Human vs LLM-Generated Alt Text for STEM Image Accessibility: A Comparative Study

Marco Cardia
Department of Computer Science
University of Pisa
Pisa, Pisa, Italy
marco.cardia@di.unipi.it

Letizia Angileri
Department of Computer Science
University of Pisa
Pisa, Pisa, Italy
letizia.angileri@di.unipi.it

Camilla Poggianti
Department of Computer Science
University of Pisa
Pisa, Pisa, Italy
camilla.poggianti@phd.unipi.it

Barbara Leporini
Department of Computer Science
University of Pisa
Pisa, Pisa, Italy
ISTI-CNR
Pisa, Pisa, Italy
barbara.leporini@unipi.it

Abstract

Image descriptions are a fundamental aspect of web accessibility, allowing blind and low-vision users to access visual information through assistive technologies such as screen readers. Despite their importance, high-quality alternative text is often missing or inadequate, especially for complex images like diagrams, graphs, and other STEM-related content. Recent advances in generative artificial intelligence and large language models (LLMs) have renewed interest in automatically generating image descriptions, raising the question of whether these systems can reliably support accessibility at scale.

In this paper, we present a comparative study of human-written and LLM-generated alternative text for STEM images, focusing on accessibility-critical aspects rather than surface-level textual similarity. Using a curated dataset of images with expert-authored reference descriptions, we evaluate the outputs of state-of-the-art multimodal LLMs through a mixed-methods approach. Our evaluation combines traditional automated metrics, such as BLEU and METEOR, with a human-centered analysis targeting accuracy, informational completeness, and the presence of hallucinations.

The results show that while LLMs often produce fluent and seemingly informative descriptions, substantial gaps remain compared to human-written alt text, particularly in conveying formal semantics and essential structural details required for accessibility. We further observe that commonly used automated metrics only partially capture accessibility-relevant errors, underscoring the need for evaluation methodologies based on the needs of screen reader users. We discuss the implications of these findings for the design, evaluation, and deployment of generative AI systems in accessible web and educational contexts.

W4A '26, Dubai, United Arab Emirates

ACM ISBN 979-8-4007-2372-8/26/04
https://doi.org/10.1145/3800424.3800457

CCS Concepts

• **Human-centered computing** → **Accessibility design and evaluation methods**.

Keywords

Web accessibility, Alternative Text, Large Language Models, Image Description, STEM images

ACM Reference Format:
Marco Cardia, Letizia Angileri, Camilla Poggianti, and Barbara Leporini. 2026. Human vs LLM-Generated Alt Text for STEM Image Accessibility: A Comparative Study. In *Web4All 2026 23rd International Web for All Conference (W4A '26), April 13–14, 2026, Dubai, United Arab Emirates.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3800424.3800457

1 Introduction

In recent years, the automatic generation of alternative descriptions (alt text) for images has become a central focus in digital accessibility research, particularly for supporting access to visual content through screen readers [23]. In this paper, we use "alternative text" to refer to a textual description associated with a non-textual element (such as an image, diagram, or chart) that conveys its content and function to users who cannot perceive it visually. Traditionally, authors or accessibility specialists with domain knowledge have manually created alt text; however, this process is time-consuming, labor-intensive, and often omitted or minimized in practice, resulting in the exclusion of screen reader users from important information conveyed by visual representations [7]. This exclusion is especially problematic in educational and scientific contexts, where figures and diagrams are essential for communicating complex concepts and supporting learning.

The advent of large language models (LLMs) and recent multimodal AI systems has created new opportunities for automatically generating richer, more context-aware image descriptions. These models can interpret visual content and produce textual outputs with increasing fluency and semantic detail, inspiring a growing body of research on generative AI for accessibility. Studies have examined the use of LLMs for captioning and visual understanding

tasks, yet persistent challenges remain, especially for images with formal notations, structured layouts, and domain-specific semantics typical of STEM content [15]. While early work on vision–language models has shown promising results on everyday images, these approaches often fall short when generating reliable alt text for technical figures, where precision, clarity, and conceptual correctness are essential [1].

In web accessibility research and practice, missing or inadequate alt text remains one of the most frequently reported barriers for users with visual impairments, significantly affecting usability across devices and platforms [7]. Recent studies emphasize that the lack of meaningful image descriptions impedes access not only to general web visuals but also to scholarly and educational materials, which increasingly rely on diagrams, graphs, and other STEM images [14]. Motivated by the need for empirical evaluations of LLM-based systems in accessibility contexts, we examine the current capabilities and limitations of generative models in producing alt text for STEM images.

Building on this background, recent exploratory work [5] has begun to investigate the maturity of AI- and LLM-based tools for generating alternative descriptions of images with STEM content. These early analyses reveal a mixed picture: while generative systems show promising capabilities in producing fluent and seemingly informative descriptions, they also exhibit substantial limitations in conceptual accuracy, informational completeness, and alignment with the actual needs of screen reader users. In particular, errors, omissions, and hallucinated elements remain frequent when models are asked to describe scientific images that rely on formal semantics, symbolic conventions, and precise structural relationships.

Given the rapid evolution of multimodal LLMs, it is necessary to update and extend these findings through systematic and reproducible evaluations. This work addresses this need by examining the current capabilities of state-of-the-art LLMs in generating alternative descriptions for STEM images when guided by structured, explicit, and accessibility-oriented prompts. The study is conducted within the framework of the STEMMA project, which aims to explore methods and tools for reducing digital barriers to STEM content and supporting inclusive educational practices.

We present a comparative evaluation of automatically generated alternative descriptions for a set of STEM images from the field of computer science, assessing the alignment between model-generated descriptions and expert-defined reference descriptions. To capture different interaction paradigms commonly used by users and developers, we evaluate three multimodal LLMs under two experimental conditions. In a batch setting, each model receives a collection of images in a single input, reflecting scenarios where content creators or systems process multiple figures at once. In a complementary sequential setting, images are submitted individually, mirroring more interactive and incremental usage patterns. In both conditions, models are prompted with the same structured instructions, designed to emphasize accessibility requirements and the inclusion of essential semantic information.

For each generated alternative description, we assess descriptive accuracy and identify any hallucinations by comparing the output to a reference description that captures the essential informational content needed for accessibility. By analyzing model behavior across various input modalities and testing conditions, this work aims to clarify the reliability and limitations of current LLM-based approaches to scientific alt text generation.

The contributions of this paper are as follows. First, we introduce a procedure for evaluating alternative descriptions of STEM figures through systematic comparison with expert reference descriptions, focusing on accuracy and informational completeness in domains with formal semantics. Second, we define and apply a structured, reproducible prompt explicitly designed to guide LLMs toward accessibility-oriented descriptions. Third, we present a comparative analysis of multiple multimodal LLMs under different testing conditions, highlighting strengths, recurring errors, and implications for the practical use of generative AI in producing accessible scientific content.

The remainder of the paper is organized as follows. Section 2 discusses the accessibility challenges of scientific figures and reviews existing approaches to alt text generation in the STEM domain. Section 3 describes the image set, the construction of the reference descriptions, and the structured prompting strategy. Section 4 details the experimental protocol, evaluation metrics, and comparison methodology. Section 4 presents and discusses the results, with particular attention to common error patterns and hallucinations. Finally, Section 6 concludes the paper and outlines directions for future work.

2 Related Work

Automatic image description has traditionally relied on large collections of web images. Over the past decade, progress in image captioning has been driven by the availability of large-scale datasets such as Flickr8k/30k [11], MS COCO [17], and VizWiz [9]. These datasets have enabled the training of models that generate descriptions for natural scenes and everyday objects, resulting in substantial improvements in performance and fluency. However, when applied to the STEM domain — characterized by diagrams, mathematical expressions, symbolic notations, and structured layouts — such models show significant limitations. The main issue is the lack of training data that captures symbols, formal relationships, and domain-specific visual conventions, which are central to STEM imagery.

To address this gap, researchers have introduced datasets specifically designed for diagrammatic and technical content. RL-CSDia [20], for example, provides thousands of computer science diagrams, including trees, graphs, and flowcharts, annotated to support the learning of abstract and relational representations. Similarly, Agrawal et al. [2] introduced a dataset of hand-drawn diagrams that better reflect the variability, noise, and imprecision typical of real educational settings. Together, these datasets highlight the importance of evaluating models on both clean, computer-generated diagrams and informal, student-produced representations, as both frequently appear in instructional materials.

Research on diagram understanding shows that effective interpretation goes beyond detecting individual visual elements. Comprehension requires identifying and reasoning about the relationships that connect symbols, shapes, and spatial groupings. Kembhavi et al. formalized this concept through Diagram Parse Graphs (DPGs) [12], proposing a framework that integrates visual recognition with structural parsing and symbolic reasoning. A similar

perspective appears in studies on mathematical notation, such as those by Zanibbi and Blostein [24], which highlight how spatial organization — relative positioning, alignment, and grouping — directly contributes to semantic meaning. These findings indicate that accurate diagram descriptions must encode both visual features and underlying logical structures.

The emergence of LLMs has created new opportunities for generating richer, more context-aware image descriptions. Recent research shows that LLM-based methods can produce explanations that are more detailed, coherent, and linguistically flexible than those from earlier vision–language models [19]. However, LLMs still struggle with STEM diagrams, especially when descriptions require explicit reasoning about hierarchical structures, causal flows, or formal dependencies among elements. This limitation raises concerns about the reliability of automatically generated alt text for technical images. Recent empirical evaluations reinforce these concerns in accessibility-oriented educational contexts. Buzzi et al. [5] examined the maturity of generative AI for alternative descriptions of STEM images by evaluating several AI-powered digital assistants and accessibility-focused mobile applications on a diverse set of STEM figures, including mathematical expressions, scientific diagrams, and computational models. Their findings indicate that conversational AI systems often provide richer and more context-aware descriptions than traditional mobile applications; however, they also reveal frequent inaccuracies, omissions, and inconsistencies, particularly when diagrams require explicit reasoning about structure, relationships, or domain-specific semantics. The study also highlights the critical role of prompt formulation: progressively specifying the user's visual impairment and explicitly requesting a graphic or educational description significantly alters the level of detail and relevance of the output, yet does not guarantee correctness or pedagogical adequacy.

Research on accessibility in STEM education underscores the critical role of high-quality descriptions in ensuring equitable access to learning materials. Universal design principles and studies on inclusive educational practices [4] consistently show that students with disabilities face substantial barriers when diagrams and symbolic content lack adequate textual representations or clear organization. This research highlights that accessibility is not merely a technical challenge but also a pedagogical one, requiring descriptions that are accurate, informative, and aligned with learners' needs. Prior work on datasets, modeling approaches, and accessibility frameworks converges toward a shared objective: developing systems capable of generating meaningful and truly accessible descriptions of STEM images.

A further challenge in this area is the lack of standardized benchmarks and evaluation datasets explicitly designed for accessibility-oriented descriptions of STEM images. As noted by [6], most existing evaluations rely on datasets and metrics originally developed for generic image captioning or diagram understanding, which do not capture whether a description provides meaningful and trustworthy access for blind and low-vision users. This absence of dedicated benchmarks limits comparability across approaches and obscures accessibility-critical failure modes such as hallucination, over-abstraction, and omission of structurally relevant information. Consequently, progress in generative approaches for accessible STEM image description is constrained not only by model capabilities but also by the lack of evaluation resources aligned with accessibility and educational requirements.

3 Methodology

In this paper, we evaluate the generation of alternative descriptions for automata diagrams under two distinct usage scenarios: single-image interactive description and fully automated batch description. The adoption of both setups allows us to understand the quality of model-generated description when focusing on a single image with a single task, and the impact of providing multiple images with a single task. Across all experiments we compare model-generated descriptions with human-written references and evaluate them using a combination of automated metrics and human-centered qualitative assessment. The inclusion of qualitative assessment allows us to investigate whether standard automated metrics are sufficiently sensitive to the specific types of errors found in complex scientific imagery.

3.1 Dataset

The dataset consists of 73 images representing automata diagrams commonly used in introductory computer science and formal language theory. The images include deterministic and non-deterministic finite automata with varying structures, number of states, transition labels, loops, and accepting states. Many images are intentionally similar in visual appearance while representing distinct automata, making them particularly sensitive to descriptive errors. The images were extracted from undergraduate computer science lecture slides, representing the visual complexity that students encounter in a standard curriculum. We selected 73 images as these represent the full set of automata included in the source lecture slides.

Each image is associated with a human-written alternative description produced following accessibility-oriented principles, in particular the W3C and the Diagram Center guidelines [8, 21] which are in compliance with the Web Content Accessibility Guidelines (WCAG) 2.2 Level AA [22]. The alternative descriptions used as a baseline were written by the four authors, all of whom have expertise in accessibility or computer vision. Importantly, one of the author is also an end user who regularly relies on screen readers. The human annotations are constructed following a domain-level interpretation, describing the functional and semantic interpretation (e.g. State Q1 transits to state Q2 on input 'a') of the images, rather than a geometric level description (e.g. A circle with an arrow). Consequently, these descriptions are intended for readers familiar with automata theory. These descriptions serve as reference material for evaluation, but are not assumed to be a single canonical ground truth; rather, they represent accessibility-informed human judgments. The dataset is used exclusively for evaluation purposes and is not employed for training or fine-tuning any of the models.

3.2 Models and Generative Setup

We evaluate multiple multimodal large language models with vision-language capabilities, capable of generating textual descriptions from image inputs. Different models are used depending on the experimental setup.

For the single-image interactive setup, we evaluate the following models:

- ChatGPT 5.2
- Gemini 3 Pro
- Claude Sonnet 4.5

These models are queried individually for each image, with a new context for every request.

For the batch processing setup, we evaluate a subset of multimodal models capable of processing multiple images within a single execution context. Specifically, we consider:

- **Gemini 3 Pro**
- **Claude 3.5 Sonnet**
- **GPT-OSS 120b**

The models are required to analyze the entire dataset of images and generate alternative descriptions for all images in one structured output. All models are evaluated under identical experimental conditions to ensure comparability.

3.2.1 Single Image Generative Setup. In the single image setup, wach image is processed independently. For every image, a new request is issued to the model, and no contextual information from previous images is retained. This configuration is designed to replicate a real-time interaction scenario, such as on-demand image description within the context of assistive technology. Each model receives (i) a single image as input, (ii) a designed prompt specifying the generation of an alternative description for accessibility, (iii) no information about other images in the dataaset.

The prompt instructs the model to describe only what is visible in the image, to avoid guessing missing or unreadable elements, and to produce an accessibility-oriented description comparable to human-written alternatives. Because each image is processed in isolation, the model benefits from maximal perceptual grounding and minimal contextual interference.

This setup aims to approximate best-case usage conditions for multimodal language models in assistive technologies.

3.2.2 Batch Generative Setup. In the batch setup, models are required to generate alternative descriptions for all images in the dataset within a single, non-interactive generation process. Rather than issuing one request per image, a single instruction is used to prompt the model to analyze a collection of images and return a structured output containing one description per image. Each model is provided with (i) A fixed system-level prompt, (ii) access to the whole dataset composed of all the images, and (iii) requirement to generate a single structured textual output in response.

Descriptions are generated without per-image resets, follow-up questions, or corrective feedback. All images are processed within the same execution context, requiring the model to maintain consistency and grounding across multiple inputs.

The batch setup is designed to approximate large-scale accessibility remediation scenarios, such as automatic generation of alternative text for image repositories, learning platforms, or archival datasets. Unlike the single-image setup, this configuration introduces greater information density and potential contextual interference, allowing us to study model robustness and failure modes under realistic automation constraints.

All models are queried using exactly the same prompt, without any model-specific adaptation. The prompt is designed to elicit domain-level alternative descriptions suitable for users familiar with the relevant STEM domain, while explicitly forbidding unsupported inference or guessing.

The instructions explicitly require models to describe only what can be inferred from each image, to avoid guessing missing or unreadable information, and to mark unclear elements when necessary. All models are required to process the entire dataset within a single execution context, without resetting between images.

The full prompt used in all experiments is reported in Listing 1..

Listing 1: Full system prompt used in the batch description experiments.

```
You are an expert in accessibility and mathematical logic,
     specialized in converting STEM technical images (graphs
     , trees, automata, set/venn diagrams, logic circuits,
     plots, proofs/derivations, tables, geometry, etc.) into
      rigorous alternative text.

GOAL
Provide full cognitive access to the visual information for
     users who cannot see the image and need a precise,
     domain-level interpretation. You must produce a correct
      description for EACH specific image in the dataset
     folder.

INPUT
The dataset folder containing multiple images.

OUTPUT (MANDATORY)
Produce a single CSV file as your final output (no extra
     commentary).
The CSV must contain exactly one row per image.
Include a header row.
Use UTF-8.
Use commas as separators.
If any field contains commas or quotes, properly CSV-escape
     it (wrap in double quotes, escape internal quotes by
     doubling them).

CSV COLUMNS (MANDATORY)
image_filename
semantic_description

SEMANTIC DESCRIPTION REQUIREMENTS (MANDATORY)
Write a domain-level interpretation of what the image
     represents, assuming the reader knows the relevant STEM
      domain (e.g., for automata, you may say "states", "
     transitions", "accepting state", "alphabet symbol", "
     DFA/NFA" if supported by the image).
Describe only what can be inferred from the image itself.
Do NOT guess missing labels, unreadable text, or unstated
     semantics.
If text/labels are too small or unclear, explicitly say "
     unreadable" or "unclear" and continue with what is
     visible.
Be specific: mention the actual labels, symbols, node names,
      transition labels, axes labels, legends, numeric
     values, set names, etc., when readable.
```

```
If the diagram type is identifiable (e.g., DFA, NFA, rooted
     tree, directed graph, bipartite graph, Venn diagram,
     Karnaugh map, plot with axes, confusion matrix), state
     it.
If the diagram includes multiple subfigures/panels, describe
      each panel and how they relate.

PROCESS (MANDATORY)
For each image in the folder:
- Determine the diagram type (automaton/graph/tree/plot/etc
     .).
- Extract all readable textual elements (node labels, edge
     labels, axis titles, legends, set labels, formulae, etc
     .).
- Write an image-specific semantic alternative description
     that captures the mathematical/scientific meaning as
     shown.

FINAL RESPONSE FORMAT (MANDATORY)
Return ONLY the CSV content (starting with the header row).
     Do not include Markdown fences, explanations, or
     anything outside the CSV.
```

This prompting and execution strategy was intentionally chosen to evaluate the models under fully automated conditions, without human intervention or corrective feedback. We remark that while such a setup does not reflect best case interactive usage, it closely approximates realistic deployment scenarios in which accessibility descriptions are generated at scale through programmatic model calls. As a consequence, the evaluation focuses on: robustness to reduced perceptual grounding, adherence to explicit constraints against guessing, consistency across multiple images, and susceptibility to hallucination in accessibility-critical contexts.

We compute standard automated text-similarity metrics between model-generated descriptions and human-written references. Specifically, we report BLEU, ROUGE, and METEOR scores with a comparison between the generated text with the human written description. These metrics measure lexical overlap and paraphrasing similarity but do not directly capture accessibility-relevant qualities such as factual correctness, completeness, or the presence of hallucinated content. In the context of automata diagrams, two descriptions may differ substantially in wording while conveying equivalent information, or conversely may score highly while sharing the same incorrect inference. On the other hand, in the xontext of STEM images and in the context of automata diagrams, these metrics may be particularly misleading: a model can achieve a high score by mimicking the expected technical style and structure, yet fail on critical details such as transition labels or state names. Such fine-grained errors are catastrophic for accessibility but often have a marginal impact on n-gram-based scores.

To complement these reference-based metrics, we also incorporate CLIPScore as a reference-free evaluation measure. Unlike lexical metrics that are sensitive to exact wording, CLIPScore leverages a shared multimodal embedding space to assess the semantic alignment between the generated description and the original image. This allows us to capture the degree of visual grounding, providing an evaluation that remains independent of the specific phrasing used in human references

To assess the quality of the generated descriptions, we employed the following metrics:

- **BLEU-4**: Measures n-gram precision overlap between the generated text and the reference [18].
- **METEOR**: Accounts for synonymy and stemming, providing better correlation with human judgment [3, 13].
- **ROUGE-L**: Measures the longest common subsequence, capturing sentence-level structure [16].
- **CLIPScore**: Uses the CLIP model to measure the semantic similarity and compatibility between the original image and the generated text, without relying on the reference text [10].

3.3 Human-Centered Evaluation

To address the limitations of the automated evaluation metrics and to capture the actual utility of the descriptions, we complement the automated analysis with human-centered qualitative assessment. This manual review focuses on the semantic accuracy of the STEM components that automated metrics may overlook. The descriptions generated under both setups were independently evaluated by the four authors of the paper. Any disagreements were discussed and resolved through consensus.

To measure the functional correctness of the descriptions, we defined the following metrics, where S denotes the set of states and T denotes the set of transitions.

- **State Coverage (SC)**: The percentage of states correctly identified and counted relative to the total number of states in the ground truth.

$$SC = \frac{|S_{\text{correct}}|}{|S_{\text{total}}|} \times 100 \tag{1}$$

- **Transition Coverage (TC)**: The percentage of transitions correctly identified.

$$TC = \frac{|T_{\text{correct}}|}{|T_{\text{total}}|} \times 100 \tag{2}$$

- **Initial State Identification (ISI)**: Binary metric indicating if the start state is correctly identified (1) or not (0).
- **Accepting State Identification (ASI)**: The count of correct accepting (final) states identified.

$$ASI = |S_{final} \cap S_{identified\_final}| \tag{3}$$

- **Hallucination Count (HC)**: The number of non-existent elements (states or transitions) invented by the model.
- **Hallucination Count Normalized (HCN)**: The number of hallucinations normalized by the total complexity of the automaton.

$$HCN = \frac{HC}{|S_{total}| + |T_{total}|} \tag{4}$$

This evaluation is particularly important for identifying accessibility-relevant failures that are not captured by automated metrics, especially in cases where hallucinated information could mislead blind users.

Table 1: Results for Batch Image Description. Where SC=State Coverage; TC=Transition Coverage; ISI=Initial State Identification; ASI=Accepting State Identification; HC=Hallucination Count; HCN=Hallucination Count Normalized. HC is reported as total sum.

Model	SC (%)	TC (%)	ISI (%)	ASI	HC	HCN
Gemini[1]	98.8 ± 6.9	92.2 ± 8.2	**100.0 ± 0.0**	0.95 ± 0.16	**13**	0.08 ± 0.02
GPT OSS	**100.0 ± 0.0**	**97.6 ± 6.6**	**100.0 ± 0.0**	**0.99 ± 0.05**	20	**0.06 ± 0.02**
Claude	96.8 ± 15.1	68.7 ± 30.8	96.6 ± 15.8	0.87 ± 0.53	104	0.21 ± 0.14

Table 2: Results for Single Image Interactive Description. Where: SC=State Coverage; TC=Transition Coverage; ISI=Initial State Identification; ASI=Accepting State Identification; HC=Hallucination Count; HCN=Hallucination Count Normalized. HC is reported as total sum.

Model	SC (%)	TC (%)	ISI (%)	ASI	HC	HCN
Gemini	**100.0 ± 0.0**	**99.2 ± 4.4**	**100.0 ± 0.0**	**1.00 ± 0.00**	**10**	**0.005 ± 0.003**
GPT-5.2	100.0 ± 0.0	98.6 ± 4.3	90.4 ± 9.4	**1.00 ± 0.00**	18	0.008 ± 0.003
Claude	99.6 ± 0.2	88.2 ± 15.9	**100.0 ± 0.0**	0.86 ± 0.30	204	0.11 ± 0.011

3.4 Annotation Protocol and Reliability

Because our reference descriptions are intentionally domain-level and reflect accessibility-informed human judgments rather than a single canonical ground truth, it is important to specify how we define correctness when computing the structural measures introduced in Section 3.3. Our objective is to assess whether a generated description provides reliable cognitive access to the automaton's semantics (states, initial and accepting markers, and labeled transitions) rather than rewarding textual style or surface similarity. First, for each image, we score outputs at the level of formal elements: states (S) and transitions (T). A state or transition is counted as correct only if it is supported by the diagram and expressed with sufficient specificity to reconstruct the automaton. This element-based approach reflects the core risk highlighted in this paper: fluent text can still be structurally wrong in ways that are critical for accessibility. A predicted state contributes to $S_{correct}$ if it unambiguously refers to a state label present in the diagram; generic placeholders ('a state', 'another node') are not counted as correct. *ISI* is correct only if the output explicitly identifies the start state indicated by the incoming start arrow. *ASI* is correct only if the output identifies the accepting or final state(s) and their role (e.g., by naming them as final or accepting, or by clearly referring to the diagram's accepting marker). Instead, a predicted transition contributes to $T_{correct}$ only if it matches (i) the source, (ii) the target, and (iii) the label when present. For unlabeled edges, correctness requires not inventing a label; for self-loops, the loop must be attributed to the correct state. When a label contains multiple symbols (e.g., 'a,b'), we treat it as a single labeled transition unless the description explicitly decomposes it into an equivalent representation without altering semantics; incorrect splitting or merging is treated as label corruption. *HC* increments when the output introduces states, transitions, or labels not present in the diagram, including incorrect source or target pairs asserted as existing edges. We count hallucinations only when they alter the automaton's formal structure available to a screen reader user.

4 Results

In this paper we compare model behavior across two complementary scenarios:

- Batch, fully automated description, representing large-scale accessibility remediation.
- Single-image, interactive description, representing best-case assistive use.

By evaluating both setups, we are able to distinguish between model capabilities under optimal grounding conditions and their robustness when deployed in realistic automated pipelines.

4.1 Batch Results

Table 1 reports the performance metrics for the batch API processing task. In this setting, models were tasked with generating descriptions for the entire dataset in a single batch process. Values are reported as Mean ± Standard Deviation.

As shown in Table 1, GPT demonstrated superior performance in the batch setting, achieving perfect State Coverage. It correctly identified every state and label in the dataset, with a Transition Coverage of 97.6% (± 6.6). Gemini followed closely, with a high State Coverage of 98.8%, though slightly lower Transition Coverage (92.2% ± 18.2). Notably, Gemini produced the fewest total hallucinations (13), compared to 20 for GPT. However, its reliability was hampered by a significant refusal rate; the model failed to provide an alternative description for 38 images in the dataset, providing a description that cannot be considered alternative description. Claude exhibited significant difficulties in this mode, with a much lower Transition Coverage (68.7% ± 30.8), and a drastically higher total Hallucination Count (104). Beyond these metrics, Claude experienced 16 cases of catastrophic failure, where the output was entirely unrelated to the source diagram. These results indicate a

tendency to fabricate non-existent graph elements and a lack of robustness under realistic automation constraints.

4.1.1 Single Image Results. Table 2 presents the results for the interactive, single-image description task. This setting reflects a typical user workflow of uploading an image to a chat interface.

In the single-image interactive setting, the performance landscape shifts dramatically. Gemini achieves perfect scores in State Coverage (100%), and a near-perfect Transition Coverage of 99.2% (± 4.4). It also maintained the lowest Hallucination Count (10). GPT5.2 showed poorer performance with TC of 98.6% (± 4.3). Claude's performance drops with extremely high Hallucination Count of 204 and a Transition Coverage of 88.2%. This represents a systematic failure to accurately perceive or describe the automata structure in this interactive mode.

4.2 Automated Metrics Analysis

Table 3 and Figure 1 report the results of the automated evaluation metrics for the single-image interactive setup, including BLEU-4, METEOR, ROUGE-L, and CLIPScore. These metrics quantify different aspects of similarity between model-generated descriptions and the corresponding human-written references, or between text and image content.

Overall, absolute scores for n-gram–based metrics (BLEU-4 and ROUGE-L) remain low across all models. This result is expected given the nature of the task: alternative descriptions for complex STEM diagrams allow for substantial lexical and syntactic variation while still being correct and accessible. Consequently, low lexical overlap does not necessarily imply poor descriptive quality, nor does higher overlap guarantee accessibility correctness.

Table 3: Automated evaluation metrics for single image descriptions. Gemini achieved the highest lexical scores (BLEU-4, METEOR, ROUGE-L), while Claude achieved the highest CLIPScore. Best scores are bolded.

Model	BLEU-4	METEOR	ROUGE-L	CLIPScore
Gemini	**0.051**	**0.257**	**0.369**	20.07
GPT5.2	0.025	0.217	0.307	20.29
Claude	0.035	0.236	0.312	**22.31**

Among the evaluated models, Gemini achieves the highest scores for BLEU-4, METEOR, and ROUGE-L. This trend is consistent with the structural evaluation reported in Section 4.1.1, where Gemini also achieved near-perfect State Coverage and Transition Coverage with the lowest hallucination rates. In this case, automated metrics partially reflect genuine descriptive accuracy, as the model's outputs closely align with the content and structure of the human-written references.

However, the results for Claude illustrate a critical limitation of automated metrics. Despite exhibiting the highest hallucination count and substantially lower Transition Coverage in the human-centered evaluation, Claude achieves the highest CLIPScore among the three models. This indicates that CLIPScore captures broad semantic compatibility between the generated text and the image, such as recognizing that the image depicts a diagram with circles and arrows, but remains insensitive to fine-grained structural correctness. For accessibility purposes, this distinction is crucial: a description can be semantically "compatible" with an image while failing to convey the precise relationships required for understanding an automaton.

Similarly, BLEU and ROUGE scores fail to penalize severe structural errors. Claude's automated scores are comparable to those of GPT5.2 despite the presence of an order-of-magnitude difference in hallucination counts. This suggests that n-gram–based metrics can be satisfied by fluent, domain-appropriate language even when the described states, transitions, or accepting conditions do not correspond to the actual diagram. From an accessibility perspective, such errors are catastrophic, as they may lead blind users to construct an entirely incorrect mental model of the automaton.

The batch setup further reinforces this observation, as shown in Table 4. GPT-OSS achieves the highest BLEU-4, METEOR, and ROUGE-L scores, aligning with its strong structural performance in batch mode. In contrast, Gemini attains the highest CLIPScore despite lower lexical similarity to the references, again highlighting that CLIP-based metrics prioritize coarse image–text alignment over formal correctness.

Taken together, these findings demonstrate that automated metrics alone are insufficient for evaluating the accessibility of the descriptions of technical diagrams. While such metrics provide useful signals about linguistic fluency and general semantic relevance, they systematically fail to capture the accuracy and completeness of formal structures that are essential for STEM comprehension. In particular, hallucinations involving states, transitions, or labels often have minimal impact on automated scores despite representing severe accessibility failures.

These results underscore the necessity of combining automated metrics with domain-specific, human-centered evaluation when assessing LLM-generated alternative text for STEM images. Without explicit structural validation, automated scores risk overestimating the practical accessibility of generated descriptions.

5 Discussion

The results of this study highlight a significant disconnect between the linguistic fluency of multimodal LLMs and their structural reliability for STEM accessibility. While model outputs often appear technically authoritative, they frequently lack the semantic precision required to support functional mental models for screen reader users. In this section, we analyze the performance disparities across different usage modes, the limitations of standard automated metrics, and the broader implications for the safe deployment of generative AI in accessible STEM education.

5.1 Performance Disparities Across Settings

The results highlight a significant performance disparity between batch and interactive modes, particularly for GPT and Claude. While GPT excelled in batch processing its performance dropped precipitously in the single-image interactive setting (GPT5.2). This suggests that the model's behavior is highly sensitive to the interaction context or system-prompt variations inherent to the chat

[1] 35 images

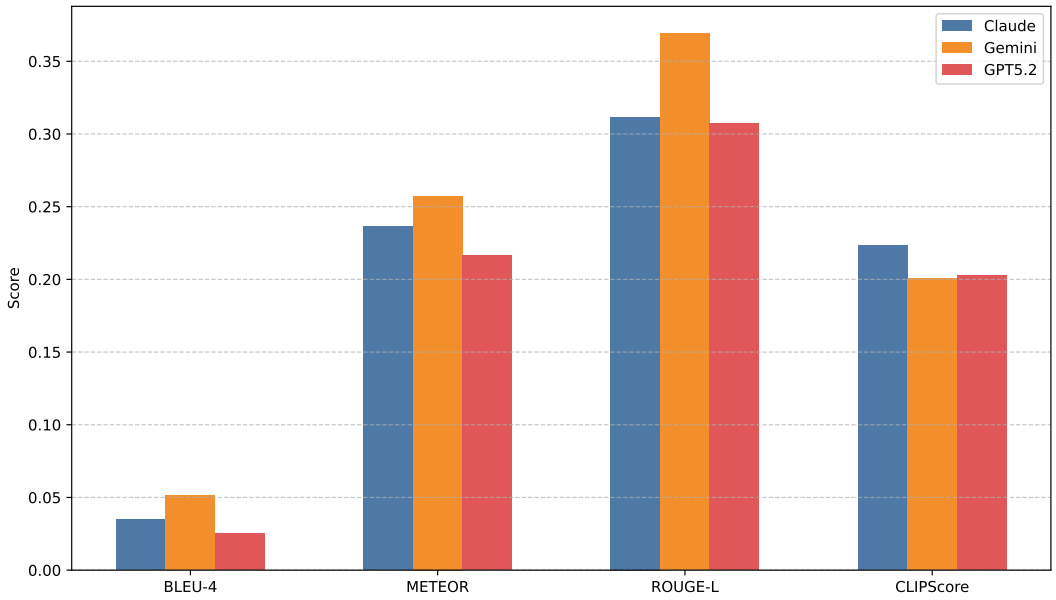

Figure 1: Comparison of Automated Metrics (BLEU-4, METEOR, ROUGE-L, CLIPScore) across models.

Table 4: Performance comparison of the three models for the batch generative setup. GPT-OSS 120b achieved the highest lexical scores (BLEU-4, METEOR, ROUGE-L), while Gemini achieved the highest CLIPScore. Best scores are bolded.

Model	BLEU-4	METEOR	ROUGE-L	CLIPScore
Gemini	0.003	0.105	0.266	**23.41**
Claude Sonnet 4.5	0.009	0.148	0.293	19.43
GPT-OSS 120b	**0.033**	**0.186**	**0.340**	20.44

interface. Gemini, by contrast, demonstrated remarkable stability and robustness, performing exceptionally well in both settings and dominating the single-image task.

5.2 Structural vs. Automated Metrics

A key finding of this study is the divergence between human-computed structural metrics (SC, TC, HCN) and automated NLP metrics (BLEU, METEOR, ROUGE, CLIPScore).

Comparing the structural results in Table 2 with the automated metrics in Table 3, we observe that **Gemini**'s dominance in structural accuracy (100% SC, 99.2% TC) is reflected in its highest scores for BLEU-4 (0.051), METEOR (0.257), and ROUGE-L (0.369). This indicates a strong alignment: the model that accurately identifies graph elements also produces text that lexically resembles the ground truth.

However, the case of **Claude** reveals a critical limitation of automated metrics. Despite failing to identify some states, and having the worst score on TC (88.2%), Claude achieved the highest CLIPScore (22.31). This effectively illustrates that CLIPScore measures broad semantic image-text compatibility but is insensitive to the fine-grained structural details required for technical diagrams. A description can be "compatible" with an image in a visual sense (e.g., "This is a diagram with circles and arrows") while being functionally useless for a blind user who needs to know exactly which state connects to which. Similarly, Claude's BLEU and ROUGE scores were comparable to GPT's despite its massive structural hallucinations (204 vs 18), suggesting that n-gram overlap metrics can be efficiently "gamed" by producing fluent, domain-relevant jargon that lacks factual grounding.

This discrepancy underscores the necessity of domain-specific structural evaluation. Automated metrics alone are insufficient for benchmarking mutlimodal LLMs on technical tasks where precision is paramount. While they provide a rough proxy for linguistic fluency and general visual relevance, they fail to capture the rigorous factual accuracy demanded by STEM accessibility applications.

5.3 Irrelevant Physical Detail

Across models and settings, we observe the frequent inclusion of physically accurate but accessibility-irrelevant details. Examples include descriptions such as "a downward arrow labeled d", "a curved arrow", or "a rightward arrow labeled d from q1 to q2" when these details do not contribute meaningfully to understanding the

automaton's structure or behavior. In some cases, spatial characterizations such as "arranged in an L shape" (e.g., Automata 39 and 40, generated by GPT) are included despite having no semantic relevance to the formal model.

While such descriptions are visually faithful, they increase cognitive load for screen reader users by foregrounding low-level geometric properties instead of higher-level relational information. From an accessibility perspective, this represents a form of over-description: information that is correct but distracts from what is necessary to build a functional mental model of the automaton. This observation reinforces the distinction between visual fidelity and accessibility usefulness, highlighting that not all visible features should be treated as equally salient in alternative descriptions.

5.4 Failure in the Batch Setting

The batch-processing setup exposes more severe and systematic failures that are not apparent in single-image interactive use.

Claude exhibits a pronounced degradation in this setting. In many cases, the generated outputs cannot be considered alternative descriptions at all: descriptions are either unrelated to the source image, internally inconsistent, or structurally incoherent. These failures go beyond isolated hallucinations and indicate a breakdown in the model's ability to ground descriptions in the actual visual input when operating under fully automated, dataset-level constraints. From an accessibility standpoint, such outputs are unusable and potentially harmful, as they provide confident but misleading information to blind users.

Gemini, by contrast, often produces descriptions that are formally correct but still fail to function as alternative descriptions. In the batch setting, we identify at least 18 images for which Gemini outputs descriptions of the following form: *DFA with compound state labels 1, 1,3, 1,4, 1,3,2. State 1 is start. State 1,3,2 is accepting. Transitions: 1 to 1,3 (a), 1,4 (b). 1,3 to 1,3,2 (a), 1,4 (b). 1,4 to 1,3 (a), 1,3,2 (b). 1,3,2 loops on a,b.* Although this description is internally consistent and resembles a valid automaton specification, it does not correspond to the image being described. Instead, it reflects a canonical or inferred automaton structure that is not visually present in the diagram. In these cases, Gemini appears to substitute image-grounded description with a plausible abstract automaton, effectively ignoring the requirement to describe the specific visual instance.

This behavior is particularly problematic for accessibility. An alternative description must provide access to what is actually shown, not to a mathematically plausible construct. For blind users, such substitutions are indistinguishable from correct descriptions and may lead to persistent misconceptions, especially in educational contexts.

5.5 Accessibility Implications

Taken together, these observations highlight two distinct but equally critical risks for generative AI in accessibility:

- Over-description of irrelevant visual features, which increases cognitive load without improving understanding.
- Under-grounded abstraction, where models replace the actual image with a generic or inferred representation.

Both phenomena are amplified in batch-processing scenarios and are only weakly reflected in automated evaluation metrics. Crucially, neither failure mode is easily detectable by blind users relying on the generated descriptions.

These findings further support the conclusion that current multimodal language models cannot be safely deployed for unsupervised, large-scale generation of alternative descriptions for complex technical diagrams. While models may perform well in interactive, single-image scenarios, accessibility-critical applications demand strict grounding, conservativeness, and human oversight.

6 Conclusion

This paper examines how effectively state-of-the-art multimodal LLMs generate alternative text for STEM diagrams compared to human-authored reference descriptions, under two complementary usage conditions: interactive single-image description and fully automated batch processing. Using a curated dataset of 73 automata diagrams with accessibility-oriented human descriptions, we combined automated metrics (BLEU, ROUGE, METEOR, and CLIPScore) with a human-centered evaluation focused on accessibility-critical elements (state coverage, transition coverage, identification of initial and accepting states) and hallucination analysis.

Across both settings, we observed that LLMs often produce fluent and structurally plausible descriptions, but fluency did not reliably translate into accessibility usefulness. In the batch scenario, representative of large-scale remediation pipelines, GPT achieved the strongest overall performance in state and transition coverage, while Gemini produced relatively few hallucinations but showed a high refusal rate that substantially limited its practical applicability. Claude exhibited the most severe robustness issues in batch mode, including frequent hallucinations and several cases of catastrophic failure where outputs were unrelated to the source diagrams.

In the interactive single-image setting, closer to optimal assistive use, performance patterns changed significantly. Gemini achieved near-perfect coverage and the lowest hallucination counts, while GPT and especially Claude showed reduced reliability, including failures on fundamental structural elements (such as initial state identification) and increased hallucination frequency. These results indicate that model behavior depends heavily on deployment conditions (batch versus interactive) and that evaluations limited to a single interaction paradigm risk overestimating real-world reliability.

Our findings also highlight significant limitations of commonly used automated similarity metrics for accessibility evaluation. N-gram-based measures (e.g., BLEU) remained low overall and did not consistently reflect the severity of accessibility-critical errors: a description can match the expected style of the reference while still omitting or distorting key diagram semantics, such as transition labels or accepting states. Conversely, CLIPScore captured some aspects of visual grounding but still failed to guarantee correctness at the level required for STEM comprehension. Together, these outcomes reinforce the need for evaluation protocols that explicitly measure diagram semantics and hallucinations, and that incorporate human-centered criteria aligned with the needs of screen reader users.

This study focuses on a single class of structured STEM figures (automata diagrams) and expert-authored reference descriptions that reflect one accessibility-informed writing style. Future work should (i) expand to additional STEM visual genres (e.g., plots, circuits, tables, proofs, derivations), (ii) include evaluations with blind and low-vision participants to directly assess task success and perceived utility, and (iii) further analyze prompt strategies and output constraints that reduce hallucinations in both interactive and batch pipelines. Future benchmarks should also distinguish between coverage and correctness of formal semantics, since minor inaccuracies can become catastrophic in educational contexts.

Acknowledgments

This work was carried out within the STEMMA PRIN project thanks to the funding by the European Union - Next Generation EU, Mission 4 Component 2 CUP B53D23019500006.

References

[1] Huda Diab Abdulgalil and Otman A. Basir. 2025. Next-generation image captioning: A survey of methodologies and emerging challenges from transformers to Multimodal Large Language Models. *Natural Language Processing Journal* 12 (2025), 100159. doi:10.1016/j.nlp.2025.100159

[2] Vanita Agrawal, Mvv Prasad Kantipudi, and Jayant Jagtap. 2025. Enhancing hand-drawn diagram recognition through the integration of machine learning and deep learning techniques. *Scientific Reports* 15 (05 2025). doi:10.1038/s41598-025-01823-4

[3] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 65–72.

[4] Sheryl E Burgstahler and Rebecca C Cory. 2010. *Universal design in higher education: From principles to practice*. Harvard Education Press.

[5] Marina Buzzi, Giulio Galesi, Barbara Leporini, and Annalisa Nicotera. 2024. Is Generative AI Mature for Alternative Image Descriptions of STEM Content?. In *International Conference on Web Information Systems and Technologies*. 274–281. doi:10.5220/0012996800003825

[6] Marco Cardia, Marina Buzzi, Giulio Galesi, and Barbara Leporini. 2025. A Descriptive Review of Image Datasets for Accessible Alternative Descriptions in STEM Domains. In *Proceedings of the 16th Biannual Conference of the Italian SIGCHI Chapter*. 1–7.

[7] Maitraye Das, Alexander J Fiannaca, Meredith Ringel Morris, Shaun K Kane, and Cynthia L Bennett. 2024. From provenance to aberrations: Image creator and screen reader user perspectives on alt text for AI-generated images. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–21.

[8] DIAGRAM Center / Benetech. [n. d.]. *Image Description Guidelines*. DIAGRAM Center. http://diagramcenter.org/specific-guidelines-d.html Accessed: 18 Jan. 2026.

[9] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3608–3617.

[10] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*. 7514–7528.

[11] M. Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *Journal of Artificial Intelligence Research* 47 (08 2013), 853–899. doi:10.1613/jair.3994

[12] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *European conference on computer vision*. Springer, 235–251.

[13] Alon Lavie and Michael J Denkowski. 2009. The METEOR metric for automatic evaluation of machine translation. *Machine translation* 23, 2 (2009), 105–115.

[14] Maurizio Leotta, Fabrizio Mori, and Marina Ribaudo. 2023. Evaluating the effectiveness of automatic image captioning for web accessibility. *Universal access in the information society* 22, 4 (2023), 1293–1313.

[15] Maurizio Leotta and Marina Ribaudo. 2024. Evaluating the Effectiveness of STEM Images Captioning. In *Proceedings of the 21st International Web for All Conference*. 150–159.

[16] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 740–755.

[18] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.

[19] João Daniel Silva, João Magalhães, Devis Tuia, and Bruno Martins. 2024. Large language models for captioning and retrieving remote sensing images. *arXiv preprint arXiv:2402.06475* (2024).

[20] Shaowei Wang, LingLing Zhang, Xuan Luo, Yi Yang, Xin Hu, and Jun Liu. 2021. RL-CSDia: Representation learning of computer science diagrams. *arXiv preprint arXiv:2103.05900* (2021).

[21] World Wide Web Consortium. Apri. *Functional Images*. Web Accessibility Initiative (WAI), W3C. https://www.w3.org/WAI/tutorials/images/functional/ Accessed: 21 Jan. 2026.

[22] World Wide Web Consortium. 2023. *Web Content Accessibility Guidelines (WCAG) 2.2*. Web Accessibility Initiative (WAI), W3C. https://www.w3.org/TR/WCAG22/ Accessed: 17 Jan. 2026.

[23] Chuqiao Yan, Hans-Peter Hutter, Felix M Schmitt-Koopmann, and Alireza Darvishy. 2025. Chart Accessibility: A Review of Current Alt Text Generation. *IEEE Access* (2025).

[24] Richard Zanibbi and Dorothea Blostein. 2012. Recognition and retrieval of mathematical expressions. *International Journal on Document Analysis and Recognition (IJDAR)* 15, 4 (2012), 331–357.