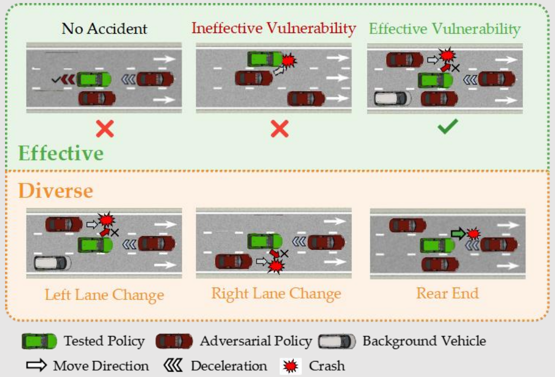
AED: Automatic Discovery of Effective and Diverse Vulnerabilities for Autonomous Driving Policy with Large Language Models

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

*Equal contribution, †Equal advising

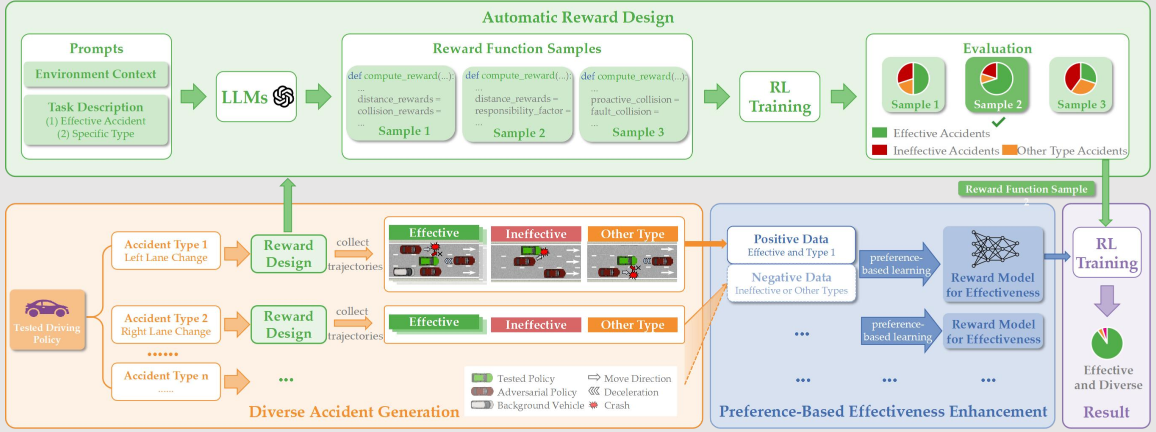
The Problem



It's challenging to *automatically* discover *effective* and *diverse* vulnerabilities in automatic driving policies.

Our Solution: AED

Leverage Large Language Models (LLMs) to automate reward design and preference-based learning to improve effectiveness



Automatic Reward Design

Prompts
Environment Context
Task Description
(1) Effective Accident
(2) Specific Type

LLMs

Reward Function Samples

```
def compute_reward(...):
  ...
  distance_rewards =
  collision_rewards =
  ...
```

Sample 1

```
def compute_reward(...):
  ...
  distance_rewards =
  responsibility_factor =
  ...
```

Sample 2

```
def compute_reward(...):
  ...
  proactive_collision =
  fault_collision =
  ...
```

Sample 3

RL Training

Evaluation
Sample 1
Sample 2 ✓
Sample 3
🟩 Effective Accidents
🟥 Ineffective Accidents 🟧 Other Type Accidents

Reward Function Sample 2

Diverse Accident Generation

Tested Driving Policy

Accident Type 1
Left Lane Change

Accident Type 2
Right Lane Change

......

Accident Type n
......

Reward Design
collect trajectories
Effective | Ineffective | Other Type

Reward Design
collect trajectories
Effective | Ineffective | Other Type

...

Tested Policy
Adversarial Policy
Background Vehicle
Move Direction
Deceleration
Crash

Preference-Based Effectiveness Enhancement

Positive Data
Effective and Type 1

Negative Data
Ineffective or Other Types

preference-based learning
Reward Model for Effectiveness

...
preference-based learning
Reward Model for Effectiveness

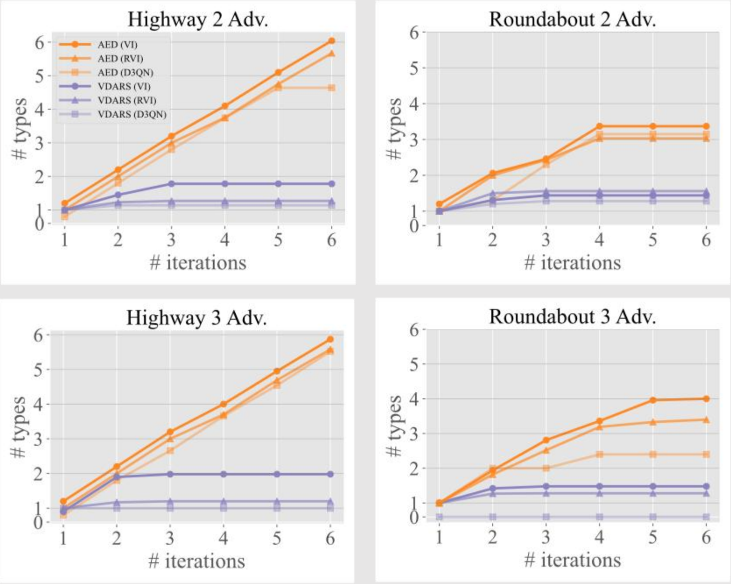
... ...

RL Training

Effective and Diverse

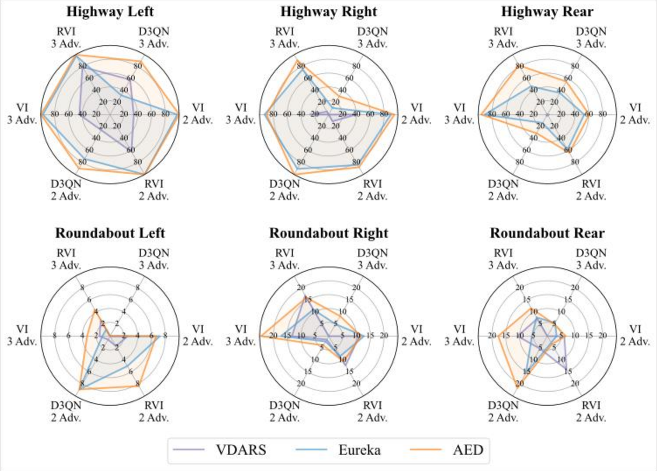
Result

Evaluation of Diversity



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

Evaluation of Effectiveness



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