Check for updates

DOI: 10.1111/cgf.70102
Eurographics Conference on Visualization (EuroVis) 2025
W. Aigner, N. Andrienko, and B. Wang
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 44 (2025), Number 3

Accessible Text Descriptions for UpSet Plots

A. McNutt[1], M. K. McCracken[1], I. J. Eliza[1]†, D. Hajas[2], J. Wagoner[1], N. Lanza[1], J. Wilburn[1], S. Creem-Regehr[1], and A. Lex[1]

[1]University of Utah, Salt Lake City, Utah
[2]Global Disability Innovation Hub, University College London

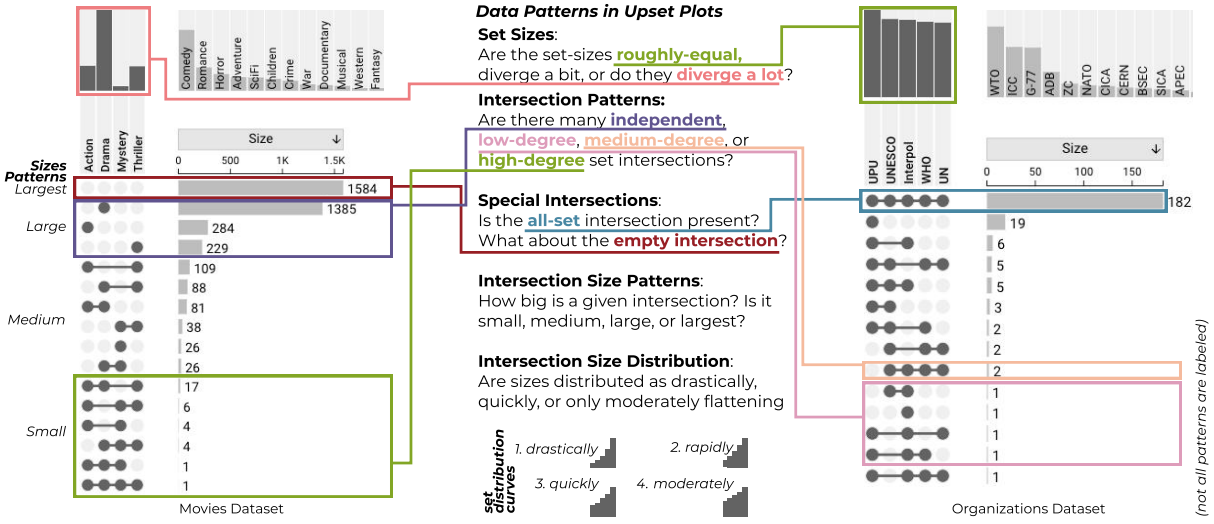


Figure 1: *UpSet plots are a prominent technique for visualizing set data. Despite their popularity, they remain inaccessible to blind or low vision (BLV) users. We seek to make this chart type accessible to all by developing a descriptive text-generation system. Here we show a subset of the descriptive data patterns that appears in their usage that enable our text descriptions.*

Abstract

Data visualizations are typically not accessible to blind and low-vision (BLV) users. Automatically generating text descriptions offers an enticing mechanism for democratizing access to the information held in complex scientific charts, yet appropriate procedures for generating those texts remain elusive. Pursuing this issue, we study a single complex chart form: UpSet plots. UpSet Plots are a common way to analyze set data, an area largely unexplored by prior accessibility literature. By analyzing the patterns present in real-world examples, we develop a system for automatically captioning any UpSet plot. We evaluated the utility of our captions via semi-structured interviews with (N=11) BLV users and found that BLV users find them informative. In extensions, we find that sighted users can use our texts similarly to UpSet plots and that they are better than naive LLM usage.

CCS Concepts

• ***Human-centered computing*** → ***Accessibility systems and tools; Visualization;***

1. Introduction

Being able to read and understand the content of information presented in data visualizations is crucial to fully grasp the content of scientific articles. Yet, most visualizations are not accessible to blind or low-vision users (BLV) [BHMG17, SL21]. A common way to make images accessible is to provide text descriptions [JMK*22] that communicate the content of a chart to screen readers [ZLL*22]. However, most scientific journals do not provide useful text descriptions for visualizations. This lack of accessibility likely contributes to an under-representation of disabled individuals in STEM [Nat23].

† The first three authors contributed equally

© 2025 The Author(s). Computer Graphics Forum published by Eurographics - The European Association for Computer Graphics and John Wiley & Sons Ltd.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

The reasons for the lack of text descriptions in scientific publications are manifold—ranging from a lack of awareness, to technical challenges when providing text descriptions in scientific papers, to adding yet another burden to the onerous publishing process, and lack of knowledge on how to write useful text descriptions. Automatic construction of text descriptions would seem then to offer a partial solution: reducing the burden of writing text descriptions and potentially forming a foundation for a human to refine.

Whereas pixel-based images are commonly analyzed and captioned using computer vision techniques (e.g., on Facebook [WWFS17]), many chart forms are computationally represented and thus amenable to automated analysis and thus description. Tang et al. [TBS23] highlight that effective text descriptions should be specific to the chart and dataset type, yet, automatic approaches for data and chart aware charts are currently limited to simple charts, like one-dimensional line and bar charts. The construction of text for more bespoke or uncommon chart types has been identified as important next steps, although efforts so far have been limited [KKK23, LPS*24].

Complicating these explorations is that usage and data patterns of complex charts are not as well as understood as those present in quotidian charts forms. For instance, widely used plots, such as scatterplots, exhibit well-understood patterns, such as clusters or outliers, that algorithms can identify to extract information and potentially generate meaningful text descriptions. Yet such patterns do not exist for exotic charts (such as necklace maps [SV10]), boutique charts (as in My Life with Long Covid [Lui23]), or even relatively well-known charts like tree maps [KKK23].

We seek to improve the process of automatic descriptive text creation, by considering the specific case of UpSet plots [LGS*14]. UpSet plots offer an intriguing test case for text description-generation, as there has been substantial adoption (particularly in the biomedical domain with over 4000 citations to the two papers that introduce them [LGS*14, CLG17]) making them a graphical form with potentially high impact if made more broadly accessible. At the same time they consider a relatively uncommon data type; compared to the ubiquity of tabular data, set data is less frequently used—making them novel compared to prior designs. They are hence an ideal exemplar of complex or scientific charts; offering fertile ground on which to explore alternative text design.

To explore text descriptions for UpSet plots, we developed a system that automatically generates descriptive texts. To do so, we first identified common patterns in UpSet plots via a survey and qualitative analysis of published UpSet figures (Sec. 3.1). Based on the results of these patterns, we conducted a design process for effective text descriptions on UpSet, consistently collecting feedback from our blind coauthor, and drawing on insights from prior work [LS22]. We then implemented text generation that automatically detects the patterns we describe and generates the description. We leverage a JSON-based description of the UpSet plot and the original dataset in the process. We implement our tool as a web-API, so that all kinds of implementations of UpSet plots can leverage our text generation, and demonstrate the process with two instances: an interactive, web-based version of UpSet and a Python library. We design our texts generically, both to be generally useful and so that they can be manually adapted to specific situations.

To evaluate the quality of our results, we conducted a semi-structured interview study with (N=11) BLV participants (Sec. 4). We find that the text descriptions gave participants a "decent sense as to like what [UpSet Plots] might look like" and that they were informative and appropriate in length. We also received suggestions to improve the descriptions, for example, by adding a glossary and by structuring the text with bullet points. More essentially, we find that participants were generally able to conduct at least simple analytic tasks using text descriptions that would be normally carried out using the visual representations.

We complement these findings with two contextualizing experiments, Sec. 5. First, we explored whether our descriptive texts contained the same level of data analytic functionality as visual UpSets. To do so, we conducted a (N=83) crowd work study with sighted users. Although this does not show that our texts are equivalent to their visual representations, we do find that sighted users perform similarly well with text and visual versions in answering a series of factual questions. We also find that preference and correctness are highest for the combined text and visual condition, indicating that providing text descriptions for complex charts could have a "curb-cut" effect, benefiting both BLV and sighted users. This also constitutes a minor contribution: this is the first user study of UpSet plots, highlighting that they can be understood with little training. Then, we compared our texts to a naive usage of LLMs. We find (via qualitative coding and expert reading from our blind coauthor) that they can produce usable texts, however the quality is variable and dependent on factors like presence of LLM training data relevant to the dataset.

By making it simpler to make scientific charts accessible we strive to bring about a world in which scientific communication can be understood and used by all. Improving UpSet plot accessibility is a single step toward that larger goal. Our materials (including our descriptive text generator) are available at osf.io/kbvs9. A live version of our implementation is available at upset.multinet.app.

2. Background and Related Work

Our work builds upon prior research on accessibility in visualization and is considered in the context of UpSet plots.

2.1. Accessibility in Visualization

There is a growing interest in making visualizations accessible [KJRK21] using a wide range of approaches including sonification, physicalization, and alternative text.

Alternative or descriptive texts (sometimes called alt text) are a means to offer a nonvisual description of an image. These texts are typically accessed by an assistive technology called a screen reader (of which JAWS and VoiceOver are common examples) that vocalizes the descriptive text. This enables BLV users to understand what is being shown in purely visual media.

Despite the promise of these affordances, however, the lack of descriptive text for images and other web-based media is prominent. For instance, L'Yi [LSLG24] find that data portals and journal websites typically do not have appropriate descriptive text. Fortunately, there has been a steady sequence of work on improving this

© 2025 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

14678659, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cgf.70102 by Ishrat Jahan Eliza - University Of Utah J. Willard Marriott Library , Wiley Online Library on [12/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

situation [WWFS17, GPM*20, CJP*19]. For instance, Bigham et al. [BKL*06] develop a system (in 2006) for automatically creating alt texts. Morris et al. [MJBC18] describe a design space of ways in which assistive technology could be improved to help users of screen readers. In the process of generating meaningful alternative texts, understanding user preferences is important. Mack et al. [MCLM21] show that users like summaries or content overviews in addition to more contextual descriptions. Our work draws on these prior streams of descriptive text automation, but in the more specific context of visualization.

In visualization, text summaries can help users understand graphically presented data. While there have been a variety of different approaches, these techniques usually center on descriptions of specific chart types (e.g., line or bar charts)—which Jung et al. [JMK*22] highlight as an essential component of descriptive texts. For instance, Moraes et al. [MSMC14] develop a system for automatically summarizing line charts. Kim et al. [KKK23] develop a system for explaining a chart based on a schema, which allows description of new chart types like treemaps. Most closely related to our work is AltGeoViz [LPS*24], which develops descriptive text for a specific chart form (choropleth maps) outside of the common set of statistical graphics which alt text work usually focuses (as in Tang et al.'s [TBS23] dataset). We focus on a less well understood chart form (UpSet plots) and data type (sets), but strive toward the same goal of making all charts accessible.

Lundgard et al. [LS22] develop a four-rlevel semantic model of the types of content that appears in alternative texts. Level 1 covers elemental visual components, such as the chart type or title. Level 2 covers descriptive statistics, such as means, outliers, and so on. Level 3 covers more complex trends identification or patterns. Level 4 covers components related to the semantics of the data being represented, such as related to the domain or contextualizing events. We make use of this model in the design of our descriptions. Chintalapati et al. [CBW22] apply this model to analyze alt text usage in HCI publications—showing accessibility coverage to be lackluster. Effective description appears to be contextual, and no one description will cover all cases—although automation efforts make exploration of information more personal via interaction. For instance, Zong et al. [ZLL*22] develop a method for traversing high- and low-level trends represented and visualizations with screen readers (later reified as Olli [BZ]).

Attempts have been made to use AI to automatically generate natural language summaries for charts [OH20] and to create figure captions [QKD*20, QKD*21], but these summaries are limited to common chart types, a limitation we address in this paper. Duarte et al. [DCBD24] use an LLM-based system to generate human quality alternative texts for simple statistical charts. Interestingly, Mack et al. [MCLM21] find that users starting from automated text craft lower quality descriptions than when starting from nothing, at least in the context of PowerPoint. A component of our approach is that a human could adapt to their local context if they wanted, but we do not require them to do so as our text offers a functional starting point on its own. Williams et al. [WdH*22] find that people often struggle to determine appropriate content for alternative texts—such as how to structure and compose their descriptions, particularly for figures with complex visual elements and relationships—highlighting the necessity for assistance in this area.

Outside of pure text descriptions there have been a number of different systems for chart accessibility. Wang et al. [WSK24] contribute a system that allows for nonvisual exploratory data analysis. More broadly, Chundury et al. [CPR*22] argue for additional sensory input modalities, such as touch (i.e. physicalizations) and sound (i.e. sonifications)—directions which have been broadly pursued. Lundgard et al. [LLS19] explore physicalizations of classic visualizations, while Engel et al. [EW18] consider tactile versions of simple statistical graphics (like bar charts). On the sonification side, Chart Reader [TMS*23] integrates sonification with alt text, whereas Erie [KKH24] builds a grammar of sonification (although it is not specific to accessibility). Zong et al. [ZPPC*24] continue this trend of blending accessibility-support modalities with their tool Umwelt, which combines visualization, sound, and text. Elavsky et al.'s Data Navigator [ENM24] broadens the scope by offering a generic accessibility interface for various modalities (including visualizations and physicalizations). These tools are important for making charts more accessible; however, we deal with the more specific problem of improving text for scientific chart forms.

2.2. UpSet Plots

UpSets plots are a form of set visualization that highlights the sizes of various intersections with those sets, as well as related statistical distributions. Originally developed by Lex et al. [LGS*14], UpSet plots are widely used across numerous implementations [Lex24].

Fig. 2 shows an example UpSet plot visualizing grand slam tennis winners. The *elements* in the dataset are tennis players, the *sets* are the four grand slam tournaments. The sets and intersections are shown in a matrix. Each column corresponds to a set, and bar charts on top show the size of the set. Each row corresponds to an *intersection*: the filled-in cells show which set is part of an intersection. There are two special cases of "intersections": first, some elements are just in one set (tennis players that have only won one of the four Grand Slam tournaments). Second, an UpSet plot can also contain an "empty intersection" that contains elements that are not in any of the sets. In the tennis example, this would be players that never won one of the tournaments, which are not included in this dataset, and hence there is no empty intersection present.

The size of the intersections are displayed as bar charts to the right of the matrix. The matrix can be sorted in various ways to surface trends. For instance, a common way is to sort by size of the intersections. Despite their prominence, there have been no previous user studies of UpSet plots. As an ancillary contribution, we provide experimental evidence that they can be used for simple tasks.

3. Building Text Descriptions for UpSets

We now describe our implementation of automated description generation for UpSet plots. To do so we first conducted a survey of extant UpSets to help identify the patterns necessary for effective description. Based on the result of these patterns we design our descriptions through an iterative co-design process with a blind coauthor. Finally, we synthesize these results as a system for generating automatic text descriptions.

© 2025 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

14678659, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cgf.70102 by Ishrat Jahan Eliza - University Of Utah J. Willard Marriott Library , Wiley Online Library on [12/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

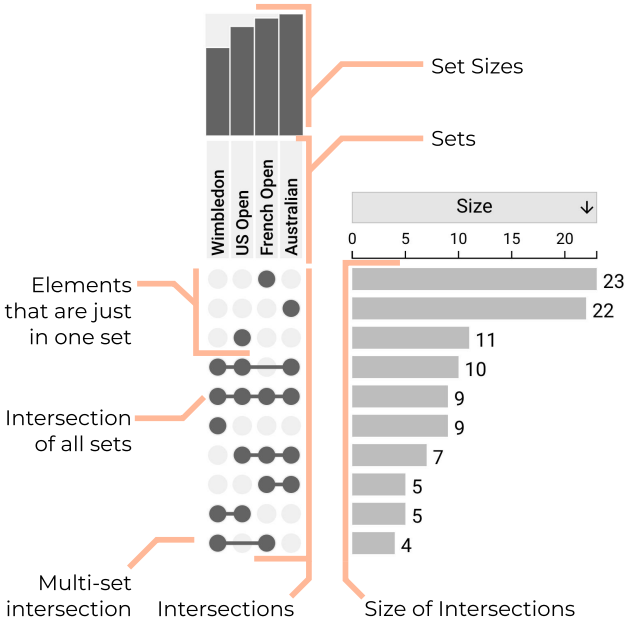


Figure 2: *An UpSet plot of the winners of major tennis tournaments, cropped to intersections with at least a size of 4. The sets are represented in the columns of the matrix; the set sizes as vertical bars; the size of the intersections as horizontal bars. Intersections are shown in the rows and identify the sets that intersect with filled in and connected circles. "Intersections" can be of only one set, here representing the players who have only won one tournament.*

3.1. Surveying UpSet Patterns

UpSet plots, like any chart type, are more than just the specific marks that they use to encode information. Instead they have patterns wedded to those marks that allow verbal or textual description. For instance, in scatterplots there is a rich language of patterns, such as outliers, clusters, bimodal distributions, a variety of scagnostics [WW08] derived terms, and so on. To form useful textual depictions of UpSet plots we need to be able to characterize the patterns present in them. To do this, we surveyed published UpSet plots and conducted a qualitative analysis of them, yielding a collection of patterns, as summarized in Fig. 1.

Seeking to understand real world usage, we gathered a collection of UpSets from published papers. The most straightforward source of such plots can be found in the papers citing the UpSet papers—Lex et al.'s [LGS*14] original description of the plot and Conway et al.'s [CLG17] UpSet R package. As of late 2024, those papers have more than 4000 citations on Google Scholar, many of which use UpSet plots for data presentation. For example, a paper on drug overdoses in the US uses UpSet plots to visualize overdoses involving multiple substances [FMB*22]. As an analysis of all papers that cite these two papers and contain UpSet plots is infeasible due to the scale. We instead drew a convenience sample that we iteratively grew until we reached saturation in our coding process, i.e., until we did not discover new types of patterns when adding additional papers to our corpus, a common method to determine when to stop adding data in qualitative research [CS14]. To this end, we collected 80 UpSet plots from 41 published papers. First, we searched for papers citing the original UpSet paper [LGS*14] in Google Scholar, resulting in 31 papers. We then searched Google for the keyword "UpSet Plot" and found another 8 papers that we included in our sample. We collected plots from 2 more papers cited in our corpus.

Using this corpus, two authors developed a collection of patterns using an iterative qualitative coding process, combining apriori codes, based on the author's knowledge of UpSet plots and emergent codes discovered during the coding process [Kin98]. First, one of them coded each plot individually based on the presence of the largest set in the largest intersection, the smallest sets in the smallest intersection (and vice versa), how the set sizes diverge, how the intersection sizes diverge, and the presence of all-set intersections (our apriori codes). They then iteratively met to revise and alter the codes based on examples they saw (the emergent codes). After two iterations, they reached a consensus and finalized the codes.

Based on this coding, we form five categories of **data patterns**, as well as a collection of **layout choices**. The characteristics of a type of data pattern can often be described along a continuous variable. For example, correlation among two dimensions in a scatterplot can be described as an R^2 value. Since we intend to generate text descriptions based on these patterns, we discretize continuous variables to ordinal labels. For the R^2 example, a value of 0.9 could be mapped to "highly correlated", whereas a value of 0.4 could be considered "moderately correlated". As a reminder, the data objects of interest in UpSets are sets and set intersections. We use UpSet plots for the Movies and Organization datasets (Fig. 1) as examples. See Table 1 for the specific heuristics for each pattern.

Set Size Patterns. Consider set size is essential when analyzing set intersections. If all sets in a dataset are of equal size, intersection sizes can be compared without considering set size. However, if set sizes diverge, an analysis of intersections also has to consider the size of the sets. For example, if 2/4 are very large while the other 2/4 are small, a large overlap between the large sets might visually dominate, even if those sets are only weakly correlated. The set sizes of the Movies dataset diverge a lot, as there are about 15 times as many Drama than Mystery movies; while in the Organizations dataset, the selected organizations are of roughly equal size. Most plots in our corpus were "diverging a lot", followed by "moderately diverging" and a few that were "roughly equal". Alarmingly, 18 of the plots we collected did not show set size, which has the potential to mislead viewers.

Intersection Patterns. There are 2^n possible intersections in a set dataset, where n is the number of sets. For most set datasets, many of the possible intersections are empty. Moreover, different data types exhibit vastly different patterns with respect to which intersections exist. The Movies dataset, for example, contains many movies with just one or two genres, where intersections between only a few sets dominate. The Organizations dataset, on the other hand, has many countries that are part of the same organizations, resulting in several nonempty intersections that involve many organizations.

Special Intersection. We treat two intersection patterns as special: the "all-set intersection", and the "empty intersection" (containing the elements that are in no set but are still in the dataset). We treat these two patterns distinctly, because we always report on

© 2025 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

14678659, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cgf.70102 by Ishrat Jahan Eliza - University Of Utah J. Willard Marriott Library , Wiley Online Library on [12/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

their presence (and their size), while we rely on size attributes of other intersection patterns to determine whether to explicitly call them out. Both the Movies and the Organizations datasets contain the all-set intersection, while only the Movies dataset contains the empty intersection.

Intersection Size Patterns. Different types of intersection patterns can be associated with different intersection sizes. We report on the association of intersection patterns with respect to their size. We distinguish between small, medium, large, and largest sizes. For example, in the Movies dataset, the "empty intersection" is largest, the "independent sets" are large, and the "high-degree" intersections are small.

Intersection Size Distribution Finally, we classify the distribution of the intersection sizes based on which of several functions they best fit (when sorted by magnitude). These include drastically flattening and rapidly flattening (exponential curves of different degrees), quickly flattening (quadratic curves), or steadily flattening (linear curves). For instance, Movies is rapidly flattening, while Organizations flattens drastically. A constant curve is possible, but we saw no evidence of such a dataset.

We also coded for **layout and presentation**. Most notably, we found that UpSet plots are commonly *sorted* by the size of the intersection, although other sorting strategies (e.g., by degree or attribute) are also possible. In our corpus, most plots were sorted by intersection size, followed by degree. We also found a few instances of attribute-driven sortings (e.g., by a group first, and by intersection size second) [FMB*22], or by a measure of deviation from expected intersection size [LSZ*22]. We frequently found custom *highlights and annotations*, and supplementary plots that show attributes or Venn diagrams to explain the intersections. They were *oriented* either horizontally or vertically, but most were horizontal.

The original UpSet implementation [LGS*14] also introduces set filters, aggregations, and queries. We found only one instance of *aggregation* in our corpus. The original UpSet enables users to interactively select which sets to include in the dataset, but we found no occurrence of sets not included in the plot in any of the figures, which makes sense given that we drew our figures from published papers, and choosing sets is only useful in an exploratory stage.

3.2. Text Description Design

Next, we used these patterns, along with general properties of the UpSet plot, to design text descriptions for UpSet plots. We approach this as a design problem, as the structure of text is just as much a UI as a traditional GUI [Yal19]. At every step of the process, we elicited feedback from our blind coauthor to ensure that our text-design is useful for the target audience. We used Lundgard and Satyanarayan's [LS22] semantic levels of context to design our text descriptions because they constitute a validated framework. We designed our text so that it could act as a launching pad for humans to improve instead of a final text to use.

Design Process. Our design process involved an iterative cycle of prototyping, abstracting, and refining. We started by having one author manually write three text descriptions for representative UpSet plots. They followed a template based on Lundgard and Satyanarayan's [LS22] semantic levels of context. This diversity of plots allowed us to explore how different patterns would be reported. We then started iterating on these descriptions also with our blind coauthor. At the same time, we looked for commonalities in descriptions with an eye toward generalizing. After reaching an initial stable point, we implemented a basic version of the text generation. However, this revealed unanticipated issues with the text. For example, our blind coauthor noted that a dataset with complex set labels (e.g., `ALL_X_citri_subsp_citri_UI6_NZ_CP008992`) resulted in hard-to-follow utterances by a screen reader. We decided to (a) encourage users of the API to use human-readable set names, and (b) for cases when we had to deal with such strings to extract a short human-readable string—'citri sub' in this case. In this phase, we also investigated three text lengths. We eventually abandoned one of the sizes, preferring two based on feedback from our collaborator, as it was difficult to distinguish between the medium and highly verbose levels when using a screen reader.

Table 1: *The pattern types and the specific names in categorization of UpSet usages. In the supplement we analyze application of these patterns to a variety of datasets. ∩ = intersection.*

Type	Name	Heuristic
Set Size	Diverging a lot	Vary by $> 30\%$
	Moderately Div.	Vary by $10\% - 30\%$
	Roughly Equal	Vary by $< 10\%$
∩ Patterns	Independent sets	∩s containing only 1 set
	Low-degree ∩s	∩s containing $2 - 3$ sets
	Medium-degree ∩s	∩s containing $3 - \frac{n}{2}$ sets
	High-degree ∩s	∩s containing $\frac{n}{2}$ to n sets
Special ∩s	All Set ∩	Check for presence
	Empty Set	Check for presence
∩ Size Patterns	Small	∩ size smaller than median of intersection sizes
	Medium	Size between median and median + 1.5* IQR
	Large	Bigger than median + 1.5* IQR but not max
	Largest	Biggest intersection size
∩ Size Distribution	Drastically Flattening	∩ sizes follow an exp. curve with high β (>0.8)
	Rapidly Flattening	Sizes follow an exp. curve with low β (<0.8)
	Quickly Flattening	Sizes follow a quadratic curve
	Steadily Flattening	Sizes follow a linear curve

Designed Text. Fig. 3 shows the resulting long description next to a web-based UpSet implementation. The *Short Description* is a brief snippet that contains the most salient trend and that acts as a high-level overview of plot, akin to image alt text. The *Long Description* is an extended description (including the short description as the introduction) designed to communicate similar amounts of information as the chart. When available, we leverage human-provided labels of the sets and items to generate more natural text. For example, in the sentence "`The largest intersection is between Drama and Comedy, with 208 items.`" we replace "`items`" with "`movies`".

© 2025 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

14678659, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cgf.70102 by Ishrat Jahan Eliza - University Of Utah J. Willard Marriott Library , Wiley Online Library on [12/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

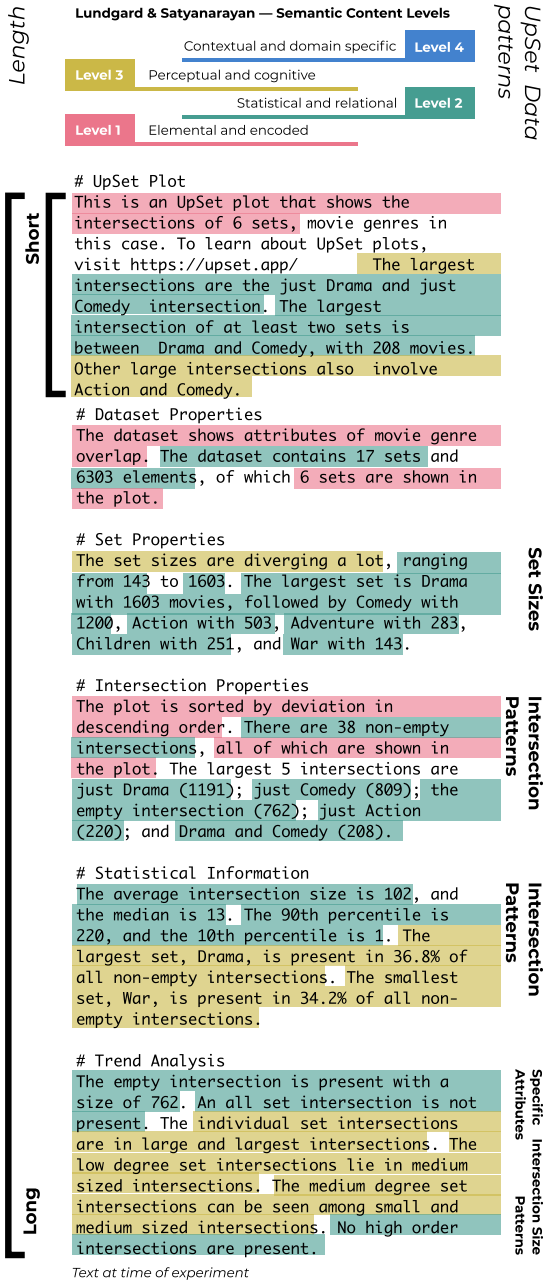


Figure 3: *Our short and long text descriptions annotated with Lundgard and Satyanarayan's [LS22] semantic text model.*

The long description includes Introduction, Dataset Properties, Set Properties, Intersection Properties, Statistical Information, and Trend Analysis. The first four sections have basic information about the type of the plot, how many sets are involved, how many intersections are seen, the count of visible sets, the title of the sets, sorting order, largest intersection name, the largest five to ten intersections, etc. For the statistical information section, we compute a variety of metrics including percentile information, the percentage of the presence of the largest set and smallest set in all the intersections, the average value, and the median divergence of the intersections. Finally, the trend analysis section describes trends that we determine to be present based on our observed data patterns. Examples of both long and short descriptions for several datasets are available at upset.multinet.app.

In designing this description we expected that we are unable to adequately generate Level 4 content (contextual and domain specific) because it requires knowledge of the data and context in which the chart is being used that is outside of the purview of our system. For instance, observations such as "The explosion of $200M+ superhero film budgets in the mid-2010s demonstrates the impact of Marvel's shared universe model, fundamentally altering how studios approach franchise development" would be difficult to generate automatically. We encourage authors of UpSet plots to edit the text descriptions and provide such context. Subsequent work could explore automation of this semantic description step through LLMs, however, as we discuss in Sec. 5, this can lead to hallucination and subtle misinterpretation.

Generating Text. Our text descriptions are generated using simple text templates, one version for the short form description and another for the long. The structure of these templates is alluded to in Fig. 3. For instance, in the long version there are sections for `Set Properties`, `Intersection Properties`, and so on. The content of these sections is formed from a descriptive JSON-based standard that summarizes the configuration of an UpSet plot—such as how Vega-Lite programs summarize the structure of the resulting visualization. It consists of the data, visible sets, plot direction, and sorting method—as well as aspects like aggregation and attribute configuration that we do not currently leverage for our descriptions. In addition to structural metadata, statistical information (such as "`The dataset contains 17 sets and 6303 elements`") is generated by analyzing the input dataset following our observed patterns using standard analytic techniques. In essence, our descriptions are simple functions that take in a configuration and dataset and return markdown containing the description. We use this design, rather than one more closely woven into a particular tool, so that any tool that generates UpSet plots might be instrumented to create these text descriptions.

Implementation. To support these goals we package our system as an open-source library github.com/visdesignlab/upset-alt-txt-gen. To test the effectiveness and abstraction of our library, we explored using it to describe charts generated by our own interactive UpSet tool (described below) and UpSetPlot [Not19]. We had to only minimally modify UpSetPlot to generate our JSON standard, so that we could use our library to generate text descriptions. To support additional libraries, we provide a REST API that accepts our standard and a dataset, and returns a text description. In this way, other versions of UpSet (including our web-based implementation) can use our library to generate text descriptions independent of the programming language used.

We also built a reference tool for creating UpSet plots (upset.multinet.app) that includes a user-provided title and caption, and our text descriptions (see supplemental video). Our tool enables users to edit the text description so that it can be tailored to the specific plot and dataset, while still providing access to the generated text. The description dynamically updates as the plot is manipulated, e.g., adding or removing sets or changing the sorting. To provide further access to BLV users, we include tabular versions

© 2025 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

14678659, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cgf.70102 by Ishrat Jahan Eliza - University Of Utah J. Willard Marriott Library , Wiley Online Library on [12/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 2: *Participant demographics. Severe low vision is classified as a Snellen visual acuity between 20/200 and 20/500, profound low vision is classified as a Snellen visual acuity between 20/500 and 20/1000 [Ame24]. IP = degree in progress; $\bar{X}$ is mean; Y is mode.*

pID	Age	Gender	Daily Hours on Computer	Screen-Reader (Years Exp.)	Described Level of Vision	Vision-Loss Level	Highest Level of Education
P_1	49	Male	8	VoiceOver (39)	Severe low vision	Lost vision gradually	Bachelor's
P_2	71	Female	12	VoiceOver (1)	Severe low vision	Lost vision suddenly	High School
P_3	24	Female	4	JAWS (16)	Light perception	Lost vision suddenly	Associate's (IP)
P_4	35	Male	8	JAWS (16)	Light perception	Lost vision suddenly	High School
P_5	24	Female	4	VoiceOver (10)	Light perception	Blind since birth	High School
P_6	27	Female	6	JAWS (14)	No residual vision	Lost vision suddenly	Master's
P_7	34	Female	10	JAWS (12)	Light Perception	Lost vision gradually	Master's
P_8	32	Female	4	NVDA (28)	No residual vision	Blind since birth	Associate's
P_9	27	Female	14	VoiceOver (8)	No residual vision	Blind since birth	Bachelor's
P_{10}	23	Male	6	VoiceOver (12)	Severe low vision	Lost vision gradually	Bachelor's (IP)
P_{11}	21	Male	7	VoiceOver (8)	Profound low vision	Lost vision gradually	Bachelor's (IP)
	$\bar{X} = 33 \pm 15$	7/11 F	$\bar{X} = 7.5 \pm 3$	Y=VoiceOver, ($\bar{X} = 15 \pm 10$)	Y= Light perception (4)	Y=suddenly(4), gradually (4)	Y= Bachelor's (4)

of the dataset. It is designed to be screen-reader accessible and was validated with our blind coauthor.

4. Evaluation: Do BLV Users Find Our Descriptions Useful?

With our ability to automatically generate text descriptions of UpSet plots now in hand, we want to ensure that these descriptions are useful to our target audience—namely BLV users. In particular, we sought to identify whether our text descriptions sufficiently portray the quantitative information of an UpSet plot to users who cannot see the corresponding visualization. To address this, we conducted a semi-structured interview study with BLV participants.

We find that, generally, people were able to use our descriptions to answer analytic questions and liked their overall structure, but they pointed out potential improvements (e.g., adding bullet points to support navigation).

4.1. Procedure

Interviews took place over Zoom using a pair-interview [AM23] technique in which two research team members were present for all interviews. One interviewer acted as the driver who was primarily responsible for interacting with the participant and asking the prescribed interview questions. The second interviewer actively listened to the discussion and offered follow-up questions. Interviews lasted in total an average of 50 ± 14 minutes. After obtaining consent, we asked demographic questions for 8 ± 3 minutes (the results of which are shown in Table 2). We conducted a pilot with our blind coauthor as a process check.

Participants. We recruited screen-reader users by circulating an IRB-approved flyer distributed at the University of Utah Moran Eye Center and social media groups. Subjects were eligible to participate if they were 18 years of age or older, proficient in English, legally blind (e.g., a visual acuity of 20/200 or greater), and use screen readers daily (as we consider users who regularly use screen-readers as the target audience of our text descriptions). Of the 11 participants, 9 completed the interview on personal laptops and two completed the interview on a smartphone. Participants received a $50 Amazon gift card as compensation. All participants provided informed consent for the procedures approved by our institution's IRB and to have the audio of the interview recorded and automatically transcribed. Transcriptions were manually corrected post hoc. We refer to our interview Participants as P_X and "quote them".

Training. Next, we directed participants to a training page sent via Zoom chat or email (vdl.sci.utah.edu/upset-alttext-example). This page includes a basic description and guide to interpretation for UpSet Plots. We then asked participants to share their screen and audio (screen reader output) to observe which part of the text they were interacting with—which we continued for the rest of the interview. Participants were allowed to reread the training and ask clarifying questions. Training lasted an average of 9 ± 2 minutes.

Text Description. Next, participants opened a new web page with an example UpSet text description. We used a Covid dataset showing co-occurrences of major symptoms (see the appendix for details). We allowed participants to read the text as many times as necessary to understand the data. We then asked participants what they learned from the description. Specific questions included: Can you describe what you learned about the dataset in your own words? How did this dataset increase your understanding of COVID-19 symptoms? We then asked for participants' feedback, including: What was difficult for you to understand about the dataset? What would have been helpful to provide additionally? Do you have comments on the style of the text description? We also asked factual questions, such as: How many sets are shown? What is the largest intersection? The part of the interview lasted 21 ± 7 minutes.

UpSet Comparison. For participants who had some residual vision, we also provided the corresponding COVID-19 symptoms UpSet visualization at this point. We asked these participants if their mental model from the alt-text matched the chart and solicited additional feedback after they had seen the corresponding visualization.

4.2. Results

We now present the findings summarized by theme.

Experience with Screen Readers. To make text accessible to BLV

© 2025 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

14678659, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cgf.70102 by Ishrat Jahan Eliza - University Of Utah J. Willard Marriott Library , Wiley Online Library on [12/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

users, it is important to understand the context in which screen readers are typically used. Most participants frequently use more than one screen reader, depending on the electronic device they use. For the interview, the most commonly used screen reader was Apple's VoiceOver, followed by JAWS (see Table 2). Prior experience with text descriptions for visualizations varied drastically. P_1, P_4, P_7, P_{10} mentioned encountering visualization text descriptions on a daily basis, P_2, P_5, P_9, P_{11} did so at least once a week, while P_3, P_6, P_8 encountered them less than once a week. Most participants encountered visualizations in news sources, for instance, P_9, noted "I come across them in news articles and magazines sometimes". Text description for visualizations were common at school and work settings. P_7 explained "I make use of graphs and charts a lot, work related". Similarly, P_3 recalled "I am taking a human biology course and that involves some graphics and data charts."

Interpretation of the Text Description. Participants sometimes misinterpreted and incorrectly used unfamiliar factual terms. For instance, P_2 highlighted the difficulty of some technical terms: "I think the word intersection might have been in this case an intersection is a person and a symptom together." Correctly understanding the term intersection was also demonstrated by using the term correctly without prompting. For instance, P_3 mentioned, "I have had COVID a couple of times myself. I could put myself in the intersection [of symptoms] I've experienced." Echoing Wang et al.'s [WHT*24] finding that BLV users draw on personal experience to understand charts. Another mistake was misidentifying the largest intersection. To wit, some participants (4/11) were unable to correctly identify anosmia and fatigue as the largest intersection.

Participants reported a mixed sense of understanding the COVID-19 data. Those with a strong grasp on it compared their knowledge with how a sighted person might perceive the UpSet plot. For example, P_3 noted that "I feel like I have a pretty decent sense as to like what this might look like if someone were to look at [the plot] visually." Others comments included having a good mental image of the corresponding UpSet plot (P_{10}) and a better understanding of the co-occurrence of COVID-19 symptoms (P_{11}). Those with a more limited understanding still felt they got the important information. P_8 noted that "if I needed specific information, I could get it from [the text description]" but had a difficult time visualizing the associated chart, a position echoed by P_6 and P_7.

Feedback on the Text Descriptions. Participants provided valuable feedback on improving our text descriptions, emphasizing clarity and accessibility. A common suggestion was to include the glossary of terms from the Introduction page within the description, such as at the top or bottom. For instance, P_4 complained that "I don't know the meanings of some of the words like elements." Providing a glossary simultaneously (alongside the description) may also reduce misinterpretations of terms. P_{10} noted concerns with terminology depending on the text description writer's training as "someone could write the text description and put it a different way"—highlighting the role of expertise and domain knowledge for both the writer (and, we add, the reader).

Participants, overall, felt the length of the description was appropriate. Many noted that it was concise (P_2, P_4) and descriptive (P_3, P_9). P_5 explained "even though there is lots of very complex data, the numbers really help me understand things better." Several participants commented on the amount of numbers in the description. For example, P_{11} recalled being unsure about the meaning of the numbers but stated, "it was clear after the second read." Participants also noted that numbers are fully read out by screen readers. P_2 explained that the number "281" might be skimmed over when reading visually while a screen reader fully articulates it as "two hundred and eighty-one". P_3 similarly noted that fully articulated numbers "takes me a little bit longer to go back and read them." We also observed that several participants slowed down the speech rate of their screen readers to receive the information at a slower pace. This suggests that numbers could be summarized with fewer significant figures with by a high fidelity table of numbers elsewhere (akin to the tables often accompanying visualizations [ZLL*22]).

Consistent with the feedback from our blind coauthor, all participants valued having headings in the alt-text. Participants stressed that headers provide an array of benefits to screen readers. For instance, organizing information by headings is "easier to navigate with screen readers" (P_5) and "helps you get right where you need very easily" (P_8). P_7 noted that the headings "helped with how I was processing what I was reading." Several participants suggested adding bullet points to the alt-text as an additional organizational tool. Others indicated that adding bullet points alongside the existing headings would help with navigating long texts (P_8, P_{11}), breaking up lengthy sentences (P_2), and organizing technical or number-intensive sections, such as statistical information (P_6, P_{10}). In addition to benefiting screen readers, organizing the information this way may be useful for other accessibility tools.

Based on these results, we updated text description to the results bullet-point and added a glossary, as in Supplementary Figure 14.

Alternative Presentations. Two participants (P_2, P_{10}) with residual vision also viewed the COVID-19 UpSet plot with screen magnification. Both participants reported that their mental model—a common practice for BLV users per Jung et al. [JMK*22]—created from the alt-text did not match the actual UpSet plot and specifically noted the matrix indicating the intersections (cf. Fig. 1). P_2 stated that, "I read it more of just the size, like the section [with the bar graphs]." Similarly, P_{10} stated, "I wasn't expecting the bubbles. The bar chart is what I was visualizing." Both participants stated that at their vision level, they would prefer visual UpSets to text because of the level of detail (P_{10}) and convenience (P_2).

Although most interviewees stated that the text description was sufficient at portraying the COVID-19 data, P_1, P_3, P_6, P_7, P_9, P_{10}, and P_{11} expressed a preference for tactile displays. We observed that participants compared tactile displays to the knowledge a sighted person would gain from a visualization. As summarized by P_{10}, "if there is a tactile element associated with it, it'll be like someone looking at it. So it would be useful." P_6 also expressed similar association as they state "I feel like that's the same, even though I can't see." Additionally, P_3 mentioned how a tactile display of an UpSet plot would help understand the layout of the visualization as "it would be nice to see what an [UpSet plot] would look like in general." In future work we intend to explore how our automated captions can tie into physicalizations.

© 2025 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

14678659, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cgf.70102 by Ishrat Jahan Eliza - University Of Utah J. Willard Marriott Library , Wiley Online Library on [12/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

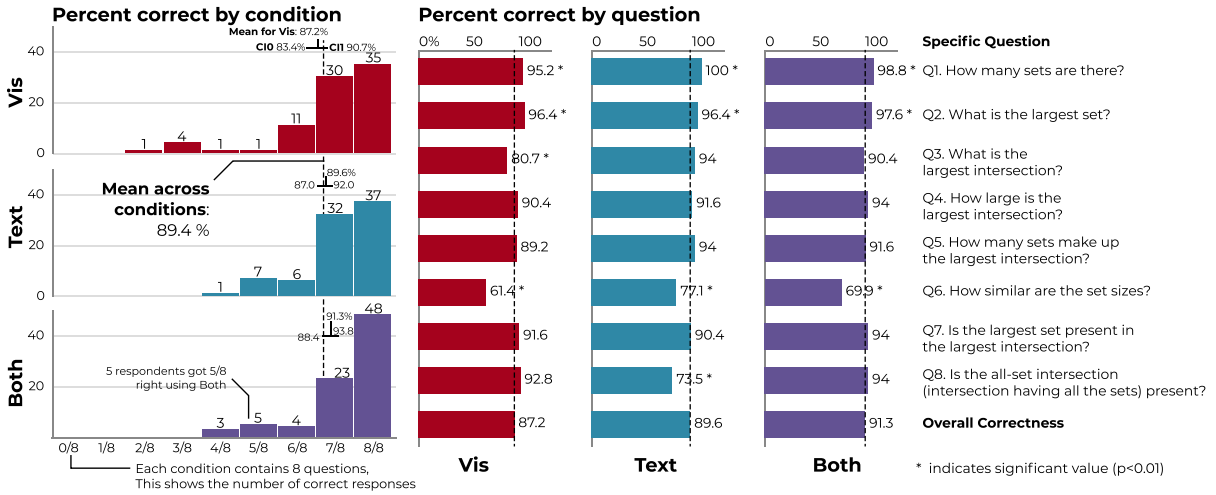


Figure 4: *Correct responses by condition (left) and by question (right). Each condition had 83 participants, all of whom saw all conditions.*

5. Contextualizing Experiments

We complement our interview study with two experiments. These studies are not key to our findings but offer context. We elide most details (see appendix for them) and present only high level findings.

5.1. Experiment 1: Utility for Sighted Users

As a simplistic quality check, we considered if the affordances present in our text descriptions support the same level of analytic functionality as visual UpSet plots, and if our text descriptions can also be useful to sighted users, indicating a curb-cut effect where assistive infrastructure can benefit all users of that infrastructure, analogous to how video subtitles are useful for more than just language barriers [LS22, PO12].

To answer these question, we conducted a within-subjects crowd-work experiment with (N=83) sighted users. Participants engaged with three conditions: just visualization (VIS), just text (TEXT), and text and visualization combined (BOTH) across three datasets arranged in a Latin square. For each dataset-condition combination we asked questions regarding each of the patterns described in Sec. 3.1. For VIS and BOTH we used static UpSet plots generated with our web tool. For TEXT and BOTH we used the long form text-descriptions. In all conditions, we included a human-provided title and an introduction that leverages facts about the dataset (what are the items and sets). The experiment was implemented with reVISit [DWS*23] and is available online.

While there are a variety of metrics of interest in this study, the key one is correctness: whether participants get answers right in each medium. We hypothesized that there will be no statistical difference for correctness between the three conditions. To support this question, we ran a mixed model. Correctness (reported as a percentage and summarized in Fig. 4) was regressed onto all three conditions. Across conditions, approximately 6% of the variance was explained by between-subject variability ($ICC = 0.06$). Thus, we included a random intercept by participant to account for these inter-individual differences. Correctness in the VIS condition was significantly ($p = 0.01$) less than in BOTH. However, correctness in text-only did not significantly differ from either other condition (BOTH $p = 0.31$; VIS $p = 0.14$). While this rejects our hypothesis, the interpretation is straightforward: having two tools for analysis improves the results. One participant explained "The combination of the text data and the visual graph is the easiest to understand in my opinion, providing most of the answers with a cursory glance but also providing the other answers with a bit deeper of a look." Suggesting that descriptions exhibit a curb-cut effect, making them useful for sighted people, as well as for a BLV audience.

The correctness parity between TEXT and VIS suggests that **our descriptions are, for sighted users, at least as good as the visual**. However, we note that the ability to answer factual questions at similar correctness levels does not necessarily mean that TEXT or VIS confer benefits typically associated with visualizations—such as being memorable or getting understanding "the gist" of the data. Finally, we note that this is the first experimental study of UpSet plot's usability and that it confirms that they are usable—as generally suggested from their widespread usage.

5.2. Experiment 2: Description Replacability

Here, we consider how well our bespoke descriptions improve over a naive replacement description. As there has not been previously been work on captions for this chart form, there is not an existing baseline by which to compare. Instead, we use LLMs to generate text, the quality of which we explore through an ablation study (see appendix). Across two studies we explore the effect of including different prompt components (such as the dataset, example descriptions, etc.), the impact of commonly known data for which context is available (like movies) vs anonymous data, and models (including OpenAI's GPT and Anthropic's Claude). We then coded the descriptions using the semantic levels model [LS22] and conducted an expert review with our blind coauthor.

We find that while LLMs can produce high-quality text descriptions (sometimes in a more narrative format), they have nontrivial variance between generations that can contain false statements. The results are also dependent on the presence of related data in the models' training set. While others [DCBD24] have previously used LLMs to successfully generate descriptive texts, we find these

© 2025 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

14678659, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cgf.70102 by Ishrat Jahan Eliza - University Of Utah J. Willard Marriott Library , Wiley Online Library on [12/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

successes come with caveats. We do not settle whether these risks are worthwhile, but find that our approach offers a consistent and reasonable quality compared to descriptions generated by LLMs.

6. Discussion

In this paper we explored how to develop accessible descriptive text for complex scientific charts in general by focusing on the specific case of UpSet plots. We developed a system for automatically generating descriptions based on a survey of extant practices based on published usage. This system is general purpose and can be used to generate descriptive texts for any UpSet whose input can be summarized through a standard that we define. Through this work we strive to make UpSet plots accessible to anyone.

6.1. Transferring our Process to Other Chart Types

Substantial study [TBS23,NM13,OH20,EMW19] has already gone into understanding effective design of descriptions for basic chart types (such as line charts, bar charts, etc.), but there has been little consideration for more bespoke chart forms. Outside of the UpSet specifics, a facet of our contribution is an example of a process for developing descriptions for complex or scientific chart forms.

We ***synthesized*** extant practices (via a survey of published charts). In particular, developing an understanding of not just the idioms of the chart type, but the patterns and practices that come from real usage were essential. These patterns and practices then can inform the choices of algorithms to surface them.

We found that it was essential to ***develop language*** around the data and layout patterns specific to UpSet plots. Many charts enjoy a substantial body of extant patterns from which to draw upon in their descriptions, but new chart forms do not come ready-made with such descriptors. Consideration of the patterns of and in chart types seems more broadly valuable, particularly as a component of inter-medium communication (i.e. between visual and text).

We ***co-designed*** with a member of the community with whom we sought to work (via collaboration with a blind coauthor). This allowed us to catch many assumptions and not overfit to our biases. We designed both our system and our descriptive text so they are ***polymorphic***; that is they are not tied to a single usage context.

We ***implemented*** our approach in a platform-independent way, using a JSON standard and a web API, so that we do not solve the problem of text generation for a chart for a single implementation, but make it accessible to implementations in different languages.

We ***validated*** our approach in interviews with BLV users and provide templates for these interviews that can be adapted to validate text descriptions for other charts in the future.

We believe that this approach can be generalized to generate text descriptions for other non-standard chart types, such as network diagrams or tree maps.

6.2. Limitations

As with any work ours has a variety of limitations. Our work centers on a single style of description. It is possible that other descriptions (such as by including the suggestions from our BLV interviewees) might have led to improved performance, however this was not examined. The goal of this work is not to perfect descriptions for UpSet plots (as such a goal is akin to writing the perfect greeting card), but instead to demonstrate they can be constructed in a consistent and useful manner. Similarly, we focused only on a single chart type. If other types of set visualization had been considered as well, our results may have changed. Our focus was on UpSet specifically due to their relative popularity, but in the future comparing with other set visualizations (such as Euler diagrams) would be useful.

While we strove to incorporate real world UpSet practices, our text descriptions only covered a subset of all possible plots. For instance, our UpSet implementation includes advanced features (e.g., aggregation, annotations) that are not currently captured by our descriptions. Future work should continue to add nuance to our descriptions to accommodate richer situations.

While we did ask some analytic questions of our interviewees, we did not perform a quantitative evaluation of BLV users' performance using our descriptions. Such studies would be valuable future work, but we emphasize that our goal was to demonstrate that they could be useful broadly, and, in doing so, we use these interviews as an opportunity to identify missing pieces in our design.

6.3. Personalization

Like any users, BLV users have varied preferences and expectations for presented information format (as Jones et al. [JPPH*24] note). Some may prefer to have topics organized by bullet points, while others may prefer iteratively expanding summaries. Similarly, this work exclusively focused on the accessibility of descriptions for BLV users, however other disabilities should be evaluated in the future—such as people with developmental disabilities [WPA*21], like ADHD, or learning disabilities, like dyslexia or dysgraphia. Moreover, additional work is likely necessary to explore how the visual representation of UpSet plots themselves might be adapted for people for whom understanding multiview visualizations is challenging. Echoing Wu et al.'s [WQW*24] findings that narrativization is useful for people with intellectual disabilities, we suggest this type of affordance is useful for any group. This introduces the possibility of false-information being induced to the output, which we should design around carefully.

6.4. Interactivity

While the web tool we built supports **interactive analysis** (including for BLV users by regenerating the descriptions on specification change), the descriptions do little to support interactivity directly. Future work might explore how user interaction might be woven into these descriptions to support more dynamic analysis. Alternatively, an approach similar to Zong et al's rich screen reader experiences [ZLL*22], where BLV users can navigate the primary chart, e.g., using data navigator [ENM24] is conceivable.

7. Acknowledgments

We gratefully acknowledge funding by Chan Zuckerberg Initiative Essential Open Source Software for Science program.

© 2025 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

14678659, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cgf.70102 by Ishrat Jahan Eliza - University Of Utah J. Willard Marriott Library , Wiley Online Library on [12/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

References

[AM23] AKBABA D., MEYER M.: "Two Heads are Better than One": Pair-Interviews for Visualization. In *IEEE Visualization and Visual Analytics (VIS)* (Oct. 2023), IEEE, pp. 206–210. doi:10.1109/VIS54172.2023.00050. 7

[Ame24] AMERICAN OPTOMETRIC ASSOCIATION: Low Vision and Vision Rehabilitation, 2024. URL: https://www.aoa.org/healthy-eyes/caring-for-your-eyes/low-vision-and-vision-rehab?sso=y. 7

[Aut] AUTHORITY C. I.: International organization participation - The World Factbook. URL: https://www.cia.gov/the-world-factbook/field/international-organization-participation/. 14

[BHMG17] BUTLER M., HOLLOWAY L., MARRIOTT K., GONCU C.: Understanding the graphical challenges faced by vision-impaired students in Australian universities. *Higher Education Research & Development 36*, 1 (Jan. 2017), 59–72. doi:10.1080/07294360.2016.1177001. 1

[BKL*06] BIGHAM J. P., KAMINSKY R. S., LADNER R. E., DANIELSSON O. M., HEMPTON G. L.: WebInSight: Making web images accessible. In *International ACM SIGACCESS Conference on Computers and Accessibility* (Oct. 2006), ACM, pp. 181–188. doi:10.1145/1168987.1169018. 3

[BZ] BLANCO M., ZONG J.: Olli. URL: mitvis.github.io/olli/. 3

[CBW22] CHINTALAPATI S. S., BRAGG J., WANG L. L.: A Dataset of Alt Texts from HCI Publications: Analyses and Uses Towards Producing More Descriptive Alt Texts of Data Visualizations in Scientific Papers. In *SIGACCESS Conference on Computers and Accessibility* (Oct. 2022), ACM, pp. 1–12. doi:10.1145/3517428.3544796. 3

[CJP*19] CHOI J., JUNG S., PARK D. G., CHOO J., ELMQVIST N.: Visualizing for the Non-Visual: Enabling the Visually Impaired to Use Visualization. *Computer Graphics Forum 38*, 3 (2019), 249–260. doi:10.1111/cgf.13686. 3

[CLG17] CONWAY J. R., LEX A., GEHLENBORG N.: UpSetR: An R package for the visualization of intersecting sets and their properties. *Bioinformatics 33*, 18 (Sept. 2017), 2938–2940. doi:10.1093/bioinformatics/btx364. 2, 4

[CPR*22] CHUNDURY P., PATNAIK B., REYAZUDDIN Y., TANG C., LAZAR J., ELMQVIST N.: Towards Understanding Sensory Substitution for Accessible Visualization: An Interview Study. *IEEE Transactions on Visualization and Computer Graphics 28*, 1 (Jan. 2022), 1084–1094. doi:10.1109/TVCG.2021.3114829. 3

[CS14] CORBIN J., STRAUSS A.: *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. SAGE Publications, Nov. 2014. 4

[DCBD24] DUARTE D., COSTA R., BIZARRO P., DUARTE C.: AutoVizuA11y: A Tool to Automate Screen Reader Accessibility in Charts. *Computer Graphics Forum 43*, 3 (2024), e15099. doi:10.1111/cgf.15099. 3, 9, 18

[DWS*23] DING Y., WILBURN J., SHRESTHA H., NDLOVU A., GADHAVE K., NOBRE C., LEX A., HARRISON L.: reVISit: Supporting Scalable Evaluation of Interactive Visualizations. In *2023 IEEE Visualization and Visual Analytics (VIS)* (Oct. 2023), IEEE, pp. 31–35. doi:10.1109/VIS54172.2023.00015. 9, 14

[EMW19] ENGEL C., MÜLLER E. F., WEBER G.: SVGPlott: An accessible tool to generate highly adaptable, accessible audio-tactile charts for and from blind and visually impaired people. In *ACM International Conference on PErvasive Technologies Related to Assistive Environments* (June 2019), pp. 186–195. doi:10.1145/3316782.3316793. 10

[ENM24] ELAVSKY F., NADOLSKIS L., MORITZ D.: Data Navigator: An Accessibility-Centered Data Navigation Toolkit. *IEEE Transactions on Visualization and Computer Graphics 30*, 1 (Jan. 2024), 803–813. doi:10.1109/TVCG.2023.3327393. 3, 10

[EW18] ENGEL C., WEBER G.: A User Study to Evaluate Tactile Charts with Blind and Visually Impaired People. In *Computers Helping People with Special Needs* (2018), Lecture Notes in Computer Science, Springer, pp. 177–184. doi:10.1007/978-3-319-94274-2_24. 3

[FMB*22] FRIEDMAN J., MONTERO F., BOURGOIS P., WAHBI R., DYE D., GOODMAN-MEZA D., SHOVER C.: Xylazine spreads across the US: A growing component of the increasingly synthetic and polysubstance overdose crisis. *Drug and Alcohol Dependence 233* (Apr. 2022), 109380. doi:10.1016/j.drugalcdep.2022.109380. 4, 5

[GPM*20] GLEASON C., PAVEL A., MCCAMEY E., LOW C., CARRINGTON P., KITANI K. M., BIGHAM J. P.: Twitter A11y: A Browser Extension to Make Twitter Images Accessible. In *CHI Conference on Human Factors in Computing Systems* (Apr. 2020), ACM, pp. 1–12. doi:10.1145/3313831.3376728. 3

[HK16] HARPER F. M., KONSTAN J. A.: The MovieLens Datasets: History and Context. *ACM Transactions on Interactive Intelligent Systems 5*, 4 (Jan. 2016), 1–19. doi:10.1145/2827872. 14

[JMK*22] JUNG C., MEHTA S., KULKARNI A., ZHAO Y., KIM Y.-S.: Communicating Visualizations without Visuals: Investigation of Visualization Alternative Text for People with Visual Impairments. *IEEE Transactions on Visualization and Computer Graphics 28*, 1 (Jan. 2022), 1095–1105. doi:10.1109/TVCG.2021.3114846. 1, 3, 8

[JPPH*24] JONES S., PEDRAZA PINEROS I., HAJAS D., ZONG J., SATYANARAYAN A.: "Customization is Key": Reconfigurable Textual Tokens for Accessible Data Visualizations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (May 2024), ACM, pp. 1–14. doi:10.1145/3613904.3641970. 10, 18

[Kan24] KANE R.: Claude vs. ChatGPT: What's the difference? [2024], 2024. URL: https://zapier.com/blog/claude-vs-chatgpt/. 18

[Kin98] KING N.: Template analysis. In *Qualitative Methods and Analysis in Organizational Research*. Sage Publications Ltd, 1998. URL: https://psycnet.apa.org/record/1999-02931-006. 4, 18

[KJRK21] KIM N. W., JOYNER S. C., RIEGELHUTH A., KIM Y.: Accessible visualization: Design space, opportunities, and challenges. *Computer Graphics Forum 40*, 3 (2021), 173–188. doi:10.1111/cgf.14298. 2

[KKH24] KIM H., KIM Y.-S., HULLMAN J.: Erie: A Declarative Grammar for Data Sonification. In *CHI Conference on Human Factors in Computing Systems* (May 2024), ACM, pp. 1–19. doi:10.1145/3613904.3642442. 3

[KKK23] KIM G., KIM J., KIM Y.-S.: "Explain What a Treemap is": Exploratory Investigation of Strategies for Explaining Unfamiliar Chart to Blind and Low Vision Users. In *CHI Conference on Human Factors in Computing Systems* (Apr. 2023), ACM, pp. 1–13. doi:10.1145/3544548.3581139. 2, 3

[Lex24] LEX A.: Implementations of UpSet, 2024. URL: https://upset.app/implementations/. 3

[LG] L'YI S., GEHLENBORG N.: Altair-Based UpSet Plots for COVID Symptoms. URL: http://upset.plus/covid19symptoms/. 14

[LGS*14] LEX A., GEHLENBORG N., STROBELT H., VUILLEMOT R., PFISTER H.: UpSet: Visualization of Intersecting Sets. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 1983–1992. doi:10.1109/TVCG.2014.2346248. 2, 3, 4, 5

[LLS19] LUNDGARD A., LEE C., SATYANARAYAN A.: Sociotechnical considerations for accessible visualization design. In *IEEE Visualization Conference* (2019), IEEE, pp. 16–20. doi:10.1109/VISUAL.2019.8933762. 3

[LPS*24] LI C., PANG R. Y., SHARIF A., CHHEDA-KOTHARY A., HEER J., FROEHLICH J. E.: AltGeoViz: Facilitating Accessible Geovisualization, June 2024. URL: http://arxiv.org/abs/2406.13853, arXiv:2406.13853. 2, 3

© 2025 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

14678659, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cgf.70102 by Ishrat Jahan Eliza - University Of Utah J. Willard Marriott Library , Wiley Online Library on [12/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

[LS22] LUNDGARD A., SATYANARAYAN A.: Accessible Visualization via Natural Language Descriptions: A Four-Level Model of Semantic Content. *IEEE Transactions on Visualization and Computer Graphics 28*, 1 (Jan. 2022), 1073–1083. doi:10.1109/TVCG.2021.3114770. 2, 3, 5, 6, 9, 15, 19

[LSLG24] L'YI S., SMITS T. C., LEX A., GEHLENBORG N.: Digital Accessibility of Life Sciences Data Portals and Journal Websites. *OSF.io* (Feb. 2024). doi:10.31219/osf.io/5v98j. 2

[LSZ*22] LU J. Y., SIMON M., ZHAO Y., ABLAEVA J., CORSON N., CHOI Y., YAMADA K. Y. H., SCHORK N. J., HOOD W. R., HILL G. E., MILLER R. A., SELUANOV A., GORBUNOVA V.: Comparative transcriptomics reveals circadian and pluripotency networks as two pillars of longevity regulation. *Cell Metabolism 34*, 6 (June 2022), 836–856.e5. doi:10.1016/j.cmet.2022.04.011. 5

[Lui23] LUIPI G.: Opinion | My Life With Long Covid - The New York Times, 2023. URL: https://www.nytimes.com/interactive/2023/12/14/opinion/my-life-with-long-covid.html. 2

[MCLM21] MACK K., CUTRELL E., LEE B., MORRIS M. R.: Designing Tools for High-Quality Alt Text Authoring. In *SIGACCESS Conference on Computers and Accessibility* (Oct. 2021), ACM, pp. 1–14. doi:10.1145/3441852.3471207. 3

[MJBC18] MORRIS M. R., JOHNSON J., BENNETT C. L., CUTRELL E.: Rich Representations of Visual Content for Screen Reader Users. In *CHI Conference on Human Factors in Computing Systems* (Apr. 2018), CHI '18, ACM, pp. 1–11. doi:10.1145/3173574.3173633. 3

[MSMC14] MORAES P., SINA G., MCCOY K., CARBERRY S.: Evaluating the accessibility of line graphs through textual summaries for visually impaired users. In *SIGACCESS Conference on Computers & Accessibility* (Oct. 2014), ASSETS '14, ACM, pp. 83–90. doi:10.1145/2661334.2661368. 3

[Nat23] NATIONAL SCIENCE FOUNDATION: *Diversity and STEM Women, Minorities, and Persons with Disabilities*. Tech. Rep. NSF 23-315, National Science Foundation, 2023. 1

[NM13] NAZEMI A., MURRAY I.: A Method to Provide Accessibility for Visual Components to Vision Impaired. *International Journal of Human Computer Interaction 4*, 1 (2013), 54–69. URL: https://espace.curtin.edu.au/handle/20.500.11937/39283. 10

[Not19] NOTHMAN J.: UpSetPlot, 2019. URL: https://github.com/jnothman/UpSetPlot. 6

[OH20] OBEID J., HOQUE E.: Chart-to-Text: Generating Natural Language Descriptions for Charts by Adapting the Transformer Model. In *International Conference on Natural Language Generation* (Dublin, Ireland, Dec. 2020), ACL, pp. 138–147. doi:10.18653/v1/2020.inlg-1.20. 3, 10

[PO12] PRASAD G. V. S., OJHA A.: Text, Table and Graph – Which is Faster and More Accurate to Understand? In *2012 IEEE Fourth International Conference on Technology for Education* (July 2012), pp. 126–131. doi:10.1109/T4E.2012.18. 9, 15, 17

[QKD*20] QIAN X., KOH E., DU F., KIM S., CHAN J.: A Formative Study on Designing Accurate and Natural Figure Captioning Systems. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Apr. 2020), CHI EA '20, ACM, pp. 1–8. doi:10.1145/3334480.3382946. 3

[QKD*21] QIAN X., KOH E., DU F., KIM S., CHAN J., ROSSI R. A., MALIK S., LEE T. Y.: Generating Accurate Caption Units for Figure Captioning. In *Proceedings of the Web Conference* (June 2021), ACM, pp. 2792–2804. doi:10.1145/3442381.3449923. 3

[SL21] SHAHIRA K. C., LIJIYA A.: Towards Assisting the Visually Impaired: A Review on Techniques for Decoding the Visual Data From Chart Images. *IEEE Access 9* (2021), 52926–52943. doi:10.1109/ACCESS.2021.3069205. 1

[SV10] SPECKMANN B., VERBEEK K.: Necklace Maps. *IEEE Transactions on Visualization and Computer Graphics 16*, 6 (nov.-dec. 2010), 881–889. doi:10.1109/TVCG.2010.180. 2

[TBS23] TANG B., BOGGUST A., SATYANARAYAN A.: VisText: A Benchmark for Semantically Rich Chart Captioning. In *Annual Meeting of the Association for Computational Linguistics* (2023), ACL, pp. 7268–7298. doi:10.18653/v1/2023.acl-long.401. 2, 3, 10

[TMS*23] THOMPSON J. R., MARTINEZ J. J., SARIKAYA A., CUTRELL E., LEE B.: Chart Reader: Accessible Visualization Experiences Designed with Screen Reader Users. In *CHI Conference on Human Factors in Computing Systems* (Apr. 2023), CHI '23, ACM, pp. 1–18. doi:10.1145/3544548.3581186. 3

[WdH*22] WILLIAMS C., DE GREEF L., HARRIS E., FINDLATER L., PAVEL A., BENNETT C.: Toward Supporting Quality Alt Text in Computing Publications. In *International Web for All Conference* (Apr. 2022), W4A '22, ACM, pp. 1–12. doi:10.1145/3493612.3520449. 3

[Wes] WEST C. E.: UpSetR: Beyond the Venn diagram. URL: https://clarewest.github.io/blog/post/2020-03-26-upsetr-beyond-the-venn-diagram/. 14

[WHT*24] WANG H. W., HOFFSWELL J., THANE S. M. T., BURSZTYN V. S., BEARFIELD C. X.: How Aligned are Human Chart Takeaways and LLM Predictions? A Case Study on Bar Charts with Varying Layouts. *IEEE Transactions on Visualization and Computer Graphics* (2024), 1–11. doi:10.1109/TVCG.2024.3456378. 8

[WPA*21] WU K., PETERSEN E., AHMAD T., BURLINSON D., TANIS S., SZAFIR D. A.: Understanding Data Accessibility for People with Intellectual and Developmental Disabilities. In *CHI Conference on Human Factors in Computing Systems* (May 2021), ACM, pp. 1–16. doi:10.1145/3411764.3445743. 10

[WQW*24] WU K., QUADRI G. J., WANG A. Z., OSEI-TUTU D. K., PETERSEN E., KOUSHIK V., SZAFIR D. A.: Our Stories, Our Data: Co-designing Visualizations with People with Intellectual and Developmental Disabilities. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility* (Oct. 2024), ASSETS '24, ACM, pp. 1–17. doi:10.1145/3663548.3675615. 10

[WSK24] WANG Y., SRINIVASAN A., KIM Y.-S.: Enabling Tabular Data Exploration for Blind and Low-Vision Users. In *Designing Interactive Systems Conference* (July 2024), ACM, pp. 1218–1233. doi:10.1145/3643834.3661609. 3

[WW08] WILKINSON L., WILLS G.: Scagnostics Distributions. *Journal of Computational and Graphical Statistics 17*, 2 (June 2008), 473–491. doi:10.1198/106186008X320465. 4

[WWFS17] WU S., WIELAND J., FARIVAR O., SCHILLER J.: Automatic Alt-text: Computer-generated Image Descriptions for Blind Users on a Social Network Service. In *Conference on Computer Supported Cooperative Work and Social Computing* (Feb. 2017), ACM, pp. 1180–1192. doi:10.1145/2998181.2998364. 2, 3

[Yal19] YALANSKA M.: UX Writing: Handy Tips on Text Improving User Experience, Mar. 2019. URL: https://blog.tubikstudio.com/user-experience-tips-ux-writing/. 5

[ZLL*22] ZONG J., LEE C., LUNDGARD A., JANG J., HAJAS D., SATYANARAYAN A.: Rich Screen Reader Experiences for Accessible Data Visualization. *Computer Graphics Forum 41*, 3 (2022), 15–27. doi:10.1111/cgf.14519. 1, 3, 8, 10

[ZPPC*24] ZONG J., PEDRAZA PINEROS I., CHEN M. K., HAJAS D., SATYANARAYAN A.: Umwelt: Accessible Structured Editing of Multi-Modal Data Representations. In *CHI Conference on Human Factors in Computing Systems* (May 2024), ACM, pp. 1–20. doi:10.1145/3613904.3641996. 3

© 2025 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

14678659, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cgf.70102 by Ishrat Jahan Eliza - University Of Utah J. Willard Marriott Library , Wiley Online Library on [12/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Appendix A: Appendix Introduction

In this appendix we include, as appendices, various details and features which did not succinctly fit in the main text. First, in Sec. B we collect all links to supplementary resources on the web. In Sec. C, we provide a full description of our contextualizing experiment in which we surveyed sighted users about their ability to user our text descriptions. Then, in Sec. D, we provide a full description of our study of LLM's ability to produce alternative texts. Next, Sec. E, we provide the instruments used in our various studies. We then, in Sec. F, provide examples of text descriptions.

Appendix B: Resources

The OSF repository containing all supplementary material:
https://osf.io/kbvs9/

The UpSet text description generation code:
https://github.com/visdesignlab/upset-alt-txt-gen

The live-version of the web-based UpSet tool with text descriptions:
https://upset.multinet.app/

The code for the web-based UpSet tool:
https://github.com/visdesignlab/upset2

The modified version of UpSetPlot that can be used to generate text-descriptions via Python API call:
https://github.com/JakeWags/UpSetPlot/tree/alt-text-generation

An example notebook illustrating how to generate text descriptions in Python:
https://github.com/visdesignlab/UpSet-Survey-Data-Analysis/blob/main/suppl_upsetplot.ipynb

Interview Study

The code for the stimuli used in the interview study:
https://github.com/visdesignlab/upset-alttext-example

The deployed version of the stimuli:
https://vdl.sci.utah.edu/upset-alttext-example/

Crowdsourced Experiment

The code defining the crowdsourced experiment:
https://github.com/visdesignlab/Upset-alttxt-study

The deployed crowdsourced experiment:
https://vdl.sci.utah.edu/Upset-alttxt-study/Upset-Alttext-User-Survey/

The code for the data analysis:
https://github.com/visdesignlab/UpSet-Survey-Data-Analysis

Appendix C: Contextualizing Experiment 1: For sighted users, are our descriptions equivalent to the UpSet plots?

As simplistic quality check, we sought to understand if the affordances present in our text descriptions are equivalent those in the visual representation of upset plots. To answer this question we conducted a within-subjects crowd-work experiment with sighted users. Participants engaged with three conditions: just visualization (VIS), just text (TEXT), and text and visualization combined (BOTH). We included the BOTH condition to see whether there are benefits of including text descriptions even when visualizations are present. Through this experiment we sought to answer the following questions:

$H_{\textbf{accuracy}}$**:** *Can participants answer factual questions about UpSet plots equally well in the three conditions?* We hypothesize that there will be no statistical difference for correctness between the three conditions.

$H_{\textbf{time}}$**:** *Does the description-only condition take significantly longer to read than the visualization conditions?* While time is not an essential attribute for most analysis tasks (if the differences are small), we speculate that the TEXT condition will lead to slower response times, because text has to be scanned linearly.

$H_{\textbf{preference}}$**:** *Which conditions will participants prefer?* We speculate participants will prefer the VIS and BOTH conditions and dislike the just TEXT condition. We expect measures of confidence, effectiveness and understandability to be lowest for TEXT.

© 2025 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

14678659, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cgf.70102 by Ishrat Jahan Eliza - University Of Utah J. Willard Marriott Library , Wiley Online Library on [12/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

$H_{\textbf{insights}}$: *Are the types of insights that are available through only* TEXT *substantially different?* We speculate that VIS/BOTH lead to "higher level" synthesis (e.g., trends or correlations) and that TEXT leads to re-stating facts.

With respect to our top-level question, we find that performance with TEXT and VIS are roughly equivalent in relevant dimensions. In addition we find that BOTH tends to perform better, suggesting that descriptions (or at least simplified alternative presentations) have value for usages other than simply making charts more accessible.

Procedure

The experiment involved four principal phases: consent, tutorial, experiment, and post-experiment survey. This study was reviewed by our institution's IRB and deemed exempt from full board review. We used reVISit [DWS*23] to develop and deploy the study.

Participants and Pilots. We recruited N=83 participants from the US (21), UK (30), and Canada (32) using Prolific. Subjects were eligible to participate if they were fluent in English and over 18. Participants took about 27 minutes on average and were compensated $8 USD, for an average hourly pay of about $17 USD. They were on average 30.7 ± 9.6 years old. 31 Female/ 52 Male.

We conducted down-the-hall pilot with (N=4) users and then a full pilot on (N=5) Prolific users. These revealed that users who did not read the introduction did poorly in the experiment. In response, we required that the training questions be answered correctly after three tries. This requirement may have led to a high return rate: 84 participants completed the study, 8 were rejected because they did not complete the study, 80 returned the study after reading the warning. During analysis, we discovered one participant was not completely logged in the raw file. The participant completed the training and first condition (BOTH), but the answers to the rest of the questions were missing and so we excluded them. to be consistent with the rest of the study.

The experiment began with an elicitation of consent and an attention check tutorial. For each dataset we asked questions regarding each of the patterns described in Sec. 3.1—for instance, whether or not the all-set intersection is present. We list these questions in Figure 4. We did not explicitly ask about intersection patterns, such as the presence and prominence of high-order set intersections because these responses require higher-level synthesis and are not easily answered succinctly. Instead, we used open-ended questions about the insights and takeaways. In addition, we asked for participants' confidence in their answers, their understanding of the information, and effectiveness of the presentation on a 1-5 Likert Scale (5 being high).

Figure 5 shows an example of the study in the BOTH condition. For the VIS and BOTH condition we used static UpSet plots generated with our web tool and sorted by size. For the TEXT and BOTH conditions, we used the text-description as provided by the long form. In all conditions, we included a human-provided title and an introduction that leverages human-provided facts about the dataset (what are the items, what are the sets).

We used four different datasets:

Movie Genres. Movies and movie genres [HK16]. 6 sets. Intersections are low-degree. Used only for the training stage.
Tennis. Tennis players who won one or multiple of the four grand slam tournaments [Wes]. 4 sets. Mix of single-set intersections and high-degree intersections. All-set is prominent. See Figure 2.
Organizations. Countries and their participation in global organizations (e.g., the WHO or NATO) [Aut]. 8 sets. Very unbalanced set sizes, lots of high-degree intersections, but no all-set intersection. See Figure 1 right.
Covid. COVID-19 symptoms [LG]. 6 sets. Unbalanced set size, many combinations of symptoms present, including all-set intersection.

To counterbalance learning effects, we used a Latin square to permute condition assignment (VIS, TEXT, BOTH) to datasets (Tennis, Organizations, Covid), and to permute the order in which the conditions would appear in the sequence (yielding a 3x3x3 square). The high return rate led to a slight imbalance in dataset-condition combinations. On average, 27.6 user saw each combination. We believe that these minor variations do not affect our results.

Post-Experiment Survey. The session concluded with a short survey covering preferences, perceived effectiveness, readability and description length (rated on a 1-5 Likert scale, 5 high), as well as open-ended general comments.

Limitations

The high return rate of our study led to a slight in-balance between groups in the Latin square, and data for one participant was lost, possibly due to connectivity issues. While unfortunate, we believe the effect of these errors to be minor and to substantially not effect the content of our analysis. The design of our survey focused on simple analytic tasks, instead of more complex exploratory-type tasks. While investigating these areas would be useful, the goal of this chart in publications is typically presentational rather than exploratory.

Results

Next we describe the results of this study incorporating both the quantitative results, as well as analysis of the free text results. We refer to our Survey Participants as SP_X and "quote them".

© 2025 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

14678659, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cgf.70102 by Ishrat Jahan Eliza - University Of Utah J. Willard Marriott Library , Wiley Online Library on [12/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

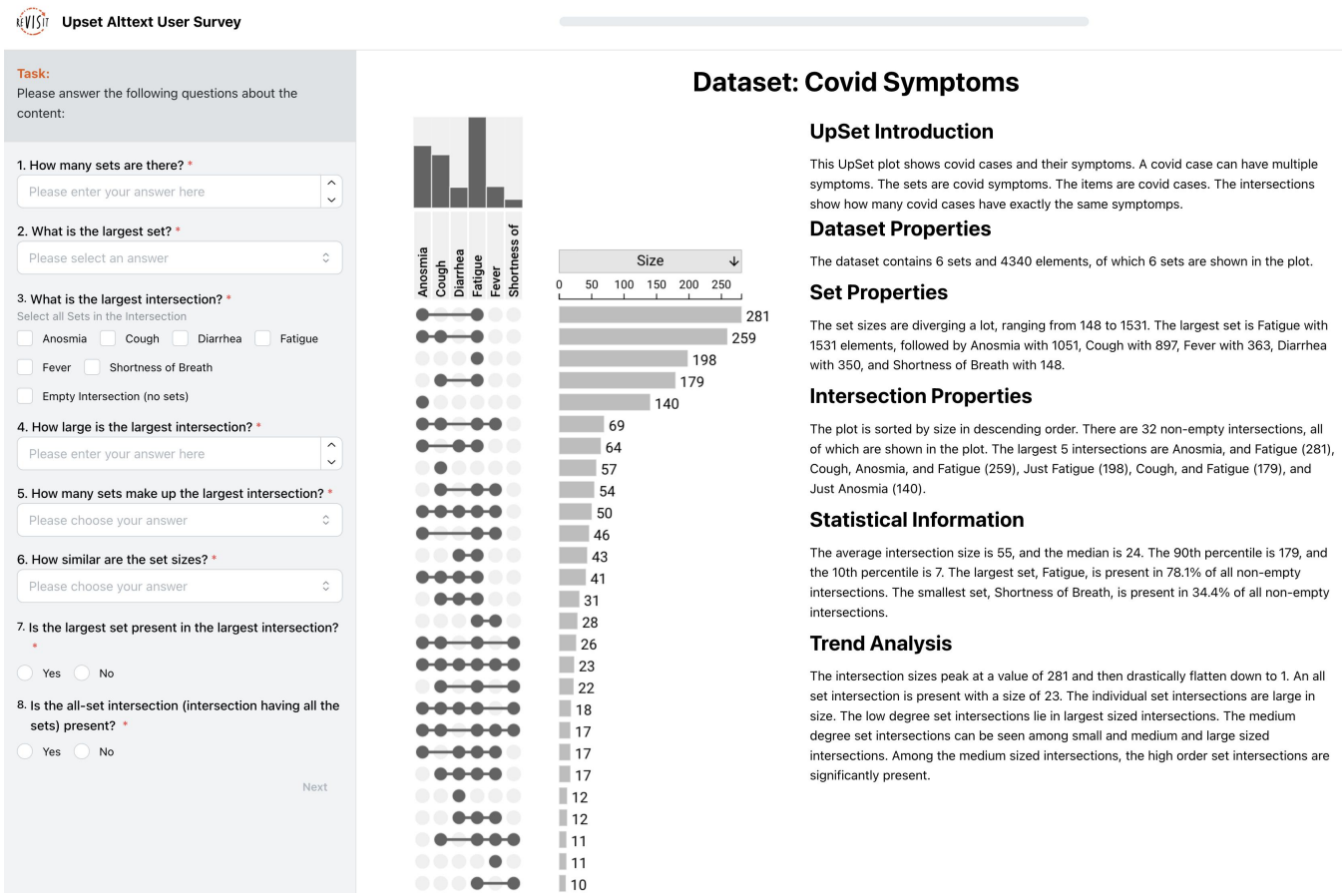


Figure 5: *Example of the study interface for the* BOTH *condition. The title (right) is human-generated, the introduction leverages a human-provided specification. See for more.*

Correctness. To evaluate whether participants were more correct in one condition over the others (H_{accuracy}), we ran a mixed model. Correctness (reported as a percentage) was regressed onto all three conditions. Across conditions, approximately 6% of the variance was explained by between-subject variability ($ICC = 0.06$). Thus, we included a random intercept by participant to account for these inter-individual differences. These correctness results are summarized in Figure 4.Correctness in the VIS condition was significantly ($p = 0.01$, $BF_{01} = 0.86$ less than in BOTH. As frequentist statistics cannot provide evidence in favor of the null hypothesis, we also report Bayes Factors (BF_{01}). BF_{01} assesses how strongly the evidence supports the null hypothesis (H_0) compared to the alternative hypothesis (H_1). A BF_{01} greater than 3 is considered strong evidence for the null hypothesis and A BF_{01} less than 0.3 is considered strong evidence for the alternative hypothesis. However, correctness in text-only did not significantly differ from either other condition (BOTH $p = 0.31$, $BF_{01} = 12.62$; VIS $p = 0.14$, $BF_{01} = 8.33$). While this rejects H_{accuracy} the interpretation is relatively straightforward: having two different tools for analysis improves the results. SP_{29} explained the benefits of using the two modes in tandem: "The combination of the text data and the visual graph is the most easy to understand in my opinion providing most of the answers with a cursory glance but also providing the other answers with a bit deeper of a look." This seems to suggest that having descriptions is useful for sighted people as well as for their expected BLV audience (which we explore in Sec. 4). Following prior works [LS22, PO12], we suggest that this is a curb-cut analogous to how video subtitles are useful for more than just those with a language barrier.

More closely related to the core of our study, the equivalence of TEXT and VIS for this measure suggests that **our descriptions are, generally, at least as good as the visual** which is a key result for this study.

Next, to examine whether the correctness across conditions were driven by a particular question, we added the eight questions (Q1-Q8) to the mixed regression. The questions were effect coded so that the regression coefficients reflect comparisons to the overall average correctness under each condition. A likelihood ratio test reveals that adding the questions as a predictor to the model significantly improves the model's fit ($\chi^2(21) = 207.48$, $p < 0.01$). This highlights that ability with a given condition is *not* linked to the type of question being asked, and is instead based on the difficulty of the question itself. This again supports a conclusion that TEXT and VIS are functionally equivalent.

© 2025 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

14678659, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cgf.70102 by Ishrat Jahan Eliza - University Of Utah J. Willard Marriott Library , Wiley Online Library on [12/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

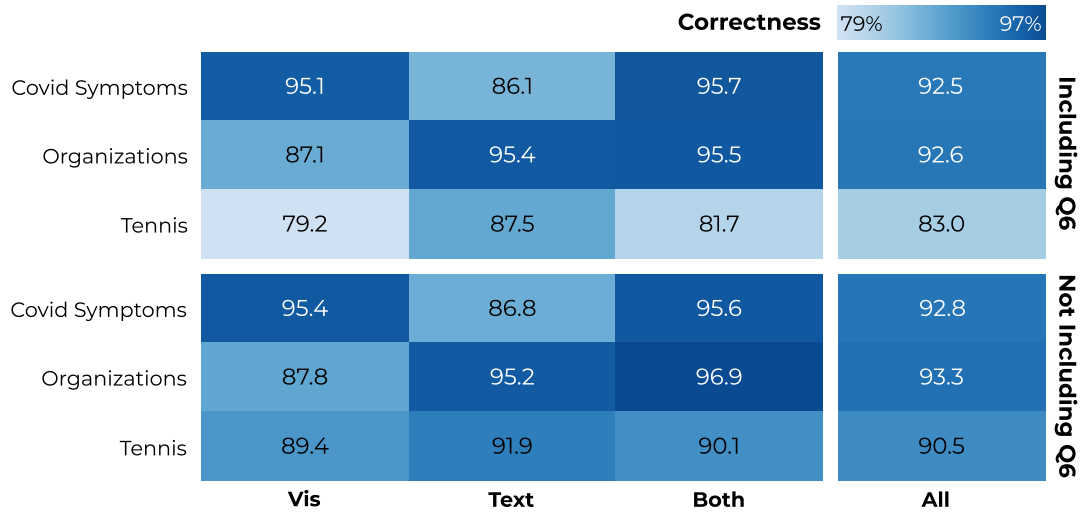


Figure 6: *Percent of questions correctness by dataset and condition. Q6, which focuses on qualitative distribution similarity, had a high influence on the average correctness by dataset.*

We evaluated several different factors to verify our results. Self-reported visualization **expertise** (using a 1-5 Likert scale) and correctness were uncorrelated (p=0.208), suggesting that this finding is broadly applicable. **Presentation** order did not significantly effect correctness. Via a regression analysis, we found that across all conditions, correctness improved by 1% on average from first block to second block (p=0.55, BF_{01} = 19.68) and $\sim$ 2% from second block to third (p=0.27, BF_{01}=11.02).

Finally, we found that **dataset** had substantial effect on correctness. In Figure 6 we see that participants performed significantly worse with the Tennis dataset across all conditions, and especially so in the conditions that showed a visualization, compared to the other datasets. This difference is caused mainly by Q6 (How similar are the set sizes) which had answers options "roughly equal", "diverging a bit", "diverging substantially" (see also Q6 in Figure 4). The tennis dataset is unique with respect to this question, because the answer for the size of sets is different from the answer for the size of the intersections. In contrast, for the other datasets the answer is (coincidentally) the same. We speculate that some participants mistakenly answered for intersections when answering based on the plot, which led to a larger number of wrong answers. This would seem to suggest a limitation of UpSet plots. As both sets and intersections can be interpreted as being on an equal visual footing, it can be easy to confuse them. At the same time, this potential for confusion is not present in the text description, likely leading to better responses. Overall, this suggests that UpSet plots are more effective in cases where that ambiguity is not relevant to the intended message or analyses.

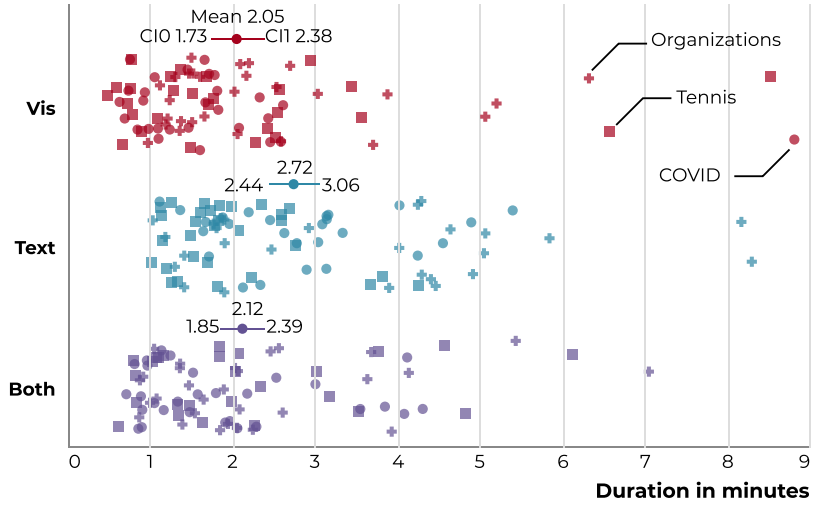


Figure 7: *Time to completion by section and condition. Participants take longer to respond in the text condition.*

© 2025 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

14678659, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cgf.70102 by Ishrat Jahan Eliza - University Of Utah J. Willard Marriott Library , Wiley Online Library on [12/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Completion Time. Next we consider the effect of condition on completion time. We again use a mixed effects model in which completion time (in minutes) was regressed onto the three conditions, with a random intercept by participant to account for high between-subject variability (*ICC* = 0.60). Per Figure 9, the average completion time was about 2 minutes across conditions. TEXT took slightly longer—averaging 2 minutes 45 seconds vs VIS's 2 minutes 3 seconds—which was a significant divergence from the other conditions (TEXT vs. VIS $p < 0.01$, $BF_{01} < 0.01$; TEXT vs. BOTH $p < 0.01$, $BF_{01} < 0.01$). The other conditions did not significantly differ (VIS vs. BOTH $p = 0.34$, $BF_{01} = 16.82$). SP_{55} summarized that the text was "easy to understand and read although missing the visual component takes away from the speed you are able to understand the information." Similarly, SP_{13} noted "Taking a little more time to crosscheck visualization with text gives greater (confidence in) accuracy and it is still very fast to parse the data." These results broadly agree with our hypothesis (H_{time}), however taking an additional 40 average seconds does not seem like an unduly high burden for a more linear medium. Prasad et al. [PO12] find for simple datasets and charts, text leads to more correct answers, but slower responses than charts or tables. Our results generally agree with this finding.

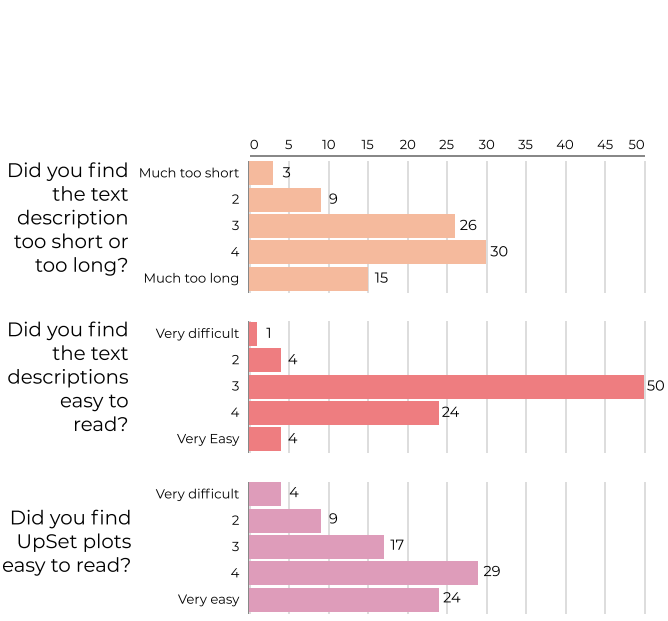


Figure 8: *Espoused preferences from survey respondents in the post-experiment survey.*

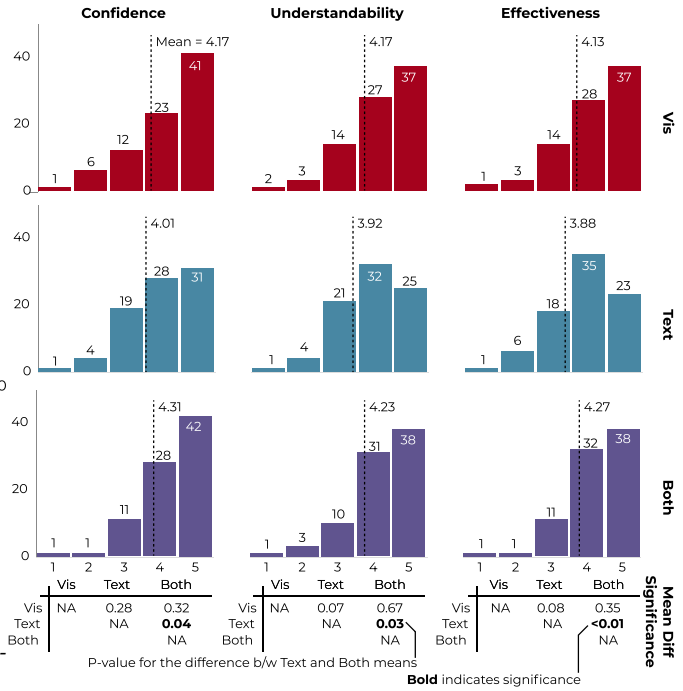


Figure 9: *Survey respondent reported confidence, understandability, and effectiveness for each of the three conditions via 1-5 Likert items (5 being high). Top rows show histogram by count, bottom shows significance of means.*

Preferences. Next, we consider participants' espoused opinions about each of the conditions. For participants ratings of confidence, understandability, and effectiveness we regress onto the three conditions (Figure 9). We find that the only significant differences among these preferences are, that for all three measures, the BOTH condition is preferred over text—echoing our correctness findings. Similarly, participants strongly preferred BOTH over the other two; 57/83 participants preferred BOTH. SP_{49} commented "Having both text description and data visualization combined works the best since the people who can read UpSet plots will use the plot to extract information, but the people who do not understand UpSet plots [...] can use the text description to help them get information about the dataset." See also Figure 8.

Connecting preference with correctness, we suggest that the lower on average correctness for VIS can be explained by these charts being unfamiliar, leading to a lack of confidence in their answers (compared to BOTH). For instance, SP_{20} noted "As this is a new way of interpreting data, I am not as confident without the written explanation alongside to support my understanding." Participants did not find UpSets to be easy (nor hard) to read, rating them 3.72 ± 1.14 out 5, with 5 being "very easy". SP_{51} explained that "some things were simple to see, such as number of sets and intersections. However I am not generally confident relying on just visualization. Although this may be because this method of displaying data is new to me." This, again, highlights the value of descriptions for all, and not just BLV users. These results partially reject $H_{\text{preference}}$ as participants seem relatively indifferent between TEXT and VIS compared to their preference for BOTH. As with correctness, this preference seems natural as BOTH offers best-of-both affordances.

Types of Data Insights. Next we consider if condition effected the types of insights (here meaning broadly facts about the data) that par-

© 2025 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

14678659, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cgf.70102 by Ishrat Jahan Eliza - University Of Utah J. Willard Marriott Library , Wiley Online Library on [12/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

ticipants had. We used a deductive coding scheme [Kin98] to code the 348 non-blank free-text responses. We met iteratively to understand the codes. We focus on those tagged *high level insight* or *low level insight*. High-level covers aspects related to the dataset. For instance SP_{43} observed, "Higher probability of being a member for Interpol, UN, UNESCO, UPU and WHO. Nobody has been a member for all organizations" concerning the Organization dataset. Low-level includes elements like counts or basic comparisons, such as in SP_{53}'s note that "fatigue is the most common symptom" for the Covid dataset. We count these insights by condition in Fig. 10. This is a relatively low sample size, and many observations in the set were not data related. Yet, it appears from this sample that lower-level commentary is more frequent, and that the BOTH had 1.6x more high-level comments than TEXT. This rejects H_{insights} as the types of insights appear to be roughly equivalent between TEXT and VIS.

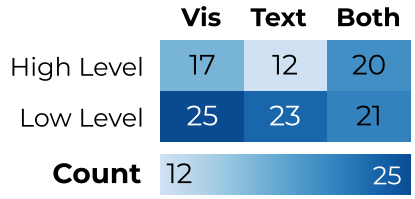


Figure 10: *Insight found condition and type of insight.*

Feedback. While participants generally found the descriptions to be informative, there was still room for improvement. SP_9 observed that "I think the text description could benefit from more formatting. It's difficult to parse the long sentences." Similarly, SP_{14} suggested that "the text description were so procedural in their style that it was difficult to tell what information it was trying to give me with the data. It seemed with that much text that there would be more explanation of the actual nature of the data." Not everyone felt it was hard to parse TEXT. For instance, SP_{44} commented that "the headings are split up nicely, so finding the section you want to look at is quite easy." SP_{30} observed that there was a slight learning curve to the text, noting "after understanding how to read the information, I found it very simple to find exactly what I needed in a short amount of time to answer the questions". Similarly, on average participants were neutral about the readability of the descriptions rating 3.54 ± 1.02 out of 5, as well as its length (averaging 3.31 ± 0.70 with 5 being too long and 1 being too short). This diversity of opinions highlights that description personalization may be valuable: what is useful for one person may not work for another (echoing Jones et al. [JPPH*24]). SP_{66} emphasized this point by noting how reformatting could improve accessibility generally: "I think overall maybe having less text and more bullet points would make it easier and quicker to get the gist of what's being depicted (especially for folks who have ADHD or other learning disabilities like myself.)" We note that evaluating description usability for people with other kinds of disabilities is valuable future work.

Appendix D: Contextualizing Experiment 2: Are our generations better than a default replacement?

Here we consider how well our bespoke descriptions improve over a naive replacement description. As there has not been previously been work on captions for this chart form, there is not a existing baseline by which to compare. Instead, we use LLMs to generate text, the quality of which we explore through an ablation study.

We find that while LLMs can produce high-quality text descriptions (sometimes of a more narrative format), they have non-trivial variance between generations (yielding inconsistent quality) that can contain false statements. The results are also dependent on the presence of related data in the models' training set. While others [DCBD24] have previously used LLMs to successfully generate descriptive texts, we find these successes come with caveats. We do not settle whether these risks are worthwhile, but find that our approach offers a consistent and reasonable quality compared to descriptions generated by LLMs.

Procedure

To answer the question how LLMs compare to our approach, we ran two comparison studies. First we ran an ablation study in which we explored the effect of several components of the generation—but kept the model and dataset fixed. We refer to runs in this study by $\mathbf{AB}_X$. Second, we ran a "best effort" study where we varied model and dataset with the best performing combination of factors from the ablation study. We refer to runs in this study by $\mathbf{BE}_X$. All studies were performed in late summer 2024. For generations the model temperature was set to zero.

Setup. For the **Ablation study** we ablated across the following components: *Image* (including an image of the UpSet plot in question), *Accessible Processed Data* (including the data being analyzed), *Example Descriptive Text* (including an example text description we created), *Strong prompt* (includes both a role play prompt focused on the creator of UpSet plots as well as guidance on effective text generation), and *Pattern prompt* (prompt summarizing the data patterns from Sec. 3.1). For the **Best** Effort study, we used an "all on" prompt strategy. We vary models between Claude 3.5 Sonnet and OpenAI's GPT-4o. These models were selected because they offer good performance [Kan24], are widely available, and can incorporate both text and images (within their respective ecosystems). See Figure 11 for model use by run. We

© 2025 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

14678659, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cgf.70102 by Ishrat Jahan Eliza - University Of Utah J. Willard Marriott Library , Wiley Online Library on [12/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

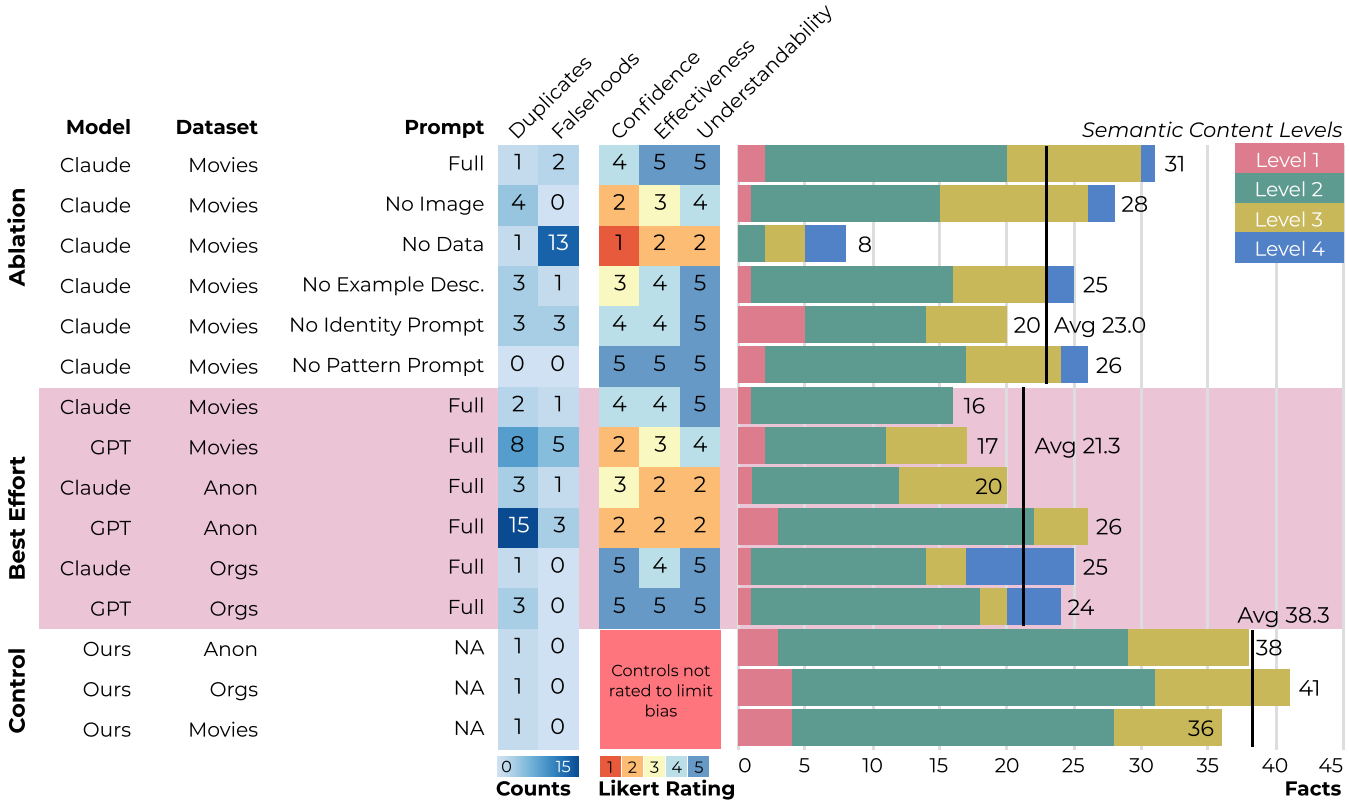


Figure 11: *A summary of our LLM study showing study instance by the number of facts generated for that description, sorted by Lundgard et al.'s [LS22] semantic content levels. We also highlight the number of hallucinated and duplicated facts in the generation—although, as they were identified manually, these values are lower bounds, as there might be additional subtleties to the data not analyzed. Claude here is Anthropic's Claude 3.5 Sonnet and GPT is OpenAI's GPT-4o. N.b. the inputs for the Best Effort Movies run and the Ablation Claude Movies were identical.*

vary datasets between Organizations, Movies, and an anonymized version of Movies called Anon-Movies (i.e. the same data with set names replaced with "set_1" and so forth). We are interested in the difference between Organizations and Movies because they have very different data distributions (as described in Figure 1), as well as the difference between Anon-Movies and Movies for the effect of context that can be inferred.

Evaluation. We took two approaches to comparing the LLM generated text and with ours: coding and expert review. First, we had two coders independently code the result of the ablation study. They counted the number of correct and incorrect facts in each generation and reviewed the content present using Lundgard et al.'s [LS22] semantic levels framework. After coding they met to discuss and resolve ambiguities, after which they adjusted their original codes. With this more robust coding model in hand the first coder then coded the best effort study independently. An example of a coding of one our generated text is shown in Fig. 3. We do not code for our identified data patterns because they are part of some prompts. Next, we had our blind collaborator, Dr. X, evaluate the LLM generated responses across the same metrics used in Fig. 9, as well as offer any qualitative commentary he had. We quote him "like so".

Limitations. Our evaluation of the LLM generated descriptions via grading may not reflect the way in which real users perform with them. This analysis was primarily centered on demonstrating that our results are not trivially replaced by an LLM call rather than the comparative effectiveness—but having users evaluate these properties is interesting future work. Our grading of the generated descriptions may have been biased. We tried to limit this effect by having two coders develop a code book, however these efforts will inevitably be incomplete. Similarly our blind collaborator's rating of the generated texts may be biased, but given the range of his ratings we feel this is unlikely. Finally, the quality of the LLM results could be improved through additional prompt engineering. The goal of this evaluation was to see if a simple or straightforward usage would be better than our handcrafted descriptions, but integrating LLMs more deeply into text generation is intriguing future work.

Results

Here we explore the results of our LLM study. Figure 11 shows a summary of our results.

© 2025 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

14678659, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cgf.70102 by Ishrat Jahan Eliza - University Of Utah J. Willard Marriott Library , Wiley Online Library on [12/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Text Content Differences. On average our descriptive text contained a substantially greater number of facts. Interestingly, our texts tended to be shorter (avg 290 words), compared to the ablation runs (avg 325) and best effort runs (avg 384)—this is unsurprising given the tendency for LLMs to be verbose. However, measuring quality through text length is noisy as LLM responses can be made more verbose or concise via prompt engineering (of which we did little, beyond the coarse component-level exploration in our ablation study). Compared to the texts we designed together, Dr. X liked "metaphorical language, such as a 'long tail distribution'" that was present in some LLM texts, noting of $\mathbf{AB}_{\textit{No Pattern}}$ that "It's the kind of alt text I'd expect to see on a news website."

However, some runs also had substantial numbers of falsehoods, some of them were obvious (and caught in our coding) while others of them might be subtler. For instance Dr. X worried that the LLM had "cherry picked certain interesting elements it wanted to highlight" or "misinterpreted" certain data. Overall GPT seemed to do worse in this regard, as it more frequently generated duplicate and false information—although it could be tuned to perform better.

Finally, the generated texts have some variance. $\mathbf{AB}_{\textit{Full}}$ and $\mathbf{BE}_{\textit{Claude-Movies}}$ were relatively different (e.g., the volume of Level 4 content) but had the same inputs. If something is to be relied upon for accessibility, its value may be limited if it changes day to day.

Knowledge is Key. The presence of data and that data being familiar seems to be an essential part of making effective LLM texts. For instance, $\mathbf{AB}_{\textit{No data}}$ generated a substantially lower number of facts and was qualitatively worse for Dr. X: "I can't build a mental picture of what the visualization might look like, or what story the data tells me." This suggests that including the dataset is necessary for providing accurate information. In addition, that data seems to need to have human-readable labels. While there was relatively little difference between the types of semantic content generated in the Anon Movies and Movies runs, the effect of reading those texts was starkly different. Dr. X observed that $\mathbf{BE}_{\textit{*-Anon Movies}}$ were "hard to follow" and "hard work to get something sensible out of it."

The LLM generated texts include some amount of Level 4 semantic content, which we intentionally excluded from our design. Notably, the $\mathbf{BE}_{\textit{*-Orgs}}$ runs included substantially more Level 4 content than compared to those using other data. For example, $\mathbf{BE}_{\textit{Claude-Orgs}}$ included observations like "`This underscores the tendency for nations to participate in multiple international bodies rather than isolating themselves to a single organization.`" This is likely because news articles and other information about organizations like the WHO are present in their training data. Dr. X specifically liked the amount of level 4 content in the Orgs runs because it provided specific examples to back up claims. Evidently, if the model being used has knowledge of the dataset under study, then it can usefully provide contextualizing information. However, this can be risky as it is not clear when a model will or will not know about a dataset.

Appendix E: Study Instruments

Here we show simplified versions of the study instruments used in the survey and interview studies

Survey Study

An anonymized version of our survey is available at . This includes a full copy of the instructions participants were shown and supports browsing each of the study conditions.

Interview Study

Demographic Information

1. What is your date of birth?
2. What is your gender?
3. What is the highest level of education you have completed or are in the process of completing?
4. Which screen reader do you use with your computer or other devices (e.g., NVDA, JAWS, VoiceOver, etc.)
5. How long have you been using a screen reader?
6. Do you use other accessibility devices or software in combination with a screen reader, such as screen magnification or a braille display? If yes, please describe.
7. What is your preferred rate of speech when using a screen reader?
8. How many hours are you on a computer each day?
9. Would you consider your career to be data-intensive or numbers-driven (e.g., regularly work with large datasets, perform statistical analyses, or make decisions based on quantitative information)?
10. How often do you interact with data visualizations, such as those for work, from news articles, in video games, etc.? And in what context?
11. How would you describe your vision-loss level?
12. How would you describe your vision level (e.g., no remaining vision, light perception, central vision, etc.)?
13. What is your corrected visual acuity in either Snellen (e.g., 20/200) or LogMAR (e.g., 1.3)?

© 2025 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

14678659, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cgf.70102 by Ishrat Jahan Eliza - University Of Utah J. Willard Marriott Library , Wiley Online Library on [12/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

UpSet Alt-Text Questions

1. Can you describe what you learned about the dataset in your own words?
2. What is the dataset about?
3. How many sets are shown?
4. What is the largest intersection?
5. How would you describe this dataset to a friend?
6. How did this dataset increase your understanding of COVID symptoms?
7. Do you feel like you have a good sense of the dataset? Why, why not?
8. What was difficult for you to understand about the dataset?
9. What would have been helpful to provide additionally? Was anything missing?
10. Do you have comments on the style of the text description? For example, was it too long, too short, too verbose?
11. Do you have any other feedback or comments that we did not touch on today?
12. In your opinion, what differentiates a great text description from other descriptions?
13. Do you have any experiences with accessible data visualizations? If so, please elaborate.

UpSet Plot Questions

1. Does your mental model from the text description match the chart?
2. Do you have any other feedback or comments now that you have seen the corresponding visualization?

Interview Stimuli

Page 1

Introduction

For this study, you will see a text description generated for an UpSet plot. An UpSet plot is a set visualization technique similar to Venn diagrams, but unlike Venn diagrams, UpSet works for more than three sets.

Our research aims to make data visualizations more accessible to people with visual impairments. We want to understand whether text can convey similar amounts of information as a chart. We first will introduce what an UpSet plot is, and how to interpret data from the plot. UpSet Explained

UpSet plots the intersections of sets in a table. Each column corresponds to a set. Bars at the top of the columns show the size of the sets. The row corresponds to an intersection: marks in the cells show which set is included in the intersection. The number of sets that participate in the intersection is referred to as degree. If there is no mark in any of the cells, then it is the intersection of no set, which is also referred to as the empty intersection, with a degree of 0. If there is a mark in every cell of a row, then it is the intersection of all sets.

UpSet plots the size of the intersections as bar charts to the right of the table. The table is also useful because it can be sorted in various ways. A common way is to sort by size, but it's also possible to sort by degree or sets.

Imagine an UpSet plot that shows movie data. Movies have genres like Drama, Comedy, Thriller, Mystery, or Crime. A movie can have a single genre, or it can have multiple genres. In this example, the genres are the sets. Some sets are bigger: there are more Drama movies than Mystery movies, for example. And some intersections will be more common: Thriller and Mystery might be a popular combination, while the combination of Drama, Comedy, and Thriller might be rare. Glossary of Terms

Here are some terms we use in the text description:

Movies that don't fall into any genre are an intersection of no set/the empty intersection. A row with only one mark for drama (movies that are just dramas and have no other genre) is an independent set intersection. A row with 2-3 marks like "Drama-Comedy" corresponds to a low-degree set intersection. A row with 3-5 marks (e.g., Mystery-Crime-Thriller) corresponds to a medium-degree set intersection. A row with even more marks is a high-order set intersection. The last case is that set containing all sets (i.e. all movie genres), which we call an all-set intersection.

That's it for the introduction! Please stop now and ask the interviewers if you might have any clarifying questions.

Next, we'll explore text descriptions generated for a variety of UpSet Plots. Click on the link below to see: UpSet Plot Description.

```
# Visualizing co-occurrence of CoVID 19 Symptoms with UpSet.

This is an UpSet plot that shows covid cases and their symptoms. A covid case can have multiple symptoms
    . The sets are covid symptoms. The items are covid cases. The intersections show how many covid
    cases have exactly the same symptoms. The plot shows intersections of 6 sets. All major
    intersections involve the set Fatigue, and Cough. The largest intersection is Anosmia, and Fatigue,
    with 281 elements. Other large intersections also involve Cough, Anosmia, and Fatigue. The
    intersection of all sets is present with 23 elements.
```

© 2025 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

14678659, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cgf.70102 by Ishrat Jahan Eliza - University Of Utah J. Willard Marriott Library , Wiley Online Library on [12/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

```
# Dataset Properties

The dataset contains 6 sets and 4340 elements, of which 6 sets are shown in the plot.

# Set Properties

The set sizes are diverging a lot, ranging from 148 to 1531. The largest set is Fatigue with 1531
    elements, followed by Anosmia with 1051, Cough with 897, Fever with 363, Diarrhea with 350, and
    Shortness of Breath with 148.

# Intersection Properties

The plot is sorted by size in descending order. There are 32 non-empty intersections, all of which are
    shown in the plot. The largest 5 intersections are Anosmia, and Fatigue (281), Cough, Anosmia, and
    Fatigue (259), Just Fatigue (198), Cough, and Fatigue (179), and Just Anosmia (140).
Statistical Information

The average intersection size is 55, and the median is 24. The 90th percentile is 179, and the 10th
    percentile is 7. The largest set, Fatigue, is present in 78.1

# Trend Analysis

The intersection sizes peak at a value of 281 and then drastically flatten down to 1. An all set
    intersection is present with a size of 23. The individual set intersections are large in size. The
    low degree set intersections lie in the largest sized intersections. The medium degree set
    intersections can be seen among small and medium and large sized intersections. Among the medium
    sized intersections, the high order set intersections are significantly present.
```

Appendix F: Text Description Examples

Here we provide examples of the text descriptions used in our studies for the Movies dataset as an example of general trend.

Movies

Short Description no Configuration

```
This is an UpSet plot which shows the intersections of 6 sets. All major intersections involve the set
    Action, and Adventure. The largest intersection is Thriller, and Action, with 104 elements. Other
    large intersections also involve Action and Thriller.
```

Long Description with no Configuration

```
# UpSet Introduction
This is an UpSet plot that visualizes set intersection. To learn about UpSet plots, visit REDACTED.

# Dataset Properties
The dataset contains 17 sets and 6303 elements, of which 6 sets are shown in the plot.

# Set Properties
The set sizes are diverging a lot, ranging from 68 to 503. The largest set is Action with 503 elements,
    followed by Thriller with 492, Adventure with 283, Children with 251, War with 143, and Western with
     68.

# Intersection Properties
The plot is sorted by size in descending order. There are 28 non-empty intersections, all of which are
    shown in the plot. The largest 5 intersections are Just the empty inter (2569), Just Thriller (349),
     Just Action (218), Just Children (160), and Thriller, and Action (104).

# Statistical Information
The average intersection size is 138, and the median is 7. The 90th percentile is 218, and the 10th
    percentile is 1. The largest set, Action, is present in 50.0\% of all non-empty intersections. The
    smallest set, Western, is present in 25.0\% of all non-empty intersections.
```

© 2025 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

14678659, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cgf.70102 by Ishrat Jahan Eliza - University Of Utah J. Willard Marriott Library , Wiley Online Library on [12/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

```
# Trend Analysis
The intersection sizes peak at a value of 2569 and then drastically flatten down to 1. Just the empty inter is the largest by a factor of 7. The empty intersection is present with a size of 2569. An all set intersection is not present. The individual set intersections are large in size. The low degree set intersections lie in small and medium sized intersections. The medium degree set intersections can be seen among medium sized intersections. No high order intersections are present.
```

Long Description with Configuration

```
# UpSet Introduction
This is an UpSet plot that visualizes set intersection. To learn about UpSet plots, visit REDACTED.

# Dataset Properties
The dataset shows attributes of movie genres and ratings. The dataset contains 17 sets and 6303 elements, of which 6 sets are shown in the plot.

# Set Properties
The set sizes are diverging a lot, ranging from 68 to 503. The largest set is Action with 503 movies, followed by Thriller with 492, Adventure with 283, Children with 251, War with 143, and Western with 68.

# Intersection Properties
The plot is sorted by size in descending order. There are 28 non-empty intersections, all of which are shown in the plot. The largest 5 intersections are Just the empty inter (2569), Just Thriller (349), Just Action (218), Just Children (160), and Thriller, and Action (104).

# Statistical Information
The average intersection size is 138, and the median is 7. The 90th percentile is 218, and the 10th percentile is 1. The largest set, Action, is present in 50.0

# Trend Analysis
The intersection sizes peak at a value of 2569 and then drastically flatten down to 1. Just the empty inter is the largest by a factor of 7. The empty intersection is present with a size of 2569. An all set intersection is not present. The individual set intersections are large in size. The low degree set intersections lie in small and medium sized intersections. The medium degree set intersections can be seen among medium sized intersections. No high order intersections are present.
```

Appendix G: Additional Figures

Finally, we include figures which that relevant to this work but were out of place in the main text. Figure 12 shows an example of our descriptive text integrated with an UpSet plot in a notebook. Figure 13 is a snapshot of our coding of in-the-wild UpSet examples. Figure 14 shows the final version of the text description, refined based on the interview results.

© 2025 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

14678659, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cgf.70102 by Ishrat Jahan Eliza - University Of Utah J. Willard Marriott Library , Wiley Online Library on [12/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

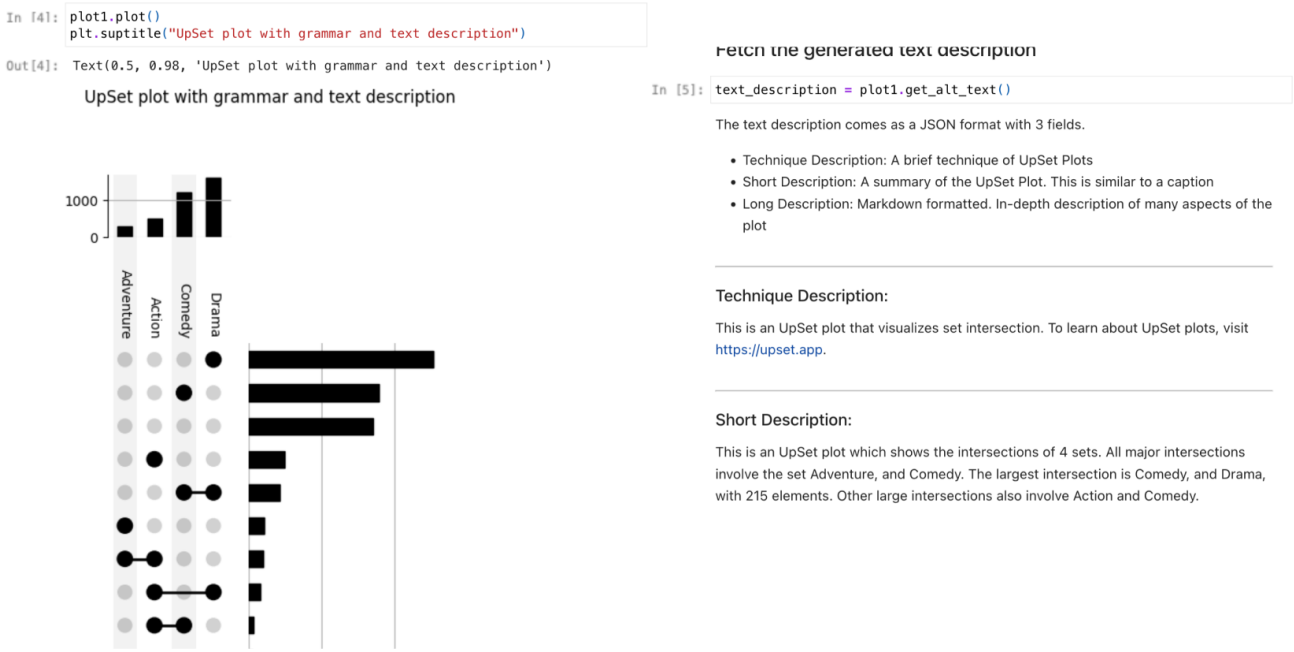


Figure 12: *A screen shot of a notebook using our automated text generation system.*

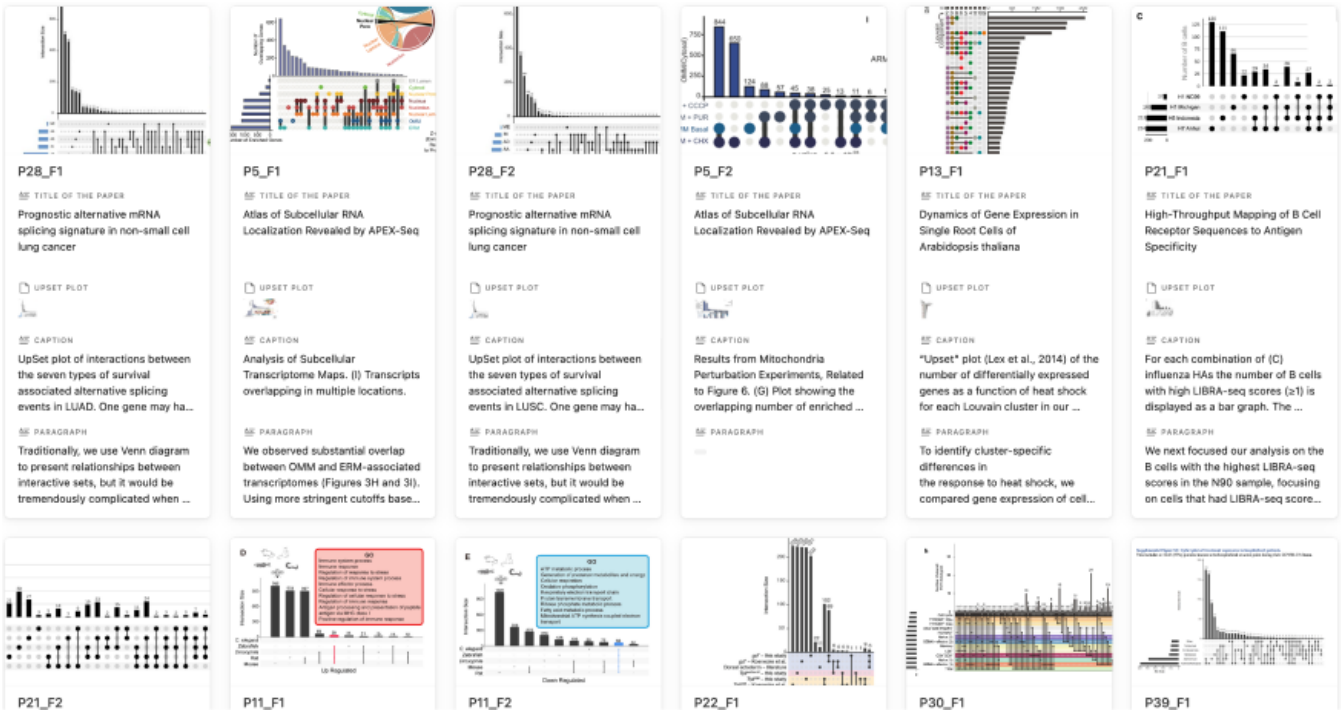
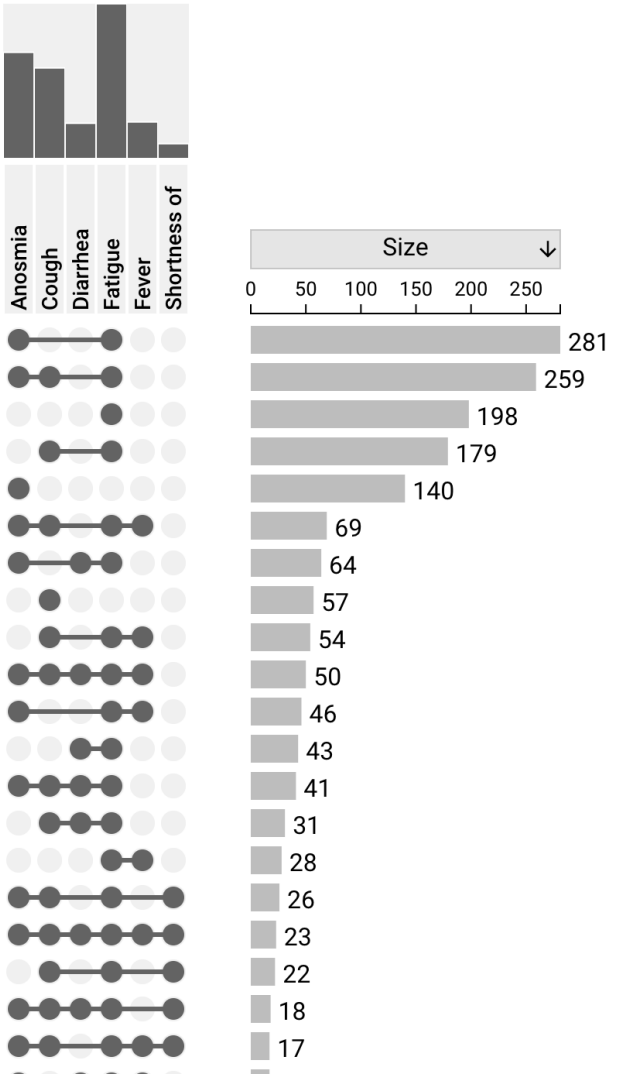


Figure 13: *Selected thumbnails of UpSet plots collected for our classification of patterns found in UpSet plots.*

© 2025 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

14678659, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cgf.70102 by Ishrat Jahan Eliza - University Of Utah J. Willard Marriott Library , Wiley Online Library on [12/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License



Statistical Information

- The average intersection size is 55, and the median is 24.
- The 90th percentile is 179, and the 10th percentile is 7
- The largest set, Fatigue, is present in 78.1% of all non-empty intersections
- The smallest set, Shortness of Breath, is present in 34.4% of all non-empty intersections.

Trend Analysis

- The intersection sizes peak at a value of 281 and then rapidly flatten down to 1.
- An all set intersection is present with a size of 23
- The individual set intersections are large in size
- The low degree set intersections are in the largest sized intersections
- The medium degree set intersections can be seen among small and medium and large sized intersections
- Among the medium sized intersections, the high order set intersections are significantly present.

Glossary

- Elements: elements are discrete entities, such as people, movies, countries. They are often the rows in tables, and we record attributes about them.
- Set: a set is an attribute of an element that expresses a “part-of”, “is-a”, “has-a” relationship, and so on. For example, a person can be in a club (a part-of relationship), or a car can be of a type, such as a convertible (an is-a relationship). They are often columns in a table.

Figure 14: *The final version of the text description, refined based on feedback received during the interviews. While the text itself is unchanged, individual statements are listed as bullet points, and a glossary is added to the end of the description.*

© 2025 The Author(s).
Computer Graphics Forum published by Eurographics and John Wiley & Sons Ltd.

14678659, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/cgf.70102 by Ishrat Jahan Eliza - University Of Utah J. Willard Marriott Library , Wiley Online Library on [12/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License