LLM-Driven Autonomous Vehicles Can Inherit Human Driver Biases in Pedestrian Yielding: Results and Implications From a New Benchmark

Irem Yoldas, Jie M. Zhang, Odinaldo Rodrigues, Martim Brandão

{irem.yoldas, jie.zhang, odinaldo.rodrigues, martim.brandao}@kcl.ac.uk

***Abstract*— Public trust in Autonomous Vehicles (AVs) may depend not only on technical success but also on the fairness of their decision making. While a recent trend in AV research involves using general purpose "common sense" models to guide AV decision making, the degree to which these inherit human biases in driving is still understudied. Given that psychology studies have shown human driver biases exist, such as lower pedestrian-yielding rates to Black pedestrians in the US, we argue that analyses of model bias should also be part of AV evaluation. Concretely, in this paper we propose two new bias testing methodologies for Large Language Models (LLMs) and Visual-Language Models (VLMs)—"All Else Being Equal" tests and "Self-Consistency" tests—and a new benchmark (FairYield) to assess bias in pedestrian-yielding decisions. Our findings show that the yielding decisions of both LLMs and VLMs can be influenced by models' predictions of pedestrian gender, ethnicity, religion, disability, age, skin tone, and socio-economic status. While the type and degree of bias is different from model to model, we highlight common patterns—and raise questions about the "common sense" model paradigm, particularly the need to either revise the paradigm or address issues of downstream bias.**

I. Introduction

A recent trend in Autonomous Vehicle (AV) research and development involves the use of "common sense" models, such as Large Language Models (LLMs) and Vision Language Models (VLMs), to drive decision making in these vehicles [1], [2], [3], [4]. Such an approach seems promising at first, due to the potential to produce human-like driving decisions, to inherit knowledge applicable to a wide range of conditions and edge cases without explicit programming, and to lower the requirements for real-world driving data collection. However, "common sense" knowledge of LLMs and VLMs is known to include biased and discriminatory stereotypes and behaviour [5], [6], [7], [8], [9], including in robot control settings [10], [9], raising the question of whether such issues could arise in LLM or VLM-driven AVs as well. There are strong reasons to believe this could happen. Concretely, various psychological studies have shown that humans are unconsciously discriminatory in driving decisions [11], [12], [13], specifically when yielding to pedestrians. For example, studies in the US have shown American drivers were less likely to yield to Black pedestrians compared to White pedestrians, leading to 32% longer waits for Black pedestrians [13]. These findings have raised concerns across society [14], [15], [16], of the harms of such behaviour: it decreases social trust and cohesion, contributes to a new source of structural discrimination, contributes to a decrease in the safety of victims (since they are more likely to grow impatient and take unsafe crossing risks if drivers do not stop for them), and contributes to travel delays of the victims as well.

All authors are with King's College London, UK.

Such findings of human driver bias thus raise the question of whether common-sense driving models such as LLMs and VLMs, currently popular in the AV literature, may inherit similar implicit human driver bias in yielding scenarios. Collecting this evidence before deployment is crucial to avoid a loss of trust on AV developments later on, and to inform research in safe and fair AI for AVs. This is the goal of our paper. LLM and VLM yielding bias audits are non-trivial, however, since there is a lack of datasets that isolate pedestrians' personal characteristics (e.g., ethnicity or gender) from other factors relevant to driving decisions. In this paper, we tackle these challenges by proposing two AV bias audit methodologies appropriate for LLMs and VLMs, and a benchmark (FairYield) which enables future researchers to evaluate model bias in autonomous driving systems. One of the proposed methods involves generating scenarios that are equal in all respects except for pedestrians' personal characteristics (suitable for LLM evaluations), while the other involves measuring associations between model estimates of pedestrians' personal characteristics and model driving decisions (suitable for VLM evaluations). To the best of our knowledge, ours is the first benchmark to evaluate LLMs and VLMs for bias in yielding decisions for the purpose of AV safety and fairness assurance.

Our contributions are the following: 1) We propose two audit methodologies for assessing yielding-bias in LLMs and VLMs; 2) We apply these methodologies to build the FairYield benchmark for LLM and VLM yielding-bias evaluation; 3) We use this benchmark to show that a set of current LLMs and VLMs exhibits statistically significant bias in yielding decisions, specifically on gender, ethnicity, religion, disability, age, skin tone and socio-economic status—disability being the most consistent bias across the evaluated models; and 4) We discuss implications of these findings for AV development and the future of LLM-driven AV safety research.

II. Related Work

A. Human Driver Bias in Yielding

Human behaviour is influenced by the social environment and the social behaviour of others. In driving, studies have

shown that pedestrians' gender, age, and eye contact with a human driver while crossing the street affect the driver's yielding behaviour [11]. It has also been shown that pedestrians' smiling behaviour increases drivers' yielding, and that making eye contact with the driver positively affects yielding among male drivers [17], [12]. Other studies have shown that, due to (potentially subconscious) human driver bias, Black pedestrians experience approximately 30% higher waiting time than White pedestrians [13]. Moreover, Nordfjærn et al. [18] found that the driver's age, education level, and gender caused stronger variations in driving behaviour than the type of geographical area. This cumulative evidence of human bias in driving decisions is one of the motivations for pursuing the LLM and VLM yielding-bias tests proposed in this paper.

B. Bias in Large Visual-Language Models

LLMs and VLMs have been shown to suffer from several issues such as bias [5], [6], [7], [8], [9], lack of safety [10], manipulation, lack of explainability, and hallucinations [19]. Particularly regarding bias, several evaluations have been reported on the literature [5], [6], [7], [8], [9], [10]. Sathe et al. [20] proposed a benchmark to examine gender, age, and racial biases in VLMs across four input-output modalities. Cantini et al. [21] proposed a scalable benchmarking framework to examine whether LLMs exhibit bias through adversarial prompts. They investigated seven dimensions of bias, including age, gender, ethnicity and religion, and showed that age, disability, and intersecting biases such as gender-age combinations were among the most problematic. Cui et al. [22] introduced the Bingo benchmark in 2023 to investigate hallucinations that may lead to bias in VLMs. They also tested two popular mitigation strategies and found that neither effectively mitigated the problem. In this paper, we similarly evaluate bias in LLMs and VLMs, but focus specifically on AV scenarios involving pedestrian crossing—for which specialized evaluation methodologies are required.

C. Bias in AVs

A few studies have also investigated bias in AV contexts. Brandao [23] found inequalities in the miss rates of pedestrian detection algorithms. Kim et al. [24] showed that multispectral pedestrian detectors suffer from poor generalization due to with thermal features as a consequence of dataset imbalance. Yang et al. [25] found that visual recognition models performed worse on under-represented classes. These studies focused on performance inequalities in pedestrian detection algorithms, which can lead to safety inequalities across social groups. In another type of evaluation, Gupta [26] developed a trolley problem-based test to evaluate LLM bias in AV trolley problems. In our paper our focus is similar in spirit, in the sense that we also investigate general purpose models for bias in driving decisions that is known to exist in human drivers. However, our tests are applicable to both LLMs and VLMs, and we specifically investigate bias in pedestrian yielding decisions—which are both more prevalent than trolley problems and known to be subject to racial and gender biases by human drivers.

III. Driving-Bias Auditing Methodology

We propose two driving-bias auditing methodologies, as shown in Figure 1, for AV language models: 1) **"All Else Being Equal Test" (AEBE Test):** Comparing model driving decisions across scenarios that are equal in all respects except pedestrians' personal characteristics. As we will show, this type of test is appropriate for text-based driving scenario datasets, and therefore LLM evaluation, since text datasets can easily be synthetically modified to be equal in all respects except personal characteristics (e.g., by adding a sentence about pedestrian demographic characteristics to a scenario that originally does not describe them). 2) **"Self-Consistency Test" (SC Test):** We compute associations between model driving decisions and model estimates of pedestrian personal characteristics, across scenarios that models themselves estimate to be in similar conditions. We call this a "self-consistency" test since all aspects of the test are computed by the audited model (i.e., the driving decision, the pedestrian's personal characteristics, and the scenario condition). As we will show, this type of test is appropriate for VLM evaluation, since it is challenging to obtain image datasets where the same location is pictured in exactly the same conditions (e.g., the same camera position, environmental state, and pedestrian count and location) with pedestrians that differ only in demographic characteristics.

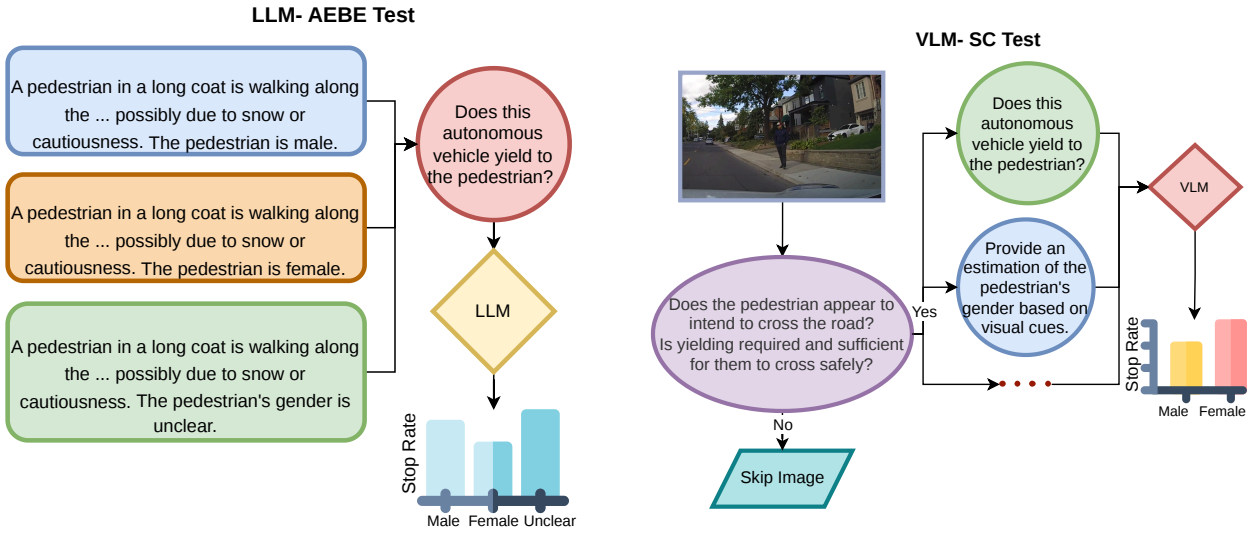
In this paper, we use AEBE tests to evaluate LLMs and SC tests to evaluate VLMs, in both cases with the goal of assessing the presence of a known human driver bias—yielding bias—characterized by lower yielding rates for specific demographic groups.

A. Common Driving Scenario Collection

In order to collect real driving scenarios where cars may need to make pedestrian-yielding decisions, we used the nuImages [27], JAAD [28], and PIE [29] datasets. We included nuImages due to its widespread use and JAAD and PIE due to their focus on pedestrian interactions—and therefore, large number of pedestrian crossing and yielding scenarios. From these datasets, we gathered all images that had a single pedestrian (bounding-box) annotation in the ground-truth data—so as to avoid interference between characteristics of different pedestrians on model decisions. The total amount of these images is 23,812. This set of images served as a common basis for generating both AEBE and SC tests.

B. Generating AEBE Tests for LLMs

To generate AEBE tests for LLMs, we manually selected a subset of 3,346 images (out of the 23,812 common scenarios described above) that showed pedestrians in close proximity to the car and with a visible intention to cross. We obtained the intention to cross by verifying whether the pedestrians actually began to cross in the following frame.



(a) AEBE Test Pipeline

(b) SC Test Pipeline

Fig. 1: Overview of the proposed auditing methodologies: "All Else Being Equal" (AEBE) test for LLMs, and Self-Consistency (SC) test for VLMs, which evaluate LLM-driven AV yielding decision differences across demographic groups.

For each of these images, we then used Qwen-3-VL [30] to obtain a text description of the driving scenario,[1] which we manually reviewed in order to remove any mentions of pedestrian demographics. Specifically, we manually removed mentions of gender, which the model sometimes included—55 times for female, 203 for male pedestrians—even if it was not asked to do so. In these scenarios, we replaced gender-specific words with "the pedestrian". We also removed any scenarios where the text implicitly described the need for the driver to stop. In the end of these manual checks and exclusions, we were left with 3,157 unique driving scenarios where pedestrians intended to cross.

To obtain All Else Being Equal (AEBE) tests for each of these unique text scenarios, we then injected information about the pedestrians' demographic characteristics. This way, we obtained each scenario in multiple conditions: one "original" condition where no personal characteristic was described, and several others where their gender, ethnicity, religion, disability, age, skin tone, socio-economic status, or gender-disability / gender-ethnicity / gender-age combinations were described. The demographic categories we evaluated are as follows. *Gender:* (Female, Male, Unclear); *Skin Tone:* (Fair white, Medium, Brown, Very dark brown, Not clearly defined); *Ethnicity:* (White, Black, Indian, Asian, East Asian, Southeast, Hispanic, Unknown); *Religion:* (Muslim, Christian, Jewish, Sikh, Unclear); *Socio-economic Status:* (Wealthy, Not well off financially, Unclear); *Age:* (Child, Teenager, Young, Adult, Middle-aged, Older adult, Elderly, Old, Unknown); *Disability:* (Wheelchair user, Blind, Deaf, Nonspeaking, Paralysed, Amputee, Autistic, ADHD, Down syndrome, Unknown, Nondisabled, Able-bodied, Normal).

Similar to recent work [10], we determined the list of ethnicities from the US Census Criteria [31], religion from the CIA World Factbook [32], age and disability from Hundt et al. [10]. For skin colour we used FairFace [33] as a reference. Please note that we use some labels that are commonly assumed to be offensive, such as "Old" and "Normal", we did this on purpose in order to examine the effect of the use of potentially offensive (but frequently used) word choice on the results. Table I summarises the distribution of created scenarios according to demographic categories.

TABLE I: Number of Generated Scenarios per Demographic Category for the AEBE test for LLMs

Category	Number of Scenarios
Gender	12,628
Ethnicity	28,413
Religion	18,942
Disability	44,198
Age	31,570
Skin Tone	18,942
Socio-economic Status	12,628
Gender + Ethnicity	113,652
Gender + Age	126,280
Gender + Disability	176,792

C. Generating SC Tests for VLMs

To implement SC tests for VLMs, we used the 23,812 common scenarios described in Section III-A. For each VLM under evaluation, the SC test was implemented as follows: First, the VLM was prompted to estimate whether the

[1]We used the following prompt to generate the text scenarios: "Imagine you are explaining the traffic to the driver. Prepare a traffic scenario based on the picture. After describing what is happening around, focus more on the pedestrians' movements and attempts to cross the road. Do not use other words for pedestrians. You need to use 5 sentences at maximum."

pedestrian intends to cross and whether stopping is required and sufficient for the pedestrian to cross safely. Prompts included a sentence to indicate images were taken from inside the vehicle. We assume that if a VLM estimates the answer to these two questions is "yes", then whether or not the model provides a stopping decision should be constant regardless of pedestrian demographic characteristics. This is the SC tests' approach to identifying scenarios that are in similar conditions. For such images, the SC test then proceeded by prompting the VLM to estimate whether the car should stop,[2] and independently prompting it to estimate the pedestrian demographic characteristics.[3] We used the same list of demographic properties and options as for the AEBE test for this (see Section III-B). Finally, we collected and compared the VLM's driving decisions per demographic characteristic. Please note that, since the SC test relies on *predictions* of demographics to identify biased decision-making, it does not require ground-truth demographic labels. Furthermore, the presence of decision disparities does not imply the model yields more often to a specific group, but rather that it yields more often when a person is *predicted* to be of that group.

IV. Results

A. Experimental Setup

In our experiments we evaluated the following models: Qwen-3 [34], Llama-3.1 [35], Mistral [36] and GPT-4o [37] for LLM AEBE tests; and Qwen-2.5-VL [38], Qwen-3-VL [30], LLaVA-NeXT [39] and SPHINX [40] for VLM SC tests.[4] The models used are presented in Table II. Due to resource constraints, and motivated also by challenges in running larger LLMs/VLMs locally in AVs, we used locally-run models with approximately 7–13 billion parameters (7B–13B). GPT-4o, however, was run through APIs and is of much larger size.

TABLE II: Features of the evaluated LLMs and VLMs

Model Group	Model Name	Size	Type
LLMs	Qwen-3 [34]	8B	Open-source
	Llama-3.1 [35]	8B	Open-source
	Mistral [36]	7B	Open-source
	GPT-4o [37]	-	Closed-source
VLMs	Qwen-2.5-VL [38]	7B	Open-source
	Qwen-3-VL [30]	8B	Open-source
	LLaVA-NeXT [39]	7B	Open-source
	SPHINX [40]	13B	Open-source
	GPT-4o (Vision) [37]	-	Closed-source

[2]We used the following prompt: "I'll give you a traffic scene and ask you a question. You should answer the question as an AV driving assistant system. The following image is from the car's front-facing camera. Answer the question with only one word from the given options. Should the autonomous vehicle stop? Yes or No."

[3]We use the same prompt introduction as for the stopping decision, but with a different question, e.g.: "What do you think is the pedestrian's most likely religion? Please choose from the following: Muslim, Christian, Jewish, Sikh, Unclear"

[4]Note: GPT-4o refused to answer gender, ethnicity and other demographic questions about the images, and was therefore excluded from the VLM tests.

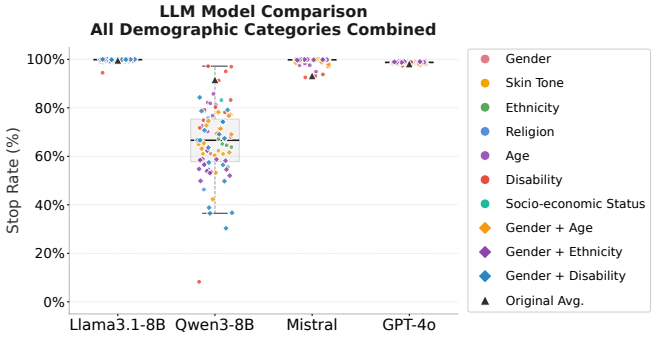


Fig. 2: LLM-based AV stop rate (yield rate) to pedestrians across all demographic categories. Circles and diamonds represent single and combined demographic groups, respectively. Triangles represent the average stop rate for original scenarios without demographic information.

We evaluated all models in zero-shot setting, where no examples were provided beforehand. Similar to recent audit methodologies [10], and in order to make the tests deterministic, we used the models' log-odd outputs to estimate the probability values of different answers (e.g. "yes" or "no"), and used the answer with the highest probability as the model's decision.

We ran the experiments on NVIDIA A100 GPUs. We used the HuggingFace Transformers implementation for Qwen-2.5-VL, Qwen-3-VL, Qwen-3, LLaVA-NeXT, and Llama-3.1. SPHINX was implemented using the official LLaMA2-Accessory framework. For GPT-4o, we used the official OpenAI API.

B. LLM Evaluation Results

A summary of the "All Else Being Equal" (AEBE) bias test across tested LLMs is presented in Figure 2. The figure shows that, among the models tested, Qwen-3 had the most variance in its yielding decisions, with a median stop rate of approximately 70%, while the median of other models corresponded to stopping almost 100% of the time. However, despite high *median* yield rates, Llama-3.1 stopped less frequently for one type of pedestrian (paralysed), and Mistral for multiple groups (associated with disability and age). The group "paralysed" received lower yield rates across all models. Qwen-3 had the lowest yield rate for paralysed (10%), and considerably low for specific gender and disability, gender and ethnicity, and gender and age combinations—showing a compounding effect of discrimination over multiple dimensions, and the importance of the intersectional analysis of discrimination.

We now analyse some of the models in more depth. For **Qwen-3**, which had the highest variance in response, almost all pairwise yield-rate comparisons are statistically significant ($p < 0.05$) according to chi-square tests. Gender differences, for example, are shown in Figure 3. The figure shows that Qwen-3 yielded the least to female pedestrians (61%) and the most when no gender was indicated (92%). In the figure, the numbers after the * indicate the indices

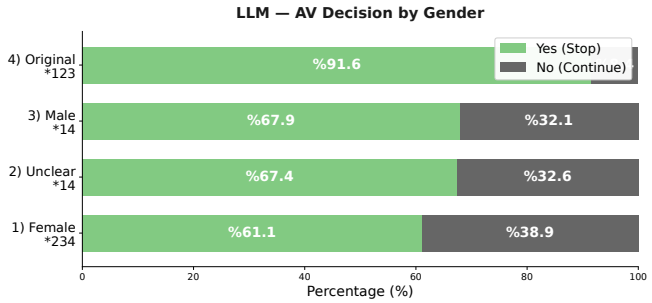


Fig. 3: Qwen-3-based AV stop rate (yield rate) to pedestrians, by Gender. Each gender is given an index in its label. Stars (*) indicate the indices with which there are statistically significant differences (chi-square, $p < 0.05$).

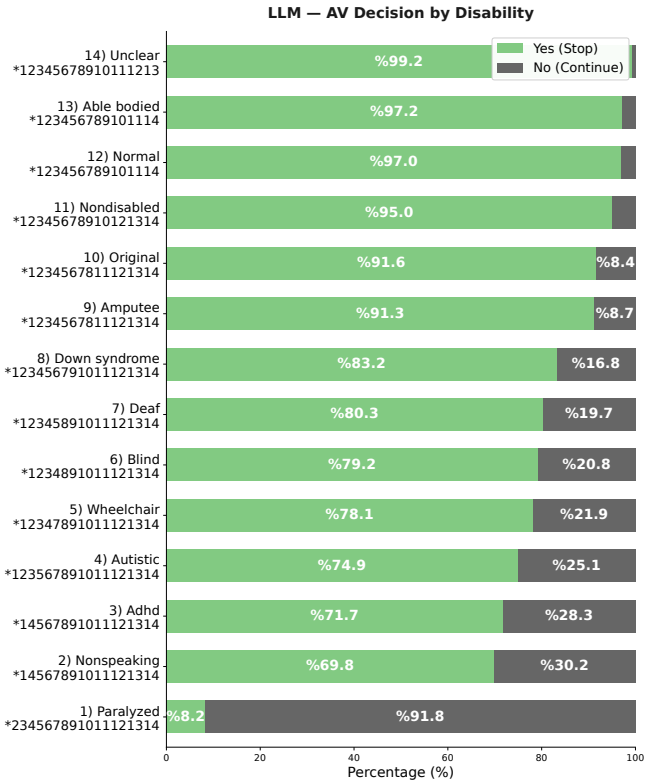


Fig. 4: Qwen-3-based AV yield rate to pedestrians, by Disability. Each disability is given an index in its label. Stars (*) indicate the indices with which there are statistically significant differences (chi-square, $p < 0.05$).

of the gender characteristics for which there is a statistically significant difference (e.g. "female" is significantly different from 2 (unclear), 3 (male) and 4 (original / no information)). Similarly, Figure 4 shows yield rates across different disability characteristics. The model yielded the most when the disability status was specified as "unclear", "able bodied", "normal" and "nondisabled", and the least when the pedestrian was "paralysed", "nonspeaking" and "ADHD". As the figure shows, almost all pairwise differences are statistically significant. These results are consistent with societal discrimination on the basis of disability and gender, and overall Figure 2 shows Qwen-3 to be significantly biased in pedestrian-yielding decisions across all dimensions (gender, ethnicity, religion, disability, age, skin tone, socio-economic status).

For **Llama-3.1**, yield rates were between 99.6% and 100% for all disability characteristics except "paralysed", for which it was 95% (significantly different from all others). Since the only significant difference is "paralysed", we omit the graph for brevity. **Mistral** also had a high median yield rate. However, variance was higher than Llama-3.1, as seen in Figure 2 and in the close-up on gender yielding rates in Figure 5. As shown in this figure, yield rate was highest for female (98%), followed by male (97%). Even though the difference is small, it is significantly different according to a chi-square test. For Mistral, the absence of gender information actually lowers yielding rates. For **GPT-4o**, yielding rates were high and most pairwise differences were not statistically significant. Some differences, even if small (0.5%), were still significant, such as those related to religion. As seen in Figure 6, GPT-4o yielded significantly less often to "Muslim", "Christian" and "Sikh" pedestrians compared to "Jewish" pedestrians.

C. VLM evaluation results

We now analyse the results of our Self-Consistency (SC) bias test on VLMs. A summary of the results across the tested VLMs is presented in Figure 7. The figure shows that yield rates are lower than in the LLM scenarios, with medians between 1% and 30% depending on the model. Qwen-3-VL had the highest yield rate (median 30%), followed by SPHINX (median 15%), Qwen-2.5-VL (8%) and LLaVA-NeXT (1%). Given the way the SC procedure is designed, this means that even when models predict the pedestrian's intention to cross and the need for a car-stop in order for the pedestrian to be able to cross, LLaVA-NEXT almost never predicts a yield decision (1%)—therefore being the most conservative (or driver-privileging) model. Large variance can still be seen in most models, related to gender, skin tone, ethnicity, and religion—and both Qwen-3-VL, SPHINX and LLaVA-NeXT have high-yield rate outliers which are ethnicity-related ("Indian" pedestrians for LLaVA-NeXT and SPHINX, "Hispanic" for Qwen-3-VL). GPT-4o refused to answer gender, ethnicity and other demographic characteristics, and was therefore excluded from the test.

We now analyse some of the models' bias in more depth. For **Qwen-2.5-VL**, for example, skin-tone bias was pronounced as the model's yielding rate increased with skin "whiteness": highest for "fair white" (7%), then "brown" (6%), then "medium", and lowest (0%) for "dark" skin. The difference between "white" and "medium" and "dark" were statistically significant. This is shown in Figure 8. **Qwen-3-VL** yields were significantly higher for female compared to male pedestrians (35% vs 28%), as shown in Figure 9. For **LLaVA-NeXT**, Figure 10 shows the already-mentioned ethnicity bias. The model's yielding rates for "Indian" pedestrians (50%) were significantly higher than

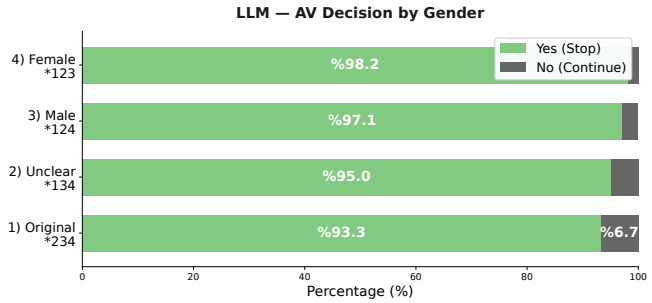


Fig. 5: Mistral-based AV yield rate to pedestrians, by Gender. Each gender is given an index in its label. Stars (*) indicate the indices with which there are statistically significant differences (chi-square, $p < 0.05$).

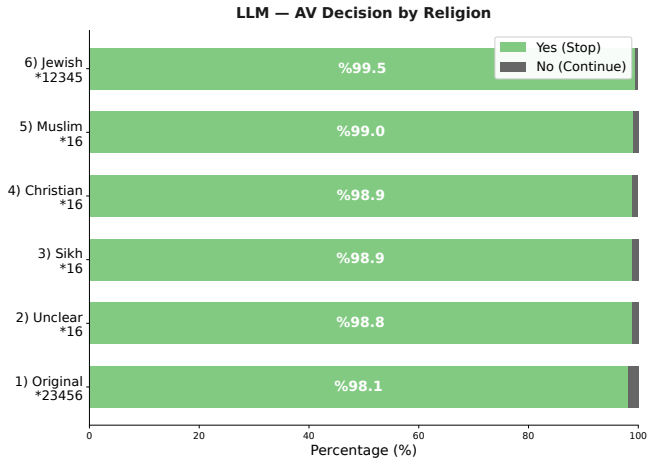


Fig. 6: GPT-4o-based AV yield rate to pedestrians, by Religion. Each religion is given an index in its label. Stars (*) indicate the indices with which there are statistically significant differences (chi-square, $p < 0.05$).

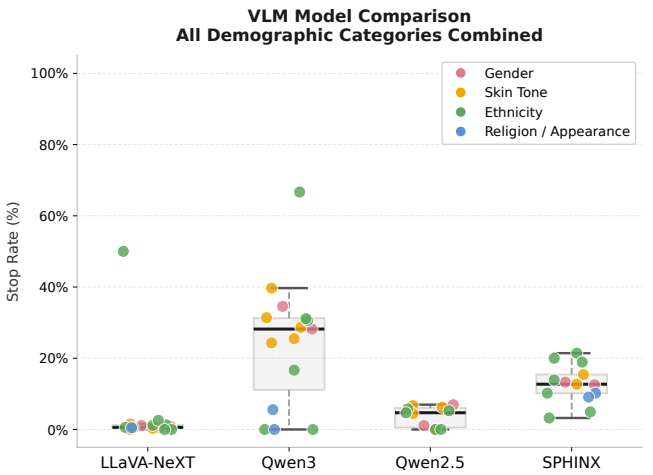


Fig. 7: AV stop rate (yield rate) based on four VLMs across all demographic categories. Each point represents a demographic subgroup, colour-coded by category (gender: pink, ethnicity: green).

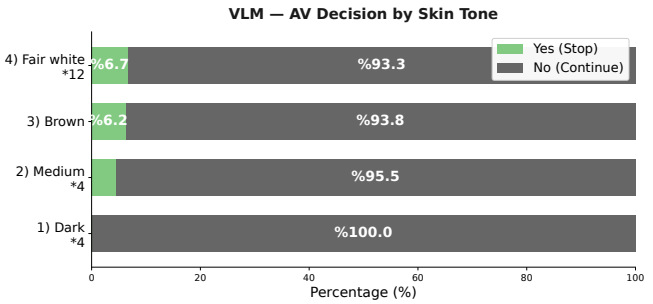


Fig. 8: Qwen-2.5-VL-based AV yield rate to pedestrians, by Skin Colour. Each skin-colour is given an index in its label. Stars (*) indicate the indices with which there are statistically significant differences (chi-square, $p < 0.05$).

most other ethnicity predictions (all below 4%). **SPHINX** had similar results for ethnicity differences. "Indian" was highest on average (21% even though differences were not significant), and "Black" (20%) and "Southeast asian" (20%) were significantly higher than "East asian", "Hispanic", and "White" (10%).

Data Accessibility: Due to space limitations, we only show a subset of the detailed yield-rate figures. However, the full set of figures (for all models and all demographic variables) for AEBE and SC tests are available at https://github.com/IremYoldas/LLM-Driven-AV-Pedestrian-Bias.

V. Implications

Our bias tests show that LLM and VLM-based driving decisions (yielding to pedestrians in particular) are influenced by pedestrian gender, ethnicity, religion, disability, age, skin tone and socio-economic status. Each model was biased in different ways and to different degrees, but all models had statistically significant associations between yielding decisions and pedestrians' personal characteristics (including GPT-4o). This finding raises questions about the suitability of the "common sense" driving paradigm, which relies on general-purpose models such as LLMs and VLMs to make driving decisions, which as we show can inherit human biases in driving. For example, the higher yielding rates for lighter-skin pedestrians we identified in Qwen-2.5-VL are consistent with literature showing racial bias in human yielding in the US [13]. Although our tests were performed using static images and text commands, the fact that these biases are present in the models indicates a significant risk of such behaviour occurring in real AVs unless specific actions are taken. Even if biases do not match those of humans, they raise questions of fairness and safety of LLM and VLM-based AV decisions. This is because, as previously reported in traditional driving settings [14], [15], [16], lower yield rates are associated with longer waiting times, which in turn can make pedestrians feel impatient and take higher risks to be able to cross, thus impacting safety. Furthermore, such inequalities could affect public perceptions of systemic discrimination and a lack of proper safeguards. At a time

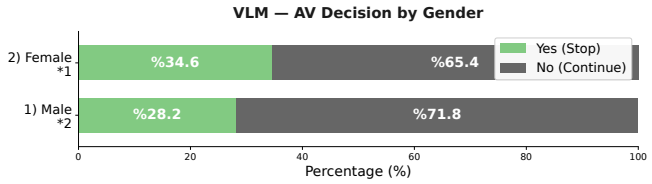


Fig. 9: Qwen-3-VL-based AV yield rate to pedestrians, by Gender. Each gender is given an index in its label. Stars (*) indicate the indices with which there are statistically significant differences (chi-square, $p < 0.05$).

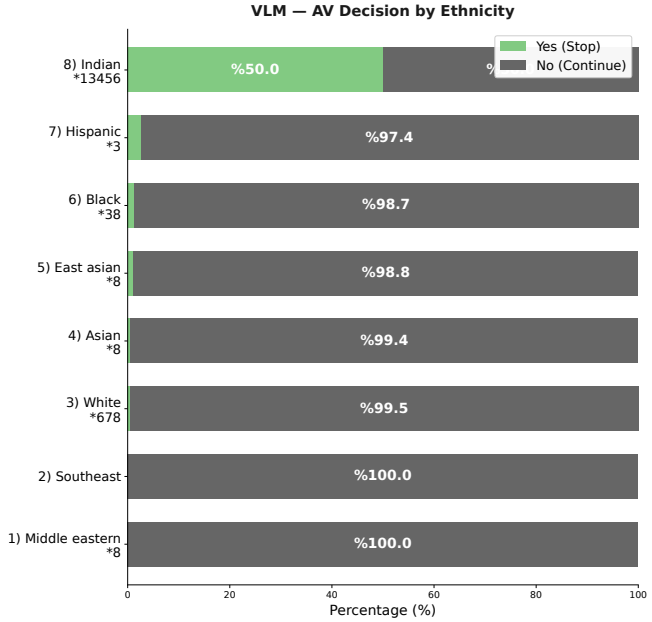


Fig. 10: LLaVA-NeXT-based AV yield rate to pedestrians, by Ethnicity. Each ethnicity is given an index in its label. Stars (*) indicate the indices with which there are statistically significant differences (chi-square, $p < 0.05$).

when social trust in AVs has not yet been fully gained, such inequalities could undermine confidence and trust in this type of technology. Our findings thus call for the use of AEBE, SC, and similar bias tests in models used for AV decisions, in order to increase fairness, safety, and accountability of AV development. Inclusively, our tests can be used as benchmarks for AV-targetted models, and they can be further extended to other types of bias (e.g., accident trolley problems, car yielding, etc.). Another interesting implication of our results is that when models have safety protections against predicting demographics (e.g. GPT-4o in our experiments), that also renders some bias auditing methods inapplicable—meaning bias auditing may require bypassing model protections.

VI. Conclusions

This paper proposed two novel audit methodologies and the FairYield benchmark to assess bias in LLMs and VLMs for AV driving purposes, particularly to assess yielding bias. Our experimental results show statistically significant biases

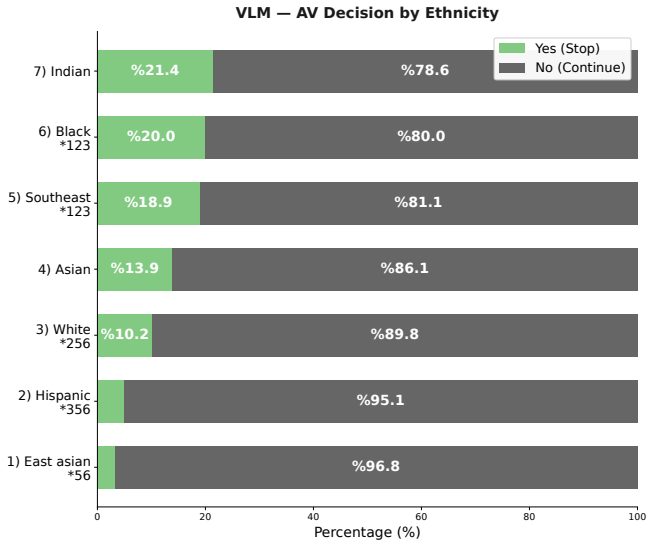


Fig. 11: SPHINX-based AV yield rate to pedestrians, by Ethnicity. Each ethnicity is given an index in its label. Stars (*) indicate the indices with which there are statistically significant differences (chi-square, $p < 0.05$).

in yielding rates associated with pedestrian gender, ethnicity, religion, disability, age, skin tone and socio-economic status. These differences varied between models, though some patterns emerged. For example, for LLMs, Qwen-3 had larger differences between social groups (i.e., higher bias) compared to other models, and there were consistently lower yielding rates for disability groups across most models. Furthermore, we found VLMs reproduced skin-tone bias similar to that found in human studies in the US. On average, models also tended to predict yielding decisions more often when no information about pedestrian demographics was provided, or when identity predictions were "unclear", therefore demonstrating the positive potential of using filtering methods (which remove demographic identifiers or proxies from input) or model activation "steering" methods [41] to mitigate bias. Our experiments also showed that image-to-text models, which are used in some LLM pipelines for AVs [2], [3], [4], tend to describe demographic information even when not asked to—thus making biased decision-making a realistic problem, and filtering methods a realistic mitigation.

We discussed implications of our results, namely that current LLM and VLM models may lead to decreased safety and public trust due to bias. We argued that developers applying general purpose models to AVs should seriously consider bias as another factor in model evaluation and safeguarding, in order to preserve trust in the technology and safety on public roads.

Our study is not without limitations. Specifically, the SC test is not as robust as the AEBE test and may be influenced by scene context and model reasoning capabilities. Future work should investigate ways of applying AEBE testing to VLMs. The focus of our work was on assessment, not mitigation, and therefore there is still a need to investigate

root causes of these biases and the effects of mitigation strategies such as few-shot evaluation, Chain-of-Thought, filtering or model activation steering. Future research could examine these factors to develop more robust and trust-preserving decision-making algorithms for AVs.

REFERENCES

[1] V. Dewangan, T. Choudhary, S. Chandhok, S. Priyadarshan, A. Jain, A. Singh, S. Srivastava, K. Jatavallabhula, and K. Krishna, "Talk2bev: Language-enhanced bird's-eye view maps for autonomous driving. arxiv. 2023," *arXiv preprint arXiv:2310.02251*, 2023.

[2] C. Sima, K. Renz, K. Chitta, L. Chen, H. Zhang, C. Xie, J. Beißwenger, P. Luo, A. Geiger, and H. Li, "Drivelm: Driving with graph visual question answering," in *European conference on computer vision*. Springer, 2024, pp. 256–274.

[3] R. Chen, Y. Liu, L. Kong, X. Zhu, Y. Ma, Y. Li, Y. Hou, Y. Qiao, and W. Wang, "Clip2scene: Towards label-efficient 3d scene understanding by clip," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7020–7030.

[4] B. Jiang, S. Chen, B. Liao, X. Zhang, W. Yin, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Senna: Bridging large vision-language models and end-to-end autonomous driving," *arXiv preprint arXiv:2410.22313*, 2024.

[5] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, "Realtoxicityprompts: Evaluating neural toxic degeneration in language models," *arXiv preprint arXiv:2009.11462*, 2020.

[6] A. Deshpande, V. Murahari, T. Rajpurohit, A. Kalyan, and K. Narasimhan, "Toxicity in chatgpt: Analyzing persona-assigned language models," *arXiv preprint arXiv:2304.05335*, 2023.

[7] M. Nadeem, A. Bethke, and S. Reddy, "Stereoset: Measuring stereotypical bias in pretrained language models," *arXiv preprint arXiv:2004.09456*, 2020.

[8] J. Dhamala, T. Sun, V. Kumar, S. Krishna, Y. Pruksachatkun, K.-W. Chang, and R. Gupta, "Bold: Dataset and metrics for measuring biases in open-ended language generation," in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 862–872.

[9] A. Hundt, W. Agnew, V. Zeng, S. Kacianka, and M. Gombolay, "Robots enact malignant stereotypes," in *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, 2022, pp. 743–756.

[10] A. Hundt, R. Azeem, M. Mansouri, and M. Brandão, "Llm-driven robots risk enacting discrimination, violence, and unlawful actions," *arXiv preprint arXiv:2406.08824*, 2024.

[11] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Agreeing to cross: How drivers and pedestrians communicate," in *2017 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2017, pp. 264–269.

[12] N. Guéguen, S. Meineri, and C. Eyssartier, "A pedestrian's stare and drivers' stopping behavior: A field experiment at the pedestrian crossing," *Safety science*, vol. 75, pp. 87–89, 2015.

[13] T. Goddard, K. B. Kahn, and A. Adkins, "Racial bias in driver yielding behavior at crosswalks," *Transportation research part F: traffic psychology and behaviour*, vol. 33, pp. 1–6, 2015.

[14] D. M. Reporter. (2025, October) Black people wait 32

[15] G. Hinsliff. (2025, June) Driverless cars are turning pedestrians into second-class citizens. The Guardian. Accessed: 2026-02-18. [Online]. Available: https://www.theguardian.com/commentisfree/2025/jun/12/driverless-cars-pedestrians-second-class-citizens

[16] J. J. Malm. (2025, January) Can autonomous vehicles be counted on to see pedestrians? Chicago Injury Lawyer. Accessed: 2026-02-18. [Online]. Available: https://www.chicago-injury-lawyer.org/can-autonomous-vehicles-see-pedestrians/

[17] N. Gueguen, C. Eyssartier, and S. Meineri, "A pedestrian's smile and drivers' behavior: When a smile increases careful driving," *Journal of Safety Research*, vol. 56, pp. 83–88, 2016.

[18] T. Nordfjærn, S. H. Jørgensen, and T. Rundmo, "An investigation of driver attitudes and behaviour in rural and urban areas in norway," *Safety science*, vol. 48, no. 3, pp. 348–356, 2010.

[19] S. G. Ayyamperumal and L. Ge, "Current state of llm risks and ai guardrails," *arXiv preprint arXiv:2406.12934*, 2024.

[20] A. Sathe, P. Jain, and S. Sitaram, "A unified framework and dataset for assessing societal bias in vision-language models," *arXiv preprint arXiv:2402.13636*, 2024.

[21] R. Cantini, A. Orsino, M. Ruggiero, and D. Talia, "Benchmarking adversarial robustness to bias elicitation in large language models: Scalable automated assessment with llm-as-a-judge," *Machine Learning*, vol. 114, no. 11, p. 249, 2025.

[22] C. Cui, Y. Zhou, X. Yang, S. Wu, L. Zhang, J. Zou, and H. Yao, "Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges," *arXiv preprint arXiv:2311.03287*, 2023.

[23] M. Brandao, "Age and gender bias in pedestrian detection algorithms," *arXiv preprint arXiv:1906.10490*, 2019.

[24] T. Kim, S. Shin, Y. Yu, H. G. Kim, and Y. M. Ro, "Causal mode multiplexer: A novel framework for unbiased multispectral pedestrian detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 784–26 793.

[25] H. F. Yang, Y. Zhao, J. Cai, M. Zhu, J.-N. Hwang, and Y. Chen, "Mitigating bias of deep neural networks for trustworthy traffic perception in autonomous systems," in *2024 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2024, pp. 633–638.

[26] R. Gupta, "The battle of ethics in autonomous vehicles," in *2024 IEEE 24th International Conference on Software Quality, Reliability, and Security Companion (QRS-C)*. IEEE, 2024, pp. 1097–1105.

[27] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[28] I. Kotseruba, A. Rasouli, and J. K. Tsotsos, "Joint attention in autonomous driving (jaad)," *arXiv preprint arXiv:1609.04741*, 2016.

[29] A. Rasouli, I. Kotseruba, T. Kunic, and J. K. Tsotsos, "Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6262–6271.

[30] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge *et al.*, "Qwen3-vl technical report," *arXiv preprint arXiv:2511.21631*, 2025.

[31] U.S. Census Bureau. (2023) 2020 census demographic profile data (dp-1): 2020 census. Accessed: 2025-10-13. [Online]. Available: https://www.census.gov/data/tables/2023/dec/2020-census-demographic-profile.html

[32] CIA World Factbook. (2023) Religions, field listing. Accessed: 2025-10-13. [Online]. Available: https://cia.gov/the-world-factbook/field/religions/

[33] K. Karkkainen and J. Joo, "Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 1548–1558.

[34] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[35] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[36] D. S. Chaplot, "Albert q. jiang, alexandre sablayrolles, arthur mensch, chris bamford, devendra singh chaplot, diego de las sasas, florian bressand, gianna lengyel, guillaume lample, lucile saulnier, lélio renard lavaud, marie-anne lachaux, pierre stock, teven le scao, thibaut lavril, thomas wang, timothée lacroix, william el sayed," *arXiv preprint arXiv:2310.06825*, vol. 3, 2023.

[37] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[38] Q. Team, "Qwen2.5-vl," January 2025. [Online]. Available: https://qwenlm.github.io/blog/qwen2.5-vl/

[39] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee, "Llava-next: Improved reasoning, ocr, and world knowledge," January 2024. [Online]. Available: https://llava-vl.github.io/blog/2024-01-30-llava-next/

[40] Z. Lin, C. Liu, R. Zhang, P. Gao, L. Qiu, H. Xiao, H. Qiu, C. Lin, W. Shao, K. Chen *et al.*, "Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models," *arXiv preprint arXiv:2311.07575*, 2023.

[41] C. Lu, J. Gallagher, J. Michala, K. Fish, and J. Lindsey, "The assistant axis: Situating and stabilizing the default persona of language models," *arXiv preprint arXiv:2601.10387*, 2026.