

AI-Native Auditing in Five Minutes

Questions, Evidence, Coverage

nisedo · Trail of Bits

DSS 2026 · Mumbai · November 1–2, 2026

TECHNIQUE 01 · THREAT MODEL

Start every agent from a threat model

> No more “malicious admin” findings when the admin is trusted.

MY RECOMMENDATION

codex-security · threat-model skill ↗

THREAT_MODEL.md

```
## Assets

## Attacker capabilities

## Trusted roles

## Security properties
```

TECHNIQUE 02 · SMALLER SCOPES

Give each agent a narrower scope

ONE AGENT ➡ WHOLE REPO

```
src/Oracle.sol
src/Liquidator.sol
src/Router.sol
src/Governance.sol
src/Staking.sol
src/lib/Math.sol
src/interfaces/…
…
```

Context overload.
Attention drifts.



BY FILE `Vault.sol`

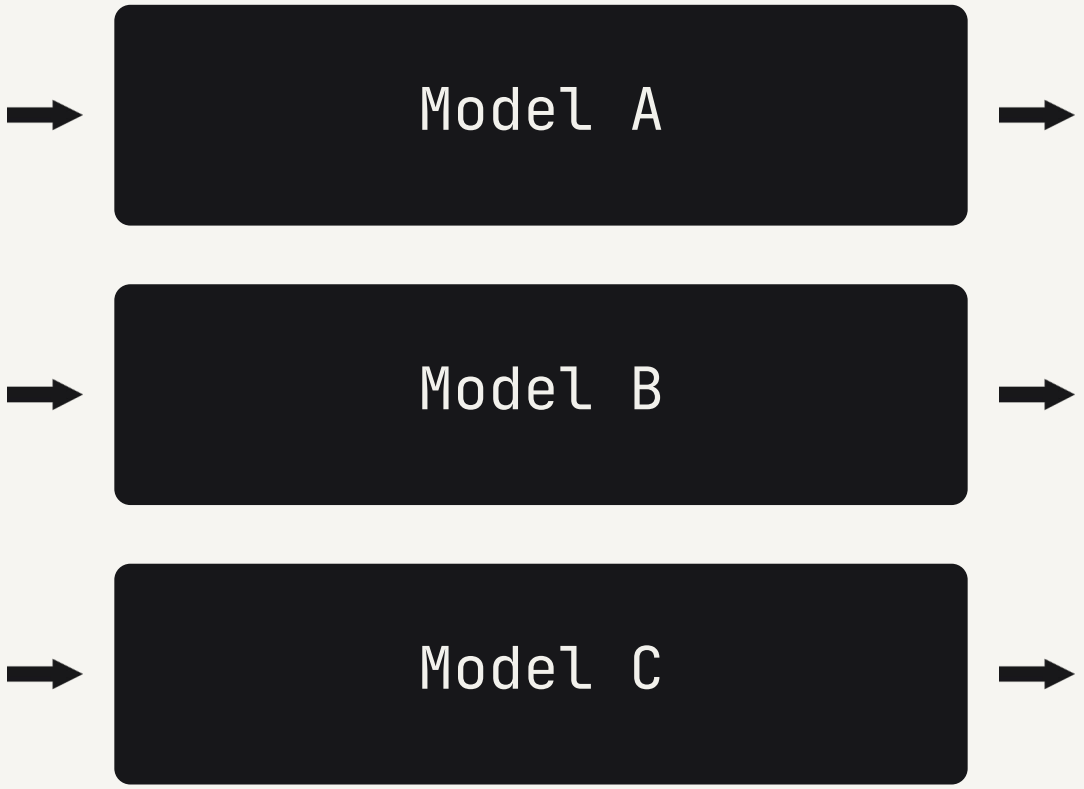
BY FEATURE **withdrawals**

BY COMPONENT **oracle**

Run several models on the same scope

Same assignment

Same threat model



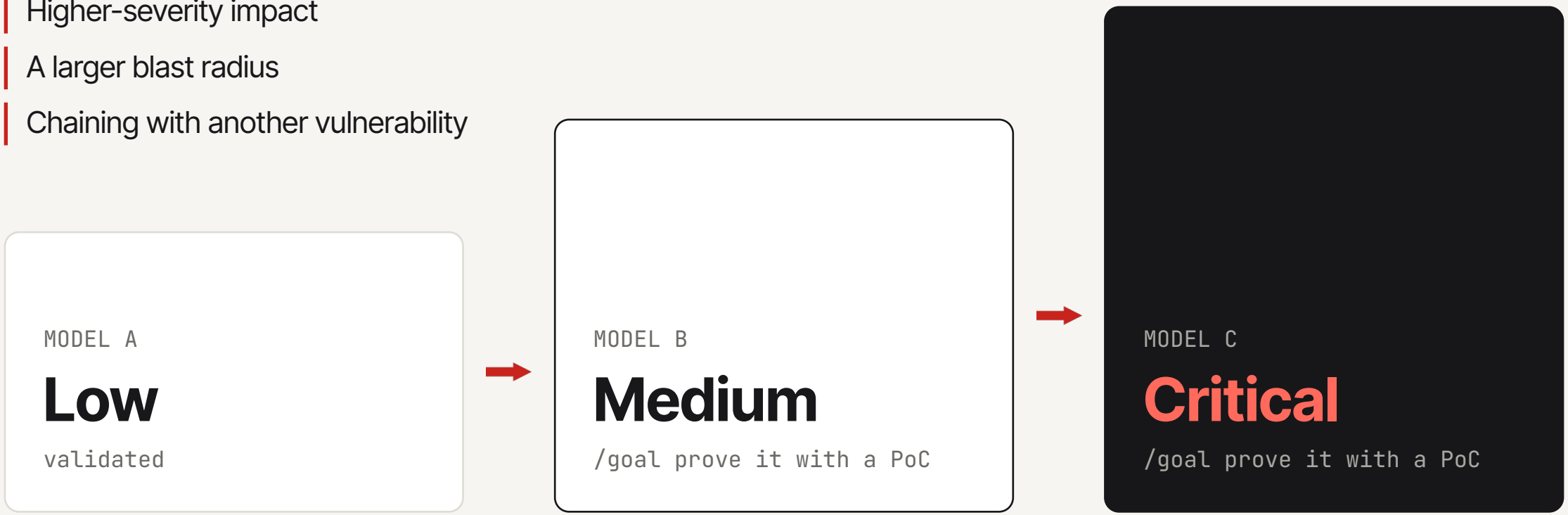
Deduplicate and merge findings

TECHNIQUE 04 · IMPACT ESCALATION

Escalate a validated Low

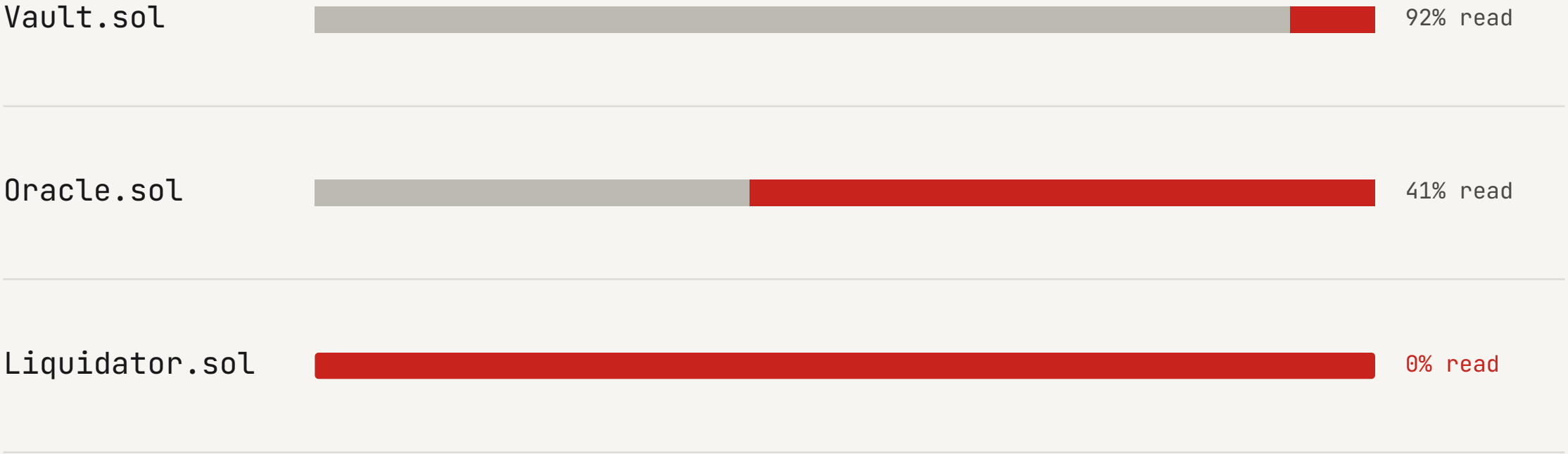
EACH MODEL LOOKS FOR

- Higher-severity impact
- A larger blast radius
- Chaining with another vulnerability



TECHNIQUE 05 · READ COVERAGE

Check what agents actually read



trailofbits/agentcov ↗

01 Start every agent from a threat model

02 Give each agent a narrower scope

03 Run several models on the same scope

04 Escalate your validated Lows

05 Check what your agents actually read

Thank you