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

3 November 2023

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

In late September 2023, OpenAI launched the much-anticipated image-to-text capabilities of ChatGPT, also referred to as GPT-4V. To date, there are few formal evaluations of GPT-4V available. In this study, we applied GPT-4V to forward-facing traffic images, where GPT-4V was prompted to arrange these images in terms of risk to the driver. A total of 210 images were ranked from low to high risk, and the correlation coefficient was determined with human ratings of the same images as established in a previous study. Across the 210 images, GPT-4V showed a strong zero-order correlation with the human risk ratings (r = 0.69). Combined with traditional computer vision features (number of detected individuals in the traffic image, average size of bounding boxes) as well as the current speed of the vehicle, the predictive value was r = 0.79. The current results suggest that GPT-4V output adds predictive value by incorporating context, something traditional computer vision methods do not incorporate. It is expected that powerful applications, such as real-time feedback systems, will become feasible if the inference time of large language models is reduced from multiple seconds to a sub-second level.

Introduction

In late September 2023, OpenAI introduced image-to-text functionality for ChatGPT Plus users. While sophisticated image-to-text software like BLIP (Li et al., 2022) and features within Bard (Google, 2023) and Bing Chat (Bing, 2023) were already available, ChatGPTs image-to-text capabilities, also known as GPT-4 Vision or GPT-4V, were highly anticipated due to the impressive quality depicted in earlier example demonstrations (OpenAI, 2023).

A small number of reports have already explored the capabilities of GPT-4V. It has been shown that GPT-4V can understand medical images and engage in object localization, scene text and chart reading, multi-sequence imaging, and processing abstract visual stimuli, including Raven's progressive matrices, amongst others (Wu et al., 2023; Z. Yang et al., 2023). In another report, a prompting strategy was detailed in which images are first segmented before being used as input for GPT-4V (J. Yang et al., 2023). Additionally, Lu et al. (2023) demonstrated that GPT-4V performs better in solving various visual mathematical problems than competing models like Bard.

However, while various demonstrations of GPT-4V usage can be found online, to date, formal evaluations of GPT-4V are still limited in number. One of the problems is that there is no API access available, and therefore it is not possible to systematically prompt GPT-4V. Moreover, ChatGPT features randomness in the output, which further complicates a formal evaluation of GPT-4V.

In this paper, we present a quantitative assessment of GPT-4V concerning the identification of risk in forward-facing photographs from the perspective of the vehicle (traffic images). Our analysis draws on a prior study (De Winter et al., 2023) wherein crowdworkers evaluated the risk of traffic images from the KITTI dataset (Geiger et al., 2013), which were taken by a camera mounted on the roof of a car while driving on German roads. In total, 210 images were rated by an average of about 650 participants per image. Based on these ratings from participants on a scale ranging from 0 (no risk) to 10 (extreme risk), an average risk score was computed for each image.

De Winter et al. (2023) also investigated whether the risk in the images was predictable based on features extracted by a pretrained YOLO object detection algorithm (Bochkovskiy et al., 2020; Redmon & Farhadi, 2018). The analysis showed that the number of people detected in the image (r = 0.33) and the average size of the bounding boxes (r = 0.54) were indicative of the level of risk. The driving speed was negatively predictive (r = -0.63), which can be explained by the phenomenon of risk compensation (a less strict variant of risk homeostasis, which posits that drivers adjust their speed to keep their perceived risk level constant; Wilde, 1982, 2013). Some situations, like empty roads, present such low risk that people can drive at the maximum allowed speed without it being high risk. Conversely, complex traffic environments, like city centers, lead people to drive slowly. This same principle underpins the concept of self-explaining roads, as applied in residential areas. Such environments do not afford fast driving and ensure that people drive slowly (Charlton et al., 2010).

Through a regression analysis, these three measures combined (number of people, size of bounding boxes, and vehicle speed when the image was taken) were found to be strongly predictive of the risk level provided by the crowdworkers (r = 0.75). Excluding the speed variable, the prediction was slightly weaker but still substantial (r = 0.62) (De Winter et al., 2023).

One might question why the prediction derived from the YOLO computer vision features was not more strongly indicative of the risk ratings determined by human evaluators, showing overall correlations of 'only' r = 0.62 (without speed) and r = 0.75 (with speed). Our prior explanation posited that the YOLO features do not account for contextual information (De Winter et al., 2023). For instance, an image of a railroad crossing was perceived as relatively hazardous by the human evaluators, whereas the YOLO object detection algorithm could not detect this, and did not understand the broader situation.

In this study, we explored whether ChatGPT Vision could more accurately gauge the risk in the traffic images compared to our earlier object detection computer vision approach. One challenge lies in the inherent variability in ChatGPT's output (which is normally modifiable via the API through the 'temperature parameters') and the fact that specific characteristics of the prompt (such as the order in which items are presented) influence the resulting output (e.g., Tabone & De Winter, 2023). Hence, we opted for a bootstrapping method, in which the same image was submitted multiple types but with the images submitted in a different order each time.

Methods

In this study, it was chosen to upload the images to the ChatGPT web interface in groups of 10, randomly selected from the total of 210 images. Due to the web interface having a limit of about 90 prompts that are accepted in a 3-hour interval, this proved to be a convenient way to process a large number of images. The decision to submit ten images at the same time was based on trial and error, aiming to maximize the number of images submitted in a single prompt while maintaining the quality of GPT-4V's output. When grouping 20 or more images, GPT-4V did not rank them logically and consistently placed the final images in the same sequence.

The 10 images were grouped into one large picture, 2 traffic images wide and 5 images high, with a total resolution of 2544×1995 pixels (including a small white margin around each individual traffic image). The speed of the car was not provided as part of the prompt or on the picture, because previous pilot tests with a total of 3 separate sets of 10 pictures each showed that this did not aid in generating a stronger prediction of the crowdworkers' perceived risk. The 10 images were grouped using MATLAB R2023a (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. 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 1. 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 batches)
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 batches)
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 batches)
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 batches)

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

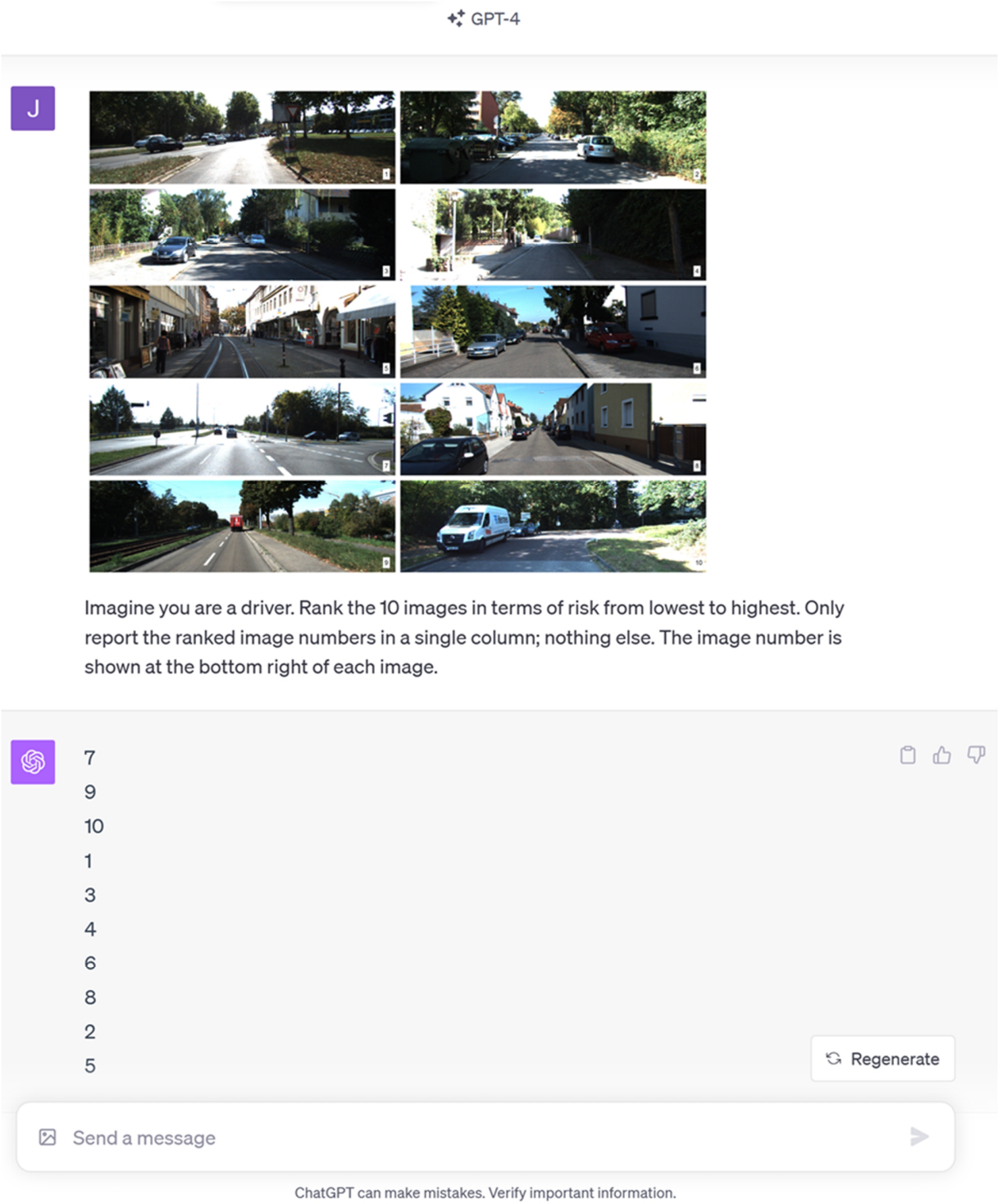

Figure 1. Example prompt and output of ChatGPT-4 Vision (GPT-4V). The prompt uses a random selection of 10 of 210 images. The 10 images were combined into a single image.

For every output, we determined the risk rank per image from 1 to 10, and calculated the mean rank per image (n = 210). The GPT-4V mean ranks 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 rating 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.

Additionally, we conducted a linear regression analysis to determine whether GPT-4V scores have added value compared to earlier computer vision metrics (number of detected people in the image, mean size of the bounding box), and vehicle speed.

Results

Analysis of the correlation between the mean ranks per image and corresponding crowdsourcing perceived risk did not show a clear difference between the four different prompts, with Pearson product-moment correlations (r) being 0.68, 0.59, 0.59, and 0.68 (n = 210) for Prompts 1–4. The mean ranks of the four prompts intercorrelated strongly, ranging from r = 0.74 (between Prompts 1 & 3) to r = 0.87 (between Prompts 1 & 4) (n = 210). Due to the lack of clear differences in output quality between the four prompting methods, it was decided to group all 974 instances where we prompted ChatGPT.

Figure 2 displays a scatter plot with 210 data points, 1 marker per image, illustrating the relationship between GPT-4V risk and human risk based on crowdsourcing. The corresponding Pearson product-moment correlation is r = 0.69 (n = 210).

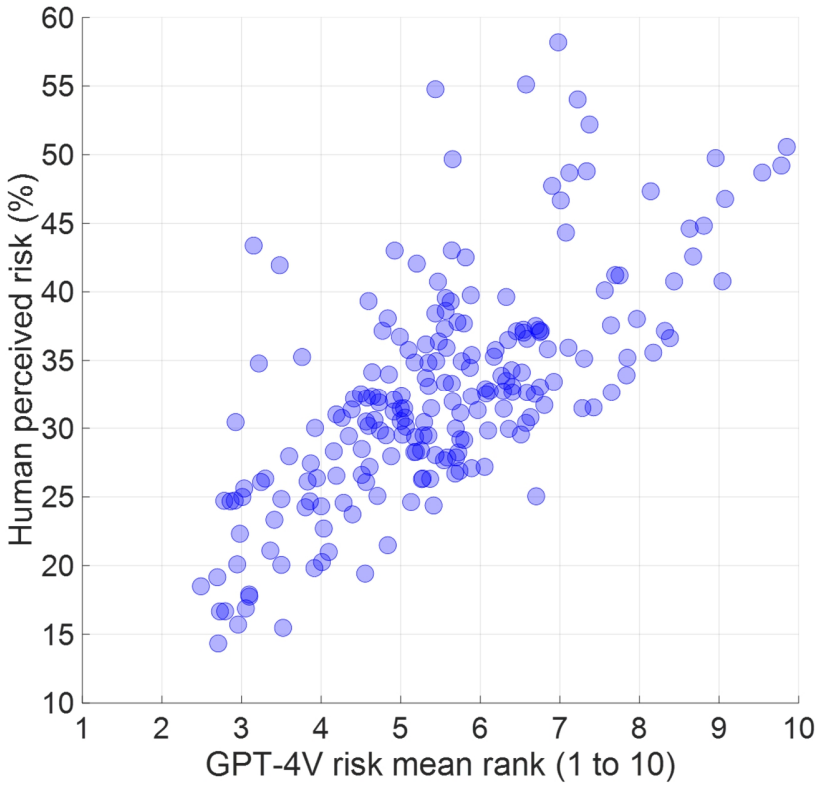

Figure 2. Scatter plot of risk in traffic images as rated by humans versus rated by GPT-4V.

Table 1 shows a correlation matrix between image-related metrics. An interesting point is that the correlation coefficient between the number of people, bounding box size, and speed of the vehicle with the GPT-4V risk (0.40, 0.49, -0.53) fairly matches the same correlation associated with the

risk assessed by humans (0.33, 0.54, -0.41). This suggests that GPT-4V made the risk assessment in a manner similar to how humans rate risk from images.

Table 1.
Pearson product-moment correlation matrix of two YOLO-based features (number of persons, bounding box size), vehicle speed, human risk scores based on crowdsourcing, and GPT-4V risk scores (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. Speed (m/s)	9.05	5.37	-0.10	-0.41		
4. Human risk (%)	32.64	8.09	0.33	0.54	-0.63	
5. GPT-4V risk mean rank (1 to 10)	5.50	1.55	0.40	0.49	-0.53	0.69

A question that arises is to what extent the risk scores produced via GPT-4V have added value in predicting human risk ratings compared to default computer vision measures (number of people, size of the bounding boxes), as well as the speed of the vehicle. For this purpose, we conducted a linear regression analysis (see Table 2), which showed that all four variables were statistically significant predictors of risk assessed by human raters, with a strong β coefficient (0.35) for the contribution of GPT-4V. The total predictive correlation of the regression model is $r = 0.788$, compared to $r = 0.746$ when the GPT-4V score was not included. The predictive correlation of the regression model is illustrated in Figure 3.

Table 2.
Regression analysis results for predicting human risk (%) from computer-vision variables, vehicle speed, and GPT-4V risk (n = 210).

	Unstandardized *B*	**Standardized β**	***t***	***p***
Intercept	24.53			
Number of persons (#)	1.271	0.15	3.06	0.003
Mean bounding box size (pixels)	0.036	0.22	4.32	< 0.001
Speed (m/s)	-0.506	-0.34	-6.45	< 0.001
GPT-4V mean rank (1 to 10)	1.830	0.35	5.85	< 0.001

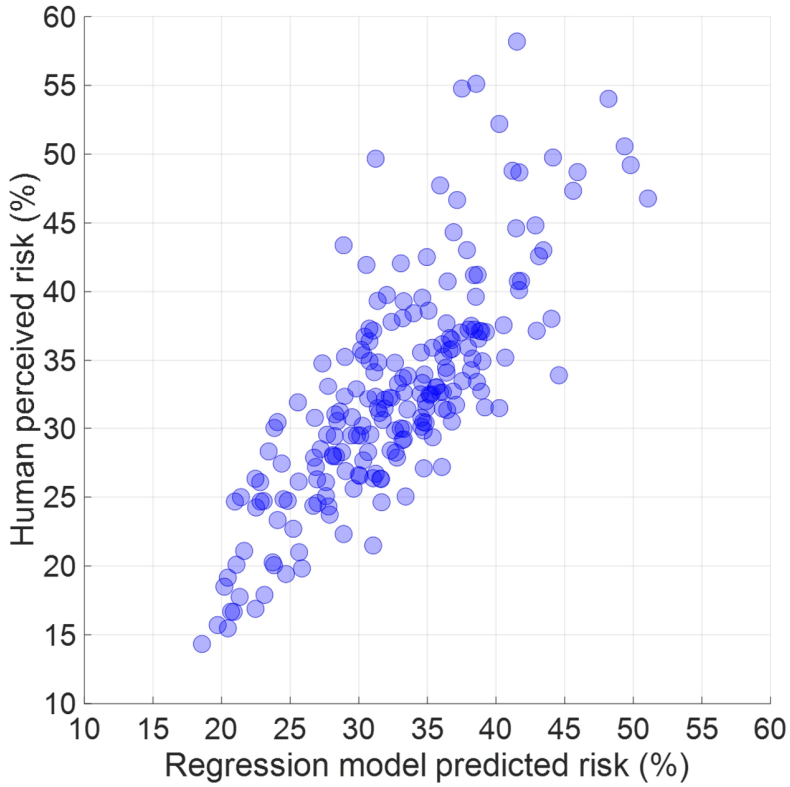

Figure 3. Scatter plot of risk in traffic images as rated by humans versus predicted through linear regression.

Discussion

This study, as one of the first, explored the potential of the image functionality of ChatGPT, also referred to as GPT-4 Vision or GPT-4V. We used GPT-4V to assess the risk associated with forward-facing road images from a previously published dataset known as KITTI (Geiger et al., 2013). It is important to note that within traffic psychology, the perceived risk level 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). Hazard perception tests, resembling the images assessed in this work, requiring the candidate to indicate the appropriate driver action (brake, release the gas, or do nothing) in a given situation, are employed in the official Dutch driving test to evaluate whether a candidate is a safe driver (CBR, 2023). The task of assessing risk is inherently challenging. However, we observed a strong prediction of GPT-4V-based risk for human risk, evident in both zero-order correlation analysis (r = 0.69) and as an incremental contribution in a multiple regression analysis (β = 0.35).

It should be noted that GPT-4V (and ChatGPT in general) is not adept at providing consistent outputs. A point of attention is that ChatGPT's output is sensitive to slight variations in the prompt, a situation exacerbated by the ChatGPT web interface that introduces deliberate randomness to its output. There were also order effects. For example, for Prompt 1, the tenth image in the batch (i.e., the one in the bottom right) was ranked as the least risky by ChatGPT in 37% (78 out of 210) of the cases, while the first image (i.e., the one in the top left) was seen as the most risky in only 3% of cases (6 out of 210), despite the images being randomly sorted within each batch. The randomness and order effects is why we applied a bootstrapping method (see Tabone & De Winter, 2023), an approach that resembles a previously published self-consistency method in the

use of GPT (Wang et al., 2022). In our study, each image was evaluated many (46 of 47) times, and the images were presented in random order in each prompt.

The results demonstrate the remarkable potential of generative AI and the future promise of general-purpose models. Our research showed that without fine-tuning, the base model generated predictive-valid risk estimates for driving scenarios. It is important to acknowledge the current limitations: the existing version of GPT-4V processes fairly slowly, with responses taking 5 to 15 seconds for a single prompt of 10 combined images. Consequently, integrating such algorithms into real-time, local systems such as dashcams or traffic warning systems is not yet worthwhile. In an additional analysis (Appendix A), we investigated if a text-only approach, where YOLO's image labels were fed directly into GPT-4's text model, could approach the performance of GPT-4V, and concluded that the performance in this approach, while also decently correlated with human ratings (r = 0.53), was not as strong as the capabilities demonstrated by GPT-4V.

Despite the aforementioned limitations, this research hints at the future. It is anticipated that with a few more years of development and the growth of local computational capabilities and potentially new model architectures that can be run locally, there will be a move towards employing general-purpose algorithms to address domain-specific and real-time challenges. Future research might also consider using large-language models, such as GPT-4V or others, that are specifically fine-tuned for the task of assessing risk from dashcam footage. It could also be investigated whether providing more explicit features, such as those related to right-of-way rules or the speed of other vehicles, enhances the prediction strength of human risk.

Acknowledgments

This research is funded by Transitions and Behaviour grant 403.19.243 ("Towards Safe Mobility for All: A Data-Driven Approach"), provided by the Netherlands Organization for Scientific Research (NWO).

References

Anand, Y., Nussbaum, Z., Duderstadt, B., Schmidt, B., & Mulyar, A. (2023). GPT4All: Training an assistant-style chatbot with large scale data distillation from GPT-3.5-Turbo [Computer software]. GitHub. https://github.com/nomic-ai/gpt4all

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://arxiv.org/abs/2004.10934

CBR. (2023). Theorie-examen auto. Leren en oefenen [Car theory exam. Learn and practice]. https://www.cbr.nl/nl/rijbewijs-halen/auto/theorie-examen-auto/leren-en-oefenen.htm [Access date: 28 October 2023]

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

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

Geiger, A., Lenz, P., Stiller, C., & Urtasun, R. (2013). Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, *32*(11), 1231–1237. https://doi.org/10.1177/0278364913491297

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

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

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.

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

Martínez Toro, I., Gallego Vico, D., & Orgaz, P. (2023). PrivateGPT [Computer software]. GitHub. https://github.com/imartinez/privateGPT

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

OpenAI. (2023). GPT-4 technical report. https://cdn.openai.com/papers/gpt-4.pdf

Redmon, J., & Farhadi, A. (2018). *YOLOv3: An incremental improvement.* arXiv. https://doi.org/10.48550/arXiv.1804.02767

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

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2022). *Self-consistency improves chain of thought reasoning in language models.* arXiv. https://doi.org/10.48550/arXiv.2203.11171

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

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 lmms: Preliminary explorations with GPT-4V(ision).* arXiv. https://doi.org/10.48550/arXiv.2309.17421

Appendix:
Assessing performance of GPT-4 on image labels produced by YOLO

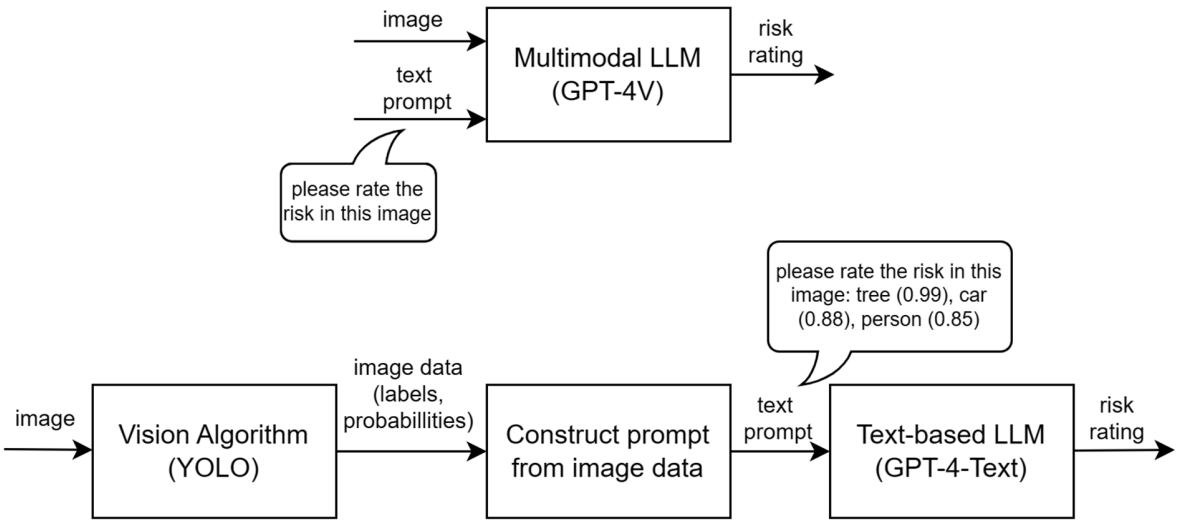

Figure A1. Conceptual diagram of GPT-4V pipeline (top) and the GPT-4-Text pipeline (bottom) proposed in this appendix. Example prompts are given in text balloons. Note that the analysis uses batched descriptions in the prompts.

In this section we investigate how GPT-4V, where images are directly fed to the model, compares to a more simple case where image labels are generated by YOLO, and risk is estimated by the GPT-4 text-only model (hereafter named GPT-4-Text). Figure A1 illustrates this concept.

Each image is described in text using the image labels and corresponding confidence scores; for example: "*car (0.99), truck (0.97), truck (0.93), …".* GPT-4's text model can then produce a risk estimate, based on this text description.

One benefit of this approach is its low computational cost for producing image labels. Moreover, only a response from a text-based Large Language Model (LLM) is needed to produce the risk estimate. A rapidly growing set of open-source projects aimed at executing LLMs locally (e.g., GPT4All by Anand et al., 2023 or PrivateGPT by Martínez Toro et al., 2023) could open ways to more mobile applications for the driving domain. Consequently, it is worth investigating to what extent text-based LLMs (in our case, GPT-4's text model) paired with computer vision models can approach GPT-4V's performance.

We accessed OpenAI's gpt-4-0613 text model through the API with the following prompt:

```
You are assisting with labeling a dataset composed of images captured from the ego-
perspective of a vehicle. These images have been processed through a YOLO image
recognition algorithm. Each line in the dataset represents an image and lists the objects
detected in the driving scene, along with the associated probabilities for each object.

Your task is to assign a risk score to each image, using a scale from 0 to 100. A score
of 0 signifies no risk, while a score of 100 indicates an extremely high-risk situation.
```

```
Consider factors such as the type of objects identified and their probabilities when
determining the risk score.

Please return the risk scores as a comma-separated list, containing only numerical values
corresponding to each image. Do not include any other text or labels in your output.

1. car (0.97), car (0.97), car (0.91), car (0.85), truck (0.67)
2. car (0.94), truck (0.78), car (0.75), car (0.74), bus (0.71), car (0.63)
3. no objects detected
(...)
25. car (0.99), car (0.97), car (0.95), car (0.88), truck (0.77), car (0.76)
```

A total of 275 prompts were submitted, with each prompt containing 25 images randomly selected from a pool of 210 (for a similar bootstrapping method, see Tabone & De Winter, 2023). On average, each image received 30.1 risk ratings, ranging from a minimum of 14 to a maximum of 45. The average of these scores per image was used for correlation analysis (Table A1) and regression analysis (Table A2).

Table A1. *Pearson product-moment correlation matrix of two YOLO-based features (number of persons, bounding box size), vehicle speed, human risk scores based on crowdsourcing, and GPT-4-Text risk scores (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. Speed (m/s)	9.05	5.37	-0.10	-0.41			
4. Human risk (%)	32.64	8.09	0.33	0.54	-0.63		
5. GPT-4V risk mean rank (1 to 10)	5.50	1.55	0.40	0.49	-0.53	0.69	
6. GPT-4-Text risk (%)	55.0	29.4	0.26	0.48	-0.47	0.53	0.70

Table A2. *Regression analysis results for predicting human risk (%) from computer-vision variables, vehicle speed, and GPT-4-Text risk (n = 210).*

	Unstandardized *B*	**Standardized β**	***t***	***p***
Intercept	32.91			
Number of persons (#)	2.080	0.24	5.03	< 0.001
Mean bounding box size (pixels)	0.047	0.29	5.30	< 0.001
Speed (m/s)	-0.639	-0.42	-7.93	< 0.001
GPT-4-Text risk (%)	0.036	0.13	2.26	0.025

We observe a correlation of $r = 0.70$ between the GPT-4-Text model scores and the GPT-4V model scores. The GPT-4-Text model has a correlation of $r = 0.53$ with human risk, which is weaker than the GPT-4V model's correlation of $r = 0.69$.

Additionally, the GPT-4-Text model risk score had a statistically significant predictive value ($\beta = 0.13$, $p = 0.025$) when integrated with computer-vision variables and speed in a regression analysis. However, its contribution is less substantial than the GPT-4V model's risk scores ($\beta = 0.35$, $p < 0.001$). The overall predictive correlation in the GPT-4-Text regression analysis was $r = 0.754$, compared to $r = 0.788$ with the GPT-4V model, and $r = 0.746$ when no GPT estimates were included.

It is concluded that the risk scores generated by the multimodal GPT-4V model more closely align with human-produced risk estimates than the GPT-4-Text model estimates do. This suggests that the GPT-4V model can offer a more comprehensive interpretation of the image displayed to it than can be achieved by simplifying the image to a sentence containing class labels. This

limitation of the text-reduction method is particularly evident for images where no objects were detected. In such cases, the text description was “no objects detected, and the GPT-4-Text model typically assigned a risk score of 0. This trend is evident by the dense grouping of points near the origin in the scatterplot shown in Figure A1. It is also noteworthy that human raters typically gave these images low ratings as well.

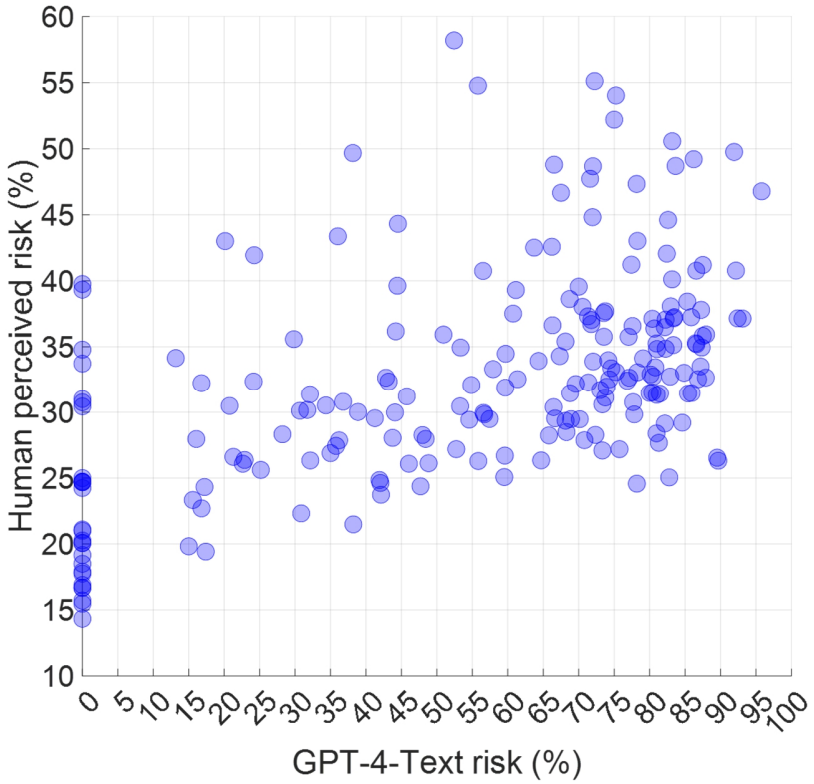

Figure A2. Scatter plot of risk in traffic images as rated by humans versus generated by GPT-4-Text

Despite its lower performance, the GPT-4-Text method may still have potential applications since its zero-order correlations are substantial. Future research could explore more advanced computer-vision algorithms that generate more detailed descriptions, and examine how local LLMs like GPT4All or PrivateGPT rate these image descriptions.