

MATCHA: LLM-Based Chart Summarization for Screen Reader Users

Lauren C. Kahn¹, Jaime Soto², Om Arya³, Audrey Michal⁴, Steven R. Wilson⁴

¹University of Texas at San Antonio, ²University of Illinois Urbana-Champaign,

³University of Maryland, College Park, ⁴University of Michigan-Flint

claurenedu@gmail.com, jaisoto00@gmail.com, om.arya0577@gmail.com,
almichal@umich.edu, steverw@umich.edu

Abstract

Online charts are ubiquitous, yet remain systematically inaccessible to screen reader users. This paper presents the process for developing the MATCHA Chart Summarizer (MCS), an interactive, browser-based system, leveraging multimodal LLM’s to generate accessible spoken chart summaries and support natural-language querying. MCS aims to enhance functional comprehension and narrow the reasoning gap for screen reader users when interacting with data visualizations. Through a three-stage mixed-methods, user-based evaluation ($N = 126$ total), we conduct a needs assessment, design our prototype, and demonstrate that LLM-generated summaries can improve task accuracy and increase user confidence, informativeness, and usability. We also quantify how visualization flaws, such as missing labels or poor contrast, induce semantic drift in generated chart summaries. We contribute an accessible assistive NLP system, an impact-oriented evaluation framework, and evidence-based design recommendations for automated chart summarization.

1 Introduction

Data visualizations are vital for interpreting complex information across science, policy, and daily life (Hjelle et al., 2024; Divecha et al., 2023; Gerela et al., 2022). Yet, they remain systematically inaccessible to screen reader users. Even when guidelines are met, charts often lack the semantic structure or interactive support necessary for meaningful comprehension (Power et al., 2012; Calvo et al., 2016; Fan et al., 2023), creating barriers to equitable participation in school, work, and civic discourse. This work aligns with NLP for Positive Impact goals by leveraging language technology to reduce information inequality (Karamolegkou et al., 2026).

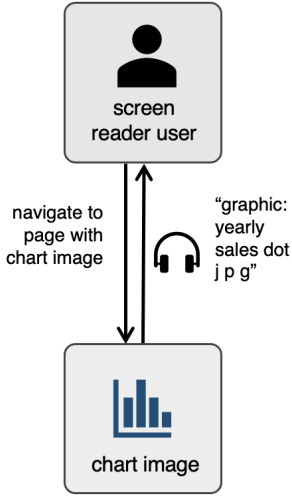
While recent NLP research has explored chart understanding (Bao et al., 2026; Kaur et al., 2026;

Xu et al., 2026), focus has primarily remained on technical capability rather than user-centered efficacy. We compare the MATCHA Chart Summarizer (MCS) against a static, fixed-prompt baseline that lacks interactivity and sensitivity to individual user needs. In contrast, MCS introduces two critical improvements: (1) adaptive granularity, allowing users to select between concise overviews and detailed breakdowns; and (2) interactive question-answering for probing specific trends. This bidirectional interaction transforms chart-reading from a passive, generic experience into an active, functional exploration, enabling users to verify graph comprehension accuracy. We focus on screen reader users self-identifying as low-vision (LV), neurodivergent (ND), or both. LV users often combine residual vision, magnification, and speech to recover visual structure, while ND users—including those with ADHD, dyslexia, or autism—rely on speech to manage attention or convert spatial information into linguistic forms (MacFarlane et al., 2010; Wu et al., 2021; Zhu et al., 2023). Consequently, a single static alt-text description cannot satisfy these diverse sensory, functional, usability requirements.

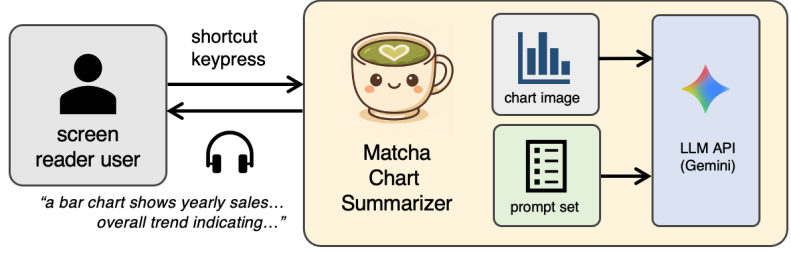
Alongside our user studies, we develop the MATCHA Chart Summarizer (MCS), a browser-based assistive tool that leverages multimodal models to generate real-time spoken chart summaries. MCS is designed to move beyond checkbox accessibility, which offers only minimal, static descriptions, toward a model of functional comprehension. The system supports dynamic summary styles and natural-language question answering, allowing users to request concise overviews, detailed numerical descriptions, or specific follow-up clarifications to support task accuracy and user confidence.

We evaluate our MCS prototype through both user-studies—measuring task accuracy, confidence, informativeness, usability, and adoption inter-

a default screen reader behavior



b descriptive summaries



c interactive chart Q&A

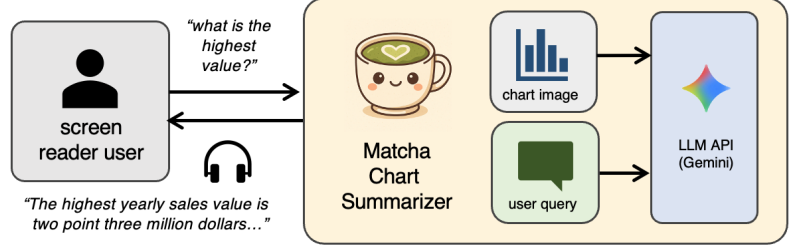


Figure 1: Comparison between (a) default screen reader behavior when interacting with data visualizations, (b) the MCS keypress-triggered chart summarization feature and (c) MCS-supported interactive speech-based chart Q&A.

est—and computational analysis. We also assess the semantic stability of model-generated summaries under chart design flaws. Both components are critical, as an assistive system may appear fluent while failing to maintain factual fidelity, potentially reproducing misinformation from poorly designed visualizations.

Our work addresses three research questions: **RQ1 (Impact on Access):** To what extent do optimized LLM-generated summaries and question answering improve screen reader users’ chart comprehension? We evaluate this in **Section 3.7** through our summative user study, with performance and usability metrics detailed in **Table 3**. **RQ2 (User-Centered Adaptation):** How do LV and ND users differ in preferred summary granularity, tone, and interaction style? We address this through our needs assessment (**Section 3.2**) and formative survey (**Section 3.4**), comparing distinct prompt styles and preferences in **Table 2**. **RQ3 (Robustness and Trust):** How do common visualization flaws affect the reliability of LLM-generated chart summaries? We evaluate this computationally in **Section 4**, quantifying semantic drift across design flaws in **Table 4**.

We find that a browser-based Vision-language model (VLM) system with interactive question-and-answering features allows for improved chart comprehension and perceived usability compared

to a baseline, fixed-prompt system. Certain informational flaws, such as missing axes labels, have the potential to significantly alter generated summaries compared to visual changes, such as contrast and aspect ratio, which VLM-based systems are fairly robust to.

We make the following contributions: (1) MCS¹, an interactive LLM-based browser extension for spoken chart summarization and question answering; (2) mixed-methods user-studies measuring whether generated language improves users’ ability to answer chart questions, confidence, perceived utility, and adoption interest; (3) an analysis of summary stability under accessibility and misleading-design flaws in real-world charts; and recommendations for adaptive, robust, and collaborative chart summarization systems for screen reader users, derived from our multistage user studies.

2 Related Work

2.1 Accessible Chart Understanding

Early chart-access systems relied on human-in-the-loop services, which offered high quality but faced scalability barriers (Bigham et al., 2010). Subsequent approaches transitioned to sonification and interactive natural-language interfaces, such as VoxLens (Sharif et al., 2022), SeeChart (Alam

¹The GitHub repository used for this work can be found here: <https://github.com/om-arya/matcha>

System	Primary Target	Technology	Tailored Tone	ND Focus
VoxLens (Sharif et al., 2022)	Blind / LV	JS / Sonification	No	No
SeeChart (Alam et al., 2023)	Blind / LV	Natural Language Interface	No	No
AutoVizuA11y (Duarte, 2024)	Screen Reader Users	LLM (GPT-3.5)	Limited	No
Alt4Blind (Jung et al., 2022)	Blind / LV	Intelligent Interface	No	No
AltChart (Thompson et al., 2023)	Blind / LV	VLM / Multi-pretext	No	No
Chart4Blind (Sharif and Ellis, 2021)	Blind / LV	Intelligent Conversion	No	No
TADA (Zhao et al., 2024)	Blind / LV	Node-Link Logic	No	No
Azimuth (Srinivasan et al., 2024)	BLV (Dashboards)	Heuristic / Rules	No	No
FigurA11y (Lundgard and Satyanarayan, 2022)	Authors / BLV	LLM (Drafting)	No	No
MATCHA (MCS)	LV & ND Users	Image-LLM	Yes	Yes

Table 1: Comparison of Chart Accessibility Systems

et al., 2023), TADA (Zhao et al., 2024), and Azimuth (Srinivasan et al., 2024). More recently, models like AutoVizuA11y, AltChart, Alt4Blind, and FigurA11y have leveraged large language and vision-language models to automate chart description (Duarte, 2024; Thompson et al., 2023; Jung et al., 2022; Lundgard and Satyanarayan, 2022).

While these systems demonstrate the potential of LLMs in accessibility, many evaluate descriptions primarily as static outputs rather than as functional interventions. Consequently, there is often a disconnect between model fluency and downstream user comprehension. Furthermore, existing research frequently insists on access to well-structured, high-quality chart data. In contrast, MCS extends this work by evaluating both measurable user impact and system robustness against flawed, real-world visualizations. A comparison between MCS and other accessible chart summarization systems can be found in (Table 1).

While chart summarization has progressed in English, models face significant performance gaps in multilingual contexts. For example, benchmarks like POLYCHARTQA demonstrate that visual-linguistic alignment remains a critical bottleneck (Xu et al., 2026). This necessitates human-centered approaches, such as the MATCHA project, which bridge the gap between complex data and accessible, intuitive summaries for all users.

2.2 Diverse Needs of Screen Reader Users

Screen reader users are not a monolithic population; accessibility needs vary across visual acuity, cognitive processing, task context, data literacy, and user preference (Wu et al., 2021). LV users often require detail sufficient to reconstruct com-

plex chart layouts, while ND users often benefit from summaries that reduce cognitive load, avoid “text walls,” and support incremental exploration (MacFarlane et al., 2010; Zhu et al., 2023). Despite these documented requirements, few chart interpretation systems explicitly allow for adaptive output. MCS directly addressed this by supporting dynamic, user-selected verbosity.

2.3 Robustness, Misinformation, and Social Impact

Visualization flaws act simultaneously as accessibility barriers and sources of misinformation. Common design oversights, such as missing axis labels, insufficient color contrast, chart clutter, truncated axes, inverted Y-axes, dual-axis scales, and biased or misleading titles, can fundamentally distort user interpretation (Cleveland, 1984; Cairo, 2016; Pandey et al., 2015). For screen reader users, these risks are significantly amplified; because these users cannot visually verify the chart’s underlying structure, they are often compelled to rely exclusively on the generated summary. Consequently, if an LLM generates a fluent, confident, yet factually incorrect summary or misleading visualization, the system may actively perpetuate or even exacerbate the harm caused by a deceptive chart design.

For this reason, assessing social impact in accessibility technology requires moving beyond traditional caption quality metrics. Accessible systems must be evaluated through a multidimensional lens: does the system measurably improve user outcomes, support user autonomy during complex data exploration, and crucially, does it avoid the overconfident transmission of misleading information?

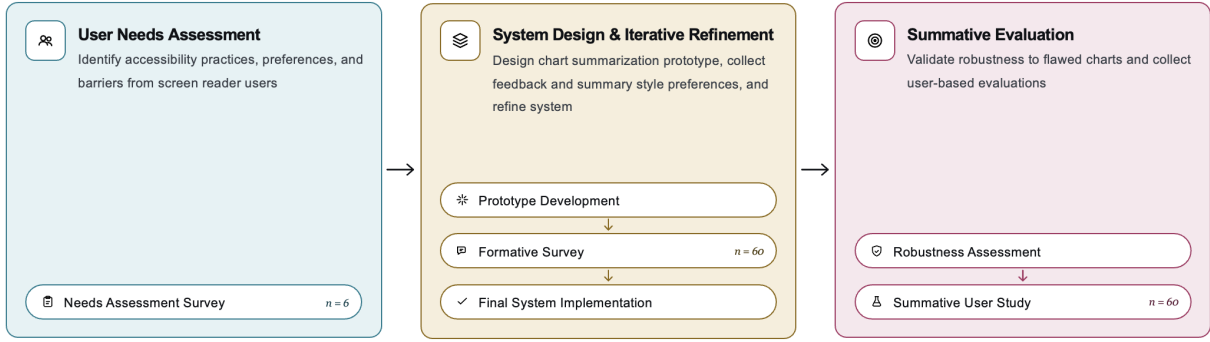


Figure 2: Overview of Methodology. Initial user needs assessment is used to inform the prototype Matcha Chart Summarizer design. The formative study evaluates user preferences and helps to refine the specific LLM prompts and features of the final system, which is evaluated based on both a technical assessment of robustness to flawed charts and a user study.

Our robustness analysis operationalizes one critical dimension of this broader impact question by testing for semantic stability. We define semantic stability as the degree to which an LLM-generated summary remains factually consistent when a chart is subjected to systematic, real-world flaws, such as the removal of a legend or the truncation of an axis. By quantifying summary drift in these scenarios, we move toward a more comprehensive evaluation framework that prioritizes the user’s right to accurate information over mere model fluency.

3 Iterative System Design

To ensure MCS was developed and validated in alignment with the lived experiences of potential users, we employed a three-stage mixed-methods design: (1) a qualitative needs assessment, (2) a formative survey, and (3) a summative experimental evaluation. See Figure 2.

3.1 Participants

Participants were recruited via the Prolific² platform with strict screening criteria requiring: (i) English language fluency, (ii) age 18 or older, (iii) active use of screen reader technology (e.g., NVDA, JAWS, VoiceOver), and (iv) self-identification as being low-vision (LV), neurodivergent (ND), or both. To ensure fair compensation, participants were remunerated at a rate of \$15 per hour, estimated based on the length required to complete each phase. To maintain data independence, participant pools for each of the three stages were recruited separately; no participant participated in more than one study phase.

²<https://www.prolific.com/>

3.2 Needs Assessment

First, we conducted semi-structured surveys with six participants who regularly use screen readers (three LV, three ND; note that this group is entirely distinct from the 60 participants who completed our later formative survey in Section 3.4) to understand current barriers to chart accessibility and to identify pain points in chart access and gather initial feature requirements. Participants reported on their habitual use of assistive technology, past chart interpretation experiences, accessibility barriers, and desired tool features. Additionally, participants performed pilot chart interpretation tasks using baseline VLM summaries and rated their answer confidence, chart understanding confidence, and ease of use on a four-point Likert scale. The full set of questions for the needs assessment are in Appendix A.

To evaluate chart comprehension, we developed a series of custom chart interpretation tasks based on a diverse selection of real-world visualizations. Rather than relying on standardized data sets, which may not reflect the accessibility barriers inherent in everyday web-based data, our research team designed the chart interpretation task to specifically probe functional understanding. These tasks required participants to derive key trends, compare data series, and identify potential visual distortions. Each question was iteratively refined by the research team to ensure clarity and target specific accessibility challenges, such as information density and axis-level chart interpretation.

Survey results revealed that accessibility barriers are often compounded by design choices; participants consistently identified low color contrast, illegible text sizing, ambiguous label placement,

and a lack of descriptive context as the primary drivers of chart abandonment. Among the various visualization formats, bar and line charts were the most frequently encountered.

While participants reported relatively high confidence in their ability to interpret charts ($M = 2.97/4$), objective performance revealed a divergence based on task complexity. Participants consistently succeeded at straightforward, fact retrieval questions (e.g., "What is the value at x?"). However, only 16.6% of participants were able to complete the computational synthesis tasks that required comparing data points across distinct regions of a visualization and performing arithmetic combinations of them.

We observed that users may perceive a generated summary as fluent and clear while simultaneously lacking the structural or comparative support required for higher-order cognitive reasoning. This discrepancy highlights the limitations of current accessibility tools that prioritize descriptive fluency over functional utility. Consequently, these findings directly informed our pivot towards interactive question answering interfacing and the explicit mapping of abstract chart features (such as axes, labels, and values) to concrete data values to better support users' synthesis process.

3.3 Prototype Development

Our pilot study helped inform our initial design of an interactive system that could summarize chart elements as well as larger trends and insights. When a chart is encountered with a screen reader, based on a user action, the chart would be passed to a Vision-Language Model (VLM) to be summarized according to a prompt. We began by implementing a basic version of this system, but several key questions remained: which mode of interaction did potential users prefer, and which prompt led to the most informative and usable summaries? We sought to answer these questions through our formative survey.

3.4 Formative Survey

Given the choice to leverage a VLM to support the chart-to-summary capability and VLM's sensitivity to instructions included in their prompts, we expanded our initial survey to include an evaluation of multiple prompting styles while also increasing our sample size to solicit more community perspectives. We therefore recruited 60 new participants from Prolific ($N = 60$, distinct from the six partic-

ipants in our earlier needs assessment in Section 3.2), balanced between LV and ND cohorts to both describe additional features, interaction models, and desired outputs from an AI-powered chart summarization tool and also to compare summaries generated from six distinct prompt styles (Table 2, full prompts in Appendix C). Each participant evaluated three varied visualizations and rated the resulting summaries for both informativeness and usability using a four-point Likert scale. This phase was instrumental in determining which styles best met user needs for clarity and cognitive accessibility, directly informing the final design of the MCS adaptive verbosity feature. The full set of questions asked of participants is listed in Appendix B.

3.5 Formative Survey Findings

The formative survey revealed a tension between informativeness and system usability (Table 2). No prompting strategy simultaneously maximized both criteria.

The "Technical and Detailed" style achieved the highest rating for informativeness ($M = 3.62$), suggesting that for data-driven tasks, participants prioritize comprehensive structural and numeric accuracy. The "Neutral and Objective" style scored highest for usability ($M = 3.45$), indicating a user preference for balanced, low-noise interfaces during initial chart exploration. Qualitative responses indicated that users valued clarity and specific data points, though some styles were criticized for being too wordy or too vague depending on the participant's individual needs. However, the usability scores ranged from 3.27–3.45, indicated only minor variation in terms of perceived usability.

Differences in preference were categorized by a slight lean toward detailed summaries among Low Vision (LV) participants, compared to a preference for conciseness among Neurodivergent (ND) participants. Independent analyses of these subgroups indicate that LV participants consistently prioritized comprehensiveness, favoring technical and insightful styles, while ND participants reported high sensitivity to cognitive load, favoring Concise, Friendly, or Neutral styles to maintain clarity. These findings confirm that accessibility requirements are not monolithic; the variance in group differences suggests that adaptive verbosity is essential to balancing informativeness and usability for diverse users.

These results informed the final system design: rather than choosing a single best style, MCS sup-

Style	Description (in prompt)	Info.	Usability
Technical and Detailed	Dense numeric and structural detail	3.62	3.29
Insightful and Analytical	Trends, comparisons, and takeaways	3.38	3.31
Neutral and Objective	Plain, balanced baseline	3.25	3.45
Friendly and Approachable	Plain, approachable language	3.12	3.35
Concise and Factual	Minimal facts and core values	3.08	3.42
Casual and Humorous	Playful wording with key facts	3.08	3.27

Table 2: Summary styles and their mean formative ratings on four-point scales.

ports adaptive summary preferences and interactive follow-up questions, which also surfaced as a recurring feature request theme (interactivity) among both populations. This allows users to balance their own informativeness and usability needs dynamically.

3.6 System Implementation

Our final version of MCS is implemented as a Chrome browser extension designed to minimize latency while maximizing accessibility for users navigating complex web interfaces. The user interface features dedicated UI toggles that allow users to dynamically switch between different prompt styles defined in Appendix C.2, instantly updating the underlying system instructions sent to the VLM.

3.6.1 Image Extraction and Processing

For broad compatibility, the MCS extension employs a multistage image extraction pipeline. When the user activates the tool, the extension scans the document object model (DOM) to identify the relevant visual elements, including standard `<img>` tags, embedded SVG containers, and `<canvas>` elements frequently used in modern data-driven web applications. Once identified, these elements are normalized and rendered into a high-fidelity image format. The resulting image is then encoded as a Base64 string for secure transmission to our backend infrastructure. This approach ensures that users can interact with diverse chart types regardless of how they are rendered by the underlying web framework.

3.6.2 Speech Synthesis and Interaction

Upon receiving the LLM-generated summary, the extension utilizes the Chrome Text-to-Speech (TTS) API to render the text as audible output. This integration allows for real-time auditory feedback, which is critical for screen reader users to maintain flow while browsing. Furthermore, the extension provides a dedicated conversational interface, al-

lowing users to engage in follow-up natural language queries. The TTS API also supports these iterative responses, allowing for a speech-based dialogue between the user and the system.

3.6.3 Design Iterations and Prompt Engineering

Our initial prototype utilized a static, zero-shot prompt that produced a single, high-level description. However, results from formative user sessions indicated that a “one size fits all” summary frequently failed to satisfy diverse user needs. Therefore, the following features were added based on findings from the formative study:

1. **Prompt Styles:** The system was updated to allow users to toggle between concise, neutral, analytical, and technical description styles.
2. **Anchoring to Chart Elements:** Prompts explicitly instruct the model to anchor descriptions to specific visual elements, including the chart type, title, axes, labels, trends, and extreme data points (peaks and lows).
3. **Interactive Question Answering:** The system was extended to allow users to ask follow-up questions for clarification or to extract specific data insights that were omitted from the primary summary.

After implementing these changes, we evaluate the system.

3.7 Summative User-based Evaluation

Our summative evaluation asked participants to use the system to perform data-centric tasks using real charts manually retrieved from the web, exemplifying a variety of chart formats from the Data Viz Project³. In addition to quantitative performance metrics, we collected qualitative feedback through open-ended post-task interviews, where participants described their interaction experiences

³<https://datavizproject.com/>

Metric	Baseline	Optimized	Change
Accuracy (%)	29.17	55.00	+88.57%
Confidence	2.79	3.15	+12.90%
Informativeness	2.54	2.86	+12.60%
Usability	2.68	2.92	+8.96%
Word count	301.50	596.32	+97.78%

Table 3: Summative comparison of baseline and optimized conditions. Confidence, informativeness, and usability are reported on four-point scales.

and the utility of specific system features. See Appendix D for a list of prompts, task-based questions, and interview guides. We recruited 60 participants for this study; demographic identifiers included 27 LV-only, 13 ND-only, and 20 participants who identified as both LV and ND.

We employed a *within-subjects counterbalanced design*. Participants completed chart comprehension tasks under two conditions:

1. **Baseline:** A static summary generated from the initial unrefined zero-shot prompt, without interactive Q&A support. Essentially, prompting the model to generate a plain text summary describing the data visualization.
2. **Optimized:** A refined summary generated via our optimized prompt, complemented by access to the interactive Q&A interface. The full prompt is presented in Appendix C.

In the analyzed summative sample, by counterbalancing the order of these conditions, we isolated the effect of our system optimizations from potential order effects. Participants performed factual and analytical “find-and-answer” tasks, followed by assessments of their confidence, summary informativeness, usability, and system informativeness.

The summative evaluation results, summarized in Table 3, demonstrate performance gains using the MCS-optimized pipeline compared to our baseline prompt, demonstrating the value of conducting our formative survey and iterating on the prompt and overall system design.

We performed paired samples t-tests comparing user performance in the baseline versus optimized conditions. For Accuracy and Confidence, the combination of optimized prompting and interactive Q&A significantly improved performance. Accuracy increased from 29.17% to 55.00% ($t(43) = 2.35$, $p < .05$, $d = 0.36$), and Confidence increased from 2.79 to 3.15 ($t(43) = 2.91$, $p < .01$,

$d = 0.44$). While descriptive means for Informativeness (+12.06 %) and Usability (+8.96%) also showed positive trends, these metrics were not subjected to formal significance testing as condition-level data was not available for these metrics in this summative dataset.

Although optimized summaries were nearly twice as long as baseline outputs (an increase of 97.78%), usability scores improved simultaneously. This suggests that increased verbosity is acceptable, and even preferred, when the information is structured and even supported by interactive Q&A. The qualitative data even reinforces this, as participants frequently cited the ability to ask follow-up questions as a critical factor in their performance, noting that interactive exploration provides a more targeted and reliable experience than relying on a single static summary.

User adoption interest was also encouraging: 53.3% of participants said they *would* use the tool in daily life, and another 31.7% said they *might*.

4 Robustness to Flawed Charts

Another key characteristic for a chart summarization system, especially one geared toward LV users, is robustness to flawed charts. To evaluate whether MCS remains reliable on imperfect real-world charts, we constructed an analysis of the summary consistency across design flaws using real open-source Matplotlib⁴ charts, identifying editable source code so that we could programmatically inject flaws. We scraped Python files containing Matplotlib code from public GitHub repositories created between January 2022 and July 24, 2025, and ran the scripts in order to confirm that charts were generated. Scripts were executed in isolated environments; scripts requiring unavailable data or human interaction were skipped.

For each of the 120 successfully rendered charts in the sampled set, we generated an original summary using Gemini 2.5 Flash. We then programmatically injected a single visualization flaw, re-rendered the chart, and manually verified that the flaw was present, as expected. If not, we re-generated that chart. Then, we generated a summary of the flawed chart, and computed the similarity between the original and perturbed summaries by embedding them with the sentenceTransformers (Reimers and Gurevych, 2019) (MiniLM-L6-v2

⁴<https://matplotlib.org/>

Flaw	N	Similarity
Missing y-label	13	0.722
Missing title	6	0.779
Figure too small	11	0.822
Font too small	10	0.833
Inverted y-axis	13	0.841
Truncated y-axis	14	0.858
Dual y-axes	8	0.861
Missing x-label	11	0.870
Missing legend	8	0.877
Low contrast	9	0.886
Tampered aspect ratio	17	0.908
Total / Average	120	0.844

Table 4: Semantic similarity between summaries of original and flaw-injected charts. Lower scores indicate greater summary drift. The reported average is weighted by the number of examples in each flaw category.

model⁵) and computing the cosine similarity. We consider higher similarity indicative of greater semantic stability, allowing us to explore the impact of various design flaws on the generated summaries. We acknowledge that cosine similarity over sentence embeddings captures high-level conceptual and semantic drift rather than complex numerical or fact-level entailment; thus, minor data value shifts may occur even under high similarity scores. We considered two flaw categories: (1) *Accessibility flaws*: missing titles, missing axis labels, missing legends, insufficient color contrast, small figure size, and small text; and (2) *Misleading design flaws*: truncated axes, inverted axes, dual y-axes, tampered aspect ratios, non-sequential ordering, and biased titles.

Table 4 indicates that missing semantic anchors drive summary drift, with missing y-labels showing the lowest similarity ($M = 0.722$). While perceptual flaws exhibit higher stability, we caution that high similarity is not always desirable; faithfully summarizing deceptive designs (e.g., truncated axes) may propagate misinformation. Our measure of similarity only accounts for semantic similarity and does not specifically account for factual alignment. Future work should balance semantic stability with critical detection of deceptive charts (Tonglet et al., 2026). Given small samples for some categories, these results remain preliminary.

⁵<https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

5 Discussion

The NLP community increasingly needs evaluation frameworks that connect model behavior to social outcomes. Our results suggest four measurable dimensions of positive impact for assistive chart summarization: (1) *Access*: whether users can obtain information that was previously unavailable through screen readers; (2) *Comprehension*: whether summaries improve task accuracy and support higher-order reasoning; (3) *Agency*: whether users can direct the system through follow-up questions and choose the level of detail they need; and (4) *Trustworthiness*: whether the system detects or communicates uncertainty about flawed or misleading visualizations. MCS improved access and comprehension, but the robustness results show that trustworthiness remains an open challenge. A system that increases confidence without detecting deceptive charts could create a harmful false sense of understanding.

5.1 Adaptive Accessibility

Our findings support an adaptive view of accessibility. LV and ND users often wanted different kinds of summaries, and individuals within each group also varied. Therefore, chart summarization systems should not optimize for one universal output. Instead, they should support user-controlled verbosity, tonal style, and interaction depth. In practice, this means pairing concise initial summaries with optional detail and question answering.

5.2 Robustness as an Accessibility Requirement

Robustness is not merely a technical property; it is an accessibility requirement. Missing labels, unclear titles, and deceptive axes can undermine the reliability of generated language. Because screen reader users may not be able to visually verify a chart, assistive NLP systems should identify common chart flaws, warn users when interpretation is uncertain, and distinguish between what the chart *visually suggests* and what the *data truly supports*. Furthermore, evaluating such systems requires looking beyond high-level conceptual drift to ensure that the underlying numerical facts are faithfully preserved.

5.3 Baseline Comparison and Technical Constraints

When evaluating the MATCHA Chart Summarizer(MCS), we utilized an internal, static baseline rather than comparing directly against existing external tools such as AltChart (Thompson et al., 2023) or AutoVizuAlly (Duarte, 2024). This methodological choice is necessitated by the fact that these external tools operate primarily as static generation pipelines or lack accessible, programmatic application programming interfaces (APIs) supporting interactive natural-language question-and-answering. Consequently, evaluating our bidirectional interaction framework requires a controlled comparison against a static baseline version of our own prompt pipeline to isolate the measurable impact of adaptive verbosity and conversational Q&A.

5.4 Design Recommendations

Based on our studies, we present the following high-level recommendations for accessible, LLM-based chart summarization systems:

- provide short initial summaries with optional expansion and interactivity;
- explicitly anchor labels, axes, legends, and data values in language;
- support natural-language follow-up questions in multiple modalities, if possible;
- allow user-controlled summary style and verbosity;
- detect, report, and explain chart flaws such as missing labels and truncated axes;
- communicate uncertainty rather than producing overconfident summaries;
- evaluate impact using user task performance, not only generation quality.

6 Conclusion

We introduced the MATCHA Chart Summarizer (MCS), an interactive, VLM-based system enhancing chart accessibility. Our evaluations confirm that adaptive, interactive summaries significantly improve comprehension and utility for screen-reader users. However, our robustness testing highlights a critical risk: models may faithfully reproduce visual distortions, necessitating a shift toward robust, detection-aware architectures. Collectively,

this work demonstrates that socially beneficial NLP for accessibility must prioritize adaptivity and robustness, ultimately advancing equitable access to data-driven information.

Limitations

This study has several limitations. First, participant classification was based on self-identification rather than clinical assessment. This choice aligns with community-centered accessibility research, but it limits claims about specific impairment severity. Second, the needs assessment sample was small ($N = 6$) and should be interpreted as formative rather than statistically generalizable. Third, the summative design may conflate prompt optimization with the presence of question answering; future work should isolate these factors. Fourth, robustness experiments used relatively small samples for some flaw types, and the chart corpus was Matplotlib-centric, which may not generalize to D3, SVG dashboards, or screenshots from web pages. Fifth, our evaluation of summary reliability, adaptive verbosity, and semantic drift metrics is constrained by its reliance on a single underlying vision-language-model, Gemini 2.5 Flash. Foundational models exhibit unique instruction-following sensitivities, training-priors, and internal spatial reasoning constraints, meaning the specific robustness thresholds and prompt-style preferences observed here may vary when deployed across alternative model architectures. Finally, our summative user evaluation couples the optimized prompt with additional interactive Q&A features; disentangling the impact of each of these improvements would require a further ablation study.

MCS is intended to advance information equity by improving access to data visualizations for screen reader users. However, assistive AI systems also introduce risks. Users may over-trust fluent summaries, especially when they cannot visually inspect the original chart. If a chart is deceptive or underspecified, an LLM may generate a confident but misleading interpretation. This risk is particularly serious in high-stakes domains such as healthcare, education, finance, and public policy.

To mitigate these risks, future deployments should include uncertainty cues, flaw detection, privacy-preserving image handling, and clear communication that generated summaries are assistive interpretations rather than authoritative ground truth. MCS should complement, not replace, the

author's responsibility to create accessible charts at the source. We also caution against using AI accessibility tools as an excuse to weaken accessibility standards or shift labor from content creators to disabled users.

Ethical and Societal Implications

This research was conducted under an institutional IRB exemption for minimal-risk human subjects research. All participant data, including screen reader type, data literacy levels, and survey responses, were pseudonymized at the point of collection, and no individual identity was reconstructed from the reported aggregate findings.

In this work, positive impact means increasing users' independent access to chart information while preserving agency, privacy, and trust. Our evaluation shows promising progress toward this goal, but also demonstrates why impact must be measured through both user outcomes and robustness to harmful inputs.

Acknowledgements

This material is based upon work supported by the U.S. National Science Foundation through the Research Experiences for Undergraduates (REU) program under Award No. 2426230.

References

- Md Zubair Ibne Alam, Shehnaz Islam, and Enamul Hoque. 2023. [Seechart: Enabling accessible visualizations through interactive natural language interface for people with visual impairments](#). In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, page 46–64. ACM.
- Jianzhu Bao, Haozhen Zhang, Kuicai Dong, Bozhi Wu, Sarthak Ketanbhai Modi, Zi Pong Lim, Yon Shin Teo, and Wenya Wang. 2026. [Learning more from less: Exploiting counterfactuals for data-efficient chart understanding](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27775–27797, San Diego, California, United States. Association for Computational Linguistics.
- Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samuel White, and Tom Yeh. 2010. [Vizwiz: nearly real-time answers to visual questions](#). In *Proceedings of the 23rd annual ACM symposium on User interface software and technology*, UIST '10, pages 333–342, New York, NY, USA. Association for Computing Machinery.
- Alberto Cairo. 2016. *The Truthful Art: Data, Charts, and Maps for Communication*. New Riders, San Francisco, CA.
- Rocío Calvo, Faezeh Seyedarabi, and Andreas Savva. 2016. [Beyond web content accessibility guidelines: Expert accessibility reviews](#). In *Proceedings of the 7th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion*, DSAI '16, pages 77–84, New York, NY, USA. Association for Computing Machinery.
- William S. Cleveland. 1984. [Graphs in scientific publications](#). *The American Statistician*, 38(4):261–269.
- Chintan A Divecha, Milind S Tullu, and Sunil R Karande. 2023. [Utilizing tables, figures, charts and graphs to enhance the readability of a research paper](#). *Journal of Postgraduate Medicine*, 69(3):125–131.
- J. Duarte. 2024. [Autovizually: A tool to automate accessibility in data visualizations for screen reader users](#). Master's thesis, University of Lisbon.
- Jialiang Fan, Huaxin Zhang, Yuran Zhang, and Guoliang Li. 2023. [Accessibility of covid-19 statistics trackers for screen reader users](#). In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*, pages 1–15, New York, NY, USA. Association for Computing Machinery.
- Pinkesh Gerela, P N Mishra, and Rajesh Vipat. 2022. [Study on data visualization: Its importance in education sector](#). *International Journal of Health Sciences*, 6(S3):6298–6305.
- Sara Hjelle, Patrick Mikalef, Najah Altwaijry, and Vinit Parida. 2024. [Organizational decision making and analytics: An experimental study on dashboard visualizations](#). *Information & Management*, 61(6):104011.
- Crescentia Jung, Shubham Mehta, Atharva Kulkarni, Yuhang Zhao, and Yea-Seul Kim. 2022. [Communicating visualizations without visuals: Investigation of visualization alternative text for people with visual impairments](#). *IEEE Transactions on Visualization and Computer Graphics*, 28(1):1095–1105.
- Antonia Karamolegkou, Angana Borah, Eunjung Cho, Sagnik Ray Choudhury, Martina Galletti, Pranav Gupta, Oana Ignat, Priyanka Kargupta, Neema Kotonya, Hemank Lamba, Sun-Joo Lee, Arushi Mangla, Ishani Mondal, Fatima Zahra Moudakir, Deniz Nazar, Poli Nemkova, Dina Pisarevskaya, Naqee Rizwan, Nazanin Sabri, and 14 others. 2026. [NLP for social good: A survey and outlook of challenges, opportunities and responsible deployment](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5110–5170, Rabat, Morocco. Association for Computational Linguistics.
- Rachneet Kaur, Nishan Srishankar, Zhen Zeng, and Sumitra Ganesh. 2026. [ChartAgent: A multimodal](#)

- agent for visually grounded reasoning in complex chart question answering. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18488–18555, San Diego, California, United States. Association for Computational Linguistics.
- Alan Lundgard and Arvind Satyanarayan. 2022. Accessible visualization via natural language descriptions: A four-level model of semantic content. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):1127–1137.
- Andrew MacFarlane, Areej Al-Wabil, Chloë R. Marshall, Nora Al-Braheem, Shaden Abdullah, and Panayiotis Zaphiris. 2010. The effect of dyslexia on information retrieval: a pilot study. *Journal of Documentation*, 66(3):307–326.
- Anshul Vikram Pandey, Katharina Rall, Margaret L. Satterthwaite, Oded Nov, and Enrico Bertini. 2015. How deceptive are deceptive visualizations? an empirical analysis of common distortion techniques. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI '15*, page 1469–1478, New York, NY, USA. Association for Computing Machinery.
- Christopher Power, André Freire, Helen Petrie, and David Swallow. 2012. Guidelines are only half of the story: accessibility problems encountered by blind users on the web. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '12*, pages 433–442, New York, NY, USA. Association for Computing Machinery.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Ather Sharif and Alizeh Ellis. 2021. Screen reader user experiences with data visualizations. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA. Association for Computing Machinery.
- Ather Sharif, Olivia H. Wang, Alida T. Muongchan, Katharina Reinecke, and Jacob O. Wobbrock. 2022. Voxlens: Making online data visualizations accessible with an interactive javascript plug-in. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI '22*, pages 1–19, New York, NY, USA. Association for Computing Machinery.
- Arjun Srinivasan et al. 2024. Azimuth: Structured navigation for complex data dashboards. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):1–11.
- John R Thompson, Jesse J Martinez, Alper Sarikaya, Edward Cutrell, and Bongshin Lee. 2023. Chart reader: Accessible visualization experiences designed with screen reader users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23*, pages 1–18, New York, NY, USA. Association for Computing Machinery.
- Jonathan Tonglet, Jan Zimny, Tinne Tuytelaars, and Iryna Gurevych. 2026. Is this chart lying to me? automating the detection of misleading visualizations. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8823–8844, San Diego, California, United States. Association for Computational Linguistics.
- Keke Wu, Emma Petersen, Tahmina Ahmad, David Burlinson, Shea Tanis, and Danielle Albers Szafir. 2021. Understanding data accessibility for people with intellectual and developmental disabilities. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21*, pages 1–16, New York, NY, USA. Association for Computing Machinery.
- Yichen Xu, Liangyu Chen, Liang Zhang, Zihao Yue, Jianzhe Ma, Wenxuan Wang, and Qin Jin. 2026. POLYCHARTQA: Benchmarking large vision-language models with multilingual chart question answering. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 44154–44186, San Diego, California, United States. Association for Computational Linguistics.
- Yichun Zhao, Miguel A. Nacenta, Mahadeo A. Sukhai, and Sowmya Somanath. 2024. Tada: Making node-link diagrams accessible to blind and low-vision people. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, CHI '24*, pages 1–20, New York, NY, USA. Association for Computing Machinery.
- Yea-Chun Zhu, Brian Lee, and Lane Harrison. 2023. The impact of visual clutter on data visualization tasks for neurodivergent users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery.

A Needs Assessment Survey Instrument

A.1 Pre-Screening and Consent

- Integrity Acknowledgement:** Instructions: Answer the questions to the best of your ability. AI and pasting are not permitted. If used, the study will be returned to you. Do you understand and agree to these terms? [Yes/No]
- Verification:** Please provide your signature or check the box to confirm your agreement. [Signature/Check box]

A.2 Section 1: Demographics and Eligibility

1. What is your age?
2. What is your highest level of education?
3. Do you identify as blind, low-vision, neurodivergent (e.g., ADHD, dyslexia), or some combination? If comfortable, please specify your diagnosis or assistive-technology needs.
4. What is your self-identified race and ethnicity?
5. **Screening Question:** Do you use assistive technologies (e.g., screen readers, Braille displays)? [Yes/No]
(Skip logic: If "No" is selected, the survey ends.)
6. Which assistive technologies do you use? Please list.

A.3 Section 2: Data Literacy and Baseline Usage

1. Which types of charts or graphs (e.g., bar charts, line graphs, pie charts, scatterplots) do you encounter most often, either at work, school, or in daily life?
2. How confident are you (on a 5-point scale: 1: Not at all → 5: Extremely) in interpreting data from charts?
3. Do you have any formal training in data literacy (e.g., coursework, workshops)? If so, could you describe it briefly?

A.4 Section 3: Accessibility Practices and Barriers

1. When presented with a chart in a document or on a website, what format do you usually prefer (e.g., CSV tables, detailed textual descriptions, tactile graphics, audio summaries)?
2. Can you share an example of a time when a chart was particularly hard to interpret? What did you do in that situation?
3. When you encounter a chart that your screen reader cannot interpret at all, what is your fallback strategy (e.g., ask a colleague, brute-force inspect the underlying data table)?
4. What aspects of a chart (e.g., colors, labels, axis scales, legends) do you find most difficult to access or understand?

5. In your experience, what information is often missing from "accessible" summaries of charts? (e.g., context, comparative values, trend descriptions)
6. When using a screen reader to interpret a chart, have you ever encountered ambiguity (e.g., when multiple data series overlap)? How did you handle it?
7. In your opinion, what is currently lacking in data visualization accessibility for someone with your needs?

A.5 Section 4: Feature Preferences and Design Implications

1. If you could redesign a bar chart, what elements would you include? (e.g., auditory cues for rising/falling trends, keyboard shortcuts to navigate series)
2. Do you prefer summaries that read out every single data point, or summaries that describe high-level trends (e.g., "Sales increased steadily from January to June")? Why?
3. How important is it for you to interactively explore a chart (e.g., press a key to hear the value of a specific data point) versus receiving a static summary? Explain.
4. If an interface let you press letter keys (like A, B, or C) to explore parts of a chart, what would you expect each key to do?
5. Have you ever used any tools (such as Python scripts, browser extensions, or assistive technology add-ons) to make charts or graphs more accessible? If yes, which ones, and what was your experience like?
6. If not, would you be open to trying a tool designed to make charts more accessible—such as something that works with screen readers or provides customized descriptions—even if it's a new kind of interface? Why or why not?

A.6 Section 5: Pilot Chart Interpretation and Robustness

1. **Interpretative Task:** Participants were presented with a series of charts (e.g., Figures 1–5) and asked to derive key trends and comparisons. We measured accuracy, response

time, and the ability of participants to identify misleading data representations (e.g., truncated axes in Figure 4).

2. **Flawed Chart Assessment:** To evaluate model and human robustness, participants were exposed to intentionally “flawed” charts—such as those with missing labels, inverted axes, or tampered aspect ratios. Participants were asked:

- Does this chart contain any visual design choices that seem intended to mislead or confuse? If so, what are they?
- How does the missing information (e.g., axis label) impact your confidence in interpreting this data?

3. **Utility Evaluation:** Following each task, participants rated the effectiveness of the AI-generated summaries on a 5-point Likert scale, specifically noting whether the summary successfully signaled the presence of the identified visual flaws.

B Formative Survey Instrument

B.1 Section 1: Eligibility and Demographics

1. Are you currently using a screen reader? [Yes/No]
2. What is your age? [18–24, 25–34, 35–44, 45–54, 55–64, 65–74, 75+]
3. Which of the following best describes your race/ethnicity? (Select all that apply)
4. What is the highest level of education you have completed?
5. Which of the following apply to you? [Blind, Low Vision, Neurodivergent, None of the above]
6. Which screen reader(s) do you use? [Open text]
7. How confident are you in interpreting data from charts? [1: Not confident at all, 4: Very confident]
8. Have you received any formal training in data literacy, data visualization, or statistics? If yes, please describe. [Open text]

B.2 Section 2: Feature Preferences

Participants were asked to imagine a chart summarization tool integrated into their screen reader and responded to the following:

1. What features would you want such a tool to have? [Open text]
2. How would you want to trigger the tool to generate a summary? (e.g., keyboard shortcut, context menu, automatic summarization) [Open text]
3. What information would you want to be able to hear individually when pressing different keys? [Open text]

B.3 Section 3: Summary Comparison Task

For each of three randomly selected charts, participants were presented with two summaries (A and B) generated using different prompt styles and were asked:

1. How informative is Summary [A/B]? [1: Not at all, 4: Very informative]
2. How easy is Summary [A/B] to understand? [1: Very difficult, 4: Very easy]
3. Which summary do you prefer? [A / B]
4. Why do you prefer that summary? [Open text]

C LLM Prompts

C.1 Formative Survey Prompts

All prompts were preceded by the instruction to start with “A [visualization type] shows...” or “A [visualization type] titled [title] shows...” and to return “N/A” if the image was not a visualization.

Concise and Factual: In 1–2 short, fact-based sentences, describe the visualization by stating the visualization type, title, key maximum(s), minimum(s), and trend(s) with no interpretation or opinion.

Friendly and Approachable: Describe it in 1–3 sentences using easy-to-understand, friendly language. Mention what kind of visualization it is, its title, any highs and lows, any overall patterns, and any relevant information for the user.

Technical and Detailed: In 2-5 sentences, include the visualization type, title, axes information (if readable), peak and low values with specific data points if available, directionality of trends, and any outliers or anomalies.

Insightful and Analytical: Give 1-2 sentences about the main features of the visualization including the title, maximum(s), minimum(s), and general trend(s), as well as any key insight(s).

Casual and Humorous: You're a screen reader with a sense of humor. Give a laid-back, fun 2-3 sentence summary that still covers the main points: what kind of visualization it is, the title, what goes up, what goes down, and any surprising twists.

Neutral and Objective: Provide a neutral, unbiased summary in 1-2 sentences. Mention the type of visualization, title, main trends, and the highest and lowest values without adding interpretation or judgment.

C.2 Optimized Prompt for Summative Study

You are a screen reader and came across this data visualization. Describe it in 1-2 sentences using simple, friendly language. Mention what kind of visualization it is, its title (if any), any highs and lows and what the overall pattern seems to be. Start with "A [visualization type] shows..." or "A [visualization type] titled [title] shows..." If it is not a data visualization, say "N/A".

D Summative Graph Interpretation Questions

Participants were shown a visualization and asked to answer a specific factual or analytical question based on its contents.

Radar Chart: Which gymnast received the highest score for Landing? (*Answer: Gymnast 3*)

Bubble Chart: Based on the chart, is the average technology salary in the SF Bay Area more than double that of Montreal? (*Answer: No*)

Stream Graph: What were the total US music sales, in billions of dollars, in 2013? (*Answer: \$7 billion*)

OHLC Chart: On how many days did the price finish higher than it started? (*Answer: 3*)

Venn Diagram: How many unique ingredients are shared between at least two of the recipes? (*Answer: 13*)

Density Plot: Which site most frequently experiences a yield of approximately 30 units? (*Answer: Duluth*)

Dot Map: What year is the data for U.S. Public Libraries from? (*Answer: 2009*)

Bar Chart: What is the count of female preferences for Vanilla flavor? (*Answer: 12*)

Flow Chart: Is it advisable to dump them by text per the provided logic? (*Answer: Yes*)

Spiral Plot: Approximate total US coal production in 1976? (*Answer: Cannot be determined*)

Radial Column Chart: Comparison of Disease deaths in February vs. Wounds in October? (*Answer: Numbers not provided*)

Tree Diagram: Win-loss record of '1 Purdue'? (*Answer: 29-5*)

Sunburst Diagram: Main ownership category for 'Peabody Energy'? (*Answer: Investor owned*)

Population Pyramid: Percentage of male population aged 0-4 in 2020? (*Answer: 4%*)

Sankey Diagram: Country receiving largest number of migrants from Wales in 2015? (*Answer: England*)

Parallel Coordinates Plot: Hourly wage for a Carpenter/bricklayer in South Africa? (*Answer: Exact number unavailable*)