Multimodal Large Language Models for Portuguese Alternative Text Generation for Images

Víctor Alexsandro Elisiário[a] and Willian Massami Watanabe[b]
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brazil

Keywords: Accessibility, Image Description, Text Generation, Multimodal Large Language Models.

Abstract: Since the creation of the Web Content Accessibility Guidelines (WCAG), the Web has become increasingly accessible to people with disabilities. However, related works report that Web developers are not always aware of accessibility specifications and many Web applications still contain accessibility barriers. Therefore, this work proposes the use of Multimodal Large Language Models (MLLM), leveraging Google's Cloud Vision API and contextual information extracted from Web pages' HTML, to generate alternative texts for images using the Gemini-1.5-Pro model. To evaluate this approach, a case study was conducted to analyze the perceived relevance of the generated descriptions. Six Master's students in Computer Science participated in a blind analysis, assessing the relevance of the descriptions produced by the MLLM alongside the original alternative texts provided by the page authors. The evaluations were compared to measure the relative quality of the descriptions. The results indicate that the descriptions generated by the MLLM are at least equivalent to those created by humans. Notably, the best performance was achieved without incorporating additional contextual data. These findings suggest that alternative texts generated by MLLMs can effectively meet the needs of blind or visually impaired users, thereby enhancing their access to Web content.

1 INTRODUCTION

In Brazil, approximately 18.6 million people are estimated to have disabilities (Instituto Brasileiro de Geografia e Estatística, 2023). Although the Brazilian Statute for Persons with Disabilities aims to allow these individuals to live independently and fully participate in all aspects of life (Brasil, 2015), the reality they encounter is marked by major challenges. In addition to ableism, which often leads to denial of rights, these individuals face daily challenges due to a lack of accessibility in daily life.

With the advancement of technology and the popularization of the Internet, digital accessibility has also become a growing concern. In this context, the Web Content Accessibility Guidelines (WCAG) were introduced in the late 1990s (Lewthwaite, 2014), with the aim of ensuring that Web navigation is accessible to everyone, including people with disabilities. Visually impaired people often need to modify the way information is presented, transforming it into more accessible formats to meet their specific needs (W3C, 2024)[1]. In this regard, the WCAG recommends that all non-text content include alternative text conveying an equivalent meaning.

However, Web developers often lack awareness of accessibility standards, resulting in Web applications that still present significant accessibility barriers (Guinness et al., 2018; Valtolina and Fratus, 2022; Inal et al., 2022). Additionally, alternative texts for non-text elements play a crucial role in the search engine ranking of Web pages (Mavridis and Symeonidis, 2015). Often, these alternative texts are used in a way that maximizes search engine scores, disregarding their intended accessibility function (Gleason et al., 2019; Sheffield, 2020).

This study aims to compare alternative texts generated by Multimodal Large Language Models (MLLMs) with alternative texts currently available for images on the Web. Specifically, it seeks to understand the state of the art in text generation by MLLMs, develop a script to perform this task, and, finally, compare the generated texts with those provided by the authors of the websites under investigation.

This approach combines computer vision tech-

[a] https://orcid.org/0009-0007-1602-3791
[b] https://orcid.org/0000-0001-6910-5730

[1] https://www.w3.org/WAI/people-use-web/abilities-barriers/visual/

Elisiário, V. A. and Watanabe, W. M.
Multimodal Large Language Models for Portuguese Alternative Text Generation for Images.
DOI: 10.5220/0013673800003985

In *Proceedings of the 21st International Conference on Web Information Systems and Technologies (WEBIST 2025)*, pages 493-501
ISBN: 978-989-758-772-6; ISSN: 2184-3252

niques for image recognition with the text generation capabilities of MLLM, enriched by contextual data extracted from HTML, to produce visual descriptions for Web images.

The remainder of this paper is structured as follows: Section 2 discusses foundational concepts of Web accessibility and specific guidelines for the use of alternative texts. Section 3 reviews related works and their applications. Section 4 details the methodology and evaluation criteria. Section 5 presents the results. Section 6 analyzes the limitations of the study. Section 7 provides a discussion of the findings, and Section 8 concludes the study and suggests directions for future research.

2 IMAGE ACCESSIBILITY

WCAG, developed by the Web Accessibility Initiative (WAI), an organization established by the W3C, serves as a foundation for Web accessibility standards. The primary objective of the WCAG is to provide a comprehensive set of recommendations to make Web content more accessible (W3C, 2023b)[2].

According to WAI, accessibility "*addresses discriminatory aspects related to equivalent user experience for people with disabilities*" (W3C, 2016)[3] and seeks to ensure that individuals with disabilities can perceive, understand, navigate, and interact with Web pages and tools without encountering barriers (W3C, 2016). Accessibility encompasses both technical requirements related to the code, as well as usability factors affecting user interaction with Web content (W3C, 2016).

The World Health Organization (WHO) states that "*vision impairment occurs when an eye condition affects the visual system and one or more of its vision functions*" including visual acuity, field of vision, contrast sensitivity, and color vision (World Health Organization, 2019). Individuals with visual impairments often rely on tools that adapt Web content to meet their needs, such as adjusting font and image sizes, using screen readers to vocalize text, or accessing audio descriptions of images and videos. For these tools to function effectively, developers must ensure that Web content is properly coded, enabling browsers and assistive technology to interpret and adapt it accordingly (W3C, 2017)[4].

The WCAG 2.2 is a set of recommendations designed to make Web content more accessible. These guidelines aim to address a wide range of disabilities, including blindness and low vision, deafness and hearing loss, limited mobility, speech impairments, photosensitivity, as well as learning difficulties and cognitive limitations (W3C, 2023b).

The WCAG is structured into four main layers. At the top is the layer of Principles, which serve as the foundation for Web accessibility. Below that, there are 13 Guidelines, which establish goals that developers must follow to make content more accessible. Although the guidelines are not testable on their own, they provide a framework and general objectives that help developers understand success criteria and implement techniques more effectively. For each guideline, the next layer includes a set of Testable Success Criteria, which can be used in contexts where requirements and conformance testing are necessary. Finally, in the last layer, there are Sufficient and Recommended Techniques, which aim to guide the implementation of solutions that meet the success criteria (W3C, 2023b).

Images fall under the first principle of WCAG 2.2. The Text Alternatives guideline suggests that text alternatives must be provided for any non-text content (W3C, 2023a)[5]. Compliance is achieved when all non-text content presented to users is accompanied by a text alternative that serves an equivalent purpose.

Despite these established guidelines for image accessibility, studies reveal significant shortcomings in practice. For instance, an analysis of the most visited Web pages, according to alexa.com, found that approximately 28% of images across 481 pages lacked alternative texts. Among the images that included alternative texts, many were of poor quality, often limited to file names or generic descriptions such as "image" (Guinness et al., 2018). Further research on municipal government websites in Italy (Valtolina and Fratus, 2022) and Norway (Inal et al., 2022) indicates widespread non-compliance with WCAG 2.0, frequently violating multiple Level A criteria. Additionally, in the context of social media, data reveals that nearly 12% of Twitter posts include images, yet only 0.1% of these images feature alternative texts (Gleason et al., 2019).

3 RELATED WORK

In recent years, various approaches have been proposed to mitigate the impacts caused by the absence

[2]https://www.w3.org/TR/WCAG22

[3]https://www.w3.org/WAI/fundamentals/accessibility-usability-inclusion/

[4]https://www.w3.org/WAI/people-use-web/abilities-barriers/visual/

[5]https://www.w3.org/WAI/WCAG22/quickref

of alternative text in Web images. One example is Twitter A11y (Gleason et al., 2020), which employs different methods to create descriptions for images without alternative text on the platform. The proposal consists of a sequence of steps for generating descriptions; if none of the methods yield a result, crowdsourcing is used. In this case, a task is created on Amazon Mechanical Turk for a person to manually generate the image description.

Crowdsourcing is widely used in generating descriptions for images (Zhong et al., 2015; Bigham et al., 2010) and is capable of producing descriptions in approximately 30 seconds (Bigham et al., 2010). However, it is an expensive solution, potentially costing over R$ 1.00 per image (Gleason et al., 2020).

With the advancement of artificial intelligence, new techniques have been explored to enhance image accessibility on the Web. Large Language Models (LLMs), which are capable of understanding human language and generating textual responses, and computer vision, an area of AI focused on analyzing and interpreting images (Faisal et al., 2022), are being used in developing tools capable of generating visual descriptions for images.

The work by Ramaprasad (2023) uses computer vision and Multimodal Large Language Models (MLLMs) to generate natural language descriptions of comic strips. Another example is an application capable of identifying the ball and players during a soccer match, interpreting on-field actions and providing real-time information to the audience through a voice synthesizer (Pavlovich et al., 2023).

In another application, GPT-3 was employed in a proof of concept for an assistive system designed for visually impaired people (Hafeth et al., 2023). The system uses captioning techniques to generate descriptions of environments from photos, providing detailed information that helps users better understand the spaces around them. The generated descriptions are analyzed by GPT-3 to determine if they indicate dangerous situations and, if necessary, suggest corrective actions.

Another study proposed an interface that allows visually impaired content creators to verify if the generated images meet their requests (Huh et al., 2023). The interface also provides additional information not initially included, as well as summaries of the similarities or differences between the generated candidate images. The descriptions generated by the tool were compared with descriptions produced by humans. The study found that while the LLM-generated descriptions were of comparable quality to human-written ones, they were able to identify more than twice as many differences between the images.

Another study evaluated the descriptions generated by an AI engine (IDEFICS) for STEM (Science, Technology, Engineering, and Mathematics) images, comparing them with those written by both untrained and trained undergraduate Computer Science students (Leotta and Ribaudo, 2024). The trained students received a brief lesson on how to create alternative texts for people with disabilities, while the untrained students participated independently without prior instruction. The study found that the descriptions generated by the AI engine were perceived as less correct, useful, and of lower overall quality compared to those written by humans, when applied to STEM related images, while this difference was less evident for non-STEM related images.

Although the use of AI for generating alternative text seems a viable and economical alternative, a study evaluating four automatic image-to-text generation services (Azure Computer Vision Engine, Amazon Rekognition, Cloudsight, and Auto Alt-Text for Google Chrome) revealed that, on average, users still prefer human-made descriptions, even when machine-generated descriptions are accurate (Leotta et al., 2023). Furthermore, another study pointed out that people with total vision loss expect visual descriptions to convey an ordered spatial notion of the items in the image, offer different levels of detail (allowing navigation among them), and include aesthetic elements, making the photos more memorable (Jung et al., 2022).

4 METHODOLOGY

This section presents the methodology adopted for generating image descriptions. The approach consists of three main stages: (1) Data Collection, (2) Image Analysis, and (3) Prompt Creation. In the first stage, a Python script extracts the image and news data from the analyzed Web page. Next, in the second stage, the image is processed using a Computer Vision application to extract relevant features. Finally, in the third stage, a prompt is created using the data collected in stage 1 and the information obtained in stage 2. Figure 1 illustrates the steps performed for each selected news page.

The following provides a detailed description of each step in the proposed methodology.

1. **Data Collection**

   On each Web page, we executed a script to extract all content within the main and article HTML tags, corresponding to headings levels h1 and h2, as well as paragraphs. Additionally, the script collected one image per page, specifically, the first

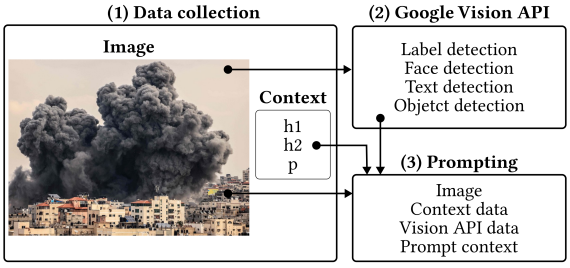

Figure 1: Methodological flow of the study, illustrating the steps of data collection, analysis with the Google Vision API and prompt generation.

image of the news article found within the same main and article tags, along with its alternative text, provided by the author. All collected images had alternative text.

2. **Google Cloud Vision API**[6]

   This step aimed to retrieve detailed information from the image to assist the Multimodal Large Language Model (MLLM) in identifying its components. To achieve this, the image obtained in the previous step was sent to the Google Cloud Vision API for the following analyses:

   (a) **Face Detection:** Identification of any faces present in the image, along with their primary attributes such as emotional state and the use of accessories.

   (b) **Label Detection:** Identification of information across various categories, including general objects, locations, activities, animal species, products and others.

   (c) **Text Detection:** Identification and extraction of textual content present within the image.

   (d) **Object Detection:** Identification and extraction of objects depicted in the image.

   All data returned by the API were accompanied by a confidence score indicating the probability of accuracy for each piece of information. These data were subsequently sent to the MLLM, which was responsible for interpreting them.

3. **Prompting and Contextualization**

   Initially, we provided contextual information, as Large Language Models (LLMs) yield higher-quality responses when the prompt includes context about the environment in which the request is made (Huh et al., 2023; Hajizadeh Saffar et al., 2024). Therefore, the prompt began with:

   "You are an efficient assistant who describes images for visually impaired individuals so they can understand what is shown in the images. Do not speculate or imagine anything that is not in the image."

   Furthermore, the MLLM was explicitly instructed not to speculate or imagine content not present in the image. This instruction ensured that the model generated its response based solely on the provided data, thereby minimizing the chances of the description deviating from the context of the news article.

   Subsequently, we added contextual data: **"Analyze the following data to formulate your response: The image was added to an online news article. The content of the article is:** *[here, the textual data retrieved from the page were added, including the h1, h2, and p content available within the main and article tags]***. A computer vision API was used to identify labels, faces, texts, and objects in the image. The results obtained were:** *[here, the data returned from the Google Cloud Vision API were added]***."**

   Finally, the prompt included the final request, along with specific guidelines: **"Provide a brief description of the image based on the image and the information above. Do not reference the provided data; just describe the image. Do not describe logos or icons, simply mention what they are. Limit your response to 25 words."**

   These specific guidelines were necessary because, during initial tests, the model would reference the supplied data (e.g., "...the vision API indicates that they are smiling") or provide overly detailed descriptions of logos and icons, such as specifying the colors of each letter in the Google logo.

After collecting and processing all the data, the request was sent to Google's *Gemini-1.5-Pro*[7] model, together with the image being analyzed, allowing the model to examine the image directly. All prompts and data were written in Brazilian Portuguese.

4.1 MLLM Approaches Evaluated

For each image, three requests were made to the MLLM, each with a distinct prompt, as described below:

A1: All available data was used, including the image, page text, Google Cloud Vision API data, and prompt guidelines.

A2: Image, Google Cloud Vision API data, and prompt guidelines were used. In this case, the text data from the page were not provided. In other words,

[6]https://cloud.google.com/vision

[7]https://gemini.google.com

the model does not receive the context in which the image is embedded and generates its response solely based on the information extracted from the image.

A3: Only the image and the prompt guidelines were used. This request aimed to verify the model's ability to generate a visual description without the aid of external data, relying solely on its image analysis capability.

4.2 Selected Topics

Ten Brazilian news pages were selected, classified into four distinct categories, as follows:

1. Israel-Palestine Conflict.
2. Floods in the state of Rio Grande do Sul, Brazil.
3. Strike at Federal Universities in Brazil.
4. Heatwaves in India.

These topics were chosen because they are accompanied by images that play a significant role in the presented context. Therefore, the text on the page, being directly related to the image, can provide better context for the MLLM to generate accurate descriptions.

4.3 Evaluation

To evaluate the proposed approaches, a proof of concept was conducted to assess the perceived relevance of the descriptions generated by each method. The original alternative text provided by the developers for each image was used as a reference for comparison.

In the created form, each section displayed the news title, the image, the alternative text provided by the author and three descriptions generated by the methods under analysis, as defined in A1, A2, and A3, respectively. However, no explicit differentiation was made between the descriptions and the alternative text in the form.

The survey involved six Master's students in Computer Science, aged between 25 and 50 years. A brief explanation about the purpose of alternative texts on the Web was provided; however, no further information was given to them about accessibility or alternative texts. Participants rated image descriptions on a scale of 1 to 5, comparing the provided descriptions with their own perceptions of how a description for the image should be. Here, 1 indicated low relevance and 5 indicated high relevance.

4.4 Hypotheses

The evaluations aim to investigate whether MLLMs can be used to generate image descriptions that serve as effective alternative texts for visually impaired individuals. The goal is to understand not only the models' ability to produce quality descriptions but also how different levels of provided context influence the quality of these descriptions. Thus, the evaluations seek to test the following hypotheses:

H1: The image descriptions generated by MLLMs can be used as alternative texts, without loss of information, for individuals with visual impairments.

H2: The use of contextual data improves the quality of descriptions generated by the MLLM.

5 RESULTS

The method that received the highest number of positive ratings was method A3, with 40% of its descriptions receiving the maximum score of 5 (Very relevant). This result highlights the effectiveness of method A3 in producing highly relevant descriptions. Additionally, only 3 ratings for method A3 received the lowest score of 1 (Not relevant). Figure 2 presents the frequency distribution of the ratings assigned to each image by the participants, according to the method employed, where ALT is the alternative text provided by the author, and A1, A2, and A3 are the descriptions generated by each method.

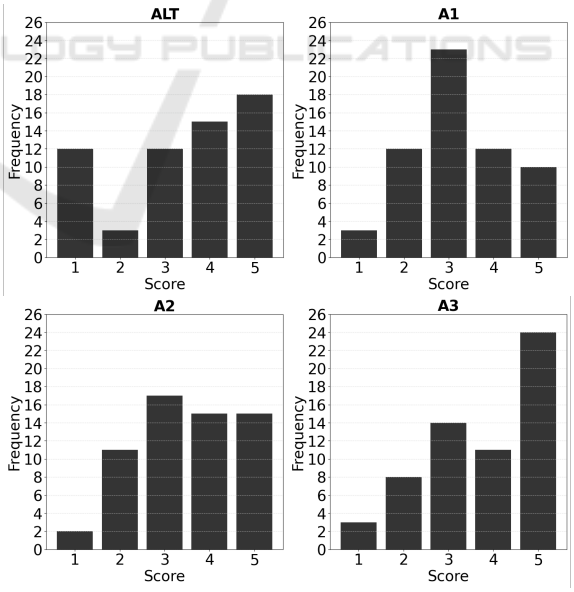

Figure 2: Frequency distribution of the ratings attributed to each evaluated method.

On the other hand, the alternative texts provided by the page authors were those that most frequently received the lowest rating (1 - Not relevant), repre-

senting 20% of the evaluations. This result reinforces the findings of Guinness et al. (2018), who observed that manually generated descriptions often lack meaningful content, thereby hindering comprehension for individuals with disabilities.

The analysis of the 10 images reveals that the proposed methods outperformed the authors' alternative texts in 5 out of the 10 cases. Moreover, in cases where the methods did not surpass the original descriptions, at least one method received a rating of 3 or higher, considered neutral on the evaluation scale. This suggests that, even in cases where the generated descriptions do not exceed the original ones, they still offer considerable relevance for content understanding. Table 1 presents the average ratings of the participants for each method applied to each image.

Table 1: Average ratings per image.

IMAGE	ALT	A1	A2	A3
1	3.2	4.0	3.0	4.2
2	1.7	3.8	4.3	4.8
3	3.5	2.3	3.5	4.5
4	4.3	3.8	3.8	3.3
5	3.0	2.8	2.8	4.2
6	4.5	3.8	3.8	4.0
7	4.2	3.0	3.8	4.0
8	4.0	3.3	3.7	4.0
9	2.0	2.8	3.2	1.8
10	3.7	2.5	3.0	2.7
AVERAGE	3.40	3.23	3.50	3.75

Methods A2 and A3 exhibited average ratings higher than those of the manually generated alternative texts, while method A1 performed worse. These results suggest that incorporating context, by using content from HTML tags such as h1, h2, and p within the main and article tags of news pages, as done in method A1, does not improve descriptions generated by MLLM. Thus, the findings go against hypothesis H2.

On the other hand, considering that all methods received scores above the neutral rating and two of the three methods outperformed the manual descriptions, the results suggest that descriptions generated by MLLM can serve as viable alternatives, particularly in the absence of alternative texts, without compromising information accessibility for individuals with visual impairments. These findings are aligned with hypothesis H1.

6 STUDY LIMITATIONS

During the research, limitations were identified in the ability of the Google Cloud Vision API and Gemini to interpret certain image contents, particularly those dependent on specific context. To better understand the limitations encountered, three images are presented in Figure 3. These images represent the third, sixth, and ninth images shown to the participants, and their respective scores are listed in Table 1. The first limitation is found in Figure 3.3, where the alternative text provided by the author is highly detailed, directly relating the image to the context of the news article "Heatwave in India kills at least 33 people". In contrast, although the description generated by the proposed method effectively captures visual details, it fails to establish a connection with the heatwave context. Only the descriptions generated by approaches A2 and A3 correctly identified that the woman in the image is lying on a bed, whereas the A1 description showed limitations in this regard. Furthermore, none of the approaches detected the presence of a handmade fan.

In Figures 3.3 and 3.6, descriptions A1 mistakenly indicated "There is a logo in the background" and "There is a logo on one of the helmets", respectively, even though such logos do not exist in the images. This may have resulted from instructions in the prompt such as "Do not describe logos or icons, simply mention what they are" initially added to avoid describing logos.

Additionally, the prompt "Limit your response to 25 words" may have constrained the ability of the MLLM to develop the sentences present in the images. For Figure 3.9, the responses from methods A1 and A2 identified the words "ANDES" and "GREVE DOCENTE FEDERAL" while A3 failed to detect any text. This limitation in A3 may be attributed to its lack of use of additional data from the Google Cloud Vision API. Despite the correct text recognition by A1 and A2, their responses conveyed only partial information. Nonetheless, when compared to the alternative text provided by the author, which not only contained a spelling mistake but also provided a largely insignificant description, the results from methods A1 and A2 were found to be adequate for providing content comprehension for individuals with visual impairments.

Insignificant alternative texts produced by the authors were identified in several images. For example, in one of the images from a news article titled "Floods ravage the population in the South of the country this weekend", the alternative text is simply "Agronômica", the name of the city where the im-

(3) (6) (9)

Figure 3: Images (3), (6) and (9) used in the study, respectively.

age was captured. Furthermore, Figure 3.6 presents an alternative text in English, despite the fact that the news article is in Portuguese, which hinders the comprehension of individuals with visual impairments. In contrast, automatic approaches generated meaningful descriptions, reinforcing the need for the use of alternative methods for the automatic generation of alternative texts.

7 DISCUSSION

Several studies have employed crowdsourcing (Gleason et al., 2020; Bigham et al., 2010; Zhong et al., 2015) to generate image descriptions. However, a key limitation of this approach is its dependence on human input, which, although effective, can result in delays and is often costly. In contrast, our approach leverages Multimodal Large Language Models (MLLMs) to automate the generation of alternative texts, providing a scalable solution for producing relevant descriptions without the need for human intervention.

Some studies (Bigham et al., 2010; Leotta et al., 2023) have reported that automatic approaches often struggle to address visual inquiries from blind users. However, our findings suggest that MLLMs can effectively generate descriptions that meet the needs of visually impaired users, potentially overcoming the limitations faced by previous methods that rely heavily on crowdsourcing. This automation is crucial in addressing the accessibility gap, particularly for images that lack any description.

A similar study to ours (Leotta and Ribaudo, 2024) was conducted on STEM (Science, Technology, Engineering, and Mathematics) images, where human-generated descriptions outperformed those produced by the AI engine (IDEFICS) in terms of quality, usefulness, and accuracy. Unlike this approach, our study focuses on images extracted from news websites, which are typically more aligned with everyday life.

Accessibility barriers are often associated with factors such as lack of awareness, time constraints, and insufficient executive support (Aljedaani et al., 2025). In this context, although human-generated descriptions outperform AI-generated ones, applications that rely solely on them may continue to face challenges in ensuring accessibility. For this reason, we adopt as our benchmark the alternative texts authored by the Web page creators themselves, rather than creating an ideal description for the images in question, as this offers a more realistic representation of the Web environment, where content creators tend to prioritize the production of the news itself rather than ensuring accessibility for people with disabilities. This is supported by our findings, where the alternative texts provided by the page authors were those that most frequently received the lowest rating in terms of relevance.

Contextual data from visual models have been used to enhance descriptions generated for comic strips, as demonstrated by Ramaprasad (2023), who employed computer vision to extract information and contextual data for image descriptions generation. Similarly, our study uses the Google Cloud Vision API to extract information from images for the generation of alternative texts. However, our results suggest that MLLMs can perform effectively even without relying on extensive contextual data. This versatility represents a significant advantage, as it enables the generation of alternative texts for standalone images.

Ramaprasad (2023) also highlighted the issue of hallucination in the generated descriptions, where the model sometimes makes up information. This issue might come from the prompt design, which did not explicitly instruct the model to base the generated description solely on the provided image and data. In contrast, our approach explicitly directed the model to generate descriptions strictly from the available visual and contextual inputs, which likely contributed to a lower incidence of hallucinations.

Although other researches (Ramaprasad, 2023; Pavlovich et al., 2023; Hafeth et al., 2023; Huh et al., 2023) have utilized LLMs and MLLMs for visual descriptions of non-textual elements, their applications did not aim to enhance Web image accessibil-

ity. Thus, this study contributes by presenting a viable alternative for generating alternative texts for images that lack any description.

Unlike other tools (Azure Computer Vision Engine, Amazon Rekognition, Cloudsight, and Auto Alt-Text for Google Chrome) examined in previous studies (Leotta et al., 2023), MLLM was assessed by sighted individuals as capable of yielding descriptions at least equivalent to those created by humans. However, further research is necessary to determine whether the method meets the expectations of people with visual disabilities, as reported in (Jung et al., 2022).

8 CONCLUSION AND FUTURE WORK

The use of MLLM for generating alternative texts for Web images shows substantial potential in enhancing accessibility for individuals with visual impairments. The study suggested that MLLM-generated descriptions could serve as valuable alternatives when human-written texts are unavailable, without compromising the information for visually impaired users. It is evident that utilizing contextual data did not produce results superior to descriptions generated solely from the image, exhibiting the method's versatility across various environments. As it requires no additional context, the method can generate alternative texts for standalone images, highlighting the potential of MLLMs in addressing accessibility challenges and fostering a more inclusive digital environment for all.

Possible future work includes: (1) testing the approach with pages of diverse topics and images; (2) validating the results with visually impaired individuals and a larger and more diverse group; (3) utilizing alternative resources for generating descriptions, such as other MLLMs like GPT-4 and more contextual data; and (4) modifying prompt parameters to assess MLLM's capacity to produce more precise results. Such research could significantly contribute to reducing Web accessibility barriers.

REFERENCES

Aljedaani, W., Eler, M. M., and Parthasarathy, P. D. (2025). Enhancing accessibility in software engineering projects with large language models (llms). In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1*, SIGCSETS 2025, page 25–31, New York, NY, USA. Association for Computing Machinery.

Bigham, J. P., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R. C., Miller, R., Tatarowicz, A., White, B., White, S., and Yeh, T. (2010). Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, UIST '10, page 333–342, New York, NY, USA. Association for Computing Machinery.

Brasil (2015). *Lei nº 13.146, de 06 de julho de 2015. Institui a Lei Brasileira de Inclusão da Pessoa com Deficiência (Estatuto da Pessoa com Deficiência)*. Brasil, Brasília, DF.

Faisal, F., Salam, M. A., Habib, M. B., Islam, M. S., and Nishat, M. M. (2022). Depth estimation from video using computer vision and machine learning with hyperparameter optimization. In *2022 4th International Conference on Smart Sensors and Application (ICSSA)*, pages 39–44, Kuala Lumpur, Malaysia. IEEE.

Gleason, C., Carrington, P., Cassidy, C., Morris, M. R., Kitani, K. M., and Bigham, J. P. (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*, WWW '19, page 549–559, New York, NY, USA. Association for Computing Machinery.

Gleason, C., Pavel, A., McCamey, E., Low, C., Carrington, P., Kitani, K. M., and Bigham, J. P. (2020). Twitter a11y: A browser extension to make twitter images accessible. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–12, New York, NY, USA. Association for Computing Machinery.

Guinness, D., Cutrell, E., and Morris, M. R. (2018). Caption crawler: Enabling reusable alternative text descriptions using reverse image search. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–11, New York, NY, USA. Association for Computing Machinery.

Hafeth, D. A., Lal, G., Al-Khafajiy, M., Baker, T., and Kollias, S. (2023). Cloud-iot application for scene understanding in assisted living: Unleashing the potential of image captioning and large language model (chatgpt). In *2023 16th International Conference on Developments in eSystems Engineering (DeSE)*, pages 150–155, Istanbul, Turkiye. IEEE.

Hajizadeh Saffar, A., Sitbon, L., Hoogstrate, M., Abbas, A., Roomkham, S., and Miller, D. (2024). Human and large language model intent detection in image-based self-expression of people with intellectual disability. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval*, CHIIR '24, page 199–208, New York, NY, USA. Association for Computing Machinery.

Huh, M., Peng, Y.-H., and Pavel, A. (2023). Genassist: Making image generation accessible. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Inal, Y., Mishra, D., and Torkildsby, A. B. (2022). An analysis of web content accessibility of municipality websites for people with disabilities in norway: Web accessibility of norwegian municipality web-

sites. In *Nordic Human-Computer Interaction Conference*, NordiCHI '22, New York, NY, USA. Association for Computing Machinery.

Instituto Brasileiro de Geografia e Estatística (2023). *Pessoas com deficiência: 2022*. Rio de Janeiro. 15 p.

Jung, J. Y., Steinberger, T., Kim, J., and Ackerman, M. S. (2022). "so what? what's that to do with me?" expectations of people with visual impairments for image descriptions in their personal photo activities. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference*, DIS '22, page 1893–1906, New York, NY, USA. Association for Computing Machinery.

Leotta, M., Mori, F., and Ribaudo, M. (2023). Evaluating the effectiveness of automatic image captioning for web accessibility. *Universal access in the information society*, 22(4):1293–1313.

Leotta, M. and Ribaudo, M. (2024). Evaluating the effectiveness of stem images captioning. In *Proceedings of the 21st International Web for All Conference*, W4A '24, page 150–159, New York, NY, USA. Association for Computing Machinery.

Lewthwaite, S. (2014). Web accessibility standards and disability: developing critical perspectives on accessibility. *Disability and Rehabilitation*, 36(16):1375–1383. PMID: 25009950.

Mavridis, T. and Symeonidis, A. L. (2015). Identifying valid search engine ranking factors in a web 2.0 and web 3.0 context for building efficient seo mechanisms. *Engineering Applications of Artificial Intelligence*, 41:75–91.

Pavlovich, R. V., Tsybulko, E. A., Zhigunov, K. N., Khelvas, A. V., Gilya-Zetinov, A. A., and Tykhonov, I. V. (2023). Soccer artificial intelligence commentary service on the base of video analytic and large language models. In *2023 31st Telecommunications Forum (TELFOR)*, pages 1–4, Belgrade, Serbia. IEEE.

Ramaprasad, R. (2023). Comics for everyone: Generating accessible text descriptions for comic strips. *ArXiv*, abs/2310.00698.

Sheffield, J. P. (2020). Search engine optimization and business communication instruction: Interviews with experts. *Business and Professional Communication Quarterly*, 83(2):153–183.

Valtolina, S. and Fratus, D. (2022). Local government websites accessibility: Evaluation and finding from italy. *Digital Government: Research and Practice*, 3(3).

W3C (2016). Accessibility, usability, and inclusion.

W3C (2017). Visual.

W3C (2023a). How to meet wcag (quick reference).

W3C (2023b). Web content accessibility guidelines (wcag) 2.2.

W3C (2024). Visual.

World Health Organization (2019). *World report on vision*. World Health Organization, Geneva.

Zhong, Y., Lasecki, W. S., Brady, E., and Bigham, J. P. (2015). Regionspeak: Quick comprehensive spatial descriptions of complex images for blind users. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 2353–2362, New York, NY, USA. Association for Computing Machinery.