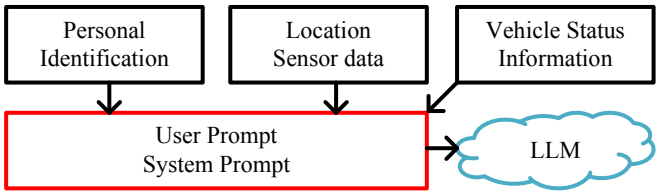
2024 IEEE Intelligent Vehicles Symposium (IV)
June 2-5, 2024. Jeju Island, Korea

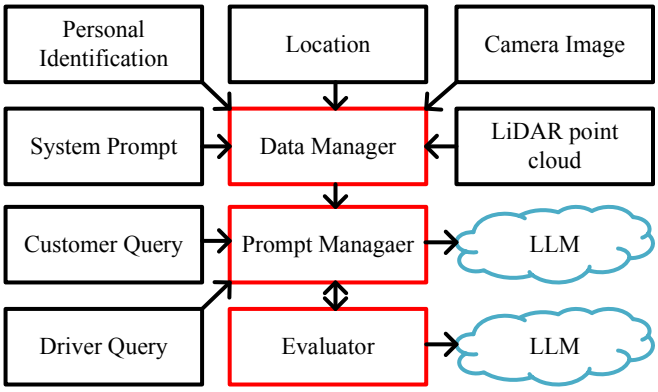
A Superalignment Framework in Autonomous Driving with Large Language Models

Xiangrui Kong[1,2], Thomas Braunl[1], Marco Fahmi[2], and Yue Wang[2,3]

***Abstract*— Over the last year, significant advancements have been made in the realms of large language models (LLMs) and multi-modal large language models (MLLMs), particularly in their application to autonomous driving. These models have showcased remarkable abilities in processing and interacting with complex information. In autonomous driving, LLMs and MLLMs are extensively used, requiring access to sensitive vehicle data such as precise locations, images, and road conditions. This data is transmitted to an LLM-based inference cloud for advanced analysis. However, concerns arise regarding data security, as the protection against data and privacy breaches primarily depends on the LLM's inherent security measures, without additional scrutiny or evaluation of the LLM's inference outputs. Despite its importance, the security aspect of LLMs in autonomous driving remains underexplored. Addressing this gap, our research introduces a novel security framework for autonomous vehicles, utilizing a multi-agent LLM approach. This framework is designed to safeguard sensitive information associated with autonomous vehicles from potential leaks, while also ensuring that LLM outputs adhere to driving regulations and align with human values. It includes mechanisms to filter out irrelevant queries and verify the safety and reliability of LLM outputs. Utilizing this framework, we evaluated the security, privacy, and cost aspects of eleven large language model-driven autonomous driving cues. Additionally, we performed QA tests on these driving prompts, which successfully demonstrated the framework's efficacy.**



(a) Insecure LLM-AD framework



(b) Proposed LLM-AD framework

Fig. 1. LLM Safety-as-a-service autonomous driving framework

I. INTRODUCTION

Large Language Models (LLMs) have gained significant attention recently, showing remarkable potential in emulating human-like intelligence [1]. A core challenge for aligning future superhuman AI systems (superalignment) is that humans will need to supervise AI systems much smarter than them [2]. The transformer-based network structure, mainly Generative Pre-trained Transformer (GPT) such as GPT-3 [3], and Llama2 [4], transfers the complexity of the data to the complexity of the network, and demonstrates powerful text reasoning and understanding capabilities. More and more autonomous systems are using LLMs as the interaction portal between humans and machines, including robots [5] and autonomous vehicles [6]. At present, the research on the interaction between LLM and unmanned systems is still in its infancy. Since LLMs need to perform inference on higher-power computing devices, the current mobile architecture cannot provide stable electrical power and computing power to support offline inference of LLMs. A common framework is to use LLMs in the cloud for inference and obtain the inference results of LLM through cloud service calls.

These LLM-driven autonomous agents has the following risks. First of all, decision-making reasoning for autonomous agent requires uploading a large amount of sensitive information such as image data, precise location, and personal information, which poses the risk of data leakage. Secondly, LLMs also face inherent challenges, such as being prone to subtle biases, arithmetic inaccuracies, and the risk of hallucinations. When LLM-driven unmanned systems interact with the environment, these built-in risks will be reflected in the real-world environment, leading to unknown consequences. Finally, the inference output results of LLM may not conform to the numerical values in specific situations, thereby violating local laws, regulations or customs, leading to a reduction in people's trust in LLMs.

The main contributions of this paper are summarized as follows:

*This work was supported in part by Australian Postgraduate Research Intern (APR.Intern) under reference number APR-2384, and INT-1256.

[1]The authors are with the Department of Electrical, Electronic and Computer Engineering, University of Western Australia, Crawley, WA 6009, Australia. E-mail: xiangrui.kong@research.uwa.edu.au, thomas.braunl@uwa.edu.au

[2]The authors are with the Department of Transport and Main Roads, Queensland Government, Brisbane, QLD 4000, Australia. E-mail: Marco.Fahmi@chde.qld.gov.au

[3]The authors are with the Center for Data Science, Queensland University of Technology, Brisbane, QLD 4000, Australia. E-mail: y355.wang@hdr.qut.edu.au

2024 IEEE Intelligent Vehicle Symposium (IV) | 979-8-3503-4881-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/IV55156.2024.10588403

- Propose a secure interaction framework for LLM, which serves as a guardrail between vehicles and cloud LLM, effectively censoring the data interacting with cloud-based LLM.
- We analyzed eleven autonomous driving methods based on large language models, including driving safety, token usage, privacy, and the alignment of human values.
- Utilizing our framework, we assessed the effectiveness of driving prompts within a segment of the nuScenes-QA dataset and compared the varying outcomes between the gpt-35-turbo and llama2-70b LLM backbones.

II. Related work

A. LLMs in Autonomous Driving

The knowledge is included in the LLMs not only for language tasks, but also for making goal-driven decisions in interactive environments [7]. LanguageMPC [8] employs LLMs to forecast vehicular dynamics, utilizing a bird's-eye view (BEV) to comprehend interactive situations or roundabout scenarios, alongside the consideration of the vehicles' current status. The Agent-Driver [9] method develops an LLM-driven framework capable of processing a variety of driving information, including images, point clouds, driving rules, and maps, which allows the LLM to access and interpret this diverse data through function calls, utilizing a chain-of-thought approach for comprehensive analysis. The DriveLLM [10] method integrates rule-based driving methods with LLMs, implementing the LLM for campus driving scenarios, and demonstrates high real-time performance within a stable network, evidenced by the efficient token processing time in GPT-3.5.

Currently, there exists a notable gap in the security research concerning the application of pre-trained large AI models in autonomous driving. Self-driving cars are at risk of potentially harmful or malicious activity when interacting with cloud systems [11]. This process entails detecting and countering attempts to jam or disrupt communication signals, discerning and addressing false or misleading information, and responding to efforts to hack or compromise the vehicle's systems [12]. The survey [13] referenced identifies various common Non-IP-based attacks on autonomous vehicles, such as position falsification [14], dissemination of false information [15], Sybil attacks [16], and privacy issues [17]. With the growing incorporation of LLMs in autonomous driving applications, the range of these attack methods is expected to expand.

B. Privacy and Alignment in LLMs

As both the model and data size increase, generative LLMs show a promising ability to understand and are capable of integrating classification tasks into their generative pipelines [18]. The safety issues related to LLMs have recently garnered widespread attention [19]. Although Differential Privacy (DP) [20] provides a theoretical worst-case privacy guarantee for safeguarded data, current privacy mechanisms considerably diminish the utility of LLMs, making many existing approaches impractical.

In the realm of LLMs, recent research has identified three safety areas of concern: prompt injection, data breaches, and model hallucinations. The phenomenon of prompt injection emerges as a significant security risk, wherein specifically crafted inputs are utilized to manipulate or exploit the natural language processing capabilities of AI systems. Moreover, LLMs are susceptible to inadvertent data breaches, where sensitive information may be leaked through model outputs, often attributed to the incorporation of confidential datasets during the training phase [21]. Additionally, a critical issue identified in these models is their tendency towards hallucination, where they generate erroneous or illogical information, often with a false sense of confidence, due to limitations in their predictive text generation algorithms [22]. These findings underscore the need for enhanced security measures and algorithmic refinements in the development and deployment of LLMs to mitigate these risks.

In the burgeoning field of artificial intelligence, the alignment of LLMs with human and organizational values presents a critical area of research, necessitating a multifaceted approach to ensure ethical and effective AI deployment [23]. In current research on LLMs, alignment of output text is primarily influenced through two methods. Firstly, the training data of the LLM significantly impacts its alignment, shaping the nature of the generated content [24]. Secondly, LLM service providers offer optional API alignment services, designed to filter out content that starkly deviates from predefined norms or standards [25]. Additionally, LLM customers often customize alignment requirements to suit their specific needs, typically employing simpler methods such as Retrieval-Augmented Generation (RAG) [26] or tailored prompting techniques.

III. Method

In order to model the behavior of LLM and alignment tasks, we follow the theoretical approach called Behavior Expectation Bounds (BEB) [23]. The behavior scoring functions are defined along a vertical axis B as $B : \Sigma^* \to [-1, 1]$. These functions evaluate a text string from an alphabet Σ, assessing how the behavior B is exhibited within the string. A score of $+1$ indicates a highly positive manifestation of B, while a score of -1 signifies a highly negative manifestation.

Given a probability distribution of language model $\mathbb{P}$ prompted with a text string s_0. After n times prompt conversation, we define the $n + 1$ behavior of the conditional probability $B_{\mathbb{P}}(s_{n+1})$ as follow:

$$B_{\mathbb{P}}(s_{n+1}) := \mathbb{E}_{s_1 \oplus \ldots \oplus s_n \sim \mathbb{P}(\cdot|s_0)}[B(s_0)] \tag{1}$$

Where $s_1 \oplus \ldots \oplus s_n \sim \mathbb{P}(\cdot\,|s_0)$ indicates sampling n continuous sentences from the conditional probability distribution $\mathbb{P}(\cdot\,|s_0)$ with the system prompt s_0.

The first important task for LLM-AD is alignment task defined as follow, for a text string s, we want $B_{\mathbb{P}}(s) \to 1$. Specifically, let $\gamma \in (0, 1]$, we say that an LLM with distribution $\mathbb{P}$ is γ-**prompt-alignable** w.r.t behavior B, if for

any $\epsilon > 0$ there exists a textual prompt $s^* \in \Sigma^*$ such that $B_{\mathbb{P}}(s^*) < \gamma + \epsilon$ where the ϵ represents a small positive number that shows how aligned the behavior values are.

The next problem is to facilitate an assessment of the extent to which sensitive data are incorporated into LLMs, we introduce the concept of probability mapping functions $D_{\mathbb{P}}(s_n)$ denoted as follow,

$$D_{\mathbb{P}}(s_n) : \mathbb{E}_{s_1 \oplus \ldots \oplus s_n \sim \mathbb{P}(\cdot|s_0, I)} \rightarrow [0, 1] \tag{2}$$

Where the context of a prompted LLM is represented as $\mathbb{P}(\cdot|s_0, I)$, where I signifies a predefined list of sensitive data. This approach allows for a systematic analysis of the LLM's interaction with and utilization of sensitive data elements in its processing and output generation.

Then we present a key aspect of our framework, an underactuated wheeled system command functions

$$C_{\mathbb{P}}(s_n) : \mathbb{E}_{s_1 \oplus \ldots \oplus s_n \sim \mathbb{P}(\cdot|s_0)} \rightarrow C_{dr} \times C_{aux} \tag{3}$$

where C_{dr} is underactuated wheeled system command space including steering angle θ and vehicle speed v. C_{aux} is auxiliary command space including other control command such as light control, catch camera images. Under these function, we define the LLM-AD safety problem under the following three conditions including driving safety, data safety, and LLM alignment. The parameters delineated in Table I denote

TABLE I

COMMAND SPACE OF C_{dr} AND C_{aux}

Space	Symbol	Range*	Meaning
C_{dr}	θ	$[-30°, 30°]$	steering angle
	v	$40km/h$	vehicle speed
C_{aux}	b_{al}	$0/1$	alarm
	b_{rp}	$0/1$	ramp
	b_{wp}	$0/1$	wiper
	b_{dr}	$0/1$	door
	b_{sp}	$string$	speaker

*Ranges vary according to different vehicle models.

the dimensions of the driving command space and auxiliary command space, with variations contingent upon distinct vehicular models. The prevailing underactuated kinematic model, commonly adopted in vehicular systems, facilitates control via manipulation of steering angle and velocity. These primary parameters collectively govern the trajectory of vehicle motion. Conversely, auxiliary instructions encompass vehicle control directives that lie beyond the scope of the kinematic model. Such instructions typically encompass functionalities such as alarm activation, wiper control, door manipulation, and in certain instances, specialized features such as ramps and speaker systems, particularly observed in public transportation vehicles.

The first condition state define the safety driving problem which is $\forall s_i,\ C_{\mathbb{P}_\phi}(s_i) \subseteq \tilde{C}$ where for all input context string s_i, the set of vehicle command states $C_{\mathbb{P}_\phi}(s_i)$ as identified by a probability distribution of a language model $\mathbb{P}_\phi$ must a subset of a safety driving space $\tilde{C}$, where $\tilde{C} := \tilde{C_{dr}} \times \tilde{C_{aux}}$. The second condition state shows the data safety problem which is $D_{\mathbb{P}_\psi}(s_i) \rightarrow 0$. We want the prompt queries have less sensitive data especially when the LLM deployed on cloud. The third condition $B_{\mathbb{P}_\omega}(s_i) \rightarrow 1$ indicates to align the LLM behaviors in natural language processing as there are conversation tasks between the LLM and passengers. For a single LLM agent structure, $\mathbb{P}_\phi$=$\mathbb{P}_\psi$=$\mathbb{P}_\omega$. These conditions collectively define a safety problem in LLM-based autonomous driving, focusing on the likelihood of encountering critical states and the model's response to such scenarios shown in Table II.

TABLE II

QUALITATIVE ANALYSIS OF LLM-AD TASK EXAMPLES

LLM-AD Task	Sensitive data usage	Related drive	Value alignment
Passenger tutorial	Low	N/A	High
Traffic light analysis	Low	High	High
Driving Instruction	Medium	High	Medium
Lane keeping	Medium	High	N/A
Incident record	High	Low	Low
In-car conversation	High	N/A	High
Route suggestions	High	Medium	High
Pedestrian detection	High	High	Medium

IV. EXPERIMENTS

Currently LLM-driven driving methods adopt the framework depicted in Figure1a, which involves setting predefined prompts and using tokenized image information to limit the scope of the LLM agent's reasoning. Furthermore, during follow-up conversations, all necessary information for reasoning is relied on the agent textually. In the evaluation of LLM-based autonomous driving methods, a multifaceted approach is necessary to assess performance across several critical dimensions.

A. Implement details

We evaluated system prompts from eleven LLM-driven autonomous driving research papers, creating an evaluation framework using AutoGen [27]. Initially, gpt-35-turbo and llama2-70b-chat were used to perform an overall evaluation of driving prompts, including aspects such as driving safety, token quantity, sensitive data usage, and alignment. Afterwards, 250 question-answer pairs were chosen from the nuScenes-QA dataset for simulated evaluation, comparing binary scale results, token consumption, and response time.

B. Evaluation of Safety Capabilities

Our experiment examines the latest eleven studies that have integrated LLM into autonomous driving methods. TableIII provided outlines a comparative analysis of system prompts in various LLM-AD methods, utilizing metrics that include token cost, driving safety rates, sensitive data usage, and alignment ranking. The token count is determined using the *cl100k_base* tokenizer. Driving safety metrics are based on experimental outcomes reported in the respective studies. We've tracked the usage of various sensitive data in the system prompt, which includes current speed, precise locations, historical movement patterns, traffic updates, obstacle detection, weather reports, energy consumption, vehicle health

status, sign information, and emergency services. Alignment measures how closely the driving habits described in the system prompt match those of human drivers, using a scale from 0 to 100, where the values are whole numbers. Both the assessment of sensitive information usage and the alignment evaluation are conducted with the assistance of GPT-4-turbo.

TABLE III
EVALUATION OF LLM-AD METHOD SYSTEM PROMPT

Method	Model	Token↓	Safety*↑	Sens.↓	Align.↑
DLAH [28]	gpt-3.5	673	>60%	20	65
SurrealDriver [29]	gpt-4	310	81.4%	25	85
DriveGPT4 [30]	LLaVa	469	87.97%	30	50
DILU [31]	gpt-3.5	384	93%	25	60
WayveDriver [32]	gpt-3.5	186	83.9%	20	55
LanguageMPC [8]	gpt-3.5	1426	80%	25	70
DriveLLM [10]	gpt-4	427	66.6%	30	75
Agent-Driver [9]	gpt-3.5	429	99.13%	30	80
ADriver-I [33]	gpt-3.5	226	91.3%	35	45
GPT-Driver [34]	gpt-3.5	265	95.7%	25	70
DriveMLM [35]	gpt-3.5	494	78%	17	92

Notably, the 'Agent-Driver' [9] method demonstrates exemplary safety performance with a 99.13% rating and a high alignment score of 80, indicating robust adherence to safety and ethical standards. On the other hand, the method proposed by wayve showcases exceptional efficiency, evidenced by the lowest token count of 186, suggesting a streamlined processing capability. When considering the balance between performance metrics, 'SurrealDriver' and 'DriveLLM', both employing the GPT-4 model, offer substantial safety assurances with over 65% safety ratings, though 'DriveLLM' has a reduced alignment score in comparison to 'SurrealDriver', signifying a potential compromise between safety and ethical alignment. As the only method in the table with road trials, the method of DriveLLM does not directly report collision rates but instead examines the LLM's response time.

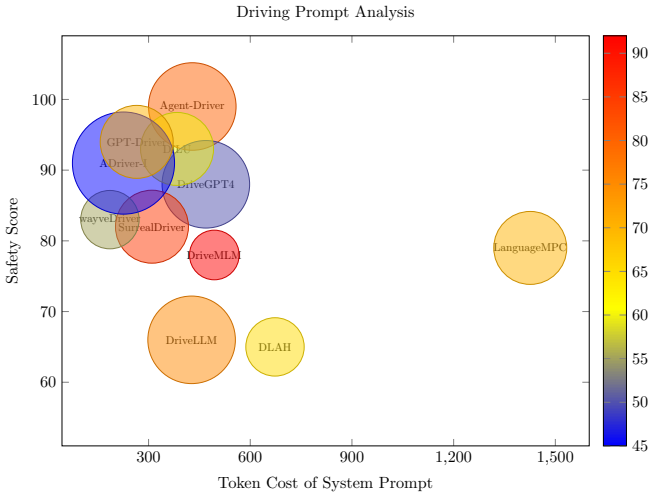


Fig. 2. LLM-AD system prompt analysis

Figure 2 provides a graphical representation of Table III. The x-axis shows the average token count of the system prompts featured in the literature, while the y-axis indicates the evaluators' ratings for safe driving. Larger circle radii indicate a greater use of sensitive data. Additionally, the lighter the color of the circle, the more closely it aligns with the driving standards of human drivers, and the opposite is also true.

In order to further analyze the vehicle sensitive data used by each method, we counted the occurrence times of various types of data in the system prompt, and the visual results after normalization for each model are shown in the Figure 3. We examined a series of sensitive data labels comprising: current speed (SC), precise location (PL), waypoints (WP), traffic conditions (TF), obstacle detection (OD), weather conditions (WT), energy consumption metrics (EC), vehicle health status (VH), signage information (SI), and emergency services (ES).

	CS	PL	WP	TF	OD	WT	EC	VH	SI	ES
LanguageMPC	10	0	20	60	0	0	0	0	0.1	0
Agent-Driver	0	0	50	10	30	0	0	0	10	0
DriveLLM	12	0	0	12	25	25	0	12	0	12
DriveGPT4	25	0	0	25	25	0	0	0	25	0
SurrealDriver	38	0	0	12	25	0	0	0	25	0
DLAH	20	20	0	20	20	0	0	0	20	0
DILU	33	0	0	67	0	0	0	0	0	0
WayveDriver	20	20	0	20	20	0	0	0	20	0
ADriver-I	50	0	0	50	0	0	0	0	0	0
GPT-Driver	17	0	17	17	17	0	17	17	0	0
DriveMLM	20	0	0	20	20	0	0	0	20	20

Fig. 3. LLM-AD system prompt analysis of sensitive data usage

C. Perception Capabilities Evaluation

To delve deeper into the safety of these models, we selected 50 questions from each category in the nuScenes-QA dataset [36]. This natural language queries of dataset fall into five groups: existence, count, object, status, and comparison. These queries are great for gauging an AD models environmental perception capabilities around vehicles. We evaluated those autonomous driving prompts using two major large language models, gpt-3.5-turbo and llama2-70b-chat. Our method involved checking if the Prompt could handle the nuScenes-QA queries and then averaging the scores of both models, using weights derived from their performance in the LLM boxing competition [37].

Table IV and Table V shows the result of those driving prompts including accuracy, token cost and time cost in different question types evaluated by gpt-35-turbo and llama2-70b-chat respectively. In Table IV evaluated by GPT-3.5, the models exhibit a range of accuracy in different question types, from a low of 14.0% (DILU in Comparison) to a high of 96.0% (Agent-Driver in Object). The overall

accuracy (Acc) also varies significantly, with Driver Like A Human (DLAH) achieving 88.8%, marking it as one of the most effective models in this evaluation. Table V evaluated by LLaMa2 indicates that ADriver-I excels with the highest accuracy reported, peaking at 97.0% in Comparison and 99.0% in Object queries. In contrast, several models like WayveDriver and DriveGPT4 show markedly lower performance, with overall accuracies of 22.8% and 16.4%, respectively.

Currently, the assessment of prompts using LLMs is linked to their linguistic capabilities. Typically, models with more advanced processing power yield more credible evaluations. Consequently, we performed a weighted summation of the Driver prompt's accuracy, taking into account the language skills of GPT-3.5 and LLaMa2, as illustrated in Figure 4.

Figure 5 illustrates how different prompt models perform in answering various types of questions in the nuScenes-QA dataset. It's evident that these models are generally more adept at responding to question types of exist, object, and status, as opposed to those involving counting and comparisons.

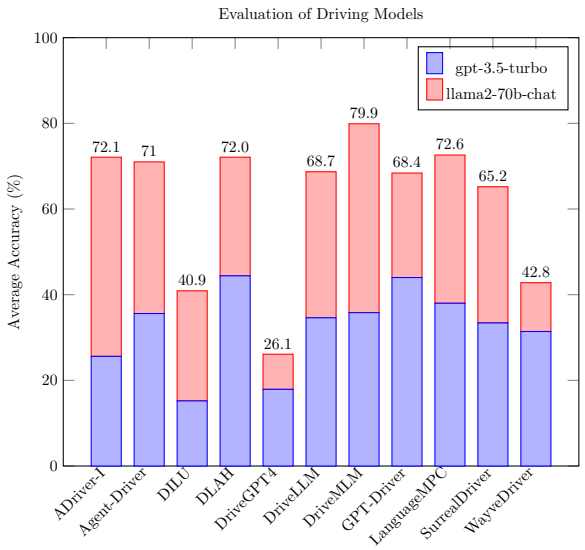


Fig. 4. Overall accuracy in nuScenes-QA dataset

V. CONCLUSION

We've developed a secure LLM driven autonomous driving framework, broadening the theoretical application of LLMs in AD safety. We evaluated the leading LLM-driven AD approaches in terms of driving safety, sensitive data usage, Token consumption, and alignment scenarios. Recognizing that these prevailing LLM-AD methods overlook key safety aspects during driving, our paper introduces a comprehensive LLM safety assessment framework based on a multi-agent system. This framework enhances the conventional structure by integrating a safety assessment agent, ensuring both vehicular safety and proper alignment.

TABLE IV

PERFORMANCE OUTCOMES OF VARIOUS MODELS ON THE CURATED NUSCENES-QA TEST DATASET EVALUATED BY GPT-3.5-TURBO

Model	Comparison			Count			Exist			Object			Status			Acc
	Acc↑	Token↓	Time↓	Acc	Token	Time	Acc	Token	Time	Acc	Token	Time	Acc	Token	Time	
ADriver-I	24.0%	6.0	0.34	16.0%	6.0	0.38	84.0%	6.0	0.42	76.0%	6.0	0.36	56.0%	6.0	0.37	51.2%
Agent-Driver	54.0%	6.2	0.35	48.0%	7.1	0.36	94.0%	7.3	0.41	**96.0%**	5.9	0.38	64.0%	6.6	0.39	71.2%
DILU	14.0%	6.2	0.38	12.0%	6.0	0.36	42.0%	4.9	0.37	42.0%	6.0	0.36	42.0%	5.4	0.37	30.4%
DLAH	84.0%	6.0	0.37	**84.0%**	6.0	0.38	**100%**	6.0	0.41	84.0%	6.0	0.36	**92.0%**	6.0	0.37	**88.8%**
DriveGPT4	75.0%	8.0	0.41	16.0%	7.9	0.40	46.0%	7.1	0.38	22.0%	7.6	0.36	20.0%	7.6	0.38	35.8%
DriveLLM	52.0%	6.1	0.35	38.0%	6.4	0.37	96.0%	7.0	0.41	88.0%	6.0	0.37	72.0%	6.0	0.37	69.2%
DriveMLM	64.0%	8.0	0.40	48.0%	8.9	0.41	94.0%	8.8	0.45	84.0%	8.8	0.41	68.0%	8.6	0.41	71.6%
GPT-Driver	**86.0%**	6.0	0.42	**84.0%**	6.1	0.38	90%	6.1	0.39	90%	6.0	0.35	90%	6.0	0.38	88.0%
LanguageMPC	56.0%	10.1	0.44	82.0%	2.6	0.33	96.0%	6.3	0.40	74.0%	10.6	0.39	72.0%	6.5	0.36	76.0%
SurrealDriver	44.0%	6.2	0.38	24.0%	6.1	0.35	94.0%	6.9	0.37	**96.0%**	6.5	0.37	76.0%	6.3	0.41	66.8%
WayveDriver	50.0%	6.0	0.35	18.0%	6.0	0.37	92.0%	6.0	0.37	80.0%	6.0	0.35	74.0%	6.0	0.38	62.8%

TABLE V

PERFORMANCE OUTCOMES OF VARIOUS MODELS ON THE CURATED NUSCENES-QA TEST DATASET EVALUATED BY LLAMA2-70B-CHAT

Model	Comparison			Count			Exist			Object			Status			Acc
	Acc↑	Token↓	Time↓	Acc	Token	Time	Acc	Token	Time	Acc	Token	Time	Acc	Token	Time	
ADriver-I	**97.0%**	12.3	7.66	81.0%	11.5	7.36	**95.0%**	12.6	8.22	**99.0%**	10.7	6.89	**93.0%**	12.1	7.88	**93.0%**
Agent-Driver	80.0%	10.8	7.24	59.0%	9.3	6.31	58.0%	9.9	6.68	78.0%	8.3	5.74	79.0%	10.0	6.79	70.8%
DILU	45.0%	12.4	8.10	33.0%	11.3	7.42	54.0%	11.4	7.50	67.0%	11.6	7.63	58.0%	11.9	7.86	51.4%
DLAH	57.0%	11.8	8.81	67.0%	11.5	8.65	43.0%	12.1	8.97	56.0%	12.1	8.91	54.0%	11.4	8.47	55.4%
DriveGPT4	13.0%	12.7	8.43	26.0%	12.8	8.50	16.0%	12.4	8.25	9.0%	12.8	8.52	18.0%	12.7	8.48	16.4%
DriveLLM	70.0%	11.6	7.84	44.0%	10.7	7.25	74.0%	11.3	7.63	82.0%	10.9	7.36	71.0%	10.9	7.37	68.2%
DriveMLM	94.0%	12.1	8.36	**84.0%**	11.5	7.86	84.0%	11.2	7.79	90.0%	11.2	7.68	89.0%	11.9	8.10	88.2%
GPT-Driver	45.0%	12.8	8.61	49.0%	12.6	8.52	51.0%	12.7	8.71	49.0%	12.8	8.68	50.0%	12.8	8.79	48.8%
LanguageMPC	68.0%	11.8	8.59	74.0%	11.5	8.41	65.0%	11.6	8.53	68.0%	12.2	8.91	71.0%	11.6	8.61	69.2%
SurrealDriver	79.0%	10.4	7.73	41.0%	10.0	7.47	68.0%	10.4	7.75	72.0%	10.0	7.43	58.0%	10.1	7.51	63.6%
WayveDriver	22.0%	11.7	7.55	15.0%	12.6	8.10	26.0%	10.7	6.98	23.0%	11.6	7.46	28.0%	12.4	7.94	22.8%

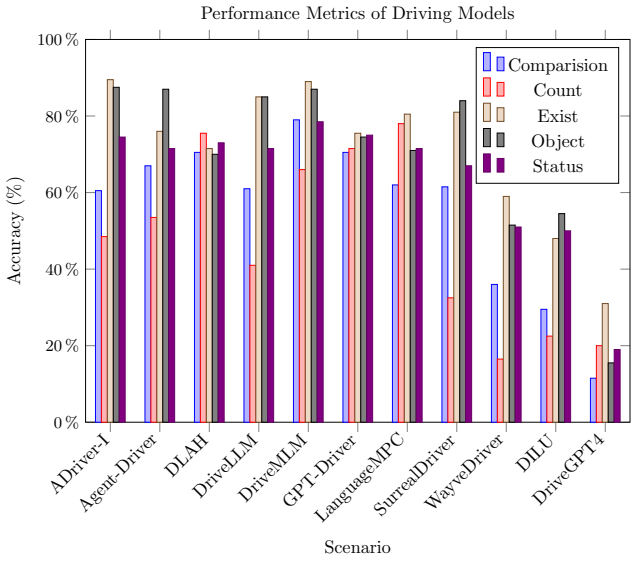


Fig. 5. Results of different models on five question types in nuScenes-QA dataset

ACKNOWLEDGMENT

The authors would like to thank all the Renewable Energy Vehicle Project (REV) sponsors for their support on this project. The authors thank Queensland Government Customer and Digital Group for their invaluable contributions and support.

REFERENCES

[1] C. Cui, Y. Ma *et al.*, "A survey on multimodal large language models for autonomous driving," *arXiv preprint arXiv:2311.12320*, 2023.

[2] C. Burns *et al.*, "Weak-to-strong generalization: Eliciting strong capabilities with weak supervision," 2023. [Online]. Available: https://cdn.openai.com/papers/weak-to-strong-generalization.pdf

[3] L. Floridi and M. Chiriatti, "Gpt-3: Its nature, scope, limits, and consequences," *Minds and Machines*, vol. 30, pp. 681–694, 2020.

[4] H. Touvron *et al.*, "Llama 2: Open foundation and fine-tuned chat models," 2023.

[5] D. Driess *et al.*, "Palm-e: An embodied multimodal language model," *arXiv preprint arXiv:2303.03378*, 2023.

[6] Y. Cui *et al.*, "Drivellm: Charting the path toward full autonomous driving with large language models," *IEEE Transactions on Intelligent Vehicles*, pp. 1–15, 2023.

[7] W. Huang, P. Abbeel, D. Pathak, and I. Mordatch, "Language models as zero-shot planners: Extracting actionable knowledge for embodied agents," 2022.

[8] H. Sha *et al.*, "Languagempc: Large language models as decision makers for autonomous driving," *arXiv preprint arXiv:2310.03026*, 2023.

[9] J. Mao, J. Ye, Y. Qian, M. Pavone, and Y. Wang, "A language agent for autonomous driving," 2023.

[10] Y. Cui *et al.*, "Drivellm: Charting the path toward full autonomous driving with large language models," *IEEE Transactions on Intelligent Vehicles*, 2023.

[11] M. L. Bouchouia *et al.*, "A survey on misbehavior detection for connected and autonomous vehicles," *Vehicular Communications*, vol. 41, p. 100586, 2023.

[12] V. L. Thing and J. Wu, "Autonomous vehicle security A taxonomy of attacks and defences," in *2016 ieee international conference on internet of things (ithings) and ieee green computing and communications (greencom) and ieee cyber, physical and social computing (cpscom) and ieee smart data (smartdata)*. IEEE, 2016, pp. 164–170.

[13] A. Boualouache and T. Engel, "A survey on machine learning-based misbehavior detection systems for 5g and beyond vehicular networks," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 2, pp. 1128–1172, 2023.

[14] S. So, P. Sharma, and J. Petit, "Integrating plausibility checks and machine learning for misbehavior detection in vanet," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2018, pp. 564–571.

[15] P. K. Singh, M. K. Dash, P. Mittal, S. K. Nandi, and S. Nandi, "Misbehavior detection in c-its using deep learning approach," in *Intelligent Systems Design and Applications: 18th International Conference on Intelligent Systems Design and Applications (ISDA 2018) held in Vellore, India, December 6-8, 2018, Volume 1*. Springer, 2020, pp. 641–652.

[16] J. Kamel, F. Haidar, I. B. Jemaa, A. Kaiser, B. Lonc, and P. Urien, "A misbehavior authority system for sybil attack detection in c-its," in *2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*. IEEE, 2019, pp. 1117–1123.

[17] A. Uprety, D. B. Rawat, and J. Li, "Privacy preserving misbehavior detection in iov using federated machine learning," in *2021 IEEE 18th annual consumer communications & networking conference (CCNC)*. IEEE, 2021, pp. 1–6.

[18] R. Colin *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[19] D. Tjondronegoro, E. Yuwono, B. Richards, D. Green, and S. Hatakka, "Responsible ai implementation: A human-centered framework for accelerating the innovation process," 2022.

[20] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[21] A. Namer, J. Miller, H. Vagts, and B. Maltzman, "A cost-effective method to prevent data exfiltration from llm prompt responses," 2023.

[22] A. Martino, M. Iannelli, and C. Truong, "Knowledge injection to counter large language model (llm) hallucination," in *European Semantic Web Conference*. Springer, 2023, pp. 182–185.

[23] Y. Wolf, N. Wies, O. Avnery, Y. Levine, and A. Shashua, "Fundamental limitations of alignment in large language models," 2023.

[24] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.

[25] B. Smith. (2023) How do we best govern ai? [Online]. Available: https://blogs.microsoft.com/on-the-issues/2023/05/25/how-do-we-best-govern-ai/

[26] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[27] Q. Wu *et al.*, "Autogen: Enabling next-gen llm applications via multi-agent conversation framework," 2023.

[28] D. Fu *et al.*, "Drive like a human: Rethinking autonomous driving with large language models," *arXiv preprint arXiv:2307.07162*, 2023.

[29] Y. Jin *et al.*, ": Designing generative driver agent simulation framework in urban contexts based on large language model," 2023.

[30] Z. Xu *et al.*, "Drivegpt4: Interpretable end-to-end autonomous driving via large language model," *arXiv preprint arXiv:2310.01412*, 2023.

[31] L. Wen *et al.*, "Dilu: A knowledge-driven approach to autonomous driving with large language models," 2023.

[32] L. Chen *et al.*, "Driving with llms: Fusing object-level vector modality for explainable autonomous driving," *arXiv preprint arXiv:2310.01957*, 2023.

[33] F. Jia *et al.*, "Adriver-i: A general world model for autonomous driving," 2023.

[34] J. Mao, Y. Qian, H. Zhao, and Y. Wang, "Gpt-driver: Learning to drive with gpt," 2023.

[35] W. Wang *et al.*, "Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving," *arXiv preprint arXiv:2312.09245*, 2023.

[36] T. Qian, J. Chen, L. Zhuo, Y. Jiao, and Y.-G. Jiang, "Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario," 2023.

[37] C. Holtz. (2024) Llm boxing. [Online]. Available: https://https://llmboxing.com/