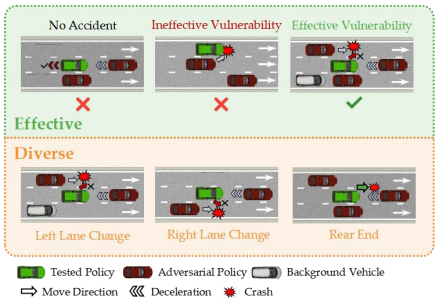
AED: Automatic Discovery of Effective and Diverse Vulnerabilities for Autonomous Driving Policy with Large Language Models

Le Qiu*, Zelai Xu*, Qixin Tan*, Wenhao Tang, Chao Yu†, Yu Wang†

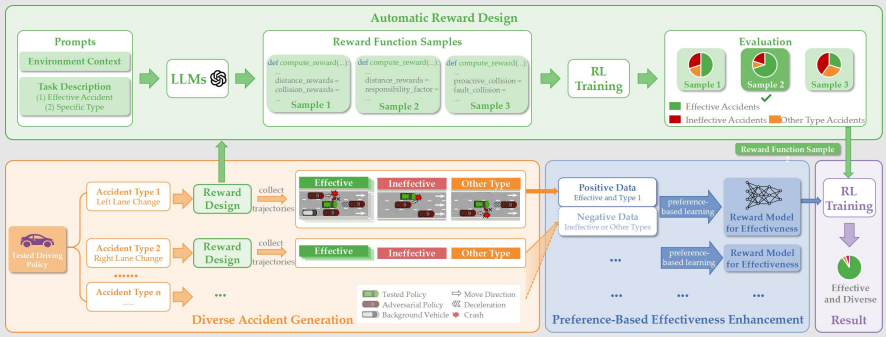
*Equal contribution, †Equal advising

The Problem



It's challenging to *automatically* discover RL-based vulnerabilities [1] that are both *effective* and *diverse*.

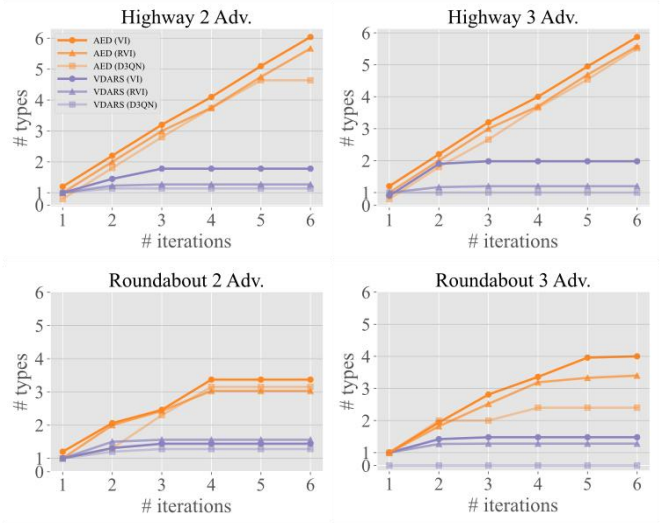
Our Solution: AED



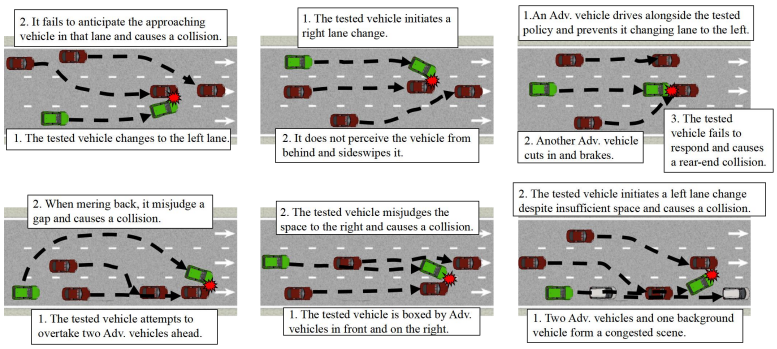
Idea: We leverage the in-context learning and code-synthesis capabilities of Large Language Models (LLMs) to automate reward design.
Approach: Given environment descriptions, LLMs synthesize reward functions [2] for diverse accident types. A preference-based reward learning module filters out noisy LLMs-generated rewards to improve effectiveness.

Evaluation of Diversity

AED consistently discovers a broader set of distinct vulnerability types than VDARS across different traffic environments.

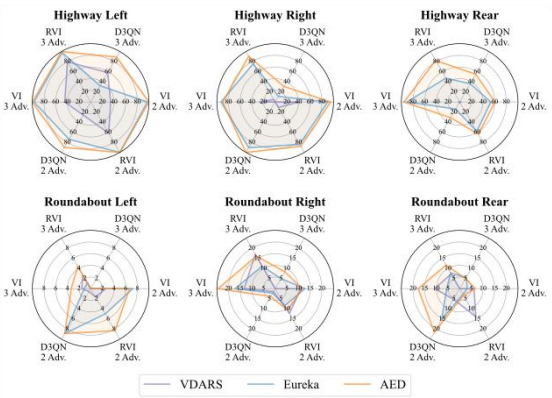


Qualitative examples of distinct vulnerability types discovered by AED in the Highway environment.



Evaluation of Effectiveness

AED consistently achieves the highest effective vulnerability rates across different traffic environments.



Key takeaways

- We propose an LLMs-based framework that uses reinforcement learning to *automatically* discover *effective* and *diverse* vulnerabilities in autonomous driving policies.
- We demonstrate the preference-based reward learning can denoise LLMs-generated rewards for autonomous driving safety evaluation.

References

[1] Mu, Ye, et al. "Multi-agent vulnerability discovery for autonomous driving policy by finding av-responsible scenarios." 2024 IEEE 20th International Conference on Automation Science and Engineering (CASE). IEEE, 2024..
[2] Ma, Yecheng Jason, et al. "Eureka: Human-level reward design via coding large language models." arXiv preprint arXiv:2310.12931 (2023).