital assistant.

To address the gap in knowledge about proactive Artificial Intelligence, we pose the following research questions:

(1) What are the differences in verbal interaction and outcomes when using a proactive digital assistant versus a non-proactive digital assistant during data exploration tasks?
(2) What are the benefits of interacting with a proactive digital assistant during data exploration?

To explore these questions, we developed two versions of a data visualization application called ArticulatePro. The first version features a digital assistant capable of generating charts only when explicitly asked. The second version features a digital assistant capable of generating charts when either the user explicitly asks for it or when it feels an opportunistic moment to be proactive. Both digital assistants continuously listen to conversations and only operate differently in the timing of chart generation. We evaluated our approach through a within-subject user study, where dyads of participants interacted with both a proactive digital assistant and a non-proactive version of the assistant. By comparing the interactions of the participants under both conditions, our study aims to uncover the potential benefits and challenges of a proactive digital assistant.

This paper makes the following contributions:

(1) A working prototype of a proactive system designed for data exploration tasks.
(2) An evaluation of the effectiveness of our proactive digital assistant.
(3) A comparison of interactions between a proactive and a non-proactive digital assistant.
(4) A discussion of the usability challenges encountered when working with a proactive digital assistant for data exploration tasks.

2 Relevant Works

In this section, we describe the relevant research in natural language for data exploration applications, proactive digital assistants, and Pragmatic theories.

2.1 Natural Language for Data Exploration

Over the past decade, significant advancements have been made in the research of systems that facilitate data exploration tasks using natural language voice commands. This focus enables users to perform data exploration more efficiently, allowing them to concentrate on the exploration process rather than on the generation of charts. Sun et al. developed and evaluated Articulate [48], one of the first data exploration systems enabled by natural language [43]. They found that users, on average, created charts 12 times faster than when using Excel, showcasing the efficiency gains that can be achieved with a natural language system.

Prior work explored hybrid voice and GUI interactions for data exploration [14, 41]; we focus primarily on voice-driven chart creation.

In 2016, Kumar et al. conducted a Wizard-of-Oz study, revealing that only 15% of user utterances were direct queries, while

the remaining 85% provided context for queries during data exploration tasks [19]. Queries lacking sufficient context often failed to generate the intended visualizations. From 2018 to 2023, researchers including Hoque et al., Srinivasan et al., Setlur et al., and Tabalba et al., designed context-supported data exploration systems [17, 41, 47, 50, 52]. These studies demonstrated how incorporating context could help repair incomplete or imprecise queries, emphasizing the importance of context-supported natural language systems, one of the key aspects of Pragmatics, specifically the theory of relevance [56], where people aim to communicate in a way that maximizes relevance.

Between 2017 and 2023, Aurisano et al., Kumar et al., and Bhattacharya et al. iterated on a conversational data exploration system, focusing on resolving co-reference resolution [4, 5, 7, 20]. However, with the advent of OpenAI's Large Language Model (LLMs), models like GPT3 have been shown to perform exceptionally well in addressing co-reference resolution [13], showcasing how LLMs' generalized knowledge can be leveraged to solve previously challenging NLP problems.

2.2 Proactive Applications

The need for proactive systems has been widely discussed. McMillan et al. hired designers to analyze daily conversations, identifying opportunities for proactive AI, such as launching apps like Uber Eats when users mentioned "hungry" or "food" [29]. Volkel et al. surveyed 205 participants and found strong support for proactivity: 87% wanted more suggestions, 83% preferred automatic scheduling, and 80% advocated for assistant-led dialogue [55].

Reicherts et al. explored perceptions of proactive voice assistants through storyboards, finding them useful, appropriate, and pleasant [36]. Similarly, Zargham et al. introduced the "Proactivity Dilemma," noting that proactive NLIs can be both helpful and intrusive depending on context [57]. While we do not evaluate intrusiveness directly, we recognize the need to balance helpfulness and disruption, focusing instead on comparing proactive and non-proactive assistants during data exploration.

Andolina et al. developed a system for proactively retrieving Google search results to reduce query formulation effort [3], while Shi et al. built a brainstorming system that retrieved images based on conversation, reducing lulls and encouraging idea generation [44]. Balaraman and Bernardo developed SimDial, a proactive retrieval system based on Pragmatic theories, reducing user-system dialogue turns during simulated conversations [6]. In contrast, our work extends these ideas to real-world, collaborative data exploration, applying theories of relevance and conversational dynamics to guide a fully operational, always-listening assistant.

2.3 Pragmatics

Li et al. provided an overview of Natural Language Processing (NLP) within the context of Information Science, highlighting semantics, syntax, and pragmatics as core areas [24]. While their work calls for greater attention to pragmatics in Natural Language Interfaces (NLIs), it focuses primarily on linguistic aspects such as context and grounding. In contrast, Skantze offers a detailed account of turn-taking — a foundational aspect of pragmatics — emphasizing its role in coordinating spoken interaction and dialog system design [45].

Our work builds on these perspectives by focusing on proactivity: determining whether the assistant should take a turn based on inferred user intent. This reflects a complementary application of pragmatic reasoning aimed at supporting a natural, context-aware interaction in collaborative data exploration.

Pragmatics, a branch of linguistics, focuses on the nuances of natural language conversations, with context as a central element. According to Cutting, context includes the physical setting, shared background knowledge, and mutual understanding between participants [10]. Context heavily shapes meaning, as illustrated by Grice's concept of "implicatures," where sentences convey implied intent beyond their literal meaning [15]. For example, "The turkey is ready to eat!" could imply either that dinner is ready or that a turkey needs feeding, depending on the setting. Inspired by this, our approach detects user utterances that hint at interests, enabling proactive generation of relevant visualizations rather than relying solely on isolated queries.

Grice further developed the Cooperative Principle, outlining four conversational maxims: quantity, quality, relevance, and manner. Wilson and Sperber later proposed that relevance alone governs effective communication, emphasizing that context determines what is meaningful in a conversation [56].

ArticulatePro builds on these pragmatic theories by leveraging implicatures and relevance to guide proactive behavior. Setlur and Tory [42] explored how Gricean Maxims inform analytical chatbot design, focusing on cooperative behavior and intent interpretation. Extending this work, ArticulatePro implements a fully functional, always-listening assistant that proactively generates visualizations in response to conversational cues. By applying relevance theory to identify opportune moments for intervention, ArticulatePro reduces users' cognitive burden and fosters a more dynamic, context-aware interaction during data exploration.

3 ArticulatePro

In designing our ArticulatePro, we emphasize context by implementing two things. First, we program the assistant to continuously listen to conversations, capturing all utterances it hears. Second, we program the assistant to know what visualizations the users asked for, what visualizations they selected, and what visualizations they most recently interacted with. Inspired by Wilson and Sperber's principle of relevance, we design prompts that guide a series of LLMs to use context in order to extract relevant details from the user's intent to create a visualization.

In the following sections, we first describe an overview of how users interact with ArticulatePro. Then we describe the technical components on how the system works internally.

3.1 An Overview of the System

ArticulatePro is implemented as an application in SAGE3 (Smart Amplified Group Environment[1]), a web application designed to support information-rich collaboration between large display walls and laptop devices [16, 49]. SAGE3 is developed using JavaScript for the frontend and Node.js and Python for the backend. We chose

[1] https://sage3.app

SAGE3 as the foundation for ArticulatePro because it allowed us to leverage its features, such as managing content on an infinite canvas, while simplifying the development stack needed to build the application.

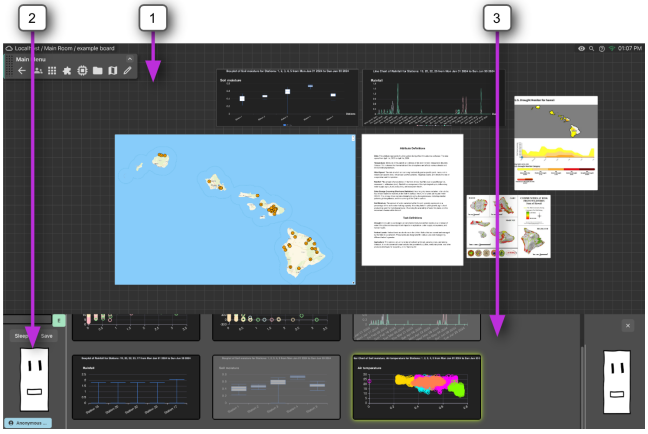

Figure 2: (1) General workspace for moving, resizing, and deleting selected visualizations. (2) The digital persona of ArticulatePro that creates visualizations. (3) A visualization conveyor belt (inspired by a sushi conveyor belt) that displays generated visualizations from the digital assistant. The assistant creates visualizations and displays them on the conveyor belt, where users can select charts to move them to the general workspace for further analysis.

The ArticulatePro application (Figure 2) enables users to explore data through interactions with a digital assistant, Arti, using natural language voice commands. Designed with a Pragmatic approach, Arti continuously listens to user conversations and tracks interactions with charts to leverage context—an essential element in Relevance Theory, a central concept in Pragmatics [10]. This allows Arti to resolve incomplete or ambiguous queries and determine the best moments to generate visualizations — whether explicitly requested or proactively inferred — akin to how humans respond in conversations. In proactive cases, Arti relies on implicatures, or implied meanings, to interpret what users might need even when they haven't directly stated it. Below are descriptions of an explicit request and how Arti proactively creates charts based on such pragmatic cues:

Explicit Request When the user explicitly asks for a chart, Arti behaves similarly to current digital assistants like Siri, Alexa, and Cortana, but without the need for wake words. Arti listens continuously and generates a chart whenever it detects a command. For example, the user might say, "Generate a chart on COVID risk and diabetes rate." In this case, the user directly requests a chart that shows the two variables.

Proactive In proactive instances, Arti generates a chart even when the user hasn't explicitly requested one. For example, if two users are exploring medical data and one says, "I don't get why COVID rates are higher in the East than in the West," Arti might generate a chart comparing medical factors between the East and West. Although the user does not directly ask for a chart, this utterance carries an implicature — a pragmatic inference — that they are seeking an explanation. By responding with a relevant visualization, Arti leverages this pragmatic cue to assist the user's understanding. This illustrates how Arti can be proactive during data exploration tasks by responding to implied intentions rather than only explicit commands

The charts generated by Arti appear on a visualization conveyor belt (Figure 2, Box 3), where users can hover to preview and, if valuable, click on them to move it to the workspace (Figure 2, Box 1) for further analysis, including resizing, comparison, or deletion.

The following sections outline our proactive design process. We first identify opportunities for proactivity in data exploration and then present our solution, rooted in Pragmatic communication theories.

3.1.1 Investigating Proactive Opportunities. To identify opportunities for a digital assistant to act proactively, we conducted an analysis on a prior study[52], where dyads of participants used an always-listening assistant during data exploration tasks. This analysis revealed moments where a digital assistant could intervene proactively—an area not previously explored in the literature. We reviewed utterances from that study and developed six key classifications based on moments where user speech implied, but did not explicitly state, a desire for system support. These pragmatic cues, often realized through conversational implicatures [15], reflect how users naturally communicate in collaborative tasks.

Discovery Users verbalize findings, such as, "As COVID rates increase, so does poverty." The assistant could confirm or challenge these insights.

Disagreement Users express differing views, e.g., "I don't think that's true." The assistant could provide data to clarify disagreements.

Preference Users indicate chart preferences, such as, "I think we should use more maps." The assistant could prioritize similar visualizations.

Criticism Users critique charts, e.g., "I can't read that text." The assistant could adjust the chart's format.

Curiosity Users state future exploration plans, like, "Let's focus on small cities." The assistant could generate relevant charts proactively.

Confusion Users express uncertainty, e.g., "Do doctors count as a resource?" The assistant could offer clarification or guidance.

Our classifications align with prior work on proactive assistants [36, 57], though ours are specific to data exploration. Among these, Discovery was the most prevalent. Focusing on this single classification allowed us to maximize user interactions with the proactive assistant while keeping the study's scope manageable.

3.1.2 Pragmatic Approach for Proactive Opportunities. To guide the development of our proactive assistant, we applied Wilson and Sperber's theory of relevance, a simplified version of Grice's theory on cooperative principles. We incorporate the theory of relevance by implementing a history component that tracks the ongoing conversation, user interactions, and previously generated charts, enabling

the system to produce relevant visualizations while minimizing redundancy (see Section 3.2.3 for implementation details).

3.2 Technical Overview

In this section, we describe the internal components of the ArticulatePro application. Figure 3 displays the architecture of ArticulatePro.

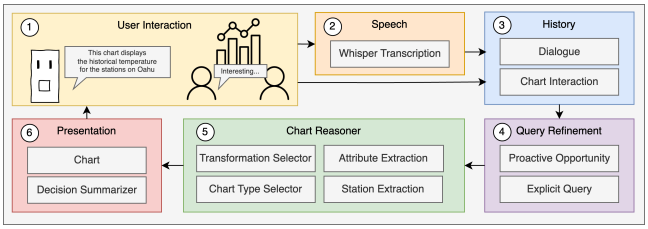

Figure 3: The architectural design of ArticulatePro. (1) User interaction consists of the user's speech and interaction of charts. (2) Speech module transcribes the user's speech using OpenAI's Whisper model [34]. (3) History consists of a history of the user's utterance and chart interaction. (4) Query Refinement creates a succinct query based on the history component. (5) Chart Reasoner extracts and decides what chart to construct and what attributes to visualize (6) Presentation component for chart construction and response generator.

Each component in Figure 3 represents a step in the pipeline, with output from one stage becoming input for the next. Below, we describe each step in more detail.

3.2.1 User Interaction. The system captures user speech, chart selections, and interactions. The assistant presents charts and accompanying textual summaries. When noise or speech is detected, the raw audio is sent to the Speech Module for transcription.

3.2.2 Speech. The speech component is designed to transcribe the collected raw audio from the user interaction component to text. To do this, we process the speech using OpenAI's powerful speech recognition model, Whisper. Once Whisper transcribes the speech, the text is sent to the history component. Any noise or speech that is not transcribed by the speech recognition model is not further processed by the system.

3.2.3 History. ArticulatePro maintains two types of histories:

(1) Dialogue History
(2) User Interactions, which include:
   - Charts that the user selected
   - Charts that the assistant generated
   - The last chart that the user interacted with

According to Relevance Theory, keeping track of the context is crucial for building a successful pragmatic natural language application. Here are specific implementation details on how the system keeps track of history:

Dialogue History ArticulatePro limits the dialogue history to the last five utterances. We experimented with longer context lengths, but this led to the model overemphasizing irrelevant context. For example, early in a data exploration task, the user might say, "We need to focus on rainfall for renewable energy." Later, they might decide that wind speed and solar energy are more important. If the model retains the initial focus on rainfall, it could generate less relevant content. In our prior work, we found found a five-utterance history strikes a good balance between relevance and filtering out outdated content [50, 52]. While other dialogue management strategies could be used, this approach meets our needs.

Chart Interaction History The user interaction history keeps track of the charts that the user interacts with and what the digital assistant generated. This history provides context that allows the system to generate relevant charts and avoids generating duplicate charts. We keep track of the charts by storing the titles of the charts. Following the five-utterance history, we also limit the history to the most recent 5 charts the user selected, most recent 5 charts that the assistant generates, and the last chart that the user selected.

This history is then utilized in the query refinement component.

3.2.4 Query Refinement. The Query Refinement component in ArticulatePro rewrites user queries by incorporating the context from the History component. This ensures that the system generates more relevant charts, corrects incomplete or inaccurate queries, and avoids redundancy by recognizing previously generated visualizations.

For example, if the History component contains utterances about Oahu, such as "Show me places in Oahu that receive the most rainfall. Yeah, click on that one. Okay…", the system may rewrite a follow-up query like "Now show me the soil moisture in that area" as "Generate a chart on the island of Oahu representing soil moisture in areas with the highest rainfall." Similarly, if the History contains references to Kauai, the query would be rewritten to reflect the user's interest in Kauai instead.

ArticulatePro determines whether a user's utterance is an Explicit Query or a Proactive Opportunity using a custom-trained neural network. We opted to train our own lightweight model to meet the system's domain-specific requirements, ensuring real-time performance without the computational overhead of large language models (LLMs). While LLMs could be used for this task, their probabilistic nature often leads to unpredictable outputs, which we sought to avoid to maintain consistency and task alignment. Details on our neural network training are provided in Section 3.2.5.

Below, we describe how ArticulatePro handles different query types:

Explicit Query Explicit queries are utterances where the user directly asks for a chart. Here are some examples:
   - Generate a chart on the solar energy for the Big Island.
   - Show us a graph of the air temperature on Oahu.
   - Display a chart of the highest recorded rainfall measurement in Hawaii.

Proactive Opportunity Using a classification model that we trained, ArticulatePro continously listens to the conversation and classifies each user utterance. If the system classifies an utterance as a discovery/finding, it will try to proactively generate a chart. Once a proactive opportunity is detected, the system creates a new query based on the history. Here

are a few examples of findings or discoveries that the system would detect:

- Station 4 on Oahu has the most rainfall.
- So as fuel efficiency increases, so do sales.
- There are a lot of affordable properties in rural areas.

Non-Query If a query is not detected as either explicit or proactive, it is labeled as a non-query. ArticulatePro does not perform an action on non-queries.

For both Explicit Queries and Proactive Opportunities, ArticulatePro refines the user's query by leveraging the History component to produce relevant visualizations. This ensures that the assistant remains contextually aware and minimizes redundant outputs.

3.2.5 Neural Network Training. ArticulatePro detects query types using a custom-trained neural network. The training data was sourced from our prior work [51], where utterances were manually labeled as Explicit Query, Proactive Opportunity, or Non-Query. Due to the limited number of utterances, we augmented the dataset by generating 700 examples per class using GPT-4.0, a common NLP practice for dataset expansion [38].

The model is a simple feedforward neural network consisting of an input layer (n-dimensional embeddings), a hidden layer with 128 units (ReLU activation), and an output layer with three units. It was trained with the Adam optimizer (learning rate 0.001) and CrossEntropyLoss, using a 60/20/20 train/validation/test split over 20 epochs with a batch size of 32. Model weights were saved at each epoch, selecting the best based on validation loss.

The final model achieved 93.3% accuracy on the test set, with training and validation losses (0.676 and 0.670, respectively) indicating no overfitting.

3.2.6 Chart Reasoner. The Chart Reasoner generates the details required to construct the user's chart. Lee et al. introduced the concept of a "Macro-Query", defined as a broad, high-level request to a system that lacks explicit instructions on which data attributes to retrieve or which transformations to apply [21]. Unlike simple syntactic retrievals, Macro-Queries involve semantic reasoning and implicit information retrieval.

In our work, we address this notion of Macro-Query by leveraging a series of Large Language Models (LLMs) to reason through various sub-tasks involved in chart generation. These tasks collectively assist the system in understanding user queries and constructing relevant visualizations.

Here is a list of tasks the system addresses:

Attribute Extractor Identifies and extracts relevant data attributes from the user's query, such as temperature, rainfall, or time periods.

Station Extractor Retrieves data from the Hawaii Climate Data Portal, filtering for stations that match the user's query criteria (e.g., geographic locations like Kauai or specific climate stations).

Transformation Selector Determines and applies any necessary filters or transformations to the data, such as calculating averages, aggregating by time period, or filtering by location.

Chart Type Selector Chooses the most appropriate chart type to represent the user's query, based on the attributes and transformations identified. For example, a scatter plot requires two numerical attributes, while a bar chart can represent a single categorical variable.

#	Llama3-70b	Llama-8b	Mixtral-8x7b	Gemma-7b	GPT-4o	GPT-4-turbo
1	9.35s	3.44s	DNF	3.07s	12.71s	32.69s
2	5.67s	DNF	2.76s	3.22s	17.58s	30.16s
3	4.23s	3.66s	DNF	DNF	13.63s	32.36s

Table 1: Llama3-70b and GPT-4o took the shortest time to generate a result without encountering any errors that would have caused it to not finish.

The Attribute Extractor, Station Extractor, and Transformation Selector can be executed in parallel, as they do not depend on each other. The only task that is order-dependent is the Chart Type Selector, which requires the outputs from the other steps to determine the appropriate chart type. For instance, a scatter plot requires two numerical attributes, while a time series chart requires a time dimension.

We used different LLMs for each extraction and selection task, prioritizing speed and accuracy. The following models were evaluated for performance across tasks: GPT-4o, Llama3-70b, Llama3-8b, Mixtral-8x7b, Gemma-7b, and GPT-3-turbo-preview. Based on our experiments, we selected Llama3-70b for tasks that required speed and GPT-4o-2024-05-13 for tasks that required greater reliability. Table 1 shows the average completion time (in seconds) for each model to process the query: "Show me a chart on air temperature for Kauai" over three attempts.

We also employed a prompt engineering technique that includes adding a reasoning step, as described in the paper [37]. The reasoning step was added for two reasons:

(1) Performance Improvement: Prior research has shown that asking the LLM to justify its answers can increase the system's performance.
(2) User Feedback: At the end of the decision-making process, the reasoning steps, along with the decisions, were summarized by a specialized agent. This summary was then presented to the user as 1 or 2 sentences.

3.2.7 Presentation. In the final component, the system produces two outputs: the chart for the user and a summary of what it has generated. ArticulatePro uses the information gathered from the Chart Reasoner component to construct the chart.

First, the system uses the stations extracted by the Station Extractor and the attributes from the Attribute Extractor to fetch the necessary data. Next, it applies the transformations selected by the Transformation Selector. Finally, based on the chart type chosen by the Chart Type Selector, the system utilizes pre-defined code to construct the chart and fills in the required details. For this process, we use the ECharts library by Apache [23].

This final component also generates a summary of its decisions for the user. It uses the information gathered from the reasoning step in the Chart Reasoner component. We use the LLM Llama3-70b-8192 to summarize these decisions.

4 Evaluation

In this section, we outline our evaluation methods to address the following research questions:

(1) What are the differences in the number of verbal interaction and outcomes when using a proactive digital assistant versus a non-proactive digital assistant during data exploration tasks?
(2) What are the benefits of interacting with a proactive digital assistant during data exploration?

To answer these questions, we conducted a comparative within-subject user study where participants interacted with two conditions: P where the proactive version of ArticulatePro was used and NP, where the non-proactive version of ArticulatePro's digital assistant was used. In the non-proactive version of the system, we simply turn off the system's ability to proactively generate a chart. So, in the non-proactive version, the application will only generate charts when the users explicitly ask for one. To simplify these conditions to the participants, the agent in the P condition was referred to as "Arti", and the agent in the NP condition was referred to as "Marti". Using these names helped the participants differentiate between the two conditions as well as added to the anthropomorphization of the agents.

4.1 Answering Research Question 1

To evaluate interaction for Research Question 1, we defined interaction metrics as:

(1) Total number of utterances
(2) Total number of task-relevant keywords spoken

To evaluate outcomes for research question 1, we defined an outcome metric as the total number of "good utterances" that occurred during the session. We defined the meaning of a "good utterance" using a codebook developed with three researchers from our data visualization laboratory. We defined a "good utterance" as reflecting a discovery, insight, finding, or decision related to the dataset. See Section 5.3 (Discovery Analysis) for more details.

4.2 Answering Research Question 2

To evaluate Research Question 2, we designed a semi-structured post-interview for the participants. The interview allowed us to gain insight on what benefits the proactive assistant had. In our discussion, we interpreted the participants' feedback in conjunction with the metrics from Research Question 1 and our own observations. This allowed us to verify the insights that we gathered with quantitative evidence that we measured.

4.3 Participants

We recruited 24 participants, grouped into dyads to form 12 groups. The reason we chose to use dyads in the study was to foster a comfortable and natural environment where participants could freely interact with each other, allowing us to capture their genuine behavior. The use of dyads was also adopted in similar studies [3, 51, 52]. Participants were aged 20 - 47, with backgrounds in Computer Science, healthcare, and engineering. This study was approved by an ethics board and all participants consented to participate in the study and be audio and video recorded. Participants were compensated with a $30 Amazon gift card.

4.4 User Study Tasks

We used a subset of the Hawaii Climate Data Portal (HCDP) dataset, a climate repository developed by the University of Hawaii at Manoa [25–28]. The HCDP [2] provides high-quality climate data, including temperature and rainfall, collected from sensor stations across Hawaii. Our subset included six months of data (Jan 1–Jun 31, 2024) from 33 stations: 3 in Kauai, 3 in Oahu, 1 in Molokai, 10 in Maui, and 16 in Hawaii. Participants could query rainfall, temperature, soil moisture, solar radiation, and wind speed. The SAGE3 board was preloaded with drought and historical fire risk images to enhance realism.

We used HCDP for its real-world applicability, ensuring a practical data exploration scenario. Participants completed two tasks using this dataset:

(1) Imagine you are a farmer in Hawaii looking for good agricultural land to grow crops and raise cattle. Based on what charts the digital assistant generates, identify areas that you would consider for good agricultural land to grow crops and raise cattle.
(2) Imagine you are responsible for planning land usage in Hawaii for renewable energy. Based on the charts that the digital assistant generates, identify areas in Hawaii that may be good for wind farms and solar panel energy.

Each task was allotted 30 minutes, and participants prepared answers with supporting visualizations. To counterbalance the study, task and condition order were alternated. Task order switched every two sessions (e.g., groups 1–2: Task 1 → Task 2; groups 3–4: Task 2 → Task 1). Condition order followed a similar pattern: odd-numbered groups experienced P (Arti) first, then NP (Marti), while even-numbered groups started with NP, then P. This ensured a balanced evaluation of task and condition order effects.

4.5 User Study Procedure

In this section, we describe the user study procedure.

Participants had individual workstations with desks, microphones, and materials. They shared a large tiled display wall (81 × 205 inches). This setup supported natural collaboration, allowing gestures, pointing, and fluid turn-taking. Only one mouse is active at a time.

The ArticulatePro interface within SAGE3 (See Figure 1) featured a Visualization Conveyor Belt, where charts were rendered based on the active agent's behavior. Participants could hover for previews or click to open visualizations as apps on the SAGE3 board.

Participants sat 30 inches apart to ensure clear communication for both the assistant and researcher. They wore headset microphones, and the researcher introduced the study, agent, and task. Participants were reminded to discuss observations with each other and the assistant. After 30 minutes, they presented their solution with supporting visualizations before switching to the next condition and task.

[2] https://www.hawaii.edu/climate-data-portal/

Before starting, participants completed a training session. They were shown example charts (bar, line, scatter, box plot, and pie) using a known car dataset [46] and practiced interacting with the system by:

- Moving an application by clicking and dragging its window
- Resizing an application by adjusting its corners
- Panning the board by clicking and dragging the background
- Zooming in and out using the scroll wheel
- Selecting an app to interact with it
- Deleting an app

We deliberately avoided training participants on how to ask for charts to capture their natural interactions, as Srinivasan et al. [46] found that providing example queries can limit users to those examples.

4.6 Participant Feedback

After the completion of both tasks, participants were asked to individually fill-out a brief questionnaire assessing their attitude towards Arti and Marti. This questionnaire asked about their personal experience with each of the digital assistants; They were asked to rate on a scale from 1 to 5 whether the assistant was annoying when it interjected with charts, whether the assistant produced relevant content, and whether the assistant produced useful content. They were also asked to choose their preferred agent, Arti or Marti.

This was followed by a joint oral semi-structured interview. Participants were asked for their demographics (including age, major, and experience with visualization) and were guided in providing qualitative feedback based on their individual rates and preferences.

5 Results

This section presents the quantitative and qualitative results from the study and their method of collection.

5.1 Total Number of Utterances

We measured the number of utterances during each session, defined by our speech recognition module. When noise is detected, the speech recognition starts recording. After a 1.5-second pause, the recorded audio is transcribed. If the transcription results in written text, it is counted as an utterance.

We examined whether participants engaged more with the proactive assistant (Arti) than the non-proactive assistant (Marti), and whether the order in which participants encountered these assistants influenced their level of engagement. To do this, we conducted a mixed two-way repeated-measures ANOVA with assistant type (Marti vs. Arti) as a within-subjects factor and condition order (NP→P vs. P→NP) as a between-subjects factor. The ANOVA revealed a significant main effect of assistant type, $F(1, 10) = 21.55$, $p = .0009$, indicating that participants produced significantly more utterances when interacting with Arti than with Marti. There was no significant main effect of order group, $F(1, 10) = 0.20$, $p = .662$. A significant interaction was found between assistant type and order group, $F(1, 10) = 7.31$, $p = .022$, suggesting that the effect of assistant type on engagement varied depending on the order in which participants encountered the assistants.

Given the significant interaction, we conducted sample t-test to examine the effect of assistant type within each condition order.

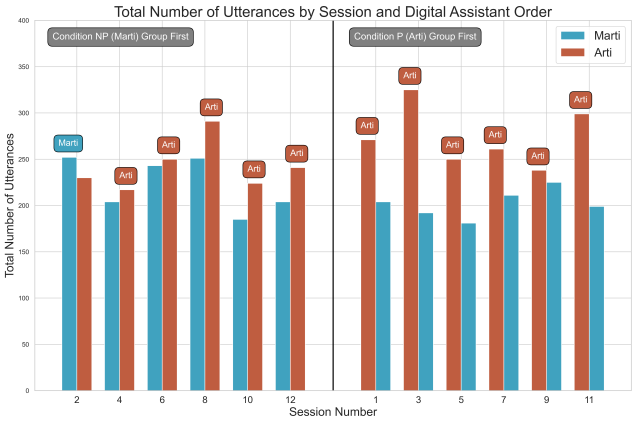

Figure 4: This chart displays the total number of utterances for each group. The chart is separated between the two groups P (Arti) first and NP (Marti) first. In almost every session but 2, users talked more with Arti.

For participants in the P→NP condition (Arti first, then Marti), a right-tailed paired t-test showed a significant difference in utterance counts: participants spoke significantly more to Arti ($M = 274.0$, $SD = 32.5$) than to Marti ($M = 202.0$, $SD = 15.3$), $t(5) = 4.32$, $p = .008$. For the NP→P condition (Marti first, then Arti), no significant difference was found: utterances toward Marti ($M = 223.2$, $SD = 29.0$) and Arti ($M = 242.2$, $SD = 26.7$) were not significantly different, $t(5) = 1.91$, $p = .059$. These results suggest a strong order effect: participants who started with the proactive assistant (Arti) showed more verbal engagement overall, even when switching to the non-proactive assistant. This supports the idea that early exposure to proactive assistance may influence subsequent interaction behavior and expectations.

5.2 Keyword Analysis

The keywords chosen for the keyword analysis include the dataset's variable names: temperature, wind, rainfall, solar, and soil. We also included task-related keywords such as station, fire, drought, farm, and agriculture. Based on our internal discussions in our visualization laboratory, these keywords were deemed most relevant to the task. By performing a keyword analysis, we can determine whether the additional recorded utterances were relevant to the task.

Figure 5 shows the total number of keywords mentioned during the tasks. We conducted a mixed two-way repeated-measures ANOVA to examine the effects of assistant type (Marti vs. Arti) and condition order (NP→P vs. P→NP) on the number of keywords produced. The ANOVA revealed a significant main effect of assistant type, $F(1, 10) = 9.66$, $p = .011$, indicating that participants mentioned more keywords when interacting with Arti ($M = 165.2$, $SD = 49.9$) compared to Marti ($M = 126.6$, $SD = 31.2$).

There was no significant main effect of condition order, $F(1, 10) = 1.54$, $p = .242$, and no significant interaction between assistant type and order group, $F(1, 10) = 0.05$, $p = .824$. These results suggest that the proactive assistant (Arti) led to greater use of task-relevant

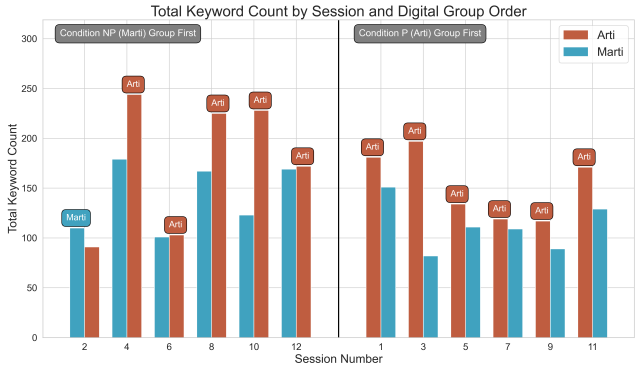

Figure 5: This chart displays the total number of keywords for each group. The chart is separated between the two groups P (Arti) first and NP (Marti) first. Similarly in the total utterance section, users used more task-relevant keywords in almost every session but session 2.

language across sessions, regardless of whether participants first interacted with Marti or Arti.

5.3 Discovery Analysis

In this section, we aim to measure the outcomes of the data exploration task. We defined outcomes as the number of "good utterances" that occurred during the session. Defining a "good utterance" is subjective, so we developed a code-book with three researchers from a data visualization laboratory to establish criteria for this. To create the code-book, we selected a random subset of utterances from all sessions. A random subset (355 utterances, 95% CI) was coded following standard sampling methodology.

The three researchers then met to discuss commonalities in their definitions and agreed on the following criteria for a "good" utterance:

"The utterance reflects a discovery, insight, finding, or decision related to the dataset. For example, a user may indicate that a station, area in Hawaii, or an entity (even if referred to using a pronoun such as 'that' or 'it') is sufficient, lacking, or comparable across any of the task-related variables (e.g., rainfall, temperature, soil moisture). This criterion is based on the likelihood that such utterances resulted from users examining a chart generated by the digital assistant."

For simplification, we will refer to a "good" utterance as a "discovery."

Here are examples of discoveries that participants made during the sessions:

- "I'm just gonna set solar energy, so 18. Well, 18 for highest solar, yeah."
- "The foundation 20 has a better average than 14 does. Cause if you go—Oh yeah, there's some lows."
- "I feel like station 20 should be a good pick. Right? But it has lower soil moisture."
- "The variability is lower in 23."
- "But 21 has the highest solar and wind speed. I think we saw that somewhere."

To qualify as a discovery, an utterance must contain enough contextual information to be understood as conveying a meaningful observation, insight, or decision related to the dataset. Discoveries typically demonstrate that the user has engaged with the charts and is making comparisons, judgments, or drawing conclusions from the visualized data.

In contrast, non-discoveries are utterances that lack sufficient context to be interpreted as conveying a specific finding or insight. These utterances may refer to the task or dataset in a vague manner but do not provide enough information to demonstrate that the user has reached a meaningful conclusion or made an actionable observation.

Here are examples of non-discoveries due to insufficient context:

- "So it could be used for..."
- "Rainfall for Maui stations."
- "There's only one site."

Unlike discoveries, non-discoveries lack specific references to variables, trends, or comparisons that would indicate the user has drawn a meaningful conclusion from the charts. Without additional context, it is unclear whether the user has made a relevant finding or is simply commenting on the task.

To account for possible inaccuracies introduced by the speech recognition system, researchers ignored minor transcription errors when labeling utterances. Once the final codebook was established, one researcher manually labeled all 5,500 utterances. To minimize bias during the labeling process, utterances were randomized, and the researcher had access only to the utterances themselves without accompanying metadata or context from the sessions.

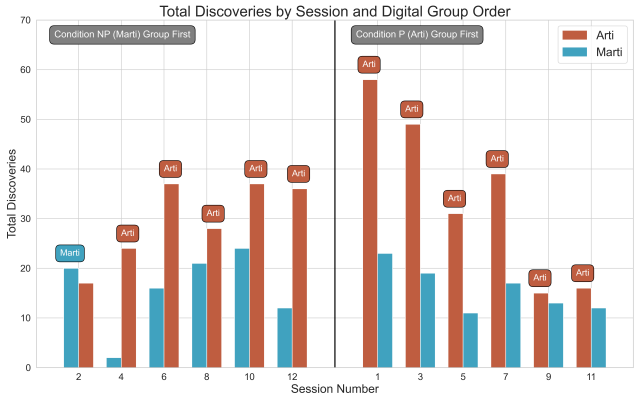

Figure 6: This chart displays the total number of discoveries for each group. The chart is separated between the two groups P (Arti) first and NP (Marti) first. In almost all sessions but session 2, the participants made more discoveries with Arti than with Marti.

Figure 6 shows the total number of "good utterances". As seen in the figure, in every session except session 2, users used more keywords with Arti than with Marti.

We conducted a mixed two-way repeated-measures ANOVA to examine the effects of assistant type (Marti vs. Arti) and condition order (NP→P vs. P→NP) on the number of discoveries made. The ANOVA revealed a significant main effect of assistant type, $F(1, 10) = 22.26$, $p = .0008$, $\eta_p^2 = 0.69$, indicating that participants made significantly more discoveries when interacting with Arti compared to Marti.

There was no significant main effect of condition order, $F(1, 10) = 0.22$, $p = .646$, and no significant interaction between assistant type and order group, $F(1, 10) = 0.48$, $p = .503$. These results suggest that the proactive assistant (Arti) substantially increased the number of discoveries participants made, regardless of whether they encountered Arti or Marti first.

5.4 Delta Time of First Explicit Request

In this section, we observed a notable difference in how users began their tasks between groups (P->NP) and (NP->P). Table 2 shows the time it took users to explicitly request for their first chart after the initial utterance.

Table 2: This table illustrates how participants who started with Marti, were hesitant to explicitly request for their first chart, delaying their data exploration analysis.

Session #	Delta Time (m:s)	Session #	Delta Time (m:s)
2	1:13	1	1:08
4	1:10	3	0:57
6	3:29	5	0:28
8	0:46	7	0:24
10	2:15	9	0:59
12	1:44	11	1:02

While this may not seem like a significant finding at first, the discussion reveals its importance in Section 6.1. We suspect that participants had a quicker and smoother experience getting accustomed to the system when they started with Arti. Those who interacted with Arti appeared to dive into their analysis immediately, whereas participants using Marti showed hesitation in getting started.

5.5 Post-Interview

This section presents feedback gathered from participants during the post-interview, including their individual ratings and open responses from the semi-structured interviews. Figure 7 shows the individual ratings from the participants where we asked the participants how they felt when they interacted with each digital assistant.

5.5.1 Ratings. Here are the specific questions that we asked the participants to rate each assistant:

(1) On a scale from 1 to 5, with 1 being "extremely annoying" and 5 being "not annoying at all", please rate your experience when the Arti/Marti interjected with charts
(2) On a scale from 1 to 5, with 1 being "irrelevant" and 5 being "very relevant", please rate your experience when Arti/Marti interjected with a chart.

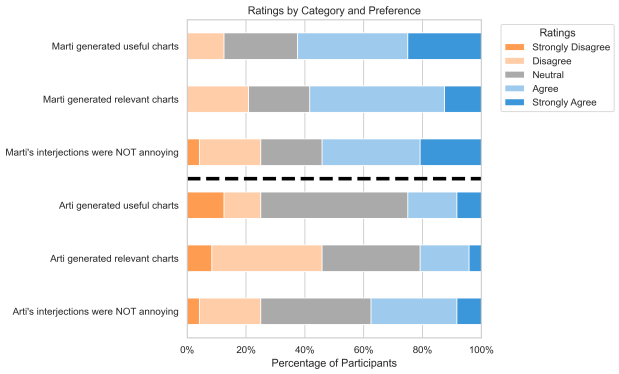

Figure 7: The top 3 bars represent the participants' feedback on Marti and the bottom 3 bars represent the participants' feedback on Arti. The figure indicates that users seemed to prefer their interaction with Marti.

(3) On a scale from 1 to 5, with 1 being "not useful" and 5 being "very useful", please rate your experience when Arti/Marti interjected with a chart.

We asked the participants which digital assistant they preferred, Arti or Marti. Figure 8 shows the participants' responses.

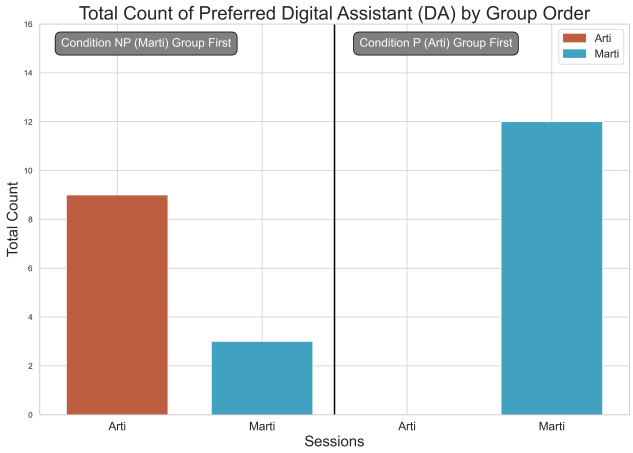

Figure 8: This chart displays the participants' digital assistant preference. The chart is separated between the two groups P (Arti) first and NP (Marti) first. Interestingly, all participants who interacted wtih Arti first, seemed to prefer Marti. And almost all of the participants preferred Arti when they interacted with Marti first.

We indicate the division into the two by-order groups (NP->P) and (P->NP) due to the significant order effects some of our data indicated. Out of 24 participants, 15 preferred Marti, while 9 favored Arti. Interestingly, 9 of the participants who preferred Arti initially interacted with Marti first.

5.5.2 Open-Ended Question. Although most participants preferred Marti, those who liked Arti were predominantly from the group that used Arti second (NP->P). We also asked participants, "If you were to conduct a data exploration task, which digital assistant would you choose, and why?" The responses revealed a more nuanced preference, with several participants expressing interest in using both assistants. Below, we summarize the feedback into three categories:

Participants who strictly chose Arti These participants appreciated Arti's proactive approach, noting that it provided charts to examine while they waited for requested ones. One participant stated, "Arti took the thinking away from the chart generation. It kept throwing information at us, and we could focus on what the data looks like." (s10p1) Another mentioned that "Marti seemed like it was not generating anything useful... Arti seemed like a faster and smarter assistant." (s8p1) Participants also felt that Arti helped them get started: "Didn't know how to start with the system. Arti helped familiarize the user about its capabilities." (s10p2) Others appreciated Arti's conversational presence: "Arti seemed more part of the conversation... listening to what we were saying and generating things that were relevant." (s8p2)

Participants who strictly chose Marti Some participants preferred Marti because they found Arti's proactivity disruptive, breaking their train of thought. As one participant put it, "The charts that Arti was generating broke our train of thought." (s3p1) Another added, "Arti was bombarding us with charts that were irrelevant... it was too much at first because we were getting sidetracked." (s5p2) Others emphasized Marti's pacing: "Marti gave us more time to think and generate the chart we actually wanted." (s3p2) and "Marti was better because it demanded less of our attention." (s7p1)

Participants who chose both A few participants expressed a desire to work with a "toned down" (s11p1) version of Arti. One participant noted, "I would use Arti or Marti for different use cases. I would choose Arti if we were starting from ground zero." (s4p2) and emphasized the need to switch modes mid-session: "Once I get the ideas from Arti, we can nail down what we want to actually look at." (s4p1) Another participant elaborated on their ideal workflow: "Arti was good for exploring, but when we're ready to drill down, we want Marti." (s3p2), explaining that Arti was distracting, but able to generate more ideas. In collaborative settings, Arti was more difficult to manage: "I didn't like the proactive nature of Arti. In a collaborative setting, I would work wih Marti, but indivdually, I could work with either." (s1p1) One participant proposed an integrated solution as a pun: "Or we can put Arti and Marti together and have a Parti!" (s7p2)

6 Discussion

Our results section highlighted the differences in the verbal interaction and outcomes when using a proactive digital assistant versus a non-proactive one during data exploration tasks (Research Question 1). In this section, we present our interpretation of the results in light of Research Question 2: "What are the benefits of interacting with a proactive digital assistant during data exploration?" We identified three key benefits when users engaged with Arti, the proactive assistant: improved system learnability and effectiveness, enhanced reliability of participants' findings, and greater variability in chart types used.

We close the section with an exploration of participants' acceptance of proactive agents and a discussion of potential future directions for the development of proactive digital assistants.

6.1 Improved System Learnability and Effectiveness

The results show that participants made significantly more utterances when interacting with the proactive assistant, Arti. However, a strong order effect emerged: participants who used Marti first spoke slightly more to Arti than Marti, but not significantly so, while participants who used Arti first spoke significantly more to Arti. This raises questions about whether experiencing Marti first restricted users' verbal engagement, whereas experiencing Arti first encouraged freer interaction.

Regarding utterance quality, participants produced significantly more task-relevant utterances—those containing keywords or discoveries—when interacting with Arti, independent of order. This suggests that the increase in utterances contributed meaningfully to task performance. As participants noted (Section 5.5.2), Arti's tendency to proactively generate data offered more opportunities for exploration, aligning with the concept of lateral thinking (De Bono, 1970) [11], where exposure to new information stimulates insight.

Our observations further suggest that participants who started with Arti found it easier to "get started" with analysis. Arti's proactive chart generation helped participants quickly engage with the data, even before issuing explicit requests. In contrast, participants who started with Marti showed a notable delay: in sessions 6 and 10, participants took over 2 and 3 minutes, respectively, to request their first chart, compared to a maximum of 1 minute and 8 seconds for participants starting with Arti. This delay seemed to impact users' confidence and exploration throughout the session.

Overall, in the context of speech-assisted data exploration tasks, proactive systems may have the potential to increase effectiveness (by encouraging unrestricted verbal interaction) and system learnability, a concept related to how users first interact with a system. This enables users to become quickly acquainted with using the system, allowing them to focus more on understanding the data and task at hand.

6.2 Enhanced Reliability of Findings

The proactive assistant, Arti, excelled at presenting data in multiple chart formats, helping users confirm or challenge their hypotheses. For example, in session 12, participants initially misinterpreted a box plot of rainfall in Kauai, concluding there was no rainfall. Arti then proactively generated a scatter plot showing rainfall and air temperature, prompting participants to re-evaluate and correct their conclusion. This illustrates how a proactive system can stimulate "users' cognitive environment", as described by Sperber and Wilson [56].

Proactive systems may thus help users gain confidence in their findings. A similar effect was observed by Reicherts et al., who

developed "ProberBot," a chatbot that prompted users to justify investment decisions, leading to more deliberate and logical choices [35]. However, Reicherts et al. also noted that excessive prompting could bias users. Similarly, while Arti can encourage re-evaluation, it is crucial that proactive systems remain neutral and avoid unduly influencing users' conclusions.

6.3 Greater Usage of Different Chart Types

Arti broadened users' exploration of chart types. A common approach among users was to rely on a single chart type, with line charts being the most prevalent. In session 2, participants primarily used line charts with Marti in the first session, briefly switching to histograms near the end. However, when they interacted with Arti in the second session, Arti quickly generated a box plot—something the users had not explored with Marti. They continued using box plots throughout the session. A similar effect was observed in session 6, where participants initially used only line charts with Marti but explored scatter plots, box plots, histograms, and line charts with Arti, greatly expanding their chart usage. This variety helped participants draw more insights from the data.

In a previous study, Pins et al. examined how people interact with home voice assistants [32]. They found that systems like Google Assistant and Amazon Alexa are often underutilized, with users sticking to simple commands, especially after failed interactions. One issue is that these assistants' capabilities are either listed in a user manual or voiced upon request, making it difficult for users to remember them. In contrast, a visual display can remind users of a system's capabilities. Our study shows the potential of proactive systems to suggest new features, helping users fully utilize the system's capabilities.

6.4 Users' Acceptance of a Proactive Assistant

We found that users did not prefer to start with the proactive assistant. As shown in Figure 8, none of the participants who began with Arti chose to continue working with it. Post-interview feedback suggested that starting with Arti was overwhelming, as participants had to simultaneously learn how to communicate with the assistant, coordinate with their partner, and navigate an unfamiliar environment. In contrast, those who started with Marti had time to familiarize themselves with the system and data before transitioning to Arti, making the shift smoother.

For developers of proactive digital assistants, we advise caution: excessive proactivity early on can overwhelm users. One participant noted that Arti was "too fast," generating charts before they were ready. While starting with a proactive assistant may accelerate learning, users prefer a gentler onboarding. Future work could explore adaptive proactivity based on user signals, building on prior research into context-aware systems using computer vision, sensors, and user behavior tracking [9, 12, 39].

Interestingly, users who ended with the proactive assistant also showed signs of reduced acceptance. Participants who first used Marti often developed workflows that persisted into the second task, making them less receptive to Arti's proactive suggestions. This was particularly evident in sessions 4 and 8, where participants stuck to their established strategies despite Arti's input. In contrast, participants who started with Arti were more open to engaging with its proactive behavior throughout the study.

6.5 Future Directions for Proactive Assistants

A future implementation that may improve the experience working with a proactive assistant may be to separate the content of what the proactive assistant generates and what users explicitly asked for. The participants in session 10 mentioned that it was difficult to differentiate between what the digital assistant generated and what they explicitly asked for. By separating what the assistant proactively generates and what the users explicitly ask for, the user can either focus on their data exploration without being interrupted by what the assistant generates. Then they are given the option to see what the assistant generates for further exploration. Andolina et al., question the possibility of combining proactive content and explicitly generated search results in the same interface [3]. Our study reveals that in the context of data exploration, having proactively generated charts and explicitly generated charts in the same interface may cause some confusion.

7 Conclusion

In this paper, we describe the implementation and evaluation of a proactive data exploration assistant designed to generate relevant data visualizations and support users during data analysis tasks. Our approach was motivated by Pragmatics, which posits that effective utterances are those that are relevant to the conversation. To our knowledge, this paper presents the first computational implementation of a proactive voice assistant for data exploration. Our findings support this theory, as users made more task-related utterances and discoveries with the proactive assistant compared to the non-proactive one. Additionally, our results demonstrate how a proactive assistant can improve system learnability, enhance the reliability of users' findings, and increase verbal interaction with the system's charting capabilities. Specifically, a between-subject design would be more suitable for thoroughly assessing the potential efficiency benefits. Interestingly, while users were more efficient when using the proactive assistant, many expressed a preference for less proactivity. This suggests a delicate balance between offering assistance and allowing users to maintain control over their workflow. When the proactive assistant became too intrusive, users tended to ignore its input in favor of their own strategies. However, reducing proactivity too much could limit the assistant's ability to prompt new discoveries and guide users toward exploring more diverse chart types. Our findings suggest that the proactive assistant positively impacted both task efficiency and the depth of user exploration. As technology continues to advance, there is an opportunity to refine natural language systems to more closely emulate human communication, as described by linguistic theories like Pragmatics. This could enable more intuitive and seamless interactions between users and digital systems.

Acknowledgments

This project is funded in part by National Science Foundation awards 2149133 and 2004014.

References

[1] Zahra Abbasiantaeb, Yifei Yuan, Evangelos Kanoulas, and Mohammad Aliannejadi. 2024. Let the llms talk: Simulating human-to-human conversational qa via zero-shot llm-to-llm interactions. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. 8–17.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[3] Salvatore Andolina, Valeria Orso, Hendrik Schneider, Khalil Klouche, Tuukka Ruotsalo, Luciano Gamberini, and Giulio Jacucci. 2018. Investigating proactive search support in conversations. In *Proceedings of the 2018 Designing Interactive Systems Conference*. 1295–1307.

[4] Jillian Aurisano, Abhinav Kumar, Alberto Gonzales, Khairi Reda, Jason Leigh, Barbara Di Eugenio, and Andrew Johnson. 2015. Show me data": Observational study of a conversational interface in visual data exploration. In *IEEE VIS*, Vol. 15. 1.

[5] Jillian Aurisano, Abhinav Kumar, Alberto Gonzalez, Jason Leigh, Barbara DiEugenio, and Andrew Johnson. 2016. Articulate2: Toward a conversational interface for visual data exploration. In *IEEE Visualization*, Vol. 8.

[6] Vevake Balaraman and Bernardo Magnini. 2020. Proactive systems and influenceable users: Simulating proactivity in task-oriented dialogues. In *Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue-Full Papers, Virually at Brandeis, Waltham, New Jersey, July. SEMDIAL*.

[7] Abari Bhattacharya, Abhinav Kumar, Barbara Di Eugenio, Roderick Tabalba, Jillian Aurisano, Veronica Grosso, Andrew Johnson, Jason Leigh, and Moira Zellner. 2023. Reference Resolution and New Entities in Exploratory Data Visualization: From Controlled to Unconstrained Interactions with a Conversational Assistant. Association for Computational Linguistics.

[8] Martin Juan José Bucher and Marco Martini. 2024. Fine-Tuned'Small'LLMs (Still) Significantly Outperform Zero-Shot Generative AI Models in Text Classification. *arXiv preprint arXiv:2406.08660* (2024).

[9] Rainara Maia Carvalho, Rossana Maria de Castro Andrade, Káthia Marçal de Oliveira, Ismayle de Sousa Santos, and Carla Ilane Moreira Bezerra. 2017. Quality characteristics and measures for human–computer interaction evaluation in ubiquitous systems. *Software Quality Journal* 25 (2017), 743–795.

[10] Joan Cutting. 2005. *Pragmatics and discourse: A resource book for students*. Routledge.

[11] Edward De Bono and Efrem Zimbalist. 1970. *Lateral thinking*. Penguin London.

[12] Renato Ferrero, Mohammad Ghazi Vakili, Edoardo Giusto, Mauro Guerrera, and Vincenzo Randazzo. 2019. Ubiquitous fridge with natural language interaction. In *2019 IEEE International Conference on RFID Technology and Applications (RFID-TA)*. IEEE, 404–409.

[13] Yujian Gan, Massimo Poesio, and Juntao Yu. 2024. Assessing the Capabilities of Large Language Models in Coreference: An Evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 1645–1665.

[14] 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.

[15] HP Grice. 1975. Logic and conversation. *Syntax and semantics* 3 (1975).

[16] Jesse Harden, Nurit Kirshenbaum, Roderick S Tabalba Jr, Jason Leigh, Luc Renambot, and Chris North. 2023. Sage3 for interactive collaborative visualization, analysis, and storytelling. In *Companion Proceedings of the 2023 Conference on Interactive Surfaces and Spaces*. 50–52.

[17] Enamul Hoque, Vidya Setlur, Melanie Tory, and Isaac Dykeman. 2017. Applying pragmatics principles for interaction with visual analytics. *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 309–318.

[18] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, Vol. 1. 2.

[19] Abhinav Kumar, Jillian Aurisano, Barbara Di Eugenio, Andrew Johnson, Alberto Gonzalez, and Jason Leigh. 2016. Towards a dialogue system that supports rich visualizations of data. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 304–309.

[20] Abhinav Kumar, Barbara Di Eugenio, Jillian Aurisano, Andrew Johnson, Abeer Alsaiari, Nigel Flowers, Alberto Gonzalez, and Jason Leigh. 2017. Towards multimodal coreference resolution for exploratory data visualization dialogue: Context-based annotation and gesture identification. In *The 21st Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2017–SaarDial)(August 2017)*, Vol. 48.

[21] Christopher J Lee, Giorgio Tran, Roderick Tabalba, Jason Leigh, and Ryan Longman. 2024. Macro-Queries: An Exploration into Guided Chart Generation from High Level Prompts. *arXiv preprint arXiv:2408.12726* (2024).

[22] Stephen C. Levinson. 1983. *Pragmatics*. Cambridge University Press.

[23] Deqing Li, Honghui Mei, Yi Shen, Shuang Su, Wenli Zhang, Junting Wang, Ming Zu, and Wei Chen. 2018. ECharts: a declarative framework for rapid construction of web-based visualization. *Visual Informatics* 2, 2 (2018), 136–146.

[24] Yan Li, Manoj A Thomas, and Dapeng Liu. 2021. From semantics to pragmatics: where IS can lead in Natural Language Processing (NLP) research. *European Journal of Information Systems* 30, 5 (2021), 569–590.

[25] Ryan J. Longman, Mathew P. Lucas, Jared Mclean, Sean B. Cleveland, Keri Kodama, Abby G. Frazier, Katie Kamelamela, Aimee Schriber, Michael Dodge, Gwen Jacobs, and Thomas W. Giambelluca. 2024. The Hawai'i Climate Data Portal (HCDP). *Bulletin of the American Meteorological Society* 105, 7 (2024), E1074 – E1083. https://doi.org/10.1175/BAMS-D-23-0188.1

[26] Jared Mclean and Sean Cleveland. 2023. Design and Implementation of Web APIs for Supporting Data Product Visualization and Dissemination In Science Gateways. *International Conference on System Sciences* (2023).

[27] Jared McLean, Sean B Cleveland, Michael Dodge, Matthew P Lucas, Ryan J Longman, Thomas W Giambelluca, and Gwen A Jacobs. 2023. Building a portal for climate data—Mapping automation, visualization, and dissemination. *Concurrency and Computation: Practice and Experience* 35, 18 (2023), e6727.

[28] Jared H McLean, Sean B Cleveland, Matthew Lucas, Ryan Longman, Thomas W Giambelluca, Jason Leigh, and Gwen A Jacobs. 2020. The Hawai 'i Rainfall Analysis and Mapping Application (HI-RAMA): Decision support and data visualization for statewide rainfall data. *Practice and Experience in advanced research computing* (2020), 239–245.

[29] Donald McMillan, Antoine Loriette, and Barry Brown. 2015. Repurposing conversation: Experiments with the continuous speech stream. In *Proceedings of the 33rd annual ACM conference on Human factors in computing systems*. 3953–3962.

[30] Christian Meurisch, Cristina A Mihale-Wilson, Adrian Hawlitschek, Florian Giger, Florian Müller, Oliver Hinz, and Max Mühlhäuser. 2020. Exploring user expectations of proactive AI systems. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–22.

[31] Pat Pataranutaporn, Valdemar Danry, Lancelot Blanchard, Lavanay Thakral, Naoki Ohsugi, Pattie Maes, and Misha Sra. 2023. Living memories: AI-generated characters as digital mementos. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 889–901.

[32] Dominik Pins, Alexander Boden, Britta Essing, and Gunnar Stevens. 2020. " Miss understandable" a study on how users appropriate voice assistants and deal with misunderstandings. In *Proceedings of Mensch und Computer 2020*. 349–359.

[33] Ghulam Jilani Quadri, Arran Zeyu Wang, Zhehao Wang, Jennifer Adorno, Paul Rosen, and Danielle Albers Szafir. 2024. Do You See What I See? A Qualitative Study Eliciting High-Level Visualization Comprehension. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 204, 26 pages. https://doi.org/10.1145/3613904.3642813

[34] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*. PMLR, 28492–28518.

[35] Leon Reicherts, Gun Woo Park, and Yvonne Rogers. 2022. Extending chatbots to probe users: Enhancing complex decision-making through probing conversations. In *Proceedings of the 4th Conference on Conversational User Interfaces*. 1–10.

[36] Leon Reicherts, Nima Zargham, Michael Bonfert, Yvonne Rogers, and Rainer Malaka. 2021. May I interrupt? Diverging opinions on proactive smart speakers. In *Proceedings of the 3rd conference on conversational user interfaces*. 1–10.

[37] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927* (2024).

[38] Maximilian Schmidhuber and Udo Kruschwitz. 2024. Llm-based synthetic datasets: Applications and limitations in toxicity detection. *LREC-COLING 2024* (2024), 37.

[39] Albrecht Schmidt, Michael Beigl, and Hans-W Gellersen. 1999. There is more to context than location. *Computers & Graphics* 23, 6 (1999), 893–901.

[40] Rob Semmens, Nikolas Martelaro, Pushyami Kaveti, Simon Stent, and Wendy Ju. 2019. Is now a good time? An empirical study of vehicle-driver communication timing. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

[41] 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.

[42] 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* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 29, 17 pages. https://doi.org/10.1145/3491102.3501972

[43] 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.

[44] Yang Shi, Yang Wang, Ye Qi, John Chen, Xiaoyao Xu, and Kwan-Liu Ma. 2017. IdeaWall: Improving creative collaboration through combinatorial visual stimuli. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative*

Work and Social Computing. 594–603.

[45] Gabriel Skantze. 2021. Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. *Computer Speech & Language* 67 (2021), 101178. https://doi.org/10.1016/j.csl.2020.101178

[46] Arjun Srinivasan, Nikhila Nyapathy, Bongshin Lee, Steven M Drucker, and John Stasko. 2021. Collecting and characterizing natural language utterances for specifying data visualizations. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–10.

[47] 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.

[48] Yiwen Sun, Jason Leigh, Andrew Johnson, and Sangyoon Lee. 2010. Articulate: A semi-automated model for translating natural language queries into meaningful visualizations. In *Smart Graphics: 10th International Symposium on Smart Graphics, Banff, Canada, June 24-26, 2010 Proceedings 10*. Springer, 184–195.

[49] R. Tabalba, N. Kirshenbaum, J. Harden, M. Rogers, A. Nishimoto, E. Christman, A. Yu, R. Theriot, L. Long, L. Renambot, M. Belcaid, C. North, A. Johnson, and J. Leigh. 2023. SAGE3 - the Smart Amplified Group Environment,. *Science Gateways 2023 Annual Conference* (2023). https://par.nsf.gov/biblio/10502367

[50] 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.

[51] 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.

[52] Roderick S Tabalba, Nurit Kirshenbaum, Jason Leigh, Abari Bhattacharya, Veronica Grosso, Barbara Di Eugenio, Andrew E Johnson, and Moira Zellner. 2023. An investigation into an always listening interface to support data exploration. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 128–141.

[53] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[54] Niels Van Berkel, Mikael B Skov, and Jesper Kjeldskov. 2021. Human-AI interaction: intermittent, continuous, and proactive. *Interactions* 28, 6 (2021), 67–71.

[55] Sarah Theres Völkel, Daniel Buschek, Malin Eiband, Benjamin R Cowan, and Heinrich Hussmann. 2021. Eliciting and analysing users' envisioned dialogues with perfect voice assistants. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–15.

[56] Deirdre Wilson and Dan Sperber. 1986. On defining relevance. *Philosophical grounds of rationality: Intentions, categories, ends* (1986), 243–258.

[57] Nima Zargham, Leon Reicherts, Michael Bonfert, Sarah Theres Voelkel, Johannes Schoening, Rainer Malaka, and Yvonne Rogers. 2022. Understanding circumstances for desirable proactive behaviour of voice assistants: The proactivity dilemma. In *Proceedings of the 4th conference on conversational user interfaces*. 1–14.