# 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

This study examines the effectiveness of using large language model-based personas to evaluate external Human-Machine Interfaces (eHMIs) in automated vehicles. 13 different models namely BakLLaVA, ChatGPT-4o, DeepSeek-VL2, Gemma 3: 12B, Gemma 3: 27B, Granite Vision 3.2, LLaMA 3.2 Vision, LLaVA-13B, LLAVA-34B, LLaVA-LLaMA-3, LLaVA-Phi3, MiniCPM-V, and Moondream were used to simulate pedestrian perspectives. Models assessed vehicle images with eHMI, assigning scores from 0 (completely unwilling) to 100 (fully confident) regarding crossing decisions. Each model was run 15 times across the full set of images, both with and without prior conversational context. The resulting confidence scores were then 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, DeepSeek-VL2 and BakLLaVA gave similar scores regardless of context, while LLaVA-LLaMA-3, LLaVA-Phi3, LLaVA-13B, and Moondream produced only limited-range outputs in both cases.

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

Additional Key Words and Phrases: Vision 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. 1, 1 (June 2025), 23 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 Introduction

In 2017, Google researchers proposed a work titled "`Attention is all you need`" [25], introducing the attention mechanism [2] to significantly enhance sequence-to-sequence (seq2seq) models [20]. This innovation paved the way for encoder-only architectures such as the Bidirectional Encoder Representations from Transformers (BERT) [9] 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 [17] and document processing [31], as well as general purpose models such as DeepSeek-VL2 [29] and ChatGPT [1]. 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 imitate human-like responses in rigorous conversational evaluations inspired by the Turing test. For example,

*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

ChatGPT-4.5 and LLaMA-3.1-405B have been judged as human in 73% and 56% of cases, respectively, under specific experimental protocols [12]. However, it is important to note that the interpretations of the turing test vary and there is no universally accepted version or passing criterion. Further highlighting their ability, Stengel et al. (2024) reported that Google's Bard LLM [21] 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 [19]. 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).

Automated vehicles (AV) may be equipped with external Human-Machine Interfaces (eHMIs), which are displays designed to communicate vehicle intentions to other road users, such as pedestrians [3], cyclists [27], and manually driven vehicle drivers [15]. The development of effective eHMIs is critical to ensuring traffic safety and fostering public trust in AV technology. However, evaluating the effectiveness of eHMI designs remains a significant challenge. Researchers such as Bazilinskyy et al. [4], Cumbal et al. [8], etc have relied on large-scale crowdsourced experiments to assess eHMI concepts based on human judgment. Although these methods can provide valuable information, they are often resource intensive, time consuming and susceptible to inconsistencies caused by participant variability, lack of diversity, inattention, or lack of domain expertise [5].

Recent studies indicate that LLMs and vision language models (VLMs) can closely approximate human opinions on subjective tasks, and in some annotation contexts, even outperform crowdsourced human raters in reliability and consistency [28]. If these models can accurately simulate human decision making in AV-pedestrian scenarios, they offer the potential for rapid and cost-effective prescreening of eHMI designs, reducing the dependency on extensive human data collection [11], particularly valuable in the early stages of interface development.

### 1.1 Aim of Study

The aim of this study is to evaluate the capability of VLMs to simulate pedestrian crossing decisions in response to eHMI messages displayed on an AV. The investigation encompasses a comparative analysis of 13 different LLM architectures, evaluating their interpretative precision with and without access to conversational history. The study further examines the alignment between model-generated confidence scores and crowdsourced human ratings, employing statistical correlation as a benchmark. Through this approach, the study aims to determine the effectiveness of LLM-based personas as reliable and scalable tools for the prescreening and assessment of eHMI designs in the AV context.

## 2 Method

This study used a crowdsourced dataset compiled by Bazilinskyy et al. (2022), involving 1,438 participants who evaluated 227 distinct textual eHMIs displayed on an AV indicating their willingness to cross via a slider scale ranging from 0 (absolute unwillingness) to 100 (complete confidence) [4]. The dataset, which served as a human benchmark for evaluating the interpretability of eHMI messages using contemporary VLM, comprised 227 standardised JPEG images (1024 × 598 pixels), each of which displayed a different textual eHMI message on an AV.

A total of 13 VLMs were evaluated in this study. The model files were obtained from Ollama (https://ollama.com) and Hugging Face (https://huggingface.co/models), while inference for ChatGPT-4o was performed through the OpenAI ChatGPT API (https://platform.openai.com/docs/overview). Table 1 summarises the evaluated models, detailing their base architectures and acquisition or deployment method. The use of the ChatGPT API incurred a cost of 20, while all other VLMs were accessed free of charge.

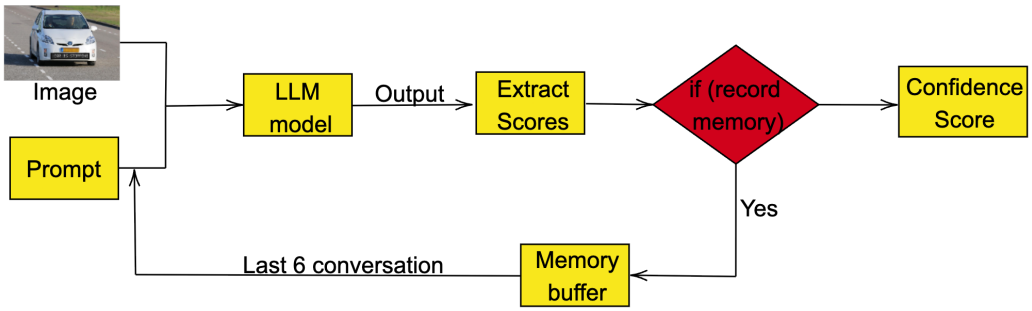


Fig. 1. Flow diagram of the system architecture showing image processing, prompting, model querying, response correction, and downstream analysis.

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

| **Model Name** | **Architecture** | **Platform** |
|---|---|---|
| BakLLaVA [18] | LLaMA 2 + BLIP | Ollama |
| DeepSeek-VL2-Tiny [29] | DeepSeek VLM | Hugging Face |
| Gemma 3: 12B [22] | Gemma 3 vision-capable model | Ollama |
| Gemma 3: 27B [22] | Gemma 3 vision-capable model | Ollama |
| Granite Vision 3.2 [23] | IBM Granite vision model | Ollama |
| ChatGPT-4o [14] | Multimodal transformer | OpenAI (ChatGPT API) |
| LLaVA:13B [13] | LLaMA + CLIP | Ollama |
| LLaVA:34B [13] | LLaMA + CLIP | Ollama |
| LLaVA-LLaMA-3 [6] | LLaMA-3 + Visual Adapter | Ollama |
| LLaVA-Phi3 [7] | Phi-3 + Vision module | Ollama |
| LLaMA3.2-vision [24] | LLaMA 3.2 + Vision | Ollama |
| MiniCPM-V [30] | CPM-based | Ollama |
| Moondream [26] | Distilled CLIP variant | Ollama |

For each VLM, the 227 images were presented one at a time in a consistent predetermined sequence. The evaluation was performed under two separate conditions: one in which the models operated without conversational history (memory) and another where the conversational context was maintained across images. In the no-memory condition, each model received a standardised prompt instructing him to interpret the vehicle's displayed message, infer its meaning for pedestrian safety, and generate a confidence score from 0 to 100. The same prompt was also used as the initial instruction in the conversational history condition.

> *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 as used for trials without memory. From the second image and onwards, contextual memory was incorporated using LangChain (https://www.langchain.com), preserving the previous 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 [16].

(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 1). To ensure consistency and prevent interference between trials, all model interactions were conducted sequentially. The environment was reset after each trial to eliminate any residual state or memory effects. With the exception of ChatGPT (ChatGPT-4o), which was accessed exclusively through the OpenAI API and was processed remotely on OpenAI's servers, all other models, including DeepSeek-VL2, were deployed and executed locally. Local deployments were managed through Ollama (supporting batch mode image processing), while DeepSeek-VL2 was run using code and weights obtained directly from Hugging Face (https://huggingface.co/deepseek-ai/deepseek-vl2-tiny). API-based models such as ChatGPT-4o were accessed via synchronous HTTP POST requests with JSON payloads, ensuring that all data processing occurred server-side for these models.

After evaluation, the responses were analysed to extract numerical confidence scores, using the local deepseek-r1:14b model (https://ollama.com/library/deepseek-r1:14b). This step was necessary to address the variability in verbosity between different model outputs. The prompt used to extract the confidence score was as follows:

> *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>"*

A regular expression extractor was then used to extract a numeric value from the model's output. If no valid number could be extracted, the value was recorded as NaN to ensure consistency in the dataset. The responses were stored for subsequent statistical analysis, allowing comparison with human benchmarks and assessment of the interpretative precision of each model with respect to pedestrian crossing decisions based on eHMI messages.

The model outputs were collected and subsequently filtered: Only responses with values between 0 and 100 (inclusive) were retained for further analysis. These filtered values were then used to calculate the average response, while any values that fell outside this range were excluded from consideration.

## 3 Results

The average responses for each model are summarised in the appendix. For each image, the confidence scores generated by each model were averaged and compared to crowdsourced results reported in [4]. Models without memory, such

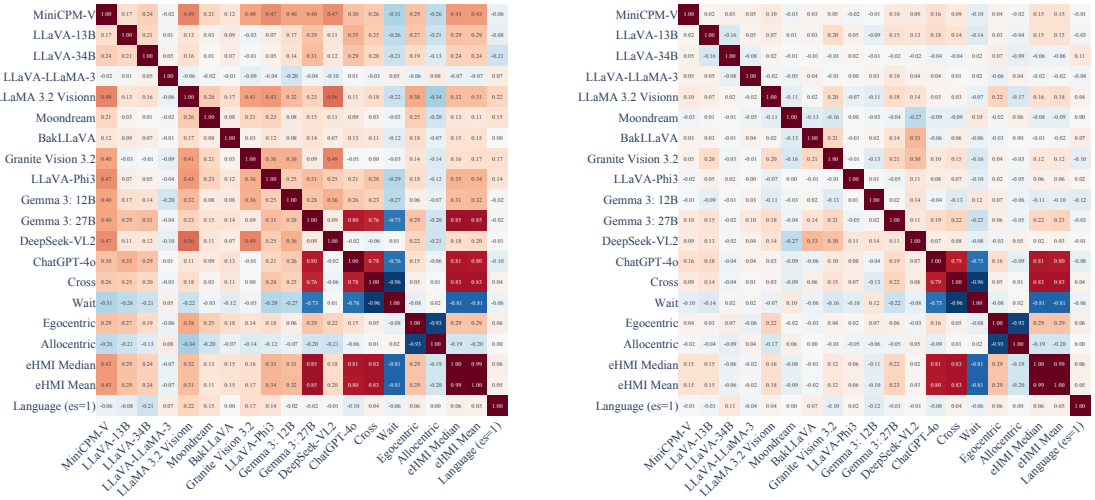


Fig. 2. Spearman correlation matrices of model outputs, behavioral features, and language encoding. The **left** image shows results *without memory*, and the **right** image shows results *with memory*. Binary behavioral features ("Cross", "Wait", "Egocentric", and "Allocentric") are encoded as 0 (absent) and 1 (present). The language variable is encoded as 0 for English ("en") and 1 for Spanish ("es"). All values represent pairwise Spearman correlation coefficients; positive correlations are shown in warmer colors, negative in cooler colors.

as BakLLaVA, predominantly returned a value of 0 for 97 cases. LLaVA-LLaMA3, on the other hand, produced a maximum confidence score of 36.66 for the prompt "PROCEED TO CROSS NOW," whereas the corresponding average crowdsourced response was 77.03. The Moondream model yielded average responses ranging between 18.60 and 56.15, while the LLaVA-13B outputs were between 31.00 and 77.23. The DeepSeek-VL2 model produced some outputs within the 0 to 100 range; however, most of its responses were restricted to 0, 75, or 90. Notably, for the prompt "YOU CAN WALK," DeepSeek-VL2 returned a value of 100, but it also assigned a high confidence score of 90 for the prompt "CAR WILL NOT STOP."

In contrast, when memory was enabled, certain models exhibited broader distributions in their output. For example, BakLLaVA with memory generated values predominantly between 0 and 13.33, slightly broader than its output without memory. DeepSeek-VL2, when using conversation history, shifted to providing consistently higher values, with a minimum response of 92. Moondream, in the memory-enabled setting, reported values in the range of 0 to 21.17.

Among the models evaluated without conversation history, Gemma3 27B exhibited the strongest alignment with human responses, yielding high correlation coefficients with the mean (r = 0.84) and median (r = 0.85) of crowdsourced data. ChatGPT-4o also showed strong correlations, with coefficients of r = 0.80 (mean) and r = 0.81 (median). In contrast, MiniCPM-V showed a substantially lower correlation (r = 0.43) with the mean and median values, indicating a weaker agreement with human judgments. The remaining ten models did not show a significant correlation with human responses; for example, Gemma3 12B, LLaVA-Phi3, and LLaMA3.2 Vision reported correlation coefficients of 0.32, 0.34, and 0.31, respectively. Models such as LLaVA-LLaMA, Moondream, and Granite Vision 3.2 showed the weakest correlations, with coefficients of -0.07, 0.11, and 0.17, respectively.

When conversation history was incorporated, ChatGPT-4o maintained the highest correlation with crowdsourced responses, with correlation coefficients for mean and median responses remaining stable at r = 0.80 and r = 0.81, respectively. In contrast, Gemma3 27B exhibited a substantial decrease in performance, with the mean correlation dropping from r = 0.84 to r = 0.23, and the median from r = 0.85 to r = 0.22. MiniCPM-V also showed a marked reduction, with mean and median correlations falling from r = 0.43 to r = 0.15. Similar trends were observed in the other models: LLaMA 3.2 Vision's mean correlation decreased from r = 0.31 to r = 0.18, and its median from r = 0.32 to r = 0.16. LLaVA 34B demonstrated a decrease in mean and median correlations from r = 0.24 to r = -0.06. LLaVA-Phi3 decreased in the mean from r = 0.34 to r = 0.06 and in the median from r = 0.35 to r = 0.06. Across all models, the inclusion of conversation history consistently reduced alignment with the mean and median human responses.

## 4 Discussion

The comparative analysis of 13 multimodal LLMs in this study reveals marked differences in how architectural design and training paradigms govern one-shot risk assessment. In memory-free scenarios, models with high parameter counts and specialised fusion layers (e.g. ChatGPT-4o, Gemma 3: 27B) produce human-aligned confidence distributions ($r = 0.80˜0.85$), reflecting robust cross-attention that tightly integrates image and text features. In contrast, lighter or less tuned architectures (BakLLaVA, LLaVA-LLaMA3) default to extreme low or capped scores (97 zero-confidence outputs; max 36.66 vs human mean 77.03), indicating insufficient nuance in feature extraction. Models like DeepSeek-VL2 and Moondream show quantised outputs, limited to discrete values due to categorical confidence buckets rather than finely ranked estimates. The weak correlation of MiniCPM-V ($r < 0.43$) further underscores that without extensive multimodal pre-training, the interpretation of signals remains coarse and misaligned with human benchmarks.

With conversational history, the memory management strategy becomes central, explaining the widespread performance degradation. ChatGPT-4o retains strong alignment ($r = 0.80˜0.81$) by leveraging hierarchical attention and explicit context window control to preserve the importance of new inputs. Others over-attenuate or overweight historical dialogue: Gemma 3:27B's correlation drops to $r = 0.22$ as a simpler recurrence allows prior exchanges to dominate, while MiniCPM-V and LLaVA variants lose calibration due to static context encoding. Distribution shifts—BakLLaVA expanding only to 13.33, DeepSeek-VL2 inflating to 92, Moondream contracting between 0 and 21.17—highlight how uncalibrated memory modules compress risk sensitivity or inflate unwarranted confidence, risking unsafe crossing suggestions, and emphasising the need for finely tuned memory gating in safety-critical AV eHMI interpretation.

These findings have significant implications for the development and application of VLMs in UX evaluation. VLMs dramatically reduce time and cost compared to large-scale human trials, offering reproducible scalable methods for validation. Driessen et al. (2024) showed that GPT 4V can achieve a population-level correlation of risk assessment $r = 0.83$ with humans in traffic images, provided that the prompts are varied and supplemented with object detection [10], underscoring how visual prompting can approximate collective human perception. However, while VLMs theoretically reach a global participant pool, in practice, even crowdsourcing struggles for true diversity due to platform restrictions or low regional adoption. Caution is warranted: Top-tier models may mirror average ratings but struggle with nuanced or culturally dependent perceptions, so VLMs are best as adjuncts, not replacements, in UX validation.

Addressing our core question, *Can VLMs replace human participants, especially in crowd-sourced UX?*, the evidence is nuanced. For straightforward and visually clear cases, models like ChatGPT-4o match aggregate judgments and are justified for pre-selection and benchmarking. As complexity grows, especially with a conversational or culturally rich context, alignment drops. These limits echo traffic risk studies, where even ChatGPT-4o required enhanced object detection for high validity.

## 5 Limitations and Future Work

This study has several limitations that must be acknowledged. First, the rapidly evolving landscape of VLMs 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.

A further limitation relates to the grounding of visual-textual information in context. For example, although models such as LLaVA-34B can accurately read and extract text from images, they may fail to properly ground this information in relation to the pedestrian's intention or immediate decision-making. Specifically, when presented with a vehicle displaying the message "I WILL NOT STOP" the model responded: *The autonomous vehicle displays a message on the windshield stating, "I WILL NOT STOP" which implies that its system has control and will likely follow traffic rules including stopping for pedestrians. However, caution should still be exercised when crossing roads even in this situation.* Similarly, when the vehicle displayed the message "OK TO PASS", the model responded: *The text displayed on the vehicle reads OK TO PASS, which implies that it is safe for pedestrians to cross the street in front of this vehicle at this moment. However, since there could be other factors that influence pedestrian safety such as traffic conditions, weather, or visibility issues, a cautious approach would still be advisable when deciding whether to wait for another opportunity to cross the road safely considering potential changes in vehicle status or other unforeseen circumstances.* These examples demonstrate that, although the model can extract and read textual information, it does not always adequately ground that information in the specific moment or align with the real-time intentions of a pedestrian who wishes to cross. Instead, the model's reasoning can be generic or disconnected from the immediate decision context. This limitation could potentially be addressed by refining prompts or by submitting multiple images that provide clearer cues about the pedestrian's intent.

Future research should consider several promising avenues. First, subsequent studies might investigate the impact of custom instructions 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.

## Supplementary Material

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

## References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[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. 2019. Survey on eHMI concepts: The effect of text, color, and perspective. *Transportation Research Part F: Traffic Psychology and Behaviour* 67 (2019), 175–194. https://doi.org/10.1016/j.trf.2019.10.013

[4] 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

[5] 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. https://doi.org/10.1115/1.4029065

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

[7] 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.

[8] Ronald Cumbal, Didem Gurdur Broo, and Ginevra Castellano. 2025. Crowdsourcing eHMI Designs: A Participatory Approach to Autonomous Vehicle-Pedestrian Communication. *arXiv preprint arXiv:2506.18605* (2025).

[9] 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

[10] 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

[11] 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

[12] 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]

[13] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems* 36 (2023), 34892–34916. https://doi.org/10.48550/arXiv.2304.08485

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

[15] Michael Rettenmaier, Deike Albers, and Klaus Bengler. 2020. After you?!–Use of external human-machine interfaces in road bottleneck scenarios. *Transportation research part F: traffic psychology and behaviour* 70 (2020), 175–190. https://doi.org/10.1016/j.trf.2020.03.004

[16] 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

[17] 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.

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

[19] 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. https://doi.org/10.1016/j.bas.2024.102765

[20] 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. https://doi.org/doi/10.5555/2969033.2969173

[21] 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).

[22] 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).

[23] 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).

[24] 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).

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6000–6010. https://doi.org/doi/10.5555/3295222.3295349

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

[27] Willem Vlakveld, Sander van der Kint, and Marjan P Hagenzieker. 2020. Cyclists' intentions to yield for automated cars at intersections when they have right of way: Results of an experiment using high-quality video animations. *Transportation research part F: traffic psychology and behaviour* 71 (2020), 288–307. https://doi.org/10.1016/j.trf.2020.04.012

[28] 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).

[29] 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).

[30] 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).

[31] 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).

## Appendix

### A Scatter Plots Without Memory: Model Predictions vs Human Responses

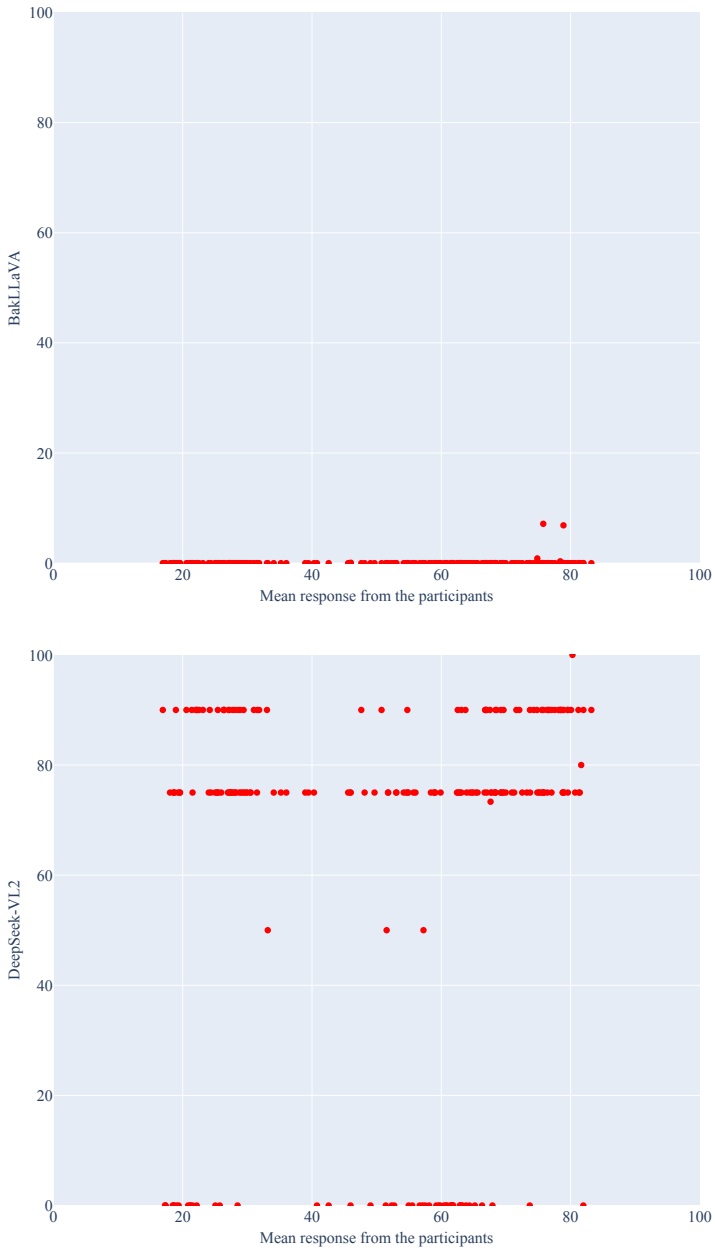


Fig. 3. Scatter plots comparing model predictions to human participant responses, **without memory**. The **top** image shows results for the *BakLLaVA* model; the **bottom** image shows results for the *DeepSeek-VL2* model.

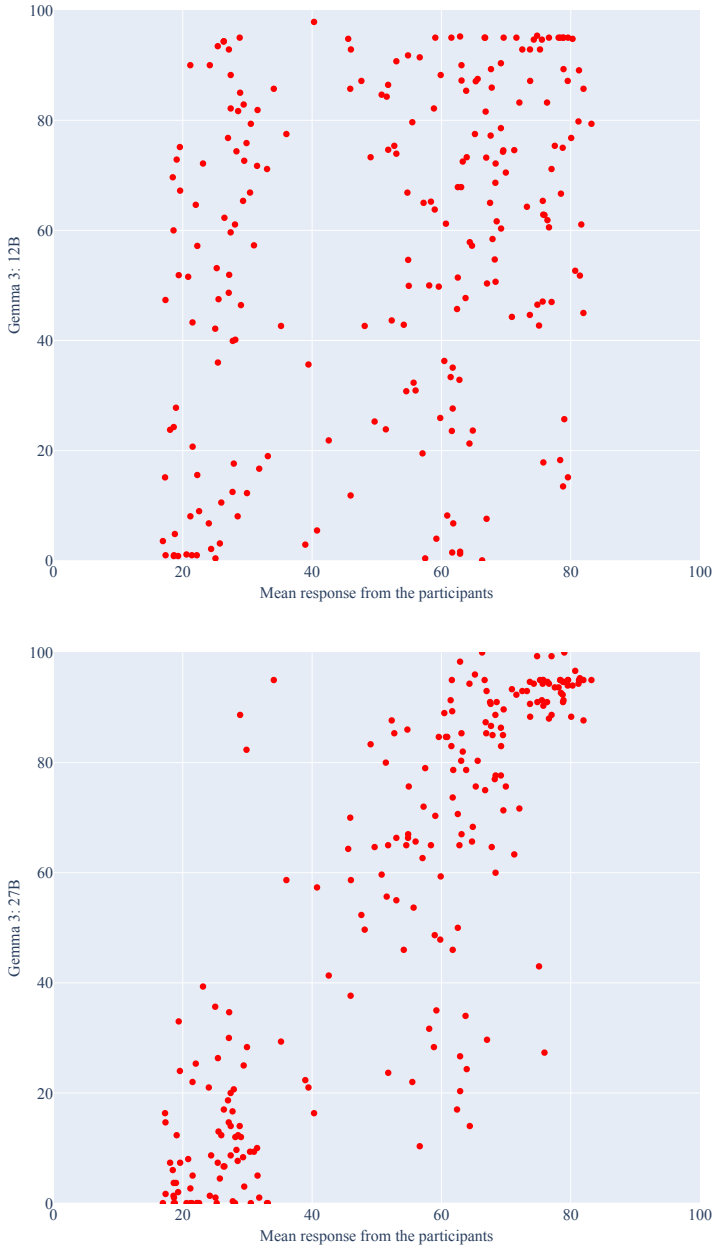


Fig. 4. Scatter plots comparing model predictions to human participant responses, **without memory**. The **top** image shows results for the *Gemma3:12b* model; the **bottom** image for the *Gemma3:27b* model.

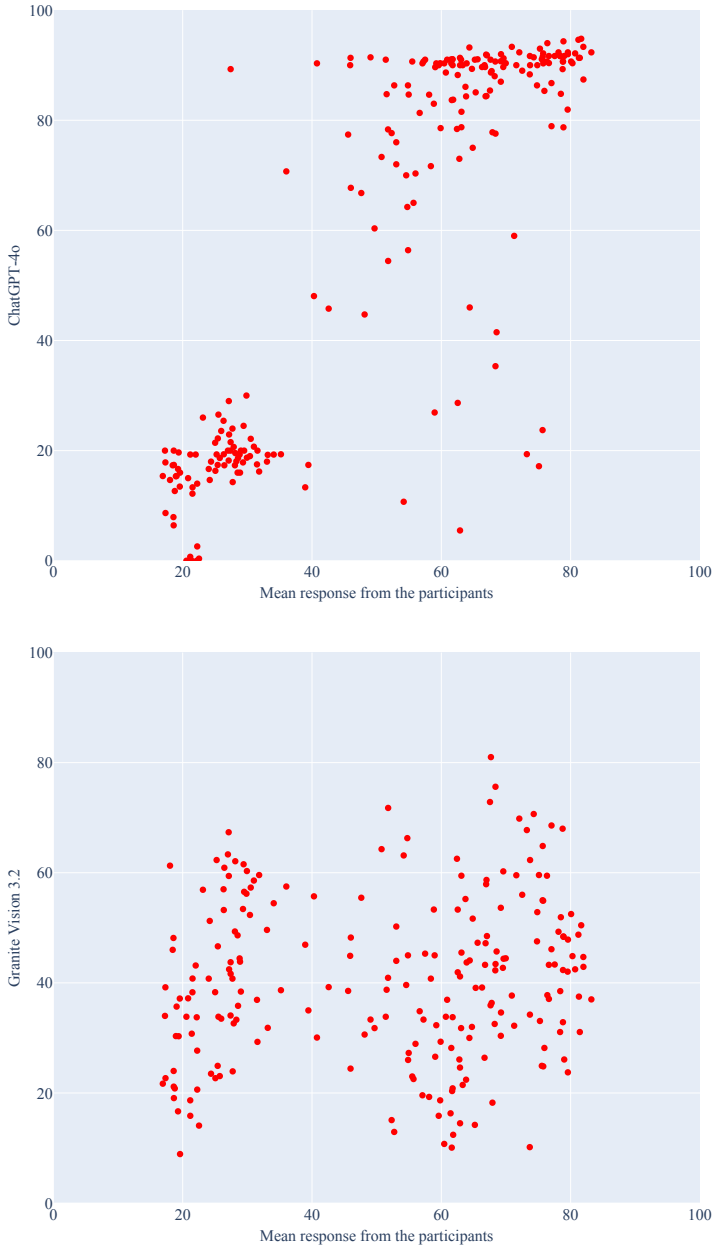


Fig. 5. Scatter plots comparing model predictions to human participant responses, **without memory**. The **top** image shows results for the *ChatGPT-4o* model; the **bottom** image for the *Granite Vision 3.2* model.

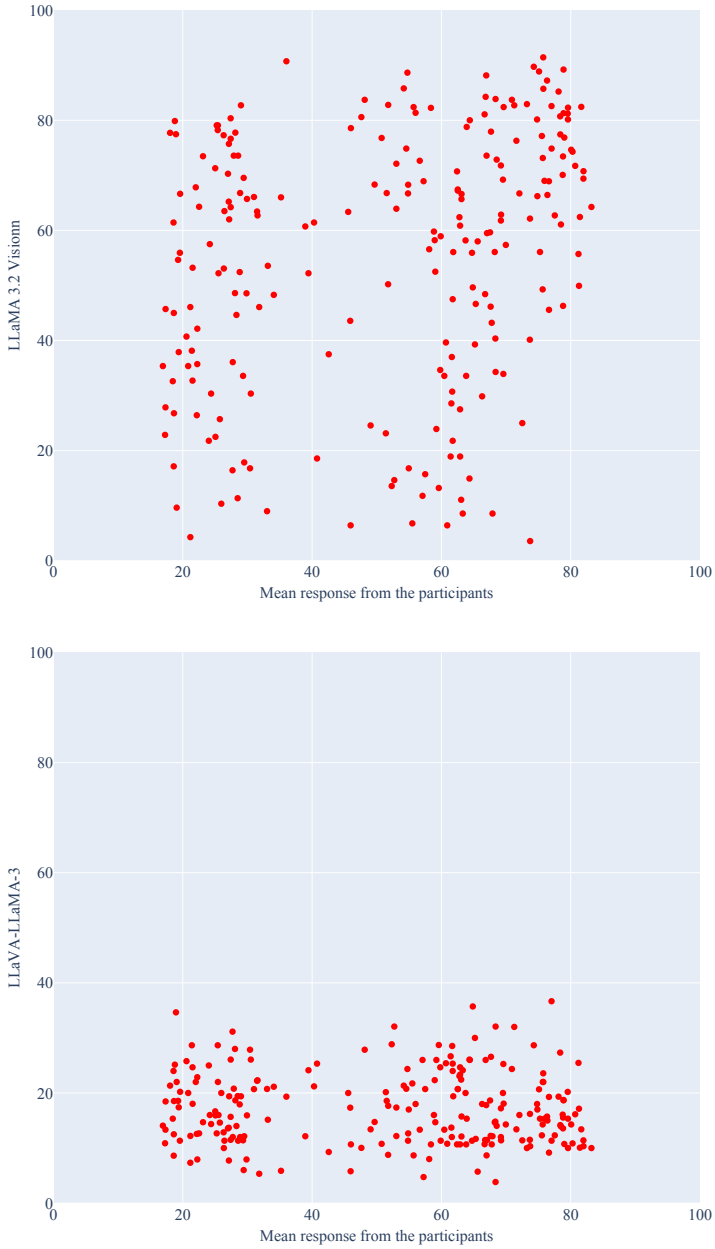


Fig. 6. Scatter plots comparing model predictions to human participant responses, **without memory**. The **top** image shows results for the *LLaMA 3.2 Vision* model; the **bottom** image for the *LLaVA LLaMA3* model.

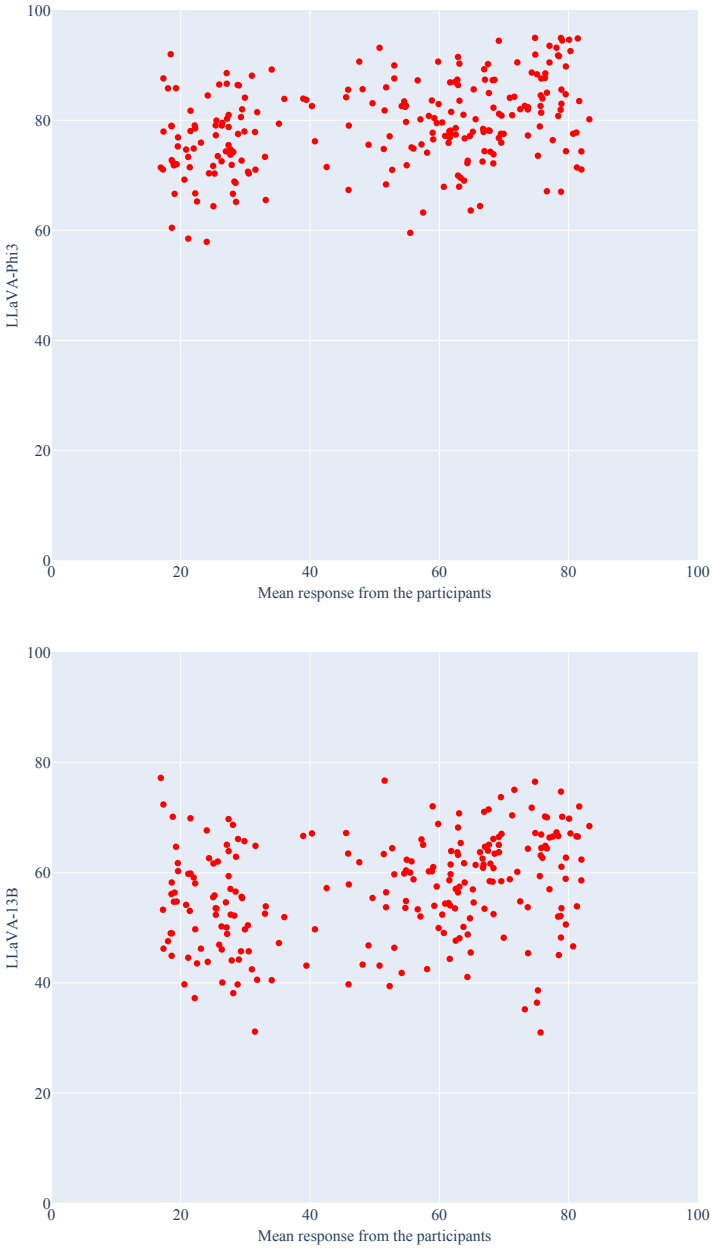


Fig. 7. Scatter plots comparing model predictions to human participant responses, **without memory**. The **top** image shows results for the *LLaVA Phi 3* model; the **bottom** image for the *LLaVA:13b* model.

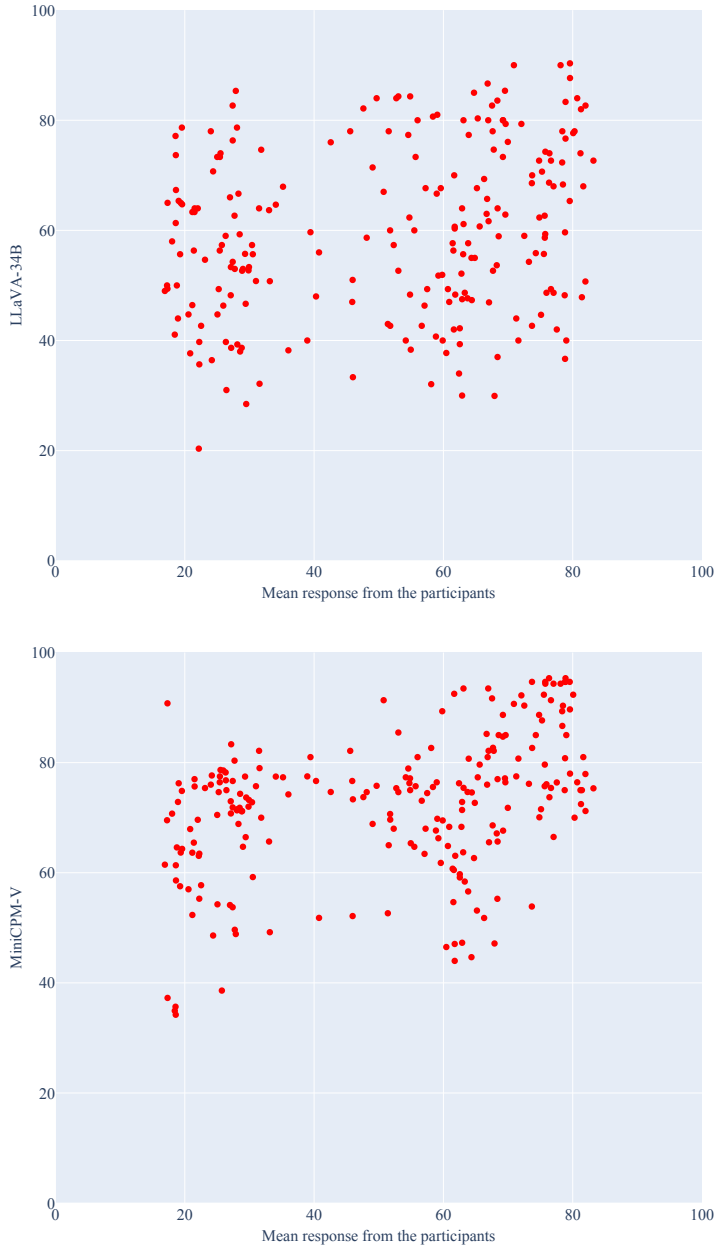


Fig. 8. Scatter plots comparing model predictions to human participant responses, **without memory**. The **top** image shows results for the *LLaVA:34b* model; the **bottom** image for the *Mini CPM-V* model.

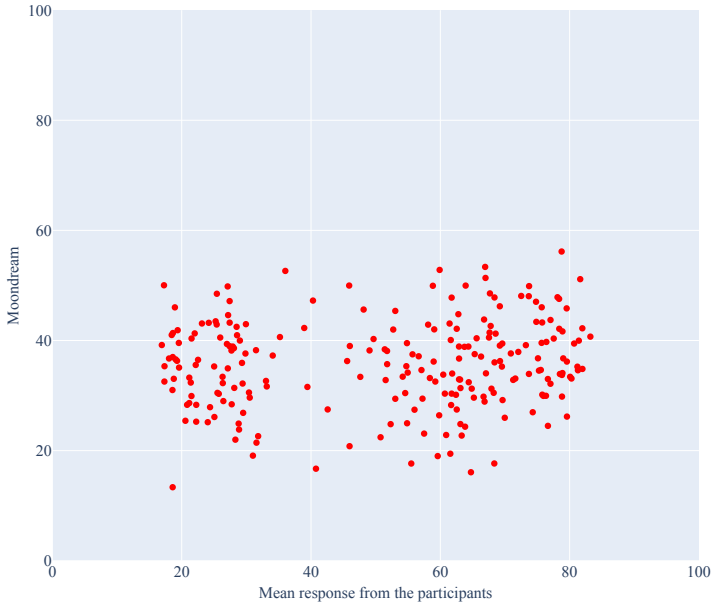


Fig. 9. Scatter plot comparing model predictions to human participant responses, **without memory**. The image shows results for the *Moondream* model.

## B Scatter Plots With Memory: Model Predictions vs Human Responses

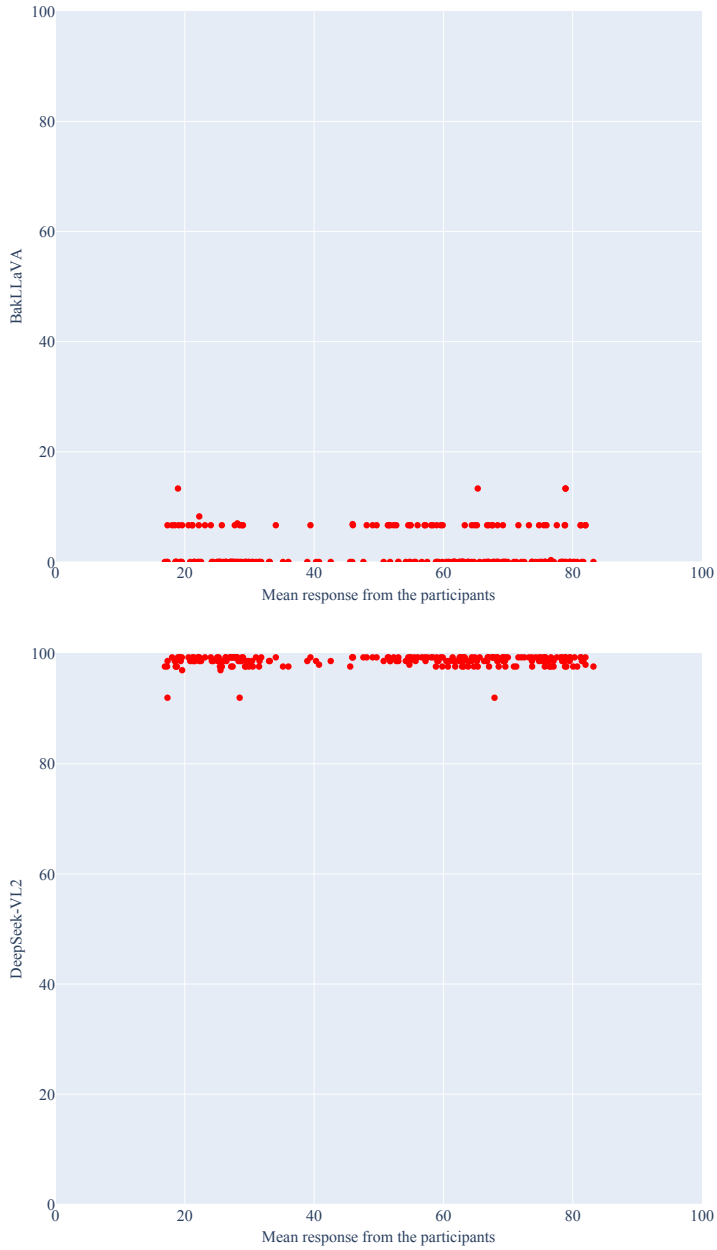


Fig. 10. Scatter plots comparing model predictions to human participant responses, **with memory**. The **top** image shows results for the *BakLLaVA* model; the **bottom** image for the *DeepSeek-VL2* model.

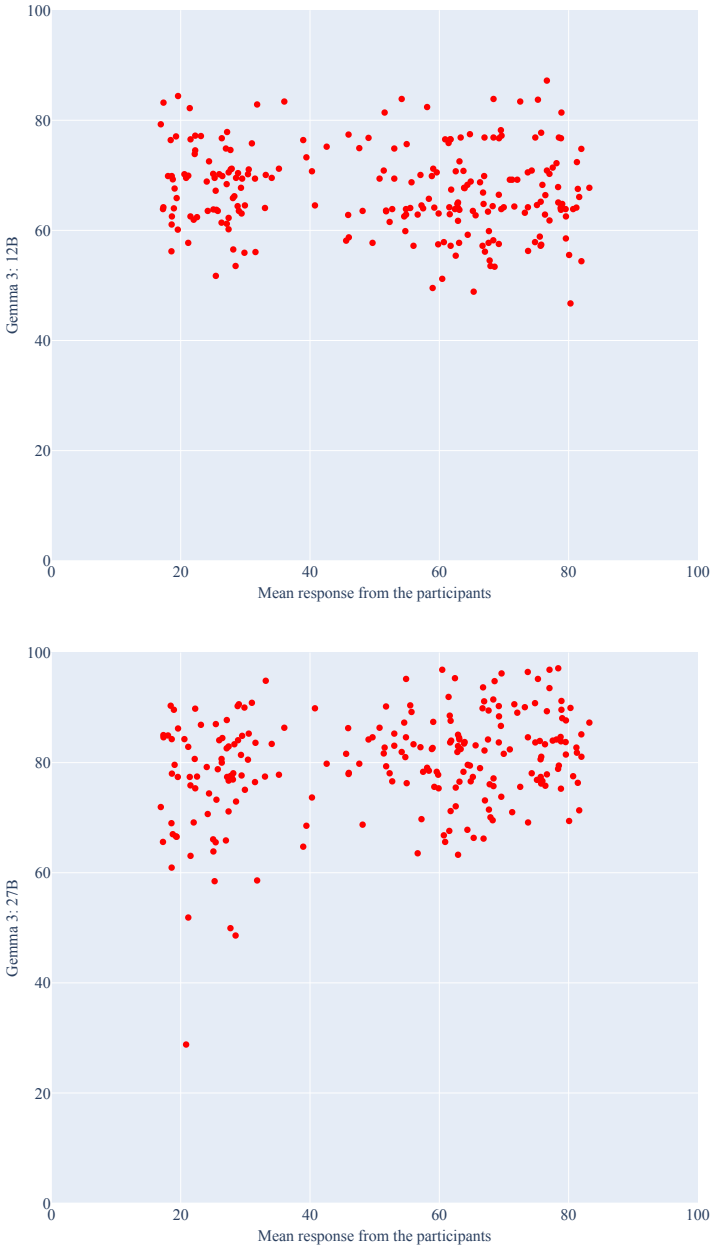


Fig. 11. Scatter plots comparing model predictions to human participant responses, **with memory**. The **top** image shows results for the *Gemma3:12b* model; the **bottom** image for the *Gemma3:27b* model.

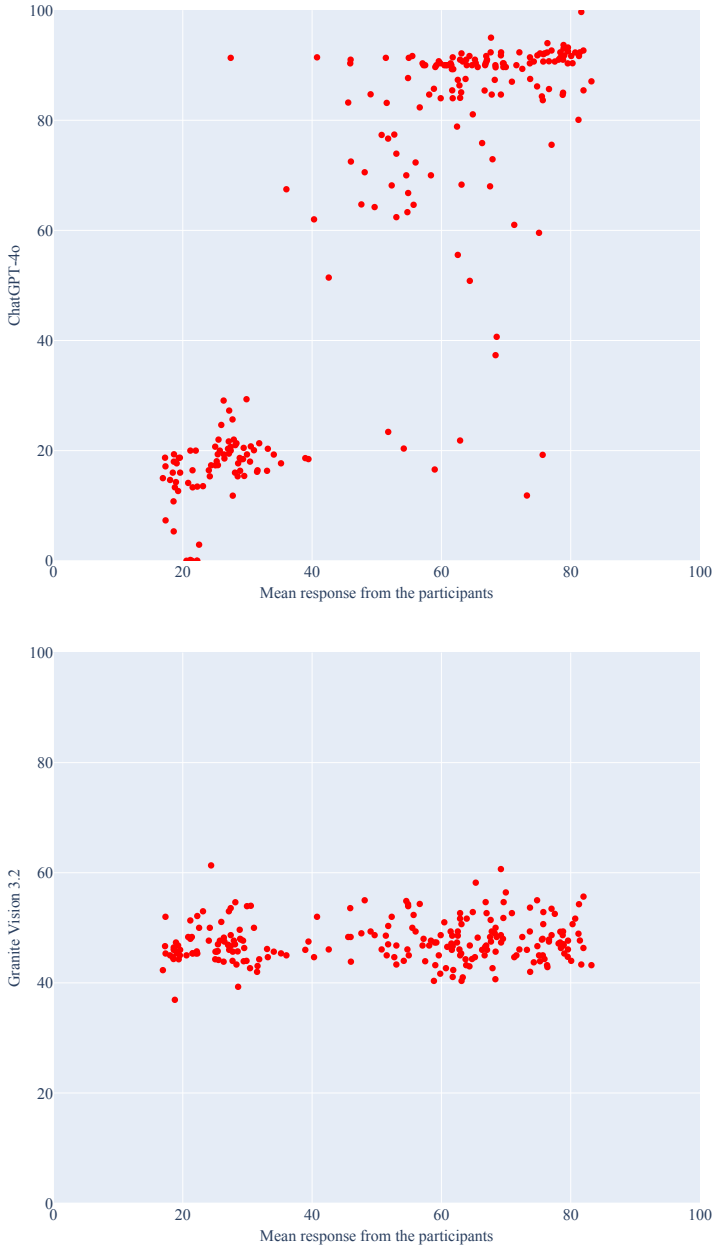


Fig. 12. Scatter plots comparing model predictions to human participant responses, **with memory**. The **top** image shows results for the *ChatGPT-4o* model; the **bottom** image for the *Granite Vision 3.2* model.

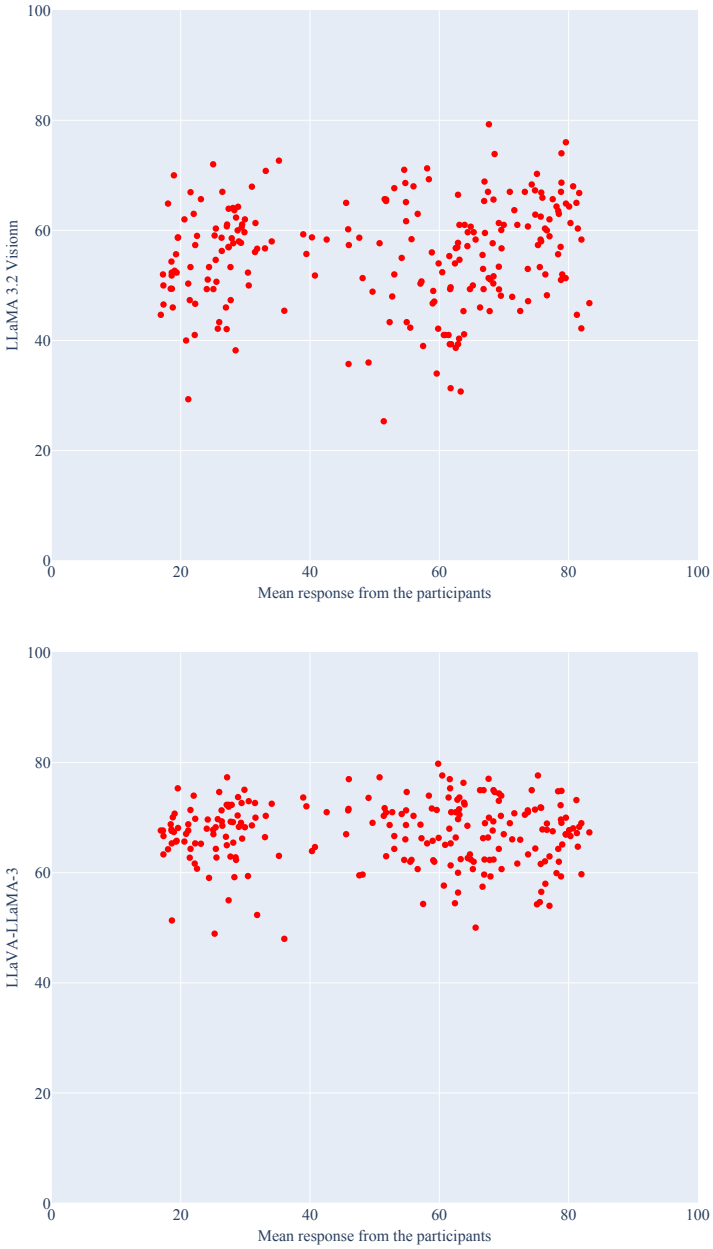


Fig. 13. Scatter plots comparing model predictions to human participant responses, **with memory**. The **top** image shows results for the *LLaMA 3.2 Vision* model; the **bottom** image for the *LLaVA LLaMA3* model.

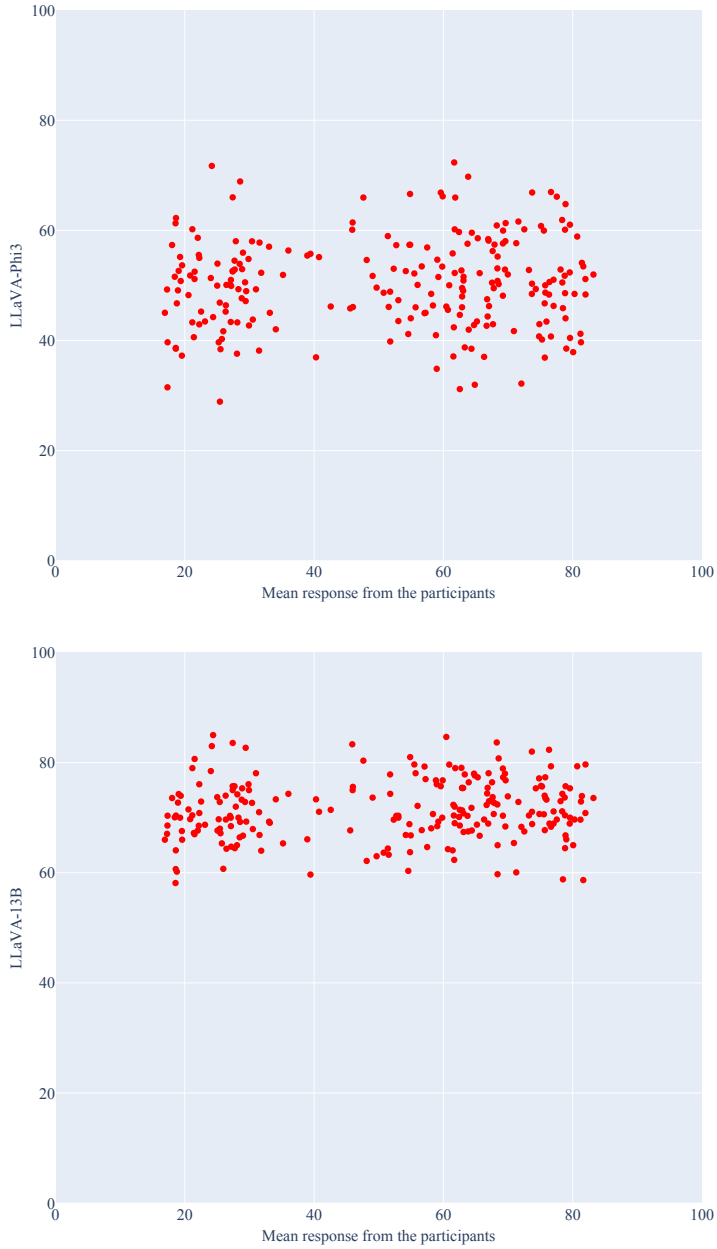


Fig. 14. Scatter plots comparing model predictions to human participant responses, **with memory**. The **top** image shows results for the *LLaVA Phi 3* model; the **bottom** image for the *LLaVA:13b* model.

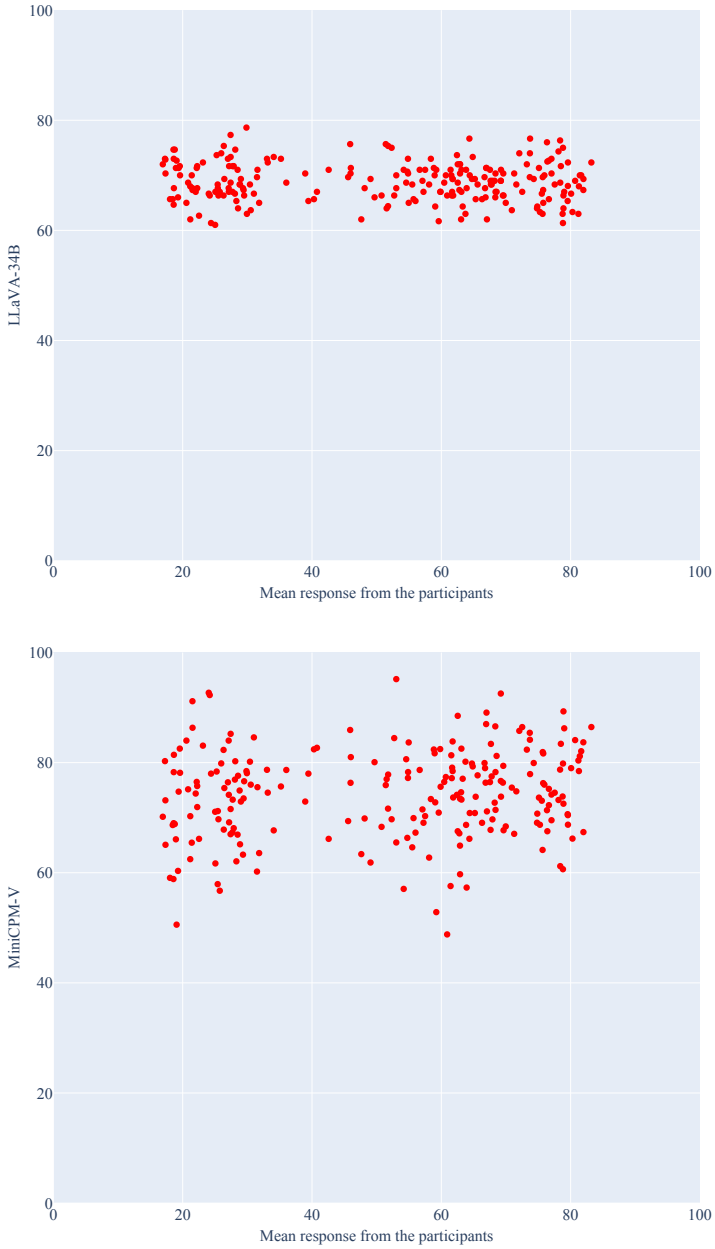


Fig. 15. Scatter plots comparing model predictions to human participant responses, **with memory**. The **top** image shows results for the *LLaVA:34b* model; the **bottom** image for the *MiniCPM-V* model.

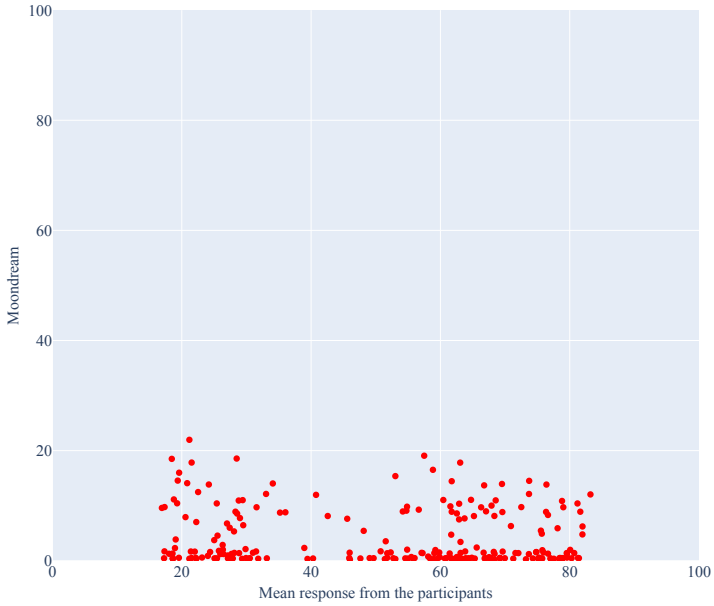


Fig. 16. Scatter plot comparing model predictions to human participant responses, **with memory**. The image shows results for the *Moondream* model.