2025 IEEE International Conference on Robotics and Automation (ICRA)
May 19-23, 2025. Atlanta, USA

# A Comprehensive LLM-powered Framework for Driving Intelligence Evaluation

Shanhe You, Xuewen Luo, Xinhe Liang, Jiashu Yu, Chen Zheng and Jiangtao Gong✉

***Abstract*— Evaluation methods for autonomous driving are crucial for algorithm optimization. However, due to the complexity of driving intelligence, there is currently no comprehensive evaluation method for the level of autonomous driving intelligence. In this paper, we propose an evaluation framework for driving behavior intelligence in complex traffic environments, aiming to fill this gap. We constructed a natural language evaluation dataset of human professional drivers and passengers through naturalistic driving experiments and post-driving behavior evaluation interviews. Based on this dataset, we developed an LLM-powered driving evaluation framework. The effectiveness of this framework was validated through simulated experiments in the CARLA urban traffic simulator and further corroborated by human assessment. Our research provides valuable insights for evaluating and designing more intelligent, human-like autonomous driving agents. The implementation details of the framework[1] and detailed information about the dataset[2] can be found at the provided links.**

## I. INTRODUCTION

The field of autonomous driving evaluation has evolved significantly, transitioning from traditional, single-layer mechanical metrics to more comprehensive, multi-dimensional evaluation frameworks [1]–[3]. Early approaches relied heavily on mathematical models to simulate driving scenarios, but these have now expanded to incorporate AI-annotated machine learning techniques [4]–[6]. Despite these advances, current evaluation metrics remain rigid, lacking the flexibility necessary to adapt to the complexities of real-world driving [7]. A critical shortcoming of these traditional methods is their inability to effectively assess the intelligence of autonomous systems, as they fail to align with the cognitive processes humans use when evaluating driving performance.

Large language models (LLMs) have demonstrated superior capabilities in complex reasoning and knowledge application [8], opening new possibilities in autonomous driving for path planning [9]–[12] and decision interpretability [13]–[15]. LLMs offer potential solutions to current evaluation limitations by mirroring human thought processes [16]. Effective evaluation requires careful selection of key factors as prompts [17], incorporating both fundamental driving elements and higher-level reasoning. Besides, LLM-as-a-judge approaches [18] have shown promise in generating neutral, interpretable evaluations across various domains [16], [19], with recent advancements in unsupervised evaluation [20], [21] and domain knowledge integration [17], [22], [23]. However, LLMs still lack sufficient understanding of complex, domain-specific knowledge, limiting their ability to fully evaluate critical driving information.

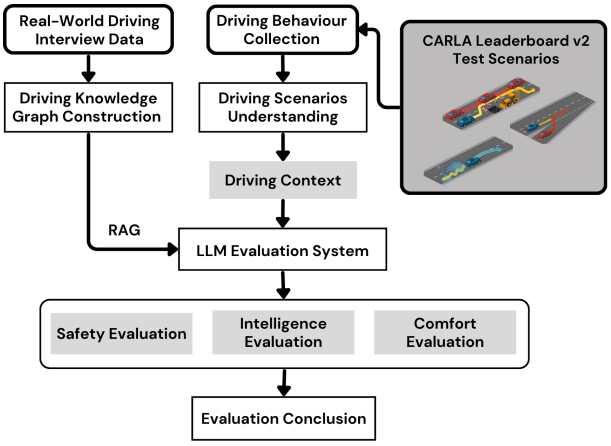



Fig. 1. A comprehensive framework for evaluating driving intelligence using LLMs. We use real-world driving interview data to construct a driving evaluation knowledge graph, which, together with driving behaviors collected from the simulator, generates driving context. Then, it generates assessments on safety, intelligence, and comfort, leading to a comprehensive evaluation of the driving performance.

To this end, we propose a novel LLM-based evaluation framework to assess driving behavior intelligence in complex traffic environments, addressing the gap in comprehensive evaluation methods. As shown in Fig. 1, we developed this framework by first constructing a natural language evaluation dataset from human professional drivers and passengers through naturalistic driving experiments and post-driving behavior evaluation interviews. We then developed an LLM-powered driving evaluation framework. To ensure its effectiveness, we validated the framework through simulated experiments in the CARLA urban traffic simulator and further corroborated the results with human assessment. The contributions of this paper can be summarized as follows:

1) We collected and constructed a driving evaluation dataset through naturalistic driving experiments and post-experiment interviews with drivers and passengers.
2) Based on in-depth analysis of the driving evaluation data, we developed a comprehensive evaluation framework for assessing autonomous driving behaviors.
3) We implemented this framework using LLM and validated it through human assessment with data collected in the driving simulator.

The authors are with the Institute for AI Industry Research, Tsinghua University, Beijing, China. Corresponding Email: gongjiangtao@air.tsinghua.edu.cn

[1]https://github.com/AIR-DISCOVER/Driving-Intellenge-Evaluation-Framework

[2]https://github.com/AIR-DISCOVER/Driving-Evaluation-Dataset

2025 IEEE International Conference on Robotics and Automation (ICRA) | 979-8-3315-4139-2/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICRA55743.2025.11128380



Authorized licensed use limited to: Tsinghua University. Downloaded on April 01,2026 at 14:01:02 UTC from IEEE Xplore. Restrictions apply.

## II. Related Work

**Autonomous Driving Evaluation:** Evaluation metrics in autonomous driving studies encompass basic [7], [24], etiquette [3], comfortableness [1], [3], [6], [25], [26], risk [2], [5], [27], and trust [28]–[31] categories. Kim et al. [29] investigated the relationship between driving scenarios and takeover events, defining DRI for safety evaluation. Kühnlenz et al. [31] proposed that users' perceptions of system intelligence are influenced by prior information and continuous self-explanations, enhancing transparency and trust. However, intelligent assessment based on specific driving scenarios remains at the analytical level, and human-like evaluation methods reflecting driving philosophy are needed. The field lacks comprehensive methods that integrate scenario-specific assessments with human-like decision-making processes.

**LLM in Driving:** LLMs have shown promise in driving by integrating common-sense knowledge [11] with machine behavior [9]. They combine objective domain knowledge (e.g., traffic rules [32]) with subjective human-generated data [14], [15], [33], [34], enabling nuanced understanding of human-like driving behavior. LLMs effectively manage motion planning tasks [9], [11], [35]–[37] through techniques like temporally causal decoders [9]. The prompting-reasoning-finetuning strategy [10], [34] enhances decision-making capabilities. LLMs offer explainability [13] by providing transparent reasoning for driving decisions, promoting human-aligned [14], [38]–[41], knowledge-driven approaches [38], [42]. However, they lack embodied intelligence, focusing on general knowledge rather than specific driving expertise, which limits their understanding of complex driving scenarios.

**LLM-as-a-judge:** LLMs generate neutral, interpretable, and scalable evaluations across diverse domains [16], [18], [19]. Recent research has employed advanced methodologies like META-REWARDING [20], [21] for unsupervised evaluation, enabling LLMs to autonomously refine their judgment capabilities. LLMs can incorporate domain knowledge [22] and assess corner cases [17] or driving skills [23]. They provide QA-based driving behavior explanations [12], [15], [33], [43]–[46], enhancing interpretability. However, LLMs often lack sufficient understanding of complex, domain-specific knowledge, especially in automated driving. This deficit hampers their ability to fully and accurately evaluate critical driving information, particularly in nuanced or rapidly changing scenarios.

## III. Driving-Evaluation Dataset

### A. Driving Experiment and Interview Data Acquisition

To collect real-world driving evaluation data from humans, we designed and implemented real-world driving experiments. In the experiment, we conducted in-depth interviews with every driver and passenger to understand their evaluations of this driving performance comprehensively. A total of 24 drivers and 48 passengers were invited to participate in this experiment. These participants were divided into expert drivers, novice drivers, and passengers. Expert drivers (10 males, aged 35-48) had 12-28 years of experience and annual mileage ranging from under 20,000 km to over 100,000 km. Novice drivers (11 males, 3 females, aged 20-25) had 1-4 years of experience with mileage up to 40,000 km. The passenger group (48 people, aged 20-50) included 32 licensed individuals with limited driving experience and 16 without a license.

We first arranged for each driver to drive an urban road and two passengers to ride this vehicle, which covered 13 different driving conditions and a total distance of 5.7 kilometers. We installed a 360-degree panoramic camera on top of the vehicle to record the vehicle's surroundings in real time and captured the driver's behavior through multiple motion cameras inside the vehicle, which enabled us to accurately review the entire driving maneuvers and performance in the follow-up interviews. In addition, we recorded the vehicle's real-time CAN bus data to capture the vehicle's driving condition, including throttle, brakes, steers, etc.

During the interviews, we used the collected videos to recreate the driving process that the drivers and passengers had just experienced. For each driving context and action, participants were asked to verbalize their thought processes during the interview (while watching the video) and to consider how they evaluate driving behavior. Drivers shared their driving preferences and assessed whether driving behaviors were in alignment with their daily habits by looking back at the videos and reflecting on the logic behind the decisions and actions made while driving and the reasons for the differences. Passengers, for their part, reviewed comfortable and uncomfortable driving segments and expressed their feelings and riding habits under different driving conditions.

These interview data constitute our original dataset, focusing on the driver's actions, decisions, and mindset to achieve the driving and the passenger's perception, expectation, and evaluation of the riding experience.

### B. Construct Driving-Evaluation Dataset

To bridge the gap in driving-specific domain knowledge currently lacking in LLMs, we extracted key messages from the interview data and constructed a knowledge database of languages in the field of autonomous driving.

Firstly, we refined the interview data to ensure both accuracy and consistency. Because of the prevalence of colloquial expressions in the raw interviews, directly utilizing this unprocessed data could potentially hinder subsequent analysis and database development. To address this issue, we employed GPT-4 as a tool to transform the informal spoken content into more formal and concise textual data. This process not only enhances the readability of the dataset but also mitigates errors caused by linguistic redundancy, thereby ensuring a higher level of consistency and structure for further data processing stages.

Next, we classified and processed the refined data based on the distinct roles of drivers and passengers, aiming to analyze and understand the divergent knowledge and evaluations of driving behavior between these two groups. The driver



Authorized licensed use limited to: Tsinghua University. Downloaded on April 01,2026 at 14:01:02 UTC from IEEE Xplore. Restrictions apply.

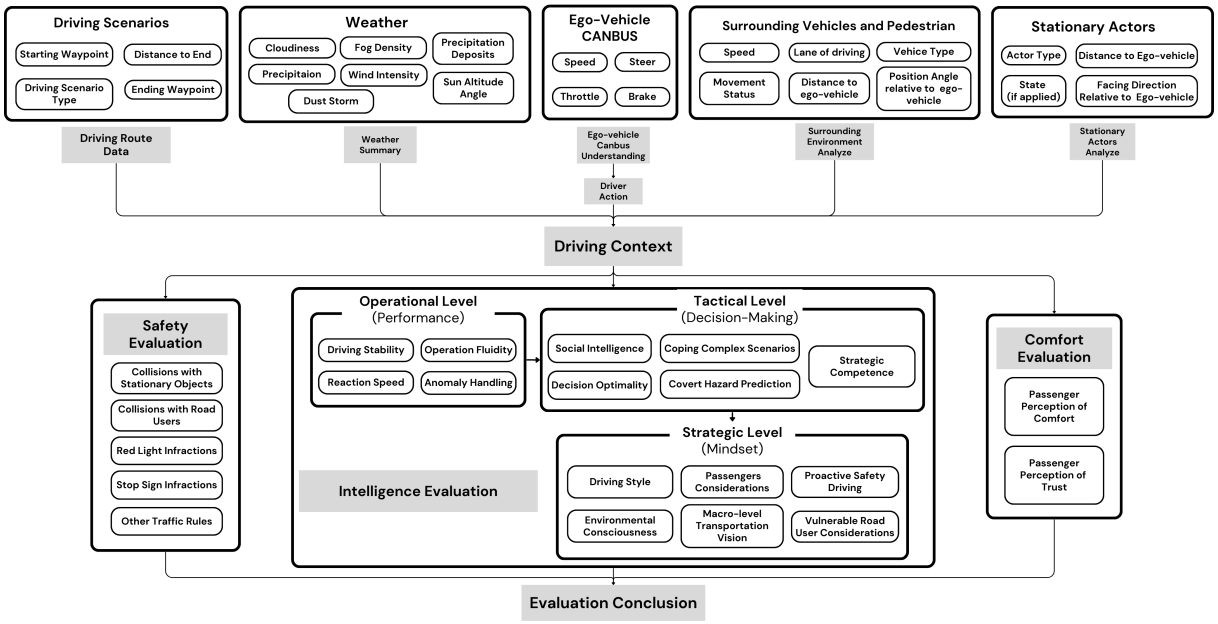



Fig. 2. Framework Structure for Driving Intelligence Evaluation

data was processed into knowledge units according to five key elements: "Context", which encapsulates the driving environment and traffic interactions; "Driver Mindset", reflecting the driver's cognitive state and motivations; "Driving Decision", which elucidates the reasoning behind the driver's choices and actions in specific contexts; "Driver Action", denoting the concrete operational behaviors executed by the driver; and "Driver Evaluation", summarizing the driver's assessment and reflection on their driving performance.

Passenger data units followed a different analytical framework: "Context", describing the driving environment from the passenger's perspective; "Passenger Mindset", which identifies the passenger's core beliefs, concerns, and emotional state during the journey; "Expectations", referring to the passenger's anticipated behavior from the driver; "Passenger Perception", capturing the passenger's actual experiences and reactions to the drive; and "Passenger Evaluation", reflecting the passenger's overall assessment of the driving experience.

By employing this structured categorization, we can effectively capture the distinct perspectives and cognitive processes of both drivers and passengers, providing a more nuanced and multi-dimensional foundation for building a knowledge database.

Finally, after processing the data, we convert it into JSON format to build an external knowledge base, including 700 Driver Knowledge Units and 760 Passenger Knowledge Units. This knowledge base enhances our system's evaluation using the Retrieval-Augmented Generation (RAG) mechanism. Providing a structured driving-related knowledge dataset, allows LLMs to generate more accurate and well-informed assessments, improving their reasoning and evaluation in this field.

## IV. DRIVING EVALUATION FRAMEWORK

To address the limitations of traditional driving performance evaluation methods, which often focus on isolated quantitative metrics and overlook the complexities of real-world scenarios, contextual nuances, and decision-making intricacies, we propose an LLM-based autonomous driving evaluation system (Fig. 2). This system processes both quantitative and contextual driving data, providing a comprehensive assessment encompassing safety, intelligence, and comfort. Our approach ensures that evaluations are accurate and reflective of nuanced driving environments encountered in practice.

### *A. From driving data to understandable driving context*

In autonomous driving, interpreting vast amounts of complex real-world data is crucial for safe and efficient algorithm operation. To enhance evaluation stability, we transform raw driving data into structured driving contexts. Our system categorizes this data into five types:

- Driving Scenarios: Information about the route setup
- Weather Conditions: Factors influencing driving (e.g., fog density affecting visibility).
- Ego-vehicle CAN BUS Data: Real-time vehicle metrics.
- Surrounding Vehicles and Pedestrians: Data on nearby vehicles (two closest in the same lane, closest in adjacent lanes, and closest opposing traffic), pedestrians, and special vehicles like firetrucks.
- Stationary Actors: Relevant objects like traffic lights and signs facing the ego-vehicle.

Our approach summarizes the route and weather conditions, followed by collision information and CAN BUS data. We then format detailed descriptions of surrounding vehicles and driving-relevant objects.



Authorized licensed use limited to: Tsinghua University. Downloaded on April 01,2026 at 14:01:02 UTC from IEEE Xplore. Restrictions apply.

### B. LLM-powered Driving Evaluation Framework

Building upon our driving evaluation dataset, we construct a comprehensive evaluation framework encompassing three primary dimensions: Safety Evaluation, Intelligence Evaluation, and Comfort Evaluation:

*1) Safety evaluation:* The safety dimension, being the most mature and quantifiable aspect of traditional evaluations, is a key component of our framework. Our safety evaluation focuses on quantifying various risk factors in the driving environment and assessing the driver's adherence to traffic regulations and ability to navigate potential hazards. This assessment encompasses common autonomous driving evaluation metrics, including collision avoidance, appropriate handling of traffic signs, and adherence to traffic lights.

*2) Intelligence Evaluation:* The intelligence dimension represents the most critical metric extracted from our Driving-Evaluation dataset. In our framework, this evaluation assesses the driver's adaptive behaviors, decision-making processes, and strategic planning in complex traffic situations. We employ a Chain-of-Thought (CoT) methodology, encompassing three hierarchical driving levels: Operational, Tactical, and Strategic.

Our hierarchical evaluation implements cascading reasoning - operational outputs inform tactical analysis, which subsequently feeds strategic assessment. At each level, the LLM dynamically retrieves relevant cases from our structured knowledge base via RAG, providing domain-specific knowledge.

*a) Operation Level:* The operational level evaluates the driver's actions with the driving context, focusing on the driving operation performance, and evaluating whether the vehicle is well-controlled through the testing criteria:

- Driving Stability. Vehicle's ability to maintain stable control, focusing on steady speed and consistent direction without unnecessary adjustments or swerves.
- Operation Fluidity. Smoothness and coherence of the vehicle's maneuvers, such as lane change and overtake.
- Anomaly Handling. Vehicle's capacity to detect anomalies, make timely decisions, and navigate safely through scenarios like road constructions or accidents.
- Reaction Speed. Reaction time of the vehicle when confronted with sudden events.

*b) Tactical Level:* Following the operational assessment, both the driving context and the operational evaluations will be passed to the tactical level evaluation. The tactical level evaluation delves into the decision-making processes of the vehicle, encompassing:

- Social Intelligence. The vehicle's ability to communicate intentions and follow social driving norms, such as turning the vehicle heading to the target direction, and providing great predictability for other road users.
- Coping Complex Scenarios. The vehicle's problem-solving abilities in challenging environments like urban intersections or pedestrian-heavy areas, reflect the vehicle's adaptability to make quick and appropriate decisions to navigate safely and highly efficiently.
- Strategic Competence. Vehicle's ability to anticipate other road users' actions. Address the ability to handle the situation when multiple road users have conflicting interests, and make optimal choices to avoid conflicts while maximizing passing efficiency.
- Covert Hazard Prediction. Vehicle's proficiency in anticipating hidden risks, focusing on the vehicle's predictive ability and preparedness for potential hazards.
- Decision Optimality. The consistency and quality of the vehicle's decision-making processes across various driving scenarios, assessing its ability to make optimal choices in diverse situations.

*c) Strategic Level:* Enhanced by the result from the previous two levels and driving context, the Strategic Level assesses the vehicle's higher-level understanding of the driving environment. This level of evaluation focuses on the mindset of the vehicle and analyzes its strategy of driving in these aspects:

- Macro-level Transportation Vision. Understand and predict broader traffic patterns, assessing its capacity to contribute to overall traffic flow optimization.
- Driving Style. Risk tolerance and driving style between aggressive and cautious, evaluating how it balances safety and efficiency in various driving contexts.
- Vulnerable Road User Consideration. Adherence to ethical driving principles and social norms, assessing the ability to prioritize vulnerable road users like bikes.
- Passengers Consideration. The ability to provide a smooth and pleasant driving experience, reflects its capacity to balance performance with passenger comfort.
- Environmental Consciousness. The vehicle's ability to optimize its driving patterns for energy efficiency, reflects its contribution to environmental sustainability.
- Proactive Safety Driving. Ability to anticipate and prevent potentially dangerous situations, assessing its proactive approach to safety.

*3) Comfort Evaluation:* The comfort dimension, a well-established metric in traditional vehicle assessment, is integrated into our framework by extracting passenger evaluation logic from the Driving Evaluation dataset. Focusing on the human experience within the vehicle. It recognizes that transportation aims not only for mobility but also for user comfort and perceived safety. We address two key perceptions:

- Passenger Perception of Comfort: Evaluates comfort based on vehicle actions (speed, throttle, braking, steering) and efforts to ensure a comfortable ride at tactical and strategic levels.
- Passenger Perception of Safety: Passenger feeling about vehicle actions, particularly during sudden events.

*4) Evaluation Conclusion:* After the sequential evaluations across these levels and dimensions, a summary of the driving behavior is produced. This assessment identifies areas for improvement and culminates with an overall driving score, providing a holistic view of the vehicle's performance. For the complete evaluation prompt used in this framework, please refer to the link provided in the abstract.



Authorized licensed use limited to: Tsinghua University. Downloaded on April 01,2026 at 14:01:02 UTC from IEEE Xplore. Restrictions apply.

## V. EXPERIMENT

### A. Experiment Environment Setup

In our experiment, we build the CARLA Autonomous Driving Leaderboard 2.0 platform on a Lenovo ThinkPad T15g Gen 1 Laptop with Ubuntu 22.04.4 LTS operation system, as the driving simulator for data collection. The leaderboard, different from the original CARLA, provides a series of predefined routes, which include 6 types in total 24 traffic scenarios. The predefined routes cover urban routes, freeways, residential regions, rural settings, and a variety of common weather conditions. In the experiment, we leverage OpenAI's GPT-4o APIs for performing our driving evaluation framework.

### B. Experiment Driving Data Collection

To comprehensively evaluate the capabilities of our assessment framework, we generated a diverse collection of driving data using CARLA Leaderboard. We collect all driving data by manual driving for better controllability. In order to retain the diversity of driving situations and weather conditions provided by predefined routes, cover as many different scenarios as possible, and avoid the problem of drowsy driving, we split the predefined driving routes into shorter routes by every five driving scenarios. By computing the combination of 8 short routes that cover most combinations of different driving scenarios, driving situations, and weather, we end up covering all types of driving scenarios, driving situations of urban routes, freeways, residential regions, and weather of daylight, rain, and fog. The data were collected once per second. Our experimental conditions combined two driving styles (Cautious/Aggressive) and two performance levels (Good/Bad): *Cautious-Good; Cautious-Bad; Aggressive-Good; Aggressive-Bad*. During the experiment, the four types of driving conditions are collected according to the guidelines, part of the criteria provided below:

*1) Criteria for good and bad performance:* **Good driving performance** include: adherence to traffic rules, maintaining speed consistency and the absence of collisions, etc. **Bad driving performance** include: frequent collisions, poor speed consistency, excessively reckless driving (in aggressive styles), excessive hesitation (in cautious styles), etc.

*2) Criteria for aggressive and cautious driving styles:* **Aggressive driving style**: Strong acceleration and deceleration, minimal use of the vehicle's natural speed reduction to decelerate, and maintaining small following distances, etc. **Cautious driving style**: Smoother acceleration and deceleration, slowing down to observe during complex road conditions, and maintaining a greater following distance.

These class-specific data also enable basic hallucination detection when LLM evaluations violate expected value ranges, for example, a high score in Proactive Safety Driving for Aggressive driving data is rare.

### C. System Evaluation

In our system setup, the three dimensions and three levels share equal weight during the corresponding evaluation, demonstrating the framework's baseline performance. The weighting factors are parameterizable through prompts to accommodate domain-specific requirements.

To understand the ability to evaluate driving conditions, we examine the accuracy of our evaluation results in the aspects of identifying between two driving styles, and distinguishing between the two driving performances. By randomly selecting one scenario from each of the eight test routes, and performing evaluations on the four driving conditions, the system correctly identifies 30 out of 32 scenarios in driving style, accounting for an accuracy of 93.75%. In the driving performance evaluation, the result correctly identifies 26 out of the 32 scenarios, accounting for an accuracy of 81.25%.

Furthermore, we performed a Spearman correlation analysis between the total intelligence scores assigned by our system and the driving scores of each scenario computed by the CARLA leaderboard. The analysis revealed a moderate and statistically significant correlation ($\rho = 0.561$, $p < 0.001$), suggesting a consistent evaluation trend between the two scoring systems. This finding indicates that our scoring method aligns reasonably well with CARLA's established assessment criteria, further supporting the validity and reliability of our evaluation approach.

### D. Human Evaluation

*1) Procedure:* Targeting to understand whether the human evaluation is consistent with our LLM-generated evaluation, we designed a survey questionnaire to collect human drivers' evaluations of the performance of our evaluation framework, which was presented to participants in the questionnaire through video clips of the driving footage. We initially collected basic information from participants. The participants evaluated the results of the framework with the corresponding video clip in randomized order. Each trial was accompanied by five rating questions and one reasoning question. Three of the five questions were designed to assess the degree of agreement on different levels of intelligence evaluation, while the remaining two addressed comfort evaluation and the overall result of the evaluation. These rating questions require participants to rate the level of agreement of the evaluation results from 0 to 10. The basic format of the rating questions in the questionnaire is as follows.

*"Based on the descriptions of the following five aspects, please rate your degree of agreement with the generated content. A score of 10 indicates complete agreement, while a score of 0 indicates complete disagreement. Lower scores represent lower levels of agreement."*

The reasoning question, which requires the participants to explain their ratings, appears as the final question for each experiment trial.

Additionally, to filter out carelessly completed questionnaires, we set a minimum answering time and included trap questions in the questionnaire, which required participants to select a particular option.

*2) Participants:* Our research team recruited participants through social media platforms, attracting over 60 individuals with a compensation of approximately $2.08 per valid questionnaire. After verification, 33 participants with valid



Authorized licensed use limited to: Tsinghua University. Downloaded on April 01,2026 at 14:01:02 UTC from IEEE Xplore. Restrictions apply.

| Driving Condition | Operation Level | Tactical Level | Strategic Level | Comfort Level | Overall |
|---|---|---|---|---|---|
| Aggressive Good | Mean: **7.05** SD: 1.97 | Mean: **7.03** SD: 2.06 | Mean: **6.88** SD: 2.25 | Mean: **5.77** SD: 2.97 | Mean: **7.03** SD: 2.24 |
| Aggressive Bad | Mean: **7.11** SD: 2.88 | Mean: **7.27** SD: 2.61 | Mean: **7.55** SD: 2.41 | Mean: **6.86** SD: 3.14 | Mean: **7.59** SD: 2.59 |
| Cautious Good | Mean: **7.95** SD: 1.55 | Mean: **7.78** SD: 1.70 | Mean: **7.83** SD: 1.67 | Mean: **7.92** SD: 1.68 | Mean: **7.72** SD: 1.87 |
| Cautious Bad | Mean: **6.67** SD: 2.37 | Mean: **6.59** SD: 2.53 | Mean: **6.76** SD: 2.47 | Mean: **6.89** SD: 2.55 | Mean: **6.77** SD: 2.71 |
| Overall | Mean: **7.19** SD: 2.30 | Mean: **7.17** SD: 2.30 | Mean: **7.25** SD: 2.27 | Mean: **6.86** SD: 2.75 | Mean: **7.28** SD: 2.41 |

Complete Disagreement 0 — Complete Agreement 10

Fig. 3. Human Agreement Score on System Evaluation

driving licenses were included in the final analysis. The sample comprised 17 male (52.22%) and 16 female (47.78%), with ages ranging from 19 to 54 years. Participants' driving experience varied widely. The majority had relatively low driving experience, with 33.3% having 0-2 years and 39.4% having 3-5 years of experience. More experienced drivers were also represented, with 12.1% having 6-10 years, 9.1% having 11-20 years, and 6.1% having over 21 years of driving experience. An interesting distribution emerged regarding self-reported driving styles. The majority (66.7%) of participants described their daily driving style as cautious. Nearly a third (30.3%) reported a neutral or situational driving style, adapting their approach based on circumstances. Only one participant identified the daily driving style as aggressive.

*3) Results:* In the evaluation of our LLM-powered framework, human drivers assessed various dimensions, including operational, tactical, strategic, comfort, and overall performance. Each dimension was rated on a scale from 0 to 10, with the results indicating a generally positive reception. The summary of human evaluation results is shown in Fig. 3.

### E. Findings

Overall, we found that human drivers provided favorable comprehensive evaluations of our system (7.28 out of 10). Notably, the acceptance levels for the three intelligence-related levels were remarkably consistent (7.19, 7.17, 7.25 out of 10), validating the rationality of our chain-of-thought design and its alignment with human driver evaluations.

Participants appreciated the objectivity and thoroughness of the safety assessment, as evidenced by the following comments: *'I concur with the low safety level assessment. The system's evaluation is highly objective and comprehensive.' (aggressive, bad condition) 'I agree with the identified issues in autonomous driving operations continuity. The system's evaluation is exceptionally thorough, analyzing perspectives of both the autonomous vehicle and other vehicles.' (cautious, bad condition);* However, the comfort evaluation received lower acceptance (6.86 out of 10). One participant remarked: *'I disagree with the comfort assessment. The vehicle's sudden acceleration during turns presents safety risks and low comfort. The system should strengthen its passenger comfort evaluation criteria.' (aggressive, bad condition)* Comfort perception varied considerably among participants, with some finding the comfort evaluation appropriate for the same video clip.

Among the four different driving conditions, our Cautious Good condition currently receives the highest level of acceptance (7.72 out of 10). In contrast, both the Cautious Bad and Aggressive Good conditions currently have relatively lower levels of acceptance (6.77 and 7.03 out of 10, respectively). This indicates that, comparatively, participants exhibited a lower degree of approval for the LLM's ratings in these two driving conditions. This disparity may indicate that driving styles characterized as cautious or aggressive are inherently coupled with driving performance (good or bad), resulting in greater difficulty in achieving consensus in evaluations for these conditions [47]. The human validation study also serves as a critical hallucination detection mechanism, with low acceptance showing possible hallucination happening or incorrect evaluation occurring.

## VI. CONCLUSIONS AND FUTURE WORK

This paper introduces a novel evaluation framework for assessing driving intelligence in complex traffic scenarios, addressing a critical gap in autonomous driving research.

Our evaluations demonstrate the framework's potential to provide nuanced, context-aware assessments of driving behavior intelligence, contributing significant insights to the field. However, limitations include reliance on simulated environments, potential biases in the natural language dataset, the possible inconsistency in manual driving data collection, and challenges with rare scenarios. The real-world driving data accounts for different driving levels and road conditions, but safety-critical scenarios, such as accidents or extreme situations, are rare. For example, in the two most recent years (2022 and 2023), in 22 countries, the number of fatalities rarely exceeded 8 per billion vehicle-kilometers, with a median of 4-5 cases [48]. This suggests that safety-critical scenarios may not be adequately represented. Future work could address this limitation by integrating high-quality simulation results into the knowledge dataset. Further exploration could also consider the framework's adaptability to diverse driving cultures, integration into real-world testing protocols, multi-modal data incorporation, edge case analysis, and the development of explainable AI. Additionally, conducting longitudinal studies, evaluating ethical decision-making, and exploring human-AI collaboration in semi-autonomous vehicles present promising research directions. These efforts aim to enhance the framework's applicability and effectiveness in advancing autonomous driving technology, ultimately leading to more sophisticated, human-like autonomous driving agents and comprehensive, human-centric evaluation systems.

## ACKNOWLEDGMENT

This research is supported by the National Key R&D Program of China (2022YFB4300300), Beijing Municipal Science and Technology Project (Nos. Z231100010323005), and Beijing Natural Science Foundation (Grant No.L233033).



Authorized licensed use limited to: Tsinghua University. Downloaded on April 01,2026 at 14:01:02 UTC from IEEE Xplore. Restrictions apply.

## REFERENCES


[1] Y. Hou, C. Wang, J. Wang, X. Xue, X. L. Zhang, J. Zhu, D. Wang, and S. Chen, "Visual evaluation for autonomous driving," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 1, pp. 1030–1039, 2021.

[2] J. Guo, U. Kurup, and M. Shah, "Is it safe to drive? an overview of factors, challenges, and datasets for driveability assessment in autonomous driving," *arXiv preprint arXiv:1811.11277*, 2018.

[3] C. Xu, W. Ding, W. Lyu, Z. Liu, S. Wang, Y. He, H. Hu, D. Zhao, and B. Li, "Safebench: A benchmarking platform for safety evaluation of autonomous vehicles," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 667–25 682, 2022.

[4] G. Lou, Y. Deng, X. Zheng, M. Zhang, and T. Zhang, "Testing of autonomous driving systems: where are we and where should we go?" in *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ser. ESEC/FSE 2022. New York, NY, USA: Association for Computing Machinery, 2022, p. 31–43. [Online]. Available: https://doi.org/10.1145/3540250.3549111

[5] M. Erdoğan, İ. Kaya, A. Karaşan, and M. Çolak, "Evaluation of autonomous vehicle driving systems for risk assessment based on three-dimensional uncertain linguistic variables," *Applied Soft Computing*, vol. 113, p. 107934, 2021.

[6] Y. Wang, Q. Zhang, L. Zhang, and Y. Hu, "A method to automatic measuring riding comfort of autonomous vehicles: Based on passenger subjective rating and vehicle parameters," in *Design, User Experience, and Usability. Application Domains: 8th International Conference, DUXU 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part III 21*. Springer, 2019, pp. 130–145.

[7] S. Gao, S. Paulissen, M. Coletti, and R. Patton, "Quantitative evaluation of autonomous driving in carla," in *2021 IEEE Intelligent Vehicles Symposium Workshops (IV Workshops)*. IEEE, 2021, pp. 257–263.

[8] J. Wei, X. Wang, D. Schuurmans, M. Bosma, b. ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 24 824–24 837. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf

[9] A. Seff, B. Cera, D. Chen, M. Ng, A. Zhou, N. Nayakanti, K. S. Refaat, R. Al-Rfou, and B. Sapp, "MotionLM: Multi-agent motion forecasting as language modeling." [Online]. Available: http://arxiv.org/abs/2309.16534

[10] J. Mao, Y. Qian, J. Ye, H. Zhao, and Y. Wang, "GPT-driver: Learning to drive with GPT." [Online]. Available: http://arxiv.org/abs/2310.01415

[11] S. P. Sharan, F. Pittaluga, V. K. B. G, and M. Chandraker, "LLM-assist: Enhancing closed-loop planning with language-based reasoning." [Online]. Available: http://arxiv.org/abs/2401.00125

[12] 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." [Online]. Available: http://arxiv.org/abs/2312.14150

[13] É. Zablocki, H. Ben-Younes, P. Pérez, and M. Cord, "Explainability of deep vision-based autonomous driving systems: Review and challenges," *International Journal of Computer Vision*, vol. 130, no. 10, pp. 2425–2452, 2022.

[14] J. Yuan, S. Sun, D. Omeiza, B. Zhao, P. Newman, L. Kunze, and M. Gadd, "RAG-driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model." [Online]. Available: http://arxiv.org/abs/2402.10828

[15] L. Chen, O. Sinavski, J. Hünermann, A. Karnsund, A. J. Willmott, D. Birch, D. Maund, and J. Shotton, "Driving with LLMs: Fusing object-level vector modality for explainable autonomous driving." [Online]. Available: http://arxiv.org/abs/2310.01957

[16] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, *et al.*, "A survey on evaluation of large language models," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, pp. 1–45, 2024.

[17] Y. Li, W. Zhang, K. Chen, Y. Liu, P. Li, R. Gao, L. Hong, M. Tian, X. Zhao, Z. Li, *et al.*, "Automated evaluation of large vision-language models on self-driving corner cases," *arXiv preprint arXiv:2404.10595*, 2024.

[18] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, *et al.*, "Judging llm-as-a-judge with mt-bench and chatbot arena," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[19] M. Gao, X. Hu, J. Ruan, X. Pu, and X. Wan, "Llm-based nlg evaluation: Current status and challenges," *arXiv preprint arXiv:2402.01383*, 2024.

[20] W. Yuan, R. Y. Pang, K. Cho, S. Sukhbaatar, J. Xu, and J. Weston, "Self-rewarding language models," *arXiv preprint arXiv:2401.10020*, 2024.

[21] T. Wu, W. Yuan, O. Golovneva, J. Xu, Y. Tian, J. Jiao, J. Weston, and S. Sukhbaatar, "Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge," *arXiv preprint arXiv:2407.19594*, 2024.

[22] Q. Pan, Z. Ashktorab, M. Desmond, M. S. Cooper, J. Johnson, R. Nair, E. Daly, and W. Geyer, "Human-centered design recommendations for llm-as-a-judge," *arXiv preprint arXiv:2407.03479*, 2024.

[23] Z. Tang, J. He, D. Pei, K. Liu, and T. Gao, "Testing large language models on driving theory knowledge and skills for connected autonomous vehicles." [Online]. Available: http://arxiv.org/abs/2407.17211

[24] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*. PMLR, 2017, pp. 1–16.

[25] D. Fu, W. Lei, L. Wen, P. Cai, S. Mao, M. Dou, B. Shi, and Y. Qiao, "Limsim++: A closed-loop platform for deploying multimodal llms in autonomous driving," *arXiv preprint arXiv:2402.01246*, 2024.

[26] M. Roshdi, N. Nayeer, M. Elmahgiubi, A. Agrawal, and D. E. Garcia, "A unified evaluation framework for autonomous driving vehicles," in *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2020, pp. 1277–1282.

[27] W. M. D. Chia, S. L. Keoh, C. Goh, and C. Johnson, "Risk assessment methodologies for autonomous driving: A survey," *IEEE transactions on intelligent transportation systems*, vol. 23, no. 10, pp. 16 923–16 939, 2022.

[28] H. Azevedo-Sa, S. K. Jayaraman, C. T. Esterwood, X. J. Yang, L. P. Robert Jr, and D. M. Tilbury, "Real-time estimation of drivers' trust in automated driving systems," *International Journal of Social Robotics*, vol. 13, no. 8, pp. 1911–1927, 2021.

[29] H. Kim, Y. Jo, M. Kim, C. Oh, and S. Lee, "Analyzing take-over event of autonomous vehicle for driving safety evaluation," in *2024 IEEE Intelligent Vehicles Symposium (IV)*, 2024, pp. 1727–1734.

[30] K. Kühnlenz, U. Schmid, and B. Kühnlenz, "A video-based study on perceived intelligence, subjective performance and trust under variation of prior information given to users in autonomous driving," in *2022 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO)*, 2022, pp. 1–4.

[31] K. Kühnlenz and B. Kühnlenz, "Study on the impact of situational explanations and prior information given to users on trust and perceived intelligence in autonomous driving in a video-based 2x2 design," in *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2023, pp. 1509–1513.

[32] Y. Li, K. Katsumata, E. Javanmardi, and M. Tsukada. Large language models for human-like autonomous driving: A survey. [Online]. Available: https://arxiv.org/abs/2407.19280v1

[33] W. Han, D. Guo, C.-Z. Xu, and J. Shen, "DME-driver: Integrating human decision logic and 3d scene perception in autonomous driving." [Online]. Available: http://arxiv.org/abs/2401.03641

[34] Z. Xu, Y. Zhang, E. Xie, Z. Zhao, Y. Guo, K.-Y. K. Wong, Z. Li, and H. Zhao, "DriveGPT4: Interpretable end-to-end autonomous driving via large language model." [Online]. Available: http://arxiv.org/abs/2310.01412

[35] X. Tian, J. Gu, B. Li, Y. Liu, Y. Wang, Z. Zhao, K. Zhan, P. Jia, X. Lang, and H. Zhao, "DriveVLM: The convergence of autonomous driving and large vision-language models." [Online]. Available: http://arxiv.org/abs/2402.12289

[36] Y. Huang, J. Sansom, Z. Ma, F. Gervits, and J. Chai. DriVLMe: Enhancing LLM-based autonomous driving agents with embodied and social experiences. [Online]. Available: https://arxiv.org/abs/2406.03008v1

[37] W. Wang, J. Xie, C. Hu, H. Zou, J. Fan, W. Tong, Y. Wen, S. Wu, H. Deng, Z. Li, H. Tian, L. Lu, X. Zhu, X. Wang, Y. Qiao, and J. Dai, "DriveMLM: Aligning multi-modal large language models with behavioral planning states for autonomous driving." [Online]. Available: http://arxiv.org/abs/2312.09245




Authorized licensed use limited to: Tsinghua University. Downloaded on April 01,2026 at 14:01:02 UTC from IEEE Xplore. Restrictions apply.

[38] L. Wen, D. Fu, X. Li, X. Cai, T. Ma, P. Cai, M. Dou, B. Shi, L. He, and Y. Qiao, "DiLu: A knowledge-driven approach to autonomous driving with large language models." [Online]. Available: http://arxiv.org/abs/2309.16292

[39] J. Zhang, Z. Huang, A. Ray, and E. Ohn-Bar, "Feedback-guided autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 000–15 011.

[40] Y. Sun, N. S. Pargoo, P. J. Jin, and J. Ortiz, "Optimizing autonomous driving for safety: A human-centric approach with LLM-enhanced RLHF." [Online]. Available: http://arxiv.org/abs/2406.04481

[41] C. Pan, B. Yaman, T. Nesti, A. Mallik, A. G. Allievi, S. Velipasalar, and L. Ren, "Vlp: Vision language planning for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 760–14 769.

[42] K. Ananthajothi, G. S. Sudarshan, and J. Saran, "Advancing autonomous driving with large language models: Integration and impact," in *2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*. IEEE, 2024, pp. 1–6.

[43] H. Shao, Y. Hu, L. Wang, S. L. Waslander, Y. Liu, and H. Li, "LMDrive: Closed-loop end-to-end driving with large language models." [Online]. Available: http://arxiv.org/abs/2312.07488

[44] C. Cui, Z. Yang, Y. Zhou, Y. Ma, J. Lu, L. Li, Y. Chen, J. Panchal, and Z. Wang, "Personalized autonomous driving with large language models: Field experiments." [Online]. Available: http://arxiv.org/abs/2312.09397

[45] S. Wang, Z. Yu, X. Jiang, S. Lan, M. Shi, N. Chang, J. Kautz, Y. Li, and J. M. Alvarez. OmniDrive: A holistic LLM-agent framework for autonomous driving with 3d perception, reasoning and planning. [Online]. Available: https://arxiv.org/abs/2405.01533v1

[46] A. M. Marcu, L. Chen, J. Hünermann, A. Karnsund, B. Hanotte, P. Chidananda, S. Nair, V. Badrinarayanan, A. Kendall, and J. Shotton, "Lingoqa: Visual question answering for autonomous driving," 2023.

[47] R. Yang, X. Zhang, A. Fernandez-Laaksonen, X. Ding, and J. Gong, "Driving style alignment for llm-powered driver agent," *arXiv preprint arXiv:2403.11368*, 2024.

[48] International Transport Forum (ITF), "Road safety annual report 2024," Paris, 2024.



Authorized licensed use limited to: Tsinghua University. Downloaded on April 01,2026 at 14:01:02 UTC from IEEE Xplore. Restrictions apply.