

# SmartCaption AI - Enhancing Web Accessibility with Context-Aware Image Descriptions Using Large Language Models

Gia Ky Huynh

*School of Computer Science and Tech.  
Algoma University  
Brampton, ON, Canada  
ghuynh@algonau.ca*

Wenjun Lin\*

*Digital Healthcare Innovation Lab  
School of Computer Science and Tech.  
Algoma University  
Brampton, ON, Canada  
randy.lin@algonau.ca*

**Abstract**—The Internet provides vast amounts of information, services, and products. However, blind individuals and those with severe vision impairments face significant challenges in navigating web content, especially with understanding images. This paper introduces SmartCaption AI, an innovative solution that leverages Large Language Models (LLM) to generate descriptive text for images on web pages. By summarizing the content of a web page, SmartCaption AI provides relevant context for the LLM to produce accurate and meaningful image descriptions. These descriptions are seamlessly integrated into the web page's structure, allowing text-to-speech software to read them aloud to visually impaired users.

SmartCaption AI offers several key contributions to web accessibility. It ensures the generated descriptions are contextually relevant, enhances the browsing experience by integrating real-time descriptions, and provides a universally accessible solution through a Chrome extension. This approach addresses the critical issue of missing or inadequate alternative text for images, thereby bridging the digital divide between sighted and visually impaired individuals.

The results of our experiment demonstrated the effectiveness of SmartCaption AI, with an average score of 8.3/10, significantly outperforming state-of-art solutions: ImageToText (1.7/10) and AI-MCS (3.6/10). The source code of the tool is available on GitHub.

**Index Terms**—image caption; Large Language Model; multi-agent; alternative text; alt text; vision impairment;

## I. INTRODUCTION

The Internet offers people better opportunities each day as they go about their lives to acquire large amounts of information, services, and products. However, for blind individuals, who numbered around 43.3 million globally in 2020 and are predicted to reach 61 million by 2050, navigating the web presents significant challenges. This growing demographic, which also includes 295 million people with moderate to severe vision impairment (projected to increase to 474 million by 2050) [1], often struggles with understanding and interpreting images on websites.

One of the most pressing issues for visually impaired users is the inability to perceive visual content on web pages.

Images, which often convey crucial information or context, remain inaccessible without text alternatives [2]. According to the report 2024 from WebAIM Million, there was 54.5% of missing alternative text for images of homepages [3]. In addition, 14.6% of images with alternative text had questionable or repetitive alternative text [3]. These limitations not only hinder the user experience but also restrict access to potentially vital information, creating a digital divide between sighted and visually impaired individuals.

To address this challenge, we propose SmartCaption AI - a unique and innovative solution that leverages Large Language Models (LLM) to generate descriptive text for images on web pages. Our approach summarizes the web pages's content and uses the summarization as the relevant context for LLM to analyze and describe images, producing a result that is more accurate and meaningful when compared to existing solutions. These descriptions are then seamlessly integrated into the web page's structure, allowing text-to-speech software to read them aloud to visually impaired users.

Our solution contributes to web accessibility in several ways. First, by summarizing the web page content, the descriptions generated are more relevant and contextually appropriate compared to manually describing images without context. Second, the image descriptions are generated and integrated into the web page in real-time as users navigate, enhancing the browsing experience. Third, SmartCaption AI is delivered as a Chrome extension, supporting any device with a Chrome browser, and making it widely accessible. Our code is publicly available at GitHub <sup>1</sup>.

The rest of this paper is organized as follows. Section II describes the literature review. Section III describes the methodology. Section IV presents experimental results. The discussion is provided in Section V. Finally, Section VI concludes the paper with possible directions for future works.

## II. LITERATURE REVIEW

A popular topic in accessibility studies is image captioning. Our work is related to prior efforts in automatically generating

\*Corresponding Author

<sup>1</sup><https://github.com/hgky95/smart-caption-ai>

captions for people with vision impairments. Alternative text is essential for these individuals, but its absence poses a significant barrier [2].

Various automated approaches have been used to address this challenge. For example, Ganesan, Jothi, et al. [4] proposed an intelligent reader system using deep learning like Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. The CNN is implemented for detecting features from the images and the LSTM network is used as a captioning tool to describe the detected text from images. These captions are then converted into voice messages using a Text-To-Speech API. Image captioning aims to generate machine-generated descriptions of image content, but traditional methods often replicate frequent phrases, missing unique image aspects. Liu, Xihui, et al. [5] proposed an enhanced framework combining an image captioning module with a self-retrieval module, where a CNN extracts image features and an LSTM decodes them. The self-retrieval module evaluates caption quality by assessing the similarity between captions and images versus distractor images, enhancing performance through text-to-image retrieval error, and leveraging unlabeled images without additional annotations.

LLMs have proven effective in improving visual language tasks through collaboration with other specialized models. Zhu, Deyao, et al. [6] proposed a novel automatic-questioning method where an LLM-based agent asks BLIP-2, a vision question-answering model, and a series of questions about images. This iterative process gathers new visual information, enabling this method to create more detailed image descriptions. Their results demonstrate that the captions are significantly more informative, receiving three times as many votes from human evaluators.

The above approaches have been built as proof of concept and are not yet user-ready. To make it more approachable for the user, Jeong, Hyeonhak, et al. [7] proposed a web application to collect missing alt text and employ GRIT - a transformer to suggest image descriptions based on the dataset to help human alt text authors improve a web page's accessibility. Microsoft Research unveiled a browser extension that searches the DOM of the active webpage for image tags and background images, sending them to the server to retrieve captions. Once a caption is found for an image, it is added to the corresponding image element [8]. To increase accessibility on social media platforms for people with vision impairment, Gleason, Cole, et al. [9] developed a browser extension for finding or generating new alternative text through a process that includes six different methods, returning a result early if one is successful. The browser extension then dynamically inserts it into the alt text tag for the image. AI-Microsoft Cognitive Service (AI-MCS) <sup>2</sup> provides the WebUI where users can upload the image directly and then it will generate the caption for the image. ImageToText <sup>3</sup>, a chrome extension

that provides the UI where user can add their image and then the tool will generate the caption automatically.

Despite the power of deep learning, these approaches face challenges like object hallucination, contextual understanding, and referring expressions [10]. For instance, in Fig. 1, the ImageToText generates the caption 'a woman and child are playing in the snow' which inaccurately describes the image. The scene depicts a chemical foam, not snow, and the individuals are performing rituals, not playing.

Instead of pursuing a new paradigm in Deep Learning as the above approaches, we choose to leverage the power of LLM to provide relevant context, from which LLM can create more relevant and accurate captions. The details of the approach will be discussed in the following section.



Fig. 1. Hindu devotees perform rituals in Yamuna river. Source: Adapted from [11]

### III. METHODOLOGY

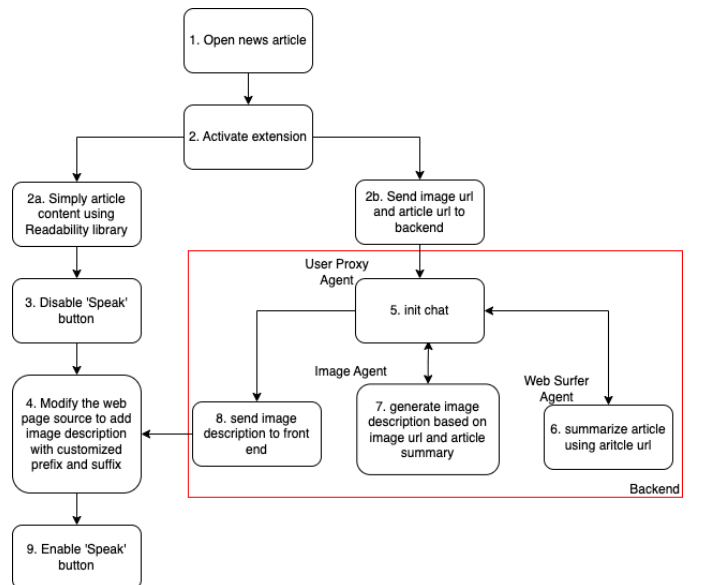


Fig. 2. SmartCaption AI workflow

<sup>2</sup><https://portal.vision.cognitive.azure.com/demo/image-captioning>

<sup>3</sup><https://chromewebstore.google.com/detail/image-to-text-using-ai/mgffgcnoocmekdiplebhipfnhijap?hl=en>

LLM has been found very promising in aiding people with disabilities [12]–[15]. However, image captioning using LLM often lacks contextual information, resulting in descriptions that may not adequately represent relationships between visual components [16]. To address this challenge, SmartCaption AI was developed to generate image captions with related contextual information while simultaneously simplifying webpage content.

The Fig 2 describes a workflow that simplifies news articles and generates LLM-powered captions for images within the article. Upon activation, it processes the article content, summarizes it, creates image captions based on the summary, and then displays these captions beneath the images and in their alternative text, while also enabling a text-to-speech feature for the entire content. Whenever a user opens a news article (step 1) and clicks the Chrome extension icon (step 2), the application simplifies the article content using the Readability library (step 2a) and disables the ‘Speak’ button (step 3). Simultaneously, it sends the image URLs and the article URL to the server (step 2b).

To enhance LLM capabilities, we use multi-agents, allowing LLM Agents to work together for better results. The User Proxy Agent controls the backend workflow and, upon receiving the information (step 5), engages the Web Surfer Agent to summarize the article content (step 6). This summary and the image URLs are then sent to the Image Agent, which generates image captions based on the summary (step 7). Once the captions are generated, the backend server responds to the User Interface (UI) (step 8).

Upon receiving the image captions, the UI appends a customized prefix and suffix to indicate that AI-generated the captions and displays them under the images and also adds to the alternative text of the image (step 4). Finally, the ‘Speak’ button is enabled to activate the Text-To-Speech (TTS) functionality (step 9). The ‘Speak’ button is initially disabled to ensure all images are captioned before TTS activation.

There are existing studies that use LLM-based solutions to generate captions from images, as mentioned in the literature review section. However, our work uniquely enhances this process by having the LLM-based agent first summarize the article and then use this context in the Image Agent to describe images. This approach addresses the critical issue of missing contextual information in traditional image-to-text conversions, resulting in more relevant and accurate image captions. This significantly benefits visually impaired users by providing accurate descriptions and more relevant information about images. The Experiment section will detail this process further.

In addition, user privacy is prioritized by not collecting or storing any user information. To avoid potential intellectual property concerns, it is important to note that this tool is designed as an assistive technology and is not distributed for commercial purposes. Additionally, after simplifying the article and appending the image’s caption, the content continues to be displayed on the original website.

## IV. EXPERIMENT

### A. Experiment Setup

We implemented SmartCaption AI as a browser extension, utilizing a tech stack comprising HTML, CSS, and JavaScript on the frontend, and Python, the Autogen library <sup>4</sup>, and the Flask framework on the backend. For frontend article simplification, we employed the Readability.js library <sup>5</sup>. On the backend, we leveraged the Autogen library for multi-agent tasks, including article summarization and image-to-text conversion using ‘GPT-4o’, a multimodal flagship model from OpenAI <sup>6</sup>. Testing was conducted on a Chrome browser version 126.0.6478.127, running macOS Monterey with 8 GB of memory and an Intel Core i5 2.3 GHz processor. To evaluate the performance of the tool, two existing tools from the Related Work section were selected for comparison, AI-MCS and ImageToText.

### B. Image Captioning Qualification

We evaluated our tool by comparing its image captions to those from existing research and tools. Using 10 images from various sources such as Aljazeera <sup>7</sup>, Huffpost <sup>8</sup>, Spiegel <sup>9</sup>, and BusinessInsider <sup>10</sup>. Evaluation of image descriptions was conducted through feedback collected from three male participants (P1, P2, and P3), aged 25-40, who had no prior knowledge of the news articles or images under assessment. Before exposure to the captions, participants were presented with a summary of the relevant article. The evaluators were then tasked with scoring each image description on a scale of 1 to 10. A score of 1 means it is a poor description; the image’s description is not relevant to the article, or objects are described incorrectly, such as describing a woman instead of a man. A score of 10 indicates a highly accurate description; the image’s description is relevant to the article, and objects are described correctly. The first four results are highlighted in Table I, with the remaining results provided in the appendix.

For Image 1, AI-MCS described it as ‘a man carrying a child with Kasubi Tombs in the background.’ There was a mistaken hut that was described as the Kasubi Tomb. This could erase an essential part of African culture and identity, as huts provide shelter and protection for specific communities [19]. ImageToText described it as ‘a man standing next to a yellow and white surfboard’ but the person in the image is clearly a woman, not a man, and she is walking, not standing, and there is no yellow and white surfboard. In contrast, our tool accurately describes the person as a woman carrying a child and walking past a hut. In addition, our tool conveys to the reader knowledge about the daily struggles of humans in the Central African Republic, such as the lack of water.

<sup>4</sup><https://microsoft.github.io/autogen>

<sup>5</sup><https://github.com/mozilla/readability>

<sup>6</sup><https://openai.com/>





<sup>7</sup><https://www.aljazeera.com/>

<sup>8</sup><https://www.huffpost.com/>

<sup>9</sup><https://www.spiegel.de/>

<sup>10</sup><https://www.businessinsider.com/>

TABLE I  
CAPTION FROM AI-MCS , IMAGE TO TEXT, AND SMARTCAPTION AI, ALONG WITH THE ACCURACY SCORES

Image	AI MCS	ImageToText	SmartCaption AI (our tool)	Score
Image-1. Source: Adapted from [17] 	man carrying a child with Kasubi Tombs in the background	a man standing next to a yellow and white surf-board	The image shows a woman carrying a child on her back and holding a yellow jerry can, walking past a traditional thatched hut. This scene reflects the daily struggles of displaced individuals in the Central African Republic, highlighting the lack of basic necessities such as water and the burden of care placed on women. The traditional hut and rural setting underscore the impact of the ongoing humanitarian crisis on local communities, emphasizing their resilience and the harsh conditions they face.	1.AI-MCS: Avg. score: 3 2.ImageToText: Avg. score: 1 3.SmartCaptionAI: Avg. score: 7.33
Image-2. Source: Adapted from [17] 	A man holding a bamboo object	a man with a black beard is holding a piece of food	The image depicts a person with their arm raised to their face, adorned with a makeshift arm guard made from sticks and fabric. The background shows traditional thatched huts, suggesting a rural or village setting. This visual captures the resilience and resourcefulness of individuals in the Central African Republic amidst the ongoing humanitarian crisis, highlighting their efforts to protect themselves in a context of severe insecurity and violence.	1.AI-MCS: Avg. score: 1.67 2.ImageToText: Avg. score: 1 3.SmartCaptionAI: Avg. score: 5.67
Image-3. Source: Adapted from [11] 	a person and a child in the snow	a woman and child are playing in the snow	The image depicts Hindu devotees standing in toxic foam in the Yamuna River during the Chhath Puja festival in New Delhi. Despite the significant health risks posed by the severe pollution of the river, the devotees are seen performing ritualistic offerings as part of their religious observance. The thick layer of toxic foam highlights the environmental challenges faced by the Yamuna River, exacerbated by industrial and domestic pollution.	1.AI-MCS: Avg. score: 2 2.ImageToText: Avg. score: 1 3.SmartCaptionAI: Avg. score: 8.33
Image-4. Source: Adapted from [18] 	a person and a child riding on a vehicle with a child on the back	a man riding a motorcycle on a dirt road	The image depicts a woman and a child seated on a cart loaded with various items, including blankets and bags, in a dry, arid landscape. This scene likely represents the rural communities in Senegal involved in the Great Green Wall initiative, which aims to combat desertification and enhance food security through sustainable agricultural practices like the Tolou Keur gardens. The presence of donkeys and other people in the background underscores the traditional and community-based nature of this effort.	1.AI-MCS: Avg. score: 3.67 2.ImageToText: Avg. score: 1 3.SmartCaptionAI: Avg. score: 10

The average score of AI-MCS is 3, ImageToText is 1 and SmartCaption AI is 7.33.

In Image 2, AI-MCS describes it as 'A man holding a bamboo object' while ImageToText generates the description 'a man with a black beard is holding a piece of food'. However, the image does not clearly show a black beard, nor is the person holding any food. This indicates that existing tools are experiencing object hallucination. In contrast, SmartCaption AI correctly describes the person's action, noting their arm raised to their face, and accurately depicts the background with thatched huts. The average scores for this image are 1.67 for AI-MCS, 1 for ImageToText, and 5.67 for SmartCaption AI.

For Image 3, AI-MCS generates the description 'a person and a child in the snow' while ImageToText describes it as 'a woman and child are playing in the snow'. Both incorrectly identify chemical foam as snow, leading to a misinterpretation of the scene. SmartCaption AI, however, accurately describes the substance and provides additional context by indicating that the individuals are performing a ritual in the river. This contextual information enhances users understanding of the image's meaning. The average scores for this image are 2 for AI-MCS, 1 for ImageToText, and 8.33 for SmartCaption AI.

In Image 4, AI-MCS generates the description 'a person and a child riding on a vehicle with a child on the back', while ImageToText produces 'a man riding a motorcycle on a dirt road'. While these captions are reasonable and demonstrate a decent performance by the models, they lack certain nuances present in the human-generated caption. The human-generated caption specifically describes the vehicle as a cart. Although 'cart' and 'vehicle' are related terms, their meanings differ significantly. A vehicle is defined as a piece of 'mechanized equipment' that may convey an inaccurate impression when mistaking a cart for a vehicle [19]. In contrast, SmartCaption AI accurately describes the conveyance as a cart, potentially leading to a more precise understanding of the image. The average scores for this image are 3.67 for AI-MCS, 1 for ImageToText, and 10 for SmartCaption AI.

We calculated the average scores for each tool across the ten images: AI-MCS (3.63), ImageToText (1.67), and SmartCaption AI (8.29). Based on these average scores and the qualitative descriptions provided in the table, SmartCaption AI consistently outperforms existing image captioning tools like AI-MCS and Image-To-Text across multiple test images. It demonstrates superior accuracy in identifying subjects, actions, and contextual elements, avoiding common pitfalls such as object hallucination and contextual information. SmartCaption AI's ability to provide relevant contextual information, such as cultural insights or environmental details, contributes to its higher average scores compared to other tools.

### C. Time Processing

Table II describes the processing times of AI-MCS, ImageToText, and SmartCaption AI on the same ten images. The processing time is captured the first time the user requests the article. Generally, the average time to process an image

with SmartCaption AI is 11.11 seconds, which is quite long compared to the other tools. However, the processing time becomes competitive on subsequent loads by enabling the caching feature. For example, on the second attempt to process Image 1, SmartCaption AI only needs 2.54 seconds.

TABLE II  
DESCRIBE THE PROCESSING TIMES OF AI-MCS, IMAGETOTEXT AND SMARTCAPTION AI

Image	AI-MCS (second)	ImageToText (second)	SmartCaption AI (sec- ond)
Image-1	1.26	3	11.34 Summarization: 5.89 Convert Image to text: 5.45
Image-2	1.04	4	12.75 Summarization: 6.53 Convert Image to text: 6.22
Image-3	1.08	3	10.1 Summarization: 4.74 Convert Image to text: 5.36
Image-4	1.2	3	9.38 Summarization: 4.61 Convert Image to text: 4.74
Image-5	1.25	3	9.55 Summarization: 5.13 Convert Image to text: 4.42
Image-6	1.07	3	7.43 Summarization: 3.78 Convert Image to text: 3.65
Image-7	1.07	3	12.21 Summarization: 5.59 Convert Image to text: 6.62
Image-8	1.41	3	11.69 Summarization: 4.71 Convert Image to text: 6.98
Image-9	1	3	19.36 Summarization: 12.47 Convert Image to text: 6.89
Image-10	1.01	4	14.22 Summarization: 7.95 Convert Image to text: 6.27

## V. DISCUSSION

The evaluation of SmartCaption AI against existing tools, AI-MCS and ImageToText, demonstrates significant improvements in accuracy and contextual relevance, it achieves an average score of 8.29 compared to 3.63 for AI-MCS and 1.67 for ImageToText. The tool's capacity to provide relevant contextual information, including political aspects, cultural insights, and environmental details, contributes substantially to its higher average scores. However, the initial processing time of SmartCaption AI (11.34 seconds on average) is notably longer than that of other tools, which could be a potential limitation in real-time applications. Nevertheless, the implementation of a caching feature significantly reduces subsequent processing times, as evidenced by the reduction to 2.54 seconds for Image 1 on the second attempt.

Besides, two limitations have been identified in the tool, 1) Using the LLM to summarize article content provides context for generating image captions but can be time-consuming. 2) Readability.js is employed to extract text and images, but it sometimes extracts incorrectly due to the iframe structure. Addressing these issues will be the focus of future work.



## VI. CONCLUSION AND FUTURE WORKS

In this paper, we presented SmartCaption AI, an innovative solution that leverages Large Language Models (LLMs) to generate accurate and contextually relevant image captions, enhancing web accessibility for visually impaired users. The tool summarizes web page content and uses it as context for converting images into text. SmartCaption AI produces more accurate and meaningful descriptions compared to existing solutions. Our experiment demonstrated an average score of 8.3 out of 10, significantly outperforming AI-MCS (3.6/10) and ImageToText (1.7/10).

Future work should focus on optimizing processing times for initial loads and improving text and image extraction accuracy from complex web structures. Additionally, cost optimization is important, so switching to open-source LLMs such as LLaMA3, Phi3, or Llava is necessary. Furthermore, expanding the tool's capabilities to handle a wider range of web content types, including videos, could further enhance its utility.

SmartCaption AI represents a significant step forward in making visual web content accessible to visually impaired users, contributing to a more inclusive digital environment. As the population of visually impaired individuals is projected to grow, tools like SmartCaption AI will play an increasingly crucial role in ensuring equal access to online information and services.

## ACKNOWLEDGMENT

The authors wish to express their sincere gratitude to Al-goma University Research Fund for their generous support of this work. Their commitment to fostering innovative research has been invaluable to the success of this project. We would also like to thank Gaici Lin and Kyle Gauthier for their helpful feedback on the Experiment section.

## REFERENCES

- [1] R. Bourne, J. D. Steinmetz, S. Flaxman, P. S. Briant, H. R. Taylor, S. Resnikoff, R. J. Casson, A. Abdoli, E. Abu-Gharbieh, A. Afshin *et al.*, "Trends in prevalence of blindness and distance and near vision impairment over 30 years: an analysis for the global burden of disease study," *The Lancet global health*, vol. 9, no. 2, pp. e130–e143, 2021.
- [2] W3C, "Visual — web accessibility initiative (wai) — w3c," <https://www.w3.org/WAI/people-use-web/abilities-barriers/visual/>, [Online; accessed 02-July-2024].
- [3] WebAIM, "WebAIM: The WebAIM Million - The 2024 report on the accessibility of the top 1,000,000 home pages," <https://webaim.org/projects/million/>, [Online; accessed 02-July-2024].
- [4] J. Ganesan, A. T. Azar, S. Alsenan, N. A. Kamal, B. Qureshi, and A. E. Hassanien, "Deep learning reader for visually impaired," *Electronics*, vol. 11, no. 20, p. 3335, 2022.
- [5] X. Liu, H. Li, J. Shao, D. Chen, and X. Wang, "Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 338–354.
- [6] D. Zhu, J. Chen, K. Haydarov, X. Shen, W. Zhang, and M. Elhoseiny, "Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions," *arXiv preprint arXiv:2303.06594*, 2023.
- [7] H. Jeong, M. Chun, H. Lee, S. Y. Oh, and H. Jung, "Wataa: Web alternative text authoring assistant for improving web content accessibility," in *Companion proceedings of the 28th international conference on intelligent user interfaces*, 2023, pp. 41–45.
- [8] D. Guinness, E. Cutrell, and M. R. Morris, "Caption crawler: Enabling reusable alternative text descriptions using reverse image search," in *Proceedings of the 2018 chi conference on human factors in computing systems*, 2018, pp. 1–11.
- [9] C. Gleason, A. Pavel, E. McCamey, C. Low, P. Carrington, K. M. Kitani, and J. P. Bigham, "Twitter ally: A browser extension to make twitter images accessible," in *Proceedings of the 2020 chi conference on human factors in computing systems*, 2020, pp. 1–12.
- [10] T. Ghandi, H. Pourreza, and H. Mahyar, "Deep learning approaches on image captioning: A review," *ACM Computing Surveys*, vol. 56, no. 3, pp. 1–39, 2023.
- [11] Aljazeera, "Hindus bathe in India's sacred Yamuna covered with toxic foam," <https://www.aljazeera.com/gallery/2021/11/11/india-hindus-yamuna-river-pollution-chhath-puja>, [Online; accessed 26-July-2024].
- [12] M. Xia, B. Huang, Y. Yan, and W. Lin, "Transforming patient experience in underserved areas with innovative voice-based healthcare solutions," in *Information and Communication Technology: ICICT 2024*. Springer, Feb 2024, pp. 1–10.
- [13] A. Ghosh, Y. Yan, and W. Lin, "Adaptive user interface framework powered by a large language model for culturally sensitive virtual healthcare applications," in *IEEE International Conference on Biomedical and Health Informatics (BHI'23)*, Pittsburgh, Pennsylvania, United States of America, 2023.
- [14] A. Ghosh, B. Huang, Y. Yan, and W. Lin, "Enhancing healthcare user interfaces through large language models within the adaptive user interface framework," in *Information and Communication Technology: ICICT 2024*. Springer, Feb 2024, pp. 1–10.
- [15] I. Rosario, B. Huang, Y. Yan, and W. Lin, "Enhancing telehealth patient experience with emotion-sensitive large language models," in *Information and Communication Technology: ICICT 2024*. Springer, Feb 2024, pp. 1–10.
- [16] A. Jamil, K. Mahmood, M. G. Villar, T. Prola, I. D. L. T. Diez, M. A. Samad, I. Ashraf *et al.*, "Deep learning approaches for image captioning: Opportunities, challenges and future potential," *IEEE Access*, 2024.
- [17] Aljazeera, "Families forced into a deadly spiral in Central African Republic," <https://www.aljazeera.com/gallery/2021/3/18/families-forced-into-a-deadly-spiral-in-central-african-republic>, [Online; accessed 26-July-2024].
- [18] —, "Senegalese plant circular gardens in Green Wall defence," <https://www.aljazeera.com/gallery/2021/7/29/senegalese-plant-circular-gardens-in-green-wall-defence>, [Online; accessed 26-July-2024].
- [19] H. Sarhan and S. Hegelich, "Understanding and evaluating harms of ai-generated image captions in political images," *Frontiers in Political Science*, vol. 5, p. 1245684, 2023.
- [20] Huffpost, "Children Facing Humanitarian Crisis Reveal What They Need Most," [https://www.huffpost.com/archive/ca/entry/children-facing-humanitarian-crisis-reveal-what-they-need-most\\_b\\_10131362](https://www.huffpost.com/archive/ca/entry/children-facing-humanitarian-crisis-reveal-what-they-need-most_b_10131362), [Online; accessed 26-July-2024].
- [21] Aljazeera, "Niger hosts prestigious camel race in the Sahara," <https://www.aljazeera.com/gallery/2021/9/21/ingall-niger-hosts-camel-race-sahara-desert>, [Online; accessed 26-July-2024].
- [22] —, "'Saved by coal': Far from COP26, another reality in India," <https://www.aljazeera.com/gallery/2021/11/11/photos-climate-crisis-saved-by-coal-far-from-cop26-another-reality-in-india>, [Online; accessed 26-July-2024].
- [23] Spiegel, "Berlin's Poor Collect Bottles to Make Ends Meet," <https://www.spiegel.de/international/germany/pensioners-in-berlin-collecting-deposit-bottles-to-supplement-income-a-823409.html>, [Online; accessed 26-July-2024].
- [24] B. Insider, "France's Macron thinks 'the worst is yet to come' in Ukraine after talking with Putin, reports say," <https://www.businessinsider.com/macron-thinks-the-worst-is-yet-to-come-in-ukraine-after-putin-call-2022-3>, [Online; accessed 26-July-2024].

# APPENDIX

TABLE III  
IMAGE CAPTION FROM AI-MCS, IMAGE TO TEXT, AND SMARTCAPTION AI, ALONG WITH THE ACCURACY SCORES

Image	AI MCS	ImageToText	SmartCaption AI (our tool)	Score
Image-5. Source: Adapted from [17] 	A child leaning on a wooden rail	a young boy is sitting on a wooden bench	The image shows a young child sitting alone, leaning against a wooden surface, with a somber expression on their face. The child appears to be in a state of distress, reflecting the severe impact of the humanitarian crisis in the Central African Republic. This visual poignantly conveys the vulnerability and hardship faced by children in conflict zones.	1.AI-MCS: Avg. score: 5 2.ImageToText: Avg. score: 2.33 3.SmartCaptionAI: Avg. score: 9
Image-6. Source: Adapted from [20] 	a group of children playing drums	a young girl is holding a small child and a bowl	The image shows a group of children standing in a line, holding containers, likely waiting for food or aid distribution. Their serious and somber expressions reflect the hardships and struggles they face in a humanitarian crisis. This visual underscores the urgency of addressing basic needs such as food and protection for children in emergency situations.	1.AI-MCS: Avg. score: 2.67 2.ImageToText: Avg. score: 1.33 3.SmartCaptionAI: Avg. score: 9
Image-7. Source: Adapted from [21] 	a group of people in traditional dress	a large group of people are standing in a line	The image depicts a large group of individuals dressed in vibrant traditional attire, gathered at the Cure Sale festival in Ingall, northern Niger. They are seen atop a vehicle and camels, showcasing the cultural heritage of the Tuareg and Wodaabe nomads. This annual festival celebrates their traditions through various activities, including camel racing, music, dances, and rituals, providing a joyful respite for the nomadic communities amidst their daily challenges.	1.AI-MCS: Avg. score: 5.33 2.ImageToText: Avg. score: 3.33 3.SmartCaptionAI: Avg. score: 8.33
Image-8. Source: Adapted from [22] 	a group of people fetching water	a woman sitting on a hill with a bunch of water bottles	The image depicts a group of women in Dhanbad, India, collecting water in large containers, likely near a coal mining area. This scene illustrates the harsh realities faced by communities dependent on coal for their livelihoods, highlighting the daily struggles for basic necessities such as water. The juxtaposition of their efforts against the backdrop of coal heaps underscores the complex challenges of balancing economic survival with environmental sustainability in developing regions.	1.AI-MCS: Avg. score: 5.33 2.ImageToText: Avg. score: 2.33 3.SmartCaptionAI: Avg. score: 9.67
Image-9. Source: Adapted from [23] 	a person carrying a bag	a woman sitting on a bench with a cell phone	The image depicts a group of women in Dhanbad, India, collecting water in large containers, likely near a coal mining area. This scene illustrates the harsh realities faced by communities dependent on coal for their livelihoods, highlighting the daily struggles for basic necessities such as water. The juxtaposition of their efforts against the backdrop of coal heaps underscores the complex challenges of balancing economic survival with environmental sustainability in developing regions.	1.AI-MCS: Avg. score: 4.67 2.ImageToText: Avg. score: 1.33 3.SmartCaptionAI: Avg. score: 9.33
Image-10. Source: Adapted from [24] 	men with suits sitting at a table	a man sitting at a table with a white tablecloth	The image depicts a meeting between two high-ranking officials, seated at opposite ends of a long, ornate table in an opulent room. This setting underscores the formality and gravity of the discussion, likely concerning significant geopolitical issues. The physical distance between the two individuals may symbolize the diplomatic and ideological divide between them.	1.AI-MCS: Avg. score: 4 2.ImageToText: Avg. score: 2 3.SmartCaptionAI: Avg. score: 6.33