Leveraging Large Language Models for Alt-Text Evaluation in E-Commerce: A Data-Driven Study

Nicola Leonardi*
CNR – ISTI, HIIS Laboratory
Pisa, Italy
University of Pisa
Pisa, Italy
nicola.leonardi@isti.cnr.it

Marco Manca
CNR – ISTI, HIIS Laboratory
Pisa, Italy
marco.manca@isti.cnr.it

Fabio Paternò
CNR – ISTI, HIIS Laboratory
Pisa, Italy
fabio.paterno@isti.cnr.it

Abstract

Generative artificial intelligence opens new opportunities for accessibility validation. An interesting case is the assessment of alternative text (alt-text) for images. This paper investigates the use of a large language model (LLM) to analyse alt-texts in real e-commerce websites, a domain in which images play an important role and have specific requirements. We present a novel data-driven method and an associated tool that employs tailored prompting strategies to incorporate contextual information when generating and evaluating image descriptions. The approach also supports systematic comparison between human-authored and LLM-generated alt-text. We conducted a user study (N = 16) involving 494 assessments of 157 images, and their corresponding alt-texts extracted from real e-commerce websites. The results show that the proposed solution can provide valid results, supporting its possible integration into existing accessibility validation workflows and authoring tools.

CCS Concepts

• **Human-centered computing** → Accessibility; Accessibility design and evaluation methods; Accessibility; Accessibility systems and tools; Accessibility; Empirical studies in accessibility.

Keywords

Accessibility, LLM, User validation

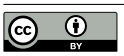
ACM Reference Format:
Nicola Leonardi, Marco Manca, and Fabio Paternò. 2026. Leveraging Large Language Models for Alt-Text Evaluation in E-Commerce: A Data-Driven Study. In *Proceedings of the 2026 International Conference on Advanced Visual Interfaces (AVI '26), June 08–12, 2026, Venice, Italy*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3811427.3811431

1 Introduction

The proliferation of e-commerce platforms has created an urgent need for accessible web experiences, yet many online retailers continue to face significant barriers in maintaining accessibility standards across their digital storefronts. In this perspective, one key point is to provide a meaningful alternative description of the visual images presented. While alternative text (alt-text) is important across all web contexts, its significance is substantially greater in e-commerce environments, where visual information plays a fundamental role in online purchasing decisions. Unlike informational websites, where images often serve supplementary or illustrative purposes, e-commerce platforms rely on product imagery as the primary medium through which customers evaluate, compare, and ultimately decide to purchase items. This visual-centric paradigm creates a unique accessibility challenge: users with visual impairments are systematically excluded from the core shopping experience when alt-text is absent or inadequate [1]. A study [2] validated e-commerce websites using an automated tool and found that links and complementary text descriptions often matched the alt-text, causing the screen reader to read the same text twice. Thus, the proper implementation of alt-text for images remains a persistent challenge, particularly given the characteristics, the scale and dynamic nature of modern e-commerce catalogues. The economic implications of inadequate alt-text in e-commerce are also more pronounced than in other domains. Inaccessible product images directly translate to lost revenue, as users with visual impairments, representing approximately 2.2 billion people globally according to the World Health Organisation [3], either abandon purchases due to insufficient information or avoid inaccessible retailers entirely. In addition, the legal and reputational risks associated with inaccessible e-commerce sites have intensified in recent years, with retailers facing increasing controversy under disability discrimination laws in various jurisdictions. Unlike content-focused websites, where accessibility issues might affect information access, inaccessible e-commerce sites can be construed as denying equal access to goods and services. This distinction carries greater legal weight and potential liability, as demonstrated by the recent collective legal case targeting several large private e-commerce retailers in France for digital inaccessibility [4].

On e-commerce websites, evaluating the original alt-text associated with an image requires a comprehensive analysis that encompasses both the image itself and, crucially, the textual context surrounding it. This is because optimal alt-text is not meant to replicate or duplicate information already present in the surrounding context; rather, it should generate complementary textual content that integrates seamlessly with the existing text by describing the visual characteristics of the product image while deliberately avoiding repetition of what has already been written. The functional role of the majority of images found on e-commerce websites (at least those designated for product overview and selection purposes) can be appropriately classified according to the taxonomy established

*Corresponding author



AVI '26, Venice, Italy

ACM ISBN 979-8-4007-2342-1/2026/06
https://doi.org/10.1145/3811427.3811431

by the Web Accessibility Initiative (WAI) standards [5] under the category of 'informative images', and more precisely within the subcategory defined as 'Images used to supplement other information' [6]. According to WAI guidelines, for images in this subcategory, the accessibility requirement is clearly defined: "A short text alternative is sufficient to describe the information that is displayed visually but is not explained in the text". This definition aligns with the fact that product images in e-commerce contexts typically visually reinforce or complement the textual product information already on the page. This principle is particularly important because assistive reading technologies, such as screen readers, process both the visible text on the page and the alt-text sequentially.

It is therefore clear that continuous monitoring of these sites is necessary to report any problems and then correct them promptly. Manual accessibility auditing presents several fundamental limitations that impede effective accessibility maintenance. The temporal demands of manual validation are prohibitive: e-commerce sites frequently contain thousands of product images distributed across multiple page types, making comprehensive manual review resource-intensive and time-consuming. Effective accessibility evaluation requires specialised expertise that may not be readily available across all teams responsible for content creation and management.

Automated accessibility validation tools can help address these limitations. Their potential near-real-time continuous monitoring capabilities can enable organisations to detect and address accessibility issues as they arise, rather than discovering them during periodic audits. This approach not only reduces the accumulation of accessibility debt but also lowers the barrier to entry for accessibility compliance by enabling team members without deep accessibility expertise to identify and flag issues for remediation. However, such validation tools are currently able to identify whether an alternative description is present but have difficulty indicating whether it is a meaningful representation of the image content. In such cases, they limit to indicate that a manual check is necessary for this purpose.

In the last few years, Large Language Models (LLMs) have demonstrated remarkable capabilities in processing and understanding multimodal content, including the simultaneous analysis of textual and visual information. Their advanced linguistic flexibility and capacity to contextualise information across different modalities make them promising candidates for addressing such accessibility validation challenges. Some studies (e.g. [7], [8]) have begun to explore their potential for accessibility validation. This work aims to expand on these studies by examining the capability of modern LLMs to evaluate existing alternative text descriptions and create new alt-text, with all results benchmarked against real user assessment. The research questions that this study aims to address, considering the e-commerce domain, are:

RQ1: To what extent can Large Language Models (LLMs) be effectively employed to evaluate the accessibility quality of alternative text (alt-text) descriptions for images?

RQ2: Can Large Language Models (LLMs) generate valid and descriptive alternative texts that meet accessibility guidelines?

The answers to the first question can indicate the LLM's effectiveness in supporting accessibility validation tools, while those to the second question can indicate their relevance for their inclusion within authoring tools. To address these research questions, a user study was conducted with N=16 participants who were asked to evaluate the alt-text of images from real e-commerce websites and to write their own alt-text. The same task was submitted to an LLM, and the results, both assessment and generation, were compared, revealing that the LLM is a reliable tool to assess the quality of image text alternatives and that users evaluated the alt-texts generated by the LLM better than the original ones.

2 Related Work

Several recent studies have investigated the application of LLMs to automatic web accessibility validation, revealing both promising capabilities and notable limitations. Research work [9] tested GPT-3.5's ability to identify accessibility problems in HTML elements, including images. A fundamental limitation was that the text-only input format used by the considered model precluded assessing whether alternative text appropriately conveyed image content. In another study [10], researchers tested whether GPT-3.5 could automatically correct WCAG 2.1 violations using prompt engineering and direct DOM modification. A similar approach was developed in another study [11], which deployed a web application for the automatic, real-time correction of web accessibility problems. A contribution [12] focused on alternative text presented a specialised Chrome extension paired with a backend infrastructure designed to generate contextually appropriate image descriptions utilising the GPT-4V vision-language model. The system examines images and their current alternative text alongside webpage contextual elements, producing refined alt text that highlights pertinent visual details, informed by spatial layout and content relationships. The user test revealed that these tailored descriptions significantly improved perceived quality, relevance, and accuracy compared to context-free alternatives. However, their system prompt was minimised to its most basic elements and reduced to a single, brief instruction, while we think a more structured prompting technique can provide more meaningful results.

More recent research [7] explored the application of the multimodal model GPT-4o for automated evaluation of aspects that require semantic and contextual interpretation. The study included an investigation of image alternative text assessment but lacked empirical validation; a comparison with human-generated outcomes could have facilitated a better understanding of the results. In another work [13], the authors evaluate LLM capabilities for detecting accessibility code issues and generating accessible HTML snippets, showing promising results while identifying clear limitations. GenA11y [14] is one of the most comprehensive LLM-based accessibility checkers available, covering several criteria through element extraction and tailored prompts. However, its reliance on semi-synthetic data and validation with only home pages raises concerns about ecological validity and performance on dynamic real-world websites. Furthermore, the study is not specific to e-commerce sites but considers different types of sites.

This work distinguishes itself from previous studies through its focus on real e-commerce web pages and its methodological framework, which employs a rigorous comparative analysis between human users and Large Language Models to both evaluate original alternative text and generate new alt-text descriptions. The research derives actionable insights from empirical data by systematically comparing outcomes using comprehensive descriptive

statistics, correlation analyses, and domain-specific metrics that assess textual complexity, syntactic and semantic alignment, and the relative focus on visual information in the generated alternative text.

3 The Proposed Study

This study focuses on the analysis of the ability of modern LLMs to evaluate one of the most important WCAG techniques that currently requires manual verification, which is the Technique G94 [15] belonging to the success criterion 1.1.1 Non-text Content: Providing short text alternatives for non-text content that serves the same purpose and presents the same information as non-text content. For this purpose, we designed and developed a tool that has two main goals: first, to identify an effective prompting technique for evaluating existing alternative text of images and generating new alternative texts; second, to allow users to compare and assess both the current alternative text and the new text generated by the LLM, and to enter that they would consider the most appropriate.

Product images on e-commerce websites generally aim to visually reinforce or complement the textual product information already available on the page. Therefore, to generate optimal alt text with an LLM, it is essential to provide the LLM with both the image and its surrounding context. The implemented solution operates at the Document Object Model (DOM) level, identifying elements that are siblings or children positioned within a certain distance from the image element. This structural approach leverages the hierarchical organisation of HTML and respects the semantic structure intended by developers, which often groups related content together in the DOM hierarchy. The depth of the hierarchy to be analysed is a key factor in determining the extent of context to consider. We performed a heuristic analysis by selecting random images from e-commerce websites; we found that contextual information obtained by considering 5 DOM levels of depth generally includes the most relevant text and provides a good trade-off. Furthermore, to utilise the most informative text possible, only specific HTML tags were considered for extraction. These tags were carefully selected based on their typical content-bearing nature and include 'p', 'span', 'h1', 'h2', 'h3', 'h4', 'h5', 'h6', 'a'. By limiting the extraction to these tags, the system focuses on elements that conventionally contain meaningful textual content while excluding structural, navigational, or purely presentational HTML elements that would add noise rather than semantic value to the analysis.

The developed tool is a web-based application designed to facilitate the evaluation of alternative text for images on e-commerce platforms. Users can select an e-commerce website from a predefined list. Once a site is chosen, the system automatically extracts all images along with their contextual information and displays the corresponding alt-text. The tool enables users to select a subset of images from the extracted collection; assess the quality of the existing alt-text by providing a rating on a 1–5 scale, propose a revised alt-text for the selected images, invoke the LLM to generate a rating of the original alt-text and provide an assessment of the LLM-generated alt-text. Figure 1 shows its software architecture; the system has been designed according to microservices principles, implementing a clear separation of concerns through two distinct and independent components that communicate via well-defined interfaces: a Gradio (open-source Python library [16]) frontend to manage user interaction and a FastAPI (web framework for building APIs with Python [17]) backend for processing requests and saving logs. This architectural choice was driven by the intention to establish a robust software infrastructure that can be effectively managed and maintained. Furthermore, given that the backend service exposes standard REST HTTP APIs, its functionalities can also be accessed and utilised by external applications and systems that may need to integrate with it in the future. Invoking the service endpoint triggers a comprehensive set of tasks. From the target page, the system extracts the page title, all heading elements, meta tag keywords and page descriptions, and, for each image on the page, gathers both the alt-text attribute (when available) and the surrounding contextual information. Once this information has been collected and aggregated, the system constructs both a system prompt and a user prompt based on the extracted data. These prompts are then used to invoke a Microsoft Azure API service that provides inference capabilities of the GPT-4o model. We set up the following model parameters: Temperature = 0.7, which controls randomness in output generation (range 0.0 to 2.0). 0.7 provides a balanced mix of creativity and consistency; Top-p = 0.95, it considers tokens with top cumulative probability (range 0.0 to 1.0). The value 0.95 is quite permissive, allowing for diverse word choices while filtering out very unlikely options; Max tokens = 800, which defines the maximum number of tokens the model can generate in its final output. This setup creates a moderately creative LLM model instance that produces medium-length responses. It is generally well-suited for content generation and scenarios where natural, somewhat varied outputs without excessive constraints are desired.

The user prompt and system prompt serve distinct but complementary roles in this process. The user prompt carries and organises the contextual information alongside the image, presenting the specific data extracted from the web page in a structured manner. The system prompt, on the other hand, defines the task the LLM should perform, such as evaluating the original alt-text and generating new alt-text in accordance with the W3C WCAG technique G94 [15]. We provide the system and user prompt in the supplementary materials. This prompt output specification provides clear guidelines on structure, length, and content requirements, ensuring that the LLM produces the desired output schema systematically. Specifically, the LLM is configured to generate a structured JSON response including the following fields[1]: original alt-text assessment score: an evaluation performed by the LLM on a scale from 1 (poorest quality) to 5 (highest quality), providing a quantitative measure of the existing alt-text effectiveness; generated improved alt-text: a newly composed alt-text description that the LLM determines to be most appropriate given both the surrounding textual context extracted from the page and the visual content of the image itself, adhering to best practices for accessibility and informational value.

[1]Some output fields are omitted here for brevity. The complete list, along with additional information (e.g., system prompts), is provided in the supplementary materials

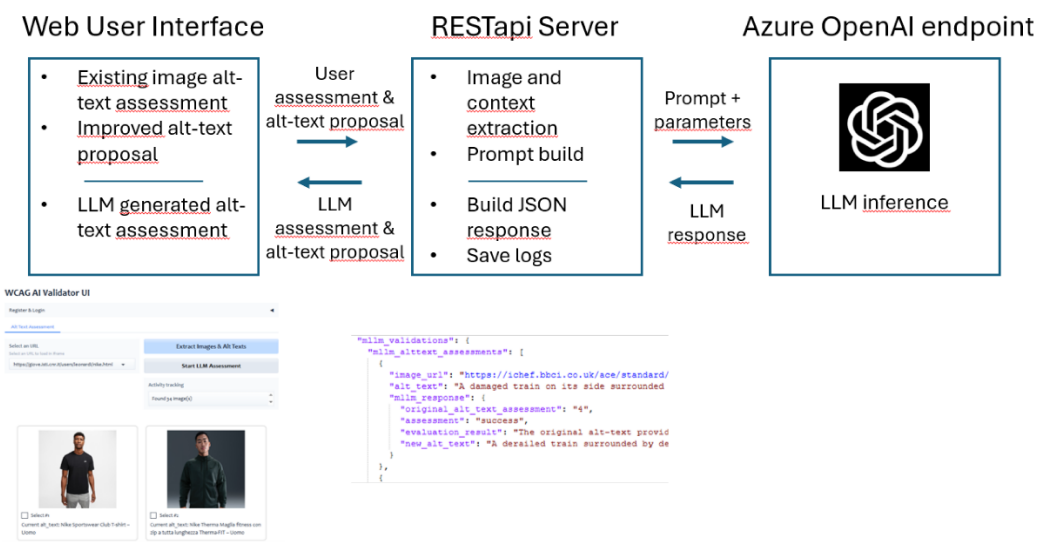


Figure 1: The software architecture of the tool supporting user and LLM testing

4 The User Test

The user study was conducted in mid-December 2025 and involved 16 participants (10 females, 6 males), all students in a master's program in digital humanities. Their ages ranged from 22 to 42 years (average = 25.7, SD = 4.7), and their proficiency in programming was quite varied (7 reported low, 4 medium and 5 good programming experience).

The test was conducted at the end of a university class that included a dedicated 1.5-hour accessibility lecture. Therefore, while participants were not accessibility specialists, they received some background information on WCAG guidelines and accessibility validation tools. Additionally, in preparation for the test, all participants were provided with a document explaining how to use the tool, the test's purpose, the tasks to be completed, and the description of the relevant W3C technique. The exercise was then performed in autonomy and remotely.

The study considered five diverse Web e-commerce applications (eBay, Amazon, Etsy, Decathlon, Nike) to assess a broad spectrum of international online retail environments, encompassing different shopping categories including apparel, artisan goods, and electronic products. Given that the testing phase was anticipated to span roughly two weeks, the considered Web pages were archived in their state as of December 3, 2025. This solution prevented potential variations in content and product listings during the study period. These archived pages, along with their associated JavaScript files, were subsequently hosted on a dedicated web server to maintain consistent conditions throughout the testing duration.

For each e-commerce website, participants were assigned a predefined subset of at least 6 images to analyse. This resulted in at least 30 images per user across all five websites. This systematic assignment strategy was implemented to build a comprehensive dataset of assessed alt-texts, with each image evaluated by multiple participants. Such an approach enabled the subsequent analysis of inter-rater reliability and consistency across different users' assessments.

In concrete terms, the users were required to complete the following tasks: first, select the URL of the page to be analysed and initiate the automated extraction process for images and their original alt-text. Based on their assigned identifier, they then had to select the associated images and evaluate the quality of the existing alt-text on a scale of 1-5, as well as write what they considered to be the most appropriate alt-text. Finally, they were asked to initiate the LLM assessment task, review the results generated by the LLM, and provide a rating on a scale of 1-5 for the new alt-text proposed by the LLM.

5 Results

A total of 494 image alt-texts were validated in this study, corresponding to 157 distinct images. The resulting dataset, therefore, includes multiple evaluations of the same images by different evaluators. The average number of ratings per image is 3.15, with most images receiving at least 3 ratings (86.6%). The users rated on 1-to-5 scale the LLM-generated alt-texts and the original alt-texts. The results can be grouped into three categories:

- No improvement observed: In 181 cases (36.64% of the total), there was no variation between the user's assessment of the original alt-text and the LLM-generated alt-text. This indicates that in more than one-third of instances, users found that the LLM intervention did not result in any measurable improvement.
- Positive improvement observed: In 220 cases (44.53% of the total), there was a positive variation between the user's assessment of the original alt-text and the LLM-generated version. This indicates that the LLM successfully generated improved alt-text that users rated higher than the original.

- Negative variation observed: In 93 cases (18.83% of the total), there was a negative variation, meaning the LLM-generated alt-text was perceived as worse than the original version.

Three cumulative assessment scores were calculated:

- Total original user assessments sum: 1312 – This represents the aggregate of all student evaluations of the original alt-text present on web pages.
- Total LLM assessments sum: 1636 – This represents the aggregate of all LLM evaluations of the same original alt-text.
- Total user LLM-assessments sum: 1619 – This represents the aggregate of all student evaluations of the LLM-generated alt-text.

The difference between the LLM's assessments (1636) and the students' assessments (1312) of the original alt-text indicates that the LLM tends to evaluate the original alt-texts more favourably than human evaluators do. In other words, the LLM appears to be less stringent or more lenient in its judgment criteria compared to the students. This discrepancy warrants further analysis to understand whether this reflects a systematic bias in the LLM's evaluation framework or differences in the interpretation of accessibility standards. Notably, the LLM's optimistic bias is also reflected in both mean (=3.31) and median ratings (=4), which exceed those of human evaluators (mean=2.65, median=3). The standard deviation values, SD = 1.24 for humans versus SD = 1.48 for the LLM, reveal similar rating behaviour patterns, with no evidence of central tendency bias (the tendency to avoid using the full scale and instead cluster ratings around the middle values) for either. The LLM demonstrates greater willingness to differentiate sharply between poor and excellent alt-text quality through more polarised scoring.

The difference between the users' assessments of LLM-generated alt-text (1619) and their assessments of the original alt-text (1312) demonstrates an overall improvement of 307 points in aggregate scoring. This positive differential indicates that, when evaluated by human users, the LLM-generated alt-text is generally perceived as superior to the original versions, suggesting that the system provides meaningful value in enhancing image accessibility, despite cases where no improvement or degradation was observed. It is important to note, however, that while this improvement is statistically observable, the difference is limited. In fact, 307 points corresponds to a 23.4% improvement over the original student assessments, and the results are still distant from the theoretical maximum achievable scores. This indicates that while the LLM-generated alt-texts demonstrate meaningful improvement over original alt-texts in many cases, there remains room for enhancements. This suggests that further refinements to the prompting strategy and context extraction methodology could yield additional improvements in alt-text quality and user satisfaction.

We report three examples extracted from our analysis. In one case, the original alt-text was "*unspecified-8553119*" and received a user rating of 1, while the LLM produced a detailed description "Quechua Men's MH100 Waterproof Mid Hiking Boots, black and blue, waterproof design" scored 5. In another case, both the original "Kiprun Men's Run 100 Dry Running T-Shirt" and the LLM version "Kiprun Men's Run 100 Dry Running T-Shirt for outdoor activities" received an equal rating of 4. In the third case, the original alt-text "Red and gray backpack on a white background" scored 4, while the LLM alternative "Simond Men's MT100 Easyfit 70L Backpacking Pack, red and gray design" received a slightly lower rating of 3.

This section has provided an overview of the assessment data. The subsequent section will advance to a more detailed analysis through two key phases: examining the correlation patterns between human and LLM evaluations on the original alt-texts; and conducting a comparative textual analysis to identify similarities and discrepancies between human-authored and LLM-generated alt-texts.

6 Data analysis

6.1 Inter-user agreement analysis

Since participants in the user test evaluated overlapping image subsets, it is possible to measure the degree of consistency among their judgments. To calculate pairwise correlations between all users, a minimum availability threshold of at least three assessments of the same image was established as a prerequisite for computing the indicator. As discussed previously, this set covers 86.6% of the total alt-text assessed. Inter-rater agreement was evaluated using two complementary measures: Spearman's correlation coefficient (0.48) to assess monotonic association between ratings, and weighted Cohen's Kappa with quadratic weights (0.31) to quantify agreement beyond chance.

These relatively low inter-user coefficients suggest that the alt-text assessment task is highly subjective and poses challenges for the consistent application of criteria. This finding is attributable to several factors inherent in the task design. First, the evaluation requires subjective judgments regarding the roles of images and their relationship to the surrounding context, areas where human evaluators may naturally and legitimately disagree. Second, the provided assessment guidelines and background information are susceptible to varying interpretations across different evaluators, leading to divergent rating behaviours. The modest correlation values thus reflect not only the complexity of the assessment criteria but also the nuanced nature of determining alt-text quality, where multiple valid perspectives may coexist depending on individual interpretation of accessibility standards and contextual relevance.

6.2 LLM intra-agreement

We also calculated LLM intra-agreement (LLM self-consistency) metrics under identical experimental settings to those used for inter-user agreement analysis. Thus, the LLM validated images with the same repetitions as human participants, so these metrics reveal how consistently the LLM evaluates the same image alt-texts across multiple assessments, essentially measuring the model's self-consistency and internal reliability. This assessment is meaningful, considering the non-deterministic nature of LLMs.

The substantially higher correlation coefficients (Spearman= 0.66) and weighted (quadratic) Cohen's Kappa (0.65), compared to the inter-user values, demonstrate that the LLM exhibits considerably greater self-consistency than human evaluators achieve amongst themselves. This suggests more systematic and uniform application of evaluation criteria by the model, while human assessors exhibit greater variability reflecting the subjective nature of alt-text quality judgment.

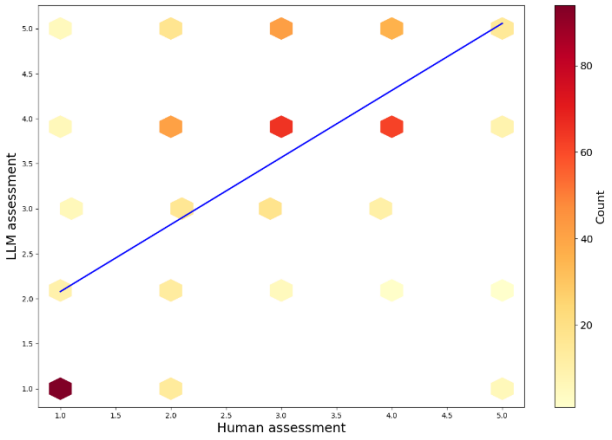


Figure 2: Human VS LLM assessments on the original alt-text

6.3 User-LLM agreements analysis

The evaluation of the LLM against human assessments shows moderate-to-strong consistency. Specifically, Spearman's rank correlation (0.59) indicates a good monotonic alignment between the two groups. Furthermore, the quadratically weighted Cohen's Kappa (0.55) confirms moderate agreement beyond chance.

Figure 2 shows this relationship. The hexagonal bins indicate the concentration of data points where user and LLM assessments coincide. Darker/redder colours represent areas where more assessment pairs fall, revealing patterns of agreement. The blue line represents the best-fit linear relationship between assessments, or in other words, it shows the overall trend and direction of the relationship.

This figure aligns with and complements the findings described before. It reveals the LLM's tendency to assign more generous scores compared to human evaluators, as evidenced by the concentration of darker points in the upper portion of the graph. The regression line's deviation from the 45-degree diagonal confirms that the LLM systematically scores higher than human assessors. However, the plot also demonstrates strong agreement at the lower end of the scale, where both LLM and human evaluators consistently identify and concordantly rate poor-quality alt-texts with low scores. Putting all together, the data reveal a compelling finding: users agree more with the LLM than with each other. In other words, the LLM (when provided with proper prompts and context) serves as a more reliable alt-text evaluator than individual humans, likely because the task's high difficulty and subjectivity lead human assessors to apply inconsistent or unclear evaluation criteria. In essence, the LLM has successfully learned to navigate the subjectivity of alt-text assessment by identifying patterns that represent collective human judgment better than those captured by selected individuals. These findings suggest that with the right settings and prompts, a modern LLM can reliably pre-screen image alternative text. Notably, the quality of these automated assessments surpasses the reliability of individual human evaluators, making them viable for practical implementation.

6.4 LLM Classifier

To evaluate the LLM's performance as a classification system, a binary classification framework has been constructed. The ground truth for each image was established by averaging all user assessments and converting the 1-5 rating scale into binary categories (0-1), where ratings of 1-2 were classified as inadequate alt-text (0) and ratings of 3-5 as adequate alt-text (1). To characterise the LLM's behaviour as a binary alt-text quality classifier, we report both the standard performance metrics and the underlying confusion matrix values. The confusion matrix summarises the classification outcomes as follows: True Positives (TP) = 107, True Negatives (TN) = 39, False Positives (FP) = 6, and False Negatives (FN) = 5 (total N=157). These counts indicate that the model correctly identifies the majority of high-quality and low-quality alt-texts while maintaining low rates of both false alarms and missed positives. The resulting performance metrics (Accuracy= 93.0%, Precision= 94.7%, Recall= 95.5%, F1 Score= 95.1%, Cohen's Kappa= 82.8%) demonstrate strong performance by the LLM as a binary alt-text quality classifier. These results provide a comprehensive answer to RQ1: they validate the LLM as a highly reliable automated tool for alt-text quality assessment. The model could be confidently deployed at scale for automated quality screening, flagging inadequate alt-texts for human review.

6.5 Human and LLM-generated alt-texts analysis

To assess whether an LLM can generate valid alt-texts that meet accessibility standards (RQ2), a systematic comparison was conducted between participant-generated and LLM-generated alt-texts across multiple dimensions: text length, language fluency, syntactic and semantic similarity, and image adherence (accuracy of visual content description). This multi-indicator framework can reveal both convergences and divergences in how humans and LLMs approach alt-text creation, providing insights into the LLM's capacity to replicate human authorship patterns and identifying systematic differences in description strategies and priorities.

The rationale behind analysing authentic human-generated alt-text rather than the original one is tied to two main reasons: in several cases the original alt-text was of poor quality, and more broadly, it was not possible to determine whether the existing alt-text had been written by humans or generated by automated tools.

6.5.1 Lengths statistics. Table 1 presents length statistics for generated alt-texts, measured in character count. This metric is particularly relevant because WCAG Technique G94 [15] identifies brevity as a core characteristic of effective alt-texts, distinguishing it from more verbose formats like captions or detailed image descriptions. Effective alt-text must balance informativeness with conciseness to optimise accessibility.

The statistics shown in Table 1 demonstrate strong alignment between human and LLM alt-texts across central metrics (mean, median, 75th percentile), indicating comparable typical lengths and shared adherence to brevity conventions. The most significant discrepancy appears in minimum values: humans assign empty strings when identifying decorative images, a valid accessibility practice where null alt-text (alt="") allows screen readers to skip

Table 1: Human and LLM proposed alt-text length descriptive statistics

Statistics	Human alt-text length	LLM alt-text length
Mean	78	80.5
SD	42.4	28.3
Min	0	11
25%	51	62
50%	72	75
75%	96.8	93
Max	401	193

non-functional imagery. Conversely, the LLM never makes null assignments, tending to provide at least minimal descriptions even for potentially decorative images. This suggests some misalignments distinguishing decorative from functional content[2].

The weak positive correlations (Pearson: 0.27, Spearman: 0.28, Kendall: 0.19) suggest that while human and LLM alt-text lengths show some correspondence, the two sources make largely independent decisions about appropriate verbosity, likely prioritizing different aspects when determining detail levels.

6.5.2 Readability Metrics. In line with WCAG Technique G94, which emphasizes clarity and equivalence of alternative text, we assessed the readability of LLM-generated descriptions using Flesch Reading Ease [18] and Gunning Fog Index [19]. These metrics provide an objective measure of linguistic complexity, ensuring that alt-text remains accessible to users with varying cognitive abilities. The Flesch Reading Ease Index evaluates how much a text is easy to read, it scores text from 0 to 100 based on sentence length and syllables per word, where higher scores mean easier reading. Scores above 60 represent conversational, accessible text suitable for general audiences, while scores below 30 indicate complex academic or professional writing. It is widely used [20] to ensure content matches the reading level of intended audiences. The Gunning Fog Index estimates the years of formal education needed to understand text on first reading. It calculates a grade level based on sentence length and percentage of complex words (three or more syllables). A score of 8 means eighth-grade readability, while scores above 14 indicate college or graduate-level complexity. Unlike Flesch, higher scores indicate harder text, with the number directly corresponding to educational grade level. Importantly, both indices measure only structural mechanics (sentence length, syllables, word complexity) and do not assess logical coherence, narrative flow, or semantic clarity. They evaluate surface-level complexity rather than whether the content is actually clear, well-organized, or meaningful.

Regarding Flesch Reading Ease, both human (mean: 66.07) and LLM (mean: 59.82) alt-texts fall within the "standard/conversational" readability range (60-70), though the LLM produces slightly more difficult text on average. The medians (67.14 vs 61.21) confirm this pattern. Mean and median values for Gunning Fog Index are aligned to Flesch Reading Ease results, showing that the LLM produces consistently more complex text. The moderate Flesch Reading Ease positive correlations between human and LLM (Pearson: 0.44, Spearman: 0.47, Kendall: 0.32) reveal an interesting pattern. These correlations are notably higher than the length correlations, suggesting that while humans and LLMs make independent decisions about how much to write, they show moderate agreement on how complexly to write: both recognize which images justify simpler language rather than more technical language. In a nutshell, humans and LLMs diverge more on verbosity (what to include/exclude) than on linguistic complexity (how to express it).

6.5.3 Lexical and semantic analysis. To address the inherent limitations of readability indices the study conducted a complementary lexical and semantic analysis of both LLM-generated and human-authored alt-texts. This complementary approach examines vocabulary choices, semantic patterns, and conceptual content, revealing whether both author types employ similar descriptive strategies and meaning-construction approaches beyond surface-level complexity. The work employed three different similarity measurement techniques: lexical similarity, semantic similarity through the use of a pre-trained neural language model and BERTscore similarity [21] that computes token-level similarity between texts, matching words based on their contextual meaning rather than exact matches.

Lexical similarity measures word-level overlap between two texts using Term Frequency-Inverse Document Frequency (TF-IDF) vectorization followed by cosine similarity calculation. The texts are converted into numerical vectors where each dimension represents the importance of a specific word or n-gram (word level with n-gram=1 is used in the work). Cosine similarity then compares these vectors, producing a score between 0 and 1 where higher values indicate greater lexical overlap. This approach captures whether texts use similar vocabulary but ignores word meaning, synonyms, or semantic relationships. Semantic similarities technique uses a pre-trained neural language model (all-MiniLM-L6-v2 in the analysis) to capture deep semantic meaning rather than just word overlap. The model encodes each text into a dense embedding vector that represents its meaning in a high-dimensional semantic space, where semantically similar texts are positioned closer together. Unlike lexical methods, this approach understands context and meaning. BERTscore similarity follows a similar approach to semantic similarity methods but goes deeper by finding optimal alignments between tokens in the two texts based on their contextual embeddings, providing a more nuanced semantic comparison than simple cosine similarity of sentence embeddings.

Human and LLM alt-texts show relatively low lexical overlap, with an average similarity of only 0.38. This indicates that humans and LLMs frequently choose different vocabulary and phrasing when describing the same images, they may describe different visual elements, use distinct terminology, or structure their descriptions differently. Despite limited lexical overlap, semantic similarity is substantially higher at 0.67, with a tighter distribution (SD: 0.20). The median of 0.71 indicates that in most cases, humans and LLMs convey similar underlying meaning even when using different words. BERTScore produces the highest and most consistent similarity scores (mean: 0.69, SD: 0.15), with the distribution concentrated in the 0.62-0.79 range. This contextualised token-matching approach reveals even stronger alignment than sentence-level embeddings, suggesting that when accounting for contextual meaning

[2]These cases represent edge situations where an image with adjacent text is part of the same link. Technically, such images are in general classified as decorative.

Table 2: Human and LLM proposed alt-text readability descriptive statistics

Statistics	Human Flesch Reading Ease index	LLM Flesch Reading Ease Index	Human Gunning Fog index	LLM Gunning Fog index
Mean	66.07	59.82	16.27	17.96
SD	25.3	20.38	7.32	6.1
Min	-96.26	-31.35	0	0.8
25%	53.13	47.06	11.6	13.2
50%	67.14	61.21	16.67	18.56
75%	80.14	73.64	21.35	22
Max	129.05	114..09	41.2	35.73

Table 3: Human and LLM proposed alt-text similarities descriptive statistics

Statistics	Lexical Similarity	Semantic Similarity	BERTscore Similarity
Mean	0.38	0.67	0.69
SD	0.24	0.2	0.15
Min	0	0	0
25%	0.19	0.57	0.62
50%	0.35	0.71	0.71
75%	0.55	0.83	0.79
Max	1	1	1

at the word level, humans and LLMs demonstrate substantial conceptual agreement in their alt-text generation strategies. The progression from lexical (0.38) → semantic (0.67) → BERTScore (0.69) reveals a valuable insight: humans and LLMs describe images with comparable semantic content but divergent surface-level expression. They identify, merge and communicate similar information but through different linguistic pathways, demonstrating convergent understanding despite stylistic differences. It is noteworthy that careful manual examination of the generated outputs reveals no evident signs of LLM hallucinations. In fact, the implementation functions as a Retrieval Augmented Generation (RAG) system in which generation is well-grounded in the provided multimodal context (image and text).

6.5.4 CLIP-score analysis. In the context of evaluating the accessibility of e-commerce websites, where images often provide supplementary information beyond the textual content, it is essential to assess how effectively alternative texts generated by humans or LLM convey the visual information contained in those images. To address this, we use CLIP (Contrastive Language-Image Pre-training [22]) scores as a proxy for semantic alignment between the image and its associated alt text. Specifically, CLIP scores [3] have been used to measure how effectively human and LLM alt-texts capture visual information from the source images. CLIP quantifies the semantic alignment between image and text in a shared embedding space, where higher scores indicate better correspondence between visual content and textual description. This metric is particularly valuable for accessibility assessment, as effective alt-text must faithfully convey salient visual information. Comparing CLIP scores reveals which author type more successfully identifies and communicates the visually relevant elements essential for accessible image descriptions.

To enrich the analysis, the LLM was evaluated under three conditions corresponding to different prompt inputs: (1) both image and context (default), (2) image only, and (3) context only. During the user test, we provided the LLM both the image and the context, and afterwards we repeated the same evaluation by providing only the image or only the context. This design serves dual purposes: comparing human-LLM differences and revealing the LLM's internal generation patterns, specifically how it weights and integrates visual versus textual information when producing alt-text. By systematically removing input modalities, it is possible to observe how the LLM's output changes, exposing how it prioritises and synthesises multimodal inputs. Crucially, this controlled ablation is only possible with the LLM. Human participants received both modalities simultaneously, making their internal weighting mechanisms impossible to decompose. Table 4 shows the results.

To quantify distributional similarity between CLIP scores across conditions from multiple perspectives, four complementary metrics have been computed on the three conditions: Kolmogorov-Smirnov statistic, Wasserstein distance, Jensen-Shannon divergence, Pearson correlation. The condition most similar (KS = 0.06, WD = 0.52, JS = 0.15, $\rho = 0.51$) to the human CLIP score was the context-only (original alt-text and the surrounding text) LLM condition[4]. This finding suggests that human alt-text generation strategies are more context-driven than vision-driven. The next objective was to determine under which condition the LLM produced a CLIP score distribution most similar to the default configuration (context+image). The four metrics defined above consistently identified the

[3]Calculated as the dot product of the image and text embeddings multiplied by the model's learned logit scale. The used model is openai/clip-vit-large-patch14.

[4]Details are available in the supplementary material

Table 4: Human and LLM proposed alt-text CLIP scores descriptive statistics.

Statistics	alt-text Human	alt-text LLM(context+image)	alt-text LLM(only context)	alt-text LLM(only image)
Mean	24.72	26.11	24.92	26.07
SD	5.34	4.13	5.35	4.39
Min	0*	10.38	5.36	12.74
25%	21.99	23.72	22.17	23.23
50%	25.61	26.53	25.91	26.52
75%	28.06	28.64	28.27	29.10
Max	38.11	38.11	39.11	38.34

*0 means no alt-text provided, considered a decorative image

alt-text configuration (image only) as most similar (KS = 0.05, WD = 0.39, JS = 0.14, $\rho = 0.65$).

This pattern in the experiment data suggests that the LLM's multimodal generation process is vision-dominant: when both image and context are provided, the output more closely resembles image-only generation than context-only generation, indicating visual information carries greater weight in the model's decision-making. The context may help with narrative coherence but does not substantially change what visual elements the LLM identifies and describes. These results are validated with some statistical tests (significance tests, effect sizes, multiple comparison corrections) that reveal significant differences among conditions ($p<0.0001$), though effect sizes are small (Cohen's d = 0.22-0.28), indicating modest practical differences (1-1.5 point difference in a 24-26 CLIP scores range). Although the results confirm that LLMs are vision-dominant in multimodal settings and that humans align more with context-only than image-only LLM behaviour, the differences, while statistically significant, are modest in practical terms, with all approaches achieving reasonable visual alignment.

Overall, to summarise, the improvement rate reported by participants, whereby LLM-generated alt-text was preferred over original descriptions in approximately one-quarter of cases (23.4%), leads to a positive answer to RQ2: LLMs can generate alt-text that is both valid (semantically faithful to the image) and descriptive (rich enough to aid understanding). However, the data analysis shows that the LLM produces longer texts with higher reading complexity, and despite low lexical overlap with human-authored descriptions, both semantic similarity measures and BERTScore indicate strong alignment in meaning, suggesting that LLMs capture the salient visual content even when it is phrased differently.

7 Conclusions and Future Work

This study examines the potential of modern LLMs to both evaluate existing alternative text descriptions and generate new alt-text, with all results systematically benchmarked against authentic user perspectives. A user study was conducted in which participants evaluated and generated alt-text for images from actual, widely used e-commerce websites, while an LLM completed identical tasks under comparable conditions. The results on the effectiveness of the LLM as an alt-text evaluator indicate its strong performance as a binary alt-text quality classifier. For LLM-generated alt-texts, the data indicate positive results that can be further improved based on the analysis performed.

The most significant contributions of this work are as follows: first, it provides clear statistics that illustrate how well an LLM, when appropriately instructed, can serve as a valid alternative to human users in validating existing alt-text. Second, it describes specific indicators that highlight the alignment between alt-text produced by humans and that generated by LLMs, considering factors such as syntax, semantics, length, and readability. Additionally, for each proposed new alt-text-image pair, it generates the CLIP score index, which measures the importance of the visual component in relation to the surrounding context, as evaluated by both humans and the LLM. Finally, the work releases a dataset of approximately 500 examples including evaluations from both users and LLM, the newly produced alt-text, and the human rating of the LLM proposed alt-text. This dataset can be employed in subsequent studies for supervised fine-tuning (instruction tuning) and for reinforcement learning (RL) purposes.

A limitation of this study is that our reference baseline relied on students with basic knowledge of accessibility rather than on experts. As future work, it would be valuable to replicate the experiment with accessibility specialists and/or BLV users. Finally, we plan to extend the study and methodological approach to different website contexts and accessibility techniques to provide broader insights into the LLM's performance across different digital environments and task types.

Acknowledgments

We acknowledge the use of Generative AI tools to improve the grammar, style, and readability of this manuscript. These tools were used exclusively for text editing and played no role in the data analysis, interpretation, or generation of the core findings presented.

References

[1] S. Hamid, N. Z. Bawany and K. Zahoor, "Assessing Ecommerce Websites: Usability and Accessibility Study," 2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Depok, Indonesia, 2020, pp. 199-204, doi: 10.1109/ICACSIS51025.2020.9263162.

[2] Gonçalves, R., Rocha, T., Martins, J. *et al.* Evaluation of e-commerce websites accessibility and usability: an e-commerce platform analysis with the inclusion of blind users. Univ Access Inf Soc 17, 567–583 (2018). https://doi.org/10.1007/s10209-017-0557-5

[3] World Health Organization. (2023, August 10). Blindness and vision impairment. https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment. Accessed January 2026.
[4] Intérêt à Agir. (2025, November 12). Assignation en référé des entreprises Auchan, Carrefour, E. Leclerc et Picard Surgelés. https://www.interetaagir.org/assignation-en-refere-des-entreprises-auchan-carrefour-e-leclerc-et-picard-surgeles/
[5] World Wide Web Consortium. (n.d.). Web Accessibility Initiative (WAI). https://www.w3.org/WAI/. Accessed January 2026.
[6] World Wide Web Consortium. (n.d.). Informative images. Web Accessibility Initiative (WAI). Accessed January 9, 2026, from https://www.w3.org/WAI/tutorials/images/informative/#example-2-images-used-to-supplement-other-information
[7] Paternò, F., Vinci, M., Manca, M., & Iannuzzi, N. (2025). How an LLM can improve automatic web accessibility validation? In Proceedings of the 16th Biannual Conference of the Italian SIGCHI Chapter (CHItaly 2025). Association for Computing Machinery. https://doi.org/10.1145/3750069.3750310
[8] López-Gil, JM., Pereira, J. Turning manual web accessibility success criteria into automatic: an LLM-based approach. Univ Access Inf Soc 24, 837–852 (2025). https://doi.org/10.1007/s10209-024-01108-z
[9] Delnevo, G., Andruccioli, M., & Mirri, S. (2024). On the Interaction with Large Language Models for Web Accessibility: Implications and Challenges. 2024 IEEE 21st Consumer Communications & Networking Conference (CCNC), Las Vegas, pp. 1-6. https://doi.org/10.1109/CCNC51664.2024.10454680
[10] Huang, C., Ma, A., Vyasamudri, S., Puype, E., Kamal, S., Garcia, J. B., Cheema, S., & Lutz, M. (2024). ACCESS: Prompt engineering for automated web accessibility violation corrections. arXiv. https://doi.org/10.48550/arXiv.2401.16450
[11] Dash, S. K. (2024). AI-powered real-time accessibility enhancement: A solution for web content accessibility issues. JOIN (Journal Online Informatika), 9(1), 70–79. https://doi.org/10.15575/join.v9i1.1310
[12] Gubbi Mohanbabu, A., & Pavel, A. (2024). Context-aware image descriptions for web accessibility. Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '24), Article 62, 1–17. https://doi.org/10.1145/3663548.3675658
[13] Bassi, B., Delnevo, G., Franco, M., Gaggi, O., Gatto, S., Mirri, S., & Olaiya, K. (2025, September 3–5). An assessment of LLM-based auditing and validation for web accessibility. In GoodIT '25: Proceedings of the 2025 International Conference on Information Technology for Social Good (pp. xx–xx). Association for Computing Machinery. https://doi.org/10.1145/3748699.3749805
[14] Ziyao He, Syed Fatiul Huq, and Sam Malek. 2025. Enhancing Web Accessibility: Automated Detection of Issues with Generative AI. Proc. ACM Softw. Eng. 2, FSE, Article FSE101 (June2025), 24 pages. doi:10.1145/3729371
[15] World Wide Web Consortium. (n.d.). G94: Providing ad hoc alternative text for non-text content. https://www.w3.org/WAI/WCAG21/Techniques/general/G94
[16] Gradio. (n.d.). Gradio documentation. Accessed January 14, 2026, from https://www.gradio.app/docs
[17] FastAPI. (n.d.). FastAPI documentation. Accessed January 14, 2026, from https://fastapi.tiangolo.com
[18] Flesch, R. (1948). A new readability yardstick. Journal of Applied Psychology, 32(3), 221–233. https://doi.org/10.1037/h0057532
[19] Gunning, R. (1952). The technique of clear writing. McGraw-Hill
[20] Microsoft. (n.d.). Get your document's readability and level statistics. Microsoft Support. https://support.microsoft.com/en-au/office/get-your-document-s-readability-and-level-statistics-85b4969e-e80a-4777-8dd3-f7fc3c8b3fd2
[21] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. Proceedings of the 8th International Conference on Learning Representations. https://openreview.net/forum?id$=$SkeHuCVFDr
[22] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning (Vol. 139, pp. 8748–8763). PMLR. https://proceedings.mlr.press/v139/radford21a.html