Personalised and Contextualised Image Captioning: Human-in-the-Loop Design and Evaluation

Aliki Anagnostopoulou
DFKI & Carl-von-Ossietzky University of Oldenburg
Oldenburg, Germany
aliki.anagnostopoulou@dfki.de

Sara-Jane Bittner
DFKI
Oldenburg, Germany
sara-jane.bittner@dfki.de

Lavanya Govindaraju
DFKI
Oldenburg, Germany
lavanya.govindaraju@dfki.de

Hasan Md Tusfiqur Alam
DFKI
Oldenburg, Germany
hasan.alam@dfki.de

Daniel Sonntag
DFKI & Carl-von-Ossietzky University of Oldenburg
Oldenburg, Germany
daniel.sonntag@dfki.de

Abstract

Image captioning is an AI-complete task that bridges computer vision and natural language processing. Its goal is to generate textual descriptions for a given image. However, general-purpose image captioning often does not capture contextual information, such as information about the people present or the location the image was shot. To address this challenge, we propose a web-based tool that leverages automated image captioning, large foundation models, and additional deep learning modules such as object recognition and metadata analysis to accelerate the process of generating contextualised and personalised image captions. The tool allows users to create personalised and contextualised image captions efficiently. User interactions and feedback given to the various components are stored and later used for domain adaptation of the respective components. In a user study comparing our system to a proprietary baseline, the latter received slightly higher scores; however, our system demonstrated competitive performance while offering greater transparency and user support. Our ultimate goal is to improve the efficiency and accuracy of creating personalised and contextualised image captions.

CCS Concepts

• **Human-centered computing** → *Interactive systems and tools*; *User studies*; • **Computing methodologies** → **Natural language generation**; **Computer vision tasks**.

Keywords

image captioning, interactive machine learning, user-based evaluation, contextualisation and personalisation

Mensch und Computer 2025 – Workshopband, Gesellschaft für Informatik e.V., 31. August - 03. September 2025, Chemnitz, Germany

https://doi.org/10.18420/muc2025-mci-ws04-272

1 Introduction

Image captioning involves automatically generating textual descriptions for visual images, leveraging advancements in computer vision and natural language processing. Recent advances, particularly large-scale, pretrained models, have led to significant improvements in generating factual and syntactically correct captions [26, 29]. These systems have enabled applications ranging from accessibility for visually impaired users to automotive scene understanding and digital content generation [16, 24].

Despite strong performance on general benchmarks, image captioning models often struggle to incorporate user-specific or situational context not directly encoded in the image [9]. This challenge is not unique to captioning, but reflects broader limitations of large foundation models such as ChatGPT when used in isolation from real-world context [12]. In practical scenarios—such as generating captions for social media posts, personal photo albums, or assistive tools for users with diverse cognitive needs—contextual relevance and personalisation are often essential. This limitation is particularly pertinent when integrating user-specific details or external context, prompting the consideration of *interactive* and human-in-the-loop approaches that engage human participation [4].

Building on this line of work, we present CUTIE, which stands for *Contextual Understanding and Tailoring for Image Explanations*, a novel interactive system for contextualised image captioning. CUTIE combines deep learning-based object detection, optional metadata extraction, and large language models within an intuitive, photobook-style user interface that supports user-driven input and caption refinement. We introduce an interactive captioning tool that enables users to incorporate contextual and personal information into AI-generated captions. We propose a hybrid interface design that blends automated caption suggestions with editable, user-guided input. We conduct a user study comparing CUTIE with a proprietary baseline (GPT-4o-based tool) to evaluate usability, creativity, and user experience, of which we derive insights from qualitative feedback to inform future research on context-aware captioning interfaces. The findings, discussed in detail later, suggest that while the baseline system received slightly higher quantitative scores, CUTIE offered unique strengths in user support and scaffolding for caption refinement.

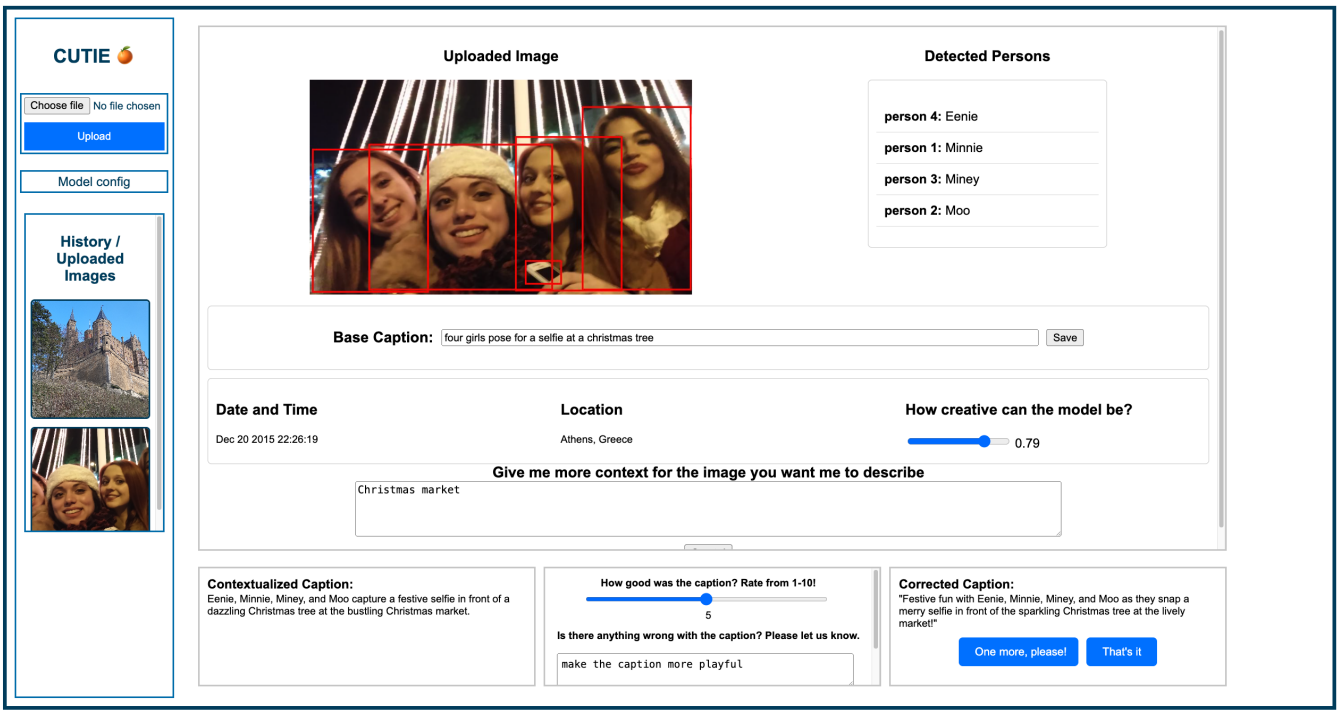

Figure 1: Screenshot of our intelligent user interface for personalised/contextualised image captioning.

2 Related work

Previous approaches in *interactive image captioning* have focused on improving general-use captions by integrating various interactive components: Peris and Casacuberta [19] present an interactive-predictive system for generation tasks, including image captioning, which considers user feedback and integrates online learning for adaptation. Jia and Li [17] involve the human-in-the-loop by providing incomplete sequences as input, in addition to each image, during inference time. Biswas et al. [8] extend the *Show, Attend, and Tell* [28] architecture by combining high-level and low-level features, which provide explainability and beam search during decoding time. Anagnostopoulou et al. [4] propose an interactive image captioning pipeline integrating data augmentation and continual learning to avoid overfitting and catastrophic forgetting during repeated training. Wang et al. [27] integrate interactive prompts for improved caption inference. More recently, Cai et al. [13] extend LLaVA by creating a model that allows users to mark images and interact with them with visual prompts.

Contextualised image captioning considers additional context to generate an image caption that describes the image's content and includes relevant external information. The context provided is, in most cases, in text form. Biten et al. [10] and Tran et al. [25] use news articles as context; the former uses a template-based architecture, and the latter uses an end-to-end architecture, considering additional features such as face and object detection. A modified version of the model proposed by Tran et al. [25] is used in Nguyen et al. [18] for image captioning on Wikipedia [23]. However, these methods typically operate in a fully automated manner and do not incorporate user interaction. They are designed to generate captions based on a predefined image–text pair (e.g., an article or snippet) without accounting for user-driven input or contextual adaptation. In contrast, CUTIE is explicitly designed to support user agency, allowing individuals to influence and tailor the contextual inputs considered during caption generation.

3 System design

We demonstrate a web-based tool for interactive image captioning. Human-in-the-loop is essential for generating personalised and contextualised image captions. The tool allows users to process images in a photo-editing-like interface (Figure 1). We integrate various deep learning modules to extract information that the user needs to provide. Contextual information and user feedback are incorporated via large language models (LLMs) and stored for fine-tuning the deep learning components (Figure 2).

The interface was designed through an internal, iterative process guided by our team's experience and the specific requirements of the project No-IDLE [22] for which the system was developed. Rather than employing a participatory design methodology, we based our initial prototype on anticipated user needs, informed by prior work from our department related to user interfaces and human-AI interaction [5, 7], common practices in image captioning workflows [19], and the practical constraints of the project context. Design decisions were made with an emphasis on usability, responsiveness, and seamless integration with captioning models and LLMs. Informal internal evaluations and ongoing adjustments helped refine the interface throughout the development process.

Although formal user studies were not conducted at this stage, the design aligns pragmatically with the goals and constraints of the intended use case.

User interface. The user interface includes four main components, as seen in Figure 1: the left bar for uploading new images or selecting old ones for captioning, as well as choosing the models for image caption generation and contextualisation; the top central box, showcasing clickable object detections, which the user can then use to enter person names; the middle central box for metadata and temperature selection; and the bottom central boxes for the manual addition of context information, caption rating, and feedback incorporation. The generation of a contextualised image caption occurs in three stages. In the *first stage*, the user uploads an image (users can also re-caption existing images) and selects a model combination for captioning and contextualisation. The image is then processed for (a) object detection and (b) image captioning. In the *second stage*, the user can provide more information for personalisation and contextualisation, as well as feedback: The uploaded image is displayed on the interface, along with detected objects marked with a red bounding box. Users can click on detected persons to initiate annotation. After selecting a detected person, a text input field appears in the designated annotation panel on the right. Users can then enter the name of the person being annotated. Each time a new person is selected for annotation, an additional text input field is dynamically generated within the annotation panel. This allows multiple persons to be annotated simultaneously. The base caption generated by the image captioning component is displayed below. The user can edit and save the improved version if the initial caption contains errors. The detected metadata, namely date, time, and location, are shown in the central component. Users can adjust the generation temperature on the right part before generating the personalised and contextualised caption. Additionally, they can provide additional information relevant to the captioning process. During the *third stage*, personalised and contextualised image captioning occurs, based on person names, base captions, metadata, and further context. The initial generated caption is displayed in the left section of the bottom central component. Users can rate the quality of the generated caption on a scale from 1 to 10 and propose improvements, which are incorporated into the updated caption shown in the bottom-right section of the interface.

Implementation. Our presented tool employs multiple deep learning components to generate personalised and contextualised image captions. The two main components are an image captioning system, which extracts visual information from the input image in the form of a *base caption*, and an LLM, which leverages contextual information to transform the base caption into a *personalised/contextualised caption*. We follow the two-step contextualised caption generation procedure proposed by [3], with additional components to extract and elicit relevant information not present in the image. While this two-stage approach can, in theory, be substituted by using visual/multimodal LLMs, we argue that it provides increased controllability and interpretability and lower inference costs. Initially, the input image is processed by both the object detection component and the image captioning one. For object detection, we utilise a Faster R-CNN model[1] [20] provided by Torchvision. For image captioning, the user can select between two pre-trained models: `BLIP-2`[2] and `ViT-GPT2`[3], both provided by Huggingface. Furthermore, if the image file contains metadata, this is extracted using the `EXIF` library in Python3. To convert the information for latitude and longitude into an exact location, `Geopy` is additionally used. After the user inputs information about the people present, the correctness of the base caption, the necessity of metadata in the caption, and the temperature for generating the caption, the user feedback is used as input into the LLM chosen by the user. The user can choose between GPT-4o, provided by the OpenAI API, and llama3 [14], provided by Ollama[4]. An initial caption is generated, conditioned on the image description from the image captioning component, people's names, and additional information inferred from the image metadata or manually entered by the user. The user can rate the quality of the caption and suggest improvements or changes. The first version of the caption is passed to the LLM, along with the proposed changes, and an updated caption is generated. In parallel, user input and corrective feedback are stored in the backend. In the future, this information can be used to fine-tune the deep learning components individually. To improve scalability and performance, the system parallelises computations using a `ThreadPoolExecutor`, which reduces redundant tasks with `Flask-Caching` backed by an in-memory cache, ensuring faster response times for multiple simultaneous image processing requests.

4 Evaluation: User Study

The proposed interface was used as a testbed for studying caption personalization and contextualization. We selected OpenAI's ChatGPT (GPT-4-turbo[5]) as a baseline for comparison. To imitate our tool's two-step approach, users first had to generate a description of the given image and then formulate a prompt to generate a contextualised and personalised caption.

4.1 Participants

We conducted a small study with seven users. The users were between 21 and 46 years old (M=30.3, Std=8.40). Two participants were female (28.6%) and five participants were male (71.4%).

The participants rated their subjective experience with generative AI as average to low, with a mean of M=2.86 on a scale from 1 (minimal experience) to 5 (a lot of experience) (SD=1.35). The participants reported how often they use generative AI: *Never*, *less than once a month*, *a few times a month*, *a few times a week*, or *almost every day*. The participants' experience with generative AI is balanced, with two participants reporting that they use it *less than once per month*, three who use it *a few times a month*, and two who even use it *a few times per week*.

The participants' attitude towards AI was measured using an adapted version of the 9-item Affinity for Technology Interaction (ATI) scale [15]. The term *technical systems* was adapted to *AI system* to fit the context of our tool. The participants showed an average to

[1] https://pytorch.org/vision/main/models/generated/torchvision.models.detection.fasterrcnn_resnet50_fpn_v2.html
[2] https://huggingface.co/Salesforce/blip2-opt-2.7b
[3] https://huggingface.co/nlpconnect/vit-gpt2-image-captioning
[4] https://ollama.com/
[5] https://chat.openai.com, last access: July 23rd 2025.

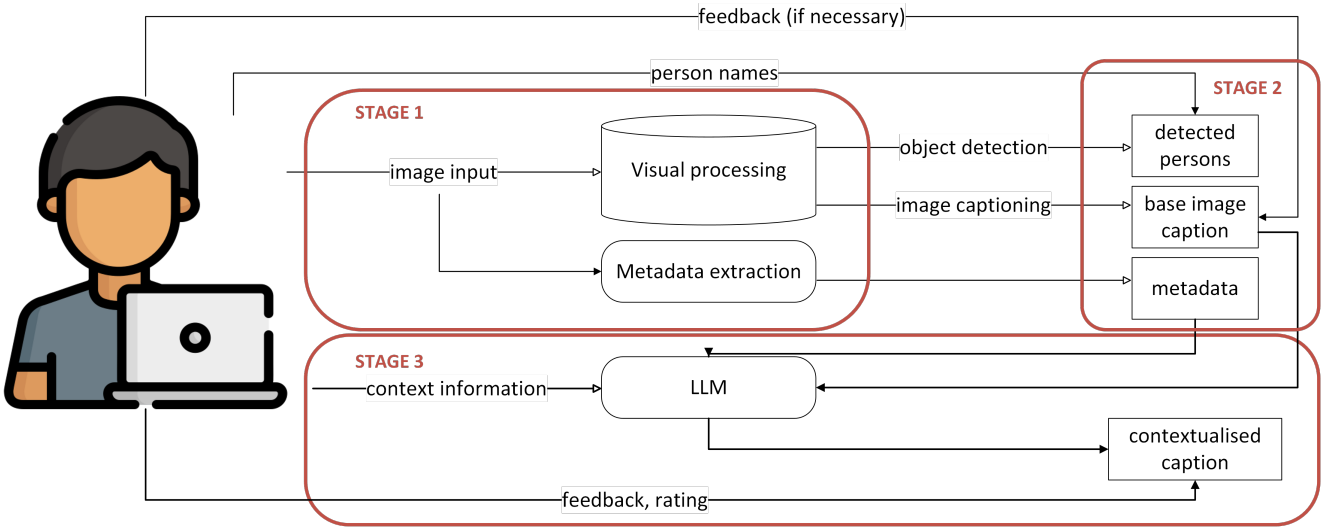

Figure 2: Overview of the architecture of our interactive image captioning system.

slightly higher affinity towards AI systems with a mean of M=3.70 (SD=0.72). The results suggest that participants tend to interact with such systems naturally and are open-minded towards them.

4.2 Methods & Measures

This study included measures to observe participants' perception of the quality and creativity of the generated captions, as well as their user experience and perceived usability when interacting with CUTIE, the caption generation tool, and a baseline tool. The perceived quality and creativity of captions were assessed with a 5-point Likert scale (1 = very low; 5 = very high). The *SUS* was used to assess the subjective usability of the caption generation tool [11]. The *SUS* contains a total of 10 items, which are measured by a 5-point Likert scale. An example item is: *"I found the system unnecessarily complex"*. The score ranges from 0 to 100, with higher values representing better usability. The short version of the *User Experience Questionnaire* (UEQ-S) was used to assess the subjective user experience of the caption generation tool [21]. The *UEQ-S* contains a total of 8 items on a 7-point Likert scale. While the mean value of the eight items will be given as an overall UX value, the questionnaire can be split into two sub-scales: The first four items form the *pragmatic* quality scale, which focuses on goal-orientation, and the last four items represent the *hedonic* quality scale, which focuses on how interesting, and stimulating the users' experience with the tool was. A higher mean-score represents a better user experience. Furthermore, the study included a questionnaire with open-ended questions to gain further insights into the users' experiences and perceptions of the baseline and our caption generation tool, CUTIE. To analyze the open-ended questions, an affinity diagram was applied.

4.3 Results

The findings from our user study are reported in Table 1 and elaborated upon in the subsequent discussion.

Metric	CUTIE	GPT-4o
Quality	M=3.29 (SD=0.62)	M=4.00 (SD=0.50)
Creativity	M=3.66 (SD=0.52)	M=4.00 (SD=0.50)
Usability	M=64 (SD=16.63)	M=81 (SD=10.5)
Overall UX	M=4.61 (SD=0.92)	M=5.79 (SD=0.64)
- Pragmatic Sub-Scale	M=5.00 (SD=1.27)	M=6.14 (SD=0.67)
- Hedonic Sub-scale	M=4.21 (SD=1.02)	M=5.43 (SD=0.81)

Table 1: Comparison of CUTIE and GPT-4o across user evaluation metrics

Quality & Creativity. Regarding the perceived quality captions generated by CUTIE received a medium rating. In comparison, the captions generated by the baseline tool received a rating of slightly above medium. These findings align with the results of the open-ended questionnaire. The quality of captions generated by CUTIE was described as mixed by participants, rating them *"sometimes good, sometimes bad"* (P3), or *"Okay. Not bad. Could be far better"* (P4). Three participants highlighted that after adapting the caption, the quality was *"good"* (P2, P4, P5). However, the captions were critiqued with P3 stating that *"some captions don't make sense and sometimes"* (P3), that they do not *"reflect the emotional state"* the participant would like to have (P4), or that they were *"too detailed"* (P7) with including all the contextual information. In comparison, the quality of the captions generated by the proprietary tool were rated *"very high"* (P1, P3, P4, P5, P6, P7), and described the captions as being *"not to detailed"* (P7) and *"That it was close enough with what [they] had on [their] mind"* (P4). One participant stated that it *"could be more fitting"*.

Furthermore, in terms of perceived creativity, the captions generated by CUTIE received a medium to above medium rating, while the captions generated by the baseline tool received a slightly above

medium rating. These findings align with the results of the open-ended questionnaire. The creativity of captions generated by CUTIE, was rated mixed by the participants. While some participants describe a *"very high level of creativity"* (P1, P2, P5), others felt overwhelmed by the amount of creativity. Yet another participant shows an opposite opinion by describing the captions as *"bland"* (P4). In comparison, the creativity of captions generated by the proprietary tool was rated *"very high"* (P1, P3, P4, P5, P6, P7). At the same time, two participants highlighted that *"creativity decreased after the adaptation"* (P2) and that it *"could be more creative"* (P6).

Usability. The *SUS* score of CUTIE, the caption generation tool, is M=64. Based on Bangor et al. [6], a score between 52 and 73 represents "ok" usability. Thus, CUTIE receives a marginal low score. The baseline tool receives a *SUS* score of 81 and reaches a "good" score, which is defined from 73 up to 85 points [6]. These findings align with the results of the open-ended questionnaire. The baseline tool was described as *intuitive* and *simple* to use (P3, P4, P5). The complexity of the tool was perceived as *not complex at all* (P1, P3, P4, P5, P6, P7), and participants required little to no time to become familiar with the tool. For CUTIE, participants showed mixed thoughts: Most participants share that they did not experience difficulties with the tool and had a good initial understanding (P1, P3, P4, P5), and most participants described the tool *"not too complex"* (P1, P2, P3, P4, P5, P6). It was highlighted that *"the order and the specific things you can adapt make sense"*. However, some participants were not sure *"what aspects need to be adjusted"* (P2), and how to use it initially (P1, P3, P4, P7), or disliked that they *"had to cl[i]ck the save and create buttons every time"* (P7).

User Experience. The overall user experience with CUTIE was rated medium to rather high based on the *UEQ-S*. In comparison, the participants rated the user experience with the baseline tool rather high. This pattern can be seen for the pragmatic and hedonic subscales as well: CUTIE receives rather high ratings for the pragmatic and a medium to high rating for the hedonic sub-scale The baseline tool achieves scores that are approximately one point higher for both scales.These findings align with the results of the open-ended questionnaire. Participants expressed that the baseline tool *"adapted well"* (P2, P3) to the orders. Especially the communicative feature of the tool was highlighted: P7 stated that *"it felt nice, because it asked you questions back"* and P1 liked that it *"also [said] what we can make better"*. However, it was pointed out that *"the does not support [them] actively"* (P3) and that manually inputting the prompts felt like *"[not] having enough support"* (P4). In comparison, participants user experience with CUTIE was more mixed. Some participants pointed out that *"the tool was intuitive"* (P2), that *most of the times* it *"reacts in a useful way"* (P3). They *"liked the ability to have a quick, prewritten captions"* and that *"it had some nice Ideas of describing a scene"* (P7). Other participants reported that the tool only helped *minimally* (P4) and that *"some captions don't make sense"* (P3).

5 Discussion

The results of our study yield meaningful insights into both the strengths and limitations of our proposed approach, and highlight promising avenues for future work. Notably, the relatively small sample size may limit the generalisability of our findings. Additionally, participant feedback occasionally revealed inconsistencies, particularly regarding subjective evaluations. These limitations are addressed in the following discussion.

Participants rated the baseline tool, where users inputted their own prompts, as more conducive to creative expression. In contrast, CUTIE received mixed qualitative feedback, suggesting that system-generated prompts may not always align with users' expectations or mental models. This points to the need for future iterations to explore *adaptive or user-influenced prompt generation mechanisms*.

Usability evaluations also presented divergent perspectives. Several participants reported confusion regarding the expected sequence of operations. These findings highlight a need for *clearer interaction affordances* and keyboard-first design considerations.

While the baseline tool was described as lacking adequate support, CUTIE's inclusion of prewritten captions was positively received. However, participants noted that some auto-generated captions lacked coherence or relevance. To address this, we propose e*nabling iterative refinement and increased user control over generated content* in future versions of the system.

6 Conclusion

We designed and implemented CUTIE, a novel tool for AI-assisted caption co-creation that integrates deep learning-based image analysis with an intuitive user interface to support collaborative human-AI interaction. The system generates initial captions based on object detection outputs, while allowing users to refine and adapt these captions through direct feedback iteratively. By reducing the manual effort associated with image annotation, CUTIE aims to enhance both the efficiency and accessibility of the creative process. To evaluate its effectiveness, we conducted a user study comparing CUTIE to a proprietary baseline system (ChatGPT). While the baseline tool, which enabled free-form prompt entry, was rated slightly higher in terms of perceived creativity, CUTIE was favorably received for its supportive features, including prewritten captions that some participants described as helpful scaffolds. Despite the modest participant sample size, qualitative feedback suggests that CUTIE provides tangible user support during the captioning process—particularly for users who benefit from structured starting points rather than open-ended generation. To build on these findings, future work will focus on (a) developing a prompt manipulation interface that suggests refinements for improved caption generation and (b) enabling repeated caption editing to better support iterative workflows. Additional avenues include the integration of structured concept-based reasoning and retrieval-augmented pipelines, as shown by Alam et al. [1, 2], to improve interpretability and domain alignment in specialised captioning contexts such as medical imaging.

Acknowledgements

This work was funded by the German Federal Ministry of Education and Research (BMBF) under grant numbers 01IW23002 (No-IDLE) and 01IW24006 (NoIDLEChatGPT), as well as by the Endowed Chair of Applied AI at the University of Oldenburg.

References

[1] Hasan Md Tusfiqur Alam, Devansh Srivastav, Md Abdul Kadir, and Daniel Sonntag. 2025. Towards Interpretable Radiology Report Generation via Concept Bottlenecks Using Multi-agentic RAG. In *Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part III* (Lucca, Italy). Springer-Verlag, Berlin, Heidelberg, 201–209. doi:10.1007/978-3-031-88714-7_18

[2] Hasan Md Tusfiqur Alam, Devansh Srivastav, Abdulrahman Mohamed Selim, Md Abdul Kadir, Md Moktadirul Hoque Shuvo, and Daniel Sonntag. 2025. CBM-RAG: Demonstrating Enhanced Interpretability in Radiology Report Generation with Multi-Agent RAG and Concept Bottleneck Models. In *Companion Proceedings of the 17th ACM SIGCHI Symposium on Engineering Interactive Computing Systems*. 59–61.

[3] Aliki Anagnostopoulou, Thiago Gouvea, and Daniel Sonntag. 2024. Enhancing Journalism with AI: A Study of Contextualized Image Captioning for News Articles using LLMs and LMMs. arXiv:2408.04331 [cs.CL] https://arxiv.org/abs/2408.04331

[4] Aliki Anagnostopoulou, Mareike Hartmann, and Daniel Sonntag. 2023. Towards Adaptable and Interactive Image Captioning with Data Augmentation and Episodic Memory. In *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing, SustaiNLP 2023, Toronto, Canada (Hybrid), July 13, 2023*, Nafise Sadat Moosavi, Iryna Gurevych, Yufang Hou, Gyuwan Kim, Young Jin Kim, Tal Schuster, and Ameeta Agrawal (Eds.). Association for Computational Linguistics, 245–256. doi:10.18653/V1/2023.SUSTAINLP-1.19

[5] Melis Aslan, Maximilian Bosse, Daniel Ehlers, Marlon Hinz, Philipp Olschewski, Jannik Podszun, Elias Scharlach, Leon Selzer, Yukun Wu, Aliki Anagnostopoulou, and Daniel Sonntag. 2025. TextVision: A more efficient way to work with research. In *Joint Proceedings of the ACM IUI 2025 Workshops co-located with the 30th Annual ACM Conference on Intelligent User Interfaces (IUI 2025), Cagliari, Italy, March 24th, 2025 (CEUR Workshop Proceedings, Vol. 3957)*, Dorota Glowacka, Carmen Santoro, and Ziang Xiao (Eds.). CEUR-WS.org, 430–443. https://ceur-ws.org/Vol-3957/MIND-paper05.pdf

[6] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining what individual SUS scores mean: adding an adjective rating scale. *J. Usability Studies* 4, 3 (May 2009), 114–123.

[7] Michael Barz, Omair Shahzad Bhatti, Hasan Md Tusfiqur Alam, Duy Minh Ho Nguyen, and Daniel Sonntag. 2023. Interactive Fixation-to-AOI Mapping for Mobile Eye Tracking Data based on Few-Shot Image Classification. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces, IUI 2023, Sydney, NSW, Australia, March 27-31, 2023*. ACM, 175–178. doi:10.1145/3581754.3584179

[8] Rajarshi Biswas, Michael Barz, and Daniel Sonntag. 2020. Towards Explanatory Interactive Image Captioning Using Top-Down and Bottom-Up Features, Beam Search and Re-ranking. *Künstliche Intell.* 34, 4 (2020), 571–584. doi:10.1007/S13218-020-00679-2

[9] Ali Furkan Biten, Lluís Gómez, Marçal Rusiñol, and Dimosthenis Karatzas. 2019. Good News, Everyone! Context Driven Entity-Aware Captioning for News Images. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 12466–12475. doi:10.1109/CVPR.2019.01275

[10] Ali Furkan Biten, Lluís Gómez, Marçal Rusiñol, and Dimosthenis Karatzas. 2019. Good News, Everyone! Context Driven Entity-Aware Captioning for News Images. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 12466–12475. doi:10.1109/CVPR.2019.01275

[11] John Brooke. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7. https://www.taylorfrancis.com/chapters/edit/10.1201/9781498710411-35/sus-quick-dirty-usability-scale-john-brooke Publisher: London, England.

[12] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* (2023).

[13] Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P. Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. 2024. Making Large Multimodal Models Understand Arbitrary Visual Prompts. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[14] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The Llama 3 Herd of Models. *CoRR* abs/2407.21783 (2024). arXiv:2407.21783 doi:10.48550/ARXIV.2407.21783

[15] Thomas Franke, Christiane Attig, and Daniel Wessel. 2019. A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. *Int. J. Hum. Comput. Interact.* 35, 6 (2019), 456–467. doi:10.1080/10447318.2018.1456150

[16] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. Captioning Images Taken by People Who Are Blind. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVII (Lecture Notes in Computer Science, Vol. 12362)*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer, 417–434. doi:10.1007/978-3-030-58520-4_25

[17] Zhengxiong Jia and Xirong Li. 2020. iCap: Interactive Image Captioning with Predictive Text. In *Proceedings of the 2020 International Conference on Multimedia Retrieval* (Dublin, Ireland) *(ICMR '20)*. Association for Computing Machinery, New York, NY, USA, 428–435. doi:10.1145/3372278.3390697

[18] Khanh Nguyen, Ali Furkan Biten, Andrés Mafla, Lluís Gómez, and Dimosthenis Karatzas. 2023. Show, Interpret and Tell: Entity-Aware Contextualised Image Captioning in Wikipedia. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, Brian Williams, Yiling Chen, and Jennifer Neville (Eds.). AAAI Press, 1940–1948. doi:10.1609/AAAI.V37I2.25285

[19] Álvaro Peris and Francisco Casacuberta. 2019. A Neural, Interactive-predictive System for Multimodal Sequence to Sequence Tasks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Marta R. Costa-jussà and Enrique Alfonseca (Eds.). Association for Computational Linguistics, Florence, Italy, 81–86. doi:10.18653/v1/P19-3014

[20] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 6 (2017), 1137–1149. doi:10.1109/TPAMI.2016.2577031

[21] Martin Schrepp, Jörg Thomaschewski, and Andreas Hinderks. 2017. Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S). *International Journal of Interactive Multimedia and Artificial Intelligence* 4, 6 (12/2017 2017), 103–108. doi:10.9781/ijimai.2017.09.001

[22] Daniel Sonntag, Michael Barz, and Thiago S. Gouvêa. 2024. A look under the hood of the Interactive Deep Learning Enterprise (No-IDLE). *CoRR* abs/2406.19054 (2024). arXiv:2406.19054 doi:10.48550/ARXIV.2406.19054

[23] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 2443–2449. doi:10.1145/3404835.3463257

[24] Matteo Stefanini, Marcella Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara. 2021. From Show to Tell: A Survey on Deep Learning-Based Image Captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45 (2021), 539–559. https://api.semanticscholar.org/CorpusId:244772950

[25] Alasdair Tran, Alexander Patrick Mathews, and Lexing Xie. 2020. Transform and Tell: Entity-Aware News Image Captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 13032–13042. doi:10.1109/CVPR42600.2020.01305

[26] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (Eds.). PMLR, 23318–23340. https://proceedings.mlr.press/v162/wang22al.html

[27] Yiyu Wang, Hao Luo, Jungang Xu, Yingfei Sun, and Fan Wang. 2024. Text Data-Centric Image Captioning with Interactive Prompts. *CoRR* abs/2403.19193 (2024). arXiv:2403.19193 doi:10.48550/ARXIV.2403.19193

[28] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell:

Neural Image Caption Generation with Visual Attention. In *Proceedings of the 32nd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 37)*, Francis Bach and David Blei (Eds.). PMLR, Lille, France, 2048–2057. https://proceedings.mlr.press/v37/xuc15.html

[29] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, Nanyun Peng, Lijuan Wang, Yong Jae Lee, and Jianfeng Gao. 2023. Generalized Decoding for Pixel, Image, and Language. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 15116–15127. doi:10.1109/CVPR52729.2023.01451

A User study results

We include the results of our user study for the scales we used in our user study, along with the questions from the open-ended questionnaires.

A.1 Rating scales: Quality and creativity, SUS, ATI

Table 2 and Table 3 show the results for perceived quality and creativity of the captions generated with CUTIE and the proprietary baseline tool, accordingly, while Table 4 and Table 5 show the results for the SUS [6]. Finally, Table 6 presents the participants' responses to the ATI scale.

Condition: CUTIE										
Participant-ID	Q1	C1	Q2	C2	Q3	C3	Q4	C4	Q5	C5
1	+	0	0	+	-	0	-	+	+	+
2	+	0	0	0	++	++	+	+	0	+
3	-	++	-	+	+	+	–	+	++	++
4	+	+	0	0	-	-	+	+	-	-
5	++	+	0	(n.s.)	++	(n.s.)	++	++	++	++
6	+	0	-	+	+	+	-	-	0	-
7	0	0	-	+	+	+	-	-	-	+

Table 2: Quality (Q) and creativity (C) results for each test image (1-5) caption using CUTIE. P. = participant. ++ = very high, + = high, 0 = neutral, - = low, – = very low.

Condition: proprietary (gpt-4o)										
Participant-ID	Q1	C1	Q2	C2	Q3	C3	Q4	C4	Q5	C5
1	++	+	+	++	+	-	++	+	(n.s.)	(n.s.)
2	+	0	0	-	+	++	0	-	0	+
3	++	++	++	++	0	++	++	++	++	+
4	+	+	++	++	0	0	++	++	–	–
5	++	++	++	++	++	++	+	+	++	++
6	+	+	0	+	0	0	+	+	+	0
7	+	+	+	+	+	+	+	+	-	+

Table 3: Quality (Q) and creativity (C) results for each test image (1-5) caption using gpt-4o. P. = participant. ++ = very high, + = high, 0 = neutral, - = low, – = very low.

A.2 Qualitative evaluation: Open-end questionnaire

We include open-ended questions following each condition and scale evaluation, as well as the final questionnaire, which assesses demographics, technological affinity, and additional general questions.

Following the quality/creativity rating scales:

(1) Please describe your experience with adapting the caption with the tool.
(2) Does the tool support you with adapting the caption? If yes: Why so? If no: Why not?
(3) What did you like in the caption adaptation tool?
(4) What did you dislike in the caption adaptation tool?
(5) How did you perceive the quality of the captions generated by the captioning tool after your adaptation?
(6) How did you perceive the creativity of the captions generated by the captioning tool after your adaptation?

Following the SUS:.

(1) Did you experience difficulties in adapting the image caption? If yes: Why?
(2) How complex is the tool for you and what is causing this complexity for you?
(3) How was your initial understanding of using the captioning tool?
(4) Did you need time to understand the design? Why?
(5) Did you find any inconsistencies in the captioning tool? If, yes: Which ones?

Final questionnaire:

(1) How old are you?
(2) As which gender do you identify?
(3) How do you rate your experience with generative AI?
(4) How often do you use AI?
(5) In your opinion, how suitable is the captioning tool CUTIE for the adaptation of image captions?
(6) In general, how does the captioning tool CUTIE compare to captioning with traditional tools like ChatGPT in the adaptation of image captions?
(7) Would you like to add something?

Condition: CUTIE							
Participant-ID	1	2	3	4	5	6	7
1. I think that I would like to use this system frequently.	1	3	1	1	2	3	0
2. I found the system unnecessarily complex.	1	1	1	0	2	1	4
3. I thought the system was easy to use.	3	3	2	4	3	3	1
4. I think that I would need the support of a technical person to be able to use this system.	0	2	1	1	1	3	1
5. I found the various functions in this system were well integrated.	2	4	3	3	3	2	0
6. I thought there was too much inconsistency in this system.	1	1	1	0	3	2	3
7. I would imagine that most people would learn to use this system very quickly.	4	3	3	4	3	3	2
8. I found the system very akward to use.	1	0	1	0	3	1	4
9. I felt very confident using the system.	3	3	2	4	2	3	2
10. I needed to learn a lot of things before I could get going with the system.	1	3	1	0	2	1	1

Table 4: SUS questionnaire results for CUTIE. The numbers range from 0 (strongly disagree) to 4 (strongly agree).

Condition: proprietary (gpt-4o)							
Participant-ID	1	2	3	4	5	6	7
1. I think that I would like to use this system frequently.	3	2	3	3	4	3	3
2. I found the system unnecessarily complex.	1	1	0	0	0	1	0
3. I thought the system was easy to use.	4	3	4	4	4	3	3
4. I think that I would need the support of a technical person to be able to use this system.	0	0	0	0	0	3	0
5. I found the various functions in this system were well integrated.	3	2	3	2	4	3	1
6. I thought there was too much inconsistency in this system.	2	1	0	1	1	1	1
7. I would imagine that most people would learn to use this system very quickly.	4	3	4	4	3	3	4
8. I found the system very awkward to use.	1	0	0	0	0	1	1
9. I felt very confident using the system.	3	2	3	4	4	3	0
10. I needed to learn a lot of things before I could get going with the system.	0	3	0	0	1	1	1

Table 5: SUS questionnaire results for gpt-4o. The numbers range from 0 (strongly disagree) to 4 (strongly agree).

Participant-ID	1	2	3	4	5	6	7	AVG	SD
1. I like to occupy myself in a greater deal with AI systems.	4	4	2	2	4	5	6	3,86	1,46
2. I like testing the functions of new AI systems.	5	5	4	2	5	5	6	4,57	1,27
3. I predominantly deal with AI systems because I have to.	2	4	5	6	5	5	5	4,57	1,27
4. When I have a new technical system in front of me, I try it out intensively.	4	2	3	4	4	4	5	3,71	0,95
5. I enjoy spending time becoming acquainted with a new AI system.	2	2	5	5	5	3	5	3,86	1,46
6. It is enough for me that an AI system works; I don't care how or why.	1	2	4	2	4	2	2	2,43	1,13
7. I try to understand how an AI systems exactly works.	2	4	3	2	4	5	4	3,43	1,13
8. It is enough for me to know the basic functions of an AI system.	3	2	2	3	4	5	2	3,00	1,15
9. I try to make full use of the capabilities of an AI system.	3	3	3	2	5	5	6	3,86	1,46
AVG	2,89	3,11	3,44	3,11	4,44	4,33	4,56	3,70	0,72

Table 6: ATI scale results. The replies range from completely disagree (1) to completely agree (6). Average scores (AVG) and standard deviation (SD) are included.