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

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

Abstract

This study evaluates the effectiveness of large language model-based personas for assessing external Human-Machine Interfaces (eHMIs) in automated vehicles. 13 different models namely BakLLaVA, ChatGPT-4o, DeepSeek-VL2-Tiny, Gemma3:12B, Gemma3:27B, Granite Vision 3.2, LLaMA 3.2 Vision, LLaVA-13B, LLaVA-34B, LLaVA-LLaMA-3, LLaVA-Phi3, MiniCPM-V and Moondream were tasked with simulating pedestrian decision making for 227 vehicle images equipped with eHMI. Confidence scores (0-100) were collected under two conditions: no memory (images independently assessed) and memory-enabled (conversation history preserved), each in 15 independent trials. The model outputs were compared with the ratings of 1,438 human participants. Gemma3:27B achieved the highest correlation with humans without memory (r = 0.85), while ChatGPT-4o performed best with memory (r = 0.81). DeepSeek-VL2-Tiny and BakLLaVA showed little sensitivity to context, and LLaVA-LLaMA-3, LLaVA-Phi3, LLaVA-13B and Moondream consistently produced limited-range output.

CCS Concepts

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

Keywords

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. In *17th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI Adjunct '25), September 21–25, 2025, Brisbane, QLD, Australia.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3744335.3758477

1 Introduction

In 2017, Vaswani et al. [31] proposed a work titled "`Attention is all you need`", introducing the attention mechanism [5] to significantly enhance sequence-to-sequence (seq2seq) models [26]. This innovation paved the way for encoder-only architectures such as the Bidirectional Encoder Representations from Transformers (BERT) [14] and subsequently decoder-only models such as the Generative Pre-trained Transformer (GPT). Although statistical language models, often comprising millions of parameters and trained in corpora of tens to hundreds of millions of words, have existed since at least the 1990s, including n-gram–based systems [10] and large-scale statistical machine translation models such as IBM Models 1–5 [8], the transformer era marked a step change in scalability, adaptability, and multimodal capability. Since then, numerous Large Language Models (LLMs) have emerged, tailored for specific tasks such as medical analysis [23] and document processing [37], as well as general purpose models such as DeepSeek-VL2 [35] 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, ChatGPT-4.5 and LLaMA-3.1-405B have been judged as human in 73% and 56% of cases, respectively, under specific experimental protocols [18]. 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. [25] reported that Google's Bard LLM [27] 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. 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 [6], cyclists [33], and manually driven vehicle drivers [21]. The development of effective eHMIs is critical to ensuring traffic safety and fostering public trust in AV technology. Research on user-centric interfaces in automated mobility has shown that providing clear and intuitive information is essential for passenger confidence and effective navigation [4]. However, evaluating the effectiveness of eHMI designs remains a significant challenge [16]. Synthetic and simulated traffic scenarios are increasingly being used to test AV systems and human responses, with recent advances in deep learning enabling the creation of realistic and varied traffic scenes [2, 3]. To evaluate eHMI concepts within these scenarios, researchers such as Bazilinskyy et al. [7] and Cumbal et al. [13] have relied

*Corresponding Author

AutomotiveUI Adjunct '25, Brisbane, QLD, Australia

ACM ISBN 979-8-4007-2014-7/2025/09
https://doi.org/10.1145/3744335.3758477

on large-scale crowdsourced experiments that gather human judgment and feedback. 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 [9].

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 [34]. 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 [17], 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

Figure 1: Example stimulus image used in the study, showing an automated vehicle equipped with an external Human-Machine Interface displaying the message "CAR IS STOPPING". Participants and models evaluated images of this format to interpret vehicle intent and pedestrian safety.

This study uses crowdsourced results compiled by Bazilinskyy et al. [7], involving 1,438 participants who evaluated 227 distinct textual eHMIs displayed on an AV. Participants indicated their willingness to cross via a slider scale ranging from 0 (absolute unwillingness) to 100 (complete confidence). 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.

In this study, a total of 13 VLMs were evaluated, namely *Bak-LLaVA, ChatGPT-4o, DeepSeek-VL2-Tiny, Gemma3: 12B, Gemma3: 27B, Granite Vision 3.2, LLaVA:13B, LLaVA:34B, LLaVA-LLaMA-3, LLaVA-Phi3, LLaMA3.2-vision, MiniCPM-V* and *Moondream*. The DeepSeek-VL2-Tiny model was obtained from Hugging Face (https://huggingface.co/models), and inference for ChatGPT-4o was conducted using the OpenAI ChatGPT API (https://platform.openai.com/docs/overview). All remaining models were downloaded from Ollama (https://ollama.com). A summary of the models evaluated, including their base architectures and acquisition or deployment methods, is provided in Table 1. In particular, the use of the ChatGPT API incurred a cost of 20€, whereas all other VLMs were accessed at no cost.

Two different evaluation conditions were defined for each VLM: a *no memory* condition and a *memory-enabled* condition. In the no-memory condition, the model processed each image independently, without access to any prior conversational history or previous responses; each image was treated as a standalone input. In the memory-enabled condition, the conversational context was preserved and provided incrementally to the model across the sequence of images, so that each prompt (from the second image onwards) included the structured history of prior prompt-response pairs, allowing the model to utilise the accumulated context when generating responses.

For each condition, 15 independent trials were conducted, corresponding to random seed values from 0 to 14. In each trial, the full set of 227 images was presented sequentially in a fixed order determined by the respective seed, ensuring that the sequence of images was identical for both conditions across all trials. At the beginning of each trial (irrespective of whether there is no memory or memory enabled condition), a standardised prompt was given to the model, instructing it to interpret the vehicle's displayed message, infer its implications for pedestrian safety, and assign a confidence score on a scale from 0 (certainly unsafe) to 100 (certainly safe). For every new trial, the first image was introduced with the following prompt:

> *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].*

In the memory-enabled condition, from the second image onwards, the prompts additionally included the conversation history, which was preserved as structured JSON data using LangChain (https://www.langchain.com). Models were explicitly instructed to consider previous responses when evaluating pedestrian safety. To ensure manageability and comparability, each trial was limited to a maximum of six historical prompt-response pairs, as suggested by See et al. [22] to reflect a reasonable window for maintaining a coherent conversational context. The sequence of prompts used for

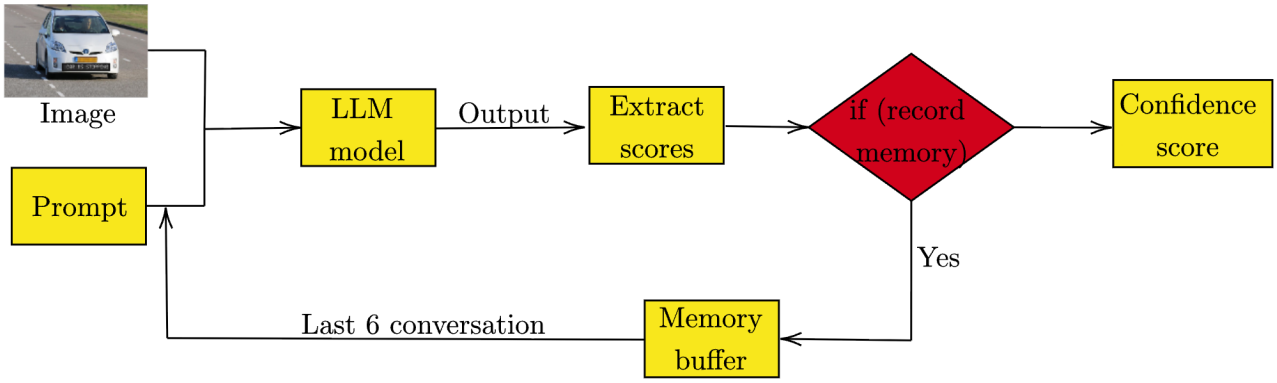

Figure 2: Flow diagram of the system architecture showing image processing, prompting, model querying, response correction, and downstream analysis.

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

Model Name	Architecture	Platform
BakLLaVA [24]	LLaMA 2 + BLIP	Ollama
ChatGPT-4o [20]	Multimodal transformer	OpenAI (ChatGPT API)
DeepSeek-VL2-Tiny [35]	DeepSeek VLM	Hugging Face
Gemma3: 12B [28]	Gemma 3 vision-capable model	Ollama
Gemma3: 27B [28]	Gemma 3 vision-capable model	Ollama
Granite Vision 3.2 [29]	IBM Granite vision model	Ollama
LLaVA:13B [19]	LLaMA + CLIP	Ollama
LLaVA:34B [19]	LLaMA + CLIP	Ollama
LLaVA-LLaMA-3 [11]	LLaMA-3 + Visual Adapter	Ollama
LLaVA-Phi3 [12]	Phi-3 + Vision module	Ollama
LLaMA3.2-vision [30]	LLaMA 3.2 + Vision	Ollama
MiniCPM-V [36]	CPM-based	Ollama
Moondream [32]	Distilled CLIP variant	Ollama

each image *from the second image onwards* in the memory-enabled condition is detailed below:

(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 2). 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 through the OpenAI API and processed remotely on OpenAI's servers, all other models, including DeepSeek-VL2-Tiny, were deployed and executed locally. Local deployments were managed through Ollama (supporting batch mode image processing), while DeepSeek-VL2-Tiny was run using code and weights obtained directly from Hugging Face (https://huggingface.co/deepseek-ai/deepseek-vl2-tiny). All local experiments were conducted on a workstation equipped with an AMD Ryzen 9 7950X3D 16-core processor, 32 GB RAM, and a 16 GB NVIDIA GeForce RTX 4080 graphics card. 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.

Once all model responses were collected, we used the local deepseek-r1:14b model (https://ollama.com/library/deepseek-r1:14b) to extract the numerical confidence scores from the output. This step was necessary because some models provided very short and direct answers (for example, just a number), while others gave

much longer and more detailed explanations—sometimes including additional context or commentary along with the score. To ensure consistency and accurately capture the intended confidence scores, we prompted the model to identify and extract only the numerical value from each response. The following prompt was used for this extraction:

> *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 simple number-extraction function was then used to find the first number mentioned in the model's response. This function scans the text for the first instance of a number (either integer or decimal) and returns it as a numerical value. If no valid number was found in the text, the response was recorded as NaN (not a number) to keep the dataset consistent. All extracted values were saved for later statistical analysis, allowing us to compare the model's confidence scores with human benchmarks and evaluate how well each model interpreted pedestrian crossing decisions based on eHMI messages.

After collecting the model output, we filtered the results to include only responses with values between 0 and 100 (inclusive). Any responses with values outside this range were excluded from further analysis. The remaining filtered values were then used to calculate the average confidence scores for each model.

3 Results

The mean confidence scores for each model are presented in the Appendix (see Table A1 and A2). For each image, the average confidence scores generated by each model were compared to crowdsourced results as reported by Bazilinskyy et al. [7].

Without conversational memory, models such as BakLLaVA overwhelmingly produced a confidence score of 0 for 222 out of 227 cases (mean = 0.07, SD = 0.66). By contrast, LLaVA-LLaMA-3 (mean = 17.24, SD = 6.56) produced a maximum score of 36.67 for the prompt "PROCEED TO CROSS NOW", while the corresponding crowdsourced average was 51.58. The Moondream model (mean = 35.75, SD = 8.05) generated scores ranging from 13.35 to 56.16, while LLaVA-13B (mean = 57.00, SD = 9.51) provided outputs between 31.00 and 77.23. LLaVA-34B reported a mean of 59.79 (SD = 15.32), with values ranging from 90.33 ("PROCEED TO CROSS PLEASE") to 20.36 ("WILL NOT STOP").

DeepSeek-VL2-Tiny produced some scores in the 0 to 100 range, but the majority were fixed at 0 (54 times), 75 (99 times), or 90 (65 times). For example, this model assigned a score of 100 to "YOU CAN WALK," but returned 0 for prompts such as "DO NOT WALK", "I WON'T STOP", and "NO CRUCE". The message "CAR WILL NOT STOP" received a score of 90.

MiniCPM-V yielded a mean score of 72.27 (SD = 12.80), with prompts like "I'LL PROCEED" and "I'M ACCELERATING NOW" receiving average scores of 82.14 and 75.67, respectively. Gemma3: 27B had a mean of 53.39 (SD = 35.95), assigning a maximum value of 100 to texts such as "CAR WILL NOT MOVE" and "CROSS NOW." ChatGPT-4o achieved a mean score of 59.09 (SD = 34.10), with its highest (94.8) given to the eHMI message "SEGURO PARA CRUZAR".

With conversational memory enabled, some models exhibited a broader distribution in their output. BakLLaVA (mean = 2.10, SD = 3.36) generated values primarily between 0 and 13.33, showing a slightly wider spread than in the memory-free condition. LLaVA-LLaMA-3 (mean = 67.13, SD = 5.82) produced a maximum of 79.80 ("I WON'T DRIVE") and a minimum of 48.00 ("PARAR"). Moondream (mean = 4.22, SD = 5.07) reported scores from 0.20 ("EL VEHÍCULO SE DETIENE") to 21.96 ("DO NOT CONTINUE"). LLaVA-34B (mean = 68.87, SD = 3.52) produced values from 78.67 ("I'LL MOVE") to 61.00 ("ESPERE POR FAVOR").

DeepSeek-VL2-Tiny (mean = 98.67, SD = 1.00) shifted to consistently high output with memory enabled, with a minimum score of 92 for messages such as "NO CRUCE", "NO REPRESENTO NINGÚN PELIGRO", and "DO NOT GO". MiniCPM-V had a mean of 74.26 (SD = 8.11), with prompts such as "I SEE YOU" and "AFTER YOU" receiving average scores of 95.17 and 70.73, respectively.

Gemma3: 27B increased to a mean of 79.80 (SD = 9.38), with a maximum value of 97.13 ("PLEASE WALK") and a minimum of 28.8 ("DO NOT WALK"). ChatGPT-4o reported a mean of 59.38 (SD = 34.04), with its highest score (99.67) also assigned to "SEGURO PARA CRUZAR".

Among the models evaluated without conversation history, Gemma3: 27B exhibited the strongest alignment with human responses, producing high correlation coefficients with the mean ($r = 0.85$) 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 $r = 0.32$, $r = 0.34$ and $r = 0.31$, respectively, with the mean response. Models such as LLaVA-LLaMA-3, Moondream, and Granite Vision 3.2 showed the weakest correlations, with coefficients of $r = -0.07$, $r = 0.11$, and $r = 0.17$, 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 $r = 0.85$ 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.

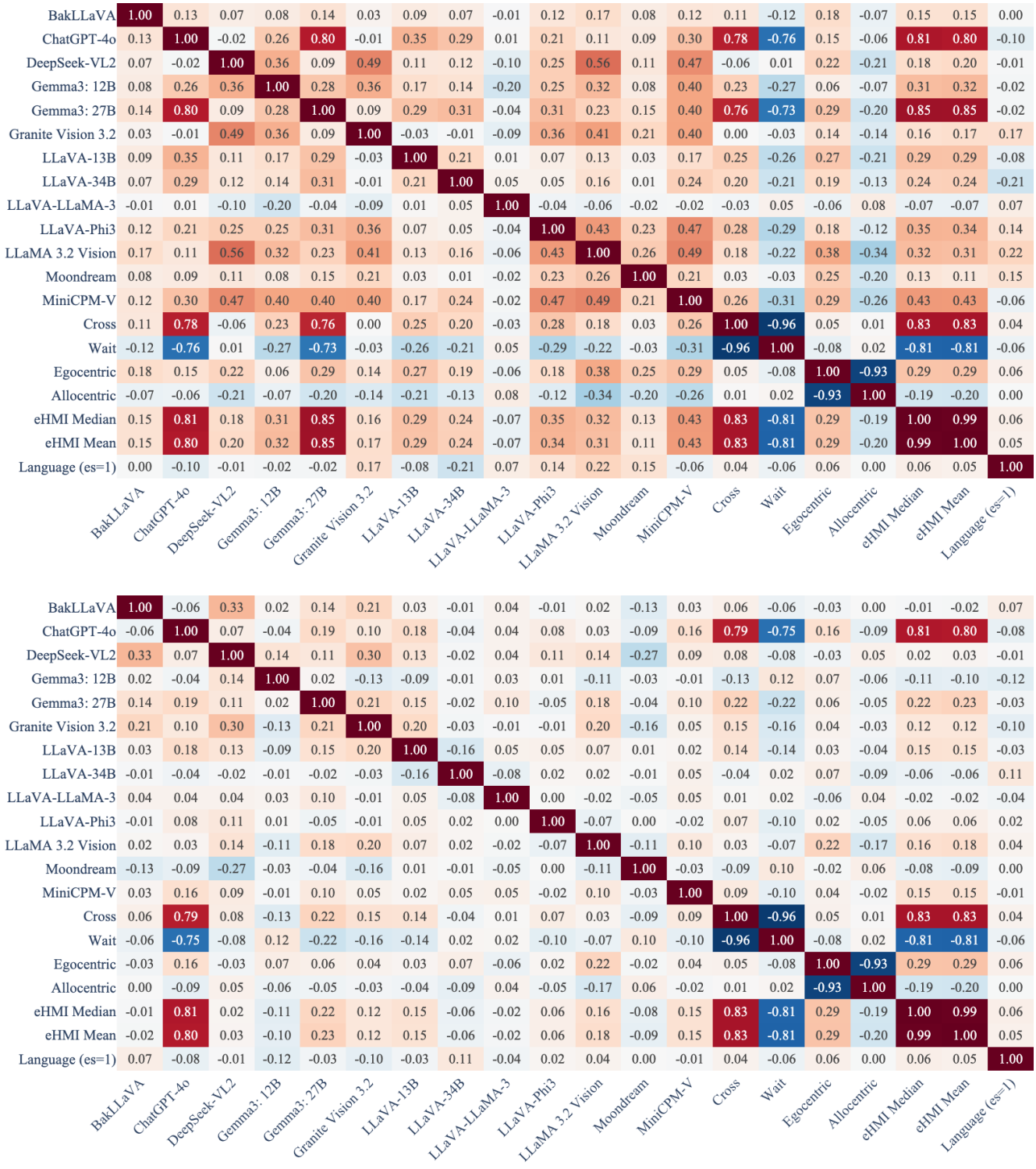

Figure 3: Spearman correlation matrices of model outputs, behavioural features, and language encoding. The top image shows results *without memory*, and the bottom image shows results *with memory*. Binary behavioural features ("Cross", "Wait", "Egocentric", and "Allocentric") are encoded as 0 (absent) 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 colours, negative in cooler colours.

4 Discussion

The results indicate that the alignment of the VLM with human crossing judgments is primarily determined by the architectural sophistication of the model, the multimodal pretraining, and the strategy for handling the conversational context. High-capacity models such as ChatGPT-4o and Gemma3:27B achieved the strongest correlation with human ratings, whereas most other models deviated substantially – especially once prior dialogue context was introduced. In fact, Gemma3:27B obtained a near-human correlation (about $r = 0.85$) when evaluating images independently (memory-free), while ChatGPT-4o uniquely maintained high alignment (peaking around $r = 0.81$) even with conversational history. In contrast, smaller or less specialised VLMs showed significantly lower and less stable agreement with human judgments.

The superior performance of ChatGPT-4o and Gemma3:27B can be attributed to their scale and advanced cross-modal attention mechanisms, which allow them to effectively integrate textual eHMI cues with the surrounding visual context. These models not only read explicit messages on the vehicle (e.g., "CROSS NOW" or "SEGURO PARA CRUZAR") but also interpret them in light of the scene, much like a human would. For example, Gemma3:27B correctly interpreted an eHMI display reading "I HAVE SEEN YOU" by assigning a moderately confident crossing score (65) and explaining that the car has detected the pedestrian by stating *The message indicates the vehicle has detected a pedestrian, suggesting it is aware of their presence and potentially adjusting its behavior accordingly. However, it does *not* guarantee the pedestrian's safety, as the vehicle's actions still depend on its programming and the specific situation.* Similarly, ChatGPT-4o, presented with "SEGURO PARA CRUZAR" (Spanish for "SAFE TO CROSS"), returned a full confidence score (100) after recognising the phrase as explicit and providing reason as *The text "SEGURO PARA CRUZAR" translates to "SAFE TO CROSS," indicating that it is safe for a pedestrian to cross the street.* These examples illustrate how large, well-trained models leverage both textual content and environmental context to produce nuanced, context-aware risk assessments.

In contrast, smaller or less specialised models (e.g., BakLLaVA or certain LLaVA-based variants) tended to lack this nuance. They often defaulted to extreme or invariant confidence scores, suggesting that they rely on rigid input-output mappings rather than truly understanding the scenario. For example, LLaVA-LLaMA-3 frequently produced low confidence (20) even for an ostensibly positive signal like "VEHICLE STOPS," stating *The car is an autonomous vehicle that is driving down the street and has a sign on the front that says "VEHICLE STOPS". This implies that the pedestrian should be cautious and make sure they do not cross the street when the sign is displayed.* This behaviour implies that the model misinterpreted or ignored the reassuring cue, instead applying a one-size-fits-all rule (e.g., always err on the side of caution) due to limited feature extraction capabilities. Such outcomes underscore the brittleness of smaller VLMs: Without sufficient architectural complexity or targeted training, they struggle to adjust their judgments to different eHMI messages and situational details.

A particularly notable pattern was observed with DeepSeek-VL2-Tiny, which quantised its output into only a few discrete levels. This model assigned most scenarios a confidence of 0, 75 or 90, indicating a coarse category-based decision process rather than a continuous evaluation. This output binning likely reflects architectural simplifications or a training paradigm emphasising categorical judgments, leading to a loss of granularity. Moreover, DeepSeek-VL2-Tiny often failed to correctly interpret the eHMI text. For example, when shown "ESTOY ACELERANDO AHORA" ("I am accelerating now"), it still gave high crossing confidence (90) with an irrelevant rationale about the automation features of the vehicle stating *The message "ESTOY ACELERANDO AHORA" translates to "I am accelerating now," which suggests that the car's system is actively engaged in its driving mode, potentially indicating advanced safety features such as adaptive cruise control or lane-keeping assistance. This implies that while there may be some level of automation involved, the driver remains responsible for safe navigation through the intersection. Therefore, the confidence score would likely be around 90, reflecting moderate confidence in the system's ability to ensure pedestrian safety.* In reality, a message indicating that the car is accelerating should reduce the confidence of a pedestrian to cross. The counterintuitive response of the model suggests that it did not properly integrate the warning meaning of the text, perhaps because its training did not cultivate a fine-grained understanding of how linguistic cues relate to pedestrian risk.

Conversational memory (i.e., providing the model with a running history of previous questions and answers) had a striking effect on performance. Counterintuitively, adding memory reduced model-human alignment for almost all models except ChatGPT-4o. Most VLMs, when given a sequence of prior interactions, appeared to overweight the earlier exchanges, allowing previous answers to bleed into subsequent ones. Instead of treating each crossing scenario independently, these models smoothed or biased their confidence scores based on what came before. This behaviour contrasts with human pedestrians, who reset their expectations and make decisions based on the current context rather than the last situation. For example, the Gemma3:27B correlation with human ratings plummeted from about $r = 0.85$ without memory to $r = 0.22$ once the conversational context was included. In one trial, after several Q&A rounds, Gemma3:27B was shown a new image with the eHMI message "WARNING, I'M DANGEROUS." Instead of lowering its confidence in crossing as expected for a danger warning, the model hallucinated a prior message ("I'LL WALK") from the conversation history and responded with an unjustified interpretation of safe crossing saying *The message "I'LL WALK" indicates the vehicle is permitting a pedestrian to cross the road, suggesting it is safe to do so, but one should still exercise caution and check for other traffic.* Similarly, LLaVA-LLaMA-3, when faced with "CRUZANDO" ("crossing") in a context-rich session, outputs a confidence of 0 accompanied by an irrelevant statement about the vehicle not being authorised to operate(*Autonomous vehicle has not been granted the right to operate*). In both cases, the presence of earlier examples in memory led to clearly inappropriate output: The models became confused by or fixated on the past context, failing to fully process the new visual cue. These errors illustrate how naïvely incorporating memory can impair a model's decision consistency by diluting the influence of immediate evidence.

The exception is ChatGPT-4o, which maintains strong human alignment even when conversation history is enabled. This robustness is likely due to hierarchical attention mechanisms and explicit

context gating, which enable the model to preserve sensitivity to new information and selectively downweight older, less relevant context. This finding underscores that conversational memory can enhance performance, but only when memory is handled in a sophisticated and human-like manner.

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. [15] 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, 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.

4.1 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 as model capabilities improve. 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. Most models tended to perform worse when provided with conversational history, in contrast to humans who typically evaluate each crossing decision independently. This suggests that current strategies for providing context may not accurately reflect how people process sequential decisions and that further work is needed to develop more sophisticated context handling approaches.

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 that displays the message "I WILL NOT STOP" the model rated confidence of 80 and responded: *The message displayed on the vehicle, "I WILL NOT STOP," implies that the car is an autonomous vehicle and it will not halt for pedestrians or other road users, which may suggest a lack of safety in certain situations. Pedestrians should exercise caution while crossing the street as they might need to rely on their own judgement instead of waiting for the vehicle to stop.* Similarly, when the vehicle displayed the message OK TO PASS", the model rated the confidence as 60 and responded: *Meaning: The message displayed on the vehicle says "OK TO CROSS", which implies that it is safe for a pedestrian to cross the street in front of this autonomous vehicle since the car has acknowledged their presence and intends to stop. However, as with any road situation, there are always inherent risks involved, so my confidence level is moderate rather than absolute certainty.* These examples demonstrate that, although the model can extract and read textual information, it does not always adequately ground that information at 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.

Another limitation concerns the generalisability of the findings across languages and cultures. Although the study included English and Spanish eHMI messages, it did not systematically test the model's abilities to interpret nuanced, culturally specific, or less common linguistic cues. This may limit the applicability of the results in more diverse or region-specific settings.

Finally, this study benchmarks VLM responses against crowdsourced human ratings, rather than actual pedestrian behaviour in real-world or highly realistic simulation environments. As such, the ecological validity of these findings remains to be established through direct behavioural comparison.

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. 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. In addition, improved methods for providing and handling conversational contexts, such as specialised training, architectural enhancements, or better context window management, should be explored to more closely mirror human reasoning. In addition, real-world validation experiments that compare LLM-derived decisions directly with actual pedestrian behaviour could improve the ecological validity of these findings. Expanding the evaluation to cover a wider range of traffic scenarios, including interactions with cyclists, other vehicles, and more complex environmental cues, may also provide a deeper understanding of the applicability of VLM. Finally, studies involving a more diverse set of languages and cultural contexts are needed to assess and improve the generalisability and inclusion of VLM-based UX evaluation tools in traffic safety and beyond.

Supplementary Material

The analysis code and responses of LLMs are available at https://doi.org/10.4121/cb208bd8-7cf4-42d5-ae5e-9ad2c654aeb3. A maintained version of code is available 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] Md Shadab Alam, Marieke H Martens, and Pavlo Bazilinskyy. 2025. Generating Realistic Traffic Scenarios: A Deep Learning Approach Using Generative Adversarial Networks (GANs). In *13th International Conference on Human Interaction & Emerging Technologies: Artificial Intelligence & Future Applications, IHIET-AI 2025*. AHFE International, 349–358. https://doi.org/10.54941/ahfe1005927

[3] Md Shadab Alam, Sagar Hitendra Parmar, Marieke H. Martens, and Pavlo Bazilinskyy. 2025. Deep Learning Approach for Realistic Traffic Video Changes Across Lighting and Weather Conditions. In *2025 8th International Conference on Information and Computer Technologies (ICICT)*. Hilo, HI, USA, 180–185. https://doi.org/10.1109/ICICT64582.2025.00034

[4] Md Shadab Alam, Thirumanikandan Subramanian, Marieke Martens, Wolfram Remlinger, and Pavlo Bazilinskyy. 2024. From A to B with ease: User-centric interfaces for shuttle buses. In *Adjunct Proceedings of the 16th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutoUI)*. Stanford, CA, USA. https://doi.org/10.1145/3641308.3685033

[5] 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

[6] 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

[7] 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

[8] Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.* 19, 2 (June 1993), 263–311. https://dl.acm.org/doi/10.5555/972470.972474

[9] 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

[10] Stanley F Chen and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language* 13, 4 (1999), 359–394. https://doi.org/10.1006/csla.1999.0128

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

[12] 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.

[13] 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).

[14] 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

[15] 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

[16] Ruolin Gao, Rutger Verstegen, Haoyu Dong, Pavlo Bazilinskyy, and Marieke Martens. 2024. Incorporating Multiple Users' Perspectives in HMI Design for Automated Vehicles: Exploration of a Role-Switching Approach. In *Adjunct Proceedings of the 16th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Stanford, CA, USA) *(AutomotiveUI '24 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 197–202. https://doi.org/10.1145/3641308.3685047

[17] 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

[18] 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]

[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. https://doi.org/10.48550/arXiv.2304.08485

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

[21] 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

[22] 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

[23] 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. https://doi.org/10.1186/s12911-025-02954-4

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

[25] 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

[26] 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

[27] 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).

[28] 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).

[29] 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).

[30] 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).

[31] 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/10.5555/3295222.3295349

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

[33] 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

[34] 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).

[35] 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).

[36] 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).

[37] 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).

Table A1: Confidence score statistics for human participants and each model in the memory-free condition, enabling comparison of model and human judgments without conversational context.

Model	Mean	SD	Median	Max value	Min value
Human Response	51.57	21.22	58.11	83.20	16.92
BakLLaVA	0.07	0.66	0.00	7.14	0.00
ChatGPT-4o	59.09	34.10	78.42	94.80	0.00
DeepSeek-VL2-Tiny	61.51	34.80	75.00	100.0	0.00
Gemma3: 12B	56.49	30.30	62.79	97.86	0.07
Gemma3: 27B	53.39	35.95	64.67	100	0.00
Granite Vision 3.2	40.39	14.96	39.62	81.0	8.91
LLaVA: 13B	57.00	9.51	58.07	77.23	31.00
LLaVA: 34B	59.79	15.32	59.29	90.33	20.36
LLaVA-LLaMA-3	17.24	6.56	16.00	36.67	3.83
LLaVA-Phi3	79.15	7.78	78.99	94.96	57.91
LLaVA3.2-vision	5.22	23.27	61.43	91.42	3.57
MiniCPM-V	72.27	12.80	74.58	95.33	34.21
Moondream	35.75	8.05	35.93	56.16	13.35

Table A2: Confidence score statistics for human participants and each model in the memory-enabled condition, showing the effects of conversational history on model and human alignment.

Model	Mean	SD	Median	Max value	Min value
Human Response	51.57	21.22	58.11	83.20	16.92
BakLLaVA	2.10	3.36	0.00	13.33	0.00
ChatGPT-4o	59.38	34.04	77.40	99.67	0.00
DeepSeek-VL2-Tiny	98.67	1.00	98.67	99.33	92.00
Gemma3: 12B	67.20	7.69	66.40	87.2	46.73
Gemma3: 27B	79.80	9.38	81.07	97.13	28.80
Granite Vision 3.2	47.51	3.77	46.92	61.33	36.92
LLaVA: 13B	71.41	5.46	70.79	85.00	58.13
LLaVA: 34B	68.87	3.52	68.67	78.67	61.00
LLaVA-LLaMA-3	67.13	5.82	67.80	79.80	48.00
LLaVA-Phi3	50.36	8.20	50.32	72.33	28.89
LLaVA3.2-vision	55.65	9.53	57.00	79.27	25.33
MiniCPM-V	74.26	8.11	74.73	95.17	48.80
Moondream	4.22	5.07	1.42	21.96	0.20

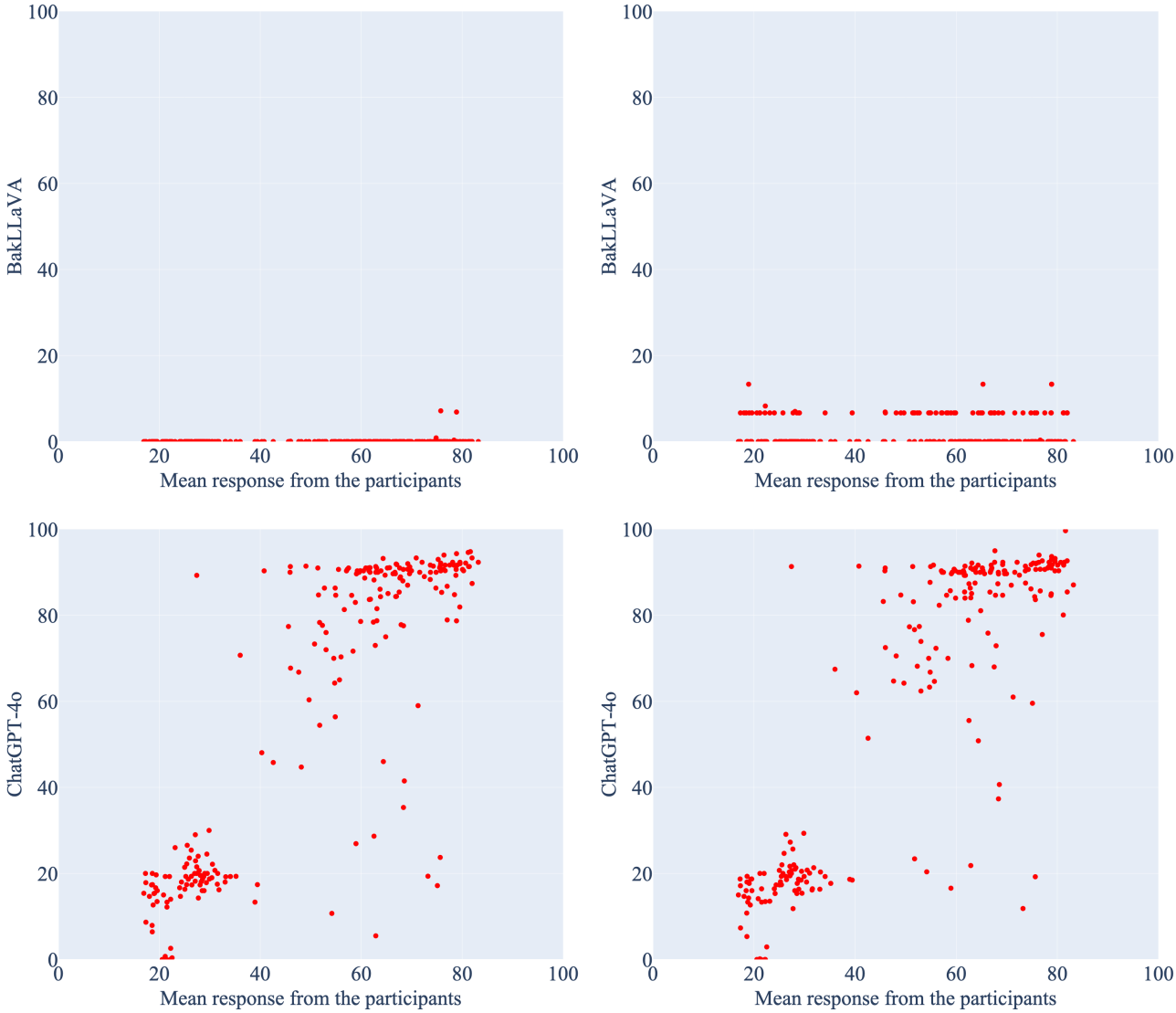

Figure A1: Scatter plots comparing model predictions to human participant responses for *BakLLaVA* and *ChatGPT-4o*. For each model, results without memory are shown on the left (*BakLLaVA*: $r = 0.15$, *ChatGPT-4o*: $r = 0.80$), and results with memory on the right (*BakLLaVA*: $r = -0.02$, *ChatGPT-4o*: $r = 0.80$).

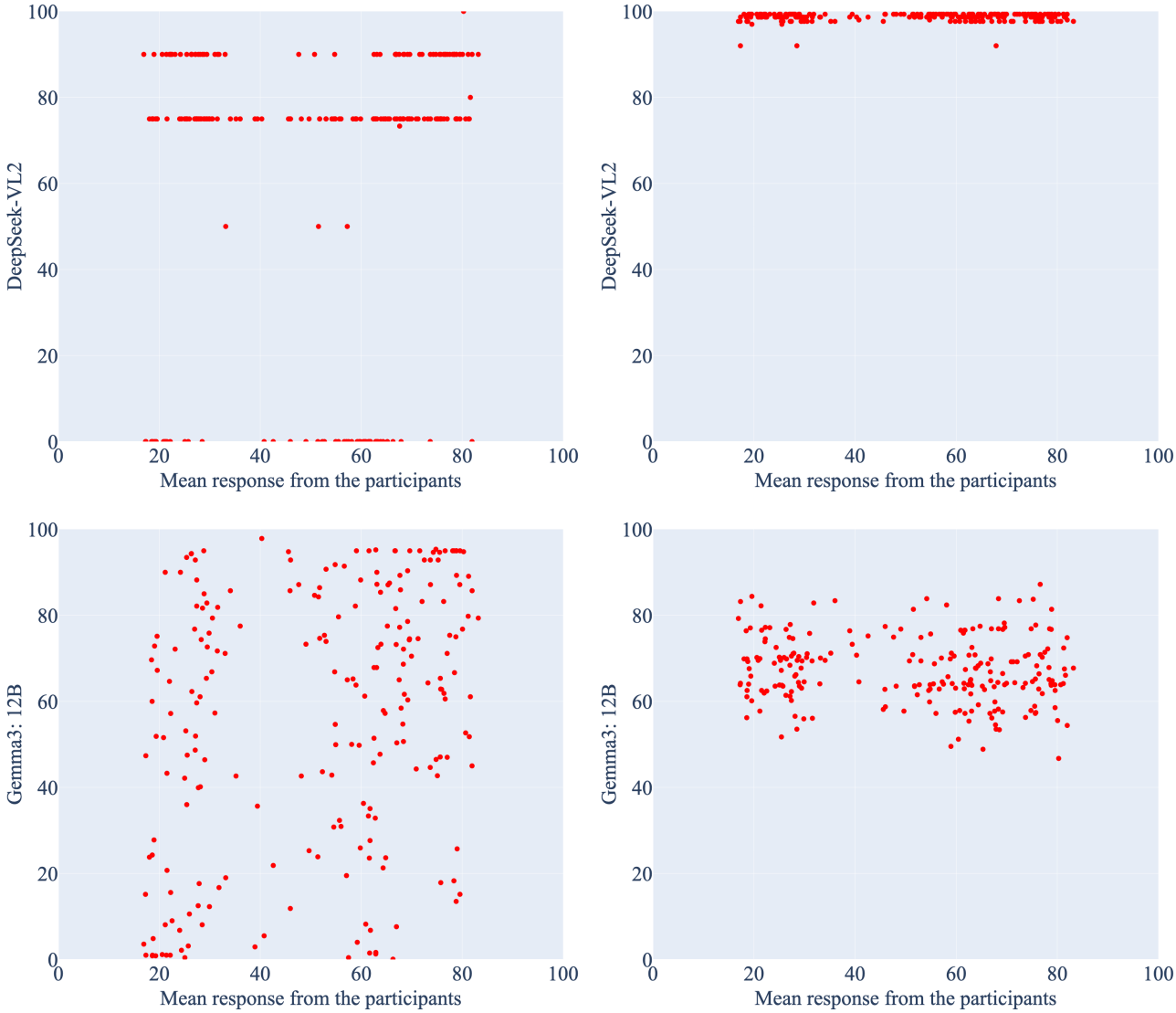

Figure A2: Scatter plots comparing model predictions to human participant responses for *DeepSeek-VL2-Tiny* and *Gemma3: 12b*. For each model, results without memory are shown on the left (*DeepSeek-VL2-Tiny*: $r = 0.20$, *Gemma3: 12b*: $r = 0.32$), and results with memory on the right (*DeepSeek-VL2-Tiny*: $r = 0.03$, *Gemma3: 12b*: $r = -0.10$).

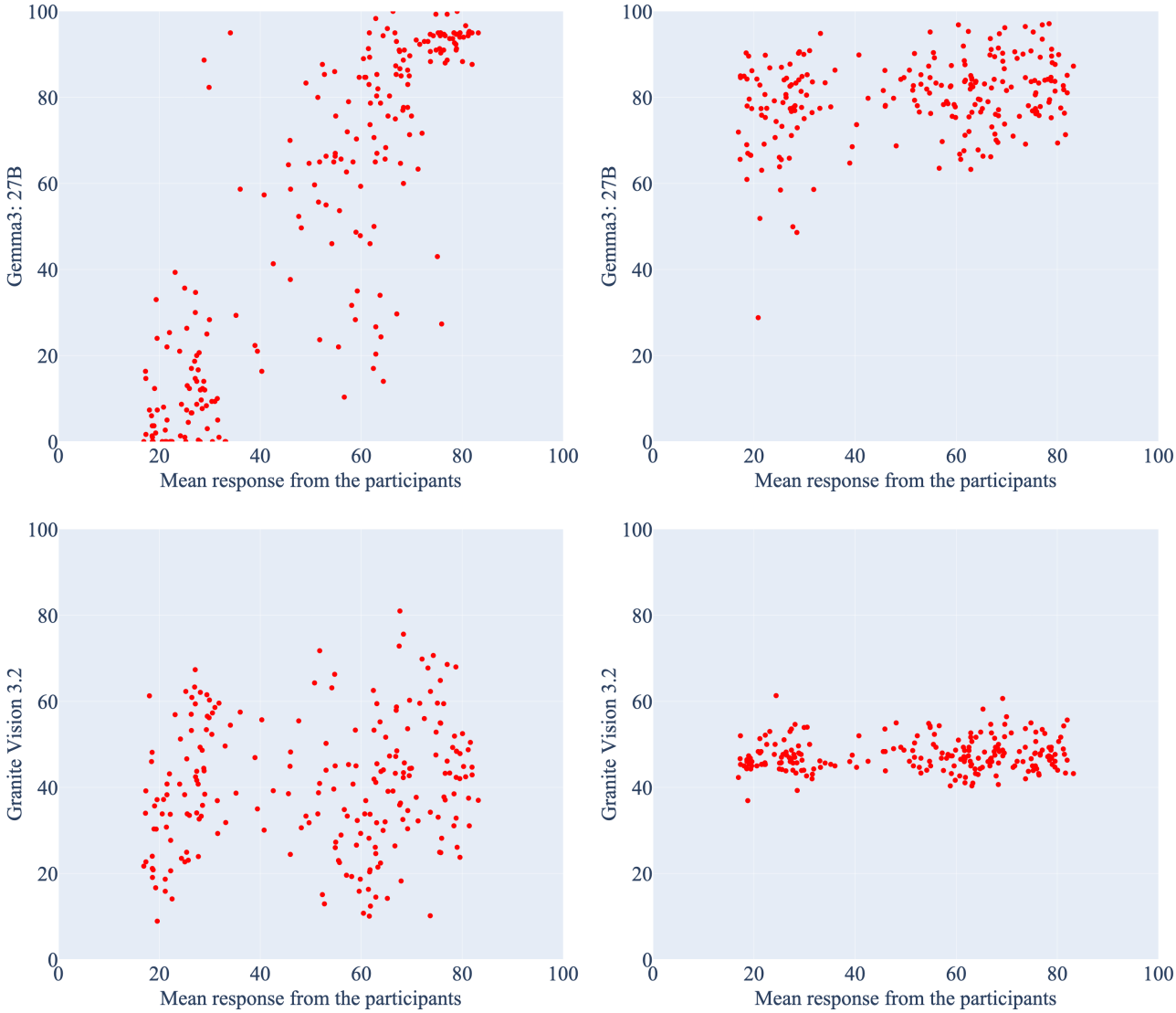

Figure A3: Spearman correlation plots comparing model predictions to human participant responses for *Gemma3: 27b* and *Granite Vision 3.2*. For each model, results without memory are shown on the left (*Gemma3: 27b*: $r = 0.85$, *Granite Vision 3.2*: $r = 0.17$), and results with memory on the right (*Gemma3: 27b*: $r = 0.23$, *Granite Vision 3.2*: $r = 0.12$).

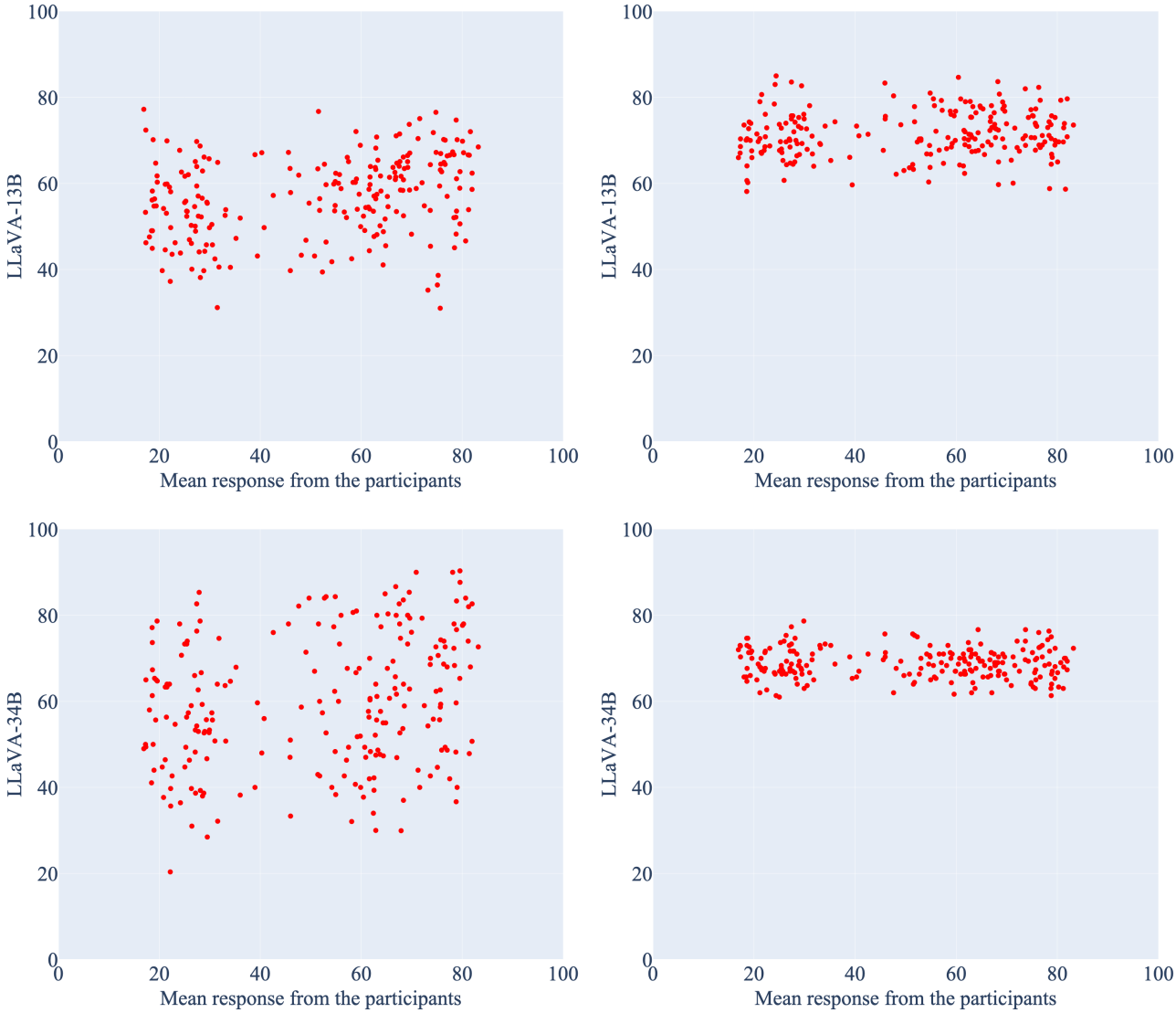

Figure A4: Scatter plots comparing model predictions to human participant responses for *LLaVA:13b* and *LLaVA:34b*. For each model, results without memory are shown on the left (*LLaVA:13b*: $r = 0.29$, *LLaVA:34b*: $r = 0.24$), and results with memory on the right (*LLaVA:13b*: $r = 0.15$, *LLaVA:34b*: $r = -0.06$).

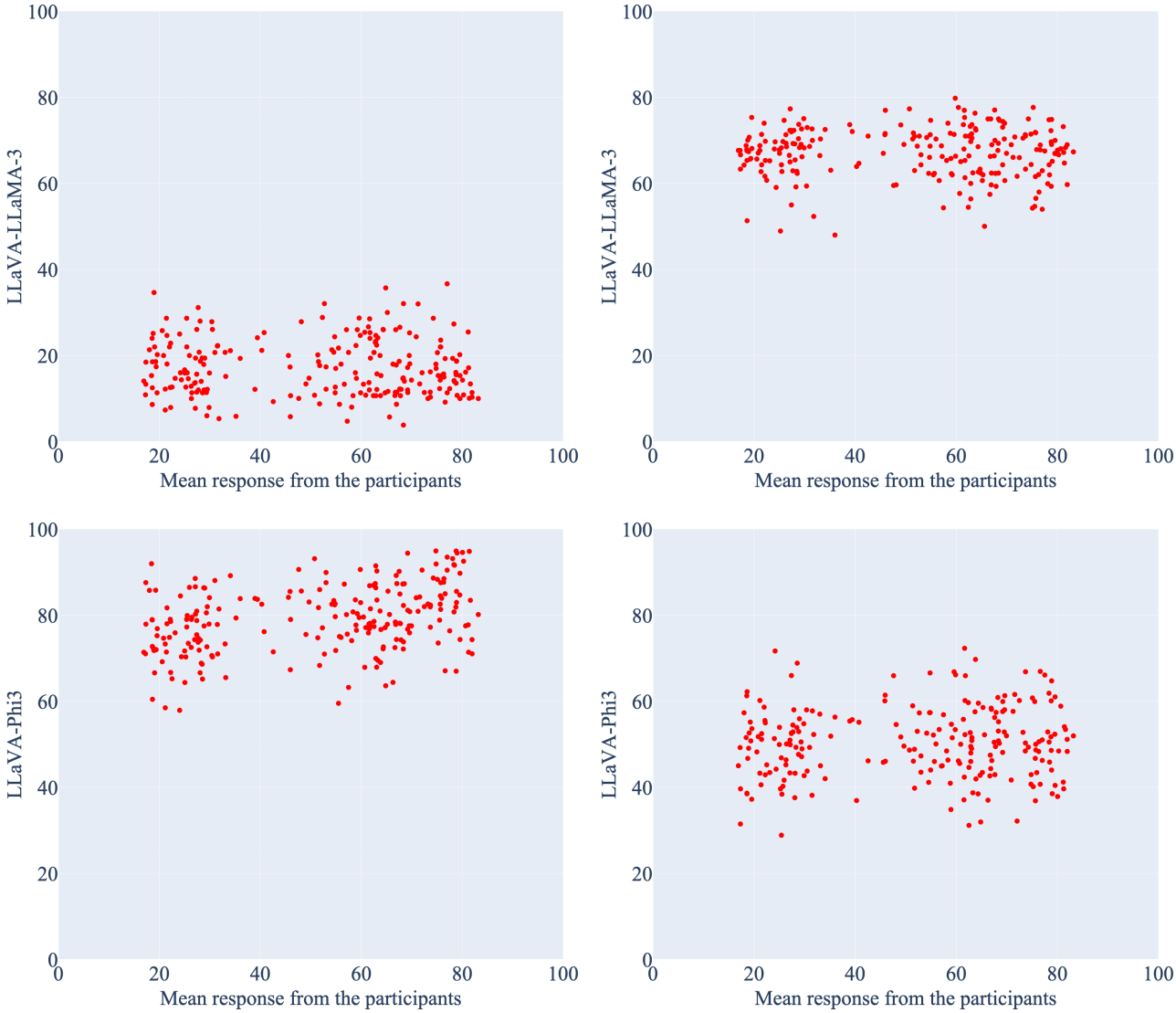

Figure A5: Scatter plots comparing model predictions to human participant responses for *LLaVA LLaMA3* and *LLaVA Phi 3*. For each model, results without memory are shown on the left (*LLaVA LLaMA3*: $r = -0.07$, *LLaVA Phi 3*: $r = 0.34$), and results with memory on the right (*LLaVA LLaMA3*: $r = -0.02$, *LLaVA Phi 3*: $r = 0.06$).

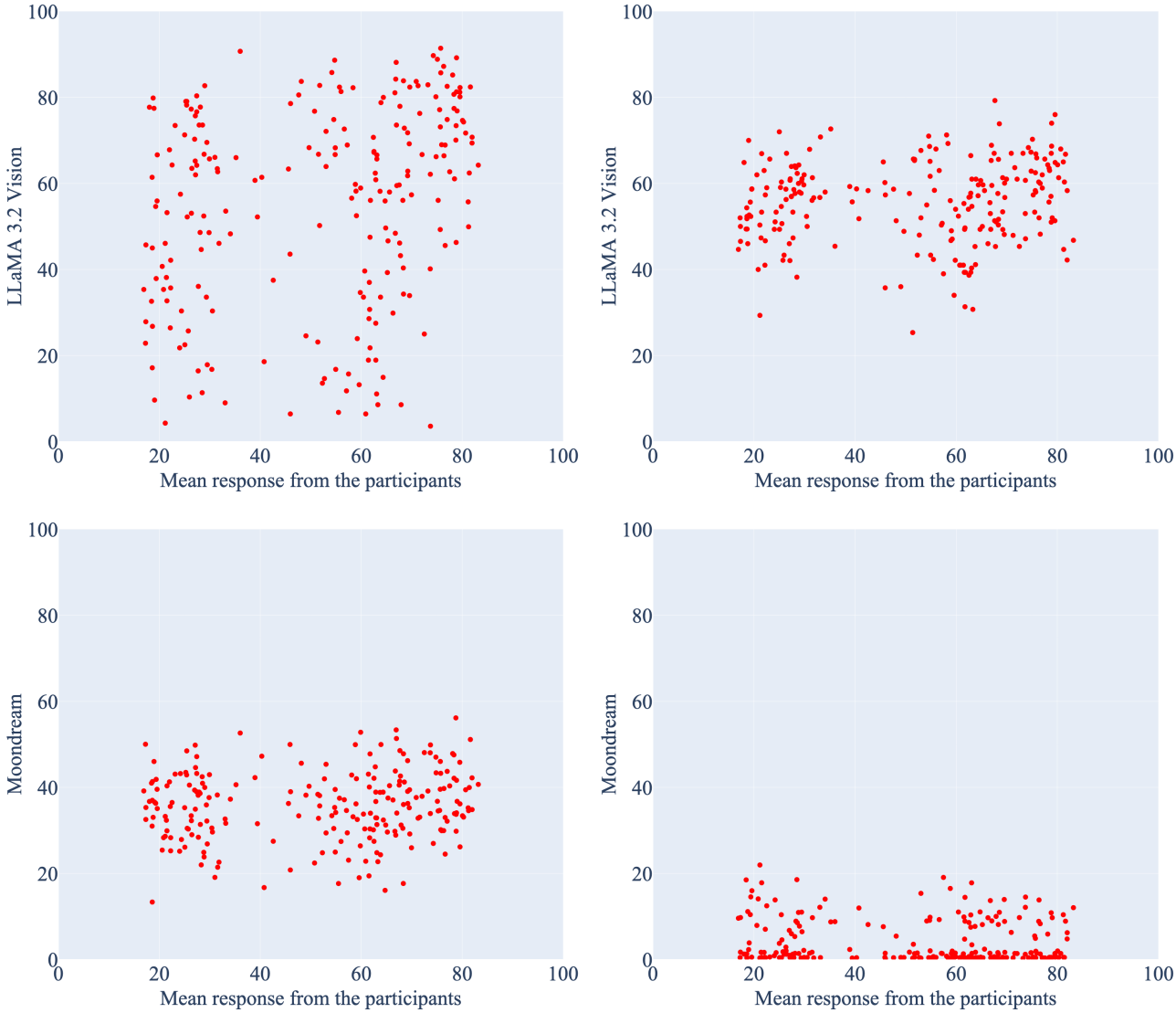

Figure A6: Scatter plots comparing model predictions to human participant responses for *LLaMA 3.2 Vision* and *Moondream*. For each model, results without memory are shown on the left (*LLaMA 3.2 Vision*: $r = 0.31$, *Moondream*: $r = 0.11$), and results with memory on the right (*LLaMA 3.2 Vision*: $r = 0.18$, *Moondream*: $r = -0.09$).

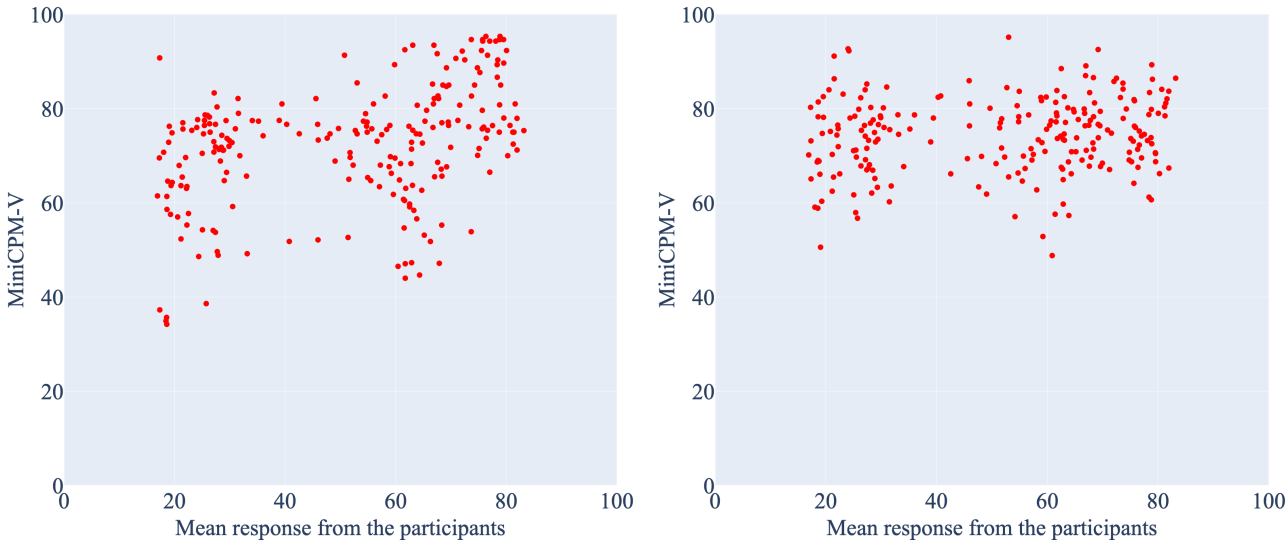

Figure A7: Scatter plots comparing *MiniCPM-V* model predictions to human participant responses. Results without memory are shown on the left ($r = 0.43$), and results with memory on the right ($r = 0.15$).