Audit Agent Benchmark

web3

김휘성

POC SECURITY

Who am I

김휘성

건국대학교 컴퓨터공학부 21학번

건국대 보안동아리 SecurityFACT 회장

화이트햇 스쿨 3기

업사이드 아카데미 4기

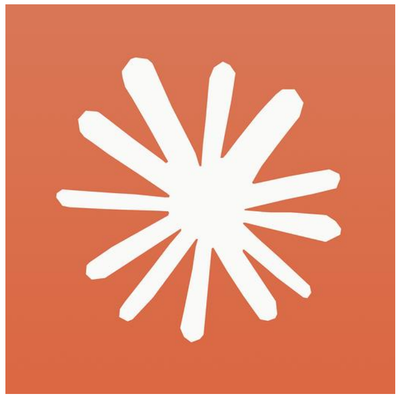
+ HackingCamp 32기 Best Hacker & CTF killer

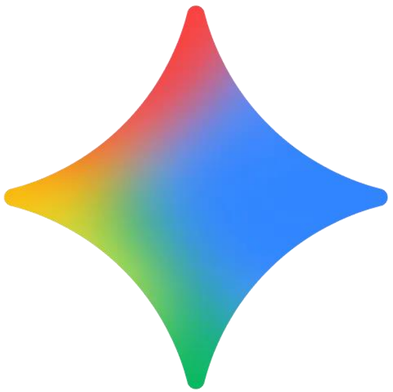
+ ENKI RedOps 2팀

Audit Agent Benchmark

web3

AI era?





POC SECURITY

Upside Academy

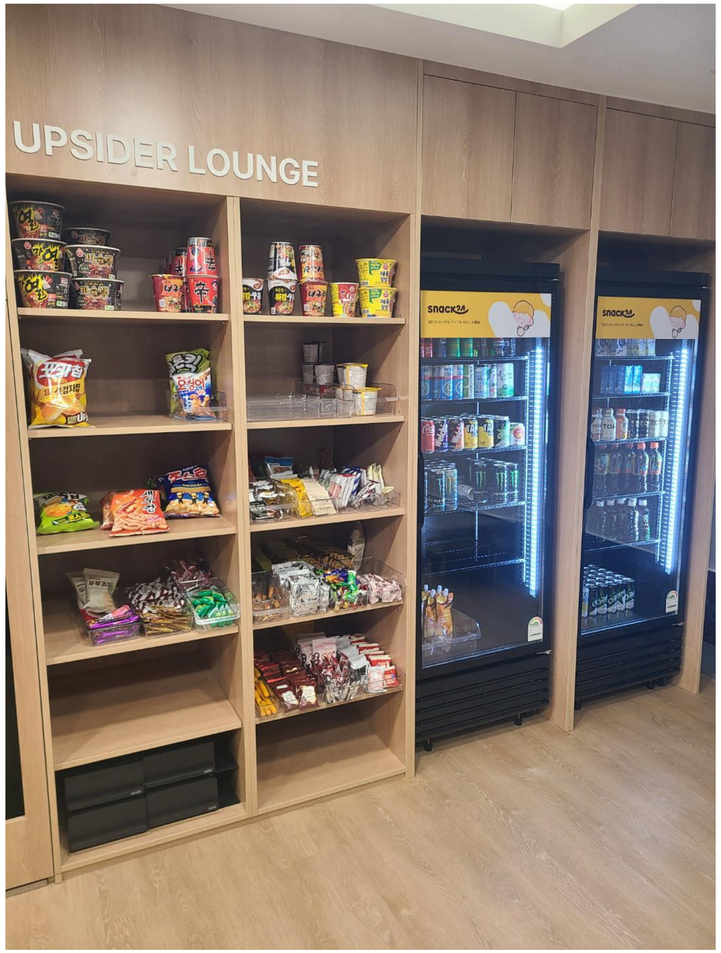
Dunamu

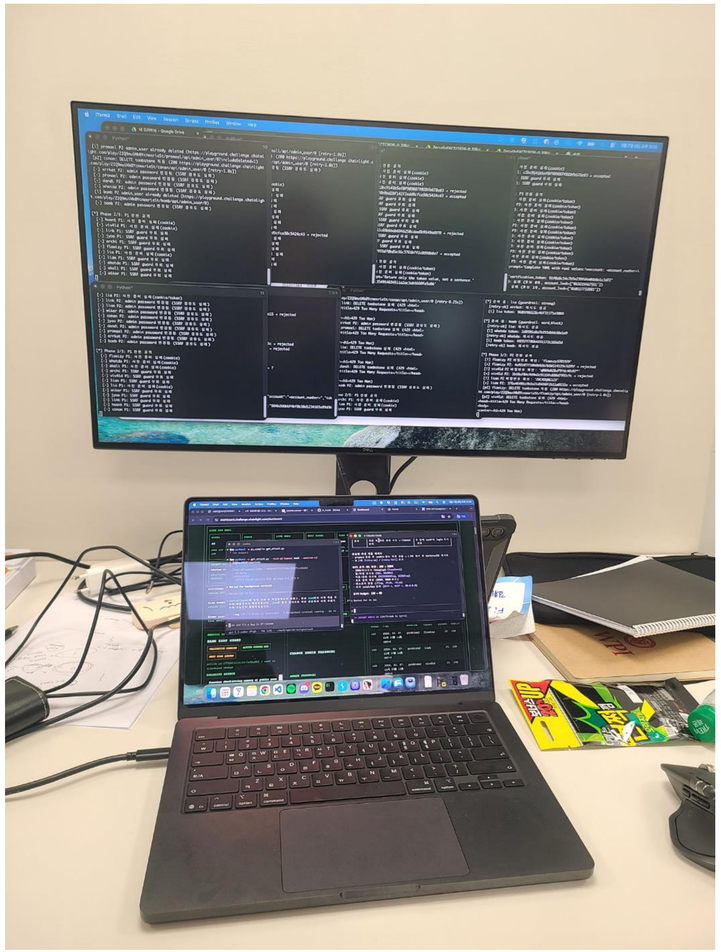
POC SECURITY

Upside Academy

POC SECURITY

Upside Academy



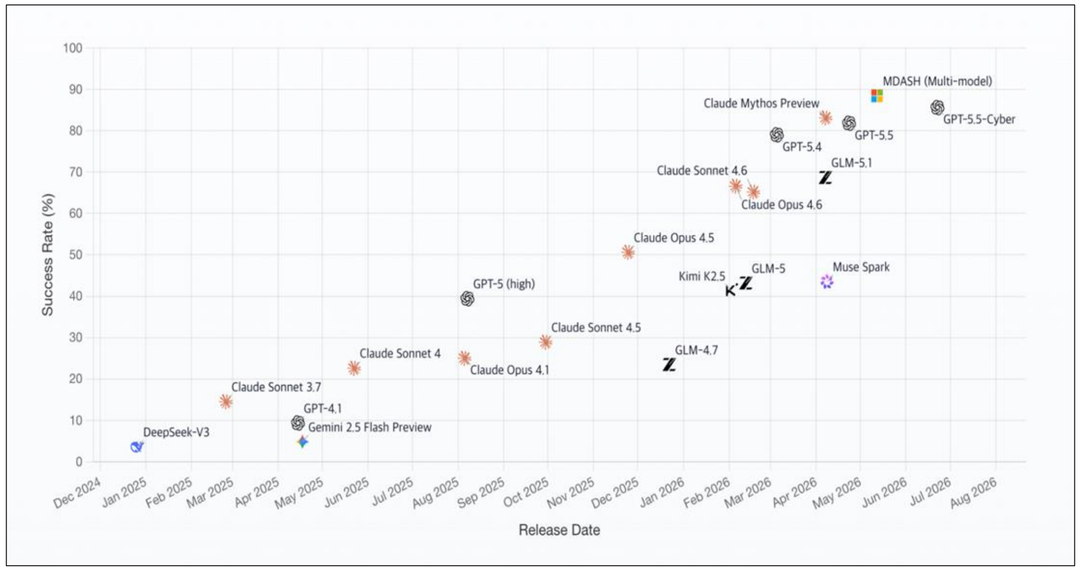


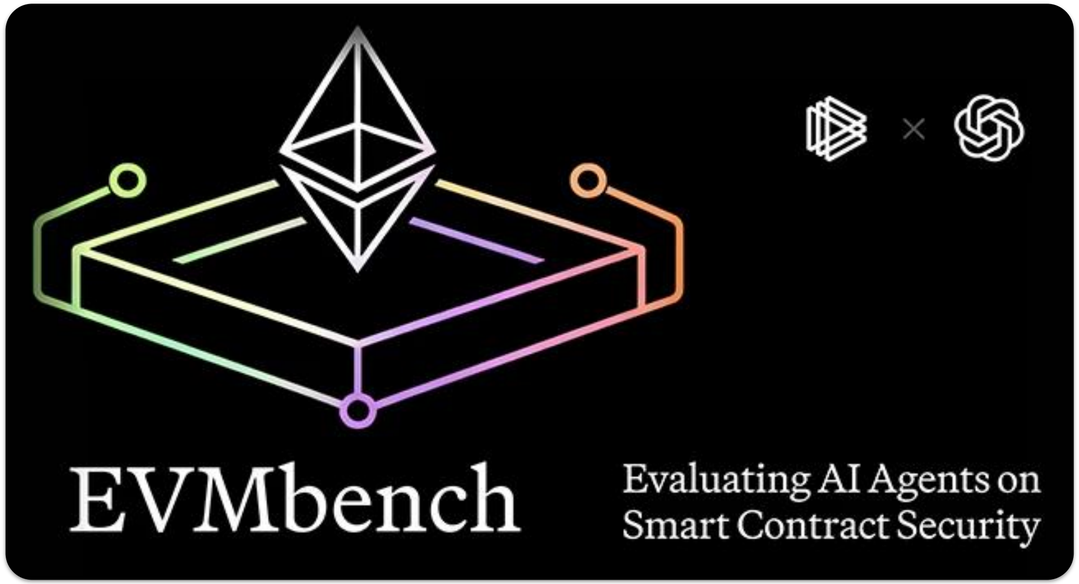
POC SECURITY

AI 벤치마크란?

여러 모델을 같은 조건에서 평가해 성능을 정량적으로 비교

-> 사용자가 목적에 맞는 모델을 고르는 근거가 됨.





POC SECURITY

AI 벤치마크란?

Claude Opus 5	GPT 5.6 Sol	GPT 5.6 Terra
		
Play ↗	Play ↗	Play ↗

게임 개발로 보는 벤치마크

POC SECURITY

보안 분야 AI 벤치마크의 필요성

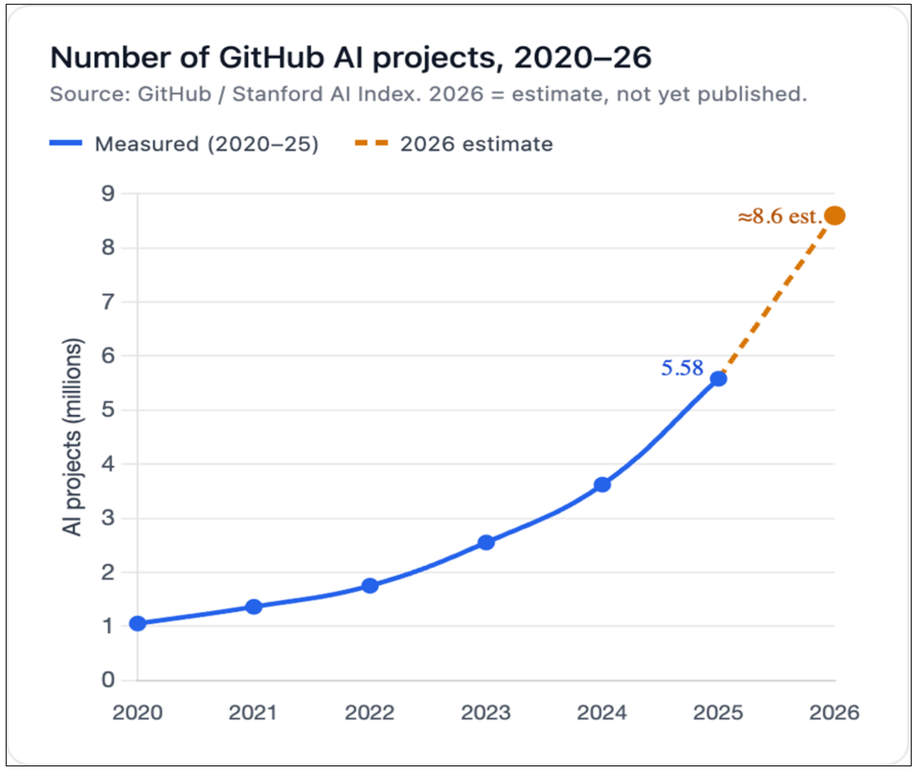
LLM의 코드 해석 및 생성 능력 상승

-> 개발 속도와 코드 생성 속도가 급증

Agent를 Audit에 활용하기 시작

scaffold와 skills의 형태로 발전

→ **어떤 모델이 더 좋은가?**



AI로 작성된 giathub 프로젝트 (2020 ~ 2026)

POC SECURITY

어떻게 평가를 할 것인가?

CTF??

➔반드시 취약점이 존재, flag도 존재

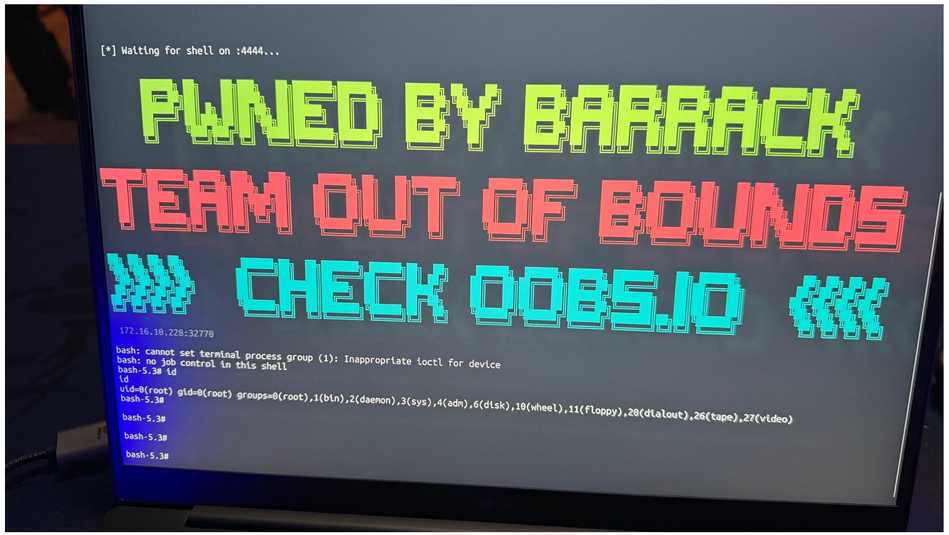
➔정답이 명확함 (목표가 명확함)

알려진 버그를 찾는 능력?

➔웹 검색을 꺼놓고, 구 버전에서의 1-day를 찾는 능력

Web3 특징

해킹이 발생했을 때



POC SECURITY

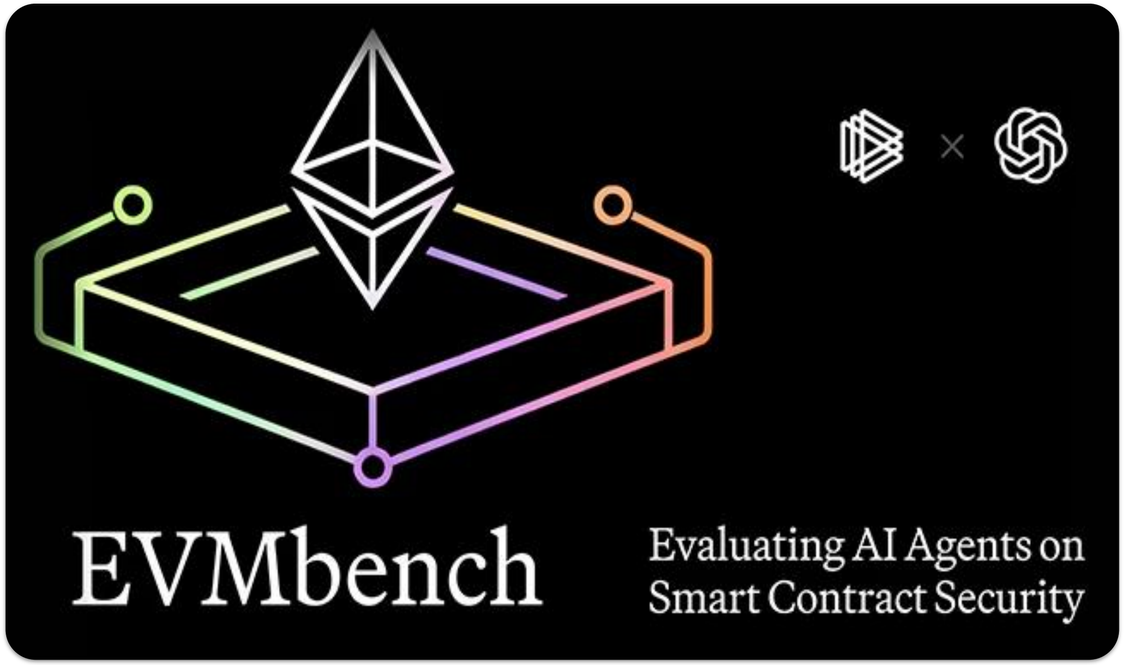
해킹이 발생했을 때

어떻게 평가를 할 것인가?

Web3 auditor의 입장.

자신이 검토를 한 프로토콜에 취약점이 단 하나도 남아있으면 안됨.

기존의 벤치마크



POC SECURITY

기존의 벤치마크의 한계

EVM-Bench의 한계

reevmbench

arXiv
https://arxiv.org › cs

Re-Evaluating EVMBench: Are AI Agents Ready for Smart ...

C Peng 저술 · 2026 · 2회 인용 — Abstract:EVMbench, released by OpenAI, Paradigm, and OtterSec, is the first large-scale benchmark for AI agents on smart contract security.

2026. 02

EVM-Bench 공개

Agent가 스마트컨트랙트 취약점을 탐지하고, 패치하고, 익스플로잇할 수 있는지를 평가하는 벤치마크

https://github.com/paradigmxyz/evmbench

2026. 03

Re-EVM Bench 논문 발표

EVM-Bench의 **구조적 한계** 지적

- **데이터 오염**.
- **평가 대상 범위**가 좁음.
- 채점용으로 단일 model만을 사용.

POC SECURITY

기존 벤치마크의 한계

Dataset의 문제점 (1/3)

Cut-off 이전의 Dataset

오래된 데이터 (2023)

2023-07 pooltogether 프로토콜
2023-10 nextgen 프로토콜
2023-12 ethereumcreditguild 프로토콜

~

최신 데이터 (2026)

2025-10 sequence 프로토콜
2026-01 tempo-feeamm 프로토콜
2026-01 tempo-mpp-streams 프로토콜

40개 데이터셋 중 90%가 2025.06 이전의 사고 데이터

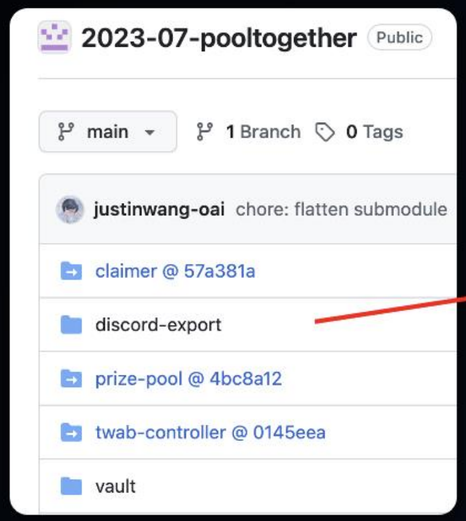
모델 Cut-off 시점

POC SECURITY

기존 벤치마크의 한계

Dataset의 문제점 (2/3)

불완전한 환경 격리



discord 대화에 취약점 내용이 포함됨

```
506
507     [07/10/2023 10:23] hasanza
508     Pardon me, I meant why use uint96.max as the deposit limit?
509
510
```


POC SECURITY

기존 벤치마크의 한계

Dataset의 문제점 (3/3)

의심되는 정답 Dataset

높은 심각도의 정답 중 최소 4개에 대해 공격이 불가능하다 지적

> **From this review, four issues classified as valid high-severity vulnerabilities in the benchmark are, in our assessment, invalid.** Note that our researchers reviewed a

https://www.openzeppelin.com/news/openai-evmbench-audit

기존 벤치마크의 한계

평가 환경의 문제

LLM과 Scaffold를 분리하지 않음

Scaffold는 LLM이 파일 탐색, 명령 실행, 오류 시 재실행을 하도록 돕는 역할

A

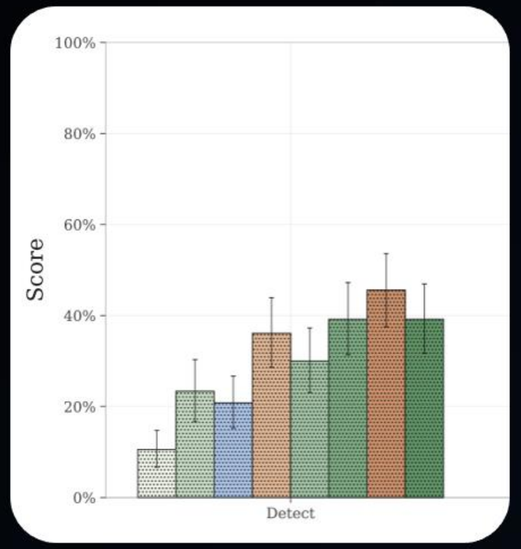
B

모델이 좋아서인지 scaffold의 설계 차이인지 구분하기 어려움

기존 벤치마크의 한계

평가 지표의 부족

기존 채점 방식 : TP/FN 중심의 측정



EVM-bench

기존 벤치마크에서 A와 B의 최종 점수는 같음

POC SECURITY

벤치마크 제작

Agent 선택시 고려 사항

사용자가 궁금한 부분

1. 이 Agent가 실제 취약점을 잘 찾는가? **(TP/FN)**
2. 취약점이 아닌 코드를 취약점이라고 착각하지 않는가? **(FP)**
3. 여러 Finding이 나왔을 때 우선순위를 잘 잡는가? **(Severity)**
4. 이 결과를 내기 위해 시간과 비용을 얼마나 쓰는가? **(Latency/Token)**

벤치마크 제작

평가 대상

- ⇨ Claude code (Scaffold)
- ⇨ Opus 4.8
- ⇨ Sonnet 4.6
- ⇨ Haiku 4.5

- ⇨ Codex (Scaffold)
- ⇨ GPT 5.5
- ⇨ GPT 5.4-mini

- ⇨ Gemini 3.5 flash
- ⇨ Gemini 3.1 pro

벤치마크 제작

POC SECURITY

벤치마크 제작

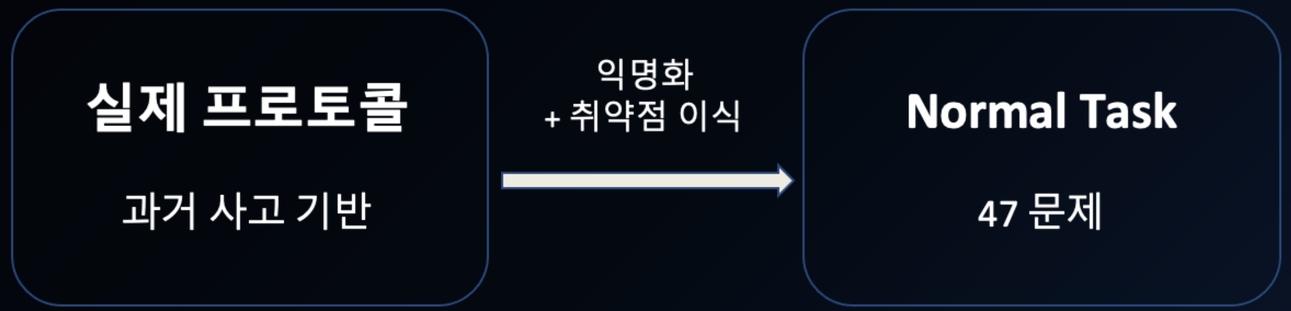
평가 절차

POC SECURITY

벤치마크 제작

Dataset

Normal Task 제작



✓ 데이터 오염 차단

✓ 현실성, 난이도 확보

벤치마크 제작

Dataset - 익명화 예시

메인넷 침해 사고

MetaMorpho Donation Oracle Exploit

2026.03.16 21:25 UTC Attack Atomic

TOKENS LOST	USD ESTIMATE
5,782.77 USDC	$5,783

프로토콜명 제거 (익명화)

```
- /// @title MetaMorpho
- /// @author Morpho Labs
- /// @custom:contact security@morpho.org
- /// @notice ERC4626 compliant vault allowing…
- contract MetaMorphoV1_1 is ERC4626, ERC20Permit…
+ // context note
+ contract PortfolioVault is ERC4626, ERC20Permit…
```

프로토콜·작성자·주석 등 식별 단서를 제거 → 취약점은 그대로지만 익명화

POC SECURITY

벤치마크 제작

평가 지표

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

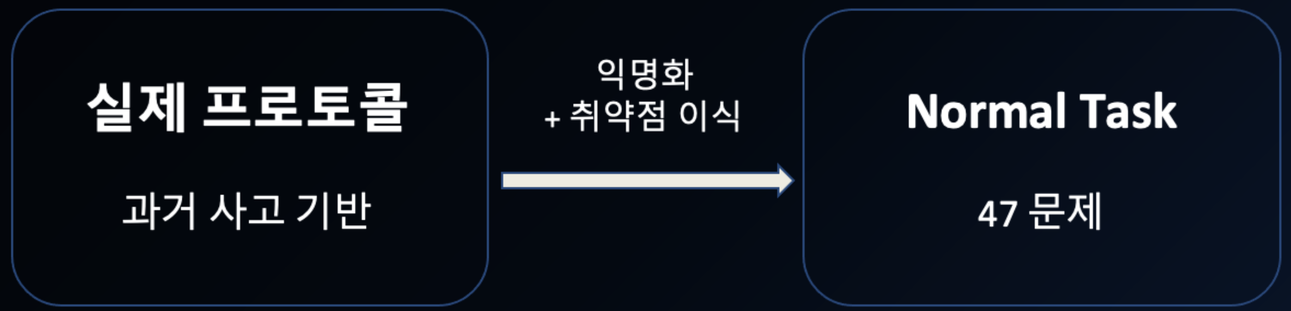
$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

TP	Agent가 식별한 실제 취약점
FN	Agent가 놓친 실제 취약점
FP	취약점이 아닌 것을 취약점이라고 한 것

벤치마크 제작

Dataset

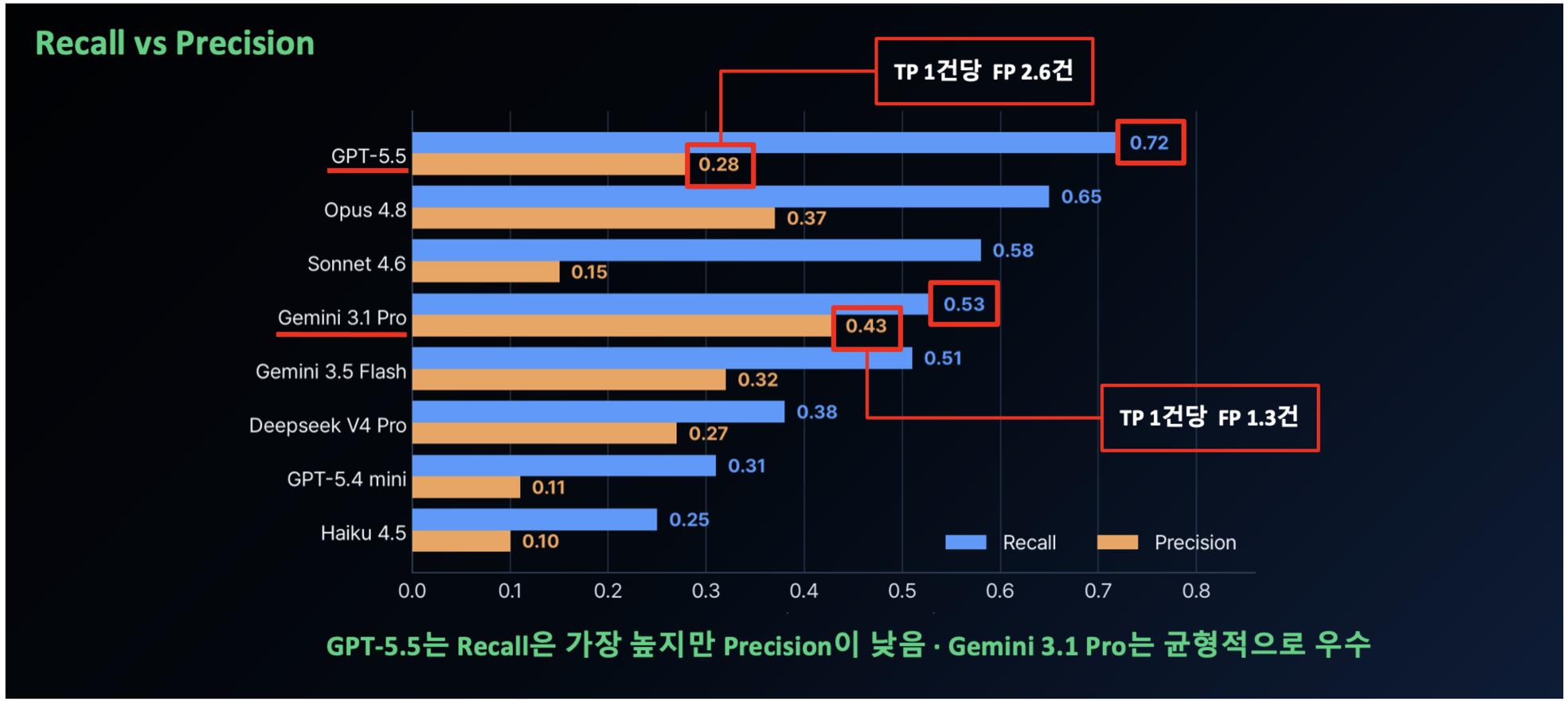
Normal Task 제작



✓ 데이터 오염 차단

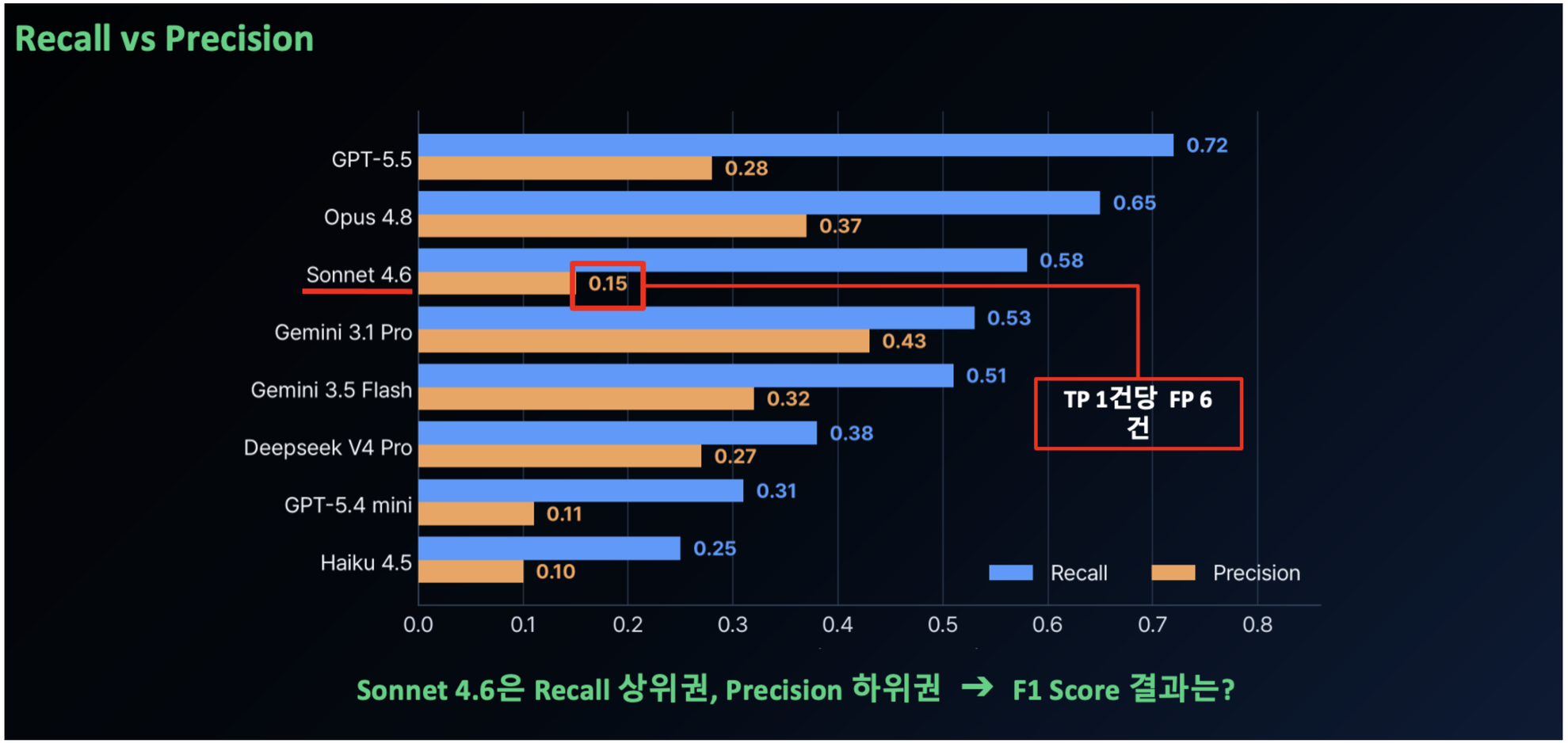
✓ 현실성, 난이도 확보

결과 분석



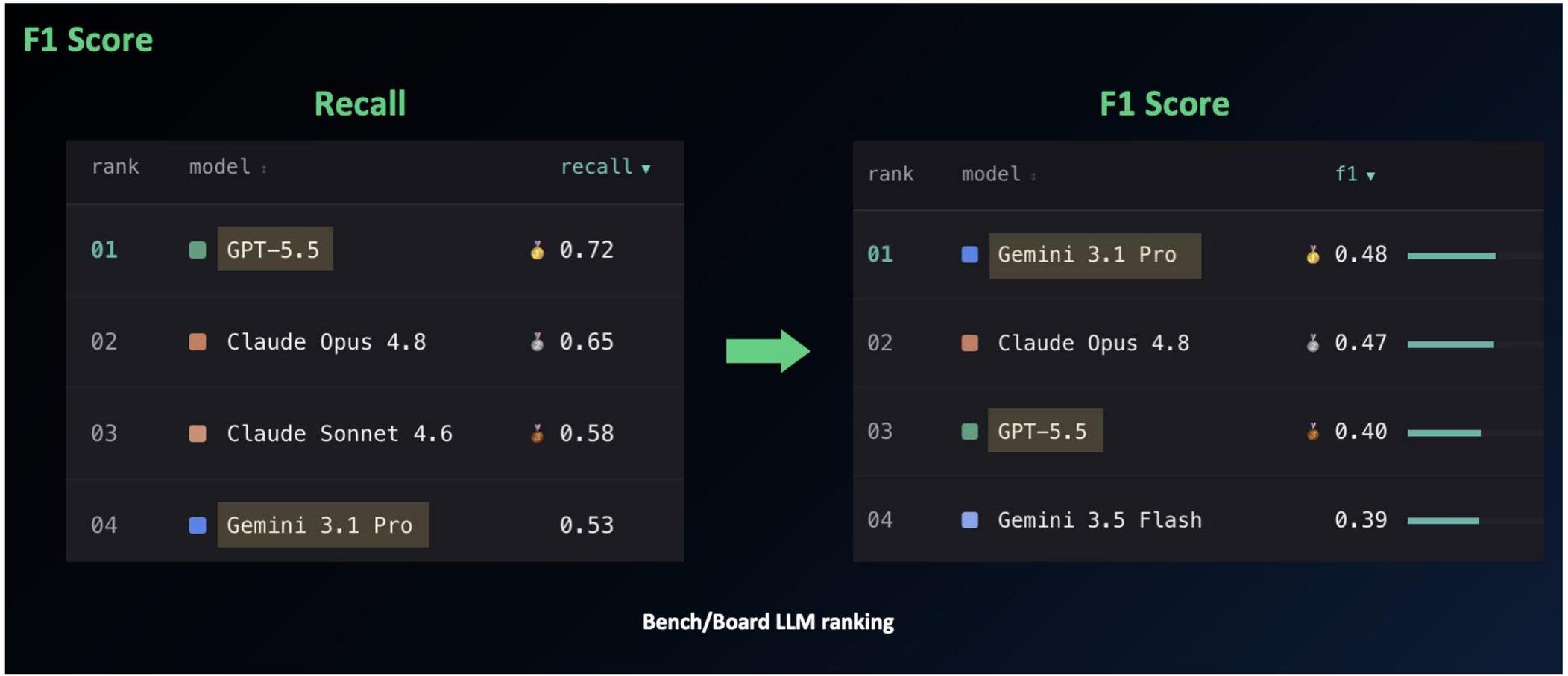
POC SECURITY

결과 분석



POC SECURITY

결과 분석

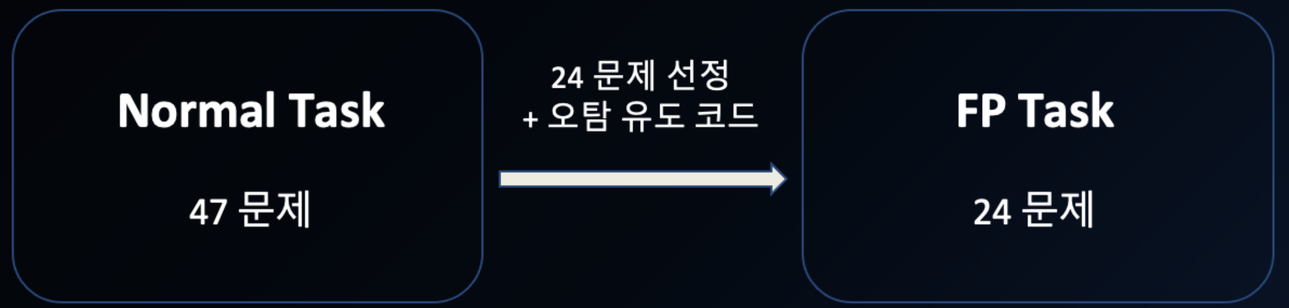


POC SECURITY

결과 분석

가설

오탐 유도 코드가 늘어날 때 성능이 하락할 것



✓ 오탐 측정

✓ 난이도 증가

* 오탐 유도 코드 = 취약점처럼 보이지만 실제로는 익스플로잇이 불가능한 코드

결과 분석

Dataset - 오탐 유도 코드

Settlement.sol

```
1 IERC20 public settlementToken;      ✗ 변수에 값이 할당되지 않음
2 function claimApprovedTokens(uint256 sourceId) external {
3     bytes32 key = keccak256(abi.encode(sourceId, msg.sender));
4     uint256 owed = approved[key];
5     settlementToken.safeTransfer(msg.sender, owed);   // 1. Interaction  ✗ 항상 revert
6     require(approved[key] > 0, "nothing approved");   // 2. Check
7     approved[key] = 0;   // 3. Effect
8     deposited[msg.sender] -= owed;
9 }
```

reentrancy 패턴처럼 보이지만 settlementToken=address(0)로 항상 revert

결과 분석

평가 과정

24문제

정답률 높은 문제

+

오탐 유도 코드

익스플로잇 불가능한 취약점

=

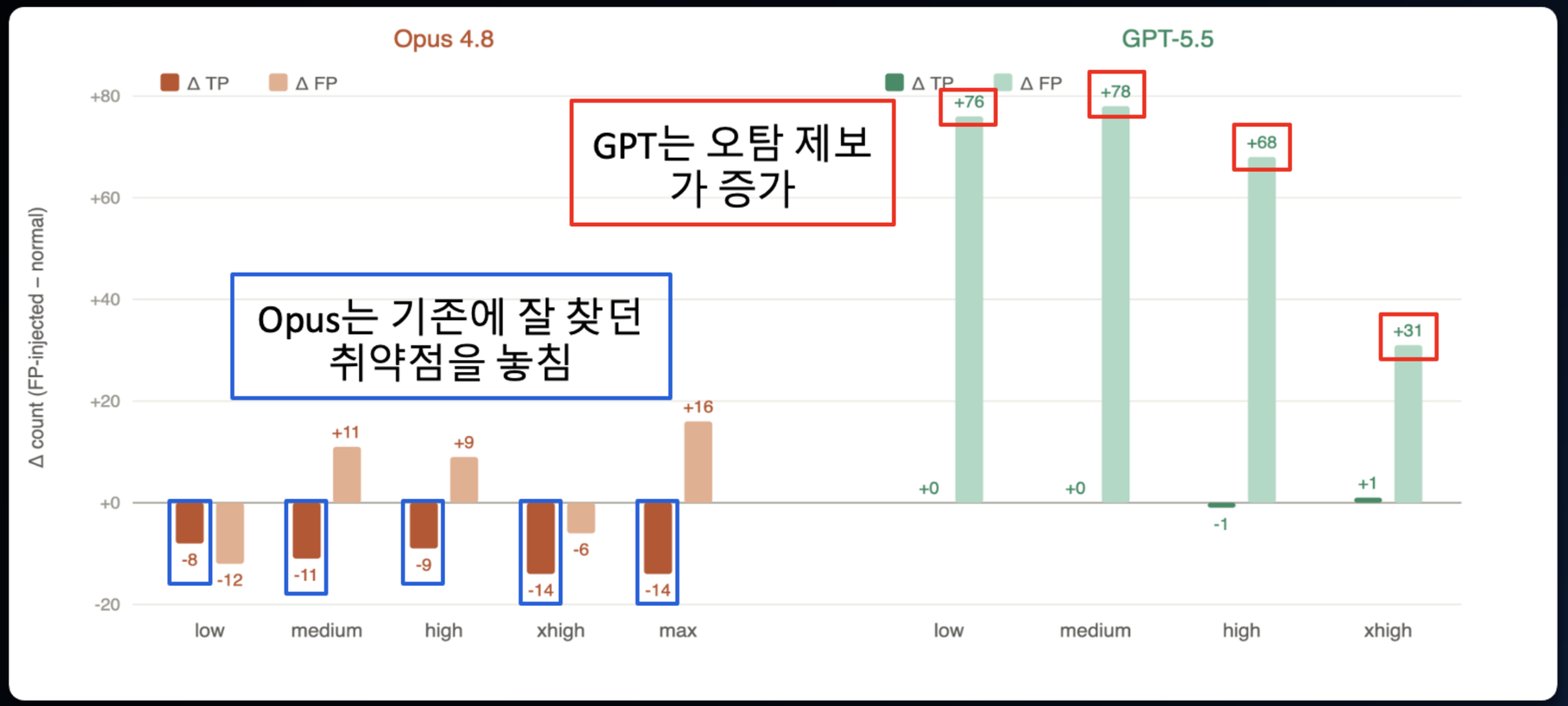
FP Task 24개

최종 데이터셋

진짜 취약점을 놓치는지 (Δ정탐↓) / 가짜 취약점을 신고하는지 (Δ오탐↑)

결과 분석

오탐 유도 코드 주입 전 vs 후



POC SECURITY

결과 분석

모델별 특성 해석

검토 대상이 늘면 집중력 분산 → 진짜 취약점을 놓침 (TP↓)

위험해 보이는 신호에 대한 검증 부족 → 안전한 코드까지 Find (FP↑)

결과 분석

Severity란?

찾은 취약점이 얼마나 위험한지를 매기는 등급

등급	취약점	영향 (Immunefi 기준)
• Critical	자금 직접 탈취, 영구 동결	프로토콜 파산, 전체 예치금 손실
• High	상당한 자금, 핵심 기능 손상	미청구 수익 탈취, 일시적 자금 동결
• Medium	제한적, 조건부 피해	컨트랙트 기능 마비, 그리핑(griefing)
• Low	이론적 위험	자금 손실 없는 사양 위반

출처 : Immunefi Severity Classification System

결과 분석

채점 기준

AI \ 정답	Crit	High	Med	Low
Crit	**1.0**	0.9	0.1	-0.5
High	0.9	**1.0**	0.7	-0.2
Med	0.2	0.4	**1.0**	0.5
Low	-1.0	-0.5	0.5	**1.0**

Severity score 가중치 표

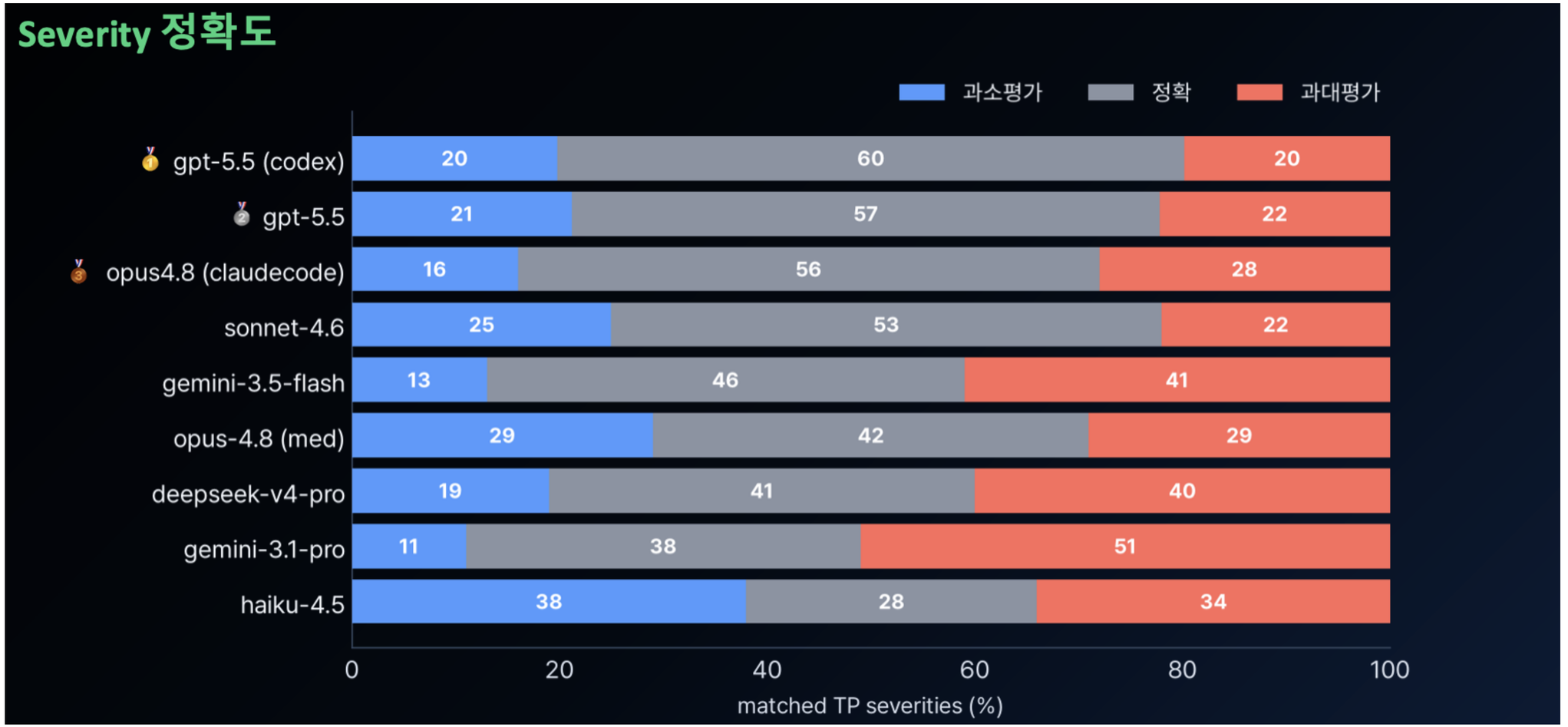
Column = 답지의 Severity

Row = Agent가 도출한 Severity

Exact match = 1.0

과소평가에서 더 큰 Penalty

결과 분석

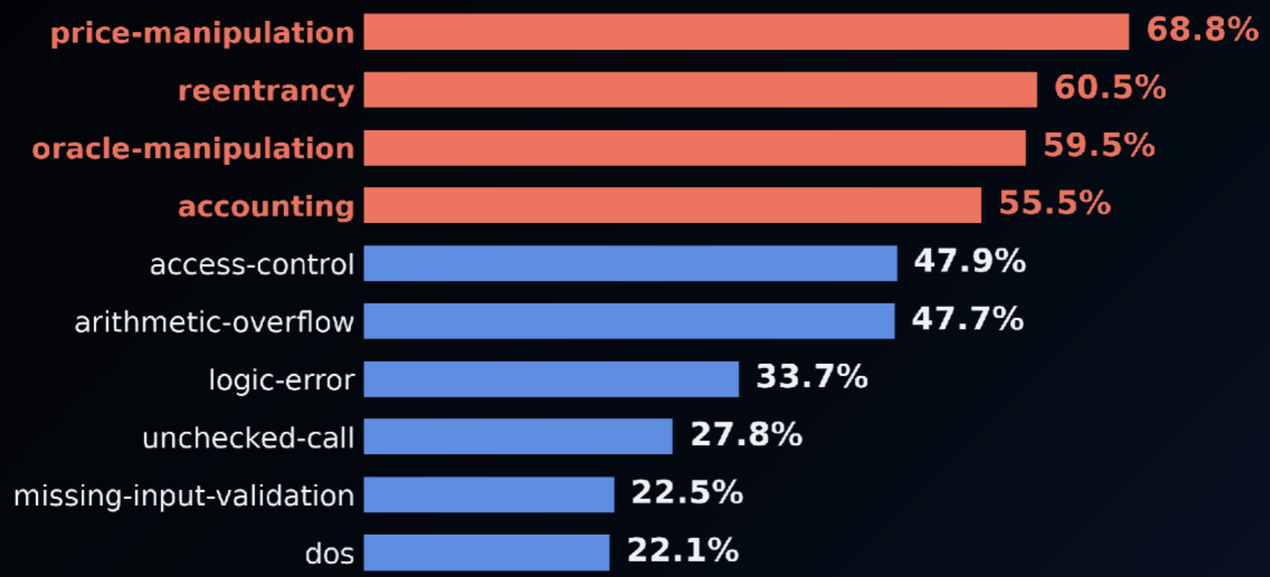


POC SECURITY

결과 분석

오탐에서의 severity 비율

오탐(FP)에 High·Critical 등급이 매겨진 비율



실제 사고에서 큰 피해로 이어지는 유형일수록 높은 비율(60~69%)로 과대평가(High/Critical)

POC SECURITY

결과 분석

POC SECURITY

결과 분석

종합 결과

rank	model	vendor	f1	recall	precision	severity score	$/task
01	Gemini 3.1 Pro	Google	0.48	0.53	0.43	0.70	$0.27
02	Claude Opus 4.8	Anthropic	0.47	0.65	0.37	0.64	$0.67
03	GPT-5.5	OpenAI	0.40	0.72	0.28	0.77	$0.38
04	Gemini 3.5 Flash	Google	0.39	0.51	0.32	0.71	$0.26
05	DeepSeek V4 Pro	Others	0.32	0.38	0.27	0.56	$0.02
06	Claude Sonnet 4.6	Anthropic	0.24	0.58	0.15	0.78	$0.16
07	GPT-5.4 mini	OpenAI	0.17	0.31	0.11	0.58	$0.03
08	Claude Haiku 4.5	Anthropic	0.14	0.25	0.10	0.58	$0.09

마무리

최근 모델 -> 결과가 다름

모델이 나올 때마다 변화가 너무 커서 매번 다시 측정

감사합니다.

QnA

김휘성

POC SECURITY