GPTLens: an adversarial two-stage detection framework

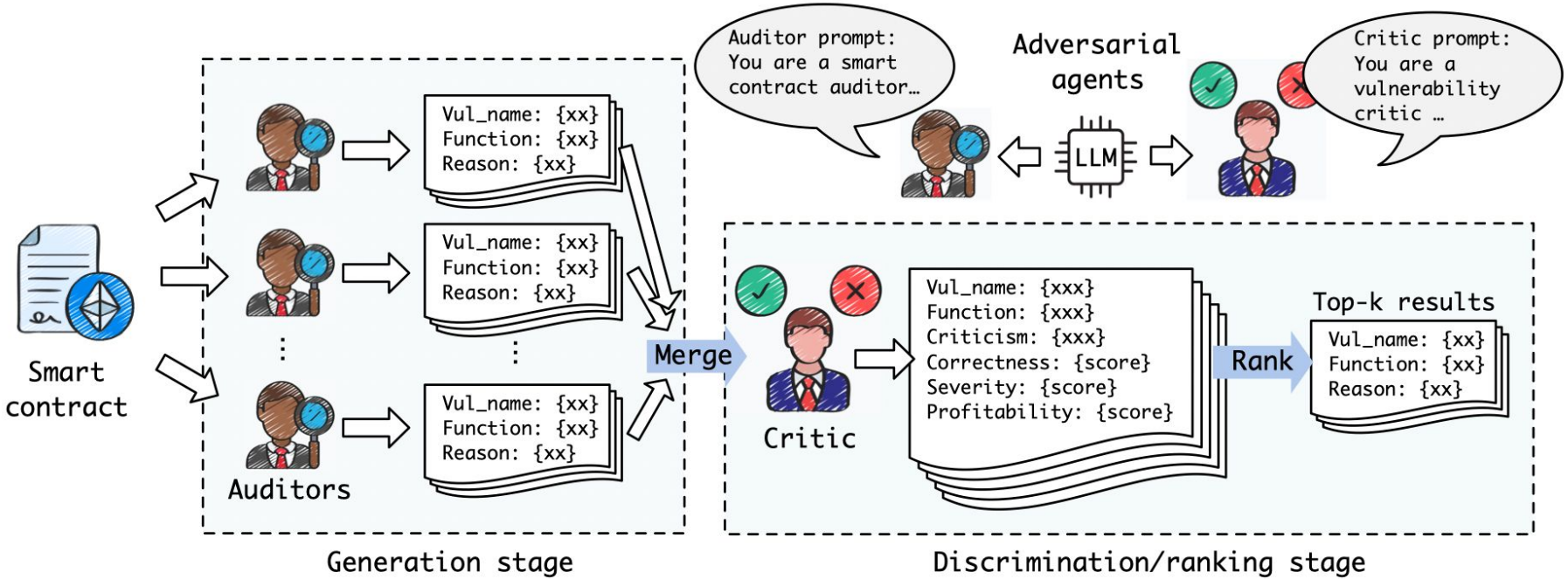


Fig. 2. GPTLens: a framework that breaks single-stage detection into the generation and discrimination stages.

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 positives introduced by generating more answers.

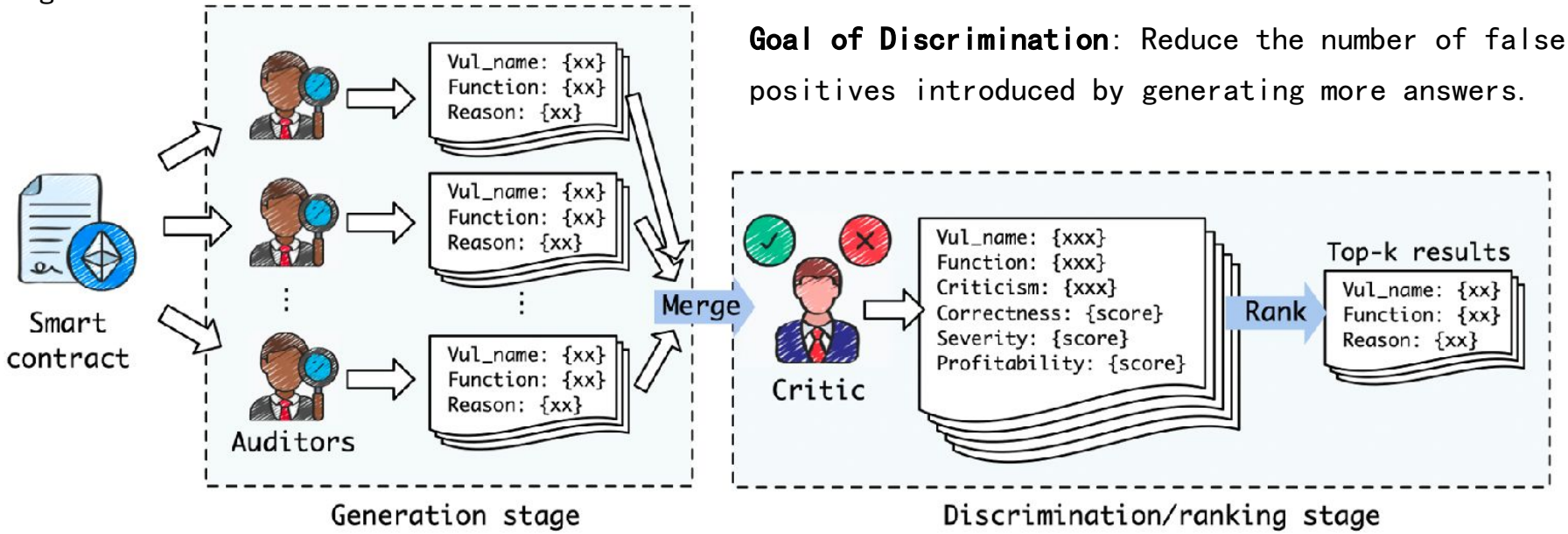


Fig. 2. GPTLens: a framework that breaks single-stage detection into the generation and discrimination stages.

Stage 1: Generation

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

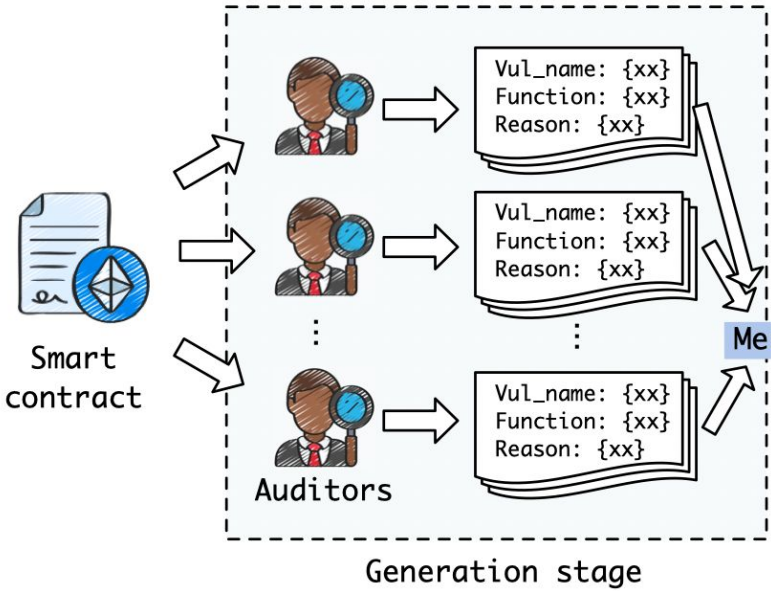


Fig. 2. GPTLens: a framework that breaks si

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

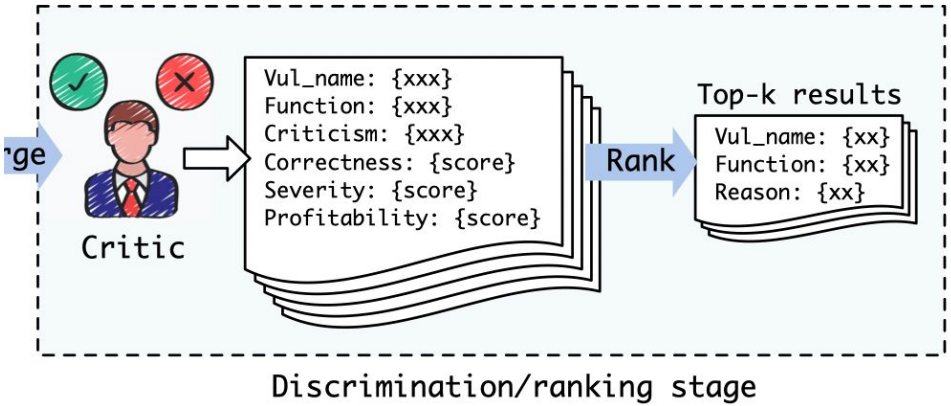
Multiple auditors and each identify multiple (k) vulnerabilities with high randomness.

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.



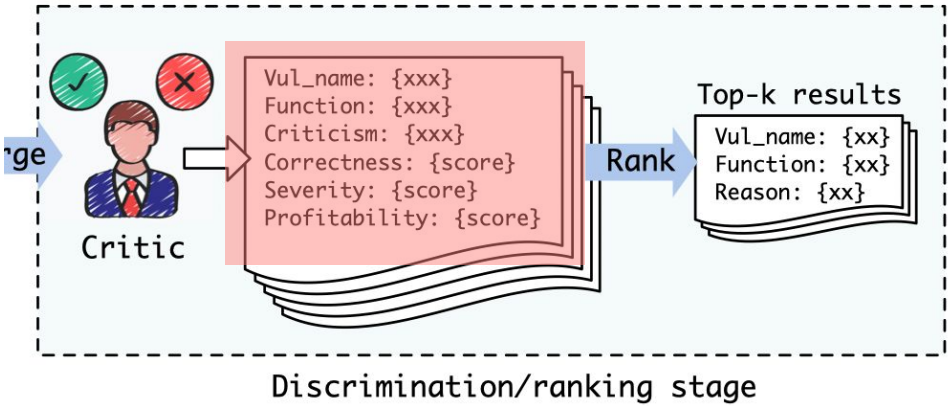
One critic with **deterministic** and **consistent** output (temperature=0).

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.

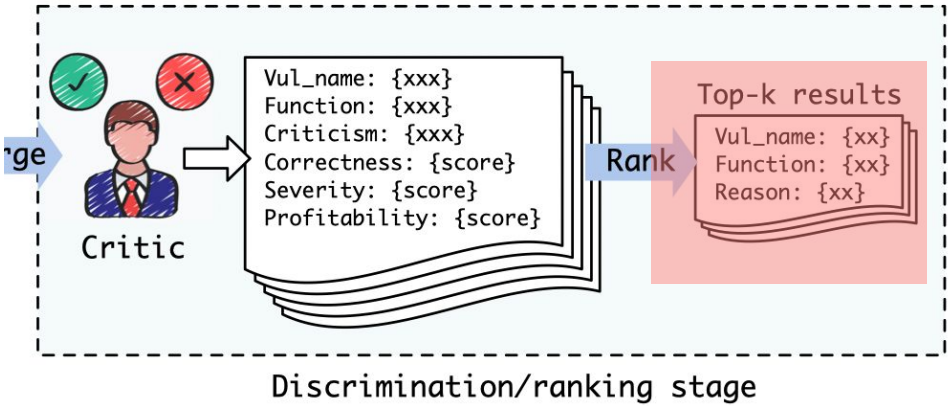
One critic with **deterministic** and **consistent** output (randomness=0).

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.