# Automated German Alt Text Generation for News Charts

**Alessia Vannini[1,+], Julia Locher[1,+], Marianne Santaholma[1,*], Claudia Amsler[1], Jonas Oesch[2], Arzu Coltekin[1,*]**

[1]University of Applied Sciences and Arts Northwestern Switzerland FHNW
[2]Neue Zürcher Zeitung, Zurich, Switzerland

[+]Shared first authorship with equal contributions [*] Shared senior authorship

**Correspondence:** marianne.santaholma@fhnw.ch and arzu.coltekin@fhnw.ch

## Abstract

We investigate whether a Multimodal Large Language Model (MLLM) can automatically generate well-formed German chart alt texts that meet the requirements of visually impaired persons while following the accessibility guidelines for chart alt texts, and match the quality of manually authored gold-standard alt texts. Focusing on bar, line, and stacked bar charts from a German-language newspaper (the *Neue Zürcher Zeitung*), we define an alt text structure, construct a gold-standard corpus, and evaluate MLLM-generated chart alt texts both quantitatively on semantic similarity and qualitatively with visually impaired persons as participants in terms of clarity, conciseness, meaningfulness, and output consistency.

## 1 Introduction

Data visualizations such as line and bar charts, and other graphical representations play a central role in communicating complex information efficiently (Çöltekin et al., 2021). Yet these visual elements are frequently inaccessible to people with visual impairments (PVIs), primarily due to the absence of alternative text descriptions, commonly referred to as alt texts. Alt texts are plain text descriptions embedded in digital documents and web pages that are rendered by screen readers, the assistive technology most widely used by PVIs (AbilityNet, 2025). Alt texts can be further rendered as large print, braille, speech, symbols, or simpler language (Accessibility Guidelines Working Group (AG WG), 2023).

The provision of alt texts for all non-text digital content has recently been formalized into a legal requirement in the EU: the European Accessibility Act (EAA), in effect since June 2025, mandates that digital media providers make non-text content such as images, charts, and graphics accessible through appropriate text alternatives (European Union, 2019). WebAIM reports that 55.5% of the one million most-visited web pages were still missing alternative text for images in 2025 (WebAIM, 2026). This gap is not simply a matter of negligence or lack of awareness. Manual alt text authoring is time-consuming and impractical, compounded by a lack of established guidelines for what constitutes a good chart description (Gleason et al., 2019). This challenge is especially acute in high-volume publishing environments such as online newspapers, where fast-paced editorial workflows leave little room for annotating every visual element.

Recent advances in natural language generation (NLG), particularly through Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs), offer a promising avenue for automating chart alt text generation (Hoque and Islam, 2025; Nylund et al., 2025). These models can process visual input alongside textual prompts to produce descriptive natural language output, potentially satisfying accessibility requirements while reducing editorial burden. However, generating high-quality alt texts for charts demands more than visual understanding: it requires accurate interpretation of quantitative data and graphical conventions, as well as clear criteria for what constitutes a good alt text per chart type (Yan et al., 2025; Kantharaj et al., 2022).

We address the above-mentioned challenges through a case study in this paper, i.e., by investigating whether an MLLM can automatically generate well-formed German chart alt texts that meet the needs of PVIs, conform to the set accessibility guidelines (Accessibility Guidelines Working Group (AG WG), 2023), and match the quality of manually authored gold-standard alt texts. The study focuses on bar, line, and stacked bar charts published in the Neue Zürcher Zeitung (NZZ), a national German-language newspaper in Switzerland. We define a structured alt text schema, construct a gold-standard corpus, and evaluate MLLM-generated alt texts against it in terms of clarity,



*Proceedings of the 11th Edition of the Swiss Text Analytics Conference (SwissText 2026) – Scientific Track*, pages 215–224
June 10, 2026 

conciseness, meaningfulness, and output consistency.

The following sections describe the related work, methodology, evaluation setup and results. We conclude with discussion and conclusions.

## 2 Background and Related Work

This section reviews what constitutes a well-formed chart alt text and the state of the art in automated chart alt text generation.

### 2.1 Chart Alt Text Structure

The alt text serves to convey the information encoded in a chart to PVIs, rendered via screen readers. However, no standard definition of alt text for different chart types exists. Researchers and accessibility organizations have developed guidelines and best practices to support digital content creators in writing effective chart descriptions (e.g., Accessibility Guidelines Working Group (AG WG), 2023; DIAGRAM Center, 2020; Pennsylvania State University, 2024; Consumer Financial Protection Bureau). Accessibility research emphasizes that effective chart alt texts rely on a clear and consistent structure that enables users to reconstruct a mental model of the chart(e.g., Belle et al., 2022; Jung et al., 2022). With respect to content ordering, prior work suggests beginning with declaring the chart type as an orienting element, followed by contextual information such as axis labels and value ranges, and concluding with the main message or trend conveyed by the data (Jung et al., 2022; Yan et al., 2025). A further structural consideration concerns the separation of short and long descriptions: since longer descriptions cannot be paused or skipped during screen reader playback, clearly distinguishing between the two allows users to decide whether to engage with the more detailed account. Beyond descriptions, providing a machine-readable data table containing all values displayed in the chart has been identified as a high-priority requirement among PVIs (Jung et al., 2022). This body of work informed our development of a structural model for chart alt texts (see Figure 1).

The model specifies the components a well-formed alt text should contain. It does not, however, prescribe how individual elements should be described across different chart types, how numerical values and comparisons should be verbalized, or how alt texts should scale with varying levels of chart complexity. The present work refines these

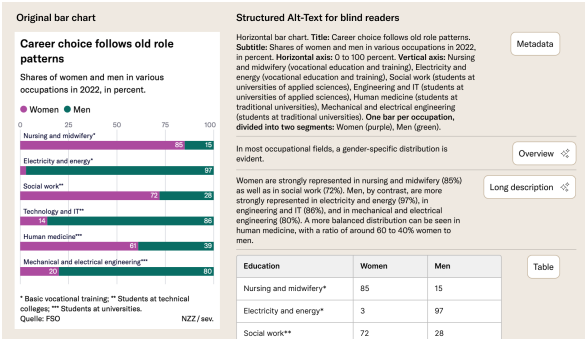


Figure 1: Chart alt text structure, illustrated with an example bar chart on the left and its structured alt text on the right (metadata, overview, long description, and data table).

structural components and their contents for specific chart types and their subcategories. We validated and iteratively refined the model in collaboration with PVIs.

### 2.2 Automated Chart Alt Text Generation

Automated chart alt text generation draws on a broader cluster of related research areas, including chart-to-text generation, chart question answering, and chart understanding (Huang et al., 2025). Chart-to-text generation focuses on producing natural language summaries of chart content, while chart question answering evaluates whether models can correctly respond to queries about specific data points or trends (Obeid and Hoque, 2020; Masry et al., 2022). Chart understanding encompasses both tasks (with data extraction and chart type classification), addressing how well models can parse the visual and semantic structure of a chart.

Early work relied on rule-based or template-driven methods that extracted data from structured chart sources and filled predefined description templates (Farahani et al., 2023). While reliable for simple chart types, these approaches lack flexibility and cannot generalize to novel or complex visualizations (Obeid and Hoque, 2020).

The emergence of LLMs and MLLMs has opened new possibilities for chart alt text generation. These models can interpret chart images, identify salient trends, and produce fluent natural language descriptions, even without having access to the underlying data (Wang et al., 2024; Yin et al., 2024; Kantharaj et al., 2022; Moured et al., 2024; Huang et al., 2025). Recent studies have demonstrated the potential of (M)LLM-based pipelines for generating alt texts that approach human-authored quality from a language-structure

and fluency point of view (Obeid and Hoque, 2020; Kantharaj et al., 2022). Nevertheless, challenges persist on multiple fronts: on the operational side, effective prompt design and robust evaluation criteria remain open problems; on the output side, issues of numerical inaccuracy, hallucination, and appropriate descriptive granularity continue to limit practical deployment (Obeid and Hoque, 2020).

Existing literature and tools we identified on automated chart alt text generation focus on English-language (such as VisText (Tang et al., 2023) and Alt4Blind (Moured et al., 2024)), presenting a gap for other languages. We thus extend this line of research to German in this study, for which, to the best of our knowledge, no comparable work currently exists, including a gold-standard corpus. Moreover, existing approaches primarily focus on training new models and benchmarking them against established datasets such as Statista (Gong et al., 2019; Statista), without taking accessibility guidelines into account (Obeid and Hoque, 2020; Belle et al., 2022; Balaji et al., 2018; Gong et al., 2019).

## 3 Methodology

The overall methodological framework and workflow are illustrated in Figure 2. The workflow begins with data preprocessing, followed by a manual construction of a gold-standard alt text corpus. The alt texts are grounded in a structured alt text template as shown in Figure 1. Alt texts are generated using an MLLM and few-shot prompting. The evaluation strategy combines qualitative feedback and Likert-scale ratings collected from PVIs and quantitative metrics including semantic similarity comparisons and an LLM-as-a-judge approach.

### 3.1 NZZ News Charts

The dataset consists of 168 charts provided by the NZZ. Each chart is provided as a PNG image accompanied by a JSON file containing metadata and underlying data values. Of the 168 charts, 150 fall within the scope of this study, covering line, bar, and stacked bar charts; the remaining 18 belong to other chart types. We categorized the charts as either *simple* or *complex* based on the number of columns in the underlying dataset: simple charts contain a single data column, while complex charts contain two or more, resulting in additional categorical dimensions such as multiple lines or grouped bars. Examples of simple and complex line, bar and stacked bar charts are shown in Figure 3.

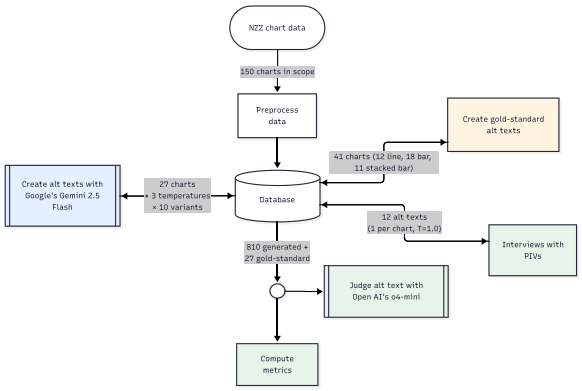


Figure 2: Overview of the alt-text generation and evaluation workflow: gold-standard creation (orange), alt-text generation with Gemini 2.5 Flash (blue), and evaluation via computed metrics, LLM-as-a-judge with o4-mini, and PVI interviews (green).

Most bar charts fall into the *simple* category (46 simple, 24 complex) and can be oriented horizontally or vertically, with either time-based or categorical axes. Line charts, by contrast, have more of the *complex* category (23 simple, 39 complex) and always display a time series on the x-axis. Stacked bar charts are the smallest group and are mostly complex (5 simple, 13 complex). They are further subdivided by whether bars sum to 100% or display absolute values, and by the number of segments. In addition to these primary structures, several charts employ supplementary visual encodings such as highlighted regions, prognosis markers, and annotated events, which add further descriptive requirements for alt text generation.

A preprocessing pipeline parses the JSON files, extracts metadata and data values, and converts them into CSV files stored in a relational database. This pipeline can be extended to other chart types.

### 3.2 gold-standard Data

At the preliminary stages of this project, we used Statista's open-source dataset (Statista) as a proxy gold-standard. However, this dataset has several shortcomings that limit its suitability for systematic benchmarking. The texts do not distinguish between short and long descriptions, chart metadata (such as chart type, axis labels, and value ranges) is frequently incomplete or absent, the text includes information that is not included in the chart, and the descriptions are assumed to function as monolithic long-form accounts rather than structured, modular alt texts. The discrepancies with accessibility

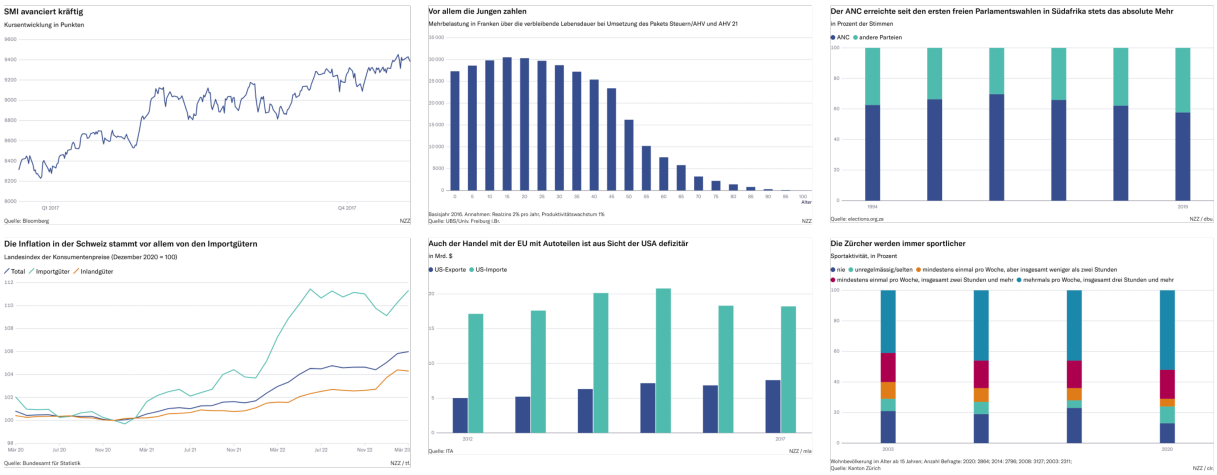


Figure 3: Representative examples from the NZZ chart dataset, organized by chart type and complexity. Columns correspond to line, bar, and stacked bar charts. The top row shows simple charts (single data column), and the bottom row shows complex charts (two or more data columns).

guidelines make the dataset ill-suited as a reference for structured chart alt text generation. Therefore, we manually created a German gold-standard corpus. The corpus covers 41 NZZ charts selected to represent a wide range of subcategories and visual encodings: 12 line charts (5 simple, 7 complex), 18 bar charts (10 simple, 8 complex) and 11 stacked bar charts (5 simple, 6 complex).

We developed the corpus through three iterative revision stages. In the first stage, we drafted initial alt texts with the assistance of Open AI's GPT 5.1 and internally validated them through a sketching exercise, in which one author produced structural reconstructions of the charts based solely on the written descriptions as a check for completeness and clarity. In the second stage, we refined the texts through a linguistic consultation with an expert in German language and simple-language writing, focusing on neutral phrasing and readability, in line with prior work (Jung et al., 2022). In the third and final stage, we presented the alt texts to five PVIs in structured interviews. They assessed how well the descriptions supported the mental reconstruction of charts across types and complexity levels.

The feedback from PVIs directly shaped the final revision. Key changes included: (a) rendering the short description as metadata in a technical rather than fluent prose style; (b) repositioning the overview summary at the end of the short description rather than at the beginning of the long description; (c) long descriptions follow a chronological progression designed to support mental reconstruction without overloading the listener with detail; and (d) reducing numerical precision where a data table was available, while retaining exact figures otherwise. More broadly, PVIs emphasized a strong preference for having access to the same information as sighted users, and for a consistent basic structure applied uniformly across all charts.

This final version constitutes the gold-standard reference data that we used for all subsequent evaluation in this study.

### 3.3 Generation of the Chart Alt Texts

We generated the alt texts using Google's Gemini 2.5 Flash Preview 09-2025 (Google DeepMind, 2025), an MLLM capable of processing both textual and visual input. The model is accessed via the OpenRouter API (OpenRouter), which supports a temperature range of 0.0 to 2.0. Temperature governs the degree of randomness in model output: a value of 0.0 yields deterministic outputs for identical inputs, while higher values increase variability. We set the temperature parameter to the model default of 1.0. Future work should evaluate the effect of temperature on alt text quality, consistency, and factual accuracy, for instance by comparing outputs at T=0.0 and T=1.0 through human evaluation. We selected Gemini 2.5 Flash as it could be integrated into the NZZ data pipeline, facilitating the possible adoption in an operational publishing context.

As inference strategy we employed **multimodal few-shot prompting**: each request supplies the model with a chart image, a CSV excerpt of the underlying data, natural language instructions, and a set of illustrative gold-standard examples, from which the model is expected to infer the desired output structure and style without any additional

fine-tuning.

The prompt design is based on the systematic trials that we conducted in our preliminary work (Locher and Vannini, 2025). We demonstrated that structured prompts with explicit formatting constraints consistently outperform free-form instructions for chart alt text generation. In particular, prompts that clearly separate short and long descriptions, impose length constraints, and adapt to chart type and complexity through type-specific instructions and examples yield the highest overall output quality.

We designed six distinct prompt variants for this study: one for each combination of chart type (line, bar, stacked bar chart) and complexity level (simple, complex). Each variant encodes chart-type-specific rules, content ordering constraints, and wording guidelines. They all follow a shared template structure comprising the following components:

**Task definition** Explicit instructions for generating the short description (metadata-based), the overview, and the long description, including word limits.

**Examples** Illustrative alt texts drawn from the gold-standard corpus. Of the 41 gold-standard alt texts available, we selected 14 as in-prompt examples. These cover the full range of chart types (12) and their subcategories. For complex bar charts, we distinguished three subcategories: 1) category within a time series, 2) time series within a category, and 3) category within a category. For simple stacked bar charts, we defined the following three variants: 1) 100% charts with multiple bars, 2) 100% charts with a single bar, and 3) charts with absolute values. Each subcategory required a dedicated example, resulting in three in-prompt examples for these two chart types rather than default two. All remaining chart types retained two in-prompt examples each.

**Data** CSV excerpt of the underlying chart data extracted during preprocessing.

**Stylistic rules** Guidelines governing tone, phrasing, and language use. The stylistic rules reflect both the accessibility requirements and the feedback that we gathered from PVIs. Descriptions are required to be concise and clear, and causal interpretations are explicitly prohibited in favor of a neutral, observational tone. Trend descriptions are restricted to the verbs *steigen* and *sinken*, avoiding evaluative or dramatic alternatives such as *erreichen* or *einbrechen*. Symbols are written out, e.g. negative values are expressed using the word *minus* rather than the minus symbol, as screen readers may misinterpret them.

**Fixed output format** A fixed output structure separates the short description, overview, and long description using defined delimiters, ensuring consistent formatting and reliable downstream parsing and evaluation.

## 4 Evaluation Setup

The evaluation strategy combines automatic reproducible metrics with qualitative human feedback. It addresses the central question of how MLLM-generated German chart alt texts compare to manually created gold-standard alt texts in terms of clarity, conciseness, meaningfulness, and output consistency. The evaluation is organized along three dimensions:

- **Semantic similarity**: the degree of semantic overlap between generated and gold-standard alt texts, assessed using sentence embeddings.
- **Text length as a proxy for conciseness**: comparison of character counts across generated and gold-standard alt texts. Text length serves as an objective and reproducible approximation of conciseness, as longer texts are more likely to contain redundant phrasing or unnecessary detail.
- **Quality judgements**: a comparison of LLM-based and human rating across shared evaluation criteria.

The evaluation is based on 27 gold-standard chart alt texts. A total of 282 MLLM-generated alt texts were produced for evaluation: 270 from ten generation runs per chart under identical conditions ($27 \times 10$, temperature $= 1.0$), plus 12 additional texts generated specifically for human evaluation. We generated these 12 additional texts in an initial pipeline validation run, using a balanced subset of the 27 benchmarking charts (two per chart type and complexity level). These texts served as stimuli in the PVI interviews (See 3.2), in which participants evaluated MLLM-generated alt texts rather than the gold-standard texts. Presenting gold-standard texts to PVIs would have introduced a methodological circularity, as the gold-standards texts were

iteratively refined based on prior PVI feedback. Alt texts are generated in a strict output format and parsed into sections. All metrics are computed separately for the full text, the short description (metadata), the short description (overview), and the long description, when relevant.

**Semantic similarity** was computed as the cosine similarity of sentence embeddings extracted using SBERT (Reimers and Gurevych, 2019). Embeddings were computed separately for four textual units: the full alt text, the short description (metadata), the short description (overview), and the long description. High cosine similarity indicates substantial semantic overlap with the gold-standard, while lower values reflect divergence in content or phrasing.

**Text length** was assessed through descriptive statistics of character and word counts per section, comparing gold-standard and generated texts across chart types and complexity levels.

**LLM-as-a-judge** evaluation (Gu et al., 2025) was conducted using OpenAI's o4-mini model accessed via OpenRouter (OpenAI, 2024) to rate each generated alt text along six criteria: clarity, conciseness, neutrality, perceived completeness, completeness, and correctness. Ratings were produced on a Likert scale and compared against human ratings on the shared subset of criteria. We treated completeness and perceived completeness as distinct criteria. Completeness assesses whether the alt text factually covers the essential elements of the chart (verifiable against the chart image and underlying CSV data), and can therefore only be evaluated by the LLM judge. Perceived completeness, by contrast, captures the subjective impression of sufficiency, i.e. whether the text appears to cover everything important without requiring visual access to the chart. We introduced this distinction to enable PVI participants to provide meaningful completeness-related judgments despite their visual impairment.

**Human evaluation** was conducted with five PVIs across 12 charts ($12 \times 5 = 60$ ratings). Prior to the evaluation, participants received a single verbal explanation of the four rating criteria (neutrality, clarity, conciseness, and perceived completeness) to ensure a shared understanding. We excluded completeness and correctness from the human evaluation as they require direct visual access to the chart and underlying data.

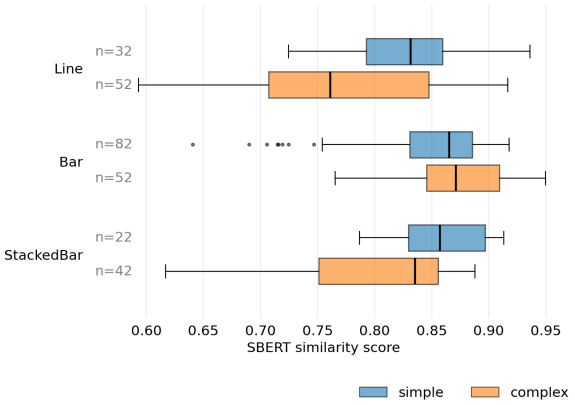


Figure 4: Overall SBERT similarity distribution across chart types and complexity levels.

## 5 Results

The results for quantitative and qualitative evaluations are detailed in the following sections.

### 5.1 Semantic Similarity

Figure 4 shows the overall similarity distribution aggregated across all alt text sections.

Across all three chart types, simple charts generally yield higher and more concentrated similarity scores than complex ones, with the exception of bar charts, where simple and complex conditions show comparable medians and interquartile ranges. Complex line and stacked bar charts show the lowest medians and widest interquartile ranges, reflecting the greater challenge of summarizing multi-layered visualizations. Simple bar charts exhibit several low-scoring outliers, though the bulk of scores remains above 0.80. When broken down by alt text section (not shown), the overview and long description sections diverge more strongly from the gold-standard and exhibit higher variance, which is expected given that these sections require abstraction and trend prioritization rather than extraction of fixed metadata.

### 5.2 Text Length and Conciseness

MLLM-generated texts are consistently longer than gold-standard alt texts and show a heavier tail of very long outputs, particularly for complex charts.

### 5.3 LLM-Judge and Human Evaluation

Overall, both evaluators assign high scores across shared criteria, indicating broad consensus that the generated chart alt texts meet essential quality standards. Median scores are closely aligned for clarity and perceived completeness, suggesting that

LLM judgements approximate human assessments on these dimensions. However, the LLM exhibits greater score variability and applies stricter criteria for conciseness and neutrality than PVIs.

Human scores show the greatest variation in **clarity**. PVIs awarded high clarity ratings when terminology was simple, structures were explicit, and trends were described sequentially. Lower scores were associated with special terminology, ambiguous references, screen reader incompatibilities, and descriptions that assumed the reader could refer back to the chart visually. No systematic directional bias was observed in LLM clarity scores relative to human ratings.

**Conciseness** is the dimension where LLM-judge and human evaluation diverge the most. The LLM applies stricter internal standards, while PVIs tend to tolerate longer descriptions if they are perceived as informative. Users lowered conciseness ratings when descriptions were numerically saturated, redundant, or structured as long enumerations. These were reported as cognitively demanding. Conversely, very short descriptions occasionally raised concerns about missing context.

**Neutrality** scores are high and closely aligned for simple charts. Minor discrepancies emerge for complex charts, where the LLM occasionally penalizes formulations that PVIs still consider neutral. User feedback suggests that neutrality concerns arose less from evaluative language than from structural choices, such as consistently pairing certain entities together in a way that implied an implicit framing.

**Perceived completeness** ratings are similarly distributed across both evaluators, though the LLM shows a slightly wider score range. For PVIs, perceived completeness depended primarily on whether the description conveyed a coherent overall picture rather than exhaustive detail. High scores were given when PVIs could mentally reconstruct the main trends and relationships; lower scores occurred when key contextual anchors, such as start or end values, time references, or explicit statements about the absence of relationships, were missing.

### 5.4 Inter-Generation Consistency

Figure 5 shows pairwise cosine similarity scores across ten generation runs per chart, aggregated across all alt text sections.

Complex charts show somewhat broader distributions, indicating greater output variability when

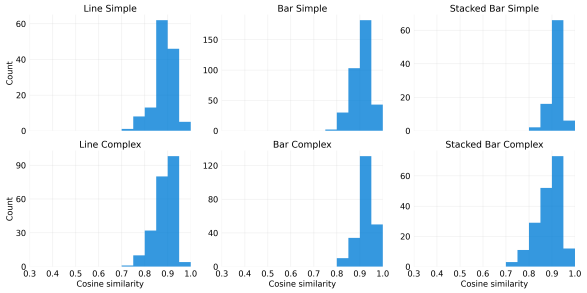


Figure 5: Pairwise cosine similarity scores across ten generation runs, broken down by chart type and complexity.

the model must describe multi-layered visualizations. The overall stability is largely driven by the metadata section, which is nearly deterministic across runs. When examined by section (not shown), the overview and long description exhibit markedly greater variability, particularly for complex charts, confirming that output instability is concentrated in the sections requiring abstraction and trend prioritization rather than in structured factual components.

## 6 Discussion

The results converge on a coherent picture of the strengths and limitations of MLLM-generated chart alt texts. Across all evaluation dimensions, the metadata section emerges as the most robust component: similarity to the gold standard is high and stable, length is well-controlled, and outputs are nearly deterministic across generation runs. This consistency reflects the structured, rule-governed nature of metadata content, which is directly derivable from the underlying chart data.

In contrast, the overview and long description sections are less stable. Semantic similarity to the gold standard is lower and more variable, the generated texts tend to be longer than their manually authored counterparts, and the output consistency across repeated runs is reduced, particularly for complex charts. These patterns are expected since both sections require the model to abstract, prioritize, and synthesize information rather than simply extract it. However, the variability they introduce poses a challenge for deployment in contexts where predictable and consistent output is required.

The quality evaluation broadly corroborates these findings. Both LLM and human evaluators rate the generated texts highly for clarity and completeness, suggesting that the outputs are generally

well-formed and informative. Conciseness is the dimension that most clearly differentiates the generated texts from gold-standard texts, with LLM-generated outputs consistently running longer, particularly for complex charts. The alignment between LLM-judge and PVI ratings is encouraging and suggests that LLM-based evaluation could be useful at least for clarity and perceived completeness. The LLM applies stricter standards for conciseness and neutrality.

A key implication of these findings concerns the role of MLLMs in the generation pipeline. The metadata section, whose content is fixed and deterministic, could be better handled by rule-based templates, which would eliminate the residual variability observed even in this structured section. MLLM generation is most valuable for the overview and long description, where flexible natural language generation adds the greatest benefit and no deterministic alternative exists. Taken together, the findings support a hybrid approach: deterministic template generation for metadata, and carefully prompted MLLM generation for higher-level descriptions. Even in its current form, however, automatically generated alt text represents a substantial improvement over its absence and brings chart accessibility within practical reach for high-volume publishing environments.

A limitation of the current study is the relatively small evaluation set and pool of PVI participants. The findings are based on 27 gold-standard charts and five PVIs, which constrains the statistical robustness of the human evaluation results, even though the effort provides a good starting point and qualitative insights are valuable for this line of work. Additionally, the scope is limited to three chart types from a single German-language publisher, which may limit generalizability, and the results should be seen as the outcomes of preliminary work on a valid, real-world case study.

## 7 Conclusions

This study investigated whether MLLMs can automatically generate well-formed German chart alt texts that meet the needs of PVIs and match the quality of manually authored gold-standards. Focusing on bar, line, and stacked bar charts from the *Neue Zürcher Zeitung* (Neue Zürcher Zeitung), we defined a structured alt text schema, constructed a gold-standard corpus in iterative collaboration with PVIs, and evaluated MLLM-generated descriptions in terms of clarity, conciseness, meaningfulness, and output consistency.

The results demonstrate that MLLM-generated alt texts are a viable and practical solution for improving chart accessibility at scale. Generated texts were consistently rated as clear and complete, and PVIs perceived them as providing meaningful added value. Performance was strongest for simpler charts, where generation was more stable and outputs more closely aligned with the gold-standard. The evaluation also identified a clear division of labor: metadata-based short descriptions could potentially be better generated via deterministic templates, while MLLMs are most effectively employed for overviews and long descriptions where linguistic generation is required. A hybrid pipeline combining both approaches is a practical and scalable strategy.

Several directions for future work emerge from this study. First, expanding the gold-standard corpus to cover a broader range of chart types, complexity levels, and source domains would improve generalizability. Second, enlarging and diversifying the pool of PVI participants would strengthen the statistical robustness of the human evaluation. Third, integrating inter-generation consistency analysis directly into the evaluation pipeline would support ongoing quality monitoring in production settings.

## Limitations

The qualitative evidence is based on a small sample of five PVIs and a limited set of charts. Thus, the findings may not represent the full diversity of screen-reader use and information needs. Furthermore, SBERT similarity measures semantic overlap, but it cannot assess structural compliance, cognitive load, or screen-reader-specific issues. Complex charts remain the main challenge: even accurate alt texts can be cognitively demanding.

## Acknowledgments

We would like to thank the *Schweizerischer Blinden- und Sehbehindertenverband* (SBV) and all participants who contributed to the refinement of the gold-standard corpus and the evaluation of the generated alt texts, as well as the *Neue Zürcher Zeitung* (NZZ) for providing the chart dataset used in this study.

## References


AbilityNet. 2025. An introduction to screen readers. https://abilitynet.org.uk/factsheets/introduction-screen-readers.

Accessibility Guidelines Working Group (AG WG). 2023. Web Content Accessibility Guidelines (WCAG) 2.2 Quick Reference. https://www.w3.org/WAI/WCAG22/quickref/.

Abhijit Balaji, Thuvaarakkesh Ramanathan, and Venkateshwarlu Sonathi. 2018. Chart-Text: A Fully Automated Chart Image Descriptor. *arXiv preprint*.

Aspen Belle, Vanessa Goh, Akshay Kumar, Richard Pranjatno, Pui Yip, Umayangani Wickramaratne, and Humphrey Obie. 2022. Alt-Texify: A Pipeline to Generate Alt-text from SVG Visualizations:. In *Proceedings of the 17th International Conference on Evaluation of Novel Approaches to Software Engineering*, pages 275–281, Online Streaming, — Select a Country —. SCITEPRESS - Science and Technology Publications.

Arzu Çöltekin, Amy Griffin, and Anthony Robinson. 2021. *Visualizations*. Oxford University Press.

Consumer Financial Protection Bureau. Data visualization guidelines. https://cfpb.github.io/design-system/guidelines/data-visualization-guidelines.

DIAGRAM Center. 2020. Specific Guidelines - Graphs.

European Union. 2019. Directive (EU) 2019/882 of the European Parliament and of the Council of 17 April 2019 on the accessibility requirements for products and services.

Ali Mazraeh Farahani, Peyman Adibi, Mohammad Saeed Ehsani, Hans-Peter Hutter, and Alireza Darvishy. 2023. Automatic Chart Understanding: A Review. *IEEE Access*, 11:76202–76221.

Cole Gleason, Patrick Carrington, Cameron Cassidy, Meredith Ringel Morris, Kris M. Kitani, and Jeffrey P. Bigham. 2019. "It's almost like they're trying to hide it": How User-Provided Image Descriptions Have Failed to Make Twitter Accessible. In *The World Wide Web Conference*, pages 549–559, San Francisco CA USA. ACM.

Li Gong, Josep Crego, and Jean Senellart. 2019. Enhanced Transformer Model for Data-to-Text Generation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 148–156, Hong Kong. Association for Computational Linguistics.

Google DeepMind. 2025. Gemini 2.5 Flash (Preview).

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A Survey on LLM-as-a-Judge. *Preprint*, arXiv:2411.15594.

E. Hoque and M. Saidul Islam. 2025. Natural Language Generation for Visualizations: State of the Art, Challenges and Future Directions. *Computer Graphics Forum*, 44(1):e15266.

Kung-Hsiang Huang, Hou Pong Chan, May Fung, Haoyi Qiu, Mingyang Zhou, Shafiq Joty, Shih-Fu Chang, and Heng Ji. 2025. From Pixels to Insights: A Survey on Automatic Chart Understanding in the Era of Large Foundation Models. *IEEE Transactions on Knowledge and Data Engineering*, 37(5):2550–2568.

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.

Shankar Kantharaj, Rixie Tiffany Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. 2022. Chart-to-Text: A Large-Scale Benchmark for Chart Summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4005–4023, Dublin, Ireland. Association for Computational Linguistics.

Julia Locher and Alessia Vannini. 2025. Assisting People with Visual Impairments to Understand Data Visualizations. Technical Report 25FS_IIT47, University of Applied Sciences and Arts Northwestern Switzerland (FHNW), School of Computer Science, Windisch, Switzerland.

Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

Omar Moured, Shahid Ali Farooqui, Karin Müller, Sharifeh Fadaeijouybari, Thorsten Schwarz, Mohammed Javed, and Rainer Stiefelhagen. 2024. Alt4Blind: A User Interface to Simplify Charts Alt-Text Creation. In Klaus Miesenberger, Petr Peňáz, and Makoto Kobayashi, editors, *Computers Helping People with Special Needs*, volume 14750, pages 291–298. Springer Nature Switzerland, Cham.

Neue Zürcher Zeitung. NZZ – Neue Zürcher Zeitung. https://www.nzz.ch.

Kai Nylund, Jennifer Mankoff, and Venkatesh Potluri. 2025. MatplotAlt: A Python Library for Adding Alt Text to Matplotlib Figures in Computational Notebooks. *Computer Graphics Forum*, 44(3):e70119.

Jason Obeid and Enamul Hoque. 2020. Chart-to-Text: Generating Natural Language Descriptions for Charts by Adapting the Transformer Model. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 138–147, Dublin, Ireland. Association for Computational Linguistics.

OpenAI. 2024. O4-mini.

OpenRouter. OpenRouter API Reference. https://openrouter.ai/docs/api/reference/parameters.

Pennsylvania State University. 2024. Charts & Accessibility.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Preprint*, arXiv:1908.10084.

Statista. Statista. https://de.statista.com/.

Benny Tang, Angie Boggust, and Arvind Satyanarayan. 2023. VisText: A Benchmark for Semantically Rich Chart Captioning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7268–7298, Toronto, Canada. Association for Computational Linguistics.

Jiaqi Wang, Hanqi Jiang, Yiheng Liu, Chong Ma, Xu Zhang, Yi Pan, Mengyuan Liu, Peiran Gu, Sichen Xia, Wenjun Li, Yutong Zhang, Zihao Wu, Zhengliang Liu, Tianyang Zhong, Bao Ge, Tuo Zhang, Ning Qiang, Xintao Hu, Xi Jiang, and 5 others. 2024. A Comprehensive Review of Multimodal Large Language Models: Performance and Challenges Across Different Tasks. *arXiv preprint*.

WebAIM. 2026. WebAIM: The WebAIM Million. https://webaim.org/projects/million/.

Chuqiao Yan, Hans-Peter Hutter, Felix M. Schmitt-Koopmann, and Alireza Darvishy. 2025. Chart Accessibility: A Review of Current Alt Text Generation. *IEEE Access*, 13:94040–94056.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.