Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

AI versus Human: Evaluating the Accuracy and Usefulness of Alt-Text Generation for Swedish Image Accessibility

Lotta Fejzullahu
Linnaeus University
Emma Johansson Olsson
Linnaeus University
Mexhid Ferati
mexhid.ferati@lnu.se
Linnaeus University

Research Article

Keywords: Alt text, artificial intelligence, digital accessibility, generative AI, image description

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11161157/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

AI versus Human: Evaluating the Accuracy and Usefulness of Alt-Text Generation for Swedish Image Accessibility

Lotta Fejzullahu, lf222yw@student.lnu.se

Emma Johansson Olsson, ej224et@student.lnu.se

Mexhid Ferati*, mexhid.ferati@lnu.se, https://orcid.org/0000-0001-8635-4069 (corresponding author)

Informatics Department, Linnaeus University, Kalmar & Växjö, Sweden

Abstract

Digital accessibility is essential for equal access to web content, which highlights the importance of alternative texts (alt texts) for people with visual impairments. The rapid development of artificial intelligence (AI) has created new opportunities for automating alt-text generation. This study investigates the extent to which AI-generated alt texts in Swedish convey accurate and relevant information compared to human-written alternatives, and identifies key opportunities and challenges of using AI for this purpose. Based on a questionnaire, 79 respondents evaluated alt texts for 15 images, each accompanied by three alternatives: one generated by ChatGPT-4o, one by Claude Sonnet 3.7, and one human-written. The questionnaire collected quantitative and qualitative data on perceived accuracy and relevance. Results indicate that AI-generated alt texts were perceived as more accurate and relevant than human-written alternatives, particularly for images depicting objects and places, and were considered sufficiently accurate to be acceptable despite minor errors. Qualitative findings highlight areas for improvement, particularly in adapting alt texts to visually impaired users' needs. We conclude that AI holds strong potential for alt-text production, though human review remains important for quality and audience adaptation; a hybrid approach is therefore recommended.

Keywords

Alt text, artificial intelligence, digital accessibility, generative AI, image description

Introduction

In today's digitalised society, web accessibility is a fundamental aspect of inclusion. For people with visual impairments, alternative image texts (alt texts) are crucial to accessing image-based information online [19]. When publishing images and illustrations, editors therefore need to take people who are blind or have low vision (BLV) into account [38]. Despite this, missing and inadequate alt text remain among the most common accessibility issues on the web today [43]. At the same time, the development of artificial intelligence (AI) has opened up new possibilities for automated alt-text generation, although its accuracy and usability remain a contested topic [7]. This study evaluates the quality of AI-generated alt text and reflects on the potential and challenges of using AI as an alt-text author.

In Sweden, disability is a protected ground for discrimination, and lack of accessibility is a common form of discrimination, as shown by the large number of reported cases related to disability and inadequate accessibility [12]. Good web accessibility is essential to ensure equal access to online information and services for all individuals, regardless of functional ability [11].

When the Agency for Digital Government (DIGG) reviewed the accessibility of various public sector websites in Sweden during 2020–2021, it found that the absence of alt text was among the six most common reasons why accessibility requirements were not met [10]. Similarly, the WebAIM Million accessibility evaluation found that more than one in four images on popular home pages had missing, questionable, or repetitive alternative text, indicating that alt-text issues remain widespread internationally [43]. This lack can be explained by several factors, including time constraints, limited expertise, and insufficient resources; writing meaningful alt text is time-consuming and requires specific accessibility-related competence, which often leads to alt text being deprioritized in development processes [13, 36]. At the same time, the rapid development of AI has sparked interest in how such technology can be used to increase efficiency across different domains [39], given that generative AI can produce new content and describe images [18]. Recent work in this journal indicates that this interest extends to accessibility evaluation, where LLM-based systems have been explored for success criteria that conventionally require manual checks, including WCAG 1.1.1 Non-text Content [28].

Since ChatGPT gained image recognition capabilities in September 2023, users have been able to upload images and inquire about their content [33]. Although ChatGPT is able to provide highly accurate descriptions of images, the use of AI to generate alt text has been subject to criticism. AI systems tasked with producing alt text often struggle to interpret contextual nuances, resulting in descriptions that may be precise yet not contextually appropriate [7]. Previous studies indicate that AI has not consistently outperformed humans in generating high-quality alt text [26].

Several studies have shown that alt text is often missing when images are published online [5, 10, 20, 43]. This issue merits renewed attention, and since the European Union's new accessibility directive came into effect on 28 June 2025 [31], addressing the current lack of accessible alt text is particularly timely.

The rapid development of AI motivates a renewed examination of its potential for alt-text generation, and of whether previous conclusions remain valid or new opportunities have emerged. Unlike prior studies, which have generally focused on English-language material, we concentrate on alt text generated in Swedish. Given the pace of development of AI models, it is necessary to investigate whether these tools can now produce alt text that is comparable to, or even surpasses, human-written alternatives.

The purpose of this study is therefore to investigate the role artificial intelligence can play in alt-text creation, and specifically to assess the extent to which AI can generate alt text of sufficient quality to be useful for BLV users, with a focus on Swedish-language alt text. To achieve this purpose, the following research questions were formulated:

RQ 1: To what extent do AI-generated alt texts convey accurate and relevant information compared to human-written alternatives, as perceived by sighted evaluators?

RQ 2: What are the key opportunities and limitations of using current-generation multimodal LLMs for automated alt-text production?

Given the numerous AI tools available, we limit our scope to testing a select number of them. Initial examination of well-known AI models revealed that not all could process images, which is essential for image-based alt-text generation; only AI tools with functional image recognition, including the ability to process images containing people, were therefore included.

Literature review

This chapter presents previous research within the areas of alt text and artificial intelligence.

Accessibility Guidelines and Alt Text

There is a global standard with guidelines to achieve good web accessibility, WCAG, and a European standard, EN 301 549, for accessibility requirements for information and communication technology (ICT) within the EU [15, 41]. Alt text is the description of images and other non-text-based information [42]. According to Guideline 1.1 Text Alternatives in WCAG 2.1, text descriptions of all informative non-text content are required to make such content accessible to individuals who cannot see the visual content [41].

Alt text is used in different circumstances, and a good alt text takes into account the context in which an image will be used. One common category on the web is informative images. These images convey a message or an idea, and the alt text should therefore convey the information the image is intended to communicate in a short phrase or sentence, without unnecessary detail. The user should get a quick understanding of the image without losing focus of the surrounding content [14, 35, 37].

Additionally, technical limitations in assistive technologies affect the length of alt text. Many screen readers truncate the text after approximately 125–140 characters [40], and braille displays are limited in the number of characters that can be presented at once, typically 40, 70, or 80 characters [1]. Although there is no official standard for the maximum length of alt text, various sources are relatively unanimous that alt text should usually be kept under 80–100 characters [17, 40, 1].

Challenges and Limitations of Alt Text Implementation

It is important for online images to have alt text, both to comply with global guidelines and accessibility requirements [15, 41], to prevent discrimination [12], and to comply with the EU Charter of Fundamental Rights [16]. Despite this, statistics show that many websites and digital platforms do not meet accessibility requirements [5, 10, 20, 43].

Back in 2016, Twitter (now named X) introduced a feature that let users write their own alt text for their images. Three years later, only three of the 50 most followed Twitter accounts had ever used the feature; furthermore, 99.9% of original images on Twitter lack alt text [20]. This significantly limits access to visual content for those in the BLV community.

Two decades ago, 60.4% of images on the 500 most visited websites lacked alt text [5]. More recently, one-fourth of images on frequently visited websites' home pages are likely to lack or have inadequate alt text [43]. This lack of alt text often forces BLV users to rely on the help and support of others, instead of being able to perform tasks independently [21].

Creating alt text is a resource-intensive process characterized by several organizational and economic challenges. The most common reasons for the absence of alt text in the publishing sector,

particularly among smaller publishers and non-profit organizations, are limited time and budget, a lack of internal expertise, and an insufficient understanding of the importance of alt text [36].

AI as an alt-text writer

Having humans write alt text for every image on a website can be challenging and time-consuming, which imposes significant effort on web developers and content creators [26]. AI-based language tools have received increased attention, as they can both automate parts of the writing process, thereby reducing writers' workload, and review the language in texts, which can increase quality [39].

Previous research has evaluated the quality of AI-generated alt text by comparing it with human-written descriptions and assessing perceived accuracy and precision using Likert-scale ratings. Leotta et al. [26] found that human-written alt text was generally rated as more accurate and informative than AI-generated alternatives. Their study shows that AI has the potential to support the automation of alt text but that it cannot yet fully replace human-written alt text.

Other studies have explored AI-generated alt text as a supportive writing aid rather than as a final output. Mack et al. [29] showed that authors appreciated having access to automatically generated alt text as a starting point, since it made the writing process faster and easier. However, the study also found that reliance on automated suggestions could negatively affect the quality of the final alt text, as authors tended to treat the AI-generated descriptions as sufficient, which resulted in shorter and less detailed alt text [29].

The development of AI's ability to understand images

A major advancement in artificial intelligence is the development of generative AI. Unlike earlier discriminative models, which are trained to classify or predict data based on predefined categories, generative AI relies on models that learn the underlying structure and distribution of data and can generate new content based on patterns in the training data [18].

Many Large Language Models (LLMs) have evolved from handling text-only input to becoming multimodal. A multimodal AI model can process and integrate multiple data modalities, including text, images, and audio, simultaneously [18]. During 2023 and 2024, Multimodal Large Language Models (MLLMs) emerged as a leading technology for tasks that combine visual and textual information, including image description, visual dialogue, and Visual Question Answering (VQA) [6]. Advanced AI models such as OpenAI's ChatGPT-4o and Anthropic's Claude Sonnet 3.7 are capable of analyzing images and generating text responses or descriptions accordingly [2, 33]. This multimodal capability is highly relevant for alt-text generation, as it enables a deeper understanding of visual content in relation to context and user needs.

Alongside this development, transformer-based image captioning systems have become the standard approach in image description tasks. Earlier image captioning systems have historically tended to produce overly literal or generic captions; transformer-based systems demonstrate improved semantic understanding, resulting in more meaningful and context-aware descriptions [6]. However, a significant challenge that remains for advanced LLMs is the phenomenon of hallucinations, in which models generate content that appears plausible but is factually incorrect, raising concerns regarding the reliability of LLMs and remaining an active area of research [22]. In MLLMs that combine text and image understanding, often referred to as Large Vision Language Models (LVLMs), this manifests as multimodal hallucinations, where generated descriptions may conflict with the visual input, such as inventing non-existent objects, misidentifying attributes, or forming incorrect relationships between elements in a scene. One reason for this is the models' tendency to rely on learned linguistic patterns, which can lead them to generate text that is linguistically correct but visually incorrect [22].

Modern MLLM-based models have surpassed previous generations in terms of performance [6] since the publication of earlier studies on AI-generated alt text (e.g., [26, 29]). It is therefore well justified to test today's systems, as they may have greater potential to generate accurate and useful alt text than their predecessors. At the same time, the problem of multimodal hallucinations remains, making it relevant to investigate the extent to which hallucinations occur in these modern systems and whether they affect the usability of alt text in practice.

Conceptual framework and positioning

We use the concepts of digital accessibility, the function of alt text, and multimodal AI as an analytical framework for understanding how AI can contribute to accessibility adaptation. The study is based on guidelines such as WCAG 2.1 [41] and previous research on user needs [37] to assess the quality and relevance of alt text alongside the collected data.

Previous studies, such as Leotta et al. [26] and Mack et al. [29], have shown that AI-generated alt text is often less accurate and informative than human-written alt text. However, these studies evaluated older AI models and focused on English-language texts, creating a knowledge gap regarding both how newer multimodal AI models perform and how they perform in other linguistic and cultural contexts.

This study contributes new knowledge about how modern AI models perform in practice and, by examining Swedish AI-generated alt text, contributes new knowledge in an under-researched area. The language dimension has received limited attention, although recent work has begun to address multilingual alt-text delivery, noting that existing systems typically provide brief descriptions that omit details such as facial expressions and body language [4]. Our study broadens the field by including the modern AI models ChatGPT-4o and Claude Sonnet 3.7, which have the capacity to generate more contextually relevant and semantically rich alt text [18, 39]. The study takes an accessibility perspective and focuses not only on the technical correctness of the text but also on its function as an accessibility tool for users with different needs.

This conceptual framework has guided our choice of methodology. Previous research by Leotta et al. [26] informed the data collection method; articles on AI informed the selection of AI tools; and [1, 17, 40] were taken into account in the development of prompts. Various studies also informed the formulation of the human-written alt text [9, 25, 40, 42]. The framework further supports our analysis and discussion of the results, ensuring that our conclusions are grounded in both previous research and established accessibility principles [6, 18, 22, 25, 26, 29, 39].

This study positions itself as a contribution to the ongoing discussion about the role of AI in promoting digital inclusion [6, 7]. It offers a practical and contextual understanding of AI's potential to improve accessibility on Swedish digital platforms.

Methods

This study employed a mixed-methods approach, combining quantitative ratings with qualitative comments to strengthen internal validity through methodological triangulation. Data were collected through a questionnaire using a five-point Likert scale, which is well suited to measuring the intensity of perceptions [23]. The purpose of the data collection was to evaluate the extent to which different AI tools could generate alt text that conveys accurate and relevant information compared to human-written alternatives. The sample comprised 79 respondents, which was broad and included adults from different age groups and backgrounds.

RQ1 partly aimed to determine how accurately AI can write alt text, and participants therefore needed to be able to clearly see the image in order to compare its content with what the alt text described.

This would have been difficult for BLV users to assess, and no particular focus was therefore placed on including them in the sample. At the same time, the perspective of visually impaired users is highly relevant in this context, as they constitute one of the main target groups for alt text. To obtain the richest data from this group, it would have been more appropriate to collect qualitative data, for example, through interviews on the suitability of the different alt texts. This, however, remains an important direction for future research.

Two multimodal AI tools, ChatGPT-4o and Claude Sonnet 3.7, were selected based on their ability to analyze images and generate textual descriptions. Other tools were excluded due to limited image support or insufficient prompt functionality. All AI-generated alt text was produced using the free versions of the selected tools during 2025.

Common image motifs on the web include people, landmarks, objects, and scenes [26, 37]. Based on this, the images used in the study were divided into three categories: objects, places, and people; with increasing levels of visual complexity both across and within categories. A total of 15 photographs were selected from Pixabay.com, all licensed for free use.

A standardized prompt was used when asking the AI tools to generate alt text for the images. Development of the prompt was an iterative process in which different formulations were tested across images of varying complexity to assess their performance. Initially, the AI was allowed to generate alt text without a character limit, which resulted in overly long and detailed descriptions. However, imposing a strict character limit of 150 caused the models to treat the maximum length as a target. A more flexible length guideline was therefore introduced, informed by recommendations from the [1, 17, 40] together with instructions encouraging the models to adapt alt-text length to suit each image.

Ultimately, the AI models were instructed to follow official guidelines for alt-text authoring without being given explicit criteria. This approach allowed the models to draw on their internalized accessibility knowledge (e.g., from the Web Accessibility Initiative) in combination with contextual understanding, without constraining them to a predefined style of alt text or limiting the sources from which accessibility principles could be drawn.

The final prompt was:

> *Skriv en alternativ bildtext för denna bild. Följ officiella riktlinjer för hur en alternativ bildtext bör skrivas. Undvik att skriva mer än 150 tecken eftersom det är gränsen för vad en del skärmläsare klarar. Du avgör själv hur lång alt-texten behöver vara men sträva efter att hålla alt-texten under 80–100 tecken, om möjligt.*
>
> (English translation: *Write an alternative image text for this image. Follow official guidelines for how an alternative image text should be written. Avoid exceeding 150 characters, as this is the limit supported by some screen readers. You determine the appropriate length of the alt text, but aim to keep it under 80–100 characters, if possible.)*

For each model, a new chat session was initiated for generating the AI-written alt text. Images were uploaded sequentially together with the standardized prompt, and this procedure was repeated until all 15 images had corresponding alt text. The same prompt was applied consistently across all images and both AI tools. Human-written alt text was produced by two of the study's authors, who first independently drafted descriptions for each image using established accessibility guidelines and national recommendations, and then collaboratively refined them. This approach was intended to represent non-expert but accessibility-aware practitioners, reflecting a realistic authoring context. The inclusion of both AI-generated and human-written alt text enabled direct comparison and methodological triangulation.

The questionnaire was administered using Microsoft Forms, selected for its compliance with the data protection standards GDPR, HIPAA, and FERPA-BAA [30]. It consisted of images paired with alt text and was divided into six sections. The first two sections provided information about the study, explained the concept of alt text, and included an age filter permitting only participants aged 18 or older to proceed; informed consent was obtained before continuing.

The third through fifth sections corresponded to the three image categories (objects, places, and people) arranged in increasing order of complexity. Objects were considered the least visually complex, followed by places, while images depicting people were considered to involve the greatest degree of interpretation and thus the highest complexity. Visual complexity also increased progressively within each category. Each category included five images and concluded with an optional open-ended reflection question.

All images were accompanied by three alt texts: two generated by the AI tools and one written by two of the study's authors. The presentation order of the three alt texts was randomized for each image to minimize order bias. Participants rated each alt text individually on a five-point Likert scale ranging from Very poor to Very good, in response to the question: "How well does the alt text convey correct and relevant information to someone who cannot see the image?" The questionnaire concluded with a closing thank-you message and an open-ended question inviting additional comments.

The questionnaire could be accessed via link or QR code and remained open for 27 days. Some participants were reached through announcements in various social media groups. It is difficult to estimate how many people saw these posts, but the Facebook groups and Reddit threads involved had several hundred thousand members in total. Current and former students and staff at the University’s Department of Technology had access to the survey via Messenger, Discord, Slack, and the university’s learning platform. Information and a QR code for the questionnaire were printed and distributed on campus and posted in a secondary school staff room. Teachers who requested a link received it by email. Links were also sent to family and friends.

In total, 79 respondents evaluated 45 alt texts, generating over 3,500 Likert-scale ratings. Sections 3–6 of the questionnaire received four to nine comments each, yielding 23 open-text comments in total. For the Likert-scale ratings, a within-subjects repeated-measures ANOVA was conducted, along with pairwise comparisons; the Friedman non-parametric test was additionally conducted as confirmation. Specific details are reported in the next section. Open-ended comments were analyzed to complement the quantitative results.

Results

This chapter presents the results of the questionnaire. The results are first presented as a whole and then organized by image category.

Differences across all images

To examine differences in alt-text quality, a 3 (Source: Claude, ChatGPT, Human) × 3 (Image Family: Things, Places, People) within-subjects repeated-measures ANOVA was conducted on participants’ mean ratings (N = 79). Because each participant evaluated all three sources across all image categories, both Source and Image Family were modeled as within-subject factors.

Mauchly’s test indicated that the assumption of sphericity was violated for both Source, $W = .474$, $\chi^2(2) = 57.53$, $p < .001$, and Image Family, $W = .790$, $\chi^2(2) = 18.15$, $p < .001$. Accordingly, Greenhouse–

Geisser corrections were applied to all repeated-measures effects. Corrected degrees of freedom are reported below.

A significant main effect of Source was observed, $F(1.31, 102.21) = 78.16$, $p < .001$, partial $\eta^2 = .50$. The partial eta squared value of .50 indicates that approximately 50% of the explainable variance in participants' ratings was attributable to the description source, representing a large effect. In practical terms, the source of the alt-text generation substantially influenced perceived quality. Throughout this section, effect sizes for the overall repeated-measures model are reported as partial η^2, which reflects the proportion of variance attributable to each factor after accounting for the other terms in the model; this differs from the generalized η^2 (η^2g) reported for the single-image comparisons in Table 4, and the two are not directly comparable in magnitude. AI-generated alt texts were on average rated as "Quite good" on the Likert scale, whereas human-written alt texts were rated closer to "Neither good nor bad."

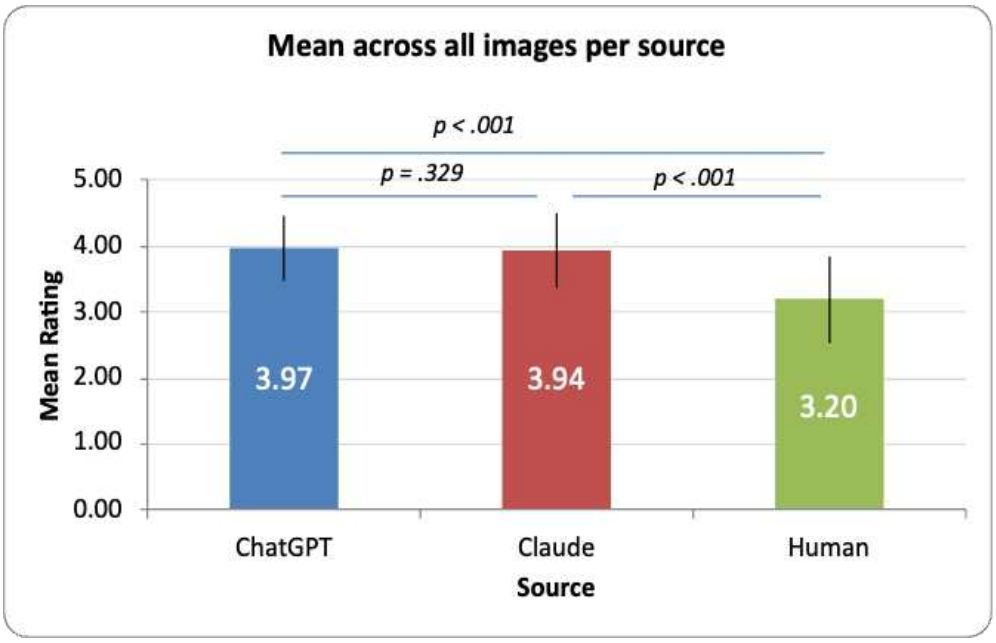

Fig. 1 Mean value comparison for all images for three sources

As depicted in Figure 1, pairwise comparisons (Holm-adjusted) showed that ChatGPT and Claude did not differ significantly, $t(78) = 0.98$, $p = .329$, with a negligible paired-samples effect size ($d \approx 0.11$). However, both AI systems significantly outperformed Human descriptions (ChatGPT vs. Human: $t(78) = 9.68$, $p < .001$, $d \approx 1.09$; Claude vs. Human: $t(78) = 8.88$, $p < .001$, $d \approx 1.00$). These effect sizes are large, indicating that the AI advantage is substantial and statistically significant.

A significant main effect of Image Family was also found, $F(1.65, 128.86) = 19.52$, $p < .001$, partial $\eta^2 = .20$. A partial eta squared of .20 represents a moderate-to-large effect, suggesting that image category meaningfully influenced ratings independent of source. The interaction between Source and Image Family was significant, $F(1.89, 147.16) = 24.36$, $p < .001$, partial $\eta^2 = .24$. This interaction indicates that differences between Claude, ChatGPT, and Human descriptions varied depending on whether the image depicted Things, Places, or People. Post hoc comparisons for image category Things and Places revealed significant results, however for People the results differed slightly.

As a non-parametric confirmation, Friedman tests corroborated the repeated-measures ANOVA results. The overall Source effect was significant, $\chi^2(2) = 80.52$, $p < .001$, and remained significant within each family. This confirms that the observed effects are valid despite deviations from normality.

Effect size magnitudes revealed a clear pattern. The use of ChatGPT, Claude or Human had a large and practically meaningful influence on ratings. AI-generated alt text was consistently preferred over human-generated descriptions, with large paired-samples effect sizes across most conditions.

The type of image family also revealed a meaningful influence. The results demonstrate that the relative strengths of ChatGPT and Claude depend on image semantics: ChatGPT showed a relative advantage for images containing things; Claude for scene-based images containing places, and both performed equally for images containing people. Thus, beyond statistical significance, the effect size indicates that AI superiority over human descriptions is substantial, whereas differences between AI sources are comparatively small and category-dependent.

Image category of Things

Within the Things category, post hoc paired comparisons revealed that ChatGPT received significantly higher ratings than Claude, $t(78) = 2.91$, $p < .05$, with a small-to-moderate effect size ($d \approx 0.33$). Both AI systems substantially outperformed Human descriptions: ChatGPT scored higher than Human, $t(78) = 10.67$, $p < .001$, with a large effect ($d \approx 1.20$), and Claude also exceeded Human performance, $t(78) = 8.70$, $p < .001$, with a similarly large effect ($d \approx 0.98$). Thus, although ChatGPT showed a modest advantage over Claude for object-centered images, the most pronounced differences were observed between AI-generated and human-generated descriptions, with large and practically meaningful effect sizes. Detailed values are shown in Figure 2.

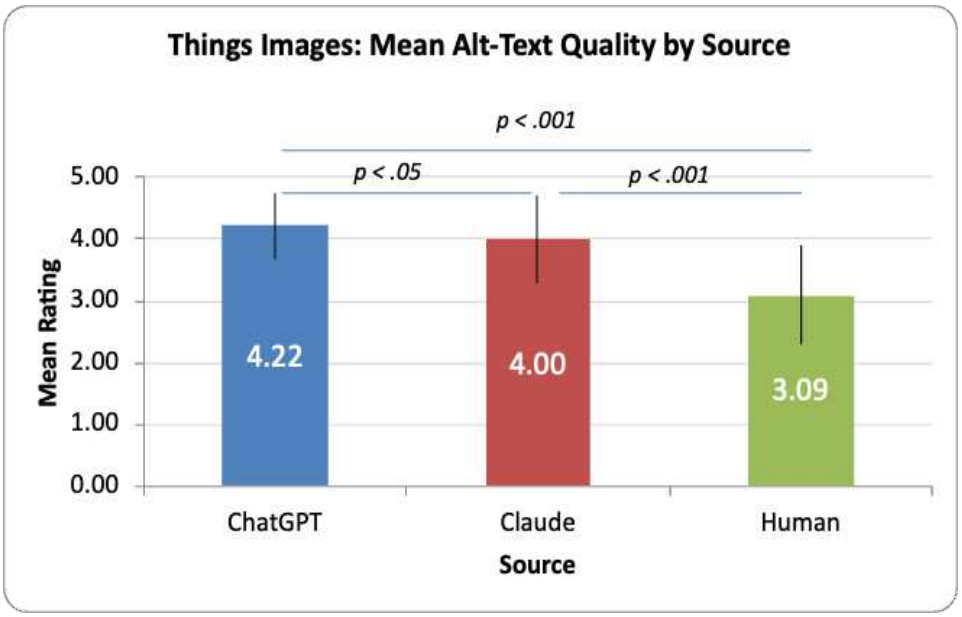

Fig. 2 Mean comparison within the image category of Things

Table 1 displays all images used in this category, along with the alt text provided by each Source.

Table 1 Images and alt texts in the Image Family Things

Image	ChatGPT-4o	Claude Sonnet 3.7	Human
 Image 1	En svartvit fotboll ligger på gräset i solljus. *(A black and white football lies on the grass in the sunlight.)*	En klassisk fotboll med svarta och vita hexagonmönster ligger på grönt gräs i solljus. *(A classic black-and-white soccer ball with a hexagonal pattern lies on green grass in sunlight.)*	Fotboll på gräsmatta. *(Football on a lawn.)*
 Image 2	En förlovningsring med en glittrande sten i en öppen grön ask med vit sammetsinredning. *(An engagement ring with a sparkling stone in an open green box with white velvet interior.)*	Diamantförlovnings-ring i vitguld placerad i en grön smyckesask med vit inredning. *(Diamond engagement ring in white gold placed in a green jewelry box with a white interior.)*	Förlovningsring i ask. Närbild. *(Engagement ring in box. Close-up.)*
 Image 3	Ett glas med rött te och en sked står på ett träbord utomhus, med suddig natur i bakgrunden. *(A glass of red tea and a spoon are placed on a wooden table outdoors, with blurred nature in the background.)*	Tidimglas-format teglas med mörkt te och tesked på en träyta, med suddig grön natur i bakgrunden. *(An hourglass-shaped tea glass filled with dark tea and a teaspoon on a wooden surface, with blurred green nature in the background.)*	Turkiskt te i glas på bord utomhus. *(Turkish tea in a glass on a table outdoors.)*
 Image 4	En person håller en smartphone med en aktiehandelsapp öppen, där en kursgraf för Dow Jones visas. *(A person holds a smartphone with a stock trading app open, displaying a Dow Jones price chart.)*	Hand håller mobiltelefon med aktiegraf för Dow Jones. En plånbok och grön kaffekopp syns i bakgrunden. *(Hand holding a mobile phone with a Dow Jones stock chart. A wallet and a green coffee cup are visible in the background.)*	Övervakning av aktiemarknad på en mobiltelefon. *(Monitoring the stock market on a mobile phone.)*



Image 5

En metallskulptur föreställande en affärsman med portfölj och mobiltelefon, skapad av muttrar och metallstänger. *(A metal sculpture depicting a businessman with a briefcase and mobile phone, made from nuts and metal rods.)*	Skulptur av affärsman gjord av metalldelar med portfölj och telefon på metallplatta. *(Sculpture of a businessman made of metal parts with briefcase and telephone on a metal plate.)*	Metallfigur föreställande en affärsman. Gjord av diverse muttrar, stålstänger och plåtskivor. *(Metal figure depicting a businessman . Made from various nuts, steel rods and metal plates.)*

The qualitative data in this category explores the respondents' opinions further. A few comments left in the Things section reflected on the balance between purpose and the level of detail in alt text descriptions. One respondent stated that “an image description should describe the purpose of the image, not explain everything in detail.” In contrast, another respondent gave failing marks for missing details. In this case, the less detailed alt texts were human-written. The respondents therefore expressed different preferences, with one favoring human-written descriptions and the other preferring AI-generated ones. A third respondent suggested that an ideal description would combine the concise structure of some alt texts with the greater level of detail found in others, effectively combining elements from both the human-written and AI-generated descriptions.

Several comments in the Things family also highlighted how the absence of contextual information caused uncertainty among some participants. One respondent noted that “without context, it is somewhat difficult to determine what is relevant to describe.” Another respondent commented on Image 3, stating that “the tea cup does not necessarily indicate the origin of the tea.” Both AI-generated alt texts for Image 3 identified the beverage only as tea, however, the human-written description also mentioned the tea’s origin (see Table 1). In addition, one respondent criticized certain descriptions for including assumptions that could not be verified from the image alone, such as identifying a ring as an engagement ring and the color of the ring box (Image 2) or that the stock chart was for Dow Jones (Image 4). Both AI-generated and human-written alt texts described the ring in Image 2 as an engagement ring. Without contextual information, this may be perceived as an assumption. Another respondent pointed out that the image quality made it difficult to confirm the claim in Image 4 that the stock chart represented the Dow Jones index, writing that they "could not see anywhere that it clearly says it is Dow Jones." Both AI-generated descriptions mentioned this detail.

Several respondents also reflected on how visual ability may influence the perception of alt-text descriptions. One respondent suggested that color might not always be as important to describe a shape, since shapes may be more easily recognized by people with visual impairments through prior sensory experience. In relation to this observation, it can be noted that the AI-generated alt texts mentioned color in at least three of the five images, whereas color was not mentioned in any of the human-written descriptions. Finally, two respondents simply commented that they found the descriptions to be good overall.

Image category of Places

Within the Places category, paired comparisons showed that Claude received significantly higher ratings than ChatGPT, $t(78) = -3.33$, $p < .001$, with a small-to-moderate effect size ($d \approx 0.38$). Claude also substantially outperformed Human descriptions, $t(78) = 10.04$, $p < .001$, with a large effect ($d \approx 1.13$), and ChatGPT similarly exceeded Human performance, $t(78) = 8.90$, $p < .001$, with a large effect ($d \approx 1.00$). Thus, while Claude demonstrated a modest advantage over ChatGPT for scene-based images, the most pronounced and practically meaningful differences were again observed between AI-generated and human-generated descriptions (Figure 3).

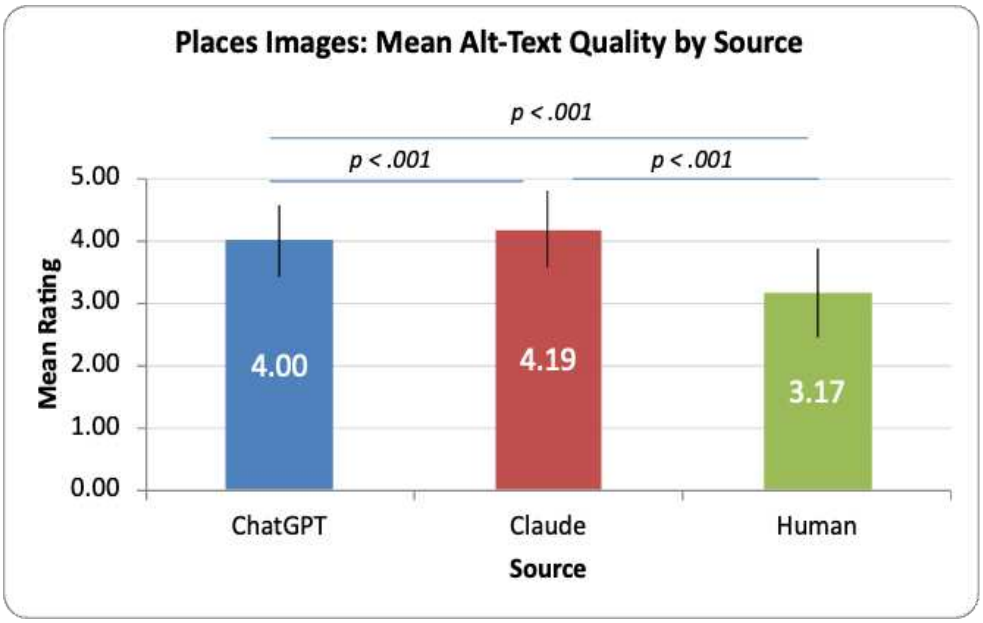

Fig. 3 Mean comparison within the image category of Places

Despite the quantitative results favoring AI-generated descriptions, open-ended feedback revealed more nuanced findings; the AI-generated descriptions received more criticism than the human-generated ones. One respondent questioned the content of several alt texts. One example was that the shop goods were described (by ChatGPT) in Image 9 (see Table 2) even though it was not possible to distinguish what goods they were. Similarly, in Image 6 and 10 (see Table 2) the identification of specific places such as Stonehenge and Santorini was also questioned, “How do we know it's Santorini? ... And there are more stone circles than Stonehenge”. They felt that these interpretations could not be confirmed visually.

Table 2 displays all images used in this category, along with the alt text provided by each Source.

Table 2 Images and alt texts in the Image Family Places

Image	ChatGPT-4o	Claude Sonnet 3.7	Human
 Image 6	Stonehenge, en förhistorisk stencirkel på en grön äng under en blå himmel med moln. *(Stonehenge, a prehistoric stone circle on a green field under a blue sky with clouds.)*	Det förhistoriska stenmonumentet Stonehenge på Salisburyslätten i England under en klar dag med en ensam vit moln på himlen. *(The prehistoric stone monument Stonehenge on Salisbury Plain in England on a clear day with an solitaire white cloud in the sky.)*	Stonehenge i öppet landskap. *(Stonehenge in open landscape.)*
 Image 7	En lång lagergång med höga hyllor fyllda med kartonger och varor, belyst av taklampor. *(A long warehouse aisle with high shelves filled with boxes and goods, illuminated by ceiling lights.)*	Lång gång i ett lagervaruhus med höga metallhyllor fyllda med kartonger och paket på båda sidor. *(Long corridor in a warehouse with tall metal shelves filled with boxes and packages on both sides.)*	Lång lagergång med höga fyllda hyllor. *(Long warehouse aisle with high, full shelves.)*
 Image 8	Ett modernt köpcentrum med rulltrappor, butiker och människor som går eller sitter vid kafébord. *(A modern shopping mall with escalators, stores, and people walking or sitting at café tables.)*	Köpcentrum med flera våningar, rulltrappor i mitten och besökare som rör sig genom de ljusa korridorerna. *(Shopping mall with several floors, escalators in the middle and visitors moving through the bright corridors.)*	Inne i ett modernt köpcentrum med flera våningar. *(Inside a modern shopping mall with multiple floors.)*
 Image 9	Liten italiensk butik med vin och lokala produkter. Korgar med pasta och delikatesser utanför. *(Small Italian shop with wine and local products. Baskets of pasta and delicacies outside.)*	Italiensk vinbutik 'Enoteca D'uu Scintu' med korgar fyllda med lokala produkter utanför entrén. *(Italian wine shop 'Enoteca D'uu Scintu' with baskets filled with local products outside the entrance.)*	Skyltfönster över liten vinbutik i Italien som säljer lokalt vin och produkter. *(Storefront of a small wine shop in Italy selling local wine and products.)*
 Image 10	Färgglada hus och vita byggnader med blå kupoler på en sluttning i Santorini, Grekland. *(Colourful houses and white buildings with blue domes on a hillside in Santorini, Greece.)*	Färgglad utsikt över Santorini med vita byggnader och blå kupoler som klättrar längs bergssluttningen under klar himmel. *(Colorful view of Santorini with white buildings and blue domes climbing along the mountain slope under clear sky.)*	Utsikt över Santorini, Grekland, där byggnaderna ligger tätt på en sluttning. *(View of Santorini, Greece, with buildings clustered along the cliffside.)*

Several people reflected on what is relevant to include in alt texts and how a person with visual impairment would perceive them. The same respondent as above questioned the relevance of mentioning locations in the alt text if the location is not clearly identifiable in the image. Another respondent questioned the relevance of describing colors, saying that "descriptions of colors are completely meaningless to someone with visual impairments." Both the AI-generated and human-written alt texts contained location information, whereas only the AI-generated texts repeatedly described colors. Another respondent observed a discrepancy between the information in the alt text and the content of the image. The respondent reflected that the alt text in isolation would have led to a different interpretation of Image 10 (see Table 2), "if I hadn't seen the pictures, I would have interpreted that all buildings, or a large majority of them, had blue domes." In this case, alt texts by both ChatGPT and Claude mentioned that the buildings had blue domes. Another participant found it "hard to know how the description is perceived once you've already seen the picture".

Finally, there was criticism regarding the language. One respondent found a misplaced modifier in "...and blue domes climbing along the mountain slope", which was found in Claude's generated alt text for image 10 (see Table 2). This points out that the participial phrase "climbing along the mountain slope" is attached directly to "blue domes," illogically attributing active physical movement to a static structure. Another respondent corrected a language error they identified in the Claude-generated alt text for Image 6 (see Table 2), "a solitary white cloud, is correct". For this image, another respondent noted that there were more clouds than mentioned in Claude's alt text. Only alt texts written by Claude were criticized for their language use. Overall, one person expressed dissatisfaction with all descriptions in this category and considered them to be poorly executed. Another participant stated that many details were missing.

Image category of People

Within the People category, ChatGPT and Claude did not differ significantly, $t(78) = 1.63$, $p = .107$, with a small effect size ($d \approx 0.18$). However, both AI systems outperformed Human descriptions. ChatGPT received significantly higher ratings than Human, $t(78) = 3.95$, $p < .001$, with a moderate effect ($d \approx 0.44$), and Claude also exceeded Human performance, $t(78) = 3.11$, $p < .05$, with a small-to-moderate effect ($d \approx 0.35$). Compared to the Things and Places categories, effect sizes in this condition were smaller overall, indicating that although AI systems maintained an advantage over human-generated descriptions, the magnitude of that advantage was more modest for human-centered images (Figure 4).

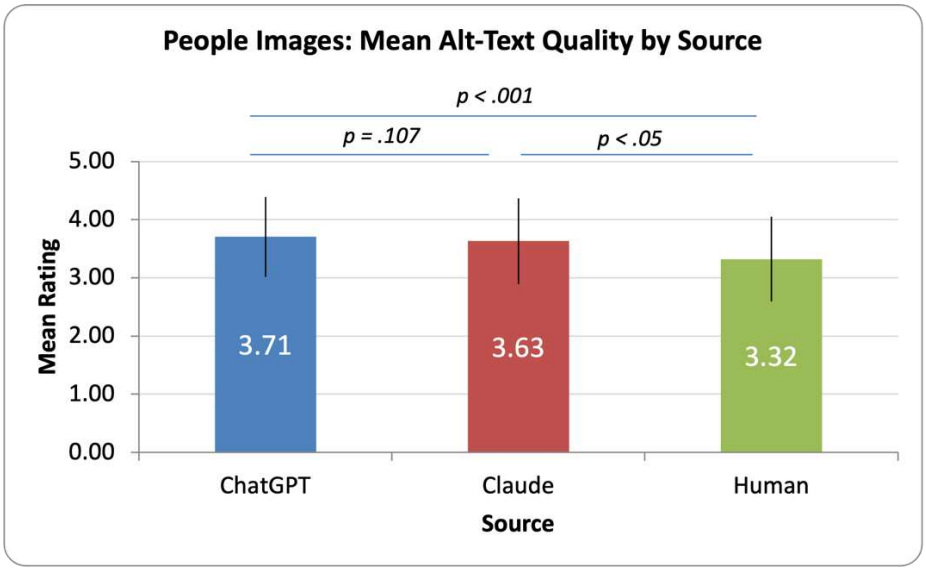

Fig. 4 Mean comparison within the image category of People

Table 3 displays all images used in this category, along with the alt text provided by each Source.

As in the previous sections of the questionnaire, respondents provided comments after evaluating the alt texts in this category. A recurring theme in these comments concerned contextual assumptions in the human-written alt texts. One respondent questioned the alt text for the image of King Charles III, in which it was suggested that he was leaving a board meeting (see Table 3), writing: “again made-up information. In the picture with the English king we do not know whether it is a corridor or whether they have just come out of a board meeting.” Similar concerns were raised for Image 14 (see Table 3), in which three boys were described as flying a kite, something that could not be clearly inferred from the image. Another respondent reflected that it can be difficult to infer underlying context solely from body posture and positioning, and therefore suggested that some alt texts should not be generated without manual verification. Finally, one comment pointed out a grammatical error in an AI-generated description of Image 13 (see Table 3).

Table 3 Images and alt texts in the Image Family People

Image	ChatGPT-4o	Claude Sonnet 3.7	Human
 Image 11	En person i keps och hoodie skapar musik vid en dator med två skärmar och ett MIDI-keyboard. *(A person wearing a cap and hoodie is creating music at a computer with two monitors and a MIDI keyboard.)*	Person med keps som arbetar på musikproduktion med MIDI-keyboard och datorskärmar i mörk studio. *(Person wearing a cap working on music production with a MIDI keyboard and computer screens in a dark studio.)*	Musikproducent som arbetar i sin studio. *(Music producer working in his studio.)*
 Image 12	Två män i kostym går bredvid varandra i en korridor, leende och i samtal, med folk i bakgrunden. *(Two men in suits walk side by side in a corridor, smiling and talking, with people in the background.)*	Två män i mörka kostymer skrattar tillsammans utanför ett styrelserum. Mannen till höger bär en röd vallmoblomma. *(Two men in dark suits laughing together outside a boardroom. The man on the right is wearing a red poppy flower.)*	Kung Charles den tredje och man i kostym på väg ut ur ett styrelsemöte. *(King Charles the Third and a man in suit leaving a board meeting.)*
 Image 13	Löpare i en hinderbana hoppar över ett vattenhinder på en blå löparbana i en stor stadion. *(Runners in an obstacle race jump over a water obstacle on a blue running track in a large stadium.)*	Löpare i hinderbana korsar vattenhinder under friidrottstävling i ett stadion med blå bana. *(Runners in obstacle course crossing water obstacle during athletics competition in a stadium with a blue track.)*	Pågående lopp i hinderlöpning där atleterna passerar hinder med vattengrav. *(Ongoing steeplechase race with athletes passing a water pit obstacle.)*
 Image 14	Tre barn står utomhus på en grönskande plats. Två av dem tittar koncentrerat framåt, medan det tredje barnet ler i bakgrunden. *(Three children are standing outdoors in a lush green setting. Two of them are looking intently ahead, while the third child is smiling in the background.)*	Tre barn utomhus tittar upp mot himlen, en skärmar ögonen med handen. Blomfält i bakgrunden. *(Three children outdoors look up at the sky, one shielding their eyes with their hand. Flower fields in the background.)*	Tre pojkar som flyger en drake. *(Three boys flying a kite.)*



Image 15

En kvinna sitter utomhus på en träveranda och ser eftertänksam ut, medan en man sitter inomhus i bakgrunden.
(A woman sits outdoors on a wooden veranda and looks pensive, while a man sits indoors in the background.)

Par sitter åtskilda i trästuga vid skymning, kvinnan på verandan och mannen inomhus, båda bär vita toppar och ljusa byxor.
(Couple sits apart in a wooden cottage at dusk, the woman on the veranda and the man indoors, both wearing white tops and light-coloured trousers.)

Ett par i en sommarstuga har blivit osams. Kvinnan sitter och tittar utåt i skymningen medan mannen hänger med huvudet i bakgrunden.
(A couple in a summer cottage have had a disagreement. The woman sits gazing out at dusk while the man in the background hangs his head.)

Individual image differences

Looking closely at each image, Table 4 shows the effect size for each image where most and least divergences occurred. At the individual-image level, the strongest divergences between sources were primarily driven by AI–Human contrasts rather than differences between AI tools. The largest effect (Image 8) reflects a substantial divergence between Claude and Human descriptions. In contrast, minimal divergence was observed for Image 13, where ratings across sources were nearly equivalent.

Unlike the overall effects reported above, effect sizes for the following per-image comparisons are reported as generalized η^2 ($\eta^2 g$) rather than partial η^2. Because each image was analyzed as an independent single-factor comparison rather than as part of the full factorial model, $\eta^2 g$ was used, as it is intended to remain comparable across differently structured analyses [32]. Readers should therefore avoid comparing $\eta^2 g$ values in Table 4 directly against the partial η^2 values reported earlier in this section.

When examined by image family, distinct performance patterns emerged. Within the Things category (Images 1–5), ChatGPT achieved the highest mean rating in four of the five images, whereas Claude led in one case (Image 1). Notably, all images in this category yielded large effect sizes, indicating strong and consistent divergence among sources. These findings suggest AI dominance for images containing things, with ChatGPT demonstrating the strongest relative advantage.

Across images, human-written alt texts were frequently shorter than their AI-generated counterparts. A general pattern was that very short alt texts, consisting of only three to eight words, tended to receive lower ratings compared to longer descriptions ranging from 10 to 21 words. Notably, the top eight images listed in Table 4, where the divergence was largest, all feature human-written alt texts that are noticeably shorter than those generated by AI tools. This suggests a possible relationship between description length and perceived quality.

Some qualitative comments addressed the balance between conciseness and descriptiveness. One respondent expressed appreciation for variation, stating that “sometimes short and concise is good, sometimes a more vivid description is beautiful.” At the same time, another respondent called for restraint in the level of detail, noting that “many descriptions were perhaps too detailed.”

Table 4 Effect size for each image and divergences

Image	Family	Winner	F	p	$\eta^2 g$	Mean Range	Magnitude
8	Places	Claude	69.41	< .001	.364	1.671	Large
2	Things	ChatGPT	51.61	< .001	.280	1.392	Large
1	Things	Claude	39.49	< .001	.240	1.291	Large
6	Places	Claude	36.55	< .001	.227	1.177	Large
4	Things	ChatGPT	31.15	< .001	.215	1.241	Large
3	Things	ChatGPT	30.57	< .001	.204	1.139	Large
14	People	ChatGPT	25.97	< .001	.178	1.114	Large
11	People	ChatGPT	26.68	< .001	.173	0.975	Large
9	Places	ChatGPT	27.21	< .001	.166	0.975	Large
5	Things	ChatGPT	25.90	< .001	.158	0.810	Large
7	Places	Claude	18.90	< .001	.137	0.861	Moderate
12	People	Human	14.97	< .001	.083	0.722	Moderate
15	People	ChatGPT	9.13	< .001	.062	0.734	Moderate
10	Places	Claude	8.33	< .001	.053	0.557	Small
13	People	Human	1.64	.196	.011	0.228	Small

Note. $\eta^2 g$ denotes generalized eta-squared, distinct from the partial η^2 reported for the overall repeated-measures ANOVA above.

In the Places category (Images 6–10), Claude exhibited its strongest relative performance, receiving the highest ratings in three images (6, 7, and 10), while ChatGPT led in one image (9). Image 8, which showed the largest overall divergence across all stimuli, strongly favored Claude. This pattern indicates that scene-based images may highlight Claude's strengths, particularly relative to both ChatGPT and Human descriptions.

In contrast, the People category (Images 11–15) displayed a more mixed pattern: this was the category in which human-written alt text performed best relative to its own results elsewhere, and correspondingly the category in which AI-generated alt text performed worst relative to its own results elsewhere. ChatGPT received the highest ratings in three images (11, 14, and 15), while Human descriptions led in two cases (12 and 13), indicating that human descriptions outperformed AI in two images within this category. Effect sizes here were generally smaller than in Things and Places, suggesting that performance differences among sources narrowed for images of people. Taken together, these results indicate that while AI systems consistently outperform human descriptions overall, the magnitude and consistency of that advantage depend meaningfully on image semantics.

In the section for additional comments at the end of the survey, several respondents provided reflections that were not linked to a specific image category, but concerned the alt texts as a whole. Three respondents felt that the suitability of the alt texts depends on the context in which the image is used. They were not provided any context so one respondent argued that it is therefore difficult to judge what is relevant or important in the descriptions. The other two respondents noted that the level of detail in an alt text may depend on the context in which the image is used. One of them also felt that some of the information seemed speculative or uncertain, possibly due to limited contextual information.

Several respondents reflected on the needs of the target audience, particularly individuals with visual impairments. A participant highlighted the difficulty of determining what information is relevant for someone who cannot see the image, while another often thought information to be irrelevant for BLV users. One respondent also suggested that a more effective approach would have been to read the descriptions first and then view the image.

Discussion

This chapter addresses the research questions, discusses the study, and analyses the main findings in relation to previous research.

Accuracy and Usefulness of AI-Generated Alt Text

The first research question in this study was: *To what extent do AI-generated alt texts convey accurate and relevant information compared to human-written alternatives, as perceived by sighted evaluators?*

The quantitative data show that both ChatGPT-4o and Claude Sonnet 3.7 were rated as better at generating alt texts that convey accurate and relevant information to someone who cannot see the image, compared to those written by humans. This suggests that AI models may have reached a level where they can compete with, or even outperform, human alt-text authors. These AI models were particularly effective at describing images of objects and places. However, it cannot be concluded that AI will always perform better, as the qualitative data revealed significant challenges. To better align AI-generated alt text with specific user needs, it could be beneficial to implement a human review process or to refine prompts with more restrictive instructions. In the following sections, the accuracy and usefulness of the alt texts are discussed separately in relation to previous research, established guidelines, and the collected qualitative data.

Accuracy

Generalisation

The empirical data in this study demonstrate that AI-generated alt texts were generally better than those written by humans. This differs from the findings reported by [26, 29], in which older AI models were used. Previous research by [6, 18, 39] indicates that newer AI models perform better than earlier generations, which may explain this difference in results. The results also revealed no significant difference in performance between ChatGPT and Claude, suggesting that many modern multimodal AI models with transformer-based image description systems perform similarly in alt-text generation tasks.

Language use

The qualitative data collected in this study suggest that there are noticeable language errors in texts written by Claude. Three respondents found language errors in three different alt texts written by Claude. Since several participants found something they considered significant enough to comment on, this indicates that the errors were substantial enough to disrupt the reading experience. Whilst ChatGPT was proficient in English, [33] has openly admitted that it may perform poorly in other languages. However, it appears that ChatGPT was better at Swedish than Claude was, and is preferable for tasks where correct grammar and linguistic ability are of greater importance.

Speculative Information

Respondents criticized certain content in the alt texts as being speculative or not fully verifiable based solely on the image; for example, the origin of the tea (Image 3) or the identification of locations (Images 6 and 10). If respondents had been provided with broader contextual information, such as an article, a headline, or the intended use of the image, it would likely have been easier for them to determine whether the alt text represented a reasonable interpretation or an incorrect assumption. In such cases, respondents could have more confidently determined whether the children in Image 14 were actually flying a kite, or what the glass in Image 3 contained.

However, one respondent criticized alt texts across all image categories for including uncertain or fabricated information. Both AI-generated and human-written alt texts received this type of criticism. In some of the examples raised, certain details could in fact be verified from the image, such as the color of the ring box in Image 2 (assuming no red–green color blindness) and that the structure in Image 6 could only be Stonehenge. This demonstrates that even accurate information was sometimes questioned. At the same time, some content in the alt text was reasonably questioned. For instance, ChatGPT's alt text for Image 9 stated: "Baskets with pasta and delicacies outside," even though the contents of the baskets could not be determined from the image (see Table 2). This indicates that the AI tool may have inferred relationships between identifiable elements in the image and how similar scenes are typically described; however, this does not necessarily mean the information is correct. This can be interpreted as an example of what Huang et al. [21] refer to as multimodal hallucinations, where LVLMs generate plausible-sounding content that lacks grounding in the visual input.

The fact that this respondent, and likely others, expressed skepticism toward more interpretative alt texts may reflect a general mistrust towards AI. Although the source of each alt text was not disclosed, some respondents may have assumed that speculative descriptions were AI-generated and therefore perceived them as less reliable. This skepticism may stem from prior experiences with AI systems providing incorrect information, or from an awareness that AI models can hallucinate. This aligns with recent findings by Cheng et al. [8], who demonstrate that AI hallucinations significantly undermine both cognitive and emotional trust. According to their model, even the perception of a hallucination can trigger a defensive shift in the user's role, causing them not to view AI as a dependable tool.

Usefulness

Purpose versus Detail

According to previous research, focusing on the purpose of the image rather than visual details is a characteristic of a good alt text [14]. However, this is not consistent with the findings from the quantitative part of this study. In the quantitative results, AI-generated alt texts consistently received higher ratings than human-written ones, despite often containing more visually oriented information. One possible explanation for this discrepancy is that the evaluations were conducted by sighted participants, and that the results therefore reflect how alt texts are perceived by individuals who can fully see the image, rather than by BLV users.

Inclusion of Color Descriptions

The qualitative data included several reflections suggesting that the needs of visually impaired users were not always taken into account. Several respondents expressed uncertainty regarding what constitutes relevant information for someone who cannot see the image. For example, one respondent argued that color is often irrelevant for BLV users. It is noteworthy that the AI-generated alt texts frequently included color descriptions, whereas color was not mentioned in any of the human-written alt texts. ChatGPT referred to color in six out of fifteen alt texts, while Claude did so in nine out of fifteen. This indicates a tendency for

AI tools to produce more descriptive or vivid image descriptions rather than focusing on functionally relevant information. A recommendation from [25] states that color should be excluded from an image description if it is not meaningful for its interpretation.

Length of alt text

A clear trend within the Objects image category was that respondents consistently preferred longer alt texts over shorter ones. The average scores reflected this, as the AI-generated texts, which were generally more comprehensive than the human-written ones, received higher ratings. Additionally, the eight images with the strongest divergence between sources (as shown in Table 4) all featured notably shorter human-written alt text. At the same time, the qualitative data show that several respondents expressed reservations about overly detailed descriptions. Despite this, over half of the respondents rated Claude's alt text, "A classic black-and-white soccer ball with a hexagonal pattern lies on green grass in sunlight", for Image 1 as "Very good." These results deviate from the length recommendations of [17, 1] suggesting that respondents' preference for longer, more detailed alt text reflects the perspective of sighted people rather than the needs of BLV users. Most respondents likely assessed how accurately and comprehensively the alt text describes the scene in the image, rather than how well it functions as an alt text. This may explain why AI-generated alt texts performed so much better than those written by humans in the Things category, which featured the least visually complex images.

Lack of Context and its Implications

According to established guidelines, image descriptions should not only describe the visual content of an image but also clarify its function and meaning within a specific context [25]. Based on this, we deliberately included information that could be perceived as interpretive assumptions in certain human-written alt texts. These formulations drew on human ability to interpret visual and contextual cues, with the aim of conveying the likely meaning of the image. For example, we described the children in Image 14 as flying a kite, even though the kite itself was not visible in the selected image. This deduction was based on other images in the same series where the kite was clearly visible, as well as the image filename. Similarly, we interpreted the individuals in Image 15 as a couple having a disagreement, assessing this interpretation as close to the likely intent of the image. However, these and other qualified judgments were not always appreciated, as reflected in respondents' qualitative comments. A limitation of this study design is therefore that respondents were not provided with the contextual information necessary to fully interpret some images.

The absence of image context in the survey, combined with AI's tendency to sometimes hallucinate [22], may have led some participants to question the reliability of the alt texts. Since the survey did not indicate the source of each description, both human-written and AI-generated alt texts were affected by this. This suggests that participants who were aware they were evaluating AI-generated alt text may have perceived subjective or non-verifiable information as particularly problematic.

Including contextual information for each image would have provided respondents with a basis for assessing the accuracy of the alt texts, and would have supported those who found the lack of context made it difficult to determine what was relevant to describe. A future iteration of the study should therefore consider including textual context to improve validity.

It is also important to adopt a critical perspective toward the collected data. One comment left in the questionnaire suggests that a strong reaction to one alt text may have influenced how critically the remaining two descriptions for the same image were evaluated. The human-written alt text for Image 3 was criticized for specifying that the tea in the glass was Turkish, with the justification that "the tea cup does not necessarily indicate the origin of the tea." Both AI-generated alt texts mentioned that the glass contained tea, and this was not questioned. The issue arose only when the origin was specified. Had the glass been

shaped differently, there would have been no clear indication it could not, for example, contain juice instead of tea. This illustrates that some assumptions were accepted while others were questioned.

The potential and challenges of AI-generated alt text

The study's second research question was: *What are the key opportunities and limitations of using current-generation multimodal LLMs for automated alt-text production?*

The results show that a sighted audience often perceived AI-generated alt text as more accurate and useful than human-written alternatives. This points to a clear potential for AI to help improve accessibility on the web. At the same time, several limitations were discovered. AI, for example, is not particularly good at appropriately adjusting the length of alt text. The following section discusses in greater detail both the potential and the challenges identified, drawing on previous research.

The Potential of AI

Previous research [26] showed that tools developed for automatic image captioning had potential, but that human-written alt text was still preferred and considered more accurate. Our study demonstrates the opposite, with current AI outperforming humans. These findings indicate that today's AI has the potential to streamline alt-text production, which in turn may increase the presence of alt text on the web, benefiting both content providers and end users.

Previous studies report that writing alt text is time-consuming and resource-intensive [13, 36]. This aligns with our own experience: the task became progressively more time-consuming and challenging as images increased in complexity. Within each image category, the final two images were perceived as particularly demanding, and images in the People category were consistently the most difficult to write alt text for. Study [36] confirms that complex images or graphics take longer to describe. This increasing time demand suggests alt-text creation is not a simple task, even for individuals with relevant training, and that by delegating this task to AI, organizations may save both time and financial resources.

At the same time, study [37] emphasize that automated solutions represent only one of several possible strategies for producing alt text. Their study identifies three main approaches: human-powered, fully automated, and hybrid. Human-in-the-loop is an example of a hybrid approach in which AI generates an initial version of the alt text, which a human then reviews, corrects, and refines. This suggests AI need not replace human labor entirely, but can instead function as a support tool. Similarly, [39] describe how AI-driven language and writing tools can help reduce authors' workload. A hybrid approach that could produce strong results today is to use AI for proofreading human-written alt text and suggesting improvements, though this is expected to be more effective for English-language texts than Swedish. This is consistent with emerging evidence that LLMs can support, rather than replace, accessibility workflows, for instance by assisting the evaluation of criteria that otherwise require manual review [28]. Building on [26], who evaluated earlier captioning tools, our findings suggest that current multimodal LLMs have reached a level where such support is practical for alt-text authoring in Swedish.

A study [29], however, found that AI-generated alt text negatively influenced the human writing process: final alt text was generally of poorer quality when authors based it on, or drew inspiration from, AI-generated text, compared to when they wrote it entirely on their own. Since that study was conducted, more advanced AI models have emerged. Newer models, including the multimodal transformer-based systems used in this study, have been shown to outperform earlier generations [6, 18, 39], suggesting that hybrid approaches today may yield better results.

The primary potential of AI lies in its ability to facilitate or automate alt-text production, making it easier to manage large volumes of images quickly and efficiently, with associated time and cost savings for organizations. It may also lead to more images online being accompanied by alt text, improving

accessibility overall. This study shows that AI performs sufficiently well in generating alt text for photographs and informative images to potentially take over this task from humans, and the rapid pace of AI development suggests current limitations are likely to diminish over time.

Challenges with AI

Despite this potential, several challenges remain in using AI to generate alt text. One of the biggest is ensuring that the correct information is included and presented usefully.

The AI tools used in this study tended to produce long, detailed descriptions regardless of image complexity. This is a significant challenge, as overly long alt text can both prevent screen readers from reading it in full and impair the user experience. Had a character limit not been included in the prompts given to ChatGPT and Claude, there is a risk the alt texts would have become even longer, as observed during earlier test runs without a limit. Finding the right balance, sufficiently informative yet brief and to the point, therefore remains a challenge when using AI tools.

For alt text to be useful, it must be tailored to the intended audience and to the specific context in which the image appears [25, 37]; however, the necessity of context largely depends on the nature of the image. Since AI cannot go beyond the data and instructions it is given, prompts for complex or context-dependent images should include background information regarding their purpose in order to achieve higher-quality alt text. Although today's AI can process detailed prompts that include the intended use or accompanying text, this requires an effort that runs counter to the very purpose of automation; AI would no longer generate alt text independently if a human had to tailor a prompt for each image. This challenge of crafting the right prompt could lead to a shift in job responsibilities, with the human role evolving from writing alt text to writing high-quality prompts for it.

Further challenges relate to the reliability and linguistic quality of AI. As noted earlier, AI can produce content that is untrue, and the occurrence of such hallucinations was confirmed in this study, meaning future AI-generated alt text risks containing inaccurate information. It was also noted that some AI-generated alt text, particularly those produced by Claude, contained linguistic errors remarked upon by respondents. Although these errors did not hinder comprehension, they can reduce readability and make the text appear less professional.

While AI has the potential to assume the role of an alt-text author, it can still make errors, some more severe than others. In this study, the errors produced by AI were relatively minor, including occasional hallucinations, linguistic inaccuracies, unnecessary color descriptions, and a tendency toward longer, more detailed alt text than required. However, instances of offensive or inappropriate content in AI-generated text have been reported elsewhere [24], emphasizing the importance of human review before such outputs are put into use. Given that both OpenAI [34] and Anthropic [3] have acknowledged that their models can make mistakes, these findings were not entirely unexpected.

Limitations and Future Research

A clear limitation of this study was that none of the respondents were confirmed to be BLV. A potential direction for future research would therefore be to include visually impaired users in the data collection process. Studies with visually impaired participants often rely on small, convenience-sampled cohorts, with interview studies typically involving around ten users [27]. A future study combining such interviews with the rating approach used here would allow triangulation across both groups. This would create two distinct data points for comparison: how sighted individuals perceive alt text in relation to the image, and how BLV users perceive alt text and its practical usefulness. This would make it possible to examine whether these groups value the same characteristics or prioritize different aspects, and to assess how well AI-generated alt text meets the actual needs of end users.

Another limitation was the absence of contextual information for the images, which made it difficult for respondents to assess whether alt text was speculative or contextually appropriate. Future research could examine how both AI and humans produce alt text when images are presented within their actual context, such as alongside a news article or caption; providing this context to both authors and respondents may enable more accurate evaluations of relevance, correctness, and how well the alt text supports the image's intended purpose.

Furthermore, future studies could focus on images in which meaning and context matter more than what is objectively visible, such as warning signs or images conveying emotion. This would provide deeper insight into AI's ability to handle subjective content and interpret an image's purpose rather than merely its visual details.

It would also be valuable to evaluate additional AI models as new tools and image-analysis features are launched. Since the AI tools used in this study are continuously updated, and earlier versions are not publicly available, exact replication of the study may be difficult in future, affecting reproducibility, but also creating opportunities for comparative studies tracking the development of AI tools over time.

Finally, future research could examine how different types of prompts influence the quality of generated alt text, helping determine whether it is more valuable for individuals to develop expertise in crafting high-quality prompts for AI, or to further refine their own alt-text writing ability.

Conclusion

In this paper, we have reported a mixed-methods study investigating the extent to which AI-generated alt text conveys accurate and relevant information compared to human-written alternatives, and the key opportunities and limitations of current multimodal LLMs for automated alt-text production.

The quantitative results indicate that both ChatGPT-4o and Claude Sonnet 3.7 perform well in generating alt text, with no significant difference between the two tools. AI-generated alt text was generally rated by sighted evaluators as more accurate and relevant than human-written alternatives, suggesting that current AI systems are capable of producing alt text that is, in many cases, perceived as highly effective in describing visual content.

However, the qualitative findings provide a more nuanced picture. While AI-generated alt text was often preferred for its detail and informativeness, respondents also raised concerns regarding its suitability for blind and low-vision (BLV) users: descriptions were at times considered overly detailed, included functionally irrelevant elements such as color, or contained information that could not be verified from the image alone. Because our evaluators were sighted, this preference for longer, more detailed text likely reflects a sighted perspective rather than the needs of screen-reader users; a distinction we consider important for interpreting the quantitative results. Claude's alt text, specifically, was also found to contain more Swedish-language errors than either ChatGPT's or the human-written alternatives, highlighting a limitation in multilingual reliability that varies by tool.

Taken together, the findings suggest that while current AI models have strong potential to support, and partially automate, the production of alt text, they still struggle with determining the appropriate level and type of information to include for a given image. This limitation affects the functional usefulness of AI-generated descriptions for accessibility purposes, even where they are well received by sighted audiences.

This study makes two main contributions. First, it provides updated empirical evidence, using modern, transformer-based multimodal models, that revises earlier findings suggesting AI-generated alt text trails human-written alt text in quality. Second, by focusing on Swedish-language alt text, it extends

this line of research beyond the English-language contexts that have dominated prior work, surfacing language-specific reliability issues that would not otherwise have been visible.

Despite the identified challenges, AI presents clear opportunities for improving the scalability of alt-text production. By reducing the time and resources required to generate descriptions, AI could contribute to a higher overall presence of alt text across the web, benefiting both content providers and end users. Given the limitations identified, however, a hybrid approach appears most appropriate at present, in which AI-generated alt text is reviewed, adjusted, and approved by a human before publication.

Future work should include BLV users directly in the evaluation process, both to assess real-world accessibility and usability outcomes and to determine whether their priorities for alt-text content align with those of the sighted evaluators in this study. Future efforts should also examine the effect of providing contextual information alongside images, and track how the reliability and linguistic performance of multimodal AI models evolve as new versions are released.

Acknowledgement

We thank all participants who contributed to this study. Their time and insights were essential to the completion of this research.

Funding

No funding was received for conducting this study.

Author contributions

L.F. and E.J.O equally contributed to the process of literature review, methodology, data collection, qualitative data analysis, Tables 1-3, and writing the main manuscript text. M.F. contributed with quantitative data analysis, Table 4, Figures 1-4, text revision and refinement, and supervision.

References

1. American Foundation for the Blind. (2024a). *Refreshable Braille Displays*. Assistive Technology Products. https://www.afb.org/node/16207/refreshable-braille-displays
2. Anthropic. (2025a, February 24). *Claude 3.7 Sonnet and Claude Code*. Retrieved June 26, 2025, from https://www.anthropic.com/news/claude-3-7-sonnet
3. Anthropic. (2025b, August 15). *Claude is providing incorrect or misleading responses. What's going on?* Anthropic Help Center. Retrieved August 16, 2025, from https://support.anthropic.com/en/articles/8525154-claude-is-providing-incorrect-or-misleading-responses-what-s-going-on
4. Balaji, V., Sonowal, G., Ramachandran, V. et al. Accessible Multilingual Audio based Image Alternative (ALT) description for the persons with visual impairment. Univ Access Inf Soc 25, 17 (2026). https://doi.org/10.1007/s10209-025-01294-4
5. Bigham, J. P., Kaminsky, R. S., Ladner, R. E., Danielsson, O. M., & Hempton, G. L. (2006, October 23). WebInSight: Making Web Images Accessible. *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility*, 181-188. https://doi.org/10.1145/1168987.1169018

6. Bucciarelli, D., Moratelli, N., Cornia, M., Baraldi, L., & Cucchiara, R. (2024, December 4). Personalizing Multimodal Large Language Models for Image Captioning: An Experimental Analysis. https://doi.org/10.48550/arXiv.2412.03665
7. Bureau of Internet Accessibility. (2023, October 26). *Be Careful When Using A.I. for Alternative Text*. Blog - Digital Accessibility. Retrieved February 26, 2025, from https://www.boia.org/blog/be-careful-when-using-ai-for-alternative-text
8. Cheng, Q., Dai, Y., Liu, X., & Peng, S. (2026). The trust crisis in artificial intelligence: AI hallucinations and human-AI collaboration. *Technology in Society, 86*, Article 103286. https://doi.org/10.1016/j.techsoc.2026.103286
9. Digg. (n.d.). *Beskriv med text allt innehåll som inte är text*. Webbriktlinjer. Retrieved February 4, 2025, from https://www.digg.se/webbriktlinjer/alla-webbriktlinjer/beskriv-med-text-allt-innehall-som-inte-ar-text
10. Digg. (2021, September 23). *De sex vanligaste bristerna när Digg granskar webbplatser*. Nyheter. Retrieved February 4, 2025, from https://www.digg.se/om-oss/nyheter/nyheter/2021-09-23-de-sex-vanligaste-bristerna-nar-digg-granskar-webbplatser
11. Digg. (2024, January 15). *Digital tillgänglighet*. Kunskap och stöd. Retrieved February 18, 2025, from https://www.digg.se/kunskap-och-stod/digital-tillganglighet
12. Diskrimineringsombudsmannen. (2025, January 8). *Funktionsnedsättning - en diskrimineringsgrund*. Diskrimineringsombudsmannen. https://www.do.se/diskriminerad/diskrimineringsgrunder/funktionsnedsattning-en-av-diskrimineringsgrunderna
13. Edwards, E. J., Gilbert, M., Blank, E., & Branham, S. M. (2023, July). How the Alt Text Gets Made: What Roles and Processes of Alt Text Creation Can Teach Us About Inclusive Imagery. *ACM Transactions on Accessible Computing*, *16*(2), 18:1-18:28. https://doi.org/10.1145/3587469
14. Eggert, E., & Abou-Zahra, S. (2019, June 17). *Informative Images*. W3C. https://www.w3.org/WAI/tutorials/images/informative/
15. European Telecommunications Standards Institute. (2021, March). *Accessibility requirements for ICT products and services* (V.3.2.1). ETSI Standards Database. https://www.etsi.org/deliver/etsi_en/301500_301599/301549/03.02.01_60/en_301549v030201p.pdf
16. European Union. (2007). Charter of Fundamental Rights of the European Union. *Official Journal of the European Union*, C 303, 1–16.
17. Evinced. (n.d.). *alt-text-is-too-long*. Retrieved May 27, 2025, from https://knowledge.evinced.com/system-validations/alt-text-is-too-long
18. Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2023, September). Generative AI. *Business & Information Systems Engineering (BISE)*, *66*(1), 111–126. SpringerLink. https://doi.org/10.1007/s12599-023-00834-7
19. Förbundet Finlands Svenska Synskadade & Näkövammaisten liitto. (2021, May 17). *Pressmeddelande inför Internationella tillgänglighetsdagen torsdagen 20.5.2021*. Utan bildbeskrivningar exkluderas personer med synnedsättning. Retrieved May 14, 2025, from https://fss.fi/aktuellt/utan-bildbeskrivningar-exkluderas-personer-med-synnedsattning
20. Gleason, C., Carrington, P., Cassidy, C., Morris, M. R., Kitani, K. M., & Bigham, J. P. (2019, May 13). “It’s almost like they’re trying to hide it": How User-Provided Image Descriptions Have Failed

to Make Twitter Accessible. *WWW '19: The World Wide Web Conference*, 549 - 559. https://doi.org/10.1145/3308558.3313605
21. Hollier, S. E. (2007, January). *The Disability Divide: A Study into the Impact of Computing and Internet-related Technologies on People who are Blind or Vision Impaired.* [Doktorsavhandling, Curtin University of Technology]. espace. https://espace.curtin.edu.au/handle/20.500.11937/214
22. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025, January). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, *43*(2), 1-55. http://dx.doi.org/10.1145/3703155
23. Jacobsen, D. I. (2017). *Hur genomför man undersökningar? introduktion till samhällsvetenskapliga metoder* (S. Andersson, Trans.; 2:1 ed.). Studentlitteratur AB.
24. Jones, D. (2021, September 4). Facebook Apologizes After Its AI Labels Black Men As 'Primates'. *NPR*. https://www.npr.org/2021/09/04/1034368231/facebook-apologizes-ai-labels-black-men-primates-racial-bias
25. Lejbro, U., & Bengtsson, E. (2023, October 19). *Riktlinjer för bildbeskrivning* (Issue MTM [2022/510]). Myndigheten för tillgängliga medier.
26. Leotta, M., Mori, F., & Ribaudo, M. (2022, August). Evaluating the effectiveness of automatic image captioning for web accessibility. *Universal Access in the Information Society*, *22*(4), 1293–1313. SpringerLink. https://doi.org/10.1007/s10209-022-00906-7
27. Liljeström, L., Bern, C. & Ferati, M. Voicing barriers: accessibility for visually impaired voiceover users on audio chat platforms. Univ Access Inf Soc 24, 3067–3083 (2025). https://doi.org/10.1007/s10209-024-01096-0
28. 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
29. Mack, K., Cutrell, E., Lee, B., & Morris, M. R. (2021, October). Designing Tools for High-Quality Alt Text Authoring. *In The 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, 14. ACM Digital Library. https://doi.org/10.1145/3441852.3471207
30. Microsoft. (2025). *Säkerhet och sekretess i Microsoft Forms*. Support. Retrieved March 13, 2025, from https://support.microsoft.com/sv-se/topic/s%C3%A4kerhet-och-sekretess-i-microsoft-forms-7e57f9ba-4aeb-4b1b-9e21-b75318532cd9
31. Myndigheten för delaktighet. (2024, September 24). *EU:s tillgänglighetsdirektiv*. EU:s tillgänglighetsdirektiv. https://www.mfd.se/utgangspunkter/tillganglighet/eus-tillganglighetsdirektiv/
32. Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods, 8*(4), 434–447. https://doi.org/10.1037/1082-989X.8.4.434
33. OpenAI. (2023, September 25). *ChatGPT can now see, hear, and speak*. OpenAI. Retrieved February, 2025, from http://openai.com/index/chatgpt-can-now-see-hear-and-speak/
34. OpenAI. (2025, August 16). *Does ChatGPT tell the truth?* OpenAI Help Center. Retrieved August 16, 2025, from https://help.openai.com/en/articles/8313428-does-chatgpt-tell-the-truth
35. Ribich, E. (2025, March 18). ALT-texter – att beskriva bilder i e-böcker. *Publit*. https://blog.publit.com/sv/alt-texter-att-beskriva-bilder-i-e-bocker/

36. Richard, B. (2023). Key Issues Affecting the Inclusion of Alt Text in Scholarly PDF Publications. *Logos*, *34*(1), 44-60. https://doi.org/10.1163/18784712-03104058
37. Stangl, A., Morris, M. R., & Gurari, D. (2020, April 23). "Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions. *CHI Conference on Human Factors in Computing Systems*, *CHI '20*, 1-13. ACM Digital Library. https://doi.org/10.1145/3313831.3376404
38. Synskadades Riksförbund. (2024, May 30). *Skriv alttexter och dölj dekorativa bilder för skärmläsare*. Retrieved August 29, 2025, from https://www.srf.nu/redaktorsmanual/grunderna-i-umbraco/skriv-alttexter-och-dolj-dekorativa-bilder-for-skarmlasare/
39. Teubner, T., Flath, C. M., Weinhardt, C., van der Aalst, W., & Hinz, O. (2023, March 13). Welcome to the Era of ChatGPT et al. *Business & Information Systems Engineering*, *65*(2), 95-101. https://doi.org/10.1007/s12599-023-00795-x
40. University of Maine System. (2025). *Guidance for Alt Text for Images on the University of Maine System Website*. Alternative (Alt) Text – Accessibility Guide. Retrieved February 27, 2025, from https://www.maine.edu/content-management/accessibility/images/alt-text/
41. W3C. (2025, May 6). *Web Content Accessibility Guidelines (WCAG) 2.1*. W3C. https://www.w3.org/TR/WCAG21/
42. WebAIM. (2021, October 19). *Alternative Text*. https://webaim.org/techniques/alttext/
43. WebAIM. (2026, March 30). *The WebAIM Million.* https://webaim.org/projects/million/