GPTLens: an adversarial two-stage detection framework

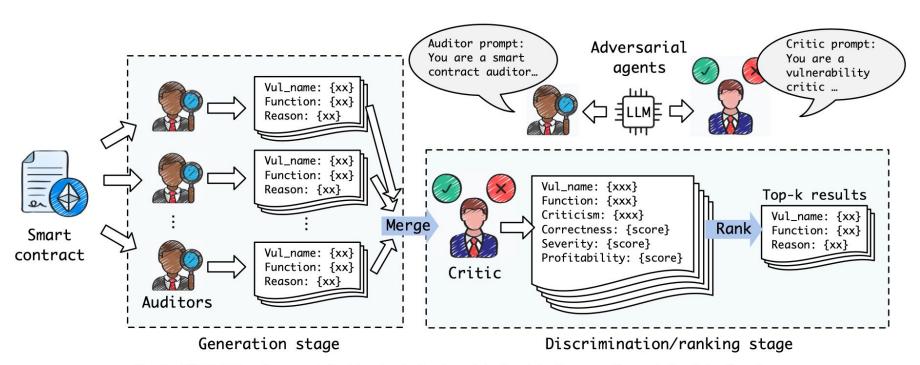


Fig. 2. GPTLENS: a framework that breaks single-stage detection into the generation and discrimination stages.

GPTLens: an adversarial two-stage detection framework

Goal of Generation: Increase the probability of the correct answer being generated. Goal of Discrimination: Reduce the number of false Vul_name: {xx} Function: {xx} positives introduced by generating more answers. Reason: {xx} Vul_name: {xx} | Function: {xx} Vul_name: {xxx} Reason: {xx} Top-k results Function: {xxx} Criticism: {xxx} Vul_name: {xx} Merge Rank Correctness: {score} Function: {xx} Smart Severity: {score} Reason: {xx} Vul_name: {xx} contract Profitability: {score} Function: {xx} Critic Reason: {xx} **Auditors**

Fig. 2. GPTLENS: a framework that breaks single-stage detection into the generation and discrimination stages.

Discrimination/ranking stage

Generation stage

Stage 1: Generation

The LLM plays the role of the **auditor** agent.

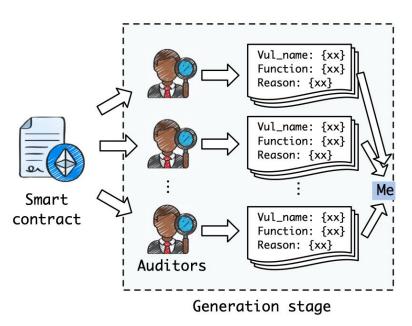


Fig. 2. GPTLENS: a framework that breaks si

Goal: Increase the probability of the correct answer being generated.

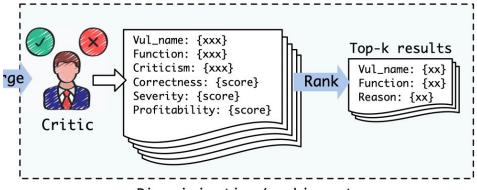
Multiple auditors and each identify multiple (k) vulnerabilities with high randomness.

Auditors not only provide the vulnerability name, but also the intermediate reasoning.

Stage 2: Discrimination

The LLM plays the role of the **critic** agent.

Goal: Reduce the number of false positives.



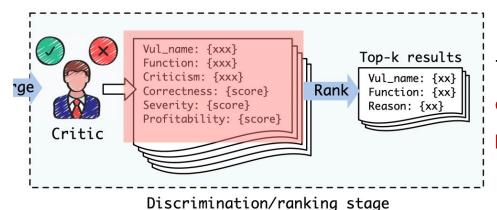
Discrimination/ranking stage

One critic with **deterministic** and **consistent** output (temperature=0). The critic evaluates both vulnerabilities and associated reasoning.

Stage 2: Discrimination

The LLM plays the role of **critic**.

Goal: Reduce the number of false positives.



The critic give scores in terms of correctness, severity and profitability.

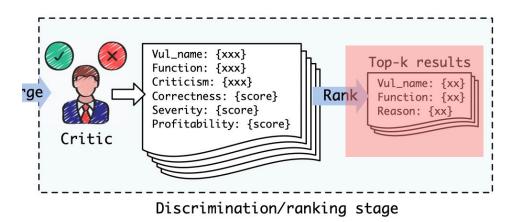
One critic with **deterministic** and **consistent** output (randomness=0).

The critic evaluates both vulnerabilities and associated reasoning.

Stage 2: Discrimination

The LLM plays the role of **critic**.

Goal: Reduce the number of false positives.



Vulnerabilities are ranked based on a composite score.

```
score = 0.5 * correctness + 0.25
* severity + 0.25 *
profitability
```

One critic with **deterministic** and **consistent** output (randomness=0).

The critic evaluates both vulnerabilities and associated reasoning.