Images Speak Volumes: User-Centric Assessment of Image Generation for Accessible Communication

Miriam Anschütz and **Tringa Sylaj** and **Georg Groh**
School for Computation, Information and Technology
Technical University of Munich, Germany
{miriam.anschuetz, tringa.sylaj}@tum.de, grohg@in.tum.de

Abstract

Explanatory images play a pivotal role in accessible and easy-to-read (E2R) texts. However, the images available in online databases are not tailored toward the respective texts, and the creation of customized images is expensive. In this large-scale study, we investigated whether text-to-image generation models can close this gap by providing customizable images quickly and easily. We benchmarked seven, four open- and three closed-source, image generation models and provide an extensive evaluation of the resulting images. In addition, we performed a user study with people from the E2R target group to examine whether the images met their requirements. We find that some of the models show remarkable performance, but none of the models are ready to be used at a larger scale without human supervision. Our research is an important step toward facilitating the creation of accessible information for E2R creators and tailoring accessible images to the target group's needs.

1 Introduction

Easy-to-read (E2R) and its German derivative *Leichte Sprache* (Easy Language) are accessibility- and readability-enhanced versions of language. They follow a strict ruleset and are targeted at people with disabilities, learning difficulties, or low literacy (DIN-Normenausschuss Ergonomie, 2023). The creation of a more accessible version of an original text is called text simplification (TS). Since this process is laborsome, previous work explored the applicability of large language models to facilitate or even automatize the creation of E2R texts (Madina et al., 2023). For German Easy language, Schomacker et al. (2023) investigated how well the currently available, text-oriented models and datasets comply with the ruleset of German Easy language and multiple open-source automatic TS models for German exist (Anschütz et al., 2023; Stodden et al., 2023).

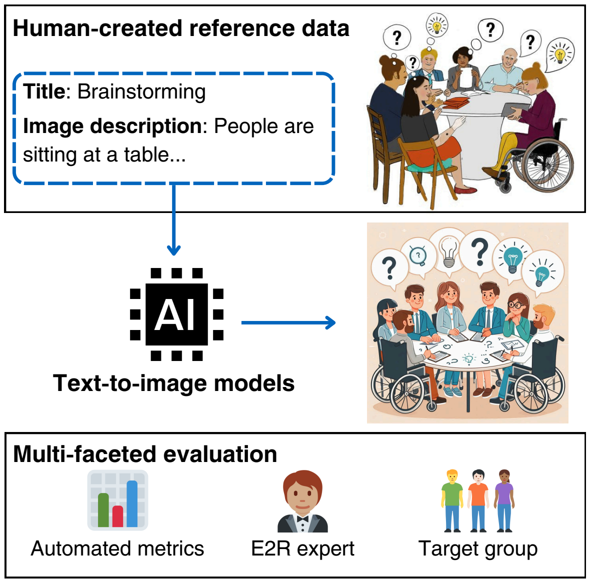


Figure 1: Overview of our approach: We selected a human-created reference dataset that was validated by the target group already. Based on the images' titles and descriptions, we used seven different text-to-image models to recreate the original images. Then, we evaluated the generated images across multiple aspects using automated and human evaluation.
© JSCHKA Kommunikationsdesign | www.jschka.de

However, one important feature of E2R texts is that they are illustrated with images that improve and facilitate the text's understanding even further. The guidelines in the DIN-SPEC 33429 (DIN-Normenausschuss Ergonomie, 2023) recommend that these images should be created specifically for each text and that they should be up-to-date and close to the target group's everyday life. Even though large image databases exist[1] that were reviewed and validated by the target group, these images were created for a general purpose and cannot be altered by the text creators. In addition, while human artists are unchallenged in creating

[1] e.g. https://www.lag-sb-rlp.de/projekte/bildergalerie-leichte-sprache

Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024), pages 27–40
November 15, 2024 ©2024 Association for Computational Linguistics

the most targeted and realistic images, their employment is financially infeasible for most E2R translators. Therefore, our work explores whether text-to-image (T2I) models can solve this problem by creating quickly available, flexible, and cheap images. An overview of our approach is presented in Figure 1.
Our contribution can be summarized as follows:

- We benchmark seven text-to-image models (four open-sourced and three closed-source) on their ability to create images for accessible communication.
- The resulting images are published as a dataset consisting of 2,217 images[2]. This dataset is relevant for text creators searching for a large and diverse image database as well as for AI researchers who want to train models or evaluation metrics for this task.
- We manually reviewed 560 images and annotated them by their closeness to the prompt, correctness, bias toward people with disabilities, and suitability for the target group. Our findings indicate that the quality of the generated images is highly dependent on the depicted content and the T2I models used and that even the best models cannot be utilized for a broader scale without further restrictions.
- We conducted a user study with seven people from the target group and report their opinions and preferences about the generated images.

2 Related work

Previous work has utilized images to enhance text accessibility, particularly in fields such as language learning. Geislinger et al. (2023) developed an iPad app for language learners that included an eye-tracking feature. When a reader focused on a word for a longer time, they retrieved a picture illustrating that word and showed it next to the text to improve the text's understanding. Similarly, Singh et al. (2023) and Schneider et al. (2021) focused on retrieving images for textbooks to improve the learning experience and make the books more appealing. To train and benchmark those retrieval models, Wang et al. (2022) published the MOTIF dataset. The dataset consists of sentences with complex words within the sentences and images that represent the context of the sentences. The complex word is highlighted within the images to give an easy-to-grasp explanation of those complex words. However, all of the previous methods only search for images in existing image databases and explore the capabilities of image retrieval methods. In contrast to this, our focus lies on the generation of new images and benchmarking models to create those images. A similar task was proposed by Kiesel et al. (2024), who tried to strengthen argumentation chains by providing images supporting the argument's premise. Nevertheless, to the best of our knowledge, this is the first work to explore image generation to automatically enhance accessible communication.

There exist multiple studies about the characteristics of image generation models, but none of them addresses their applicability to accessible communication. Mack et al. (2024) benchmarked different T2I models like DALL-E 2, Stable Diffusion, and Midjourney about how they depict disabilities. Even though the prompts described different forms of disabilities, the models mostly depicted disabled people as sitting in wheelchairs. The findings were repeated in the study from Tevissen (2024). The author investigated the latest Stable Diffusion checkpoints, SDXL and Stable Diffusion 3, DALL-E 3, and Midjourney. Again, people with disabilities were depicted very stereotypically: as old and sad people sitting in wheelchairs.

In our study, we include people from the target group and also report their perspectives on image generation models for accessible communication. Similar user studies were conducted in previous work. Huh et al. (2023) aim to make image generation as a process more accessible. They created a framework called GenAssist, in which blind and low-vision creators can ask questions about the image to determine whether the image generation models followed their prompts or whether additional content was added. A user study with the target group proved that the tool made visual creations more accessible. Another target group study was conducted by James Edwards et al. (2021), who worked with people with disabilities and asked them how their disability should be depicted in generated images. They especially focused on disability descriptions and the best level of detail for these descriptions. Similarly, Das et al. (2024) worked together with image creators and screen reader users to evaluate images' alt texts from different perspectives. They report that manually created alt texts

[2] https://github.com/MiriUll/Image-Generation-for-Accessible-Communication

are often too subjective and that prompts for T2I models cannot be used as alt text alternatives.

3 Methodology

Our study investigates whether the latest T2I models can create images suitable for E2R texts. For this, we use an open-source database of images for German Easy language and try to recreate the images based on their title and descriptions. Most of the images in E2R image databases are cartoon images since they are often easier than photo-realistic images, and the readers don't get confused by actual people. Therefore, we only focus on the generation of cartoon images as well.

3.1 Reference dataset

Our target dataset is the publicly available Leichte Sprache image gallery[3] from the LAG Selbsthilfe von Menschen mit Behinderungen und chronischen Erkrankungen Rheinland-Pfalz e.V. (State working group for self-help for people with disabilities and chronic illnesses Rhineland-Palatinate, Germany), a state-level organization uniting self-help associations and groups of individuals with disabilities or chronic illnesses and their relatives. It offers 413 images within 16 categories drawn by the artist Juliane Kriegereit[4]. The images were created for E2R texts and reviewed by the target group. An example image is shown in Figure 2. The images' license is very permissive to enable content creators to illustrate their texts. The categories are targeted to people with disabilities and cover areas like assisting technologies, diseases, and body parts. Each picture comes with a topic that is depicted and a description of the image's contents.

For our experiments, we randomly selected five images per category, yielding a dataset of $16 \times 5 = 80$ reference images in total. We translated the image titles and descriptions into English using ChatGPT (OpenAI et al., 2024) since some image generation models only work with English prompts.

3.2 Text2Image models

Our model selection featured a mix of open and closed-source models, SOTA and older models, as well as models of various sizes, culminating in a comprehensive evaluation of seven models in total. An overview of the models can be found in Table 3 in the Appendix.

[3]https://www.lag-sb-rlp.de/projekte/bildergalerie-leichte-sprache

[4]JSCHKA Kommunikationsdesign | www.jschka.de

Figure 2: Example image for the word "Inclusion" from the Leichte Sprache image gallery. The image description is "A group of very different people with and without disabilities is sitting at a table and eating together." © JSCHKA Kommunikationsdesign | www.jschka.de

We constructed the model prompts as "Cartoon picture of {title} - {description}" where we filled the placeholders with the values from the dataset and used the same prompts for all models.

For the open-sourced models, we utilized various versions of Stable Diffusion (Rombach et al., 2022) and Würstchen (Pernias et al., 2024). Stable Diffusion v1.4, v2.1 base, and v3 were employed to generate 512x512 pixel images. For SD3, we used the default parameters optimized for the output quality: *num_inference_steps* was configured to 28, defining the number of denoising steps the model takes during image generation. A higher number of inference steps generally leads to finer details and improved image quality. Additionally, the *guidance_scale* was set to 7.0, indicating the strength of the conditioning on the input text prompt. A higher guidance scale helps produce images that are more closely aligned with the given text descriptions, ensuring the semantic accuracy of the generated images.

Würstchen (Pernias et al., 2024) is another diffusion model where the text-conditional component functions within a significantly compressed latent space of images, attaining a 42x spatial compression. This enables the model to be much more time- and memory-efficient, significantly reducing training and inference time. Würstchen was used to produce higher-resolution images at 1024x1024 pixels, using a *prior_guidance_scale* set to 4.0, which similarly influences the model's adherence to the textual input.

For the closed-sourced model we focused on DALL-E-3 (Ramesh et al., 2021), Midjourney [5], and Artbreeder [6]. We accessed DALL-E-3 through the Bing Image Creator by Microsoft [7], which Microsoft states is powered by an advanced version of the DALL-E model. We used the free version, which allows 15 prompts per day. For Artbreeder, we use the Composer model, which is a GAN architecture (Goodfellow et al., 2020) incorporating elements of BigGAN (Brock et al., 2019) and StyleGAN (Karras et al., 2021).

3.3 Evaluation

We evaluated our generated images on different aspects, which include the closeness to the reference images, how well the models follow the image description in the prompt, and the image correctness.

The most popular automatic evaluation metric to measure the quality of a generated image is the Inception Score (Salimans et al., 2016). However, it compares the generated images against photo-realistic reference images from the CIFAR-10 dataset that are limited in the items they depict. Hence, the inception score is not suitable for our cartoon-style images (Proven-Bessel et al., 2021; Barratt and Sharma, 2018). To automatically assess the quality of our generated images, we used the Fréchet Inception Distance (FID). In contrast to the inception score, FID compares the generated images against a set of user-selected reference images. It estimates the distributions of the reference and generated image sets and reports the distance between the two distributions. Therefore, a lower FID score indicates better matches with the reference images and, thus, a better overall image quality. For our experiments, we used the FID implementation by PyTorch Lightning[8].

The second aspect of our evaluation is how well the generated images follow the image descriptions. For this, we evaluate two different metrics. The first metric is Contrastive Language-Image Pre-training (CLIP), which is trained to determine if an image and a text are paired together (Radford et al., 2021). It encodes the images and texts into a joint embedding space and selects the most probable pairs among them. For our experiments, we use the pre-trained CLIP ViT-L/14@336px model that achieves the highest accuracies according to the authors (Figure 10 in Radford et al. (2021)).

Our third metric, TIFA (Hu et al., 2023), also evaluates the fit between an image and its description, similar to the CLIP score, but chooses a different approach: visual question answering. For this, Hu et al. (2023) created a three-step pipeline: First, an LLM creates single-choice, multiple-choice, and free-form questions and their answers from the image descriptions. Each question is categorized by the elements it is asking for, e.g., color or location, and the number of questions and the element types vary among the different images. Then, a second question-answering model tries to answer the questions based on the image descriptions. Only questions that receive the same answers from both systems are kept for visual evaluation. Finally, a visual question-answering model answers the questions by looking at the images. The image-based accuracy of the answers indicates how faithful the image is to the image description. TIFA incorporates the accuracy metric, and thus, the scores range between 0 and 1. The authors show that TIFA has a much higher correlation with human judgments than previous metrics like CLIP (Hessel et al., 2021).

For our study, we use the pre-trained checkpoints for the different parts of the pipeline. For the question generation, we use the author's fine-tuned Llama2 (Touvron et al., 2023) model. For the question filtering, we use a UnifiedQA (Khashabi et al., 2020) model. The set of questions was only created once per image prompt and then used for all model evaluations. With this, we reduce biases in the scores that could come from non-determinism in the question generation or filtering models. Finally, for the visual question-answering, the authors compared different models. We selected the model with the highest correlation with human judgment, according to the authors, which is mPLUG-large (Li et al., 2022).

4 Results

To obtain a diverse image collection, we created up to four images per model an prompt. We investigate seven different models, and thus, expected to generate $80 \times 4 \times 7 = 2,240$ images. How-

[5] https://www.midjourney.com/ Last accessed: Jul 2024

[6] https://www.artbreeder.com/ Last accessed: Jul 2024

[7] https://www.bing.com/images/create Last accessed: Jul 2024

[8] https://lightning.ai/docs/torchmetrics/stable/image/frechet_inception_distance.html

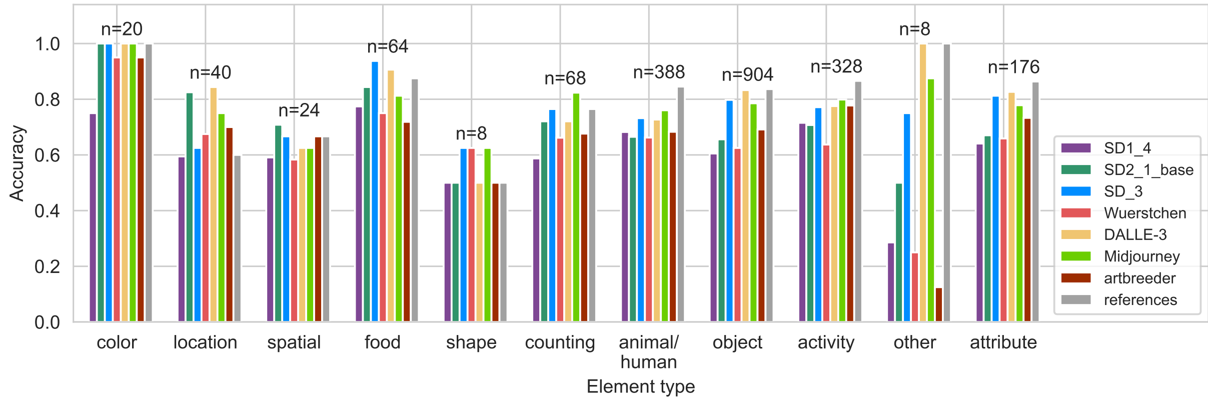


Figure 3: TIFA accuracies for different element types targeted by the TIFA questions. The performance of the models depends on the content that they have to depict. Images with all black content were filtered.

ever, we only obtained 2,217 images in total. For some descriptions, Copilot's DALL-E interface returned less than four images, resulting in only 297 instead of 320 images from DALL-E. In addition, Stable Diffusion 1 and DALL-E marked some of our image descriptions as offensive and blocked the input. For Stable Diffusion, this resulted in images that were all black. We followed this approach and added four black images for each of the five blocked inputs for DALL-E as well. This results in 51 images that are blackened.

4.1 Automatic evaluation

To further assess our generated images, we calculated metric scores as shown in Table 1. The best FID scores are achieved by Stable Diffusion 3 and Midjourney. This indicates that their style of images comes closest to the style of the reference image and that the images have similar features.

Model	FID↓	CLIP↑	TIFA↑
SD1_4	1.49	0.22	0.58
SD2_1_base	1.37	0.24	0.68
SD_3	**0.89**	**0.27**	**0.78**
Würstchen	0.90	0.24	0.65
DALL-E-3	1.26	**0.26**	0.74
Midjourney	**0.90**	**0.26**	**0.78**
Artbreeder	1.52	0.25	0.70
References	-	0.27	0.84

Table 1: Macro-averaged automatic evaluation scores to evaluate the images' distribution compared to the references (FID) and their closeness to the prompts (CLIP and TIFA). Stable Diffusion 3, DALL-E-3, and Midjourney come closest to the human-created reference images.

The CLIP and TIFA metrics evaluate how well the images align with the image descriptions. These metrics don't rely on the reference images, and hence, we calculated the scores on the references as well. We manually set the scores to 0 for all-black images with blocked contents and ignored one image title in the CLIP evaluation whose prompt was too long for the CLIP model. For both metrics, Stable Diffusion 3 has the highest scores, performing on par with the references according to the CLIP score. The other open-source models fall far behind in terms of automatic scores. For the closed-source models, Midjourney performs best, closely followed by DALL-E-3. However, none of the models can match the TIFA accuracies of the reference images. Interestingly, even the human-created images don't achieve perfect accuracy. Yet, this could be due to the shortcomings of the models in the TIFA pipeline.

The TIFA score is based on visual question answering, and the questions are categorized into the different elements that are evaluated. To dig deeper into the strengths and weaknesses of the T2I models, Figure 3 shows the models' TIFA accuracies per element type. Most of the questions are targeted toward animals or humans, activities, and especially the objects depicted. The reference images (grey bars) outperform the image generation models, especially for animals/humans, activities, and attributes. This aligns with our assumption that body parts and movements are the hardest aspects for the models to generate. In contrast, almost all models outperform the reference images in terms of location and shape, and Stable Diffusion, as well as DALL-E, outperforms the references in the food category.

Model	Prompt coherence↑	Correctness↑	Bias↓	Suitability↑
SD1_4	0.48 (± 0.80)	0.19 (± 0.45)	**0.00** (± 0.00)	0.06 (± 0.29)
SD2_1_base	0.50 (± 0.75)	0.16 (± 0.51)	**0.00** (± 0.00)	0.06 (± 0.29)
SD_3	**1.48** (± 0.98)	**0.90** (± 0.94)	0.14 (± 0.61)	**0.51** (± 0.83)
Würstchen	0.76 (± 0.82)	0.46 (± 0.84)	**0.00** (± 0.00)	0.20 (± 0.51)
DALL-E-3	**2.23** (± 0.91)	**2.19** (± 0.96)	0.21 (± 0.74)	**1.85** (± 1.12)
Midjourney	2.06 (± 0.88)	1.99 (± 0.88)	0.09 (± 0.48)	1.20 (± 1.11)
Artbreeder	1.25 (± 0.88)	1.05 (± 0.99)	**0.01** (± 0.11)	0.39 (± 0.68)

Table 2: Results from our human evaluation. The scores range from 0-3 and are averaged across all generated images. SD_3 and DALL-E created the most accurate and most suitable images.

4.2 Human evaluation

While TIFA scores have a high correlation with human judgments (Hu et al., 2023), automatic metrics can't cover all evaluation aspects. Especially for the overall correctness and simplicity of the images, there is currently no metric available. Therefore, we added a human evaluation of our generated images. For this, we asked an expert for German Easy language (one of the authors) to manually review and rate the images. Images are an essential part of German Easy language (DIN-Normenausschuss Ergonomie, 2023), and many Easy language courses also address criteria for selecting appropriate images. To reduce the overall workload, we selected one image per model and title, resulting in a dataset of 560 images. For each combination, we selected the image with the highest TIFA score. If two or more images shared the highest score, we sampled an image from among them. The images were evaluated on four different scales by asking these questions:

- *Does the image follow the prompt?*: This question checks for missing or additional content. We only focused on relevant content and ignored aspects that did not affect the meaning of the image (e.g., the prompt describing a group of nine people, but the model only drew seven).
- *Is the image correct?*: This question evaluates if the depicted content aligns with world knowledge, e.g., that people don't have three arms.
- *Does the image exhibit a bias towards people with disabilities?*: This question is targeted towards the findings of previous work (Mack et al., 2024; Tevissen, 2024) and evaluates whether the models tend to show people with disabilities as old or unhappy, even if the prompt does not define that.
- *Is the image suitable for the target group?*: For the target group, it is important that the images are not overloaded with details, text, or colors and that they align with situations familiar to the target group. These criteria are in line with the DIN SPEC for German Easy language (DIN-Normenausschuss Ergonomie, 2023). In addition, this question checks whether the image is helpful to understand the original concept.

The human annotator could choose between four possible answers to the questions: no/indeterminable, partly, mostly, and yes. We mapped these answers to a numerical scale between 0 (answer no) and 3 (answer yes). The images were blinded, i.e., we only showed the annotator the images and the descriptions but not the name of the model that generated the image.

Table 2 shows the averaged scores from the human evaluation. While Stable Diffusion 3 outperformed the closed-source models in the automatic evaluation, it can not hold up to the expectation in the human evaluation, receiving significantly worse scores across all scales. Still, it is by far the best open-source model. The bad scores for the other open-source models are mostly due to unclear and indeterminable content. Remarkably, these open-source models show the least biases. However, this is an artifact from our evaluation setup: If the image does not show any depictable content, then it also can't show biases toward people with disabilities.

During our manual review, we made additional observations. Examples of them are depicted in Figure 4. The models sometimes hallucinate additional details. For example, one of the prompts is "Cartoon picture of Security - Depicted are a

(a) Cartoon picture of **Security** - Depicted are a woman and a man lovingly embracing a child.
Created by SD3

(b) Cartoon picture of **Security** - Depicted are a woman and a man lovingly embracing a child.
Created by DALL-E-3

(c) Cartoon picture of **Fear** - A woman covers her mouth with both hands. Her eyes and mouth are wide open. Sweat runs down her forehead.
Created by SD3

(d) Cartoon picture of **Refusal of Physical Contact** - A woman tries to hug a girl. The girl looks away and resists.
Created by Midjourney

Figure 4: Example prompts and generated images.

woman and a man lovingly embracing a child.". Many models draw a policeman or officer, even though the prompt does not describe any (see Figure 4a). This indicates that the models have an inherent interpretation of world knowledge and, thus, associate security with police (Fu et al., 2024).

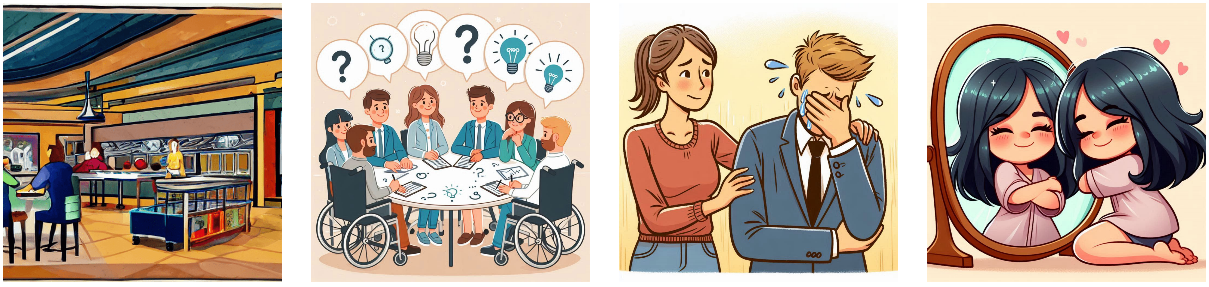
The biggest issues with the generated images arise with body parts and human motions. Examples are presented in Figures 4a, 4c and 4d where body parts such as arms or legs are missing, or too many fingers were added. Another issue is that the models don't pay enough attention to small details, and thus, important aspects are missing. For example, for the prompt "Cartoon picture of Refusal of Eye Contact - A woman stands directly in front of a man and speaks to him. The man has his arms crossed in front of his chest. He does not look at her.", all models created two people standing in front of one another, but no model could depict the refusal of eye contact properly. This could also be an issue with input token limits, i.e., that this important information was truncated. In addition, missing or misinterpreted details can change the meaning of the image. The images in Figures 4c and 4d should show the emotions of fear and the refusal of physical contact. However, in both pictures, the people look rather angry and as if they would fight one another. Especially for people struggling with reading emotion from human expressions, this could evoke wrong associations. Therefore, such images are not suitable for the target group without further restrictions.

As indicated by the low bias scores in Table 2, the models exhibit hardly any bias toward people with disabilities. The biases that we find are mostly related to hearing or vision impairments, where models tend to add an eye fold to visualize that a person is blind or draw incorrect hearing aids that look more like headsets. None of the models depicted people with disabilities as especially unhappy, except if the prompt especially stated it. On the contrary, most of them were smiling and happy.

The model with the best human evaluation scores is by far DALL-E-3. It was able to create correct images even for difficult body positions like in Yoga or hugging. In addition, the images were especially inclusive in terms of diversity: Pictures with multiple people often depicted people of color or people with glasses as parts of the groups. An example is Figure 4b, where the woman wears a head scarf, a garment only seen in minority groups in Western countries. These features were not described in the image prompts but added by the model and its world knowledge.

4.3 Feedback from the target group

In line with the UN inclusion slogan "Nothing about us without us!" (Harpur and Stein, 2017) and in accordance with the DIN SPEC recommendation that the target group should review all content, we wanted to hear the opinion about the images from the target group. Therefore, we invited seven people with different disabilities (physical, mental, and combinations of both) between the ages of 21 and 42 for a workshop at the university. They were accompanied by their living assistants and two German Easy language experts. The study participants received a compensation of 32,50€ for their effort. We conducted two types of studies: comparative voting and a free-form discussion. Direct quotes



(a) Cafeteria
Created by SD1

(b) Brainstorming
Created by DALL-E-3

(c) Empathy
Created by DALL-E-3

(d) Self-love
Created by DALL-E-3

Figure 5: Term-guessing images. The images were shown to the target group participants, and they had to guess the title or describe what the word could mean if they didn't know it.

from the participants are formatted in italic.

For the first part, we filtered the titles from the image dataset, where at least three models created suitable or mostly suitable images. Then, we selected seven titles from them. We presented all images with the same title at the same time but in a randomized order. The participants were asked to vote for all images they liked. We allowed multiple selections to account for equally good images. We used the voting platform Mentimeter[9] where every participant can participate on their smartphones and submit their votes anonymously. With this, we could collect their opinions independently without being influenced by other participants. Some participants were supported by their living assistants when interacting with the smartphones. The images and the participants' votes are presented in Table 4 in the Appendix.

For most of the images, DALL-E received the most votes, often even more than the reference images. In general, colorful images with few additional or decorative content were preferred. Some images seemed a bit abstract, e.g., the fruit depicted for the bowl of vitamins had a few mistakes or weird coloring. Nevertheless, the participants showed great creativity when naming the fruits. Therefore, if the context is clear, small mistakes don't bother too much. This also came clear when participants explained why they chose certain images for the bedroom: they chose the ones that looked the most "*cozy*". In contrast, some toothbrushes in the hygiene product images looked unrealistic, not suitable for the teeth, or were simply wrong. Multiple participants were distracted by these mistakes and started a discussion about what was wrong with the toothbrushes and how it would hurt to actually use them.

[9] https://www.mentimeter.com/

The second part of the workshop was more open to direct feedback. We presented four different images, one from Stable Diffusion 1 and three from DALL-E-3, and the participants should guess the depicted content. The images and their titles are shown in Figure 5. We chose some complex terms on purpose to assess whether the images could help with understanding them. For example, the word "Brainstorming" was unfamiliar to half of the participants. Nevertheless, they could describe and explain the word as "*people are sitting together and collect ideas*" only based on the image. This shows that the selected images are not only suitable for illustrating texts, but they also fulfill their purpose of explaining complex terms with ease.

In addition to the term guessing of the second part, the participants were also invited to express their thoughts and opinions about the presented images. We try to summarize them in the following:

- Participants did not like black&white-only images because "*it makes you depressed*".
- Different illustrations of the same objects (e.g., the light bulbs in the brainstorming image) were confusing, and participants tried to find a reason for the differences, even if there was no reason for that.
- The images should show accessible situations, i.e., suitable for people using wheelchairs or hearing aids. We had a discussion about whether the counter in SD1's cafeteria was accessible for wheelchairs and whether people would need help to reach all the offers. For this discussion, the blurry and abstract style of the image was of minor relevance.
- In the DALL-E-3 image for yoga (Table 4 in Appendix), the person using the wheelchair is very old. When the two participants who

used a wheelchair were asked whether they felt discriminated by this, they answered "*No, why should I? Even old people can do yoga!*"

5 Conclusion

In this paper, we have explored whether text-to-image models can be utilized to create illustrations for easy-to-read texts. For this, we evaluated the generated images in large-scale human studies, including seven participants from the target group. Closed-source models like DALL-E-3 and Midjourney and the open-source Stable Diffusion 3 have shown impressive performance in creating these images, sometimes creating even more favorable images than the gold-standard references. However, their performance highly depends on the depicted content, and the models struggle with difficult postures and specific body parts especially. Therefore, they cannot be used without human oversight or multiple iterations of image description optimization. In addition, the best-performing models are closed-source or very large in parameter size, meaning that text creators will still have to pay for their images. Since text creators will use the models, they have human expert oversight, and thus, every generated image will be reviewed, and erroneous images can be filtered before being shown to the target group. Finally, we believe that T2I models are especially suited for accessible communication due to their fast availability and options for tailored, customizable, and copyright-free content.

In future research, we would like to get rid of the intermediate step of explicit image descriptions and hope to see models that can create the images directly from the text paragraph. In addition, we would like to investigate their compliance with prompts in German and other non-English languages and investigate conditioning the images on the reference images during generation.

Limitations and ethical considerations

Our work presents a quite extensive comparison of different T2I images. While we did our best to include as many models with different architectures, sizes, and availabilities, we can only test the models published by the time of writing this paper. The current developments and improvements in AI are rapid, and thus, there may be newer and better models soon that we couldn't include in our study.

We tried to design this study as participatory as possible and included seven people from the target group in our human evaluation. They received 32,50€ to compensate for their effort. Nevertheless, the feedback session was moderated, and the authors pre-selected the images. A target group evaluation of all images would be infeasible and not of any help to the target group. Still, our image selection and moderation introduced a bias from the authors on them that we can not neglect. In addition, the disabilities and needs of the target group are very diverse and cannot be represented by only seven people. Nevertheless, we try to make their opinions be heard and invite all researchers in the area of accessible communication to work together with the target group.

Finally, we are aware that generative AI, whether it generates text, images, or any other modality, is being criticized for threatening jobs and content quality. The goal of our work is in no way to replace humans in the process of creating accessible content. However, we believe that the benefits of the short-time availability of simplified texts and images are important to overcome information barriers, especially on the internet. Studies such as ours can be of great help to further improve the quality of those models and to align their objectives with what is actually needed by the target group. In the end, our investigations show that the T2I models are far from being perfect and still need careful human oversight. Especially in terms of image evaluation, we could not find an automatic metric that was satisfactory in alignment with our judgment.

Lay Summary

Creating texts that are easy to read and understand is important for people with disabilities, learning difficulties, or those who have trouble with reading. These easy-to-read (E2R) texts often include pictures to help explain the information. However, it can be hard to find images that fit the specific needs of each text. Hiring artists to make customized images can be expensive, and existing image databases don't allow for easy changes to match the content of the text.

Our study looks at whether we can use artificial intelligence (AI) to generate these images quickly and cheaply. We tested seven different AI tools, called text-to-image models, which create pictures based on written descriptions. Some of these tools are open to the public, while others are not. We wanted to see if these AI-generated images could be a good solution for E2R creators.

We evaluated over 2,000 images created by these models and manually reviewed 560 of them. During the review, we looked at how well the images matched the description, if they were accurate, if they had any bias against people with disabilities, and if they were useful for the target group. Our results show that while some models produced high-quality images, none of them are ready to be used on a large scale without human oversight.

We also conducted a user study with seven people from the E2R target group to gather feedback on how well the images met their needs. It is important to include the target group and their opinions and preferences when doing research. The feedback was helpful in identifying areas where the AI models worked well and where they fell short.

Our research is an important first step toward making it easier and more affordable to create images that help make information more accessible. However, more improvements are needed before these AI tools can fully replace human involvement in creating custom images for E2R texts.

Acknowledgments

The images were provided by the Landesarbeitsgemeinschaft Selbsthilfe von Menschen mit Behinderungen und chronischen Erkrankungen Rheinland-Pfalz e.V. with the kind support of AOK Rheinland-Pfalz/Saarland. We thank them for their assistance that made our research possible in the first place!
In addition, we thank the *FÜP - FortSchritt Übersetzungs- & Prüfbüro für Leichte Sprache* (office for German Easy language translation and evaluation) by the FortSchritt Verein zur Verbreitung der Konduktiven Förderung e.V.[10] and especially their review group for the insightful feedback.
This research has been funded by the German Federal Ministry of Education and Research (BMBF) through grant 01IS23069 Software Campus 3.0 (Technical University of Munich) as part of the Software Campus project "LIANA".
This paper is based on a joined work in the context of Tringa Sylaj's master's thesis (Tringa Sylaj, 2024).

[10] https://www.fortschritt-bayern.de/angebote/leichte-sprache

References

Miriam Anschütz, Joshua Oehms, Thomas Wimmer, Bartłomiej Jezierski, and Georg Groh. 2023. Language models for German text simplification: Overcoming parallel data scarcity through style-specific pre-training. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1147–1158, Toronto, Canada. Association for Computational Linguistics.

Shane Barratt and Rishi Sharma. 2018. A note on the inception score. *Preprint*, arXiv:1801.01973.

Andrew Brock, Jeff Donahue, and Karen Simonyan. 2019. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*.

Maitraye Das, Alexander J. Fiannaca, Meredith Ringel Morris, Shaun K. Kane, and Cynthia L. Bennett. 2024. From provenance to aberrations: Image creator and screen reader user perspectives on alt text for ai-generated images. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

DIN-Normenausschuss Ergonomie. 2023. Empfehlungen für Deutsche Leichte Sprache (DIN SPEC 33429).

Xingyu Fu, Muyu He, Yujie Lu, William Yang Wang, and Dan Roth. 2024. Commonsense-t2i challenge: Can text-to-image generation models understand commonsense? In *First Conference on Language Modeling*.

Robert Geislinger, Ali Ebrahimi Pourasad, Deniz Gül, Daniel Djahangir, Seid Muhie Yimam, Steffen Remus, and Chris Biemann. 2023. Multi-modal learning application – support language learners with NLP techniques and eye-tracking. In *Proceedings of the 1st Workshop on Linguistic Insights from and for Multimodal Language Processing*, pages 6–11, Ingolstadt, Germany. Association for Computational Lingustics.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM*, 63(11):139–144.

Paul Harpur and Michael Ashley Stein. 2017. The convention on the rights of persons with disabilities as a global tipping point for the participation of persons with disabilities. In *Oxford Research Encyclopedia of Politics*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Y. Hu, B. Liu, J. Kasai, Y. Wang, M. Ostendorf, R. Krishna, and N. A. Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20349–20360, Los Alamitos, CA, USA. IEEE Computer Society.

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

Emory James Edwards, Kyle Lewis Polster, Isabel Tuason, Emily Blank, Michael Gilbert, and Stacy Branham. 2021. "that's in the eye of the beholder": Layers of interpretation in image descriptions for fictional representations of people with disabilities. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '21, New York, NY, USA. Association for Computing Machinery.

Tero Karras, Samuli Laine, and Timo Aila. 2021. A style-based generator architecture for generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(12):4217–4228.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.

Johannes Kiesel, Çağrı Çöltekin, Maximilian Heinrich, Maik Fröbe, Milad Alshomary, Bertrand De Longueville, Tomaž Erjavec, Nicolas Handke, Matyáš Kopp, Nikola Ljubešić, Katja Meden, Nailia Mirzakhmedova, Vaidas Morkevičius, Theresa Reitis-Münstermann, Mario Scharfbillig, Nicolas Stefanovitch, Henning Wachsmuth, Martin Potthast, and Benno Stein. 2024. Overview of Touché 2024: Argumentation Systems. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 15th International Conference of the CLEF Association (CLEF 2024)*, volume 14959 of *Lecture Notes in Computer Science*, pages 308–332, Berlin Heidelberg New York. Springer.

Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, He Chen, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, and Luo Si. 2022. mPLUG: Effective and efficient vision-language learning by cross-modal skip-connections. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7241–7259, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kelly Avery Mack, Rida Qadri, Remi Denton, Shaun K. Kane, and Cynthia L. Bennett. 2024. "they only care to show us the wheelchair": disability representation in text-to-image ai models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Margot Madina, Itziar Gonzalez-Dios, and Melanie Siegel. 2023. Easy-to-read language resources and tools for three european languages. In *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '23, page 693–699, New York, NY, USA. Association for Computing Machinery.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, and et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Pablo Pernias, Dominic Rampas, Mats Leon Richter, Christopher Pal, and Marc Aubreville. 2024. Würstchen: An efficient architecture for large-scale text-to-image diffusion models. In *The Twelfth International Conference on Learning Representations*.

Ben Proven-Bessel, Zilong Zhao, and Lydia Chen. 2021. Comicgan: Text-to-comic generative adversarial network. *Preprint*, arXiv:2109.09120.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. 2016. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Florian Schneider, Özge Alaçam, Xintong Wang, and Chris Biemann. 2021. Towards multi-modal text-image retrieval to improve human reading. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, Online. Association for Computational Linguistics.

Thorben Schomacker, Michael Gille, Marina Tropmann-Frick, and Jörg von der Hülls. 2023. Data and approaches for German text simplification – towards an accessibility-enhanced communication. In *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*, pages 63–68, Ingolstadt, Germany. Association for Computational Linguistics.

Janvijay Singh, Vilém Zouhar, and Mrinmaya Sachan. 2023. Enhancing textbooks with visuals from the web for improved learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11931–11944, Singapore. Association for Computational Linguistics.

Regina Stodden, Omar Momen, and Laura Kallmeyer. 2023. DEplain: A German parallel corpus with intralingual translations into plain language for sentence and document simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16441–16463, Toronto, Canada. Association for Computational Linguistics.

Yannis Tevissen. 2024. Disability representations: Finding biases in automatic image generation. *Preprint*, arXiv:2406.14993.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Tringa Sylaj. 2024. Image generation for accessible communication. Master's thesis, Technical University of Munich. Advised and supervised by Miriam Anschütz and Georg Groh.

Xintong Wang, Florian Schneider, Özge Alacam, Prateek Chaudhury, and Chris Biemann. 2022. MOTIF: Contextualized images for complex words to improve human reading. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2468–2477, Marseille, France. European Language Resources Association.

A Appendix

Model	Prompt Limitation	Flagging Content	Resources Used
SD1_4	77 tokens	Black image	T4 GPU, 15GB RAM
SD2_1_base	77 tokens	Black image	T4 GPU, 15GB RAM
SD_3	77 tokens	Black image	L4 GPU, 24GB RAM
Würstchen	77 tokens	Black image	T4 GPU, 15GB RAM
DALL-E-3	380 characters	Not processed + warning	Free via Microsoft
Midjourney	None	N/A	\$10/month for $\approx$ 200 images
Artbreeder	$\sim$ 129 tokens	N/A	Free with multiple accounts

Table 3: Comparison of the different models we investigated: Limitations, content flagging, and resource usage

Model	**Multi-family House**	**Bedroom**	**Vitamins**	**Poor Memory Performance**	**Hygiene Products**	**Yoga**	**Cafeteria**
SD_3	3 	4 	4 	**6** 	-	1 	-
Wuerstchen	2 	-	3 	-	-	-	-
DALLE-3	**6** 	**6** 	4 	**6** 	4 	**2** 	**4** 
Midjourney	4 	5 	6 	**6** 	**5** 	**2** 	**4** 
artbreeder	5 	1 	3 	-	4 	-	-
References	4 	3 	**7** 	5 	4 	**2** 	2 

Table 4: Number of votes from the target group during their review session. We only included images that the German Easy language expert deemed suitable for the target group. Thus, no images from Stable Diffusion 1 and 2 were shown.