KGs-Augmented Testsuite Generator Framework

Author: Mr. Chaipat Hainan
Advisor: Assistant Professor Dr. Areerat Trongratsameethong
Mentor: Associate Professor Dr. Natthawut Kertkeidkachorn

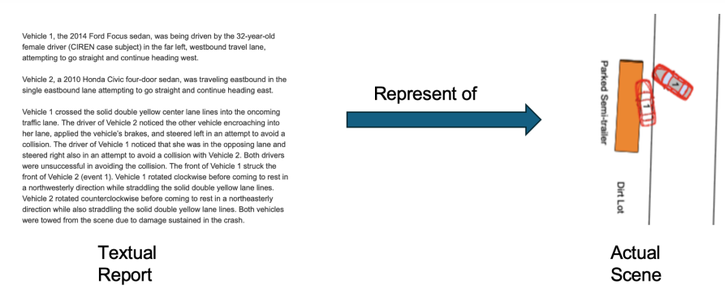
1 Abstract

Ensuring the safety of Autonomous Driving Systems (ADS) necessitates rigorous testing using simulated scenarios derived from real-world accident reports. While Large Language Models (LLMs) offer a promising approach for generating these scenarios, their direct application in discovering elusive edge-cases is often inefficient and unreliable due to inconsistent source reports and the generation of irrelevant scenarios. To address this, this study proposes a **KGs-Augmented Testsuite Generator Framework** that employs a schema-guided LLM to accurately extract structured accident data from unstructured reports. This data is modeled as a Knowledge Graph (KG), serving as a semantic backbone to infer missing information and maintain causal continuity. By integrating Operational Design Domain (ODD) constraints from JAMA directly into the KG, the framework focuses on scenarios that push boundary limits, significantly reducing reconstruction efforts needed to discover challenging edge-cases while remaining compliant with industry standards like ASAM OpenSCENARIO

2 Introduction

Safety testing of autonomous vehicles (AVs) is a top priority, yet converting real-world accident reports into simulation scenarios remains a significant challenge.

- **Narrative Complexity:** Reports are often unstructured narratives, making automatic machine-understandable conversion difficult and error-prone.
- **Irrelevant Scenarios**: Direct LLM application can generate numerous irrelevant cases, leading to inefficient reconstruction efforts before finding a novel edge-case.
- **Fidelity Gap:** Automatically generated scenarios frequently lack realism and causal continuity.

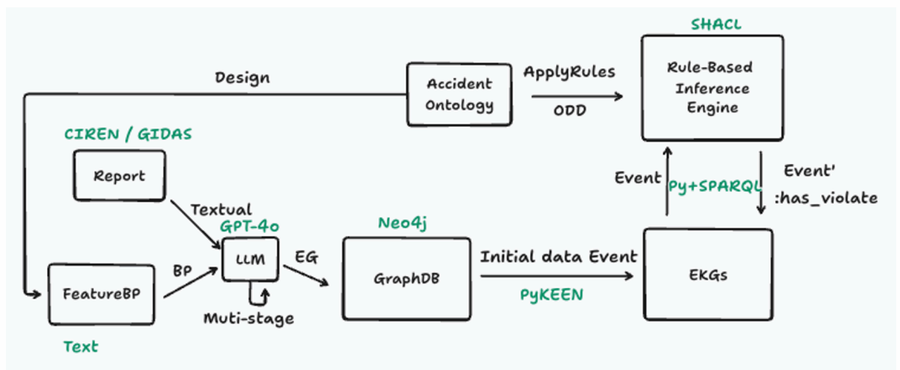


Difficulty of Transformation: Converting narratives into machine-understandable simulation scenarios automatically is challenging and error-prone.

3 Methodology

The research introduces the “KGs-Augmented Re-construct Scenario” pipeline:

- **Schema-Guided Extraction:** Uses a Large Language Model (LLM) to extract structured data from reports.
- **Divide-and-Conquer (DaC):** Decomposes complex extraction into smaller, parallel subproblems (World, Actor, and Sequence) to reduce cognitive overload.
- **Knowledge Graph (KG) Integration**: Aggregates data in a centralized KG to infer missing info and ensure causal continuity



Scenario Evolution Knowledge Graph (SEKG)

4 Technology

- **Large Language Models (LLMs):** Utilized for natural language understanding and multi-stage feature extraction. Employs a "Divide-and-Conquer" prompting technique to accurately parse complex accident narratives without cognitive overload or hallucination.
- **Neo4j:** The core semantic backbone of the framework. Used to store structured Entity-Event-Relationship data using Graph Databases enabling rule-based inference and logical consistency checks for missing data.
- **CARLA:** The target execution environment. The framework automatically translates the extracted Knowledge Graph scenarios into executable, machine-understandable scripts for 3D physics-based simulation.
- **Real-World Crash Databases:** Sources unstructured narrative data to ensure realistic edge-case generation (e.g., CIREN, GIDAS, NAIS).

5 Result

1. Scenario Generation Output

- **High-Volume Creation:** The framework successfully automated the transition from unstructured text to executable simulation scripts.
- **Diversity:** Generated **[318]** distinct, runnable test cases distributed across [21] different Modular ODD (Operational Design Domain) categories.

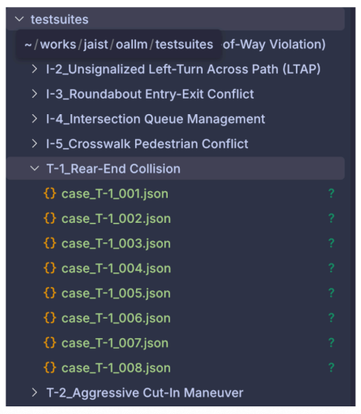
2. Fidelity & Realism Assessment

To ensure the generated simulations accurately reflected the physics and logic of the real-world accidents, the framework was evaluated using two primary metrics:

- **Scene Graph Comparison (Realism Score / R_Score):** We compared the semantic graphs of the generated test suites against the baseline Knowledge Graphs of the original accident reports. This measured how accurately the environmental conditions and actor behaviors were preserved.
- **Scenario Completeness Score:** Evaluated the functional quality of the generated scripts by verifying that all mandatory simulation components (e.g., valid coordinates, ego-vehicle instantiation, weather parameters) were present without logical conflicts.

3. Key Findings & Analysis

- **Highest Accuracy:** The system achieved the highest **R_Score** in structurally straightforward scenarios, such as **[E-6: Road Obstacles / Debris]**, where environmental variables and actor intents are highly explicit in the text.
- **Areas for Improvement:** Lower R_Scores were observed in complex, multi-actor sequential events, such as **[T-1: Rear-End Collisions],** indicating that highly dynamic spatial-temporal relationships occasionally challenge the LLM's spatial reasoning during extraction.



Generated Test case

### 1. Traffic Maneuvers (T) - High Frequency		
ODD Modular Group	Percentage (%)	Case Count (out of 318)
I-1: Signalized Conflict	6.89	22
I-2: Unsignalized Left-Turn	7.28	23
I-3: Roundabout Conflict	4.59	15
I-4: Intersection Queue Mgmt	6.3	20
I-5: Crosswalk Pedestrian Conflict	11.19	35
I-6: Four-Way Stop Misinterpretation	3.58	11
I-7: Grade/Curvature Intersection	5.93	19

### 2. Intersections / Junctions (I) - High Complexity		
ODD Modular Group	Percentage (%)	Case Count (out of 318)
T-1: Rear-End Collision	4.69	15
T-2: Aggressive Cut-In	2.68	9
T-3: Unsafe Lane Change	6.01	19
T-4: Following Too Closely	3.19	10
T-5: Wrong-Way Driving	4.28	14
T-6: Unexpected Lane Departure	2.98	9
T-7: Cyclist/Motorcyclist Close Pass	6.37	20

Sample Case group by OOD Modulars

ODD_Group_ID	Total_Cases	Avg_Realism_Score	Min_Realism_Score
T-1	8	0.9577	0.7737
E-1	5	0.9389	0.837
T-4	7	0.8814	0.7063
E-4	12	0.7888	0.633
I-1	31	0.7886	0.4425
E-2	10	0.7861	0.5477
I-6	20	0.7624	0.5477
E-3	18	0.744	0.5357
T-5	20	0.7414	0.4905
T-2	8	0.74	0.6073
E-5	9	0.7337	0.5705
E-7	17	0.7282	0.5111
I-3	9	0.7265	0.553
I-5	20	0.7239	0.5051
I-4	20	0.7143	0.4572
I-7	22	0.7114	0.5354
T-6	26	0.7065	0.4583
T-7	16	0.7018	0.579
T-3	15	0.6825	0.5146
I-2	13	0.6751	0.4391
E-6	10	0.6305	0.4258

Worst R_score — T-1_Rear-End Collision
Best R_score — E-6_Road Obstacle - Debris

R_score based on ODD group

6 Conclusion

We propose a Testsuite Generator Framework that uses schema-guided Large Language Models (LLMs) and Knowledge Graphs (KGs) to successfully construct realistic, Operational Design Domain (ODD)-based test scenarios from real-world accident reports. This approach enables highly reliable safety validation for autonomous vehicles by logically inferring missing information from text narratives and accurately assessing Automated Driving System (ADS) performance, especially in safety-critical edge cases.

References
[1] Zhou, H., et al. "Analysis of Traffic Accident Based on Knowledge Graph," Journal of Advanced Transportation, 2022.
[2] Bagschik G, Menzel T, Maurer M. "Ontology based Scene Creation for the Development of Automated Vehicles," IEEE ITSC, 2018.
[3] Khot T., et al. "Prompting Large Language Models with Divide-and-Conquer Program for Discerning Problem Solving," arXiv, 2024.
[4] National Highway Traffic Safety Administration (NHTSA). Crash Injury Research and Engineering Network (CIREN) Database.
[5] German In-Depth Accident Study (GIDAS).

Co-operative Education, 204496, 2025
Department of Computer Science, Faculty of Science, Chiang Mai University