Cross or Nah? LLMs Get in the Mindset of a Pedestrian in front of Automated Car with an eHMI

MD SHADAB ALAM*, Eindhoven University of Technology, The Netherlands
PAVLO BAZILINSKYY, Eindhoven University of Technology, The Netherlands

Fig. 1. Ghibli-style personas for LLMs (image was generated using ChatGPT-4o).

This study examines the effectiveness of using large language model-based personas to evaluate external human-machine interfaces (eHMIs) in automated vehicles. Various models including miniCPM-V, LLaVA, LLaVA-LLaMA-3, Llama3.2 vision, Moondream, Bak-LLaVA, Granite3.2 vision, LLaVA-Phi3, Gemma 3, Deepseek-vl2, and ChatGPT-4o were used to simulate pedestrian perspectives.

*Corresponding Author

Authors' Contact Information: Md Shadab Alam, m.s.alam@tue.nl, Eindhoven University of Technology, Eindhoven, The Netherlands; Pavlo Bazilinskyy, p.bazilinskyy@tue.nl, Eindhoven University of Technology, Eindhoven, The Netherlands.

Manuscript submitted to ACM

Models assessed vehicle images with eHMIs, assigning scores from 0 (completely unwilling) to 100 (fully confident) regarding crossing decisions. Each model was compared with 15 trials with randomised image sequences, both with and without prior chat context, and the results were compared with crowdsourced human ratings. The findings indicate Gemma3: 27B performed better without chat history (r = 0.85), while ChatGPT-4o was superior when the historical context was included (r = 0.81). In contrast, models such as Deepseek-vl2 and BakLLaVA provided uniform confidence scores with memory context, while Llama3.2 vision failed entirely to produce outputs.

CCS Concepts: • **Computing methodologies** → **Machine learning approaches**; **Cross-validation**; Simulation theory; Image processing.

Additional Key Words and Phrases: Large language Models, Automated Cars, eHMI, Crowdsourcing

ACM Reference Format:
Md Shadab Alam and Pavlo Bazilinskyy. 2025. Cross or Nah? LLMs Get in the Mindset of a Pedestrian in front of Automated Car with an eHMI. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

In 2017, Google researchers proposed an influential work titled "`Attention is all you need`" [30], introducing the attention mechanism [2] to significantly enhance sequence-to-sequence (seq2seq) models [25]. This innovation paved the way for encoder-only architectures such as the Bidirectional Encoder Representations from Transformers (BERT) [12] and subsequently decoder-only models such as the Generative Pre-trained Transformer (GPT). Since then, numerous Large Language Models (LLMs) have emerged, tailored for specific tasks such as medical analysis [22] and document processing [36], as well as general purpose models such as DeepSeek [34] and ChatGPT (https://chatgpt.com). With the continuous growth in available data and advancements in computational resources, the performance and capabilities of these models have steadily improved.

Researchers have shown significant interest in evaluating whether these sophisticated language models can successfully pass rigorous evaluations such as the Turing test. In particular, Chat GPT-4.5 and LLaMa-3.1-405B have recently passed this test, with GPT-4.5 judged human in 73% of cases and LLaMa-3.1-405B in 56% of cases [16]. Further highlighting their capability, Bhardwaj et al. (2025) demonstrated that models like ChatGPT 3.0 can achieve an accuracy of 48.71% on some of the world's most challenging examinations, a figure that increases substantially to 77.69% with targeted training [5]. Similarly, Stengel et al. (2024) reported that Google's Bard LLM [26] outperformed human participants by correctly answering 62% of the European Board Examination in Neurological Surgery (EANS) questions in general and 69% when excluding IB-specific questions, compared to human scores of 59% (p = 0.67) and 59% (p = 0.42), respectively [24]. In particular, LLMs consistently performed best in theoretical questions, significantly exceeding human performance with scores of 79% for ChatGPT, 83% for Bing (https://www.bing.com), and 86% for Bard, compared to a human baseline of 60% (p = 0.03).

Traditional crowdsourcing strategies, such as putting the questionnaire on a website such as Appen (https://www.appen.com) or Amazon Mechanical Turk (https://www.mturk.com) have led researchers to frequently encounter issues with regard to data quality and adherence to experiment protocols, resulting in considerable data exclusion. This removal process can significantly impact the efficiency and cost-effectiveness of research. For example, de Winter et al. [11] conducted a crowdsourcing experiment involving 1918 participants who evaluated 100 images over a period of 19 days. However, they subsequently excluded 540 responses due to participants not complying with the guidelines

Fig. 2. Car equipped with an eHMI given as input to the LLMs.

provided or submitting responses from duplicate IP addresses. Similarly, Bazilinskyy et al. (2022) conducted a crowdsourced study examining interactions between cyclists and automated vehicles, initially collecting 2,000 responses but eventually discarding 740 responses for similar reasons [4]. Burnap et al. (2015) further demonstrated that even when participants adhere to protocols, the presence of non-expert evaluators and clusters of consistently incorrect judgments can degrade the reliability of the results [7]. Their findings suggest that crowdsourcing can fail to deliver accurate insights, particularly for tasks requiring specialised knowledge or domain-specific expertise.

Recently, researchers such as Driessen et al.[13], Bubeck et al.[6], Wu et al.[33], Argyle et al.[1], H"am"al"ainen et al.[14], and Wang et al. (2021) have demonstrated that LLMs can effectively approximate human opinions on subjective tasks [32]. This capability allows them to simulate human responses to crowdsourced questionnaires and interviews with impressive accuracy. Additionally, LLMs have shown promise as annotators, often exceeding human annotators recruited through crowdsourcing in terms of reliability and consistency. For example, He et al. (2024) illustrated that GPT-3.5, when adequately guided with clear instructions and examples, outperforms typical crowdsourced annotators [15]. Furthermore, studies by Li et al. (2025) indicate that LLMs substantially improve the processing of crowdsourced test results, particularly in tasks involving the clustering and deduplication of bug reports, further emphasising their utility in crowdsourced research settings [18].

1.1 Aim of Study

This study aims to explore the effectiveness of vision language models (VLM) in interpreting text-based external Human-Machine Interface (eHMI) messages displayed on automated vehicles (AVs), emphasising their potential to simulate

pedestrian decision-making processes. Specifically, we evaluate multiple VLM architectures, comparing their ability to accurately recognise and interpret eHMI messages, assess pedestrian crossing safety, and maintain performance consistency when historical conversational context (memory) is introduced. The performance of these models is benchmarked against large-scale human responses collected through crowdsourcing [3]. By investigating the impact of contextual memory on model accuracy, this research seeks to validate the reliability of VLMs as cost-effective and consistent substitutes for human participants in human-machine interaction (HMI) evaluations, ultimately forming design guidelines for effective pedestrian-AV communication interfaces.

2 Method

This study used a crowdsourced dataset initially compiled by Bazilinskyy et al. (2022), involving 1,438 participants who evaluated 227 distinct textual eHMIs displayed on an AV [3]. Participants indicated their willingness to cross the road by adjusting a slider scale ranging from 0, representing absolute unwillingness to cross, to 100, representing complete confidence in crossing safely. This dataset provided an essential human benchmark for the evaluation of the interpretability of eHMI messages using contemporary vision-language models (VLMs).

All images used in this research were sourced directly from the dataset established by Bazilinskyy et al. (2022) [3]. The dataset comprised a total of 227 unique images, each displaying distinct textual eHMI messages on an AV. The images were standardised at a resolution of 1024 × 598 pixels and stored in JPEG format.

A variety of LLMs equipped with vision processing capabilities were evaluated. These models were deployed across three distinct platforms: Ollama (https://ollama.com), Hugging Face (https://huggingface.co/models), and the OpenAI ChatGPT API (https://platform.openai.com/docs/overview). Table 1 summarises the models employed, detailing their base architecture and respective deployment platforms. The cost of using the ChatGPT API was 20 EUR. All other LLMs were used for free.

Table 1. Overview of the models used, their architecture, and deployment platform.

Model Name	**Architecture**	**Platform**
MiniCPM-V [35]	CPM-based	Ollama
LLaVA:13B [19]	LLaMA + CLIP	Ollama
LLaVA:34B [19]	LLaMA + CLIP	Ollama
LLaVA-LLaMA-3 [9]	LLaMA-3 + Visual Adapter	Ollama
Llama3.2-vision [29]	LLaMA 3.2 + Vision	Ollama
Moondream [31]	Distilled CLIP variant	Ollama
BakLLaVA [23]	LLaMA 2 + BLIP	Ollama
Granite3.2-vision [28]	IBM Granite vision model	Ollama
LLaVA-Phi3 [10]	Phi-3 + Vision module	Ollama
Gemma 3: 12B [27]	Gemma 3 vision-capable model	Ollama
Gemma 3: 27B [27]	Gemma 3 vision-capable model	Ollama
DeepSeek-VL2-Tiny [34]	DeepSeek VLM	Hugging Face
GPT-4o [20]	Multimodal transformer	OpenAI (ChatGPT API)

Each model received an identical series of images sequentially in a controlled, standardised order. The models were evaluated under two distinct conditions: without conversational history (memory) and with conversational history. For trials without memory, each image was accompanied by a uniform prompt requesting the models to observe the vehicle's displayed message, interpret its implications regarding pedestrian safety, and assign a numerical confidence

score (0-100). Below is the prompt used for models without conversation history and the first prompt used when delegated to work with conversation history.

> *Carefully observe the image of an automated vehicle and quote the exact text displayed on the vehicle. Briefly explain what this message implies regarding whether it is safe for a pedestrian to cross the street. Then assign a numerical confidence score from 0 (certainly unsafe) to 100 (certainly safe). Respond strictly in this format: Confidence: [numeric value] Meaning: [brief explanation].*

For trials incorporating memory, the first image was introduced with the initial prompt. From the second image and onwards, contextual memory was incorporated using LangChain software (https://www.langchain.com), preserving the prior conversation history as structured JSON data. The instructions in subsequent trials explicitly instructed the models to consider previous responses when evaluating pedestrian safety. Each trial was limited to a maximum of six historical conversations to ensure manageability and comparability, as See et al. (2019) showed, that it is a reasonable window for humans to maintain a coherent context and judge factors such as fluency, listening, and engagement [21].

(1) **Pre-image prompt**:
> *You are a pedestrian deciding whether to cross the road in front of this autonomous vehicle. Pay attention to any external text or signals displayed on the car.*

(2) **History prompt**:
> *Below is the conversation history from previous interactions (your past answers might influence your decision):*

Followed by the complete prompt-response history of earlier steps.

(3) **Final image prompt**:
> *Now, based on the current image details, please respond with a number from 0 to 100 indicating your confidence to cross the road(0 = no confidence, 100 = full confidence). Respond strictly in this format: Confidence: [numeric value] Meaning: [brief explanation].*

The evaluation used a modular system architecture to process and analyse responses (see Figure 3). To ensure response consistency and avoid interference between trials, model interactions occurred sequentially. Local model deployments via Ollama supported batch mode image processing. The DeepSeek-VL2 model code and weights were sourced directly from Hugging Face (https://huggingface.co/deepseek-ai/deepseek-vl2-tiny), while API-based models, such as GPT-4o, were accessed through synchronous HTTP POST requests using JSON payloads.

All evaluations were conducted on a computing system featuring an AMD Ryzen 9 7950X3D 16-core processor with 32GB RAM and a 16GB NVIDIA GeForce RTX 4080 graphics card, ensuring high performance, reproducibility, and compatibility for local lightweight model execution.

Upon completion of model evaluations, responses were meticulously analysed to extract numeric confidence scores using deepseek-r1: 14b https://huggingface.co/deepseek-ai/DeepSeek-R1. This was necessary due to variations in model output verbosity. The following is the prompt used for the extraction of the confidence score.

> *Read the following sentence carefully and extract the number mentioned in it. Only return the number (as digits), without any additional explanation or units.*
> *Sentence: "<model's previous response>"*

In instances where direct numeric extraction was not straightforward, a correction protocol was implemented, instructing the models explicitly to return the numeric value without additional explanation. This ensured accuracy and consistency across all model outputs. The responses were systematically stored for subsequent statistical analysis,

allowing comparison with human benchmarks and assessment of the interpretative precision of each model regarding pedestrian crossing decisions based on eHMI messages.

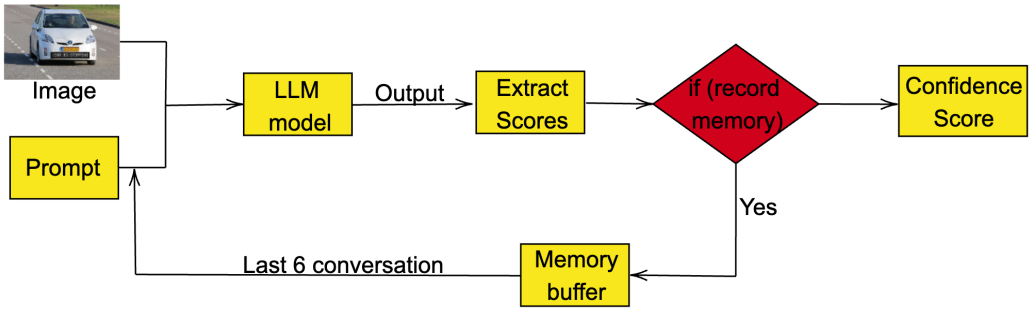

Fig. 3. Flow diagram of the system architecture showing image processing, prompting, model querying, response correction, and downstream analysis.

3 Results

The confidence outputs of each model were averaged for each image and compared with crowdsourced results [3]. Without conversation history, Gemma3 27B demonstrated the highest correlation with crowd-sourced responses, with a strong correlation with mean values (r = 0.85) and median values (r = 0.86). GPT-4o also showed high correlation coefficients, closely aligned with crowd-sourced responses of the mean (r = 0.84) and median (r = 0.85). In contrast, MiniCPM-V exhibited a lower correlation coefficient (r = 0.42), indicating weaker alignment with human scores.

When incorporating conversation history, GPT-4o retained the highest correlation coefficient (r = 0.81), indicating its superior ability to interpret eHMI cues in a conversational context. Gemma 3's correlation significantly decreased when using conversation history, suggesting that its performance was adversely impacted by prior conversational inputs. The MiniCPM-V correlation dropped drastically to -0.01, reflecting poor consistency under memory-based conditions. Additionally, models like Llama3.2-Vision did not generate output in scenarios involving conversation history, demonstrating severe limitations in handling contextual memory. Other models also exhibited peculiar behaviours under these conditions; for example, DeepSeek VL2 provided a constant confidence score of 75 regardless of the image presented, always accompanied by identical explanations. BakLLaVA consistently produced a confidence score of 0 without providing any explanatory context, indicating significant issues with interpretative consistency and precision.

4 Discussion

This study examines the nuanced performance of various LLMs in simulating pedestrian decision making in response to eHMI cues from AVs. Excluding BakLLaVA and DeepSeek (with conversation history enabled), our analysis of models including ChatGPT-4o, Gemma 3:27B, MiniCPM-V, and LLaMA3.2-vision reveals significant differences in the calibration of the risk assessment.

In particular, ChatGPT-4o and Gemma 3:27B exhibit strong alignment with human-like safety evaluations, particularly in settings without conversational history. Their superior performance in these memory-free scenarios is attributable to robust one-shot inference capabilities, enabling accurate extraction and interpretation of salient visual and textual

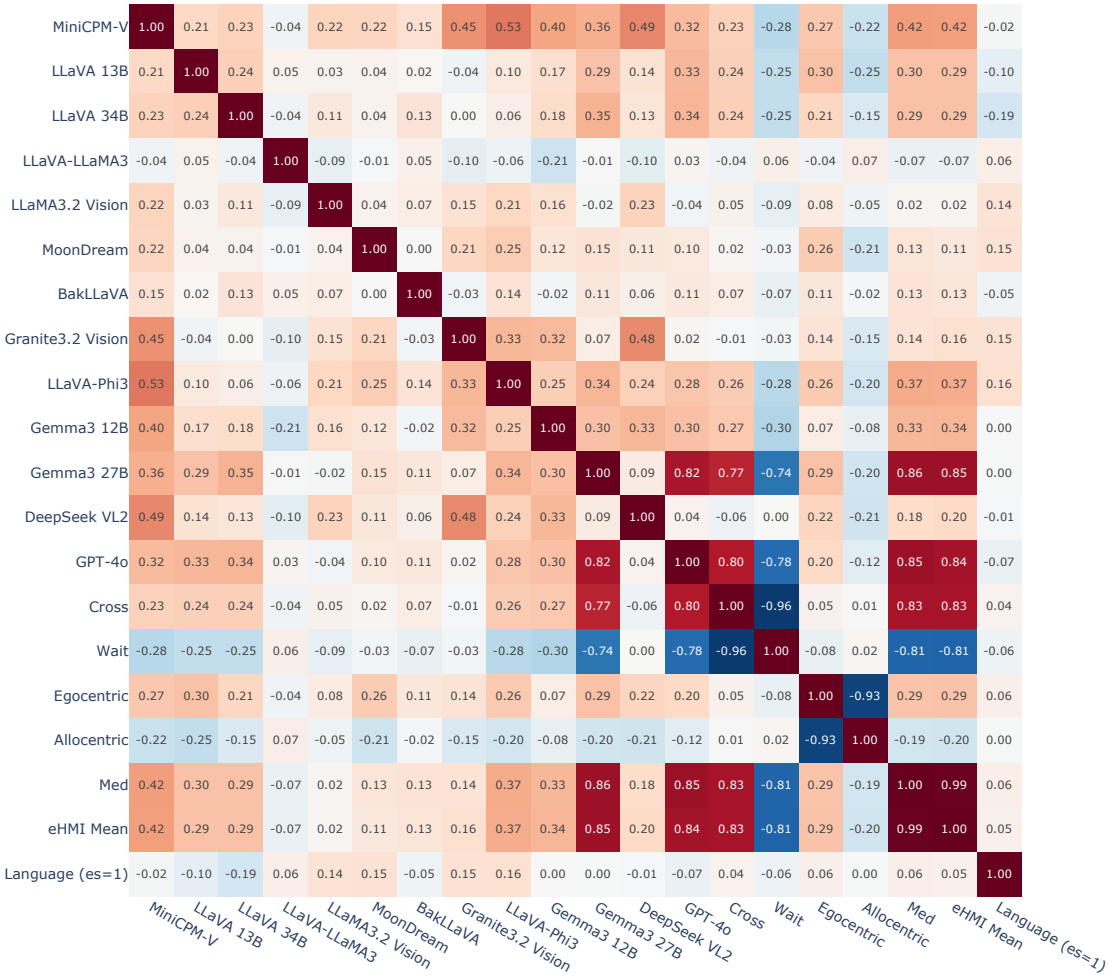

Fig. 4. Correlation matrix without memory.

cues without interference from prior dialogue [17]. The advanced architectures and high parameter counts of these models facilitate sophisticated multimodal integration, resulting in outputs that closely mirror human judgment. In contrast, other models demonstrate an overreliance on conversational memory, which often leads to context dilution and generalised or flattened responses, as they struggle to prioritise immediate inputs over preceding interactions [8]. These architectural and training differences explain the consistent performance of ChatGPT-4o and Gemma 3:27B under similar conditions.

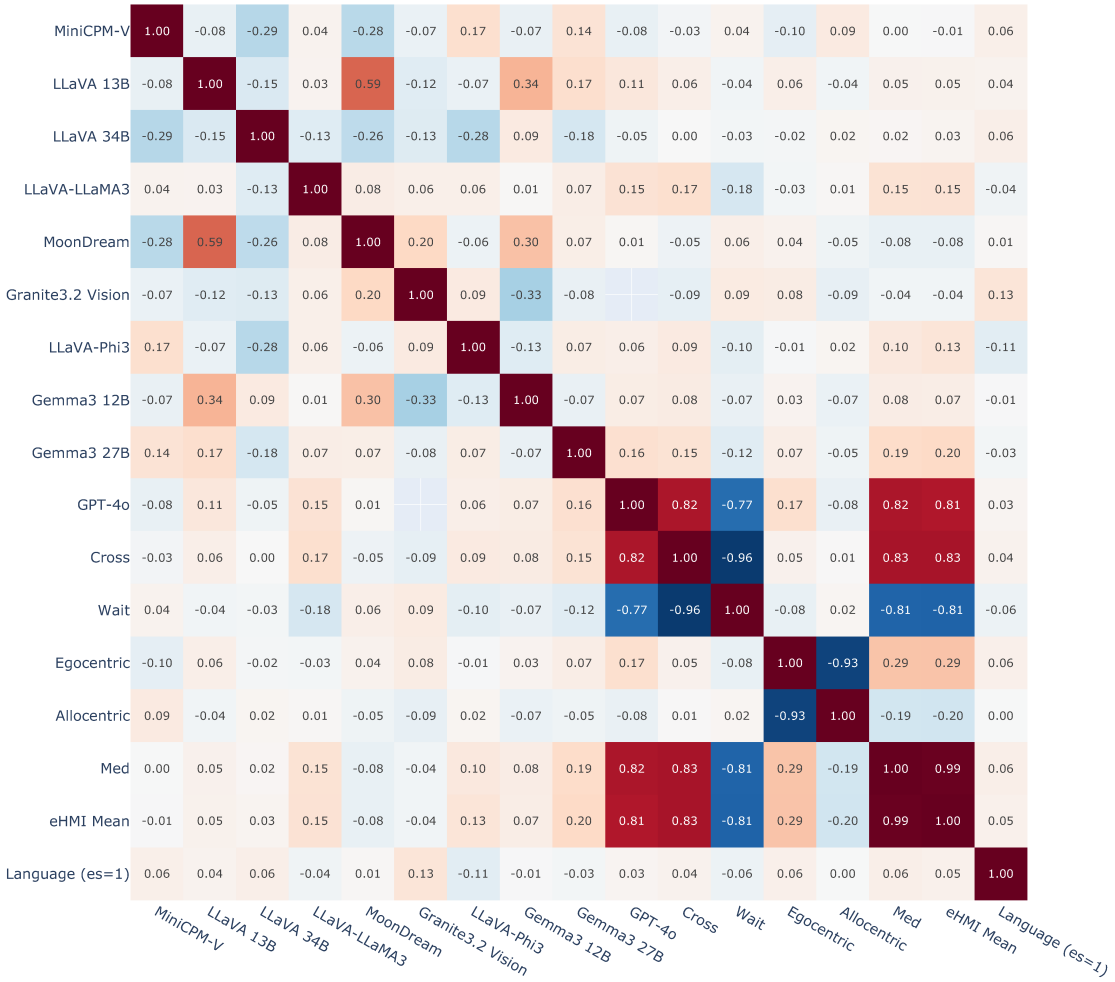

Fig. 5. Correlation matrix with memory

When conversational history is introduced, ChatGPT-4o maintains high performance, while Gemma 3:27B and several other models exhibit a noticeable decrease in predictive accuracy. This divergence appears to be the result of model-specific strategies for integrating the extended context. The architecture and training of GPT-4o are explicitly optimised for multi-turn dialogue, allowing it to incorporate previous exchanges without obscuring current visual inputs. In contrast, the performance degradation observed in Gemma 3:27B suggests limitations in its ability to balance

the historical context with new stimuli. Other models, such as those producing uniform responses under memory-based conditions, further illustrate how conversational history can overshadow relevant immediate input, diminishing performance.

Additional models, including MiniCPM-V and LLaMA3.2-vision, display increased variability in behaviour when context is introduced. MiniCPM-V experiences a marked decrease in performance, while LLaMA3.2 vision fails to generate outputs under memory-inclusive scenarios. These results suggest that in the absence of carefully calibrated memory mechanisms, some LLMs may underestimate or overestimate risks in pedestrian-vehicle interactions. Such inconsistencies underscore the importance of fine-tuning models to more accurately reflect human risk assessments in these contexts.

A further observation involves models such as LLaVA-13B, LLaVA-LLaMA3, and, occasionally, Moondream, which at times issue crossing commands even when the AV's LED or eHMI does not explicitly prompt pedestrian movement. This behaviour indicates a degree of proactivity or inference beyond the given cues, raising concerns about misaligned or anticipatory decision making in safety-critical scenarios.

5 Limitations and Future Work

This study has several limitations that must be acknowledged. First, the rapidly evolving landscape of Large Language Models (LLMs) poses a challenge, as newer models regularly emerge that may provide increasingly human-like responses. Consequently, the results presented here may quickly become outdated. In addition, a standardised prompt was uniformly applied across all models tested, potentially limiting the ability of individual LLMs to demonstrate their maximum performance. Customised prompts tailored to each model's strengths could provide more accurate assessments of their capabilities. Furthermore, the study observed significant variability in how the models handled contextual memory, with some performing poorly or completely failing when historical conversational context was included.

Future research should consider several promising avenues. First, subsequent studies might investigate the impact of customised prompts specifically designed to take advantage of each LLM's unique architecture and strengths. In addition, as new models continue to develop rapidly, systematic evaluations of emerging LLMs should be performed regularly to assess improvements and refinements in the simulation of human-like decision-making processes. Future work should also address the challenges associated with incorporating conversational history more effectively, possibly through specialised training or improved architectural designs. Furthermore, real-world validation experiments that compare LLM-derived decisions directly with actual pedestrian behaviours could enhance the ecological validity of these findings. Finally, broadening the scope of scenarios studied beyond pedestrian crossing decisions might provide deeper insights into the capabilities and limitations of LLM in various contexts of traffic and safety.

6 Supplementary Material

The code and responses of LLMs are available at https://www.dropbox.com/scl/fo/xs37ldfp72dspsrjykc2d/AEnstqB2KFDRjnnl8M0VJz8?rlkey=63vcekw3qr2c91ao38j1wtxow&e=1. The maintained code is at https://github.com/Shaadalam9/llms-av-crowdsourced.

References

[1] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis* 31, 3 (2023), 337–351.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural Machine Translation by Jointly Learning to Align and Translate. arXiv:1409.0473 [cs.CL] https://arxiv.org/abs/1409.0473

[3] Pavlo Bazilinskyy, Dimitra Dodou, and J. C. F. De Winter. 2022. Crowdsourced assessment of 227 text-based eHMIs for a crossing scenario. In *Proceedings of International Conference on Applied Human Factors and Ergonomics (AHFE)*. New York, USA. https://doi.org/10.54941/ahfe1002444

[4] Pavlo Bazilinskyy, Dimitra Dodou, Y. B. Eisma, W . V. Vlakveld, and J. C. F. De Winter. 2022. Blinded windows and empty driver seats: the effects of automated vehicle characteristics on cyclist decision-making. *IET Intelligent Transportation Systems* 17 (2022), 72–84. https://doi.org/10.1049/itr2.12235

[5] Ravindra Giriraj Bhardwaj and Harpreet Singh Bedi. 2025. ChatGPT as an education and learning tool for engineering, technology and general studies: performance analysis of ChatGPT 3.0 on CSE, GATE and JEE examinations of India. *Interactive Learning Environments* 33, 1 (2025), 321–334.

[6] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv:2303.12712 [cs.CL] https://arxiv.org/abs/2303.12712

[7] Alex Burnap, Yi Ren, Richard Gerth, Giannis Papazoglou, Richard Gonzalez, and Panos Y Papalambros. 2015. When crowdsourcing fails: A study of expertise on crowdsourced design evaluation. *Journal of Mechanical Design* 137, 3 (2015), 031101.

[8] Aolin Chen, Haojun Wu, Qi Xin, Steven P Reiss, and Jifeng Xuan. 2025. Studying and Understanding the Effectiveness and Failures of Conversational LLM-Based Repair. *arXiv preprint arXiv:2503.15050* (2025).

[9] XTuner Contributors. 2023. XTuner: A Toolkit for Efficiently Fine-tuning LLM. https://github.com/InternLM/xtuner.

[10] XTuner Contributors. 2024. llava-phi-3-mini-gguf: A LLaVA Model Fine-Tuned from Phi-3-Mini-4k-Instruct and CLIP-ViT-Large-patch14-336. https://huggingface.co/xtuner/llava-phi-3-mini-gguf. Accessed: 2025-04-07.

[11] Joost De Winter, Jim Hoogmoed, Jork Stapel, Dimitra Dodou, and Pavlo Bazilinskyy. 2023. Predicting perceived risk of traffic scenes using computer vision. *Transportation Research Part F: Traffic Psychology and Behaviour* 93 (2023), 235–247.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL] https://arxiv.org/abs/1810.04805

[13] Tom Driessen, Dimitra Dodou, Pavlo Bazilinskyy, and J. C. F. De Winter. 2024. Putting ChatGPT Vision (GPT-4V) to the test: Risk perception in traffic images. *Royal Society Open Science* 11 (2024), 231676. https://doi.org/10.4121/dfbe6de4-d559-49cd-a7c6-9bebe5d43d50

[14] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

[15] Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024. AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators. arXiv:2303.16854 [cs.CL] https://arxiv.org/abs/2303.16854

[16] Cameron R. Jones and Benjamin K. Bergen. 2025. Large Language Models Pass the Turing Test. https://doi.org/10.48550/arXiv.2503.23674 arXiv:2503.23674 [cs.CL]

[17] Yunshui Li, Binyuan Hui, Xiaobo Xia, Jiaxi Yang, Min Yang, Lei Zhang, Shuzheng Si, Ling-Hao Chen, Junhao Liu, Tongliang Liu, et al. 2023. One-shot learning as instruction data prospector for large language models. *arXiv preprint arXiv:2312.10302* (2023).

[18] Ying Li, Ye Zhong, Lijuan Yang, Yanbo Wang, and Penghua Zhu. 2025. LLM-Guided Crowdsourced Test Report Clustering. *IEEE Access* 13 (2025), 24894–24904. https://doi.org/10.1109/ACCESS.2025.3530960

[19] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems* 36 (2023), 34892–34916.

[20] OpenAI. 2023. GPT-4 with Vision. https://openai.com/research/gpt-4. Accessed: 2025-04-06.

[21] Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? How controllable attributes affect human judgments. arXiv:1902.08654 [cs.CL] https://arxiv.org/abs/1902.08654

[22] Sina Shool, Sara Adimi, Reza Saboori Amleshi, Ehsan Bitaraf, Reza Golpira, and Mahmood Tara. 2025. A systematic review of large language model (LLM) evaluations in clinical medicine. *BMC Medical Informatics and Decision Making* 25, 1 (2025), 117.

[23] SkunkworksAI. 2023. BakLLaVA-1: Mistral 7B Base Augmented with LLaVA 1.5 Architecture. https://huggingface.co/SkunkworksAI/BakLLaVA-1. Accessed: 2025-04-07.

[24] Felix C Stengel, Martin N Stienen, Marcel Ivanov, María L Gandía-González, Giovanni Raffa, Mario Ganau, Peter Whitfield, and Stefan Motov. 2024. Can AI pass the written European Board Examination in Neurological Surgery?-Ethical and practical issues. *Brain and Spine* 4 (2024), 102765.

[25] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2* (Montreal, Canada) *(NIPS'14)*. MIT Press, Cambridge, MA, USA, 3104–3112.

[26] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).

[27] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786* (2025).

[28] Granite Vision Team, Leonid Karlinsky, Assaf Arbelle, Abraham Daniels, Ahmed Nassar, Amit Alfassi, Bo Wu, Eli Schwartz, Dhiraj Joshi, Jovana Kondic, et al. 2025. Granite Vision: a lightweight, open-source multimodal model for enterprise Intelligence. *arXiv preprint arXiv:2502.09927* (2025).

[29] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[31] Vikhyat. 2025. Moondream 2: A Tiny Vision Language Model. https://github.com/vikhyat/moondream.

[32] Xuezhi Wang, Haohan Wang, and Diyi Yang. 2021. Measure and improve robustness in NLP models: A survey. *arXiv preprint arXiv:2112.08313* (2021).

[33] Tongshuang Wu, Haiyi Zhu, Maya Albayrak, Alexis Axon, Amanda Bertsch, Wenxing Deng, Ziqi Ding, Bill Guo, Sireesh Gururaja, Tzu-Sheng Kuo, et al. 2023. Llms as workers in human-computational algorithms? replicating crowdsourcing pipelines with llms. *arXiv preprint arXiv:2307.10168* (2023).

[34] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302* (2024).

[35] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. *arXiv preprint arXiv:2408.01800* (2024).

[36] Anni Zou, Wenhao Yu, Hongming Zhang, Kaixin Ma, Deng Cai, Zhuosheng Zhang, Hai Zhao, and Dong Yu. 2024. Docbench: A benchmark for evaluating llm-based document reading systems. *arXiv preprint arXiv:2407.10701* (2024).

Received 20 February 2025; revised 12 March 2025; accepted 5 June 2025