Semantic Scaffolding: Augmenting Textual Structures with Domain-Specific Groupings for Accessible Data Exploration

Jonathan Zong[12] **Isabella Pedraza Pineros**[13] **Mengzhu (Katie) Chen**[4]
Daniel Hajas[5] **Arvind Satyanarayan**[6]

Abstract. Drawing connections between interesting groupings of data and their real-world meaning is an important, yet difficult, part of encountering a new dataset. A lay reader might see an interesting visual pattern in a chart but lack the domain expertise to explain its meaning. Or, a reader might be familiar with a real-world concept but struggle to express it in terms of a dataset's fields. In response, we developed *semantic scaffolding*, a technique for using domain-specific information from large language models (LLMs) to identify, explain, and formalize semantically meaningful data groupings. We present groupings in two ways: as *semantic bins*, which segment a field into domain-specific intervals and categories; and *data highlights*, which annotate subsets of data records with their real-world meaning. We demonstrate and evaluate this technique in Olli, an accessible visualization tool that exemplifies tensions around explicitly defining groupings while respecting the agency of readers to conduct independent data exploration. We conducted a study with 15 blind and low-vision (BLV) users and found that readers used semantic scaffolds to quickly understand the meaning of the data, but were often also critically aware of its influence on their interpretation.

1 Introduction

When a lay reader encounters a new dataset on a website or news article, an important part of their exploratory process is to identify interesting groupings of data, and explain them in terms of that data's real-world meaning. Broadly, the real-world meaning of data is known as its *semantics*; domain-specific knowledge about data is how a reader determines whether a number "represent[s] a day of the month, or an age, or a measurement of height, or a unique code for a specific person, or a postal code for a neighborhood, or a position in space" [14] or other possible meanings. Effectively drawing connections between data and its semantics can be challenging for a lay reader. For instance, a reader might see an interesting visual pattern in a chart but lack the domain-specific knowledge to explain its meaning. In this case, difficulty arises because the reader does not know what they don't know — it may be difficult to know how to begin to acquire the relevant context. Or, a reader might be familiar with a real-world concept but struggle to translate it into the terms of the dataset. For example, a reader may not know how to approximate a subjective concept like "sports car" in terms of fields like `Horsepower` and `Miles_per_Gallon` — or possibly other fields in the dataset that were not in the initial visualization. These challenges can affect a reader's comprehension of the data — as prior research has shown, a reader's interpretation of a visualization is sensitive to differences in their prior knowledge and personal background [9, 16].

In response, we developed a technique called *semantic scaffolding* for using domain-specific information from large language models (LLMs) to identify, explain, and formalize semantically meaningful data groupings. Rather than use an LLM to generate summary descriptions or chat responses, we use it to return groupings as data structures that include a *name*, *explanation*, and *query predicate*. For instance, in the example cars dataset, an LLM might create a grouping named "Fuel Efficient Japanese Cars" with the following explanation: "This group represents cars from Japan that are known for their fuel efficiency, reflecting Japanese automotive engineering and consumer trends towards sustainable driving" (Figure 1). Crucially, the LLM associates this grouping with the following query predicate: $\texttt{Miles\_per\_Gallon} \geq 25 \cap \texttt{Origin} = \texttt{Japan}$. This allows a reader to connect the name and explanation with an explicitly-defined subset of data. We developed two types of interface elements for presenting semantic groupings to a reader: *semantic bins*, which segment a single field (i.e. column) into domain-specific intervals and categories; and *data highlights*, which annotate subsets of data records (i.e. rows) with their real world meaning. These two uses of semantic scaffolding in an interface serve distinct purposes. Semantic bins help a reader break down a single field into understandable pieces, facilitating navigation and exploration through a dataset. Data highlights help a reader quickly get an overview of a dataset, indicating potentially interesting subsets to begin exploring further.

We prototyped these designs via extensions to Olli [3], an accessible visualization system for screen reader users. Data accessibility is a rich context in which to evaluate our work because current tools for blind and low vision (BLV) screen reader users amplify the challenges lay readers of visualizations face. First, sighted readers often rely on their visual perception to identify interesting groupings in a visualization, but screen reader interfaces rarely afford a similar type of overview. Second, BLV readers highly value the agency to independently explore data, conduct an open-ended interpretive process, and arrive at their own conclusions about the data [13]. For instance, research has shown that screen reader users find descriptions less useful when they over-emphasize contextual and domain-specific infor-

[1] Co-first authors
[2] University of Colorado Boulder, USA, e-mail: jzong@colorado.edu
[3] MIT CSAIL, USA, e-mail: ipedraza@mit.edu
[4] MIT CSAIL, USA, e-mail: mzc219@mit.edu
[5] UCL Computer Science, UK, e-mail: d.hajas@ucl.ac.uk
[6] MIT CSAIL, USA, e-mail: arvindsatya@mit.edu

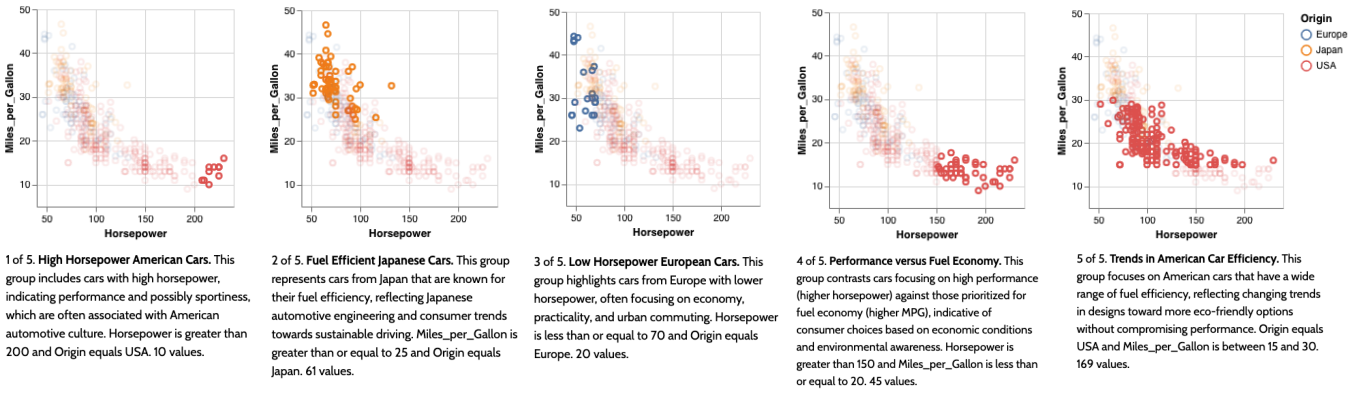

Figure 1. Data highlights generated from an example cars dataset using the *semantic scaffolding* technique. Each highlight is shown with a visualization of its query predicate, demonstrating that highlights correspond to areas of interest in the data.

mation at the expense of descriptive statistics or the data values themselves [13, 5]. These two factors require screen reader interfaces to strike a balance between making groupings available without overly prescribing a reader's interpretation. In a study with 15 blind and low-vision users, we found that textual scaffolding was able to accelerate BLV readers' understanding of a dataset's real-world meaning, but that readers were often critically aware of AI's influence on their interpretation process.

2 Related Work

2.1 Accessible Textual Data Representations

To make data visualizations accessible to screen reader users, a designer must provide descriptions that can be read as text-to-speech. Because conventional static alt text does not afford data exploration at varying levels of detail comparable to strategies sighted readers employ, researchers have turned to *structured textual descriptions* [27], which enable screen reader users to navigate along a hierarchy and move between overview and detail with textual descriptions. Accessible visualization systems that incorporate structured textual descriptions include Olli [3], Data Navigator [6], VizAbility [7], Chart Reader [25], and Umwelt [28]. This work introduces a technique for augmenting textual structures with LLM-generated groupings, and demonstrates the technique via extensions to Olli.

2.2 Data Narratives and Interpretation

Visualization researchers have used the term "narrative visualization" to describe the use of visualizations to tell stories with data [20]. Researchers studying narrative visualization have noted that the interpretation of data is not self-evident; rather, it is shaped by design and communication strategies that incorporate additional context and guide a reader through an interpretation process. For instance, "enhancing structure and navigation" and "providing controlled exploration" as important data storytelling techniques [23].

For users of textual data representations, textual annotation and interactivity (e.g. navigation) are particularly important factors that influence readers' experience of the data. To understand the different kinds of information that authors typically include in textual descriptions or annotations, Lundgard and Satyanarayan [13] define a four-level model of semantic content in descriptions: (L1) elemental and encoded details like chart types and labels; (L2) statistical and relational data such as outliers and correlations; (L3) perceptual and cognitive insights into trends and exceptions; and (L4) contextual and domain-specific knowledge. When they ran a study to find out what content is most useful, they found that BLV users' preferences diverged from sighted users in a significant way. Where sighted readers rated L4 content the most useful, BLV users found it less useful because they wanted "to have the time and space to interpret the numbers for myself before I read the analysis" [13]. So while domain-specific information can be useful in descriptions, they also influence a reader's takeaway and their feeling of agency.

Our work draws on this work to introduce interface designs for guiding a reader's navigation through a textual data representation using domain-specific information.

2.3 Generating Textual Descriptions with Language Models

Natural language generation of textual descriptions for data visualization is an area of research that has received renewed attention due to recent advances in large language models (LLMs). There are generally two interface design approaches that these systems have taken. First are systems that provide a chat-like interface with which a user can query a description by inputting a question in natural language (often known as chart question answering systems) [11, 21, 7]. Second are systems that generate standalone summary descriptions of charts, similar in format to conventional alt text written by humans. This work includes systems like Chart-to-Text [15] and DataTales [24]. The advantage of using LLMs to generate descriptions is that they can automatically incorporate contextual and domain-specific information — also known as L4 semantic content [13] — into descriptions [12]. However, a critical limitation of this approach is the possibility that generated descriptions can contain errors, including hallucinations [10]. But even if generated captions were correct all of the time, there would still be drawbacks to existing approaches. For instance, Choe et al. noted that users sometimes become overreliant on the LLM in a chart question answering system instead of developing their own interpretations of the data [5]. Similarly, summarization-based approaches have the same limitations that conventional alt texts do; namely, that they lack affordances for information granularity and limit users' ability to conduct self-guided exploration [27]. In our work, we introduce a new technique — distinct from Q&A or summarization — for using LLMs to incorporate

domain-specific information. We use an LLM to generate semantically meaningful groupings, and use that output to support structural navigation [27].

3 Semantic Scaffolding: Identifying, Explaining, and Formalizing Meaningful Data Groupings

Semantic scaffolding is a technique for using domain-specific information from a large language model (LLM) to guide a reader's understanding of a dataset's meaning. We engaged in an iterative co-design process involving a blind co-author in which we prototyped methods for incorporating domain-specific information into a user interface for data exploration. This prototyping process revealed different types of user needs that semantic scaffolding could address, which we explore via two types of interface elements: *semantic bins* and *data highlights*.

3.1 Semantic Bins

Binning is a common operation for analyzing and communicating data that involves dividing a field into equally-sized intervals that cover the extent of the field's data values. Most data visualizations implicitly use binning to generate axis ticks, and accessible textual data representations frequently use binning to structure a screen reader's navigation through data.

The conventional approach to binning does not take into account domain-specific information; computing equally-sized bins is a function that can be applied to any dataset. However, lack of semantic information could make it more difficult for a reader to understand and contextualize a field's data values. For example, in a dataset about cars, a bin function might segment the `miles_per_gallon` field by equally-sized increments of 10. However, a reader might not know what range of values is considered a low vs. high fuel efficiency, or whether an increment of 10mpg represents a large or small difference in fuel efficiency. As a result, it might be difficult for them to map the numbers onto their subjective understanding of fuel efficiency.

Semantic bins are groupings that use domain-specific information to segment a field into higher-level intervals and categories that express a dataset's meaning. Figure 2 shows an example usage of semantic binning applied to fields with a variety of Vega-Lite measure types [1] (temporal, quantitative, and nominal). For a continuous (e.g., temporal, quantitative) field, we prompt an LLM to re-bin the field, specifying in the prompt that bins should be non-overlapping intervals that cover the extent of the data. In the figure, the semantic bins map onto how a reader might make sense of years (via historical periods corresponding to agricultural practices), and amount of wheat yield (via levels from low to high relative to the typical or expected yield) (Figure 2A). For a categorical field, we prompt an LLM to group categories into higher-level groupings, specifying in the prompt that the groupings should be mutually exclusive and exhaustively cover all categories. The figure example takes the names of barley varieties and groups them into meaningful high-level categories, such as "Heritage Varieties" or "Modern Breeds" (Figure 2B).

3.2 Data Highlights

Data highlights are groupings of data records along criteria that correspond to a real-world interpretation. In contrast to a semantic bin, which is defined by a predicate involving only one field, a data highlight's predicate can involve multiple fields to select a subset of data records. For example, each data highlight in Figure 1

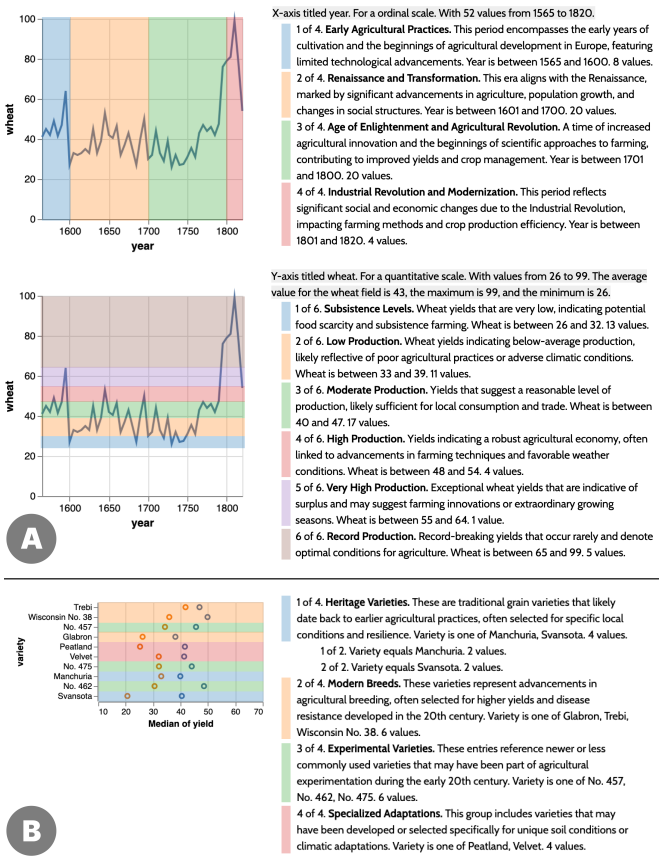

Figure 2. Semantic binning using a example wheat and barley datasets. A) the `year` field (temporal) is binned into historical periods, and `wheat` yield (quantitative) is binned into levels of low to high production. B) barley `variety` (nominal) is grouped into higher-level variety types.

represents a semantically-meaningful subset of cars in the example car dataset, with a predicate that involves two or more of the fields `Horsepower`, `Miles_per_Gallon`, and `Origin`.

Data highlights are akin to visual annotations for readers, which are a technique that designers frequently use to emphasize and draw attention certain parts of the data [18]. Indeed, data highlights have the same components as many instances of visual annotation: a defined subset of data records, and a semantically-meaningful explanation. However, we think of data highlights as independent of any specific visual representation. In our examples, we convey data highlights both as a conditional encoding in a visualization (e.g. visually annotating the included data points using color or opacity), and as descriptions in a textual structure (Figure 1).

Data highlights are designed to help lay readers understand a dataset even if they lack prior knowledge about the data domain. For example, in the cars dataset from Figure 1, a reader might observe the inverse relationship between `Horsepower` and `Miles_per_Gallon` but lack the context to know why there might be a tradeoff between the two, or how cultural factors between the three `Origin` values contribute to differences. In Figure 1, each data highlight focuses on a different region of the chart, and connects the selected data points with a real-world explanation. For example, the third data highlight in the figure is called "Low Horsepower European Cars," and explains the tendency for European cars to have lower horsepower in terms of economy, practicality, and urban com-

muting. The highlight includes a query predicate defining the grouping's inclusion criteria in terms of the `Horsepower` and `Origin` fields, allowing the reader to connect the explanation to concrete values in the dataset.

3.3 Limitations

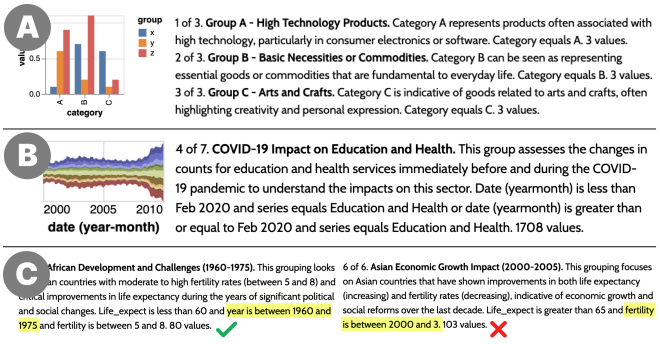

Figure 3. Examples of LLM errors encountered in our design process. A) LLM hallucinates dataset semantics when there is not enough information. B) LLM includes information that is outside of the dataset's context. C) LLM generates incorrect query predicate.

Because semantic scaffolding is a technique that relies on LLMs, it is limited by the kinds of errors LLMs tend to make. Here, we provide examples of errors we observed in our prototyping process.

Hallucinated dataset semantics. When a dataset does not have meaningful field labels, the LLM might hallucinate a domain for the data. In Figure 3A, the data has the generic field labels of `category` and `value`. Even though there's no information about what these fields mean, the LLM interprets them as consumer products.

Social context inappropriate to dataset. The LLM may incorporate domain-specific information that is not necessarily factually wrong, but is outside of the scope of the dataset. Figure 3B shows a data highlight referencing the COVID-19 pandemic, which started in 2020, for an unemployment dataset that only covers 2000–2010.

Incorrect query predicates. Figure 3C demonstrates an error where the LLM identified a reasonable grouping, but was unable to formulate a correct query predicate that matches the explanation. The figure shows one correct predicate and one incorrect predicate generated from the same data. While the `fertility` between `2000` and `3` term is syntactically valid, it is semantically incorrect.

Discussion of limitations. In our prototyping process, we explored recent techniques for LLM output validation, such as SymGen [8]. However, we found that most data-to-text validation techniques were not appropriate for our use case, because we are not merely asking the LLM to summarize a dataset. In our case, our goal is to leverage the LLM's knowledge that is external to the dataset. As a result, validating this information requires either access to an external knowledge base, or use of a proxy like the model's internal confidence [26]. Therefore, because this is still an active area of research, we were not able to incorporate reliable technical approaches to validating the output of semantic scaffolding at this time. However, in our user testing, we surfaced insights on user behaviors and strategies for validating and thinking critically about LLM output (section 5).

4 Implementation

We implemented semantic scaffolding by prompting `gpt-4o-mini` with instructions to create and return data groupings, attaching a full dataset to the prompt. To ensure that we receive a structured response from the LLM, we defined a data structure that specifies our expected response format. Each grouping identified by the LLM is returned as a data structure with three properties: a name that summarizes the content of the grouping; a longer explanation of the grouping's meaning; and a query predicate that defines the criteria for a data point to be included in the grouping. For compatibility with existing visualization tools, the query predicate is specified in Vega-Lite's existing predicate syntax [19]. Figure 4 contains a definition of the data structure in TypeScript syntax.

```
type LLMResponse = {
  groups: SemanticScaffold[];
};

type SemanticScaffold = {
  name: string;
  explanation: string;
  predicate:
  ↪ LogicalComposition<FieldPredicate>;
};
```

```
{
  "name": "AAPL Price Surge During the Tech
  ↪ Boom",
  "explanation": "This group focuses on
  ↪ Apple's stock price during the late
  ↪ 2000s and early 2010s, highlighting
  ↪ its rise in value matched with the
  ↪ smartphone revolution and innovation
  ↪ of products like the iPhone.",
  "predicate": {
    "and": [
      { "field": "symbol", "equal": "AAPL"
      ↪ },
      { "field": "price", "gte": 150 },
      { "field": "date", "range":
        [ "2008-08-31", "2012-12-31" ] }
    ]
  }
}
```

Figure 4. Type definitions and example for semantic scaffolding

5 Evaluation: User Study

To evaluate semantic scaffolding, we conducted a user study with 15 blind and low-vision (BLV) participants who use screen readers. The goal of this evaluation was to understand how our technique affects readers' data exploration and interpretation. In this section, we describe the methods we used to design our evaluation, and report survey and interview results.

5.1 Methods

The study involved 100-minute Zoom interviews with 15 BLV participants exploring three prototype implementations of semantic scaffolding in Olli [3]. We began each interview with a brief 10 minute

introduction to learn more about the user's interactions with data, data visualizations, and large language models. To familiarize participants with navigation in Olli, we did a 15 minute introduction of the existing baseline version of Olli. Then, we had users explore the *data-first* and *highlights-first* prototypes (described below), each for 20 minutes. For each prototype, we guided participants through the prototypes and asked participants to explore the data and comment on their thought process. Then, we had users complete the prototype's corresponding Likert survey. In the final 15 minutes, we discussed users' overall takeaways and conducted a post-study survey.

Prototypes. In our study, we used the default version of Olli as a baseline, and compared it to two prototypes with semantic scaffolding. These two prototypes explored different ways to present semantic scaffolds in a screen reader interface. One prioritizes showing additional domain-specific information first, while the other prioritizes direct access to the data first. We created these two variations because we were interested in whether readers would prefer to have highlights first to orient themselves, or whether they would rather explore the data directly before zooming back out for additional context. For us, this question was connected to the idea of agency — for instance, would data highlights make a reader feel like they were being told what to think rather than drawing their own conclusions? For both prototypes, we presented a scatterplot but with a different dataset each time so we could understand how a prototype helped them make sense of new datasets. We chose scatterplots because they are typically considered difficult to read for screen reader users, and because sighted readers tend to identify groupings in scatterplots using visual features like clustering. For the user study, we pre-generated and hand-curated LLM responses to ensure that they made sense, were the same for each participant, and would load quickly.

- **Olli baseline**. We use the existing version of Olli as a baseline and to familiarize participants with textual structures. This baseline uses an example movies dataset with fields `US Gross`, `IMDB Rating`, and `MPAA Rating`.
- **Data-first prototype**. This prototype prioritizes giving readers a chance to explore the data before adding additional context. It conserves the original Olli structure as much as possible, and appends highlights to the end of the encodings level of the Olli tree view. This prototype uses an example penguins dataset with fields `Flipper Length (mm)`, `Body mass (g)`, and `Species`.
- **Highlights-first prototype**. This prototype prioritizes giving an overview with data highlights first, before opening up further exploration in the data. Data highlights are presented as a subtree of annotations, prepended to the encodings level of the Olli tree view. This prototype uses an example cars dataset with fields `Horsepower`, `Miles_per_Gallon`, and `Cylinders`.

Participants. We recruited 15 blind and low-vision participants by sending a participant call to a blind programmers' community mailing list and reaching out to a local BLV community group. Each participant was compensated $60 for 100 minutes. To protect participant privacy, we've included anonymized and aggregated demographic information to provide background context. The majority (67%) of our participants were totally blind (n=10), while 27% identified as low-vision with some light perception (n=4) and 7% of participants did not respond (n=1). More than half (60%) of our participants have been blind since birth. 53% of participants were JAWS users (n=8) and 47% were NVDA users (n=7), relatively consistent with screen reader statistics [2]. Demographically, 80% of participants use he/him pronouns (n=12) and 20% of participants use she/her pronouns (n=3). Participants were based across multiple continents, including North America, Europe, and Asia. Participants self-reported their ethnicities (Asian, Black/African, Caucasian/white, Hispanic/Latinx, Other), represented a diverse range of ages (20–50+), and and had a variety of educational backgrounds (high school through to undergraduate and graduate school). With the exception of one participants that did not respond, almost all participants (n=14) self-reported as slightly, somewhat, or moderately familiar with statistical concepts. 13 participants self-reported as slightly, somewhat, or moderately familiar with data visualization methods and 2 participants did not respond. Participants reported a high variety of frequency interacting with data or visualizations, from 1-2 times/year to 3 or more times/week. Most participants (n=9) reported using data analysis tools or visualizations either outside of their professional work or sometimes, but 2 participants reported data analysis tools or visualizations being an important part of their workflow. 4 participants reported rarely use data analysis tools.

5.2 Quantitative Results

We designed a Likert survey to understand participants' preferences across the highlight-first and data-first prototypes. Participants responded on a five point scale where 1 = Very Difficult to Understand / No Influence / No Additional Context / Not Effective / Never Felt the Need and 5 = Very Easy to Understand / Significant Influence / Extensive Additional Context / Extremely Effective / Always Felt the Need (Table 1).

The highlights-first prototype consistently scored higher in terms of understandability for both highlights and bins, indicating that users found this prototype more intuitive and easier to use. Despite this, the influence of highlights and bins on data interpretation was similar between the two prototypes, suggesting that while the highlights-first prototype was easier to understand, it did not significantly change how users interpreted the data compared to the data-first prototype. However, participants felt the highlights-first prototype was more effective in explaining and justifying its data highlights and bins, as reflected by higher scores in that category. Users also felt a slightly greater need to double-check the accuracy of data highlights and bins in the highlights-first prototype, though this difference was minimal. These results suggest that participants generally preferred the highlights-first prototype, which offers better usability and clarity. However, these findings are best contextualized with our qualitative findings in the next section, especially around participants' trust in AI and desire for agency in interpretation.

5.3 Qualitative Results

Several themes emerged during our observation of participant interaction with the prototypes, and in the interviews that followed. To analyze our qualitative data, authors Zong, Pedraza Pineros, and Hajas independently conducted qualitative coding using interview transcripts, grouping participant quotes into thematic categories. Then, authors synthesized these themes into the findings reported in this section.

5.3.1 Semantic scaffolding helps readers understand and contextualize data

When encountering a new dataset, readers need to understand the data's real-world meaning first. P1 illustrated this point nicely, saying "the numbers don't mean much until I know a little bit about what

Table 1. Rating scores for each prototype (Data-first Prototype, Highlights-first Prototype) on a five point Likert scale. Median scores are shown in boldface, averages in brackets, standard deviations in parentheses.

Prompt: After understanding how the [prototype] works...	Data-first	Highlights-first
How understandable were the data highlights in the prototype?	**4** [4.13] (0.64)	**5** [4.53] (0.64)
How understandable were the bins in the prototype?	**4** [4.00] (1.20)	**5** [4.67] (0.49)
How much influence did the data highlights have on your interpretation of the data?	**4** [3.60] (1.30)	**4** [3.33] (1.45)
How much influence did the bins have on your interpretation of the data?	**3** [3.13] (1.25)	**3** [3.00] (1.36)
How much additional context did the data highlights provide about the data?	**4** [3.93] (1.10)	**4** [3.60] (1.24)
How much additional context did the bins provide about the data?	**4** [3.47] (0.92)	**4** [3.47] (0.83)
How effectively did the prototype explain and justify its data highlights and bins?	**3** [2.93] (0.96)	**4** [3.73] (0.88)
How often did you feel the need to double-check the accuracy of the data highlights and bins?	**2** [2.33] (1.18)	**2** [2.47] (1.06)

I'm reading about." However, accurately understanding the dataset's semantics can be difficult for readers. We found that participants were able to quickly identify what the data refers to using semantic scaffolds even if their initial understanding was incorrect. For instance, participants had many different initial guesses about the prototypes' data semantics. Based on the `Flipper Length (mm)` and `Body Mass (g)` fields, some participants initially thought the penguins dataset referred to fish (P1), seals (P2), or dolphins (P3). However, all these participants correctly updated their understanding to reflect information about penguins found in the data highlights. Accelerating this process of identifying data semantics can be very important to a reader's experience: as one participant put it, "I don't feel dumb looking at this data" (P4).

One participant contrasted semantic scaffolding positively with prior AI-based tools they had used. P9, who had used chart question-answering tools in the past, noted that "sometimes you don't know what questions to ask" when you are not familiar with the data. Further, they said "sometimes its unfair to expect you to ask the right questions because you don't have a clue of [...] what the data is about." Semantic scaffolding can address this "cold start" problem when participants start with limited prior knowledge.

5.3.2 *Reshaping textual structures using semantic scaffolding helps with navigation and wayfinding*

Participants found semantically meaningful groupings helpful for navigating through the data. For example, P5 said that "the groupings keep it in an organized fashion so that I get a more orderly presentation" of the data. Others similarly noted that groupings help them not "wander uselessly around," (P6) and not "have to look for things random" (P4) and instead explore within a relevant space. These comments echo previous findings that screen reader users benefit from interfaces that provide a *bounded space* [27] that helps a user locate themselves within a smaller space with known boundaries.

Other participants also found semantic scaffolds helpful for identifying where they could navigate to find more relevant information. For instance, P8 said that semantic bins helped "frame my expectations in advance to have some information of what the legend is." For them, the description of the semantic bin was providing *information scent* [17] about what could be found when navigating further into that grouping. This finding reinforces the value of nodes that compare across multiple data entries or fields, representing meaningful subsets of data, and do not require manual specification of filters by a user.

5.3.3 *Readers critically appraise semantic scaffolds using their prior knowledge*

Semantic scaffolds prompted participants to think about data in semantically meaningful ways, incorporating their own prior knowledge. For example, engaging with LLM-generated categorizations of data prompted participants to imagine other ways of thinking about the data's meaning, and other possible categorizations. When reading the categorization of cars in the example dataset, P2 noted that "one of the things that's missing is work vehicles, or pickup trucks." In other words, there was a category that they expected to see that was not present. Similarly, P8 noted that "car manufacturers don't [categorize cars with] engine sizes, they just base it on fuel economy," and suggested "utility" as a potential additional category. In this case, the participant had additional context or expertise that made the LLM's categorization insufficient for them, even if it was not technically incorrect.

Sometimes, participants did not necessarily disagree with the content of semantic groups, but were skeptical of arbitrariness and asked for justification. For example, P1 questioned the cutoffs for semantic bins, saying, "I can hear that it's grouping cars by cylinders and horsepower, but why are these the breaks? Why 101 to 200, or why eight cylinders? It seems arbitrary." Similarly, P11 questioned the subjectivity of certain categories: "I don't know how arbitrary the cutoff is. Like, what would you say a gas guzzler is? That's maybe arbitrary, maybe not. I don't know." Even when participants lacked domain knowledge to fully evaluate explanations, they were still aware of potential limitations in the groupings.

A person's cultural context can also be a factor, as we observed when P6 pointed out that "the values don't correspond to my knowledge because I use different units." Because the data used miles and gallons, which are units that not every country uses, the explanation was less useful for them.

However, people appreciated when the descriptions matched their existing knowledge, whatever level of knowledge that may be. Many applied the standard of "it sounds plausible" when evaluating groupings (P3, P9). For participants who did have more prior knowledge, they felt like they were still able to learn something new. For instance, P8 said of a semantic binning that they "know a little about

cars," but "without it telling me I would have no idea that this range is economy / mid range." They felt that they were provided with useful information they still would not have known otherwise.

5.3.4 *Trust in AI-generated descriptions is situational*

Participants had a range of attitudes about their trust in AI-generated information. Some were fairly confident in the information they were reading in the semantic scaffolds, as long as it seemed plausible. For example, participants remarked, "I didn't hear anything outlandish" (P3), and "I assume that the domain experts would agree with those descriptions" (P11). Even participants that were fairly aware of LLM hallucinations said that, even if they might double-check later, they would generally take the descriptions at face value (P1, P3, P5). For example, P3 said, "I might go to Google and re-read on these things but at least I will take the first impressions here."

On the other hand, some participants expressed more concerns about the risks of incorrect descriptions. Others said that "hallucination is very dangerous" (P6), noting that LLMs "sometimes give you really really false data with high confidence" (P13). These participants felt they were more likely to verify information with external sources. A consistent factor in people's willingness to double-check output is how important or high-stakes their data analysis felt. Many said that they would be more likely to check against other sources if they were looking at data for professional purposes (P1), making a presentation (P2), or if it was otherwise really important (P6, P8).

Participants employed a variety of strategies to manage trust and verification of LLM-generated information. They mentioned using Google or their own knowledge (P3), finding another non-automated source (P5), or refreshing the LLM output to see if it remained consistent (P1). In our study, for example, participants used the data table to compare against the description of the grouping's query predicate, to check if it "matches the description" (P11). Overall, many participants shared a pragmatic attitude. For instance, P6 said,"you have to take what is good from it, but with a grain of salt."

5.3.5 *AI-generated descriptions might risk reducing readers' agency to interpret the data*

Participants' preferences between the highlights-first and data-first prototypes often boiled down to how much they wanted to rely on semantic scaffolding to understand the data. Participants generally preferred the highlights-first prototype, echoing previous findings about the importance of an "overview first, then detail" strategy in data exploration [22, 27]. For example, P3 preferred the highlights first because "it gives an overall picture / what to expect before I dig into the data." However, P14 dissented, saying that "I'm skeptical of people interpreting my data for me, [so] I like the facts first."

These differing preferences were emblematic of different perspectives on how the AI-generated descriptions were influencing reader interpretation. The more common view was that semantic scaffolding has a negative influence on reader agency (P2, P6, P10, P14). For instance, P2 said, "I feel like my conclusions are being shaped. If someone first gives you the conclusion, it kind of shapes your perception. If you are first given the raw data, you try to make sense for yourself and after that you can compare it to the conclusions provided." Similarly, P14 drew a distinction between interpretative and factual content in descriptions, saying, "I want the AI to separate interpretation from facts. 'Body mass' is fact, 'healthy body mass' is getting into interpretation." In this instance, the participant felt a semantic bin titled "healthy body mass" and associated with a specific range was imposing a subjective interpretation on the data, rather than letting the reader decide what range is considered healthy. These concerns are consistent with prior findings that unlike sighted readers, BLV readers find interpretive content less useful in descriptions of data [13]. P6 felt that they would get "spoiled" by the tool, saying "I would lose the tendency of exploring and rely on the highlights too much." This concern is consistent with prior findings too, that users sometimes become over-reliant on LLM agents in data analysis, leading to less overall engagement with data [5].

However, there were also a number of participants who felt that semantic scaffolding had a positive influence on their interpretation process (P5, P11). For instance, P11 said of data highlights, "It's giving me ideas about what questions to ask or what relationships are worth looking at." Similarly, P5 said that semantic scaffolds "[give] me a sense of confidence now that I have access to information that I didn't have before. This information is presented in a means that I can obtain it, digest it and work with it in my own way." In this case, the participants felt that data highlights were getting them started on asking their own followup questions about the data — rather than foreclosing further interpretation.

5.3.6 *Readers expected semantic scaffolds to augment the original data*

Multiple participants expected semantic groupings to be reflected in the dataset itself, and were confused when that did not happen. For example, P10 said, "You gave me this description here of, you know, sports cars, muscle cars, sedans." When they then drilled down into the data table view, they were "a little bit frustrated that you don't specify which one is a sedan and which one is a muscle car." They then elaborated that they expected the AI to "add that information as an extra column to the table" (P10). This suggests that participants did not necessarily distinguish semantic scaffolds from the data itself, because they both describe the same real-world domain. Future work could potentially explore semantic scaffolding in conjunction with data augmentation [4] to synchronize these two views, or otherwise manage this mismatch in expectations.

5.3.7 *BLV readers also want sighted readers to have semantic scaffolding, to preserve common ground*

For many participants, the use of semantic scaffolding made them reflect on situations where they are collaborating with sighted people. For example, participants noted that semantic scaffolding provides "maybe more than a sighted person would get from such a graphic because it gets you info about the relevance of the values with knowledge you don't have" (P6) and that this "feels weird when working with sighted people" (P8). Participants generally felt that it was important to maintain parity of information access with sighted people. For example, P10 asked, "does the visual person get the same info?" Prior work has emphasized the importance of common ground between mixed-ability collaborators [28], and participants' concerns reflected the idea that it was important to them to participate in and lead conversations about data even in spaces that have a majority of sighted people. P14 explained, "whenever you can put sighted and blind individuals on the same playing field with equal footing, you've accomplished something that's really marvelous. [...] For us to be able to be literally on the same page, it's helpful because we're looking at the very same information." For P14, accessibility is a two-way street: "[blind people] have to accommodate [visual people] in a way." It was important to them to include sighted collaborators

in their data analysis process, even when using assistive tools like semantic scaffolding. These discussions led us to reflect on the distinction in our work between *disability inclusion for BLV people* and *making data understandable for lay readers*, two ideas of accessibility that can be complementary and mutually-reinforcing.

6 Conclusion

In this paper, we introduced semantic scaffolding: a technique for using an LLM to generate meaningful groupings of data to aid lay readers in understanding unfamiliar datasets. With semantic scaffolding, readers benefit from domain-specific knowledge that is incorporated into descriptions and navigational structures. We instantiated this approach in Olli, and observed that users find semantic scaffolds valuable for forming an initial understanding of a dataset while applying their prior knowledge to evaluating the trustworthiness of descriptions. Our work suggests a use for LLMs to augment user interfaces with domain-specific affordances for interpretation.

ACKNOWLEDGEMENTS

This work was supported by NSF #2341748, MIT-SERC, and MIT-Google.

REFERENCES

[1] Type | Vega-Lite, 2023.
[2] Screen Reader User Survey #10 Results, 2024.
[3] Matthew Blanco, Jonathan Zong, and Arvind Satyanarayan, 'Olli: An Extensible Visualization Library for Screen Reader Accessibility', in *IEEE VIS Posters*, (2022).
[4] Dylan Cashman, Shenyu Xu, Subhajit Das, Florian Heimerl, Cong Liu, Shah Rukh Humayoun, Michael Gleicher, Alex Endert, and Remco Chang, 'CAVA: A Visual Analytics System for Exploratory Columnar Data Augmentation Using Knowledge Graphs', *IEEE Transactions on Visualization and Computer Graphics*, **27**(2), 1731–1741, (February 2021).
[5] Kiroong Choe, Chaerin Lee, Soohyun Lee, Jiwon Song, Aeri Cho, Nam Wook Kim, and Jinwook Seo, 'Enhancing Data Literacy On-demand: LLMs as Guides for Novices in Chart Interpretation', *IEEE Transactions on Visualization and Computer Graphics*, 1–17, (2024).
[6] Frank Elavsky, Lucas Nadolskis, and Dominik Moritz, 'Data Navigator: An Accessibility-Centered Data Navigation Toolkit', *IEEE Transactions on Visualization and Computer Graphics*, 1–11, (2023).
[7] Joshua Gorniak, Jacob Ottiger, Donglai Wei, and Nam Wook Kim, 'VizAbility: Multimodal Accessible Data Visualization with Keyboard Navigation and Conversational Interaction', in *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23 Adjunct, pp. 1–3, New York, NY, USA, (October 2023). Association for Computing Machinery.
[8] Lucas Torroba Hennigen, Zejiang Shen, Aniruddha Nrusimha, Bernhard Gapp, David Sontag, and Yoon Kim, 'Towards Verifiable Text Generation with Symbolic References', in *First Conference on Language Modeling*, (2024).
[9] Jessica Hullman and Nick Diakopoulos, 'Visualization Rhetoric: Framing Effects in Narrative Visualization', *IEEE Transactions on Visualization and Computer Graphics*, **17**(12), 2231–2240, (December 2011).
[10] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung, 'Survey of Hallucination in Natural Language Generation', *ACM Computing Surveys*, **55**(12), 1–38, (December 2023).
[11] Ecem Kavaz, Anna Puig, and Inmaculada Rodríguez, 'Chatbot-Based Natural Language Interfaces for Data Visualisation: A Scoping Review', *Applied Sciences*, **13**(12), 7025, (January 2023). Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.
[12] Jiho Kim, Arjun Srinivasan, Nam Wook Kim, and Yea-Seul Kim, 'Exploring Chart Question Answering for Blind and Low Vision Users', in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pp. 1–15, New York, NY, USA, (April 2023). Association for Computing Machinery.
[13] Alan Lundgard and Arvind Satyanarayan, 'Accessible Visualization via Natural Language Descriptions: A Four-Level Model of Semantic Content', *IEEE Transactions on Visualization and Computer Graphics*, **28**(1), 1073–1083, (January 2022).
[14] Tamara Munzner, *Visualization Analysis and Design*, A K Peters/CRC Press, New York, October 2014.
[15] Jason Obeid and Enamul Hoque, 'Chart-to-Text: Generating Natural Language Descriptions for Charts by Adapting the Transformer Model', in *Proceedings of the 13th International Conference on Natural Language Generation*, eds., Brian Davis, Yvette Graham, John Kelleher, and Yaji Sripada, pp. 138–147, Dublin, Ireland, (December 2020). Association for Computational Linguistics.
[16] Evan M. Peck, Sofia E. Ayuso, and Omar El-Etr, 'Data is Personal: Attitudes and Perceptions of Data Visualization in Rural Pennsylvania', in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, Glasgow Scotland Uk, (May 2019). ACM.
[17] Peter Pirolli and Stuart Card, 'Information foraging', *Psychological Review*, **106**(4), 643–675, (1999). Place: US Publisher: American Psychological Association.
[18] Donghao Ren, Matthew Brehmer, Bongshin Lee, Tobias Hollerer, and Eun Kyoung Choe, 'ChartAccent: Annotation for data-driven storytelling', in *2017 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 230–239, Seoul, South Korea, (April 2017). IEEE.
[19] Arvind Satyanarayan, Dominik Moritz, Kanit Wongsuphasawat, and Jeffrey Heer, 'Vega-Lite: A Grammar of Interactive Graphics', *IEEE Transactions on Visualization and Computer Graphics*, **23**(1), 341–350, (January 2017).
[20] Edward Segel and Jeffrey Heer, 'Narrative Visualization: Telling Stories with Data', *IEEE Transactions on Visualization and Computer Graphics*, **16**(6), 1139–1148, (November 2010).
[21] Leixian Shen, Enya Shen, Yuyu Luo, Xiaocong Yang, Xuming Hu, Xiongshuai Zhang, Zhiwei Tai, and Jianmin Wang, 'Towards Natural Language Interfaces for Data Visualization: A Survey', *IEEE Transactions on Visualization and Computer Graphics*, **29**(6), 3121–3144, (June 2023).
[22] Ben Shneiderman, 'The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations', in *The Craft of Information Visualization*, eds., Benjamin B. Bederson and Ben Shneiderman, Interactive Technologies, 364–371, Morgan Kaufmann, San Francisco, (2003).
[23] Charles D Stolper, Bongshin Lee, Nathalie Henry Riche, and John Stasko, 'Emerging and Recurring Data-Driven Storytelling Techniques: Analysis of a Curated Collection of Recent Stories', Technical Report MSR-TR-2016-14, Microsoft Research, (2016).
[24] Nicole Sultanum and Arjun Srinivasan, 'DataTales: Investigating the use of Large Language Models for Authoring Data-Driven Articles', pp. 231–235. IEEE Computer Society, (October 2023).
[25] John R Thompson, Jesse J Martinez, Alper Sarikaya, Edward Cutrell, and Bongshin Lee, 'Chart Reader: Accessible Visualization Experiences Designed with Screen Reader Users', in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–18, Hamburg Germany, (April 2023). ACM.
[26] Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D. Manning, and Chelsea Finn, 'Fine-Tuning Language Models for Factuality', in *The Twelfth International Conference on Learning Representations*, (2024).
[27] Jonathan Zong, Crystal Lee, Alan Lundgard, JiWoong Jang, Daniel Hajas, and Arvind Satyanarayan, 'Rich Screen Reader Experiences for Accessible Data Visualization', *Computer Graphics Forum*, (2022).
[28] Jonathan Zong, Isabella Pedraza Pineros, Mengzhu (Katie) Chen, Daniel Hajas, and Arvind Satyanarayan, 'Umwelt: Accessible Structured Editing of Multi-Modal Data Representations', in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, pp. 1–20, New York, NY, USA, (May 2024). Association for Computing Machinery.