WIP: From Detection to Explanation: Using LLMs for Adversarial Scenario Analysis in Vehicles

David Fernandez, Pedram MohajerAnsari, Cigdem Kokenoz, Amir Salarpour, Bing Li, Mert D. Pesé
Clemson University
{dferna3, pmohaje, ckokeno, asalarp, bli4, mpese}@clemson.edu

Abstract

We propose a framework that leverages Large Language Models (LLMs) for adversarial scenario analysis in Autonomous Vehicles (AVs), generating interpretable explanations for anomalies and bridging the gap between detection and semantic understanding. Conventional Deep Neural Networks (DNNs) lack robustness against adversarial perception attacks and provide limited interpretability. To address these limitations, our method uses LLMs to process structured vehicular data encoded in a Domain-Specific Language (DSL), incorporating the Manual on Uniform Traffic Control Devices (MUTCD) as a formal knowledge base. Leveraging zero-shot chain-of-thought (CoT) prompting, the framework distinguishes benign sensor errors from adversarial manipulations through stepwise reasoning. We introduce **`AutoSec-X`** , a dataset of 40 MUTCD-based driving scenarios, to evaluate LLM architectures, demonstrating that larger models (e.g., Gemini) exhibit superior domain-specific reasoning, often citing relevant MUTCD sections. Results validate the effectiveness of CoT-augmented LLMs for semantic anomaly analysis in AVs without labeled training data. Future work will extend **`AutoSec-X`** and investigate multimodal inputs [1].

1 Introduction

Autonomous Vehicles rely on the autonomous driving system (ADS) to achieve autonomous driving capabilities. The ADS combines perception, planning, and actuation modules to enable autonomous navigation [33, 36]. The perception module employs deep learning methodologies such as deep neural networks (DNN) to enable processing and application of perceptual data for obstacle classification, road surface segmentation, object detection [6, 17, 28], etc.

While these DNN models can effectively detect, track, and predict the movements of nearby objects [26, 39], they are vulnerable to physical perturbations that lead to adversarial perception attacks. Malicious actors manipulate physical-world integrity — such as altering traffic signs or obscuring obstacles — to mislead DNN models and disrupt the vehicle's decision-making [11, 20]. While DNNs can effectively detect anomalies they have encountered during training or bearing close similarity to known patterns, they lack the capacity for integrated semantic reasoning. Consequently, they struggle with new or adversarial scenarios that require understanding context or meaning beyond learned patterns [18, 27]. By conducting analysis of driving scenarios using Large Language Models (LLMs) semantic reasoning, we could identify recurring environmental contexts, such as zones with dense signage or unusual signage that cause DNNs to struggle. These insights could guide the development of more robust training datasets and targeted adversarial training.

Recent research has explored semantic anomaly detection using LLMs. For instance, studies have demonstrated that LLMs can monitor vision-based policies and identify semantic inconsistencies in robotic systems through human-like reasoning [9]. Similarly, zero-shot anomaly detection techniques in time-series data [4] and industrial settings [22] highlight the potential of LLMs for detecting anomalies without explicit training on labeled anomaly data. However, these works do not address AV-specific anomaly detection using LLMs.

To bridge this gap, we propose a novel anomaly detection approach that leverages pre-trained LLMs in a zero-shot chain-of-thought (CoT) manner to reason about possible anomalies in the driving environment. Unlike traditional DNN models, which rely on predefined patterns or anomaly-labeled datasets, our method enables AVs to identify and explain anomalies without requiring examples in the training data. For instance, if a school zone sign is altered from 25 mph to 45 mph, a DNN-based model would likely fail to recognize the violation due to its lack of contextual understanding, whereas an LLM can explain it based on its broader contextual understanding.

Furthermore, we envision that this tool could be used for forensic purposes by using data from a mandatory Event Data Recorder [29] or Automated Driving System Data Logger [37] — which are becoming increasingly common in

[1] The Dataset and implementation code of this work is available at https://github.com/tigerseclab/VehicleSec25_LLM_Reasoning

Level 3 and higher AVs — to analyze driving environments. This forensic capability is particularly crucial in the context of regulatory compliance, as highlighted in ISO/SAE 21434, as well as UN Regulations No. 155 and 157 (UN R155/157) [1, 21, 41], which emphasize the need for the generation and preservation of security logs, anomaly detection, and analysis of cyber incidents post-attack.

This WIP paper makes the following contributions:

- We introduce a zero-shot CoT LLM-based semantic reasoning approach that enables pre-trained LLMs to analyze structured adversarial and benign driving scenarios using guidelines of the Manual on Uniform Traffic Control Devices (MUTCD) [12] and identify anomalies without relying on labeled training examples.
- We introduce a semantically rich dataset called **`AutoSec-X` (Autonomous Vehicle Security & Explanation Dataset)** , which currently contains 40 scenarios in total, split evenly between benign and anomalous cases. Each scenario is based on established rules and regulations to reflect real-world driving conditions accurately. This preliminary dataset will be extended more systematically with additional scenarios in the future.
- We propose a robust evaluation strategy that quantitatively and qualitatively compares the reasoning processes of different LLMs, enabling an assessment of their effectiveness in identifying, explaining, and contextualizing anomalies within vehicle systems. We employ both traditional n-gram-based metrics, such as ROUGE [23] and BLEU [32], as well as semantic similarity measures like SBERT [35]. Our results indicate that while smaller models can detect anomalies in many cases, larger models demonstrate a deeper semantic understanding and more consistent domain-specific reasoning, as reflected by higher SBERT and BERTScore values.

Unlike approaches such as ScenicNL [10] and ChatScene [19], which focus on generating scenarios from natural language or enabling conversation with 3D scenes, our framework generates and analyzes benign and threat scenarios based on MUTCD rules. This approach provides significant advantages over human analysis, such as the reduction of cognitive biases that human investigators might introduce and scaling analysis to thousands of incidents. This paper pays equal attention to dataset generation and reasoning methodologies to establish a foundation for semantic understanding of anomalies in AV environments. While our current work evaluates LLMs on structured scenario descriptions rather than real-time vehicle data, it lays essential groundwork for future applications with actual AV log files.

2 Related Work

Our work introduces a zero-shot CoT LLM-based semantic reasoning method that enables pre-trained LLMs to analyze structured driving scenarios. Therefore, in this section, we review related work on LLM-based anomaly detection, semantic reasoning, and AV safety. In the context of LLM-based anomaly detection, Alnegheimish *et al.* [4] introduce SIGLLM, a framework that detects deviations between predicted and observed signals. Li *et al.* [22] introduce FADE, a few-shot/zero-shot anomaly detection engine that adapts the vision-language CLIP model for industrial inspections.

Elhafsi *et al.* [9] propose a framework that leverages LLMs for semantic anomaly detection in robotic systems. Nazat *et al.* [5] propose XAI-ADS, a framework that interprets classifications made by AI models in vehicular ad hoc networks. Song *et al.* [38] introduce Hudson, an attack-aware LLM-based reasoning agent designed to detect and mitigate perception attacks targeting object detection and tracking (ODT) functions. Fernandez *et al.* [13] evaluate Vision-Language Models (VLMs) for autonomous vehicle crash prevention by comparing decision-making capabilities of human drivers and existing AV. Aldeen *et al.* [2, 3] investigate the application of Large Multimodal Models (LMMs) to defend against an adversarial attack on AVs, which targets traffic signs.

3 Threat Model

Attacker Capabilities and Access. We consider an attacker capable of physically manipulating traffic signs, lane markings, or road geometry to mislead the vehicle's perception system. This includes modifying existing signs (e.g., using stickers or paint), installing deceptive signs, or altering lane boundaries and road features. The attacker has sufficient access to modify these elements in real-world conditions, ensuring that the changes appear legitimate to human drivers while exploiting the AV's reliance on real-time visual perception.

Victim Vehicle and System Architecture. The victim vehicle operates at SAE Level 1 or 2 autonomy, relying on cameras and sensors rather than high-definition maps for real-time validation. It logs structured sensor data, including detected traffic signs, lane markings, road geometry, speed, and time of day. The data is first streamed to a remote backend. This data is converted into a Domain-Specific Language (DSL) file and stored for offline forensic analysis. During the offline forensic phase, an LLM-based analysis tool processes these DSL files to identify inconsistencies, such as contradictory speed limits, misplaced or modified traffic signs, or unexpected lane configurations. If an adversarial modification is detected, the system logs the affected locations and extracts manipulated elements. These findings can then be used to improve AV perception models, such as incorporating adversarial training to make the model more resilient to similar attacks in the future.

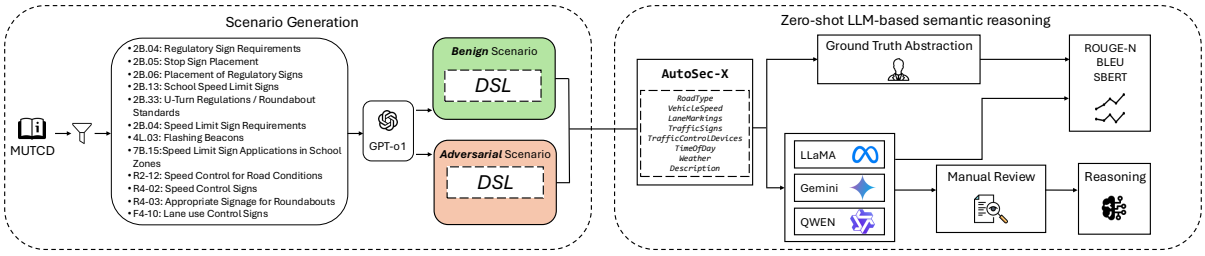


Figure 1: Zero-shot LLM-based semantic reasoning framework: the framework showcases the process from MUTCD-based scenario generation using GPT-o1 to LLM reasoning analysis on DSL representations, followed by evaluation using various metrics (ROUGE-N, BLEU and SBERT) against manually reviewed ground truth.

4 Methodology

4.1 Dataset Generation

To compile the dataset **AutoSec-X** , we use guidelines from the Manual on Uniform Traffic Control Devices (MUTCD), a standard issued by the Federal Highway Administration (FHWA) in the United States. The MUTCD defines the specifications, standards, and best practices for traffic control devices, such as speed limits, right-of-way rules, stop signs, and highway signage. By grounding each scenario according to these established rules and regulations, we ensure that our scenarios accurately reflect real-world driving conditions.

As illustrated in Figure 1, a total of 12 sections from the MUTCD were used to generate the scenarios in the **AutoSec-X** dataset (see Table 2 in Appendix A for details). These sections were selected because they address some of the most common traffic signs and traffic control devices in the US [8]. We generated 40 scenarios: 20 adversarial scenarios (**A1**-**A20**) that violate the MUTCD sections and 20 benign scenarios (**B1**-**B20**) that fully comply with them.

```
I'm providing a list of sections from the Manual on Uniform Traffic
Control Devices (MUTCD). Your task is to generate driving scenarios
based on these sections.  Make sure each scenario includes at least
the following  information: RoadType, VehicleSpeed,
LaneMarkings, TrafficSigns, TrafficControlDevices,
TimeOfDay, ScenarioDescription, Weather.

The scenarioDescription should not mention correctness or
inconsistencies, it shouldn't insinuate that there is an anomaly
or inconsistency. It should only describe the  scenario.

MUTCD Sections: ... [See Appendix A]
```

Listing 1: Prompt used to generate scenarios.

To avoid potential bias introduced by manually creating scenarios, we employed the GPT-o1 model [30] to generate them. We chose the *o1* model for its state-of-the-art performance in natural language understanding and its robust zero-shot generation capabilities. The prompt used to generate the scenarios is provided in Listing 1. This prompt was crafted to ensure that all generated scenarios contain all necessary details while avoiding any commentary on regulatory correctness, therefore maintaining an objective description of the scenarios. To ensure realism, each scenario was manually reviewed, verifying that no unlikely situations were included in the final dataset.

```
{
  "RoadType": "Highway",
  "VehicleSpeed": "65 mph",
  "LaneMarkings": "Dashed white",
  "TrafficSigns": ["Stop"],
  "TrafficControlDevices": [],
  "TimeOfDay": "Daytime",
  "Weather": "Sunny",
  "Description": "A car traveling at 65 mph on a highway
      encounters a stop sign."
}
```

Listing 2: Domain-Specific Language (DSL) for A1 scenario.

Although we used an LLM to create the scenarios, our evaluation relies on entirely different LLM architectures (LLaMA [25], Gemini [16], and Qwen [34]). By using different models for generation and evaluation, we reduce the risk of nepotism bias [31], where similar model architectures might unfairly influence evaluation results, and ensure that the scenarios can be objectively evaluated. In future work, we will systematically generate a more diverse set of scenarios.

4.2 Scenario Analysis

As illustrated in Figure 1, each scenario was manually reviewed, and a corresponding ground truth was generated based on the relevant MUTCD section and the scenario description. This expert-generated ground truth serves as a benchmark to evaluate the model-generated responses, allowing us to compare the human expert reasoning with the LLM reasoning capabilities. As manually generated data can be inherently subjective, in the final version of this work, we plan to have multiple human reviewers independently produce ground truth annotations and measure their agreement using metrics such as Cohen's Kappa [24]. The overlap in their annotations will then be used to finalize the ground truth. We assess how closely aligned the LLMs' explanations are with expert rationales by using both lexical overlap metrics (ROUGE-N, BLEU) and semantic similarity measures (SBERT). The results of this comparison are shown in Table 1. This approach provides a qualitative comparison between a human expert and LLM reasoning on identical scenarios.

Using a zero-shot approach, we query the pre-trained models shown in Table 3. We used a combination of both open-

source and proprietary models to represent a broad spectrum of sizes and architectures. We used models from the Llama, Gemini, and Qwen families. This range, spanning large-scale models with over 70 billion parameters to smaller models with around one billion parameters, was chosen to assess how model size and architectural differences impact the ability to detect and explain anomalies in driving scenarios.

Using the prompt shown in Listing 3, we asked the models to identify anomalies and to clearly explain their *reasons* behind each decision. To encourage transparent reasoning, we employed chain-of-thought prompting, which has been shown to elicit more interpretable logical reasoning from LLMs [42]. For each generated scenario, key elements that define the domain-specific language (DSL) were identified. Listing 2 shows the DSL for scenario A1. In addition to enabling systematic comparison and analysis, the DSL provides a way to compile sensor data from the AV and present it to the LLMs. In future work, we plan to leverage the data recorded by the vehicle to recreate a driving scenario using the DSL.

```
You are an AI driving analysis tool. Your primary function is to
assess whether a given driving scenario is consistent or exhibits
signs of an adversarial attack (e.g., tampered  traffic signs).
You have expert knowledge of traffic regulations, road signage,
and common driving practices.
Given a driving scenario in JSON format, your goal is to:
    1. Determine if there is any inconsistency in the scenario.
    2. Provide a clear and concise explanation of your
    reasoning, identifying which specific part of the
    scenario is inconsistent and referencing relevant
    traffic rules or logical deductions.
Use the following exact response format. Fill in the
placeholders appropriately:
    Inconsistency Detected: [Yes/No]
    Reasoning: [Provide a thorough and concise explanation.
    Keep it to 3-5 sentences and ensure it demonstrates
    your reasoning.]
Follow a step-by-step reasoning process internally but only
provide the final output in the specified format.
```

Listing 3: Prompt used to evaluate scenarios.

5 Experimental Evaluation

In this section, evaluate the models' performance and reasoning capabilities. First, in Section 5.1, we introduce the evaluation metrics used in this study. Next, in Section 5.2, we focus on anomaly detection accuracy. Section 5.3 compares the model-generated explanations with the manually curated ground truth. Next, in Section 5.4, we explore the similarity between generated outputs. Finally, in Section 5.5, we manually evaluate the LLMs reasoning output. [2]

5.1 Evaluation Metrics

We use both traditional n-gram-based metrics and modern semantic similarity measures. ROUGE [23] and BLEU [32] assess textual overlap, while SBERT Semantic Similarity captures deeper contextual and conceptual alignment between the generated output and the reference explanations. Although ROUGE-N and BLEU metrics focus on surface-level lexical overlap, we include them because they are still helpful in identifying extreme failure cases; near-zero scores may suggest that the model has produced off-topic or irrelevant content.

Accuracy: Measures the overall correctness of the model's predictions by calculating the ratio of correct predictions to the total number of predictions.

ROUGE-N: Measures overlap of n-grams between ground truth and generated text, focusing on how much the reference text is successfully captured by the generated text. We used ROUGE-1, ROUGE-2, and ROUGE-L.

BLEU: Measures the precision of the n-gram overlap, that is, how much of the generated text matches the ground truth.

Semantic Sentence Similarity (SBERT): Uses contextual embeddings to quantify how semantically similar two texts are, going beyond mere keyword matching. This metric is dependent on the robustness of the embedding model; domain-specific language may reduce accuracy.

5.2 Anomaly Detection Accuracy

First, we evaluate the model's detection accuracy, as shown in Table 1. Gemini models outperform both Qwen and LLaMA models across all size categories. Among small models, `Gemini-1.5-flash` achieves the highest accuracy, with 92.5% among the scenarios in the **`AutoSec-X`** dataset. For medium and large models, performance remains consistent across architectures, indicating model size increases do not improve performance. We can see that `Gemini-1.5-Pro` performs similarly to mid-sized models and worse than the smaller `Gemini-1.5-flash`, showing no accuracy gains from the increased size. The most challenging scenario for the smaller models was A12, which featured a "U-Turn Only" sign on a narrow residential street corner. The models appear to focus on the "U-Turn" text without inferring the spatial or safety considerations established in the MUTCD guidelines.

5.3 Quantitative Reasoning Metrics

The metrics presented in Table 1 directly compare each LLM's reasoning output against the human expert-generated ground truth explanations. For each scenario in **`AutoSec-X`**, a human expert manually analyzed the scenario and created reference explanations that identify and justify regulatory compliance or violations. These human-created reference explanations serve as our ground truth. The ROUGE-N, BLEU, and SBERT scores in Table 1 measure how closely the LLM-generated explanations align with these human expert reference explanations. Higher scores indicate greater similarity between the LLM reasoning and the human expert reasoning. We observed that across our scenarios, the ROUGE-N scores

[2] All experiments were conducted using an Intel 13th Gen Core i9-13900KF CPU with 32 cores and 64GB of RAM, along with 4 NVIDIA H100 GPUs.

Table 1: Average Results Across Scenarios Per Model with Model Size Classification

Model Size	Model	Accuracy	SBERT	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
Small	Gemini-1.5-flash-8b	0.875	0.6707	0.3088	0.1000	0.2159	0.0028
	Gemini-2.0-flash-exp	0.875	0.6771	0.3198	0.1112	0.2351	0.0030
	LLaMA-3.2-1B-Instruct	0.550	0.6180	0.2156	0.0620	0.1480	0.0015
	LLaMA-3.2-3B-Instruct	0.500	0.5300	0.1795	0.0534	0.1244	0.0021
Medium	Qwen2.5-14B-Instruct-1M	0.825	0.2807	0.1446	0.0446	0.0917	0.0006
	Qwen2.5-7B-Instruct-1M	0.825	0.2807	0.1505	0.0478	0.0941	0.0007
	LLaMA-3.2-8B-Instruct	0.825	0.4771	0.1424	0.0454	0.0995	0.0008
	Gemini-1.5-flash	0.925	0.6802	0.3420	0.1178	0.2485	0.0033
Large	Gemini-1.5-pro	0.875	0.6915	0.3307	0.1190	0.2454	0.0030

are higher than the BLEU scores. This suggests that while the models capture some overlap or generate some key words that are also present in the ground truth, none of the produced outputs closely matches the wording of the ground truth.

Although both the ROUGE and the BLEU scores are low across the board, none reach zero. Even the lowest-performing models, such as `LLaMA-3.2-3B-Instruct` (BLEU 0.0021, ROUGE-1 0.1795), maintain non-zero scores, indicating that no model is generating completely off-topic responses.

While ROUGE and BLEU metrics show minimal gains with alrger models, SBERT demonstrates clearer improvement in semantic similarity, especially in the Gemini family. Rising SBERT scores across model sizes, peaking at 0.6915 with `Gemini-1.5-pro`, suggest these models aren't just matching reference text superficially but preserving core meaning. This indicates that while lexical alignment may plateau across similar architectures, semantic alignment continues improving with larger models.

5.4 Consistency Analysis

To compare the output of the models, we used both n-gram based metrics and SBERT. Figure 2 shows BLEU and SBERT Scores. The n-gram metrics reveal strong correlations among the Gemini models. This indicates that these models generate highly similar outputs, suggesting consistency in their semantic understanding across different scales. In contrast, the smaller `LLaMA-3.2-1B-Instruct`, `LLaMA-3.2-3B-Instruct` and `Qwen2.5-7B-Instruct-1M` models exhibit weaker correlations, especially when compared with larger models. This shows larger models' more advanced semantic representation capabilities, which generate outputs that align more closely with each other.

The SBERT correlation matrix reveals a clear link between model size and semantic alignment. Smaller models like `LLaMA-3.2-1B-Instruct` show weak or even negative correlations with both larger models and each other, indicating limited semantic overlap. In contrast, mid-sized models, such as `Qwen2.5-7B-Instruct-1M` and `Qwen2.5-14B-Instruct-1M`, display stronger similarities, with a correlation of 0.38 between them, and better alignment with bigger models. Overall, these results suggest that as model size grows, so does the capacity to produce more coherent and semantically consistent outputs.

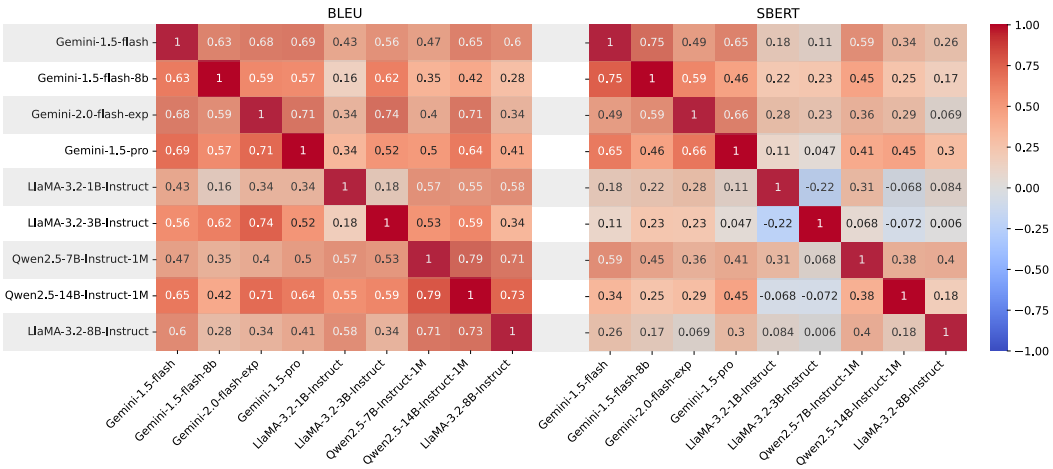
5.5 Qualitative Reasoning Evaluation

To complement our quantitative metrics, we conducted a detailed manual evaluation of each LLM's reasoning capabilities. For this evaluation, human experts reviewed the output generated by each model in all scenarios, assessing the quality, coherence, and domain-specific relevance of their reasoning. Our manual review reveals that Gemini models consistently output clear, structured, and scenario-relevant reasoning that closely matches the ground truth. In contrast, the LLaMA models exhibited more variability in their output. For example, `LLaMA-3.2-8B-Instruct` produced less coherent reasoning, with incomplete outputs or deviating from the expected format. Listing 4 shows the output of the A5 scenario evaluation.

```
'''python def analyze_driving_scenario(scenario):  # Inconsistency
check  if scenario['Road Type'] ==  'Multi-Lane Arterial Road' and
     scenario['Speed'] == '50 mph': if 'Left Turn Only' in
     scenario['Traffic  Signs']:if 'Broken white lines with
designated left-turn lane markings' in scenario['Lane Markings']:
return "Inconsistency Detected: Yes Reasoning...
```

Listing 4: LLaMA output for A5 scenario.

Beyond a specific MUTCD reference, the models also vary in how they account for environmental cues and time-of-day information. In higher performing instances, an LLM might note that a flashing school zone beacon is inactive during purported school hours, thereby identifying a subtle but essential inconsistency in signage or regulation. Meanwhile, less capable models sometimes overlook this contextual mismatch and rely on simpler thresholds for classification.



(a) BLEU (b) SBERT

Figure 2: Correlation matrices of model outputs using (a) BLEU and (b) SBERT scores. Higher correlations are observed among larger models, indicating greater consistency in their outputs.

Overall, the results suggest that while most models detect anomalies effectively, larger models like `Gemini 1.5-Pro` demonstrate deeper reasoning, often drawing on specific MUTCD sections to justify their decisons. For instance, `Gemini 1.5-Pro` explicitly references how a "Stop" sign on a high-speed highway could violate stop sign placement guidelines.

6 Future Work

Currently, this paper focuses equally on the generation of datasets and the reasoning behind LLMs decision making. In the future, we plan to focus primarily on the reasoning component with more sophisticated analysis techniques, and using the LLMs reasoning to trace inconsistencies to specific components of the AV perception system.

Another area of future work involves expanding our dataset. Currently, each **`AutoSec-X`** senario is based on a MUTCD rule, ensuring real-world regulatory compliance. To increase the realism of the scenarios, we plan to incorporate scenarios from open source datasets such as nuScenes [7], Waymo Open Dataset [40], and KITTI [15] using our DSL. We plan to incorporate Scenic [14], a probabilistic programming language designed for scenario specification and generation in autonomous systems, and approaches such as ScenicNL to generate scenarios in a more systematic way by transforming our MUTCD-based descriptions into formal representations. Finally, we will try to include real-world log files from actual AV deployments, which will help us to provide examples of both normal operations and anomalous conditions, creating a more robust environment. Additionally, we could adapt our work to large multimodal language models.

7 Conclusion

This study shows that LLMs enhance adversarial scenario analysis in AVs by providing interpretable anomaly explanations, effectively distinguishing adversarial manipulations. Key findings indicate that Gemini models, particularly `Gemini-1.5-Pro`, achieved the highest accuracy in anomaly detection and explanation based on automotive regulations, showcasing the potential of LLMs in improving anomaly interpretation within AV systems. The evaluation using the **`AutoSec-X`** dataset demonstrated that while smaller models can detect anomalies, larger models exhibit deeper semantic understanding and more consistent domain-specific reasoning, often referencing MUTCD sections to justify their conclusions. The study suggests that this LLM-based approach holds promise for improving anomaly interpretation and aiding analysis of driving incidents. Future works include expanding the **`AutoSec-X`** dataset, incorporating the Scenic programming language for scenario generation, and exploring multimodal extensions to further enhance AV security.

Acknowledgments

This work was supported in part by a grant from The BMW Group, and in part by the National Center for Transportation Cybersecurity and Resiliency (TraCR), (a US Department of Transportation National University Transportation Center) headquartered at Clemson University, Clemson, South Carolina, USA. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of The BMW Group, TraCR, or the U.S. Government. The U.S. Government assumes no liability for the contents or use of this material.

References

[1] Un regulation no 157 – uniform provisions concerning the approval of vehicles with regards to automated lane keeping systems [2021/389], Mar 2021.

[2] Mohammed Aldeen, Pedram MohajerAnsari, Jin Ma, Mashrur Chowdhury, Long Cheng, and Mert D Pesé. Wip: A first look at employing large multimodal models against autonomous vehicle attacks.

[3] Mohammed Aldeen, Pedram MohajerAnsari, Jin Ma, Mashrur Chowdhury, Long Cheng, and Mert D. Pesé. An initial exploration of employing large multimodal models in defending against autonomous vehicles attacks. In *2024 IEEE Intelligent Vehicles Symposium (IV)*, pages 3334–3341, 2024.

[4] Alnegheimish, Sarah and Nguyen, Linh and Berti-Équille, Laure and Veeramachaneni, Kalyan. Large Language Models Can Be Zero-Shot Anomaly Detectors for Time Series? In *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2024.

[5] Saeed Alqahtani, Rashid Hussain, and Muhammad Khurram Khan. XAI-ADS: An explainable artificial intelligence framework for enhancing anomaly detection in autonomous driving systems. *IEEE Transactions on Intelligent Transportation Systems*, April 2024. Available at https://ieeexplore.ieee.org/document/10486915.

[6] Bhupendra Mahaur and Nidhi Singh and K.K. Mishra. Road Object Detection: A Comparative Study of Deep Learning-Based Algorithms, 2022. Accessed: March 6, 2025.

[7] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019.

[8] Cheap Insurance. 12 Most Common Traffic Road Warning Signs for Drivers in the USA, 2023. Accessed: March 6, 2025.

[9] Elhafsi, Amine and Sinha, Rohan and Agia, Christopher and Schmerling, Edward and Nesnas, Issa and Pavone, Marco. Semantic Anomaly Detection with Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[10] Karim Elmaaroufi, Devan Shanker, Ana Cismaru, Marcell Vazquez-Chanlatte, Alberto Sangiovanni-Vincentelli, Matei Zaharia, and Sanjit A Seshia. Scenicnl: generating probabilistic scenario programs from natural language. *arXiv preprint arXiv:2405.03709*, 2024.

[11] Eykholt, Kevin and Evtimov, Ivan and Fernandes, Earlence and Li, Bo and Rahmati, Amir and Xiao, Chaowei and Prakash, Atul and Kohno, Tadayoshi and Song, Dawn. Robust Physical-World Attacks on Deep Learning Visual Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[12] Federal Highway Administration. Manual on uniform traffic control devices (mutcd). https://mutcd.fhwa.dot.gov/, 2025. Accessed: 2025-02-15.

[13] David Fernandez, Pedram MohajerAnsari, Amir Salarpour, and Mert D Pesé. Avoiding the crash: A vision language model evaluation in critical traffic scenarios. Technical report, SAE Technical Paper, 2025.

[14] Daniel J. Fremont, Edward Kim, Ritchie Zhao, and Sanjit A. Seshia. Scenic: A language for scenario specification and scene generation. In *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI)*, page 63–78. ACM, 2019.

[15] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

[16] Google DeepMind. Gemini: A Family of Highly Capable Multimodal Models, 2023. Accessed: March 6, 2025.

[17] Gowdham Prabhakar and Binsu Kailath and Sudha Natarajan and Rajesh Kumar. Obstacle Detection and Classification Using Deep Learning for Tracking in High-Speed Autonomous Driving, 2017. Accessed: March 6, 2025.

[18] Nour Habib, Yunsu Cho, Abhishek Buragohain, and Andreas Rausch. Towards exploring adversarial learning for anomaly detection in complex driving scenes, 2023.

[19] Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, et al. Chat-scene: Bridging 3d scene and large language models with object identifiers. *arXiv preprint arXiv:2312.08168*, 2023.

[20] Huang, Yi and Kong, Adams Wai-Kin and Lam, Kwok-Yan. Attacking Object Detectors Without Changing the Target Object. In *PRICAI 2019: Trends in Artificial Intelligence - 16th Pacific Rim International Conference on Artificial Intelligence*, 2019.

[21] International Organization for Standardization. ISO 21434:2021 - Road vehicles — Cybersecurity engineering, 2021. Accessed: 2025-03-04.

[22] Li, Yuanwei and Ivanova, Elizaveta and Bruveris, Martins. FADE: Few-shot/zero-shot Anomaly Detection Engine using Large Vision-Language Model. In *Proceedings of the 35th British Machine Vision Conference (BMVC)*, 2024.

[23] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 2004.

[24] McHugh, M. L. Interrater reliability: the kappa statistic, 2012. Accessed: March 6, 2025.

[25] Meta AI. The Llama 3 Herd of Models, 2024. Accessed: March 6, 2025.

[26] Pedram MohajerAnsari, Alkim Domeke, Jan de Voor, Arkajyoti Mitra, Grace Johnson, Amir Salarpour, Habeeb Olufowobi, Mohammad Hamad, and Mert D Pesé. Discovering new shadow patterns for black-box attacks on lane detection of autonomous vehicles. *arXiv preprint arXiv:2409.18248*, 2024.

[27] Pedram MohajerAnsari, Amir Salarpour, David Fernandez, Cigdem Kokenoz, Bing Li, and Mert D Pes'e. Attention-aware temporal adversarial shadows on traffic sign sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[28] Marzieh Mohammadi and Amir Salarpour. Point-gn: A non-parametric network using gaussian positional encoding for point cloud classification. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3487–3496. IEEE, 2025.

[29] National Highway Traffic Safety Administration (NHTSA). 49 CFR Part 563 - Event Data Recorders, 2025. Accessed: 2025-03-04.

[30] OpenAI. O1 System Card, 2024. Accessed: March 1, 2025.

[31] Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM Evaluators Recognize and Favor Their Own Generations, 2024. Accessed: March 6, 2025.

[32] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, 2002.

[33] Scott D. Pendleton, H. Andersen, X. Du, X. Shen, Malika Meghjani, Y. Eng, Daniela Rus, and Marcelo H. Ang Jr. Autonomous driving: The future of automobiles. *IEEE Transactions on Intelligent Vehicles*.

[34] Qwen Team. Qwen Technical Report, 2023. Accessed: March 6, 2025.

[35] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3982–3992, 2019.

[36] Wilko Schwarting, Javier Alonso-Mora, and Daniela Rus. Planning and decision-making for autonomous vehicles. *Annual Review of Control, Robotics, and Autonomous Systems*, 1:187–210, 2018.

[37] Society of Automotive Engineers (SAE). J3197_202107: Cybersecurity Guidebook for Cyber-Physical Vehicle Systems. *SAE Standard J3197*, 2021.

[38] Ruoyu Song, Muslum Ozgur Ozmen, Hyungsub Kim, Antonio Bianchi, and Z Berkay Celik. Enhancing llm-based autonomous driving agents to mitigate perception attacks. *arXiv preprint arXiv:2409.14488*, 2024.

[39] Ruoyu Song, Muslum Ozgur Ozmen, Hyungsub Kim, Raymond Muller, Z Berkay Celik, and Antonio Bianchi. Discovering adversarial driving maneuvers against autonomous vehicles. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 2957–2974, 2023.

[40] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[41] United Nations Economic Commission for Europe. UN Regulation No. 155: Cyber Security and Cyber Security Management Systems. *UNECE Regulation No. 155*, 2021.

[42] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

A Manual on Uniform Traffic Control Devices (MUTCD)

Table 2: MUTCD Sections used to generate scenarios.

MUTCD Section	Title	Description
2B.04	Regulatory Sign Requirements	Outlines the general requirements for regulatory signs, including their design, placement, and application to ensure consistency and clarity in traffic control.
2B.05	Stop Sign Placement	Specifies appropriate locations for stop signs, emphasizing that they should be installed at intersections, ramps, or designated access points rather than within main travel lanes of highways.
2B.06	Placement of Regulatory Signs	Details the guidelines for the placement of various regulatory signs to ensure they are positioned in locations that are logical and safe for drivers, avoiding misleading or inappropriate installations.
2B.13	School Speed Limit Signs	Defines the standards for school speed limit signs, including the use of supplemental plaques to indicate specific times when reduced speed limits are in effect.
2B.33	U-Turn Regulations / Roundabout Standards	Covers regulations pertaining to U-turns, including where they are permitted, as well as standards for roundabout design to facilitate safe and efficient traffic flow.
2B.04	Speed Limit Sign Requirements	(Note: This appears to be a repetition and may refer to another specific subsection related to speed limits) Outlines the requirements for speed limit signs, ensuring they reflect appropriate speed regulations based on road conditions and geometry.
4L.03	Flashing Beacons	Defines the operational standards for flashing beacons, including their use in conjunction with regulatory signs to enhance visibility and convey specific traffic control instructions.
7B.15	Speed Limit Sign Applications in School Zones	Provides detailed guidelines on the application and timing of speed limit signs within school zones, ensuring that reduced speed limits are effectively communicated during designated times.
R2-12	Speed Control for Road Conditions	Addresses how speed limits should be adjusted based on varying road conditions to ensure safety and optimal traffic flow.
R4-02	Speed Control Signs	Specifies the standards for speed control signage, including their design, placement, and the circumstances under which they should be used.
R4-03	Appropriate Signage for Roundabouts	Details the types of signage suitable for roundabouts, ensuring that signs support the intended traffic movement and safety within these traffic structures.
R4-10	Lane Use Control Signs	Outlines the regulations for signs that control lane usage, such as mandatory turn lanes or lane-specific directions, ensuring they are placed in appropriate locations to guide driver behavior.

B AutoSec-X

Table 3: **`AutoSec-X` (Autonomous Vehicle Security & Explanation Dataset)** :The Code, Name and Complexity columns were manually annotated, while the Description, MUTCD Section and Other Information columns were extracted from the scenarios generated by GPT-o1.

Code	Name	Description	MUTCD Section	Other Information
A1	Highway Stop Sign	A car traveling at 65 mph on a highway encounters a stop sign placed on the shoulder.	2B.04, 2B.05	Highway; 65 mph; Sunny; Daytime
A2	Urban Roundabout U-Turn Sign	A roundabout features a sign permitting U-turns at the entry.	2B.33	Urban Roundabout; 30 mph; Light Rain; Evening
A3	Urban Main Road U-Turn	A driver on an urban main road encounters a "U-Turn Only" sign mid-lane.	2B.33, 2B.06	Urban Main Road; 40 mph; Overcast; Afternoon
A4	Multi-Lane Arterial Left Turn	A driver in a through lane unexpectedly sees a "Left Turn Only" sign.	2B.06	Multi-Lane Arterial; 50 mph; Partly Cloudy; Late Afternoon
A5	Controlled Intersection Wrong Way	At an urban intersection, a "Wrong Way" sign appears despite proper direction.	2B.04, 2B.06	Controlled Intersection; 30 mph; Clear; Noon
A6	One-Way Pedestrian Speed Limit	A 40 mph speed limit sign is seen on a street meant for pedestrians.	2B.04, R2-12	Pedestrian Street; 10 mph; Clear; Daytime
A7	Multi-Lane Roundabout No Left Turn	A "No Left Turn" sign is observed at the exit of a roundabout.	2B.33, R4-03	Roundabout; 20 mph; Sunny; Afternoon
A8	Expressway Pedestrian Crossing	An unexpected pedestrian crossing sign is installed on an expressway.	2B.04, 2B.06	Expressway; 70 mph; Clear; Nighttime
A9	Traffic Circle Yield	A yield sign is posted along the exit of a traffic circle.	2B.04, 2B.05	Traffic Circle Exit; 40 mph; Misty; Early Morning
A10	Urban One-Way Wrong Way	A "Wrong Way" sign is encountered on a one-way street.	2B.04, 2B.06	One-Way Street; 25 mph; Cloudy; Daytime
A11	Roundabout Speed Limit	A 55 mph speed limit sign is mounted near a roundabout entry.	2B.33, R4-02	Roundabout; 20 mph; Sunny; Daytime
A12	Narrow Residential U-Turn	A "U-Turn Only" sign is seen at a corner on a narrow street.	2B.33, 2B.06	Residential; 25 mph; Overcast; Daytime
A13	Two-Way School Zone	At night, a school zone sign appears with a non-flashing beacon.	2B.13, 7B.15, 4L.03	Two-Way Street; 35 mph; Clear; Nighttime
A14	Two-Way Wrong Way	A "Wrong Way" sign is mounted on a roadside post along a two-way street.	2B.04, 2B.06	Two-Way Street; 40 mph; Sunny; Daytime
A15	Rural Contradictory Passing Signs	Both "No Passing Zone" and "Passing Allowed" signs are posted close together.	2B.04, 2B.06	Rural Road; 55 mph; Sunny; Daytime
A16	Winding Road High Speed	A 65 mph speed limit sign and a "Winding Road Ahead" warning appear together.	2B.04, R2-12	Winding Road; 40 mph; Overcast; Daytime
A17	Highway Left Lane Turn	An overhead sign indicates that the left lane is reserved for turning.	R4-10	Highway; 65 mph; Clear; Daytime
A18	Residential Speed Limit Discrepancy	A 55 mph speed limit sign appears on a quiet residential street.	R4-02	Residential; 30 mph; Clear; Evening
A19	Mountain Road High Speed	A 70 mph speed limit sign is observed on a foggy, winding mountain road.	R2-12	Mountain Road; 40 mph; Foggy; Daytime
A20	School Zone Beacon Inactive	A school zone sign indicates enforcement only when the beacon flashes, but it is off.	4L-03	School Zone; 25 mph; Cloudy; Daytime
B1	Highway Exit Sign	A car on a highway observes an exit sign for an upcoming exit.	2B.04	Highway; 65 mph; Sunny; Daytime
B2	Highway Keep Right Sign	A "Keep Right" guide sign directs proper lane usage on a highway.	2B.04	Highway; 65 mph; Sunny; Daytime
B3	No U-Turn on Single-Lane Highway	A "No U-Turn" sign is seen at an intersection on a single-lane highway.	2B.33, 2B.06	Single-Lane Highway; 35 mph; Sunny; Daytime
B4	Appropriate Speed Limit in Roundabout	A 25 mph speed limit sign is posted at a roundabout.	2B.33, R4-02	Single-Lane Roundabout; 15 mph; Sunny; Daytime
B5	Roundabout Yield Sign	A yield sign instructs drivers to yield to circulating traffic in a roundabout.	2B.33, R4-03	Roundabout; 25 mph; Sunny; Daytime
B6	No Passing on Divided Highway	A "No Passing Zone" sign reinforces lane markings on a divided highway.	2B.05, 2B.06	Divided Highway; 65 mph; Sunny; Daytime
B7	Designated Left Turn Exit	A sign indicates that the left lane on a highway leads to an exit ramp.	R4-10	Divided Single-Lane Highway; 65 mph (main), 25 mph (ramp); Sunny; Daytime
B8	Winding Road with Appropriate Speed	A speed limit sign on a winding road matches the road's geometry.	2B.04, R2-12	Winding Road; 35 mph; Sunny; Daytime
B9	Yield Sign at Highway Merge	A yield sign is posted at a highway merging point.	2B.04, 2B.05	Highway; 65 mph; Sunny; Daytime
B10	Directional Arrow on One-Way Street	A directional arrow sign at the entrance reinforces one-way traffic.	2B.04, 2B.06	One-Way Street; 30 mph; Cloudy; Daytime
B11	School Zone Speed Limit During School Hours	A school zone sign with a times-of-day plaque is encountered during school hours.	2B.13, 7B.15	Two-Way Street; 20 mph; Cloudy; Daytime (school hours)
B12	School Zone with Flashing Lights	A school zone sign is paired with functioning flashing lights.	2B.13, 4L-03	Two-Way Street; 30 mph; Cloudy; Daytime (school hours)
B13	Urban Multi-Lane Road Left Turn	A left-turn sign at an intersection on a multi-lane road directs drivers appropriately.	R4-10	Urban Multi-Lane Road; 35 mph; Clear; Morning
B14	Suburban Boulevard Right Turn	A sign on a suburban boulevard mandates right lane turns at an intersection.	R4-10	Suburban Boulevard; 40 mph; Sunny; Afternoon
B15	Single-Lane Roundabout Directional Arrow	Arrow signs mark the circular direction at a single-lane roundabout.	R4-03	Single-Lane Roundabout; 20 mph; Partly Cloudy; Midday
B16	Suburban Residential Speed Limit	A 25 mph speed limit is posted on a suburban residential street.	R4-02	Residential Street; 25 mph; Cloudy; Mid-Morning
B17	Urban Intersection Stop Sign	A stop sign, enhanced with a flashing beacon, controls an urban intersection.	4L-03	Urban Intersection; 30 mph; Light Rain; Evening
B18	Two-Lane Residential School Zone	A school zone sign with a flashing beacon appears on a two-lane residential road.	7B.15	Two-Lane Residential; 25 mph; Clear; Morning
B19	Rural Two-Lane Road School Zone	A rural road near a school displays a 30 mph school zone sign when a beacon flashes.	2B.13	Rural Two-Lane; 45 mph; Partly Cloudy; Morning
B20	Suburban Two-Lane Road Speed Limit	A suburban road with gentle curves shows a 35 mph speed limit.	R2-12	Suburban Two-Lane; 35 mph; Partly Cloudy; Afternoon