Putting ChatGPT Vision (GPT-4V) to the test: Risk perception in traffic images

April 10, 2024

Tom Driessen[1], Dimitra Dodou[1], Pavlo Bazilinskyy[2], Joost de Winter[1]

[1] Delft University of Technology, The Netherlands
[2] Eindhoven University of Technology, The Netherlands

Abstract

Vision-language models are of interest in various domains, including automated driving, where computer vision techniques can accurately detect road users, but where the vehicle sometimes fails to understand context. This study examined the effectiveness of GPT-4V in predicting the level of 'risk' in traffic images as assessed by humans. We used 210 static images taken from a moving vehicle, each previously rated by approximately 650 people. Based on psychometric construct theory and using insights from the self-consistency prompting method, we formulated three hypotheses: 1) repeating the prompt under effectively identical conditions increases validity, 2) varying the prompt text and extracting a total score increases validity compared to using a single prompt, and 3) in a multiple regression analysis, the incorporation of object detection features, alongside the GPT-4V-based risk rating, significantly contributes to improving the model's validity. Validity was quantified by the correlation coefficient with human risk scores, across the 210 images. The results confirmed the three hypotheses. The eventual validity coefficient was r = 0.83, indicating that population-level human risk can be predicted using AI with a high degree of accuracy. The findings suggest that GPT-4V must be prompted in a way equivalent to how humans fill out a multi-item questionnaire.

Introduction

GPT-4V Background

In late September 2023, OpenAI introduced image-to-text functionality for ChatGPT, also called GPT-4V or GPT4 Vision. At that time, image-to-text software, such as BLIP, and functionalities within Google's Bard and Bing Chat were already available (Bing, 2023; Google, 2023; Li et al., 2022; see Cui et al., 2024 for a survey on multimodal large language models). However, GPT-4V was highly anticipated due to the high quality of its output, as demonstrated in earlier previews (OpenAI, 2023).

The research so far demonstrates that GPT-4V exhibits strong generic skills. It can comprehend diverse stimuli such as written text, charts, graphical user interfaces, abstract visual pictures, and visual IQ tests (Ahrabian et al., 2024; Yan et al., 2023; Z. Yang et al., 2023). GPT-4V is also capable of solving visual mathematical problems, although not yet at a high level (Lu et al., 2023). As of early 2024, GPT-4V is still considered superior to a recent competitor from Google, called Gemini-Pro (M. Liu et al., 2024; Qi et al., 2023), but see proprietary evaluations of Google's largest model, Gemini-Ultra (Gemini Team Google, 2023; Yue et al., 2023).

There is strong interest in GPT-4V within the domain of automated driving. Current automated vehicles are effective at detecting objects and handling routine scenarios, but the challenge still lies in rare situations that are not included in the training data (Bogdoll et al., 2022; Jain et al., 2021). The strength of GPT-4V (and other vision language models) is its ability to understand

context, including scenarios not previously encountered (Hwang et al., 2024; Z. Yang et al., 2023; Zhou & Knoll, 2024). On the other hand, while GPT-4V is skilled in recognising unusual traffic events, it is not skilled at seemingly trivial tasks such as recognising details like the status of traffic lights, and spatial tasks such as reporting the orientation and (relative) position of road users (Wen et al., 2023; Zhou & Knoll, 2024).

Indeed, GPT-4V exhibits several limitations. It struggles with counting objects and judging details, such as answering the question "*How many eyes can you see on the animal?*" or "*Count the number of trees in the given image*", tasks that normally do not pose a challenge for humans (Tong et al., 2024; Zhang & Wang, 2024). Furthermore, although GPT-4V performs well in commonsense visual question answering, it is prone to hallucinations when world knowledge is required, such as about real-world objects (Y. Li et al., 2024), especially for objects from non-Western countries (Cui et al., 2023). A similar pattern has been observed for medical images, where GPT-4V does not seem to possess the knowledge required for making accurate diagnoses or reports (Senkaiahliyan et al., 2023; Wu et al., 2023). Guan et al. (2023) made a distinction between visual illusions, in which a visual element is misrepresented, and language hallucinations, where GPT-4V fails to recognise a feature in the image because it adheres to previously learned stereotypical responses for similar images. Guan et al. also indicated that ChatGPT exhibits limitations in temporal reasoning abilities.

Prompting Methods

Different strategies exist for improving the output of GPT-4V. This includes a prompting method where images are first segmented and marked with characters or boxes before being submitted to GPT-4V (J. Yang et al., 2023). The use of composite images (Y. Li et al., 2024), comparing images in pairs (Zhang et al., 2023), or multimodal cooperation (Ye et al., 2023) are other viable strategies. Additionally, the literature recommends chain-of-thought prompting for GPT-4V (Ahrabian et al., 2024; Hou et al., 2024; Zhang et al., 2024), a strategy also known for text-only ChatGPT (Bellini-Leite, 2023; Wei et al., 2022). Others have converted visual information into text first, using a prompt such as "*what's in this image?*"; this method is promising when processing large quantities of images that occur in a temporal sequence (Y. Liu et al., 2024).

Small variations in the prompt can lead to substantially different outputs of large language models (Huang et al., 2023; Salinas & Morstatter, 2024). For example, when a list of short phrases is submitted to GPT for sentiment analysis, but the same list is sorted in a different order, the sentiment score from GPT is usually different, even if GPT is set to produce near-zero variation through its temperature parameter (Tabone & De Winter, 2023). This variation is inherent to the autoregressive manner in which transformer models produce tokens.

A technique to mitigate this randomness is self-consistency, also referred to as bootstrapping (Tabone & De Winter, 2023; Tang et al., 2023; Wang et al., 2023): After repeating the prompting process multiple times, each time with a different permutation of the text, the modal or mean output can be extracted. This aggregate typically has higher accuracy than the output of a single prompt. Various refinements of the self-consistency method exist (Fu et al., 2023; Li et al., 2023), more recently expanded to the notion of invoking multiple different language models (J. Li et al., 2024; Lu et al., 2024).

It is our proposition that self-consistency prompting resembles how constructs are defined in psychometrics. In psychology, a construct, such as personality (e.g., extraversion), can be estimated by having the person fill out multiple questionnaire items. By averaging the results of items that have been sampled from a domain of possible items, an estimation of the construct can

be made (Cronbach et al., 1972; Little et al., 2013; McDonald, 2003; Nunnally & Bernstein, 1994; Sawaki, 2010).

Current Study

This research focuses on evaluating GPT-4V, but not as in identifying specific visual elements, a domain in which GPT-4V demonstrates limited performance. Instead, we conducted a holistic assessment by examining the ability of GPT-4V to predict 'risk' as evaluated by humans. Instead, we conducted a holistic evaluation, where we examined how well GPT-4V can predict 'risk' as assessed by humans. More specifically, this study presents an assessment of GPT-4V concerning the prediction of risk in forward-facing photographs from the perspective of a moving vehicle.

Our analysis draws on a prior study (De Winter et al., 2023), in which human crowdworkers assessed the risk of traffic images, taken by a camera mounted on the roof of a car while driving on German roads (KITTI dataset; Geiger et al., 2013). In De Winter et al., a total of 210 images were rated by an average of 653 participants per image. Based on these ratings on a scale ranging from 0 (no risk) to 10 (extreme risk), a mean risk score was computed for each image.

De Winter et al. (2023) investigated whether the images' risk level, as assessed by humans, was predictable based on features extracted by a pretrained object detection algorithm (Bochkovskiy et al., 2020; Redmon & Farhadi, 2018), see Figure A1 in the Appendix. Their analysis showed that the number of people in the image (r = 0.33) and the mean size of the bounding boxes (r = 0.54) were predictive of the human risk scores. The driving speed was negatively predictive (r = -0.63), which can be explained by risk compensation (a less strict variant of risk homeostasis; Wilde, 1982, 2013): some situations, like empty roads, allow drivers to drive at the maximum allowed speed without it being high risk. Conversely, complex traffic environments, such as city centres, lead people to drive slowly (Charlton et al., 2010). Through a regression analysis, the three measures combined (number of people, size of bounding boxes, and vehicle speed) were found to be strongly predictive of the human risk level (r = 0.75). Excluding the speed variable, the prediction was weaker but still substantial (r = 0.62) (De Winter et al., 2023).

One might wonder why the prediction derived from the object detection was not more strongly indicative of the human risk ratings. In the previous study, we hypothesised that the object detection algorithm does not account for contextual information. For example, an image of a railroad crossing was perceived as hazardous by the human evaluators, whereas the object detection algorithm could not detect this railroad and did not understand the broader situation (De Winter et al., 2023). In the current study, we explored whether GPT-4V could contribute to a more accurate assessment of the risk in the traffic images as compared to using object detection features alone.

Hypotheses

Figure 1 provides one manner in which construct validity can be interpreted for risk ratings. Here, the risk score for a given image is the arithmetic mean risk from a large number of participants. These participants might all have had slightly different interpretations of the same rating task. For example, Participant 1 might interpret the task as 'probability of an accident occurring', Participant 2 as 'difficulty of the task', etc.—interpretations that are positively correlated but not the same (Fuller, 2005). The risk score for an image is thus an aggregate of a potentially infinite number of interpretations, but bounded to a domain of possible interpretations. Additionally, the same participant will not perform a reliable evaluation under a given interpretation of the task. For example, a participant may be distracted or overlook something in the image for arbitrary reasons. Therefore, noise is present, also known as 'measurement error'.

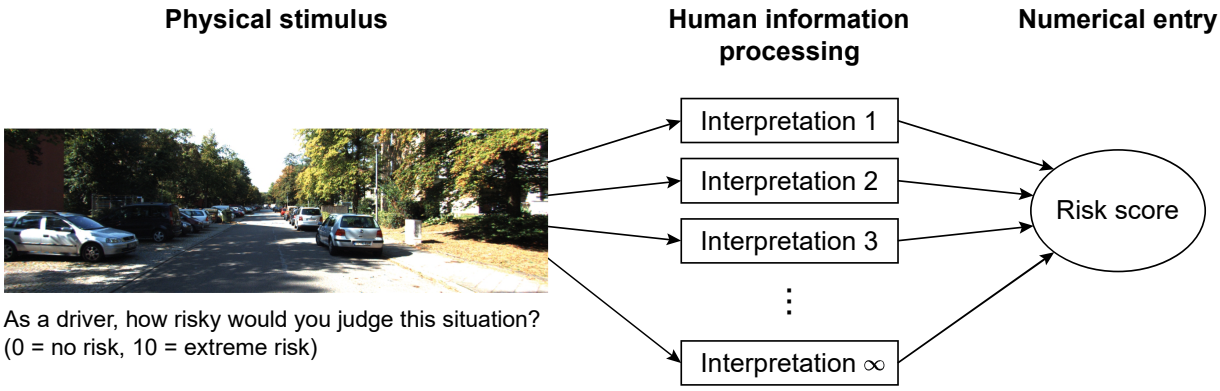

Figure 1. Causal process of how a participant generates a risk score for an image. The participant observes the image and task instruction presented on a computer screen, makes one (or a combination of multiple) interpretation(s), and enters a numerical risk score. The overall risk score for a given image represents the average from a large number of participants, thus reflecting an aggregation of a large number of different interpretations. This conceptualisation of construct validity is based on Markus and Borsboom (2013).

Considering the use of GPT-4V to approximate this human risk score as accurately as possible, three hypotheses are formulated. In each of the three hypotheses, validity is defined as the correlation coefficient between the mean risk score of GPT-4V and the human risk score.

H1: Repeating the same prompt under nearly identical conditions (in our case: keeping the images and prompt text identical, and only changing the order of the images within the same prompt) will result in higher validity as compared to using the exact same prompt.

H2: Aggregating the results of different prompts within a behavioural domain (in our case: slightly rephrasing the question) will result in higher validity as compared to using a single prompt text.

The aforementioned hypotheses are consistent with the self-consistency prompting method (Wang et al., 2023), but adapted for quantitative assessment and motivated from a psychometric perspective. Here, H1 is equivalent to the use of items in parallel forms, with the aim to reduce measurement error, while H2 is equivalent to the use of multiple items to estimate a latent construct.

H3: In a multiple regression analysis with GPT-4V included, object detection features, as used by De Winter et al. (2023), will statistically significantly contribute to predicting human risk. This hypothesis is based on the previously mentioned review, which indicated that GPT-4V possesses generic skills but may fail to recognise specific elements in images (e.g., Wen et al., 2023; Zhou & Knoll, 2024). Hence, the two different AI-based methods (vision-language model vs. object detection) were expected to have complementary value.

This study was conducted in two phases. Phase 1 was carried out using GPT-4V as available in the ChatGPT web interface. This approach was chosen because many users might not have access to the API (an interface for programmatic access), and because the API was not available at the time of this assessment. Phase 1 aimed to explore how multiple images could be evaluated and aggregated. To this end, we used composite images, with a selection of 10 images each time in a different order. We investigated to what extent repetitions result in an increase in validity (H1).

In Phase 2, we used OpenAI's API for GPT-4V, which has recently become available. We submitted images one-by-one and in batches of four, and we examined to what extent repetition

improves validity (H1) and to what extent different prompts within a plausible domain improve validity (H2). Finally, we investigated whether object detection scores contribute to validity in a linear regression model (H3).

Methods

Phase 1: ChatGPT Web Interface

The images were uploaded to the ChatGPT web interface as composites of 10, randomly selected from the total of 210 images. The individual images had a size of 1242×375 pixels. Due to the web interface having a limit of about 90 prompts in a 3-hour interval, this approach proved to be suitable for processing a large number of images. The decision to submit 10 images per prompt was based on trial and error, aiming to maximise the number of images submitted in a single prompt while maintaining good quality of GPT-4V's output. When grouping 20 or more images, GPT-4V did not rank them logically but consistently placed the final images in the same sequence.

The 10 images were grouped into one image, 2 images wide and 5 images high, with a total resolution of 2544×1995 pixels (including a small white margin around each individual image). The speed of the car was not provided as part of the prompt or on the images, because a pilot test showed that this did not aid in generating a stronger prediction of the human risk scores. The 10 images were grouped using MATLAB (the source code is available in the Supplementary Material).

Between October 26 and November 1, 2023, we prompted GPT-4V a total of 974 times, each time using a randomly selected 10 out of 210 images, and each prompt in a new chat window. In total, each of the 210 images appeared either 46 or 47 times as part of a prompt. In each prompt, we requested GPT-4V to rank or rate the 10 images. An example of a prompt is shown in Figure 2. In total, we used four different prompts:

1. "*Imagine you are a driver. Rank the 10 images in terms of risk from highest (1) to lowest (10). Only report a rank; nothing else. The image number is shown at the right bottom.*" (210 prompts)
2. "*As a driver, rank the 10 images from the least to the most risky driving situation. Each image is labeled with a letter in the bottom right corner. Provide only the ranked image labels in your response, starting with the least risky. For example: j, b, c, h, e, g, a, i, d, f*" (344 prompts)
3. "*Imagine you are a driver. Rank the 10 images in terms of risk from lowest to highest. Only report the ranked image numbers in a single column; nothing else. The image number is shown at the bottom right of each image.*" (210 prompts)
4. "*Imagine you are a driver. Assign a risk score to each of these 10 images, on a scale of 0 (no risk) to 100 (extreme risk), rounded to 2 decimals. Only report the risk percentages in a single column. Nothing else; no percentage sign either.*" (210 prompts)

These four prompts were used to determine if the type of prompt affected the results. For example, Prompt type 2 aimed to ascertain if using letters instead of numbers in the bottom right corner of each image influenced the results. We also tested how ranking from low to high (Prompt types 2 & 3) or from high to low (Prompt type 1) affected the output, and if giving a numeric rating (Prompt type 4) instead of a ranking possibly yielded better results.

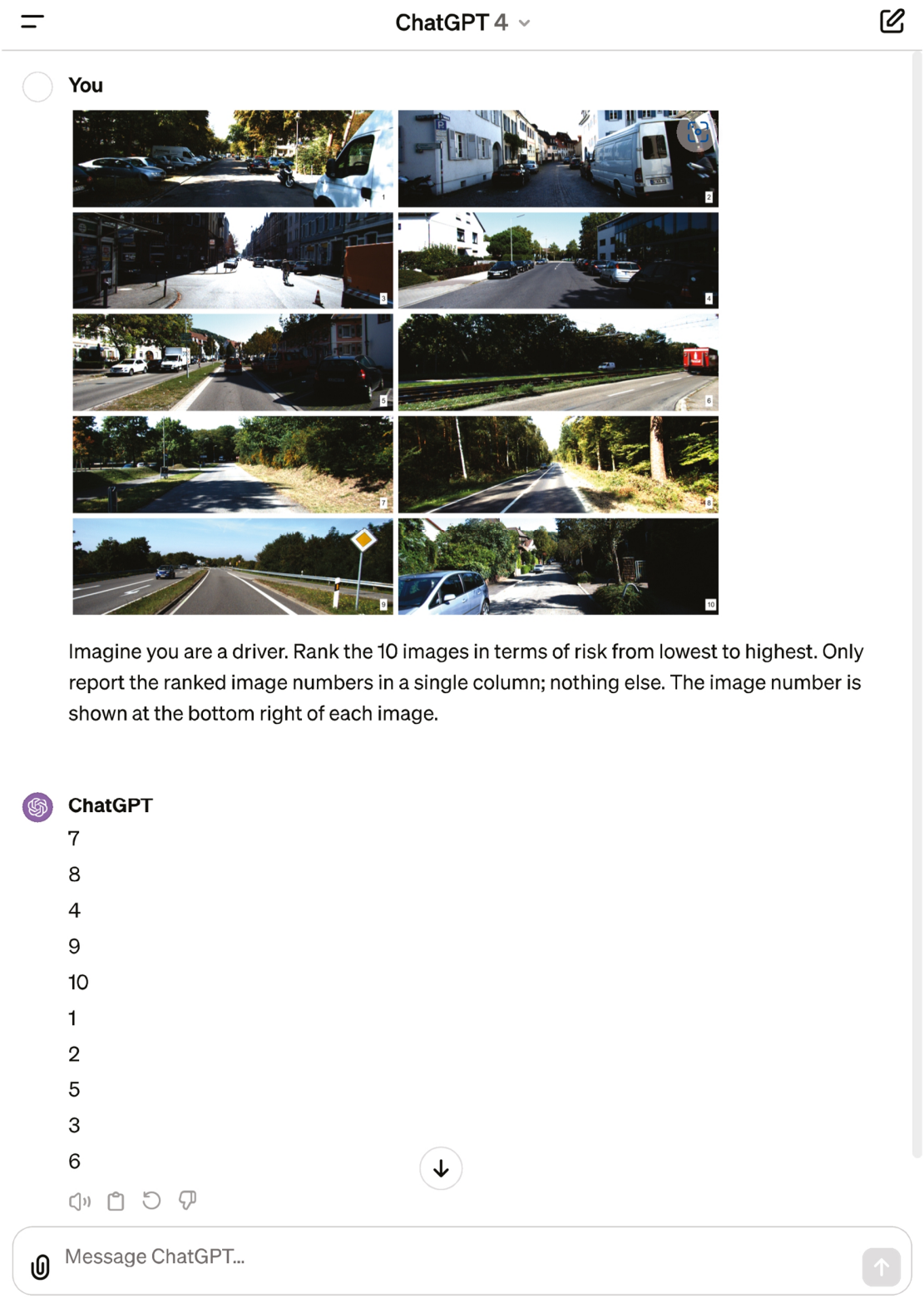

Figure 2. Example prompt and output of GPT-4V (Prompt type 3). The prompt includes a random selection of 10 of 210 images. The 10 images were combined into a single image.

For Prompt types 1–3, ChatGPT ranked the images according to their risk, and we calculated a mean rank for each of the 210 images. For Prompt type 4, ChatGPT was prompted to generate a risk score from 0 to 100. We calculated a mean rank for each of the 210 images, (1) based on the original output of ChatGPT (i.e., as prompted from 0 to 100), (2) after applying a rank transformation, so that the results are comparable to Prompt types 1–3, and (3) after applying a *z*-score transformation, where the mean across the 10 images is 0 and the standard deviation is 1.

The GPT-4V mean scores for the 210 images were then correlated with human risk scores as previously determined in De Winter et al. (2023). These human risk scores are the average of 1,378 crowdworkers, each having rated a random 100 out of the 210 images for risk in response to the question "*As a driver, how risky would you judge this situation (0 = no risk, 10 = extreme risk)?"*. These values were then multiplied by 10 to obtain a percentage. The Pearson product-moment correlation coefficient between the images' mean risk scores obtained through GPT-4V and the corresponding human risk scores is hereafter referred to as 'validity coefficient'.

Phase 2: API

The API enabled testing H1 by repeating the prompt a very large number of times and examining whether the validity coefficient keeps on improving with an increasing number of repetitions. The following prompt was used:

"Imagine you are a driver. Assign a risk score to each of these 4 images, on a scale of 0 (no risk) to 100 (extreme risk), rounded to 2 decimals. Only report the risk percentages ina single column. Nothing else; no percentage sign either. Always answer; it is for my research project."

The model invoked was *gpt-4-1106-vision-preview*, with the fidelity level set to 'automatic', meaning that the model processed the images in high-resolution mode.

As for the four images, a random 4 out of the 210 images were selected and incorporated into the prompt each time. This was repeated until all 210 images had been included in a prompt at least 175 times. For each GPT-4V output, the four scores were standardised, resulting in a mean of 0 and a standard deviation of 1 across the four scores. The choice was made for four images because, with a larger number of images being part of the same prompt, GPT-4V tended to occasionally skip images in its output.

Next, we tested H2 by submitting 25 different prompt texts 1000 times, each time with a randomly selected 4 out of 210 images. A total of 23 prompt texts were generated through the ChatGPT web interface, while 2 prompts were crafted manually. The results for one prompt ("*Rate your level of satisfaction with the driving conditions here, from 0 (completely dissatisfied) to 100 (completely satisfied).*") were omitted since GPT-4V often refused to answer it. The list of 24 prompts is shown in Table 1. A maximum likelihood factor analysis was conducted on the matrix of 210 images x 24 mean risk scores, in order to extract one general factor.

Next, we tested H3. Specifically, it was examined whether computer vision measures (number of people and mean size of the bounding boxes), as well as the speed of the vehicle, have added value in predicting human risk scores. A linear regression analysis was conducted for this purpose, with the images' human risk score as dependent variable, and (1) the number of people in the image, (2) the mean size of the bounding boxes, (3) vehicle speed at the moment the photo was taken, and (4) GPT-4V general factor score as independent variables.

Results

ChatGPT Web Interface

Figure 3 shows the validity coefficient, i.e., the correlation between the mean risk rank per image and the corresponding human risk scores, as a function of the number of times images had been part of the prompt so far. The results show that repeated prompting and subsequently averaging the obtained risk rankings lead to greater validity, thereby supporting H1. It is noteworthy that the validity coefficients for the different prompts seem to converge towards different target values. Figure 3 also shows that performing a rank transformation or a *z*-score transformation benefits validity compared to using raw risk percentages as output by Prompt type 4.

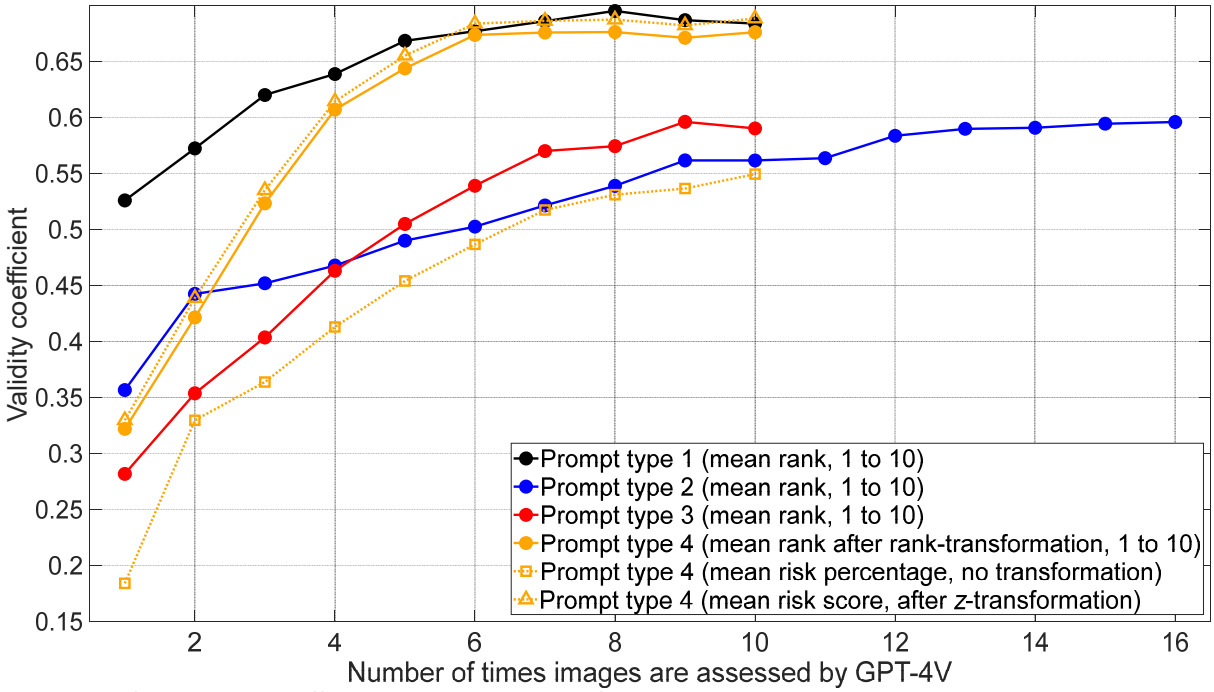

Figure 3. Correlation coefficient between mean GPT-4V-based risk rankings, as obtained using the ChatGPT web interface, and the human risk scores, for four different prompt types (see Methods). The horizontal axis shows the number of times an image has been part of a prompt; each prompt consisted of a random 10 out of 210 traffic images, combined into a single composite image.

API

Figure 4 shows the validity coefficients as a function of the number of times the images were assessed by GPT-4V. As in Figure 3, repeating the assessment was found to increase validity (i.e., higher correlation between GPT-4V mean risk and human risk, n = 210 images), supporting H1. Furthermore, although conclusive evidence cannot be obtained because there are practical and financial limits to how often a prompt could be repeated, it seems that there is convergence towards a target value, similar to Figure 3.

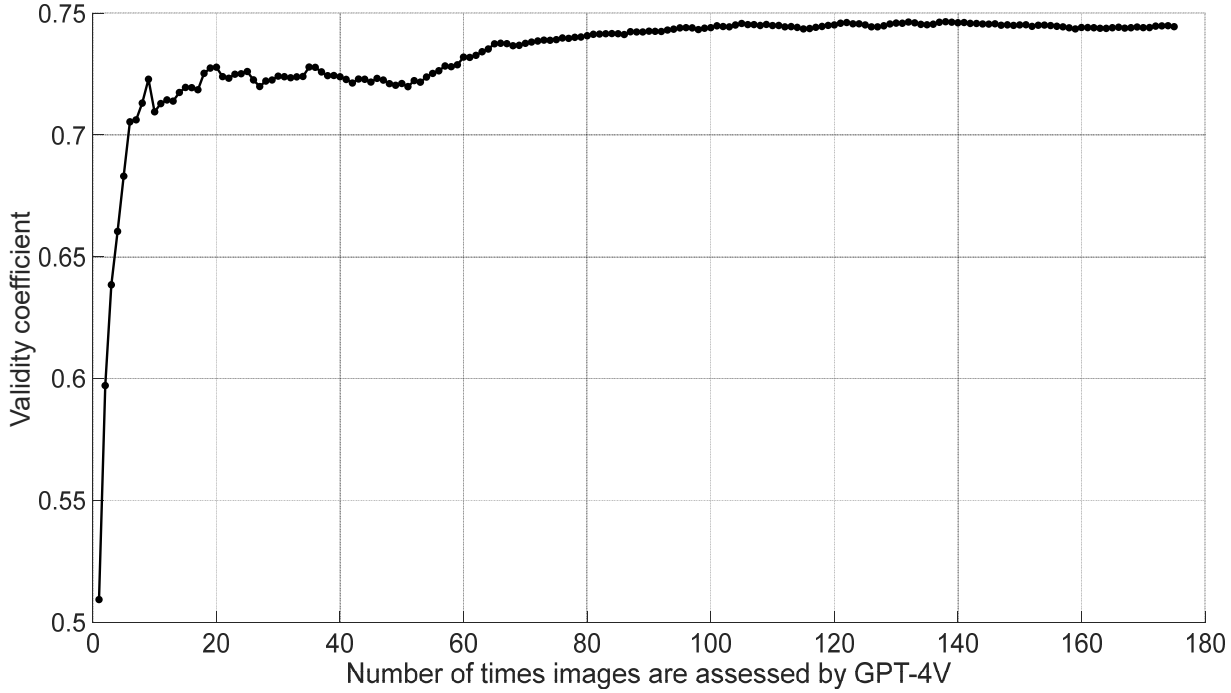

Figure 4. Correlation coefficient between mean GPT-4V-based risk rankings, as obtained using the API, and the human risk scores. For each prompt, a random 4 of 210 images were assessed. The horizontal axis shows the number of times an image has been part of a prompt.

Table 1 shows the validity coefficients (*r*) for 24 different prompt texts. Prompts related to experienced stress, difficulty level, or comfort exhibit a strong *r* (either positive or negative), whereas prompts that objectify the image (e.g., in terms of obstacles, traffic density, visibility) resulted in an *r* closer to 0. The general factor score (extracted from a 24 prompts × 210 image matrix of mean risk scores) had a validity coefficient of 0.78 (*n* = 210). This is stronger than when prompting about risk directly (see Figure 4), thereby supporting H2.

Table 1.
Prompts, validity coefficients (r) (n = 210), mean number of times an image was part of a prompt (m), and factor loading (λ). The prompts are sorted by the factor loadings in descending order.

No.	**Prompt**	***r***	***m***	**λ**
1	*Imagine you're the driver. Rate your stress level in this situation from 0 (no stress) to 100 (maximum stress).	0.73	18.8	0.96
2	*Imagine you're the driver. Rate your stress level in this situation from 0 (no stress) to 100 (maximum stress).	0.73	18.8	0.96
3	Assess the degree of difficulty you perceive in this driving situation, with 0 being no difficulty and 100 being extremely difficult.	0.73	19.0	0.95
4	On a scale from 0 (completely calm) to 100 (extremely agitated), rate your agitation level while driving in this scenario.	0.74	16.7	0.94

5	**Imagine you are a driver. Assign a risk score to each of these 4 images, on a scale of 0 (no risk) to 100 (extreme risk)	0.70	18.8	0.92
6	Evaluate your level of discomfort in this driving situation, with 0 indicating no discomfort and 100 indicating extreme discomfort.	0.71	15.4	0.90
7	On a scale from 0 to 100, how risky does this situation in the dashcam footage appear to you?	0.67	18.4	0.88
8	Rate the level of focus a driver needs in this situation, from 0 (minimal focus) to 100 (maximum focus).	0.73	19.0	0.88
9	Assess the level of distraction present in this scene, with 0 being no distractions and 100 being highly distracting.	0.67	19.0	0.87
10	Evaluate the presence of obstacles on the road, with 0 indicating no obstacles and 100 indicating many significant obstacles.	0.62	18.8	0.86
11	How probable is a collision in this scenario, on a scale from 0 (improbable) to 100 (inevitable)?	0.69	17.8	0.84
12	What threat level do you assign to this dashcam image, where 0 is no threat and 100 is extreme threat?	0.61	18.3	0.77
13	How likely is interaction with pedestrians in this scenario, from 0 (not likely) to 100 (very likely)?	0.54	18.9	0.71
14	Assess the traffic density in this image on a scale from 0 (very light) to 100 (extremely heavy).	0.42	19.0	0.60
15	Assess the condition of the road in the image, where 0 means excellent condition and 100 indicates extremely poor condition.	0.44	18.7	0.58
16	On a scale from 0 (perfect visibility) to 100 (no visibility), rate the visibility in this dashcam image.	0.54	19.0	0.57
17	Rate the risk to pedestrians in this image from 0 (no risk) to 100 (extremely high risk).	0.13	18.9	0.20
18	How quick should a driver's reaction time be in this situation, from 0 (slow) to 100 (instant)?	-0.16	19.0	-0.19
19	Perceive the speed of vehicles here, rating it from 0 (stationary) to 100 (extremely fast).	-0.18	17.2	-0.28
20	Assess your level of ease in navigating this scenario, with 0 being very uneasy and 100 being completely at ease.	-0.65	17.2	-0.80
21	**How much risk do you perceive in this scenario, on a scale from 0 (extremely risky) to 100 (no risk at all)?	-0.63	19.0	-0.83
22	*How comfortable would you feel driving in this scenario, with 0 being extremely uncomfortable and 100 being very comfortable?	-0.75	18.9	-0.91

23	On a scale of 0 to 100, where 0 is not at all confident and 100 is extremely confident, how confident would you feel about your driving skills in this situation?	-0.76	17.6	-0.92
24	*How comfortable would you feel driving in this scenario, with 0 being extremely uncomfortable and 100 being very comfortable?	-0.74	19.0	-0.92

*This prompt was used twice.
**This prompt was manually generated instead of being generated by ChatGPT.

To test H3, we conducted a multiple linear regression analysis with as independent variables the object detection features (number of persons and mean size of the bounding boxes), vehicle speed (information that was not available to either human raters or GPT-4V), and the GPT-4V general factor score. The correlations between variables are shown in Table 2, while the results of the regression analysis for predicting human risk are shown in Table 3. All four predictor variables contributed significantly ($p < 0.05$) to the human risk scores, providing support for H3. The overall predictive correlation of the regression model was $r = 0.83$, stronger than for the GPT-4V general factor score alone, as illustrated in Figure 5.

Table 2.
Pearson product-moment correlation matrix of two YOLO-based features (number of persons, mean bounding box size), vehicle speed, human risk score, and GPT-4V general factor score (n = 210).

Variable	**Mean**	***SD***	**1**	**2**	**3**	**4**
1. Number of persons (#)	0.27	0.93				
2. Mean bounding box size (pixels)	62.77	48.81	0.06			
3. Vehicle speed (m/s)	9.05	5.37	-0.10	-0.41		
4. Human risk score (%)	32.64	8.09	0.33	0.54	-0.63	
5. GPT-4V general factor score	0.00	1.00	0.37	0.49	-0.54	0.78

Table 3.
Regression analysis results for predicting human risk score from computer-vision variables, vehicle speed, and GPT-4V general factor score (n = 210).

	Unstandardised *B*	**Standardised β**	***t***	***p***
Intercept	34.23			
Number of persons (#)	0.966	0.11	2.63	0.009
Mean bounding box size (pixels)	0.029	0.18	3.84	< 0.001
Vehicle speed (m/s)	-0.406	-0.27	-5.70	< 0.001
GPT-4V general factor score	4.086	0.51	9.47	< 0.001

Note. $F(4, 205) = 115.0$, $p < 0.001$, $r = 0.83$

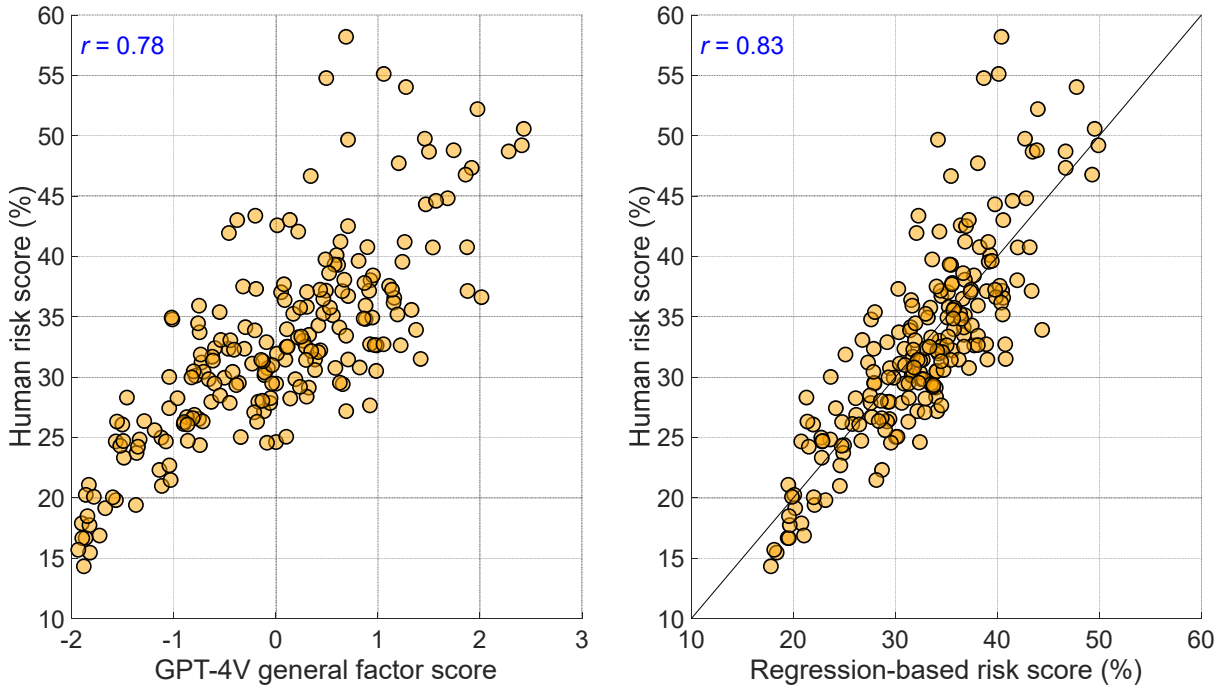

Figure 5. Scatter plot of risk in traffic images as rated by humans versus the GPT-4V general factor score (left) and versus risk predicted through multiple linear regression (right). Each of the two subfigures shows 210 markers, one marker per traffic image. The right subfigure also depicts a line of unity.

Discussion

Prior studies have demonstrated the capability of machine learning and computer vision techniques in analysing image datasets, including images from Google Street View, to predict factors such as scene complexity, safety, or poverty/wealth (Dubey et al., 2016; Fan et al., 2023; Guan et al., 2022; Nagle & Lavie, 2020; Naik et al., 2017; Zhang et al., 2018). Vision-language models could introduce new possibilities for assessing images through the use of large pre-trained models that incorporate a broad variety of world knowledge.

Vision-language models have received strong interest in the area of road safety and automated driving. This interest arises because current automated driving systems occasionally fail to understand the idiosyncrasies of certain traffic scenarios (Z. Yang et al., 2023). Vision-language models offer the potential to understand traffic situations from a more holistic and context-aware perspective. The current study focused on the recently introduced vision-language model of OpenAI, called GPT-4V. We used GPT-4V to judge the risk in forward-facing road images from a previously published dataset known as KITTI (Geiger et al., 2013).

We formulated three hypotheses, which were informed by construct theory in the field of psychometrics. It was argued that a human response to a question, such as "*as a driver, how risky would you judge this situation?*" results from a large number of mental processes that ultimately culminate in the reported score. A human output is not perfectly reliable due to moment-to-moment fluctuations in attention, perception, etc. Therefore, when measuring a construct ('perceived risk'), multiple different items must be used, and these should be administered not under slightly varied circumstances. Similarly, a language model does not produce consistent output either, and to ensure that its output is valid, the language model must be prompted multiple times, also known as the self-consistency method (Wang et al., 2023).

Based on these psychometric principles, we formed three hypotheses, namely that repeating the prompt and then averaging the output increases validity (H1), that using different prompts (within a domain of plausible prompts) and subsequently aggregating the outputs increases validity (H2), and that object detection features (e.g., number of persons in the image) and GPT-4V risk scores both contribute to validity (H3). Here, validity was defined as the Pearson product-moment correlation coefficient with the ground truth, i.e., the mean risk score of images based on a large number of human raters.

We found confirmation for all three hypotheses. Regarding H1, it was found that keeping the prompt text the same and repeating this prompt with different images contributed to a gradually increasing validity coefficient (see Figure 4). This provides support for the self-consistency method, as previously described in the literature (Tabone & De Winter, 2023; Wang et al., 2023). The inclusion of multiple images in random order induces output variability, consistent with the notion outlined in the Introduction stating that questionnaire items must be administered in parallel forms[1]. Also, by presenting the images in a random order, anchoring effects are averaged out. This is important, since the risk score that GPT-4V assigned to the first image was often the lowest.

Regarding H2, we found that different prompt texts yielded different validity coefficients (see Table 1), and that a general risk score, extracted through exploratory factor analysis, yielded a high validity coefficient of 0.78, higher than prompting about risk directly (see Figure 4). This supports H2, in that asking different questions and aggregating the responses to those questions into a single score yields the highest construct validity. A correlation coefficient of 0.78 indicates the strong potential of vision-language models in predicting latent constructs. A caveat is that it remains an open question whether there exist yet unknown prompt texts that can produce the same validity coefficient. For example, we found that outputs regarding 'confidence' strongly correlated with human risk scores (r = -0.76, see Table 1). Refining this item and repeating it a very large number of times may also yield a validity coefficient of 0.78 or stronger. An equivalent issue to 'finding the perfect prompt' exists in psychometrics. For example, in measuring the construct of human intelligence, it is common to administer a large battery of cognitive tests (Johnson et al., 2004). It is conceivable that an individual 'pure reasoning' test exists that provides a more predictive-valid measure of intelligence than an entire test battery; however, such a test has not yet been identified (Gignac, 2015).

Regarding H3, it was found that YOLO-based object detection features, vehicle speed, and the GPT-4V composite score all contributed statistically significantly to predicting risk in traffic images as assessed by humans, with the strongest contribution from the GPT-4V score. The predictive correlation of the regression model was r = 0.83. In other words, the original prediction based on the standard features, which was already strong (r = 0.75; De Winter et al., 2023), was strengthened by incorporating the GPT-4V-based assessment, thereby confirming H3.

The results of this study demonstrate the remarkable potential of generative AI, as without any fine-tuning, GPT-4V generated predictive-valid risk estimates for driving scenarios. It is important to acknowledge the limitations of the current study. Firstly, only static images were used. Future

[1]Regarding the findings in Figure 4, the most frequent risk percentage was "20", found in 17.9% of all numeric outputs. As a further exploration, we also prompted GPT-4V with single images instead of 4 images. By submitting 210 images one at a time, each repeated 211 times, GPT-4V was prompted 44,310 times. Using this method, the output "20" appeared in 73.7% of outputs. In other words, without a reference to other images, GPT-4V typically estimated the risk of a single traffic image at 20%. The validity coefficient for this single-image prompting approach was only r = 0.38, based on 211 repetitions per image.

research should use videos, so that the model can include movements of objects in its assessment. Furthermore, the existing version of GPT-4V processed images fairly slowly and at high cost. Regarding the four-image results shown in Figure 4, a total of 11,471 prompts were executed, comprising a total of 28.2 million input tokens (i.e., the images) and 0.17 million output tokens (i.e., the numeric scores). Using parallel prompting, the results were obtained in 1.8 hours, at a cost of $287.

Integrating vision-language models into real-time local systems such as dashcams or traffic warning systems is not yet feasible (but see Hwang et al., 2024). Future versions are expected to support local execution, improving inference speed and privacy, with local vision-language models, such as LLaVA, already available (Liu et al., 2023). Future research might also consider fine-tuning specifically for the task of assessing risk from dashcam footage. Future studies could also investigate whether the inclusion of additional explicit features, such as those related to right-of-way rules or the speeds of other vehicles, would enhance the ability of the model to predict human-assessed risk. The suggested capabilities of GPT-4V extend beyond merely processing camera images; options being considered in the literature include multimodality, such as evaluating and integrating Lidar data, HD maps, or other types of information flows, as well as using language models for user interaction and creating personalised driving experiences (Cui et al., 2024; Liao et al., 2024; Yan et al., 2024).

Apart from practical implications, the results in Table 1 may prove valuable for the field of psychology. Within traffic psychology, the perceived risk while driving is regarded as a key construct that underlies decision making (He et al., 2022; Kolekar et al., 2021; Näätänen & Summala, 1974; Wilde, 1982, 2013). While according to many perceived risk is a key determinant of driving behaviour (Kolekar et al., 2020; Wilde, 1982), others have argued that risk is not precisely what drivers respond to—certainly not objective risk in the form of probability of collision—but rather that the act upon perceived difficulty or effort (Fuller, 2005; Melman et al., 2018). The current results (Table 1) correspond with this and suggest that ‘confidence’ or ‘comfort’ align more closely with what drivers judge when asked to rate the risk in an image.

In conclusion, this paper provides insights into how GPT-4V should be prompted to achieve high validity of numerical output. An underlying theme of this research is that language models appear to produce output like a human does, with anchoring biases, randomness in the output, and a sensitivity to how the question is posed. Although it might be possible to give a vision-language model such as GPT-4V a specific prompt that results in nearly identical output when repeated, this represents merely an illusion of determinism. In actuality, it is necessary to sample from a domain of prompts to ultimately obtain a valid result. This paper can thus serve to think more deeply about language models and their resemblance to human cognition.

Data availability

The code used in this project can be found online at https://doi.org/10.4121/dfbe6de4-d559-49cd-a7c6-9bebe5d43d50

Acknowledgements

This research is funded by Transitions and Behaviour grant 403.19.243 (“Towards Safe Mobility for All: A Data-Driven Approach”), provided by the Dutch Research Council (NWO).

References

Ahrabian, K., Sourati, Z., Sun, K., Zhang, J., Jiang, Y., Morstatter, F., & Pujara, J. (2024). *The curious case of nonverbal abstract reasoning with multi-modal large language models.* arXiv. https://doi.org/10.48550/arXiv.2401.12117

Bellini-Leite, S. C. (2023). Dual Process Theory for Large Language Models: An overview of using Psychology to address hallucination and reliability issues. *Adaptive Behavior*. https://doi.org/10.1177/10597123231206604

Bing. (2023). Introducing the new Bing. https://www.bing.com/new

Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). *YOLOv4: Optimal speed and accuracy of object detection.* arXiv. https://doi.org/10.48550/arXiv.2004.10934

Bogdoll, D., Eisen, E., Nitsche, M., Scheib, C., & Zöllner, J. M. (2022). Multimodal detection of unknown objects on roads for autonomous driving. *Proceedings of the 2022 IEEE International Conference on Systems, Man, and Cybernetics*, 325–332, Prague, Czech Republic. https://doi.org/10.1109/SMC53654.2022.9945211

Charlton, S. G., Mackie, H. W., Baas, P. H., Hay, K., Menezes, M., & Dixon, C. (2010). Using endemic road features to create self-explaining roads and reduce vehicle speeds. *Accident Analysis & Prevention*, *42*, 1989–1998. https://doi.org/10.1016/j.aap.2010.06.006

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. R. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles.* New York: John Wiley.

Cui, C., Ma, Y., Cao, X., Ye, W., Zhou, Y., Liang, K., Chen, J., Lu, J., Yang, Z., Liao, K.-D., Gao, T., Li, E., Tang, K., Cao, Z., Zhou, T., Liu, A., Yan, X., Mei, S., Cao, J., … Zheng, C. (2024). A survey on multimodal large language models for autonomous driving. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 958–979, Waikoloa, HI.

Cui, C., Zhou, Y., Yang, X., Wu, S., Zhang, L., Zou, J., & Yao, H. (2023). *Holistic analysis of hallucination in GPT-4V(ision): Bias and interference challenges.* arXiv. https://doi.org/10.48550/arXiv.2311.03287

De Winter, J. C. F., Hoogmoed, J., Stapel, J., Dodou, D., & Bazilinskyy, P. (2023). Predicting perceived risk of traffic scenes using computer vision. *Transportation Research Part F: Traffic Psychology and Behaviour*, *93*, 235–247. https://doi.org/10.1016/j.trf.2023.01.014

Dubey, A., Naik, N., Parikh, D., Raskar, R., & Hidalgo, C. A. (2016). Deep learning the city: Quantifying urban perception at a global scale. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer Vision–ECCV 2016* (pp. 196–212). Cham: Springer. https://doi.org/10.1007/978-3-319-46448-0_12

Fan, Z., Zhang, F., Loo, B. P. Y., & Ratti, C. (2023). Urban visual intelligence: Uncovering hidden city profiles with street view images. *Proceedings of the National Academy of Sciences*, *120*, e2220417120. https://doi.org/10.1073/pnas.2220417120

Fu, Y., Peng, H., Sabharwal, A., Clark, P., & Khot, T. (2023). *Complexity-based prompting for multi-step reasoning.* arXiv. https://doi.org/10.48550/arXiv.2210.00720

Fuller, R. (2005). Towards a general theory of driver behaviour. *Accident Analysis & Prevention*, *37*, 461–472. https://doi.org/10.1016/j.aap.2004.11.003

Geiger, A., Lenz, P., Stiller, C., & Urtasun, R. (2013). Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, *32*, 1231–1237. https://doi.org/10.1177/0278364913491297

Gemini Team Google. (2023). *Gemini: A family of highly capable multimodal models.* ArXiv. https://doi.org/10.48550/arXiv.2312.11805

Gignac, G. E. (2015). Raven's is not a pure measure of general intelligence: Implications for g factor theory and the brief measurement of *g*. *Intelligence*, *52*, 71–79. https://doi.org/10.1016/j.intell.2015.07.006

Google. (2023). What's ahead for Bard: More global, more visual, more integrated. https://blog.google/technology/ai/google-bard-updates-io-2023

Guan, F., Fang, Z., Wang, L., Zhang, X., Zhong, H., & Huang, H. (2022). Modelling people's perceived scene complexity of real-world environments using street-view panoramas and open geodata. *ISPRS Journal of Photogrammetry and Remote Sensing*, *186*, 315–331. https://doi.org/10.1016/j.isprsjprs.2022.02.012

Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., Manocha, D., & Zhou, T. (2023). *HALLUSIONBENCH: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models.* ArXiv. https://doi.org/10.48550/arXiv.2310.14566

He, X., Stapel, J., Wang, M., & Happee, R. (2022). Modelling perceived risk and trust in driving automation reacting to merging and braking vehicles. *Transportation Research Part F: Traffic Psychology and Behaviour*, *86*, 178–195. https://doi.org/10.1016/j.trf.2022.02.016

Hou, I., Man, O., Mettille, S., Gutierrez, S., Angelikas, K., & MacNeil, S. (2024). More robots are coming: Large multimodal models (ChatGPT) can solve visually diverse images of Parsons problems. *Proceedings of the 26th Australasian Computing Education Conference*, 29–38, Sydney, Australia. https://doi.org/10.1145/3636243.3636247

Huang, J., Jiang, P., Gautam, A., & Saripalli, S. (2023). *GPT-4V takes the wheel: Evaluating promise and challenges for pedestrian behavior prediction. arXiv.* https://doi.org/10.48550/arXiv.2311.14786

Hwang, H., Kwon, S., Kim, Y., & Kim, D. (2024). *Is it safe to cross? Interpretable risk assessment with GPT-4V for safety-aware street crossing.* arXiv. https://doi.org/10.48550/arXiv.2402.06794

Jain, A., Del Pero, L., Grimmett, H., & Ondruska, P. (2021). *Autonomy 2.0: Why is self-driving always 5 years away*? arXiv. https://doi.org/10.48550/arXiv.2107.08142

Johnson, W., Bouchard, T. J., Jr., Krueger, R. F., McGue, M., & Gottesman, I. I. (2004). Just one *g*: Consistent results from three test batteries. *Intelligence*, *32*, 95–107. https://doi.org/10.1016/S0160-2896(03)00062-X

Kolekar, S., De Winter, J., & Abbink, D. (2020). Human-like driving behaviour emerges from a risk-based driver model. *Nature Communications*, *11*, 4850. https://doi.org/10.1038/s41467-020-18353-4

Kolekar, S., Petermeijer, B., Boer, E., De Winter, J. C. F., & Abbink, D. A. (2021). A risk field-based metric correlates with driver's perceived risk in manual and automated driving: A test-track study. *Transportation Research Part C: Emerging Technologies*, *133*, 103428. https://doi.org/10.1016/j.trc.2021.103428

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *Proceedings of the International Conference on Machine Learning*, 12888–12900.

Li, J., Zhang, Q., Yu, Y., Fu, Q., & Ye, D. (2024). *More agents is all you need.* arXiv. https://doi.org/10.48550/arXiv.2402.05120

Li, Y., Lin, Z., Zhang, S., Fu, Q., Chen, B., Lou, J.-G., & Chen, W. (2023). Making language models better reasoners with step-aware verifier. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 5315–5333, Toronto, Canada. https://doi.org/10.18653/v1/2023.acl-long.291

Li, Y., Wang, L., Hu, B., Chen, X., Zhong, W., Lyu, C., & Zhang, M. (2024). *A comprehensive evaluation of GPT-4V on knowledge-intensive visual question answering.* arXiv. https://doi.org/10.48550/arXiv.2311.07536

Liao, H., Shen, H., Li, Z., Wang, C., Li, G., Bie, Y., & Xu, C. (2024). GPT-4 enhanced multimodal grounding for autonomous driving: Leveraging cross-modal attention with large language models. *Communications in Transportation Research*, *4*, 100116. https://doi.org/10.1016/j.commtr.2023.100116

Little, T. D., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. *Psychological Methods*, *18*, 285–300. https://doi.org/10.1037/a0033266

Liu, H., Li, C., Li, Y., & Lee, Y. J. (2023). *Improved baselines with visual instruction tuning.* arXiv. https://doi.org/10.48550/arXiv.2310.03744

Liu, M., Chen, C., & Gurari, D. (2024). *An evaluation of GPT-4V and Gemini in online VQA.* arXiv. https://doi.org/10.48550/arXiv.2312.10637
Liu, Y., Wang, Y., Sun, L., & Yu, P. S. (2024). *Rec-GPT4V: Multimodal recommendation with large vision-language models.* arXiv. https://doi.org/10.48550/arXiv.2402.08670
Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.-W., Galley, M., & Gao, J. (2023). *MathVista: Evaluating mathematical reasoning of foundation models in visual contexts.* arXiv. https://doi.org/10.48550/arXiv.2310.02255
Lu, X., Liusie, A., Raina, V., Zhang, Y., & Beauchamp, W. (2024). *Blending is all you need: Cheaper, better alternative to trillion-parameters LLM.* arXiv. https://doi.org/10.48550/arXiv.2401.02994
Markus, K. A., & Borsboom, D. (2013). Reflective measurement models, behavior domains, and common causes. *New Ideas in Psychology*, *31*, 54–64. https://doi.org/10.1016/j.newideapsych.2011.02.008
McDonald, R. P. (2003). Behavior domains in theory and in practice. *Alberta Journal of Educational Research*, *49*, 212–230. https://doi.org/10.11575/ajer.v49i3.54980
Melman, T., Abbink, D. A., Van Paassen, M. M., Boer, E. R., & De Winter, J. C. F. (2018). What determines drivers' speed? A replication of three behavioural adaptation experiments in a single driving simulator study. *Ergonomics*, *61*, 966–987. https://doi.org/10.1080/00140139.2018.1426790
Näätänen, R., & Summala, H. (1974). A model for the role of motivational factors in drivers' decision-making. *Accident Analysis & Prevention*, *6*, 243–261. https://doi.org/10.1016/0001-4575(74)90003-7
Nagle, F., & Lavie, N. (2020). Predicting human complexity perception of real-world scenes. *Royal Society Open Science*, *7*, 191487. https://doi.org/10.1098/rsos.191487
Naik, N., Kominers, S. D., Raskar, R., Glaeser, E. L., & Hidalgo, C. A. (2017). Computer vision uncovers predictors of physical urban change. *Proceedings of the National Academy of Sciences*, *114*, 7571–7576. https://doi.org/10.1073/pnas.1619003114
Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory.* New York. NY: McGraw-Hill.
OpenAI. (2023). GPT-4 technical report. https://cdn.openai.com/papers/gpt-4.pdf
Qi, Z., Fang, Y., Zhang, M., Sun, Z., Wu, T., Liu, Z., Lin, D., Wang, J., & Zhao, H. (2023). *Gemini vs GPT-4V: A preliminary comparison and combination of vision-language models through qualitative cases.* arXiv. https://doi.org/10.48550/arXiv.2312.15011
Redmon, J., & Farhadi, A. (2018). *YOLOv3: An incremental improvement.* arXiv. https://doi.org/10.48550/arXiv.1804.02767
Salinas, A., & Morstatter, F. (2024). *The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance.* arXiv. https://doi.org/10.48550/arXiv.2401.03729
Sawaki, Y. (2010). Generalizability theory. In N. J. Salkind (Ed.), *Encyclopedia of research design.* Thousand Oaks, CA: Sage Publications. https://doi.org/10.4135/9781412961288
Senkaiahliyan, S., Toma, A., Ma, J., Chan, A.-W., Ha, A., An, K. R., Suresh, H., Rubin, B., & Wang, B. (2023). *GPT-4V(ision) unsuitable for clinical care and education: A clinician-evaluated assessment.* medRxiv. https://doi.org/10.1101/2023.11.15.23298575
Tabone, W., & De Winter, J. C. F. (2023). Using ChatGPT for human-computer interaction: A primer. *Royal Society Open Science*, *10*, 231053. https://doi.org/10.1098/rsos.231053
Tang, R., Zhang, X., Ma, X., Lin, J., & Ture, F. (2023). *Found in the middle: Permutation self-consistency improves listwise ranking in large language models.* arXiv. https://doi.org/10.48550/arXiv.2310.07712
Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., & Xie, S. (2024). *Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs.* arXiv. https://doi.org/10.48550/arXiv.2401.06209

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2023). *Self-consistency improves chain of thought reasoning in language models.* arXiv. https://doi.org/10.48550/arXiv.2203.11171

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in neural information processing systems, Vol. 35* (pp. 24824–24837). Curran Associates, Inc. https://doi.org/10.48550/arXiv.2201.11903

Wen, L., Yang, X., Fu, D., Wang, X., Cai, P., Li, X., Ma, T., Li, Y., Xu, L., Shang, D., Zhu, Z., Sun, S., Bai, Y., Cai, X., Dou, M., Hu, S., Shi, B., & Qiao, Y. (2023). *On the road with GPT-4V(ision): Early explorations of visual-language model on autonomous driving.* arXiv. https://doi.org/10.48550/arXiv.2311.05332

Wilde, G. J. S. (1982). The theory of risk homeostasis: implications for safety and health. *Risk Analysis*, *2*, 209–225. https://doi.org/10.1111/j.1539-6924.1982.tb01384.x

Wilde, G. J. S. (2013). Homeostasis drives behavioural adaptation. In C. M. Rudin-Brown & S. L. Jamson (Eds.), *Behavioural adaptation and road safety: Theory, evidence and action* (pp. 61–86). Boca Raton, FL: CRC Press.

Wu, C., Lei, J., Zheng, Q., Zhao, W., Lin, W., Zhang, X., Zhou, X., Zhao, Z., Zhang, Y., Wang, Y., & Xie, W. (2023). *Can GPT-4V(ision) serve medical applications? Case studies on GPT-4V for multimodal medical diagnosis.* arXiv. https://doi.org/10.48550/arXiv.2310.09909

Yan, A., Yang, Z., Zhu, W., Lin, K., Li, L., Wang, J., Yang, J., Zhong, Y., McAuley, J., Gao, J., Liu, Z., & Wang, L. (2023). *GPT-4V in wonderland: Large multimodal models for zero-shot smartphone GUI navigation.* arXiv. https://doi.org/10.48550/arXiv.2311.07562

Yan, X., Zhang, H., Cai, Y., Guo, J., Qiu, W., Gao, B., Zhou, K., Zhao, Y., Jin, H., Gao, J., Li, Z., Jiang, L., Zhang, W., Zhang, H., Dai, D., & Liu, B. (2024). *Forging vision foundation models for autonomous driving: Challenges, methodologies, and opportunities.* arXiv. https://doi.org/10.48550/arXiv.2401.08045

Yang, J., Zhang, H., Li, F., Zou, X., Li, C., & Gao, J. (2023). *Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4V.* arXiv. https://doi.org/10.48550/arXiv.2310.11441

Yang, Z., Li, L., Lin, K., Wang, J., Lin, C. C., Liu, Z., & Wang, L. (2023). *The dawn of LLMs: Preliminary explorations with GPT-4V(ision).* arXiv. https://doi.org/10.48550/arXiv.2309.17421

Ye, Q., Xu, H., Ye, J., Yan, M., Liu, H., Qian, Q., Zhang, J., Huang, F., & Zhou, J. (2023). *mPLUG-Owl2: Revolutionizing multi-modal large language model with modality collaboration.* arXiv. https://doi.org/10.48550/arXiv.2311.04257

Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., … Chen, W. (2023). *MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI.* arXiv. https://doi.org/10.48550/arXiv.2311.16502

Zhang, C., & Wang, S. (2024). *Good at captioning, bad at counting: Benchmarking GPT-4V on Earth observation data.* arXiv. https://doi.org/10.48550/arXiv.2401.17600

Zhang, D., Yang, J., Lyu, H., Jin, Z., Yao, Y., Chen, M., & Luo, J. (2024). *CoCoT: Contrastive chain-of-thought prompting for large multimodal models with multiple image inputs.* arXiv. https://doi.org/10.48550/arXiv.2401.02582

Zhang, F., Zhou, B., Liu, L., Liu, Y., Fung, H. H., Lin, H., & Ratti, C. (2018). Measuring human perceptions of a large-scale urban region using machine learning. *Landscape and Urban Planning*, *180*, 148–160. https://doi.org/10.1016/j.landurbplan.2018.08.020

Zhang, X., Lu, Y., Wang, W., Yan, A., Yan, J., Qin, L., Wang, H., Yan, X., Wang, W. Y., & Petzold, L. R. (2023). *GPT-4V(ision) as a generalist evaluator for vision-language tasks.* arXiv. https://doi.org/10.48550/arXiv.2311.01361

Zhou, X., & Knoll, A. C. (2024). *GPT-4V as traffic assistant: An in-depth look at vision language model on complex traffic events.* arXiv. https://doi.org/10.48550/arXiv.2402.02205

Appendix

Figure A1. Results of YOLOv4 for 2 of the 210 images.