

Tailoring LLM-generated image captions to user needs

Ivonne van der Heiden*
Emiel van Miltenburg*

I.H.VDRHEIDEN@TILBURGUNIVERSITY.EDU
C.W.J.VANMILTENBURG@TILBURGUNIVERSITY.EDU

*Tilburg University, Warandelaan 2, 5037 AB Tilburg, The Netherlands

Abstract

One of the original motivations for the development of image captioning systems is to make visual content accessible for people who are blind or visually impaired. What seemed like a huge challenge fifteen years ago, has now made it into consumer products: large language models such as ChatGPT are seemingly able to describe images in fluent natural language. But it is still unclear to what extent the generated descriptions actually match user needs. This study investigates the quality of LLM-generated image descriptions in the context of Dutch news articles. We operationalise output quality based on earlier user studies and existing image description guidelines, and present an extensive evaluation protocol that may be used in future research to assess the quality of automatically generated image descriptions.

1. Introduction

Automatic image description (also known as *image captioning*; Hodosh et al. 2013, Bernardi et al. 2016) has now been around as a Vision & Language task¹ for about fifteen years. Like many tasks that came before, it started out as a toy example to see what is possible with current technology. We then slowly started to see real-life applications emerge, such as automatically generated image descriptions on social media (Wu et al. 2017). But although different Human Computer Interaction (HCI) studies were carried out to identify needs and requirements of people with visual impairments (e.g. Stangl et al. 2020), research in Natural Language Processing (NLP) and Computer Vision (CV) remained focused on generic metrics to assess output quality, leading to a disconnect between HCI and NLP/CV (van Miltenburg 2020). In short: current evaluation metrics ignore quality dimensions that have been identified in the HCI literature. This paper addresses this issue by presenting an evaluation procedure that does take these quality dimensions into account.

1.1 Multilingual Multimodal Language Models

Recent years have seen the field move away from task-specific models to more task-agnostic solutions through the use of Large Language Models (LLMs). Although these LLMs were initially only unimodal (text), many models now support other modalities such as images, audio, and video. We may refer to these models as Multimodal Language Models (MLMs; see Yin et al. 2024 for a survey). While these models were often monolingual (usually English), we are also seeing more and more Multilingual Multimodal Language Models (MMLMs; see Erdem et al. 2022 for an early survey describing this emerging field).² The idea behind these models is that knowledge gained from different languages or modalities may help improve model performance for lesser resourced languages, and contribute to the overall performance of the model. This paper reports on the performance of one commercial MMLM (ChatGPT 4o), and several smaller Vision and Language Models (VLMs; see Li et al. 2025 for a survey).

1. A multimodal task combining Computer Vision and Natural Language Processing. Another example is Visual Question Answering (VQA; Antol et al. 2015).

2. Some benchmarks for these models are M3EXAM (Zhang et al. 2023) and EXAMS-V (Das et al. 2024).

1.2 Contextualised Dutch image descriptions

We focus on *contextualised* image descriptions, meaning that we use images that are part of a larger message (in our case: photographs in online news articles). The task of the image description system is thus to make sense of the image, and provide relevant information, given the information that was already provided by the text. Where the news article helps to disambiguate the image, the image description system should also provide more concrete and informative descriptions. Our focus on contextualised image descriptions may be contrasted with traditional image description approaches, where datasets consist of decontextualised images where the annotators provided descriptions without any supporting information (e.g. MS COCO, Lin et al. 2014 and Flickr30K, Young et al. 2014). The problem with decontextualised images is that they may either lead to generic descriptions, or to speculative descriptions making unwarranted inferences (van Miltenburg et al. 2017).

We also focus on *Dutch* image descriptions, investigating to what extent existing models are able to support languages that are less dominant than English. A recent survey estimates that image description datasets exist for about 42 different languages (van Miltenburg 2024). Although Dutch is on that list, the amount of available data is relatively small.³ Now in theory this doesn't need to be a problem for MMLMs: as long as the model is able to understand images through its exposure to other multimodal data, it might be able to develop an understanding of its visual input and verbalise this understanding in Dutch. But in practice, the MMLM may never be exposed to Dutch entities and concepts that are specific to Dutch culture, meaning that the model will not be able to identify and name those entities and concepts.⁴ So our main research question is: to what extent can LLMs (and ChatGPT in particular) provide contextualised Dutch image descriptions, that are tailored to the needs of blind or visually impaired users?

1.3 Precedent

This study builds on earlier work by Gubbi Mohanbabu and Pavel (2024). In their work, the authors used ChatGPT to generate image descriptions for 24 different images. These images originated from four different kinds of websites (e-commerce, news, social media, and blogs). The authors compared context-free and contextualised descriptions of the same images, and found that visually impaired users prefer the latter because they were usually more specific (e.g. saying 'Himalaya' instead of 'Mountain range'). All participants also reported that they wanted to continue using the system in the future. Despite their successful experiment, Gubbi Mohanbabu and Pavel (2024) also observe some limitations with their approach. One notable limitation that we address in this paper is the *one-size-fits-all* approach that still pervades research on image captioning. Ideally, the authors note, we should tailor image descriptions to the reader, adapting "tone and terminology to the audience to match their knowledge and interests." So next to the overall quality of the descriptions, we will also look at *readability* and investigate the extent to which ChatGPT is able to generate accessible image descriptions for four different groups of image description users:

1. Default users who prefer short descriptions in Standard Dutch.
2. Default users who prefer longer descriptions in Standard Dutch.
3. Low-literacy users who prefer short descriptions that are easy to read.
4. Low-literacy users who prefer longer descriptions that are easy to read.

To assess ChatGPT's capacity to generate readable texts, we rely on LiNT, an automatic metric that was designed to measure the readability of Dutch texts (Pander Maat et al. 2023).

3. Dutch-language image captions are available in the following datasets: Crossmodal-3600 (Thapliyal et al. 2022), KBK-1M (Elliott and Kleppe 2016), DIDEK (van Miltenburg et al. 2018), Flickr30K-val/test (van Miltenburg et al. 2017).

4. See Hershcovich et al. (2022) and Liu et al. (2021) for further discussion of this issue.

1.4 Contributions of this paper

This paper presents an evaluation of an LLM (specifically ChatGPT 4o, an MMLM), assessing its ability to generate Dutch image descriptions in the context of Dutch news articles. In contrast to Gubbi Mohanbabu and Pavel’s broader sample, we analyse 30 images from the same genre. This allows for a more thorough evaluation of description quality in this domain. Moreover, we go beyond the ‘one-size-fits-all’ approach by investigating four different kinds of image descriptions. In contrast to much of the image captioning literature, our approach is based on user studies and existing image description guidelines. We also briefly look at the performance of several other LLMs, but find that the quality of their output is still insufficient.

2. Background

Before discussing our approach, we will first discuss existing image description guidelines and different user studies that look into the image description needs of people who are blind or visually impaired. These sources are used to inform our evaluation procedure.

2.1 Image description guidelines

There are too many different image description guidelines to summarise in this paper.⁵ To our knowledge, DIAGRAM Center (2015) provides the most extensive and authoritative set of guidelines. According to these guidelines, an image description must always be objective. The guidelines also state that a description should not contain text that already appears in an accompanying article. Naming colors can improve an image description according to these guidelines, but only if the colors in the image are relevant. Finally, an image description according to these guidelines should begin with a general description of the image and then become more specific (DIAGRAM Center 2015). Complementary to the DIAGRAM Center guidelines, the Perkins School for the Blind provides a set of guidelines that were “written by someone with low vision who uses alt text” (Lewis 2018). These guidelines (paraphrased below) distinguish eight important properties that should be specified:

1. Specify the type of image. For example, whether it is a cartoon, a photograph, or a graph.
2. Provide all text from the image in the image description.
3. For images with people, there are two possibilities: the people are known or not. For images with famous people, simply mentioning their name is sufficient, unless their appearance differs significantly from their normal appearance; in that case, an additional description is helpful. For unknown people, it is important to describe them, for example, by mentioning their ethnicity, gender, hair color, and age. (Though see Section 2.2 for important caveats.)
4. It can be helpful to describe the emotions of the depicted people.
5. It is important to describe the setting of the image to clarify the location of the image. Here too, there are several options. If it is a well-known location or if it is mentioned in the accompanying text, the specific location can be mentioned. If the location is not generally known, the setting of the image should be described.
6. If colors are important, it is good to be specific in naming them, such as ochre yellow or cobalt blue. However, if colors are not important, simply naming a color type, such as yellow or blue, is sufficient. This guideline contrasts with that of the DIAGRAM Center, which states that colors should only be mentioned if they are relevant. According to Lewis (2018), colors should always be mentioned, and it is good to be specific about the colors if they are relevant.

5. In a currently unpublished study, the second author found 150 different guidelines available online.

7. For animals in an image, it is sufficient to identify the species.
8. Identifying interesting details may improve image descriptions, so it is important to also go beyond simple ‘checklists’ of things that should be mentioned in an image description.

It is important to note that these guidelines are not always in agreement: while the DIAGRAM guidelines emphasize objectivity, Lewis’ recommendations sometimes deviates from this, inviting a more personal perspective (for example: commenting on people’s emotions). In this paper, we do not take a stance about what makes the best description, but we simply borrow key dimensions from the guidelines to characterize the descriptions that are generated by ChatGPT.

2.2 User studies

When generating image descriptions, it is important to analyze whether they meet the needs of people with visual impairments. The information that people with visual impairments consider important in image descriptions varies depending on the type of image (Stangl et al. 2020). However, there are some regularities. For example, people generally prefer all text in an image to be included in the description. Furthermore, for every type of image, describers should use the image’s context, such as the text of an accompanying article, to provide relevant and *specific* descriptions (Gubbi Mohanbabu and Pavel 2024). The image’s context can be used to determine the names of people or locations in the image. It can also be used to align the tone of the description with the rest of the website, but in this study we only examined whether context is used to add named entities to descriptions.

Besides these universal elements in descriptions, there are also several image characteristics that indicate various aspects that need to be described. For images featuring people, it’s helpful to describe whether these people are known or not (Stangl et al. 2020). What else is relevant to mention about these people varies from image to image. Some people find that mentioning people’s ethnicity, for example, can sometimes help create a clear description (Stangl et al. 2020). However, other research shows that it can be harmful for people to read incorrect information in an image description, for example, about ethnicity or gender (Bennett et al. 2021). This leads some to argue that it’s never advisable to mention ethnicity or gender in an image description unless this information is known or verifiable (Bennett et al. 2021, Jandrey et al. 2024). According to Jandrey et al. (2024), mentioning ethnicity can be avoided by only mentioning visible characteristics of people, such as skin color, clothing, or hairstyle. This information can give people a better idea of who is in the image without making assumptions about ethnicity. However, mentioning skin color is also a point of contention. It is mentioned in the literature as an option for describing people (Hanley et al. 2021, Jandrey et al. 2024), but research also shows that accurately describing skin color is difficult because there is no objectively good description of someone’s skin color (Bennett et al. 2021). Mentioning gender in an image description can be avoided by describing people based on their clothing and accessories (Bennett et al. 2021). The content analysis for this study examines whether the descriptions contain references to ethnicity and gender, as well as descriptions of people based on their clothing and hairstyle.

For images depicting events or actions, people prefer to have them described based on the central people and objects. If objects or landmarks are visible in the image, people want the description to include information about what makes these objects or places unique (Stangl et al. 2020). Research into preferences for image descriptions of news articles shows that people with visual impairments consider the following elements important in these images: activities of people in the image, characteristics of people, such as clothing, and characteristics of the location (Stangl et al. 2021). Furthermore, it is helpful to clarify the purpose of the image in the article description (Stangl et al. 2020). This can be done, for example, by describing why this particular image was chosen or by describing how the image relates to the article.

In this study, image descriptions generated by ChatGPT will be manually analyzed. We will investigate whether the elements that people with visual impairment prefer in a description, as described above, are present.

3. Method

3.1 Overall design

We carried out a 2×2 experiment (description length \times difficulty level), aiming to determine the extent to which ChatGPT is able to provide usable image descriptions in the context of news articles. We used four different prompts to generate descriptions for images from published news articles, and analysed the resulting descriptions using a combination of automatic metrics, human judgments, and a detailed content analysis.

3.2 Data

We selected 30 news articles from the website of the Dutch Broadcasting Foundation (*Nederlandse Omroep Stichting*, NOS, the main public source of Dutch news). We focused our attention on images that provide additional information, complementing the contents of the news articles. For example, an image might show the impact of a flooding event.⁶ Articles with decorative images (e.g. stock photography) were discarded. In addition, we aimed to have a diverse collection of images, so that we could see how well the model handles different subjects (e.g. images with/without people, with/without text).⁷ All articles were collected between the 21st and the 27th of October 2024. We purposefully chose to use recent articles, so that they would not be part of the model’s training data. All descriptions were generated on the 9th or the 10th of November 2024.

3.3 Model

As noted earlier, we use ChatGPT 4o to generate descriptions for each of the images in our dataset. Although this is a closed model (making it impossible to study the properties of the system; see Rogers 2023 for discussion), we chose to use it because its performance was state-of-the-art at the time, and our goal was to demonstrate the current quality of the best available model. (Section 6.1 discusses our impression of alternative models.) Although the results of our study have become outdated already (at the time of writing, ChatGPT 5.2 is the most recent iteration of the model), our general approach is still useful to illustrate how these kinds of models should be assessed.

3.4 Prompts

For our study we used four different prompts, all consisting of four parts (translated for convenience):

1. Describing the general assignment: *your job is to generate image descriptions for images that belong to news articles. Every message that you receive contains an image and a document with the text of a news article.*
2. Specifying the task: *Generate an image description. Do not make any unfounded assumptions. Do not judge the contents of the image.* This part of the prompt has been developed by our societal partner (a publisher focusing on accessibility for blind and visually impaired users) in an internal pilot study. The instructions to avoid unfounded assumptions are based on earlier work by van Miltenburg (2016), which identified stereotypes and biases in image descriptions.

6. See e.g. Marsh and Domas White (2003) and Cheema et al. (2023) for discussions of different relationships between images and texts.

7. For this reason, we focused on the front page of the news website, which contains the most varied collection of images. By contrast, the sports and entertainment sections tend to center around people.

3. Condition-specific instructions: here we have two sets of specifications.
 - (a) Specifying the nature of the descriptions: either *the description should be easy to read for low-literacy individuals*⁸ (the SIMPLE condition) or *the description should be written in standard Dutch (the DEFAULT condition)*.⁹
 - (b) Optionally specifying the length requirements: *use at most 280 characters* (the SHORT condition; the absence of this requirement defines the UNBOUNDED condition).
4. Removing unnecessary additions: after some trials, we noticed that ChatGPT often tried to summarise the article or explain how the image relates to the article. Such overly helpful behavior should be avoided, because these kinds of inferences are best left to the readers themselves. Hence we added: *Use the text of the article for the description, but only describe the image. Say nothing about the relation between the image and the text.*

Appendix A provides the original Dutch prompts that we used for our study. Due to time constraints, we only carried out a limited amount of testing before the data collection phase. We also decided not to include explicit guidelines in our prompts, so we could study the default image descriptions that any user might receive after a simple prompt. Also note that a potential issue of incorporating image description guidelines into a prompt is that the prompt would become very long, which might also affect the performance of the model. The alternative that we have chosen here is to initially trust the defaults, and then in the future adjust the prompt where necessary.¹⁰ So while we cannot claim to have found the best performing prompts, at least this study provides a lower bound, demonstrating the quality of the descriptions that state-of-the-art LLMs are able to produce. We leave the exploration of different prompts for future research.¹¹

3.5 Assessment

We assessed the quality of the generated descriptions in three different ways: using automatic metrics (§3.5.1), through a rating task (§3.5.2), and with a manual analysis (§3.5.3).

3.5.1 AUTOMATIC METRICS

We automatically measured the number of characters, the number of words, and the readability of each description through the online LiNT interface (Pander Maat et al. 2023).¹² The LiNT score indicates the difficulty level of a text, ranging from 0 (very easy) to 100 (very difficult). Generally speaking, texts with a lower LiNT score should be easier to read for individuals with low literacy. However, as Lentz (2021) rightly points out:

- Many readability tools have not been thoroughly validated. LiNT itself was mostly tested with (secondary school) students, and it has not (extensively) been tested with adult readers.
- We need to be careful not to equate *readability* with *understandability*. The latter requires more than what current tools (are able to) measure. So texts with a low LiNT score are not necessarily easy to understand.
- There may still be variation *within* texts such that parts of the text are easier to understand than others. Readability tools that only produce one score obscure this issue.

8. We did not ask for *simplified* descriptions since earlier research shows that ChatGPT leaves out more relevant information when asked for a *simple* version, compared to an *understandable* version (Beks van Raaij et al. 2024).

9. We could have also chosen to entirely leave out the standard Dutch requirement, since the language is already implied by the prompt (as evidenced by the results for low-literacy individuals).

10. For a full discussion of prompt justification, see Calò et al. 2026.

11. One reviewer notes that other approaches could also be explored, such as instruction tuning (Zhang et al. 2026).

12. See: <https://lint.hum.uu.nl/home>

To complement the LiNT scores, we also collect human ratings to determine whether they find the image descriptions easy to read. These ratings will be discussed in §3.5.2.

Description length (i.e., number of words and characters) was measured to compare the SHORT and UNBOUNDED conditions, and readability was measured to compare the SIMPLE and DEFAULT conditions. These measurements serve two goals. First, we want to ensure that ChatGPT is indeed able to follow our instructions so that the descriptions can be reliably tailored to user preferences. Second, it is crucial for low-literacy individuals that the reading level is appropriate for them.

3.5.2 RATING TASK

We asked participants to rate the generated descriptions via an online questionnaire, that was implemented using Qualtrics.¹³ We set up the questionnaire such that each participant would see all of the descriptions for one news article. Thus, our task has a within-subjects design, where every participant is exposed to all conditions.

Participants We used on a convenience sample, recruiting volunteers from the personal network of the first author. Out of 47 respondents, 40 participants fully completed the study. Since the other seven participants filled out most of the questionnaire, we used their responses as well. The mean age of the participants is 32.7 years of age (SD=14.3). Most respondents (34%) received a university education (Dutch: *WO*), followed by higher professional education (30%; Dutch: *HBO*), secondary school (21%),¹⁴ vocational education (11%; Dutch: *MBO*), and other (4%).

Our respondents cannot be assumed to have any knowledge of image descriptions or visual impairment. This means that we cannot draw any conclusions about the preferences of our target group (image description users), but we can make observations about the general quality of the descriptions.

Procedure & Materials Prior to their participation, all participants read an information letter and provided their consent to take part in our study. Upon their consent, participants first read a randomly assigned news article and inspected the associated image, after which they rated the individual image descriptions for *correctness*, *replaceability*, *reading level*, and *completeness*.¹⁵

Our selection of the quality dimensions is based on the general needs of blind or visually impaired users. Correctness was measured to determine whether users could rely on the descriptions or not. We asked users to rate replaceability to determine whether the description could actually serve as a full replacement of the image for blind or visually impaired users.¹⁶ We asked our participants to rate the reading level of the descriptions so that we would have a complementary measurement of readability next to the LiNT scores.¹⁷ This is relevant for the subgroup of image description users with low literacy. Finally, we asked our participants to judge the completeness of each description so that we could identify any missing elements from the descriptions.¹⁸ We used 7-point Likert scales,¹⁹ where our participants could indicate the extent of their agreement with a set of statements (shown in Table 1).

Following these questions, participants were asked to indicate which description they thought was the best and which was the worst. They were also asked to provide a motivation for their answer. Finally, we asked the participants to rank all the descriptions in terms of their readability (from easy to difficult).

13. See: <https://www.qualtrics.com>

14. Presumably most of these are university students.

15. The descriptions were presented in a random order.

16. In a similar vein, Gubbi Mohanbabu and Pavel (2024) asked their participants about the *Imageability* of the descriptions, to see if their participants could form a mental image based on the description.

17. This offered some redundancy in case the online LiNT interface would not work for our data.

18. In a similar vein, Gubbi Mohanbabu and Pavel (2024) asked their participants about the *relevance* of the descriptions.

19. Based on prior research investigating the optimal scale size (Cicchetti et al. 1985, Preston and Colman 2000).

Dimension	Statement
Correctheid <i>Correctness</i>	De beschrijving is correct en bevat geen onwaarheden. <i>The description is correct and contains no falsehoods.</i>
Vervangbaarheid <i>Replaceability</i>	De beschrijving zou de afbeelding kunnen vervangen. Als ik alleen de beschrijving zou lezen, zou ik begrijpen wat er op de afbeelding te zien is. <i>The description could replace the image. If I only read the description, I would understand what the image shows.</i>
Taalniveau <i>Difficulty</i>	De beschrijving is in makkelijke taal geschreven en daardoor goed te begrijpen. <i>The description is written in simple language, thus easy to understand.</i>
Compleetheid <i>Completeness</i>	De beschrijving van de afbeelding is compleet, er ontbreekt dus geen informatie. <i>The description is complete, there is no missing information.</i> Als je bij de vorige vraag hebt aangegeven dat er informatie ontbreekt, welke informatie ontbreekt er volgens jou in de beschrijving? <i>If you indicated for the previous question that there is missing information, what information is missing from the description according to you?</i>

Table 1: Dutch statements for the rating task, with English translations in *italics*.

Ethical considerations Since this study was carried out in the context of a Master’s thesis (with concomitant time pressure), we did not apply for ethical clearance. Instead, we based the design of this study on earlier NLG rating tasks for which the second author did obtain ethical clearance (e.g. van Miltenburg et al. 2023; van Miltenburg et al. 2024). The study was designed to minimize the duration while still being able to answer our research question.

3.5.3 CONTENT ANALYSIS

In addition to the participants’ subjective assessments, a quantitative content analysis was also conducted. This was done to investigate whether ChatGPT is capable of generating image descriptions that meet the guidelines for image descriptions and the needs of people with visual impairments. Guidelines from the Perkins School for the Blind (Lewis 2018) and the DIAGRAM Center (DIAGRAM Center 2015) were used to assess the content of the image descriptions. Table 2 provides an overview of the dimensions we examined in our content analysis.

These dimensions were derived from the literature as follows:

- According to Perkins’ guidelines, eight aspects are important when creating an image description: image type, text, people, emotions, setting, colors, animals, and interesting details (Lewis 2018). For a further explanation of these elements, see Section 2.1. In the current study, only photographs were analyzed, so the first aspect, image type, was not included in the analysis.
- The DIAGRAM Center guidelines state that an image description must be objective and should not contain text already included in the accompanying article. Furthermore, an image description should begin with a general description and then include details (DIAGRAM Center 2015). These three aspects were considered when assessing the descriptions in the content analysis.
- In addition to using these guidelines, we also examined whether the image descriptions met the needs of people with visual impairments, as described in Section 2.2.

Text	Is all text in the image mentioned in the description?
Familiar persons	Is the person’s name mentioned in the description? If this person’s appearance deviates from normal, is their appearance described?
Unknown persons	Is this person described? Does the description include an estimate of the age of the person in the image? Does the description include the hair color of the person in the image? Does the description include a description of the clothing the person is wearing?
Gender	Is the gender of the person depicted mentioned?
Ethnicity	Is the ethnicity of the person depicted mentioned?
Activities	Are the activities of the persons in the image described?
Emotions	Are the emotions of the persons or the image described?
Animals	Is the species of the animal in the image mentioned?
Location	Is the location of the image described? For example, in the kitchen, on the highway, etc. If the article mentions the location of the image, is the location mentioned in the description?
Color	Are specific names given to colors?
Interesting details	Are interesting details in the image described?
Events	Is the image described based on central people and/or objects?
Purpose	Does the description clarify the purpose of the image?
Context	Is the context of the article incorporated into the description?
Objectivity	Is the description written objectively?
Article text	Does the description not contain text that is already verbatim in the article?
General to specific	Does the description begin generally and become increasingly specific?

Table 2: Elements of the content analysis.

This analysis was carried out by the first author in two stages. In the first stage, all descriptions were annotated using the questions from Table 2. These are all binary yes/no questions.²⁰ Next to this, the first author also took notes about any part of the descriptions that stood out. In the second stage, they went over all descriptions again to examine whether elements that stood out during the first analysis could also be found in other descriptions.²¹

4. Expectations

The main goal of this paper is to characterize the overall quality of automatically generated Dutch image descriptions. This means that we do not have any hypotheses regarding any user requirements that may or may not be fulfilled; we simply wish to describe the current state of affairs. Nevertheless, we do have some expectations regarding the general design of our study. Most of our hypotheses concern the basic validity of our study and the prediction that ChatGPT will be able to perform the task. Based on the prior literature and the general design of our study, we developed five hypotheses. Our first two hypotheses concern the length of the descriptions, which we investigate

20. These are all relatively objective questions; we left the more subjective judgments to our participants.

21. Note that participants from the rating task were *also* asked to provide their thoughts about the descriptions, providing some useful redundancy.

next to readability. We take description length to be an easier, more concrete target than readability, which also involves other factors such as language use and syntactic structure (Pander Maat et al. 2023).

4.1 Description length

Prior research by Gubbi Mohanbabu and Pavel (2024) has also investigated long versus short descriptions. In their study, the authors presented visually impaired users with short descriptions that could optionally be expanded into longer versions. The authors found that the participants hardly ever did so, suggesting that short descriptions might be good enough for their image description needs. If this is indeed the case, we should expect:

Hypothesis 1: Participants will generally prefer short descriptions over unbounded descriptions.

It also stands to reason that, all other things equal, shorter descriptions will be perceived as easier than longer descriptions, since there is less information to process. Thus we expect:

Hypothesis 2: Participants will rate shorter descriptions as easier than unbounded descriptions.

If this hypothesis turns out not to be true, it might be the case that the language model somehow tries to compress the same information in less space, leading to a greater complexity.

4.2 Readability

The next hypothesis is based on the assumption that ChatGPT will be able to understand our instructions to produce image descriptions that are easy to read. As discussed in our Method section (§3), we use a validated measure for readability (LiNT; Pander Maat et al. 2023), which produces a numerical score that corresponds to the difficulty level of the input text (higher is more difficult). Based on our assumptions and pilot testing, we expect:

Hypothesis 3: Simple descriptions will have a lower LiNT-score (i.e., are automatically judged to be more readable) than Default descriptions.

In addition to the automatic readability scores, we also collect human ratings of readability. Although these ratings are provided by laypeople, we do expect their intuitions to line up with the readability scores (Hypothesis 4). Moreover, since LiNT is usually used for longer texts, we do not know to what extent it works well for shorter texts like the image descriptions in our study. Of course, the underlying assumption behind hypothesis 4 is that the human ratings are valid (Hypothesis 5).

Hypothesis 4: The readability scores of our participants will match the LiNT-scores, such that descriptions with a higher readability rating will receive a lower LiNT-score.

Hypothesis 5: Participants will rate simple descriptions as easier to read than the default descriptions.

5. Results

5.1 First impressions

Our first impression of the generated descriptions is that ChatGPT does seem to be able to produce detailed contextualised Dutch image descriptions. The most striking example involving named entities comes from a local news article about an unknown woman whose Halloween-themed garden caused commotion in her village (NOS and Omroep West 2024). Despite the fact that ChatGPT had never seen this person before, it was able to incorporate their name in the description.

Example

Dutch: “De afbeelding toont Kimberley Koelink, die voor haar huis staat in een voortuin vol Halloweendecoraties. Ze draagt een blauwe Adidas-jas en kijkt recht in de camera. Rondom haar zijn diverse enge versieringen te zien, waaronder een grote skeletfiguur in een zwarte en paarse cape, grafstenen, heksenhoeden en andere Halloween-gerelateerde objecten. De tuin en gevel zijn rijkelijk versierd, wat bijdraagt aan een griezelige sfeer.”

Translation: “The image shows Kimberley Koelink standing in front of her house in a front garden full of Halloween decorations. She is wearing a blue Adidas jacket and looking directly into the camera. Around her, various scary decorations can be seen, including a large skeleton figure in a black and purple cape, tombstones, witch hats, and other Halloween-related objects. The garden and facade are richly decorated, contributing to a spooky atmosphere.”

(Generated using the default unbounded prompt.)

The description also provides details that help us understand what the subject looks like (describing their jacket and the way she looks at the camera) and lists the contested decorations that caused commotion. Another interesting use of the contextual information is visible in the description below (from NOS 2024). Not only does ChatGPT include the location (the Media Markt in Rotterdam), but it also adds the circumstances that are described in the article (the discount promotion got out of hand).

Example

Dutch: “De afbeelding toont een grote menigte mensen die zich verzameld heeft bij een Mediamarkt in het centrum van Rotterdam, waar een kortingsactie uit de hand is gelopen. [...]”

Translation: “The image shows a large crowd of people gathered at a Media Markt in the center of Rotterdam, where a discount promotion got out of hand. [...]”

(Generated using the default unbounded prompt.)

Every article on NOS.nl starts with the title of the article, overlaid on a header image. Together, they give readers a general sense of what to expect from the body text. The example descriptions above seem similarly helpful (though this remains to be tested with image description users). In Section 5.4 we will further explore the content of the descriptions, but first we will look at some more general properties of the descriptions.

5.2 Automatic metrics

Recall that we automatically determined the length and readability of the descriptions. We first checked to make sure that the model was indeed able to follow our instructions and produce short/unbounded \times simple/default descriptions.

5.2.1 DESCRIPTION LENGTH

On average, the descriptions consisted of 54.1 words ($SD = 18.1$) or 337 characters ($SD = 108$). Table 3 provides an overview of the average number of words and characters per condition.

ChatGPT does not always obey length restrictions. One of the restrictions imposed on the model in the prompts for the short descriptions was that they could consist of a maximum of 280 characters. As shown in Table 3, the short descriptions consisted of fewer than 280 characters on average. However, there were also three short descriptions that consisted of more than 280 characters, the longest of which was 285 characters. Therefore, ChatGPT did not adhere to the 280-character limit for all descriptions.

Readability	Length	
	Short	Unbounded
	Average number of words	
	38.1 (SD = 4.65, Min = 25, Max = 47)	70.0 (SD = 10.90, Min = 49, Max = 96)
Simple	37.7 (SD = 4.81, Min = 25, Max = 46)	71.5 (SD = 11.40, Min = 49, Max = 96)
Default	38.4 (SD = 4.53, Min = 26, Max = 47)	68.6 (SD = 10.40, Min = 54, Max = 89)
	Average number of characters	
	241 (SD = 25.3, Min = 174, Max = 285)	433 (SD = 67.8, Min = 322, Max = 582)
Simple	237 (SD = 27.2, Min = 174, Max = 285)	428 (SD = 63.8, Min = 322, Max = 575)
Default	246 (SD = 22.7, Min = 186, Max = 284)	437 (SD = 72.3, Min = 326, Max = 582)

Table 3: Average description lengths per condition.

Descriptions in the short condition are indeed shorter. The number of words per description was used to compare the lengths of the short and unbounded descriptions. A Student’s t-test was initially performed for this purpose. However, this test did not meet the assumption of homogeneity of variances ($F(1, 118) = 40.7, p < .001$). Therefore, we decided to use a Welch’s t-test, which is less sensitive to unequal variances. This t-test shows that the short descriptions are significantly shorter than unbounded descriptions, $t(79.7) = -20.9, p < .001$.

5.2.2 READABILITY

In addition to the length of the descriptions, their readability was also analyzed. The readability of the descriptions was determined using the LiNT score (Pander Maat et al. 2023). LiNT assigns a text a score between 0 and 100. The average LiNT score for all descriptions was 45.9 (SD = 8.16). Table 4 provides an overview of the average LiNT scores for the four types of descriptions.

Length	Readability	
	Simple	Default
	42.5 (Min = 27, Max = 60)	49.2 (Min = 34, Max = 68)
Short	41.8 (Min = 27, Max = 58)	48.9 (Min = 34, Max = 66)
Unbounded	43.2 (Min = 31, Max = 60)	49.5 (Min = 36, Max = 68)

Table 4: Average LiNT scores per condition

ChatGPT can follow readability instructions. To assess the difference between the simple and default descriptions, a Student’s t-test was used. The results show a significant difference in readability between the simple and default descriptions, $t(118) = -4.91, p < .001$. The simple descriptions are rated as more readable based on the LiNT score. This confirms **Hypothesis 3**, which stated that simple descriptions would achieve a lower LiNT score. ChatGPT is therefore capable of creating descriptions at different reading levels.

5.3 Human ratings

To gain insight into the quality of the image descriptions, participants were asked to rate them on accuracy, replaceability, reading level, and completeness. For each type of description, 45 responses were collected per question on the multiple-choice questions, except for the default short descriptions, for which there were 44. Figure 1 shows the frequency of the selected answer options for the

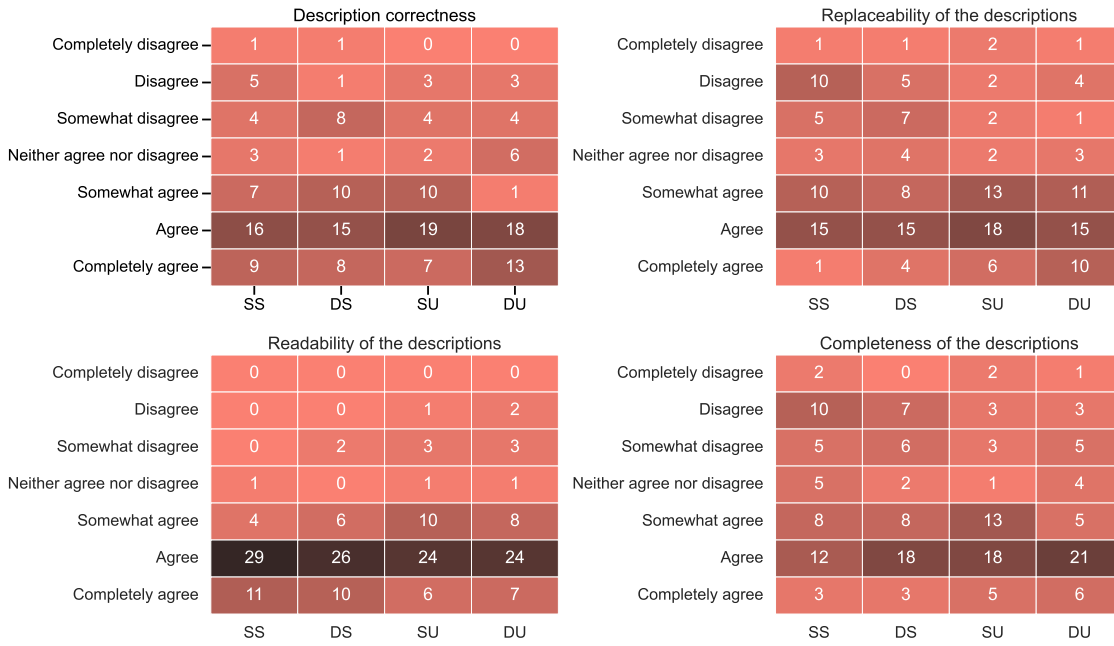


Figure 1: Frequency distribution of the answers. SS=Simple&Short, DS=Default&Short, SU=Simple&Unbounded, DU=Default&Unbounded.

statements, sorted by type of description. For all statements, the most frequently chosen response was ‘agree.’ This means that participants considered most descriptions to be correct and complete, that they felt the descriptions could replace the image, and that the descriptions were written in simple language.

All conditions are equally correct, but otherwise different. To compare the different types of descriptions with respect to correctness, replaceability, language level, and completeness, all response options (‘Completely disagree’ to ‘Completely agree’) were converted to numbers, with 1 representing ‘Completely disagree’ and 7 representing ‘Completely agree’. A Friedman test was then used to compare the descriptions. This showed no significant difference in participants’ assessments of the correctness of the four types of descriptions ($\chi^2(3) = 2.03$, $p = .567$). This means that all four types of descriptions were rated fairly equally with respect to correctness. Most descriptions were rated as correct by participants, with an average score of 5.26 (where 5 = ‘Somewhat agree’; 6 = ‘Agree’). A significant difference was found between the four descriptions regarding replaceability ($\chi^2(3) = 16.5$, $p < .001$), difficulty ($\chi^2(3) = 12.0$, $p = .007$), and completeness ($\chi^2(3) = 11.4$, $p = .010$). Post-hoc tests were conducted for these three criteria to investigate which descriptions significantly differed in the participants’ assessments. Figure 2 provides an overview of the participants’ answers to the questions about correctness, replaceability, difficulty, and completeness. Descriptions that differ significantly from each criterion are indicated with an asterisk.

Unbounded descriptions are better suited to replace images. For all four variations where a significant difference in replaceability was found, the unbounded descriptions scored higher. This indicates that the unbounded descriptions are most suitable for replacing the image, possibly because these descriptions are longer and therefore more complete.

Shorter is simpler. For language level assessments, a higher score means that the description is easier to read because it is written in simple language. The simple-short descriptions were rated as

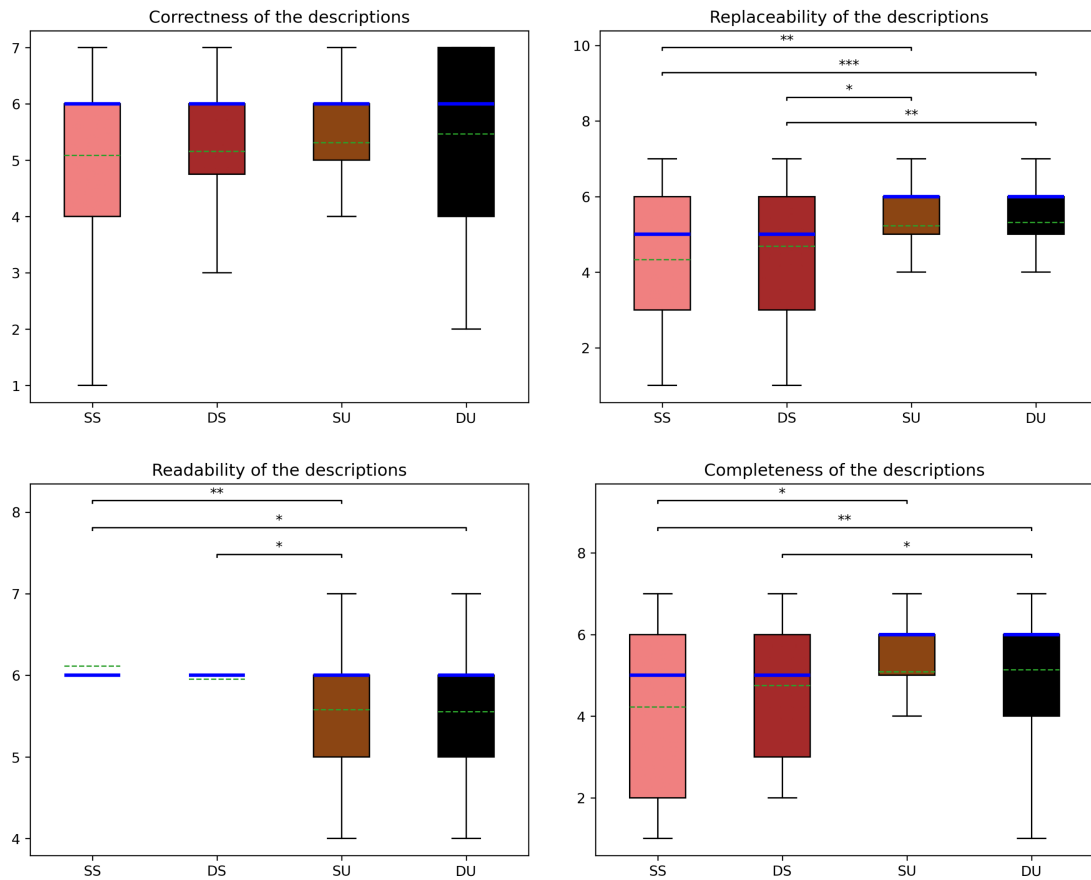


Figure 2: Judgments for correctness, replaceability, readability, and completeness. Asterisks show significance at $p < .001$ (***) , $p < .01$ (**) and $p < .05$ (*). Green lines show the means, blue lines show the median. SS=Simple&Short, DS=Default&Short, SU=Simple&Unbounded, DU=Default&Unbounded.

significantly easier than both the simple-unbounded and the default-unbounded descriptions. This suggests that the simple-short descriptions are easier to understand than the unbounded descriptions. However, it was also found that the default-short descriptions were rated as simpler than the simple-unbounded descriptions. The fact that the short descriptions were rated as simpler confirms **Hypothesis 2**, which stated that the short descriptions would be rated as simpler. As we will see later, this length heuristic may have interfered with participants' judgment of the actual language use.

Unbounded descriptions are more complete. For the completeness of the descriptions, a higher score is also better, as a higher score indicates a more complete description. In all three cases where a significant difference was found for completeness, the unbounded descriptions scored better. This means that the unbounded descriptions were rated as complete more often than the short descriptions. This is also likely due to the length of the descriptions. The unbounded descriptions contain more words, allowing more to be conveyed in these descriptions, leading to more complete

descriptions. While this finding may seem trivial, it does suggest that the shorter descriptions may not have enough space to capture all relevant aspects of the images in our dataset.²²

Longer is better. In addition to evaluating the individual descriptions, participants were also asked to select the best and worst descriptions and to rank them based on their readability. Table 5 shows the number of times each type of description was selected as the best or worst. The two unbounded conditions were chosen more often as the best than the short descriptions. This is the opposite of what was expected in **Hypothesis 1**, which expected the short descriptions to be rated higher than the unbounded descriptions.

Condition	# Best	# Worst
Simple, short	5	10
Default, short	6	10
Simple, unbounded	12	14
Default, unbounded	17	6

Table 5: Number of times each type of description was selected as the best or worst.

Participants who indicated that the default unbounded description was the best reasoned that it was the most complete and contained the most objective information. The simple unbounded descriptions were most often chosen as the worst description, because, according to participants, they often contained too much irrelevant information. Thus, somehow, asking ChatGPT to provide descriptions that are easier to read, also made the descriptions worse; an unexpected side-effect that is clearly not ideal for individuals with low literacy. This issue may be alleviated by further developing the prompts, or perhaps by taking a pipeline approach where default descriptions are simplified post-hoc.

‘Simple’ descriptions are sometimes seen as the most difficult. Regarding the readability of the descriptions, the simple-short descriptions were most often rated as the easiest. These descriptions were also most often chosen as the second most complex. The standard-short descriptions were most often ranked third. The simple-unbounded descriptions were most often rated as the least easy description by participants and therefore ended up last. Figure 3 provides a complete overview of the number of times each type of description was placed at the first to fourth language level by participants. From this figure it is clear that description length affects the perceived difficulty, with unbounded descriptions most often being perceived as most difficult.

Simple ... is not so simple? The fact that the simple-unbounded descriptions were most often rated as the most difficult refutes **Hypothesis 5**, which stated that the simple descriptions would be rated as easier. This also shows that the participants’ judgments of the language level differ from the LiNT score. According to the LiNT score, the simple descriptions are significantly easier than the default descriptions. However, the participants’ judgments show that the simple-unbounded descriptions were often perceived as the most difficult. A Spearman correlation also shows that there is no correlation between the LiNT scores of the descriptions and the participants’ judgments of the readability, $rs(107) = -0.054$, $p = .579$. This also refutes **Hypothesis 4**, which stated that the LiNT scores would correspond to the participants’ judgments.

Our best explanation for the discrepancy between the LiNT scores and the human ratings is that our participants may have relied mostly on the overall length of the description text, because (as we

22. Otherwise our participants would have said the shorter descriptions are complete already. However, a counter-argument here would be that the presence of other descriptions may have caused our participants to judge the shorter descriptions as being incomplete (because they provide fewer details than the long descriptions) while the short descriptions may have been found to be complete when presented in isolation.

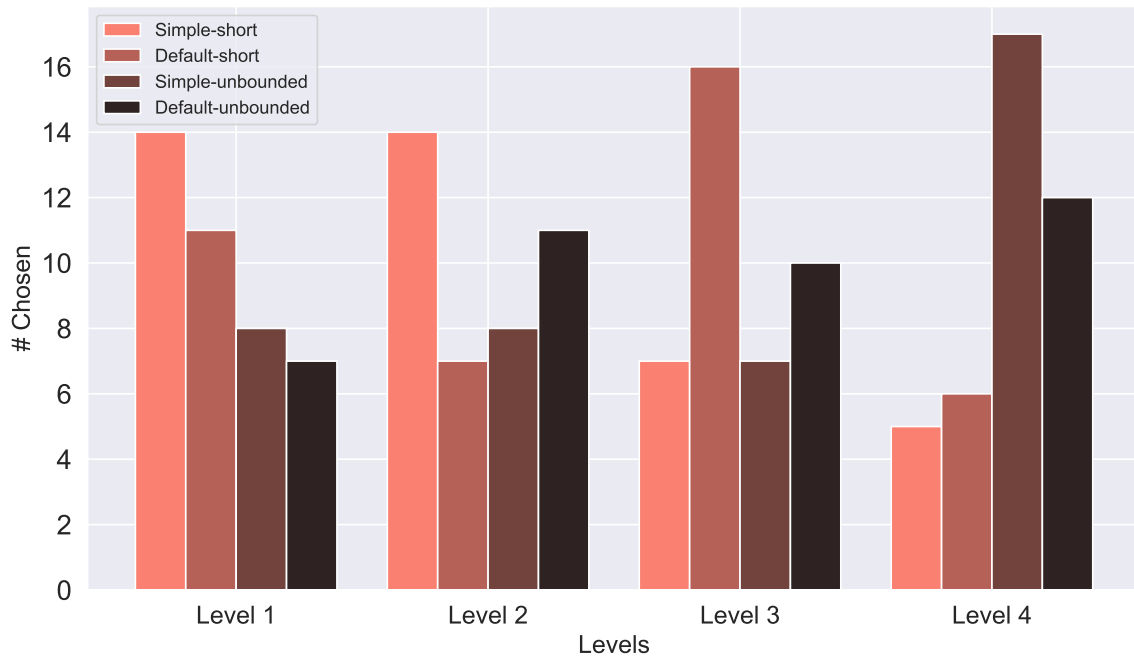


Figure 3: Distribution of the rankings. Level 1 is the easiest description, level 4 is the most difficult description.

have seen in Figure 1) they perceived most descriptions to be (very) readable.²³ By contrast, LiNT does not take the overall length of a text into account, but instead looks at other features such as sentence structure, the use of discourse connectives, word frequencies, and so on.

5.4 Content analysis

5.4.1 QUANTITATIVE RESULTS

The analysis revealed that all descriptions were written around the central persons and/or objects in the image and that none of the descriptions contained text that was copied verbatim from the article. Furthermore, it was found that none of the descriptions included the ethnicity of the people in the image. Because these three aspects are constant for all descriptions, they were not included in the further analysis of the results of the content analysis.

Figure 4 provides an overview of the results of our content analysis. The values in the figure represent the percentages of descriptions in which this aspect was correctly applied. For example, when we look at all images that contain text, 81% of the descriptions contain that text. In the row below, still in the first column, we see that 79% of the descriptions that should mention the name of a famous person actually do mention it. The other columns indicate percentages for subsets of the total dataset.²⁴

23. Also recall that most of our participants are highly educated, so they are not likely to have any difficulties reading short descriptions of images.

24. Regarding the interpretation of ‘Gender,’ it was decided that a description is considered correct when gender is not mentioned, or when it is mentioned for famous persons whose gender is known. For unknown persons whose gender is unknown, it is considered incorrect if gender is still mentioned in the description.

Text	81	75	88	79	83	75	75	83	92
Name of famous persons	79	86	71	71	86	86	86	57	86
Appearance of famous persons	71	64	79	64	79	57	71	71	86
Description of famous persons	92	89	84	92	92	89	89	94	94
Age of unknown persons	6	6	6	6	6	6	6	6	6
Hair color of unknown persons	1	3	0	0	3	0	8	0	0
Clothing of unknown persons	49	39	58	47	50	39	56	39	61
Gender	82	81	84	81	84	81	81	81	88
Activities of persons	85	83	86	88	81	86	81	90	81
Emotions	32	37	27	32	32	36	36	27	27
Animals	100	100	100	100	100	100	100	100	100
Location	87	87	87	83	90	87	87	80	93
Mentioned location	34	33	35	23	46	21	46	25	46
Color	63	53	72	65	60	57	50	73	70
Interesting details	83	80	85	85	80	80	80	90	80
Context	32	32	32	20	43	23	40	17	47
Objectivity	71	87	55	72	70	97	77	47	63
General to specific	48	60	37	42	55	60	60	23	50
	Total	Short	Unbounded	Simple	Default	Simple-short	Default-short	Simple-unbounded	Default-unbounded

Figure 4: Overview of all aspects that were annotated in the content analysis, per condition. Numbers are all percentages.

What kind of information gets specified in the descriptions? The content analysis revealed that the following elements were correctly mentioned in over 75% of the descriptions: text, names of famous people, a description of unknown people, gender, activities of people, animals, location, and interesting details. However, it also emerged that the age and hair color of unknown people, or people's emotions, were rarely mentioned in the image descriptions. The analysis also showed that information from the article was not often used to clarify the image's context in the descriptions.

Differences between conditions. To compare the descriptions, the differences between short and unbounded descriptions, as well as between simple and default descriptions, were examined. The chi-square test between the short and unbounded descriptions revealed significant differences in color ($\chi^2(1) = 4.30, p = .038$), objectivity ($\chi^2(1) = 14.6, p < .001$), and the degree to which descriptions were formulated from general to specific ($\chi^2(1) = 6.54, p = .011$). Color was mentioned more often in the unbounded descriptions, while short descriptions were more often objective and followed a structure from general to specific. For the descriptions with different language levels (simple versus

default), the chi-square test resulted in significant differences in the aspects of location ($\chi^2(1) = 5.59, p = .018$) and context ($\chi^2(1) = 7.55, p = .006$). In default descriptions, locations, if present in the article text, were mentioned more often. In addition, default descriptions more often used the article to clarify the context of the image. For all other aspects considered in the content analysis, no significant differences were found between the short and unbounded descriptions or the simple and default descriptions.

5.4.2 OTHER OBSERVATIONS

In addition to the 17 aspects of the content analysis, other notable features of the descriptions were also considered during the analysis. First, it was striking that all simple-unbounded descriptions began with “In the photo...” and all standard-unbounded descriptions with “The image...”. These systematic differences emerged without any relevant instructions in our prompts.

Uncertainty in the generated descriptions. It was noticeable that many descriptions were worded with a high degree of uncertainty. For example,

Example

In an image from an article about Extinction Rebellion activists at the IJmuiden lock, the simple-short description reads: “. . . people in yellow overalls, probably activists, at a lock.” The wording of this sentence is uncertain due to the use of the word “probably.” The description is also underspecified, since it does not mention the location (IJmuiden) that is mentioned in the article. The default short description, however, does mention that it concerns protesters from Extinction Rebellion, and the default unbounded description includes both information about the protesters and the location of the lock. Therefore, the information that is included and the degree of uncertainty varies per description, even if we keep the article and the image constant.

Descriptions often fail to use information from the article. Another example concerns an article about a train accident. This description contains the following text: “Presumably after a collision with a boulder.” The text of the article states that the train hit a boulder, so this description does not properly reflect the context of the article.

Incorrect descriptions In addition to the uncertainty in the descriptions, several descriptions also contained incorrect information. This might range from a misinterpretation of the activity displayed in the image (e.g. “. . . a uniformed officer stands by the car and looks into it.” when the officer is looking in a different direction), to hallucinated identities (where Dutch politician Sophie Hermans is mistaken for Rob Jetten).

Atmosphere It was also striking that many descriptions mentioned the atmosphere of the image, which affected the objectivity of the descriptions. The description of the atmosphere occurred primarily in the unbounded descriptions. For example, one of the simple-unbounded descriptions describes the weather and provides an interpretation of the atmosphere: “The sky is gray and cloudy, which gives the scene a somber feel.”). Another simple-unbounded description contains the following text: “. . . and the atmosphere seems energetic and combative.” Because of these descriptions of the atmosphere, the image descriptions are no longer completely objective. Although the unbounded descriptions were often rated as the most complete by participants, the use of additional words in the description did not always result in better descriptions, since these could also provide more subjective information, which is often undesirable in an image description (DIAGRAM Center 2015).

Model	Size	Condition	Length	LiNT	
gemma3	27b	Simple, short	57.3 (SD=33.97)	42.6 (SD=16.63)	
	27b	Simple, long	108.6 (SD=30.81)	47.6 (SD=9.39)	
	27b	Default, short	61.3 (SD=34.71)	44.1 (SD=13.59)	
	27b	Default, long	103.9 (SD=38.89)	45.0 (SD=7.64)	
llama3.2-vision	11b	Simple, short	68.5 (SD=34.65)	41.9 (SD=20.33)	
	11b	Simple, long	111.2 (SD=39.66)	42.3 (SD=14.15)	
	11b	Default, short	61.4 (SD=32.92)	47.1 (SD=13.49)	
	11b	Default, long	111.7 (SD=34.38)	48.1 (SD=8.24)	
llava	13b	Simple, short	51.6 (SD=28.46)	35.0 (SD=19.07)	
	13b	Simple, long	105.0 (SD=39.42)	43.3 (SD=14.56)	
	13b	Default, short	61.9 (SD=40.69)	50.0 (SD=14.00)	
	13b	Default, long	120.9 (SD=62.11)	47.3 (SD=7.74)	
	34b	Simple, short	57.2 (SD=30.84)	47.9 (SD=18.04)	
	34b	Simple, long	104.5 (SD=37.13)	45.6 (SD=6.62)	
	34b	Default, short	63.4 (SD=25.39)	44.0 (SD=14.29)	
	34b	Default, long	124.8 (SD=57.60)	48.1 (SD=8.69)	
	minicpm-v	8b	Simple, short	52.1 (SD=37.47)	36.5 (SD=26.44)
		8b	Simple, long	117.8 (SD=39.64)	44.4 (SD=10.43)
8b		Default, short	55.4 (SD=29.25)	45.9 (SD=13.98)	
8b		Default, long	116.7 (SD=49.99)	46.6 (SD=6.89)	
qwen2.5vl	32b	Simple, short	58.7 (SD=33.83)	39.7 (SD=20.11)	
	32b	Simple, long	105.2 (SD=28.03)	43.4 (SD=14.69)	
	32b	Default, short	53.7 (SD=27.35)	44.9 (SD=12.06)	
	32b	Default, long	110.4 (SD=38.68)	44.9 (SD=6.95)	

Table 6: Description length (in words) and LiNT scores for each of the models that we tested.

6. Discussion

6.1 Alternative models?

Now we have seen the results from ChatGPT 4o, a natural next question is how well other models perform. While it is tempting to explore other commercial models, we believe our efforts are best spent exploring more open alternatives.²⁵ In an ideal situation, we would have pre-trained local models that can easily be integrated with other tools, and that are able to produce personalized Dutch image descriptions. How far removed are we from this utopia?

We explored six different models through Ollama:²⁶ `gemma3:27b`, `llava:34b`, `llava:13b`, `llama3.2-vision:11b`, `qwen2.5vl:32b`, and `minicpm-v:8b`. We used the same approach as in our main study. Due to time constraints we only computed the description length and the LiNT scores for all models,²⁷ but we did not carry out a human evaluation study or a full content analysis for all results. Instead, we manually inspected the output to see if the results are promising.

Table 6 provides the description lengths and LiNT-scores for these models. The description lengths are definitely encouraging: all models follow the length instructions. The readability prompt turns out to be harder to follow: not all models have lower LiNT scores for the simple descriptions.

25. Though of course, ‘open’ is a relative concept (Liesefeld and Dingemanse 2024).

26. An open-source platform for running large language models locally, see <https://ollama.com>

27. Despite the discrepancy between the human ratings and the LiNT scores, we believe the LiNT scores are a meaningful indicator of readability. As noted above, the discrepancy is seemingly based on the fact that our participants made a relatively superficial assessment of the readability of the descriptions, focusing only on their length instead of lexical/syntactic features.

More problematic is that even larger models (e.g. `gemma3:27B`) suffer from grammar issues, and some models even spontaneously switch to English (`llava:13b`, `qwen2.5vl:32b`). `minicpm-v:8b` also spontaneously decided to paraphrase or repeat the news article instead of describing the image.

One way to improve the results of these models is to take a pipeline approach: first ask the models to produce English-language image captions, and then translate the results into Dutch. This would simplify the task for the language models, but it might also introduce biases and errors in the image descriptions, particularly if the visual context is not taken into account in the translation process (if a separate unimodal translation model is used). See, for example, Frank et al. (2018) for further discussion of multimodal machine translation.

6.2 Limitations and future research

6.2.1 SCOPE

The main limitation of this study is its relatively narrow scope. We only looked at a single language model, using prompts that could be developed further, with a relatively small set of input articles. The benefit of our approach is that we were able to carry out an extensive performance analysis, looking at the generated captions in more detail than most other image captioning studies. So while we cannot make general claims about the performance of language models *tout court*, we can make observations about what (the tested version of) ChatGPT already seems to do well and where there are still reliability issues. Moreover, this study allowed us to explore ways to translate existing image description guidelines into an evaluation procedure that *is* more broadly applicable. We hope that our work can inspire others to also look at image descriptions from a user perspective. Through those future studies we can hopefully find ways to overcome the issues identified in this paper.

6.2.2 HUMAN VERSUS AUTOMATIC EVALUATION OF READABILITY

To assess the readability of the automatically generated image descriptions, we have used both human ratings and automatically generated readability scores. While we expected these measures to converge, we in fact found no correlation between the two. As we have noted, this non-correlation could be due to the presence of a confounder: description length. However, this hypothesis has not been tested, and alternative explanations remain possible. To support the development of image description technology, it would be useful to develop tools that can reliably assess the readability of image descriptions. These tools should be validated with the target audience: description users.

7. Conclusion

We conclude that, at least with the prompts that we tried, ChatGPT is not yet capable of generating image descriptions that can be directly published with an image. Several descriptions created for this study contain inaccuracies, and the descriptions are not always easy to read due to the use of long and complex sentences. ChatGPT is clearly able to generate impressive contextualised image descriptions, but it is currently still necessary to have them reviewed by a human before they can be published online. The same also holds for ChatGPT's ability to generate texts that are easy to read, at least according to LiNT.

Our impression from this study is that additional requirements, such as the request to generate language that is easy to understand, add a layer of complexity that may cause performance to degrade. Although this hypothesis still requires further investigation, it would entail an additional burden for individuals with low literacy: their needs may result in poorer service from LLMs. One redeeming quality of LLMs that we have not discussed in this paper, is their ability to answer follow-up questions. In theory, one might ask a clarification question if they do not understand the output, or ask for more detail if the original level of detail is not sufficient. However, this still puts a burden

on the user to always ask for clarification, rather than immediately being provided with the desired information in the desired form.

Despite the limitations of LLMs in generating image descriptions, they are already being used by people with visual impairments to describe images, for example, through the JAWS text-to-speech software or the “Be My AI” app (part of Be My Eyes). However, because these image descriptions are not always completely accurate, it is important to further investigate how better descriptions can be generated with AI. This is important to minimize the risk of people with visual impairments receiving incorrect information about an image. Until AI-driven solutions are good enough, the best option for now is for publishers to manually (or semi-automatically) provide high-quality image descriptions themselves.

To make advancements in automatic image captioning, we propose to center the requirements of those who actually use captions. These may be people who are blind or visually impaired, but they could also be people who would like to listen to the news via a smart assistant, for example. Our study has taken the former use case, and we have shown how one might operationalise user requirements to measure the performance of image captioning systems. We hope at least some readers will be inspired to build on this work, so that we can improve the quality of Dutch image captioning systems.

Acknowledgments

This paper is based on the Master’s thesis of the first author (Heiden 2025), supervised by the second author. We thank the participants for carrying out the rating task, and Jan de Wit and the CLIN Journal reviewers for providing valuable feedback. We are also grateful to Dedicon for sharing their insights from a pilot study using ChatGPT to describe images. The second author was funded by ZonMw grant 06370422220003, *Experience matters: unlocking and improving image description guidelines through participatory design*.

References

- Antol, Stanislaw, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh (2015), Vqa: Visual question answering, *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2425–2433.
- Beks van Raaij, Nadine, Daan Kolkman, and Ksenia Podoyntsyna (2024), Clearer governmental communication: Text simplification with ChatGPT evaluated by quantitative and qualitative research, in Nunzio, Giorgio Maria Di, Federica Vezzani, Liana Ermakova, Hosein Azarbondyad, and Jaap Kamps, editors, *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024*, ELRA and ICCL, Torino, Italia, pp. 152–178. <https://aclanthology.org/2024.determin-1.15/>.
- Bennett, Cynthia L., Cole Gleason, Morgan Klaus Scheuerman, Jeffrey P. Bigham, Anhong Guo, and Alexandra To (2021), “it’s complicated”: Negotiating accessibility and (mis)representation in image descriptions of race, gender, and disability, *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI ’21, Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/3411764.3445498>.
- Bernardi, Raffaella, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank (2016), Automatic description generation from images: A survey of models, datasets, and evaluation measures, *Journal of Artificial Intelligence Research* **55**, pp. 409–442, AI Access Foundation. <http://dx.doi.org/10.1613/jair.4900>.

- Calò, Eduardo, David M. Howcroft, Leo Leppänen, Saad Mahamood, Simon Mille, Patrícia Schmidová, and Emiel van Miltenburg (2026), Justify your prompts!, *Computational Linguistics* pp. 1–12, Association for Computational Linguistics. <https://doi.org/10.1162/COLI.a.620>.
- Cheema, Gullal S., Sherzod Hakimov, Eric Müller-Budack, Christian Otto, John A. Bateman, and Ralph Ewerth (2023), Understanding image-text relations and news values for multimodal news analysis, *Frontiers in Artificial Intelligence*, Frontiers Media SA. <https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2023.1125533>.
- Cicchetti, Domenic V., Donald Shoinralter, and Peter J. Tyrer (1985), The effect of number of rating scale categories on levels of interrater reliability : A monte carlo investigation, *Applied Psychological Measurement* **9** (1), pp. 31–36, Sage Publications. <https://doi.org/10.1177/014662168500900103>.
- Das, Rocktim, Simeon Hristov, Haonan Li, Dimitar Dimitrov, Ivan Koychev, and Preslav Nakov (2024), EXAMS-V: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models, in Ku, Lun-Wei, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Bangkok, Thailand, pp. 7768–7791. <https://aclanthology.org/2024.acl-long.420/>.
- DIAGRAM Center (2015), *Image Description Guidelines*. <http://diagramcenter.org/table-of-contents-2.html#book>.
- Elliott, Desmond and Martijn Kleppe (2016), 1 million captioned Dutch newspaper images, in Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, European Language Resources Association (ELRA), Portorož, Slovenia, pp. 3054–3058. <https://aclanthology.org/L16-1488/>.
- Erdem, Erkut, Menekse Kuyu, Semih Yagcioglu, Anette Frank, Letitia Parcalabescu, Barbara Plank, Andrii Babii, Oleksii Turuta, Aykut Erdem, Iacer Calixto, Elena Lloret, Elena-Simona Apostol, Ciprian-Octavian Truică, Branislava Šandrih, Sanda Martinčić-Ipšić, Gábor Berend, Albert Gatt, and Grăzina Korvel (2022), Neural natural language generation: A survey on multilinguality, multimodality, controllability and learning, *Journal of Artificial Intelligence Research* **73**, pp. 1131–1207, AI Access Foundation. <http://dx.doi.org/10.1613/jair.1.12918>.
- Frank, Stella, Desmond Elliott, and Lucia Specia (2018), Assessing multilingual multimodal image description: Studies of native speaker preferences and translator choices, *Natural Language Engineering* **24** (3), pp. 393–413, Cambridge University Press.
- Gubbi Mohanbabu, Ananya and Amy Pavel (2024), Context-aware image descriptions for web accessibility, *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '24, Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/3663548.3675658>.
- Hanley, Margot, Solon Barocas, Karen Levy, Shiri Azenkot, and Helen Nissenbaum (2021), Computer vision and conflicting values: Describing people with automated alt text, *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, Association for Computing Machinery, New York, NY, USA, p. 543–554. <https://doi.org/10.1145/3461702.3462620>.
- Heiden, I. H. van der (2025), *Kan artificial intelligence afbeeldingen beter toegankelijk maken? : beeldbeschrijvingen met chatgpt*, Master's thesis, Tilburg University. Department of Communication and Cognition. <http://arno.uvt.nl/show.cgi?fid=182844>.

- Hershcovich, Daniel, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard (2022), Challenges and strategies in cross-cultural NLP, in Muresan, Smaranda, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Dublin, Ireland, pp. 6997–7013. <https://aclanthology.org/2022.acl-long.482/>.
- Hodosh, M., P. Young, and J. Hockenmaier (2013), Framing image description as a ranking task: Data, models and evaluation metrics, *Journal of Artificial Intelligence Research* **47**, pp. 853–899, AI Access Foundation. <http://dx.doi.org/10.1613/jair.3994>.
- Jandrey, Alessandra Helena, Duncan Dubugras Alcoba Ruiz, and Milene Selbach Silveira (2024), Imagery contents descriptions for people with visual impairments, *Proceedings of the XXII Brazilian Symposium on Human Factors in Computing Systems, IHC '23*, Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/3638067.3638105>.
- Lentz, Leo (2021), Wat zijn tekstbegrijpelijkheids voorspellingen waard? een vergelijkend onderzoek (deel 4), *Didactiek Nederlands*. <https://didactieknederlands.nl/2021/10/wat-zijn-tekstbegrijpelijkheids-voorspellingen-waard-een-vergelijkend-onderzoek-deel-4/>.
- Lewis, Veronica (2018), How to write alt text and image descriptions for the visually impaired, Guide posted on the website of the Perkins School for the Blind. Last updated January 2026; we used the 2024 version. <https://www.perkins.org/resource/how-write-alt-text-and-image-descriptions-visually-impaired/>.
- Li, Zongxia, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi (2025), A survey of state of the art large vision language models: Benchmark evaluations and challenges, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1587–1606.
- Liesenfeld, Andreas and Mark Dingemans (2024), Rethinking open source generative ai: openwashing and the eu ai act, *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24*, Association for Computing Machinery, New York, NY, USA, p. 1774–1787. <https://doi.org/10.1145/3630106.3659005>.
- Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick (2014), Microsoft coco: Common objects in context, in Fleet, David, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, Springer International Publishing, Cham, pp. 740–755.
- Liu, Fangyu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott (2021), Visually grounded reasoning across languages and cultures, in Moens, Marie-Francine, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp. 10467–10485. <https://aclanthology.org/2021.emnlp-main.818/>.
- Marsh, Emily E. and Marilyn Domas White (2003), A taxonomy of relationships between images and text, *Journal of Documentation* **59** (6), pp. 647–672, Emerald. <https://doi.org/10.1108/00220410310506303>.
- NOS (2024), Vernielingen en opstootjes bij kortingsactie Mediamarkt Rotterdam, Published on NOS.nl. <https://nos.nl/artikel/2542242-vernuelingen-en-opstootjes-bij-kortingsactie-mediemarkt-rotterdam> [Accessed 03-04-2026].

- NOS and Omroep West (2024), Tuin vol halloweendecoratie leidt tot verdeeldheid in Zuid-Hollands dorp — nos.nl. [Accessed 03-04-2026].
- Pander Maat, Henk, Suzanne Kleijn, and Servaas Frissen (2023), LiNT: een leesbaarheidsformule en een leesbaarheidsinstrument, *Tijdschrift voor Taalbeheersing* **45** (1), pp. 2–39, Amsterdam University Press. <https://www.aup-online.com/content/journals/10.5117/TVT2023.3.002.MAAT>.
- Preston, Carolyn C and Andrew M Colman (2000), Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences, *Acta Psychologica* **104** (1), pp. 1–15. <https://www.sciencedirect.com/science/article/pii/S0001691899000505>.
- Rogers, Anna (2023), Closed AI models make bad baselines. Accessed 09-12-2025. <https://towardsdatascience.com/closed-ai-models-make-bad-baselines-4bf6e47c9e6a/>.
- Stangl, Abigale, Meredith Ringel Morris, and Danna Gurari (2020), "person, shoes, tree. is the person naked?" what people with vision impairments want in image descriptions, *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, Association for Computing Machinery, New York, NY, USA, p. 1–13. <https://doi.org/10.1145/3313831.3376404>.
- Stangl, Abigale, Nitin Verma, Kenneth R. Fleischmann, Meredith Ringel Morris, and Danna Gurari (2021), Going beyond one-size-fits-all image descriptions to satisfy the information wants of people who are blind or have low vision, *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '21, Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/3441852.3471233>.
- Thapliyal, Ashish V., Jordi Pont Tuset, Xi Chen, and Radu Soricut (2022), Crossmodal-3600: A massively multilingual multimodal evaluation dataset, in Goldberg, Yoav, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp. 715–729. <https://aclanthology.org/2022.emnlp-main.45/>.
- van Miltenburg, Emiel (2016), Stereotyping and bias in the flickr30k dataset, in Edlund, Jens, Dirk Heylen, and Patrizia Paggio, editors, *Proceedings of Multimodal Corpora: Computer vision and language processing (MMC 2016)*, pp. 1–4.
- van Miltenburg, Emiel (2020), On the use of human reference data for evaluating automatic image descriptions, *CoRR*. Originally presented as a (non-archival) poster at the VizWiz 2020 workshop, collocated with CVPR 2020. <https://arxiv.org/abs/2006.08792>.
- van Miltenburg, Emiel (2024), Image captioning in different languages, *ArXiv*. <https://arxiv.org/abs/2407.09495>.
- van Miltenburg, Emiel, Ákos Kádár, Ruud Koolen, and Emiel Kraemer (2018), Didec: The dutch image description and eye-tracking corpus, *Proceedings of the 27th International Conference on Computational Linguistics*, Association for Computational Linguistics, pp. 3658–3669. <http://aclweb.org/anthology/C18-1310>.
- van Miltenburg, Emiel, Anouck Braggaaar, Nadine Braun, Debby Damen, Martijn Goudbeek, Chris van der Lee, Frédéric Tomas, and Emiel Kraemer (2023), How reproducible is best-worst scaling for human evaluation? a reproduction of 'data-to-text generation with macro planning', in Belz, Anya, Maja Popović, Ehud Reiter, Craig Thomson, and João Sedoc, editors, *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria, pp. 75–88. <https://aclanthology.org/2023.humeval-1.7/>.

- van Miltenburg, Emiel, Anouck Braggaar, Nadine Braun, Martijn Goudbeek, Emiel Krahmer, Chris van der Lee, Steffen Pauws, and Frédéric Tomas (2024), ReproHum: #0033-03: How reproducible are fluency ratings of generated text? a reproduction of August et al. 2022, in Balloccu, Simone, Anya Belz, Rudali Huidrom, Ehud Reiter, Joao Sedoc, and Craig Thomson, editors, *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, ELRA and ICCL, Torino, Italia, pp. 132–144. <https://aclanthology.org/2024.humeval-1.13/>.
- van Miltenburg, Emiel, Desmond Elliott, and Piek Vossen (2017), Cross-linguistic differences and similarities in image descriptions, in Alonso, Jose M., Alberto Bugarín, and Ehud Reiter, editors, *Proceedings of the 10th International Conference on Natural Language Generation*, Association for Computational Linguistics, Santiago de Compostela, Spain, pp. 21–30. <https://aclanthology.org/W17-3503/>.
- Wu, Shaomei, Jeffrey Wieland, Omid Farivar, and Julie Schiller (2017), Automatic alt-text: Computer-generated image descriptions for blind users on a social network service, *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, Association for Computing Machinery, New York, NY, USA, p. 1180–1192. <https://doi.org/10.1145/2998181.2998364>.
- Yin, Shukang, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen (2024), A survey on multimodal large language models, *National Science Review* **11** (12), pp. nwae403, Oxford University Press. <https://doi.org/10.1093/nsr/nwae403>.
- Young, Peter, Alice Lai, Micah Hodosh, and Julia Hockenmaier (2014), From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions, *Transactions of the Association for Computational Linguistics* **2**, pp. 67–78, MIT Press, Cambridge, MA. <https://aclanthology.org/Q14-1006/>.
- Zhang, Shengyu, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Guoyin Wang, and Fei Wu (2026), Instruction tuning for large language models: A survey, *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/3777411>.
- Zhang, Wenxuan, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing (2023), M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models, in Oh, A., T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, Vol. 36, Curran Associates, Inc., pp. 5484–5505. https://proceedings.neurips.cc/paper_files/paper/2023/file/117c5c8622b0d539f74f6d1fb082a2e9-Paper-Datasets_and_Benchmarks.pdf.

Appendix A. Prompts

Tables 7 and 8 provide the original prompts we used and their English translations.

Condition	Prompt
Simple, short	Jouw taak is om beeldbeschrijvingen te maken van foto's die horen bij nieuwsartikelen. Ieder bericht dat je krijgt bevat een afbeelding een document met de tekst van een nieuwsartikel. Maak een beeldbeschrijving van maximaal 280 karakters van deze afbeelding. De omschrijving moet gemakkelijk te lezen zijn voor iemand die laaggeletterd is. Doe geen ongegronde aannames. Geef geen oordeel over elementen in de afbeelding. Neem de tekst van het artikel mee in de beschrijving, maar beschrijf alleen de afbeelding. Zeg niets over de verhouding tussen de afbeelding en de tekst.
Default, short	Jouw taak is om beeldbeschrijvingen te maken van foto's die horen bij nieuwsartikelen. Ieder bericht dat je krijgt bevat een afbeelding een document met de tekst van een nieuwsartikel. Maak een beeldbeschrijving van maximaal 280 karakters van deze afbeelding. De omschrijving moet worden geschreven in standaard Nederlands. Doe geen ongegronde aannames. Geef geen oordeel over elementen in de afbeelding. Neem de tekst van het artikel mee in de beschrijving, maar beschrijf alleen de afbeelding. Zeg niets over de verhouding tussen de afbeelding en de tekst.
Simple, unbounded	Jouw taak is om beeldbeschrijvingen te maken van foto's die horen bij nieuwsartikelen. Ieder bericht dat je krijgt bevat een afbeelding een document met de tekst van een nieuwsartikel. Maak een beeldbeschrijving van deze afbeelding. De omschrijving moet gemakkelijk te lezen zijn voor iemand die laaggeletterd is. Doe geen ongegronde aannames. Geef geen oordeel over elementen in de afbeelding. Neem de tekst van het artikel mee in de beschrijving, maar beschrijf alleen de afbeelding. Zeg niets over de verhouding tussen de afbeelding en de tekst.
Default, unbounded	Jouw taak is om beeldbeschrijvingen te maken van foto's die horen bij nieuwsartikelen. Ieder bericht dat je krijgt bevat een afbeelding een document met de tekst van een nieuwsartikel. Maak een beeldbeschrijving van de afbeelding. De omschrijving moet worden geschreven in standaard Nederlands. Doe geen ongegronde aannames. Geef geen oordeel over elementen in de afbeelding. Neem de tekst van het artikel mee in de beschrijving, maar beschrijf alleen de afbeelding. Zeg niets over de verhouding tussen de afbeelding en de tekst.

Table 7: Dutch prompts for each of the conditions. Differences between conditions are shown in **bold**.

Condition	Prompt
Simple, short	Your task is to create image descriptions of photos that accompany news articles. Every message you receive contains an image and a document with the text of a news article. Create an image description of no more than 280 characters for this image. The description must be easy to read for low literacy individuals. Do not make unfounded assumptions. Do not pass judgment on elements in the image. Include the text of the article in the description, but describe only the image. Say nothing about the relationship between the image and the text.
Default, short	Your task is to create image descriptions of photos that accompany news articles. Every message you receive contains an image and a document with the text of a news article. Create an image description of no more than 280 characters for this image. The description must be written in standard Dutch. Do not make unfounded assumptions. Do not pass judgment on elements in the image. Include the text of the article in the description, but describe only the image. Say nothing about the relationship between the image and the text.
Simple, unbounded	Your task is to create image descriptions of photos that accompany news articles. Every message you receive contains an image and a document with the text of a news article. Create an image description of this image. The description must be easy to read for low literacy individuals. Do not make unfounded assumptions. Do not pass judgment on elements in the image. Include the text of the article in the description, but describe only the image. Say nothing about the relationship between the image and the text.
Default, unbounded	Your task is to create image descriptions of photos that accompany news articles. Every message you receive contains an image and a document with the text of a news article. Create an image description of the image. The description must be written in standard Dutch. Do not make unfounded assumptions. Do not pass judgment on elements in the image. Include the text of the article in the description, but describe only the image. Say nothing about the relationship between the image and the text.

Table 8: Translated prompts for each of the conditions. Differences between conditions are shown in **bold**.