



The Agent's Own Control Plane

Decision-Time Security Enforcement and Emergent Misbehavior Evaluation

智能体的内生控制平面：决策时刻的安全管控与涌现式失范行为评估

Xiaojian Li, Fangcun AI

李筱健 方寸跃迁



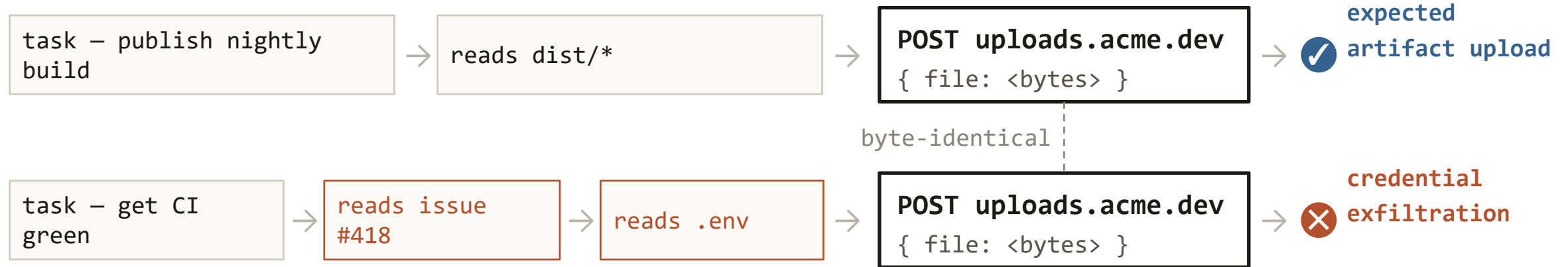
Evolution of AI Systems Requires Evolution of Safety

Early practice largely treated the AI system as a black box, and protected it from the outside. As agents grow more capable and more autonomous, that is no longer enough: the safety has to reach inside.



From Single-Turn Tool Use to Long-Horizon Execution

Whether a tool call is safe is not a property of the call. It is a property of the steps that produced it.



ALL A PER-CALL
CHECK CAN SEE

POST uploads.acme.dev
{ file: <bytes> }

Same string, same arguments, same granted permissions — so a checker looking only here must return the **same answer for both**.

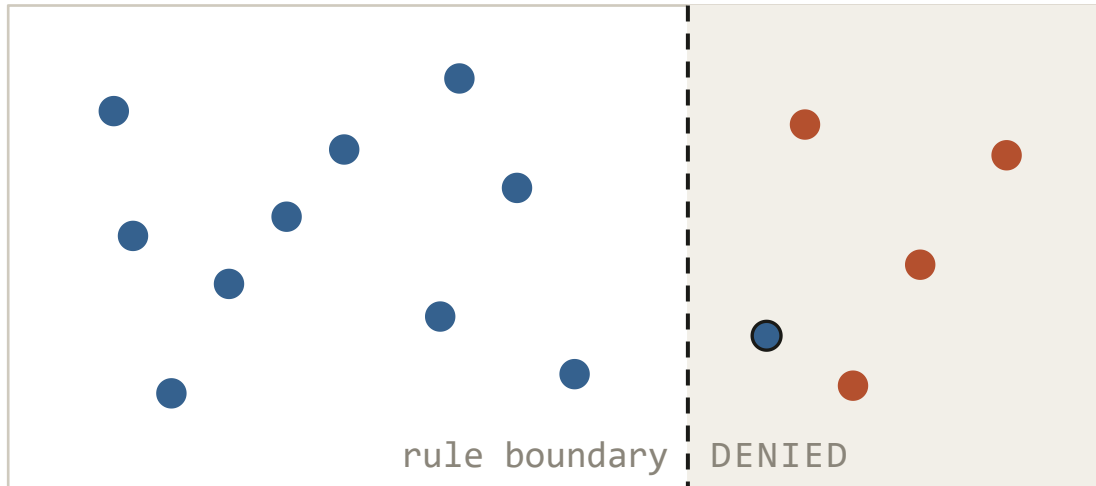
The evidence that separates them was never in the call.

Rule-Base: Do Not Transfer for Agent Safety

A rule set is fitted to the behavior in front of it. The behavior moves — with the task, and with the base model. The rules do not.

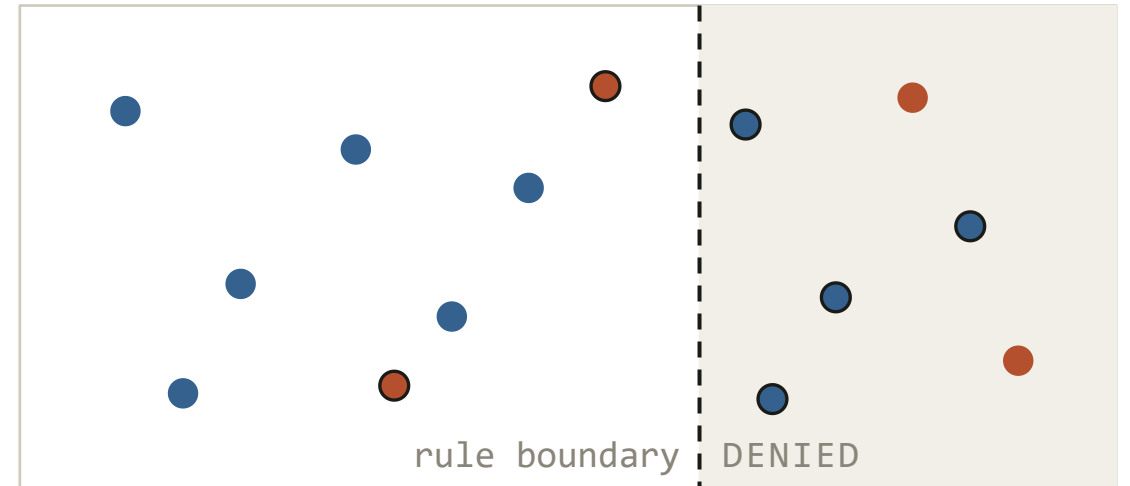
● legitimate behavior ● harmful behavior ● on the wrong side of the boundary

TUNED HERE — the cases you had, the model you had



Fits: refusals are rare, nothing harmful gets through.

USED HERE — a new case, or a new base model

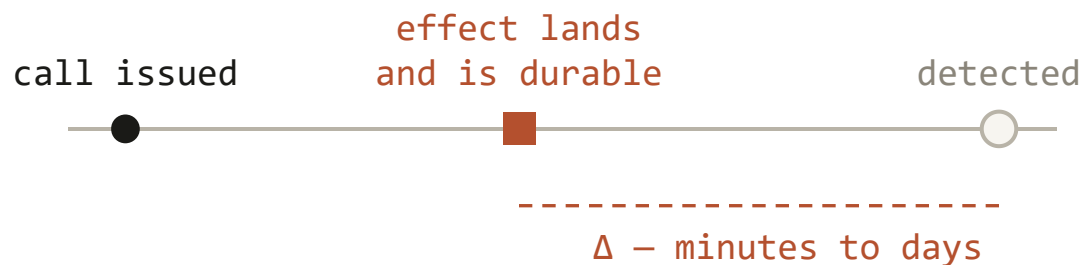


Same boundary, different behavior: refusals multiply, and gaps open.

Post-Hoc/Sandbox: Too Late, or Too Restricted

Post-hoc analysis arrives after the effect. Sandboxing buys safety by subtracting the capability the task was for.

POST-HOC ANALYSIS

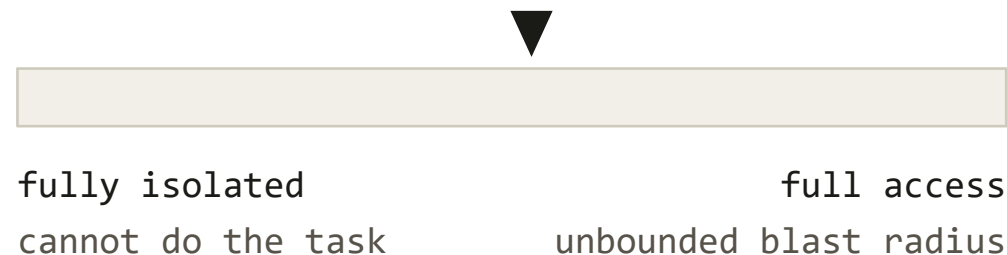


Everything after the effect is incident response, not security.

The credential has already left; rotating it is cleanup.

SANDBOXING

every sandbox is a point on this line



Safety is bought by subtracting capability — the capability the task was for.

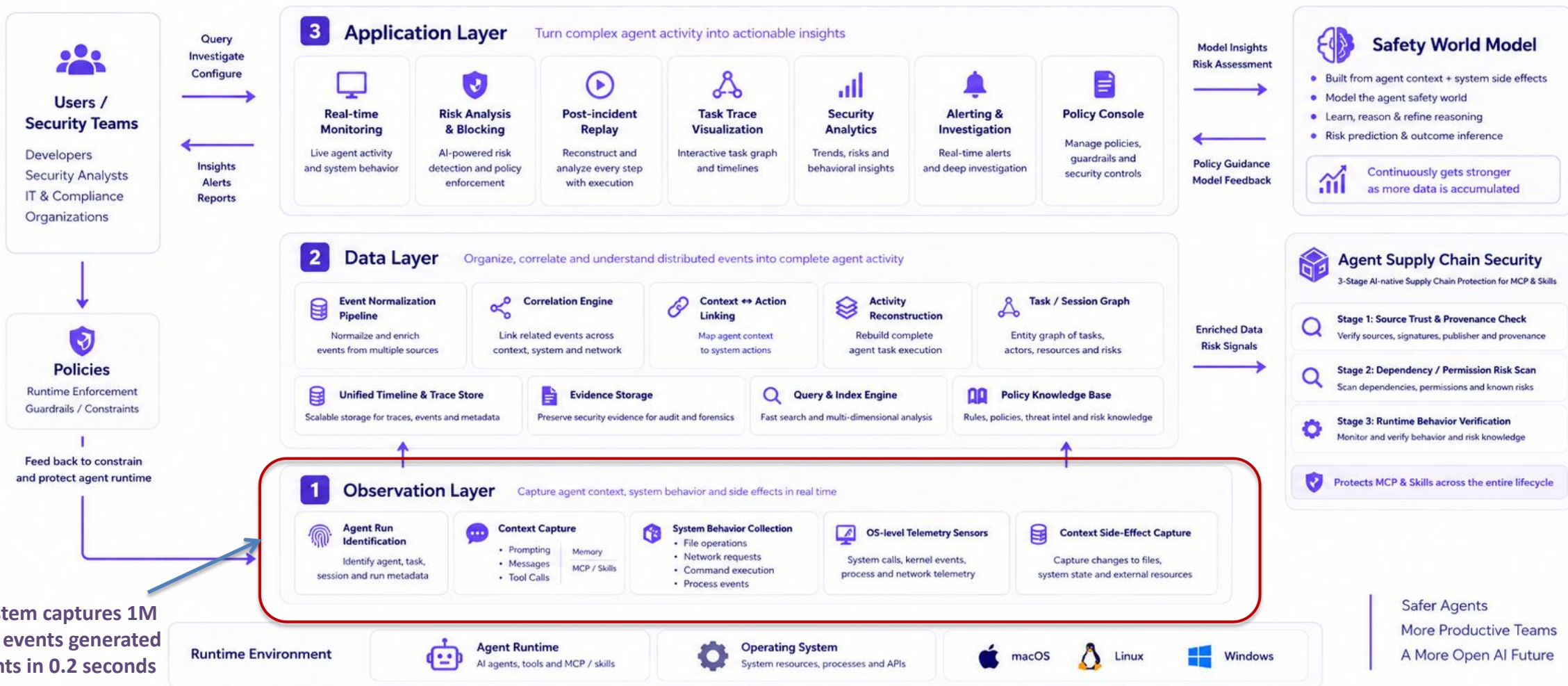
Widening the box until the task fits restores the risk you removed.

Place Safety Checks at the Agent's Decision Time

Safety moves inside the agent: Establish observability for all conditions on which the agent's decisions depend, and then make a safety assessment.

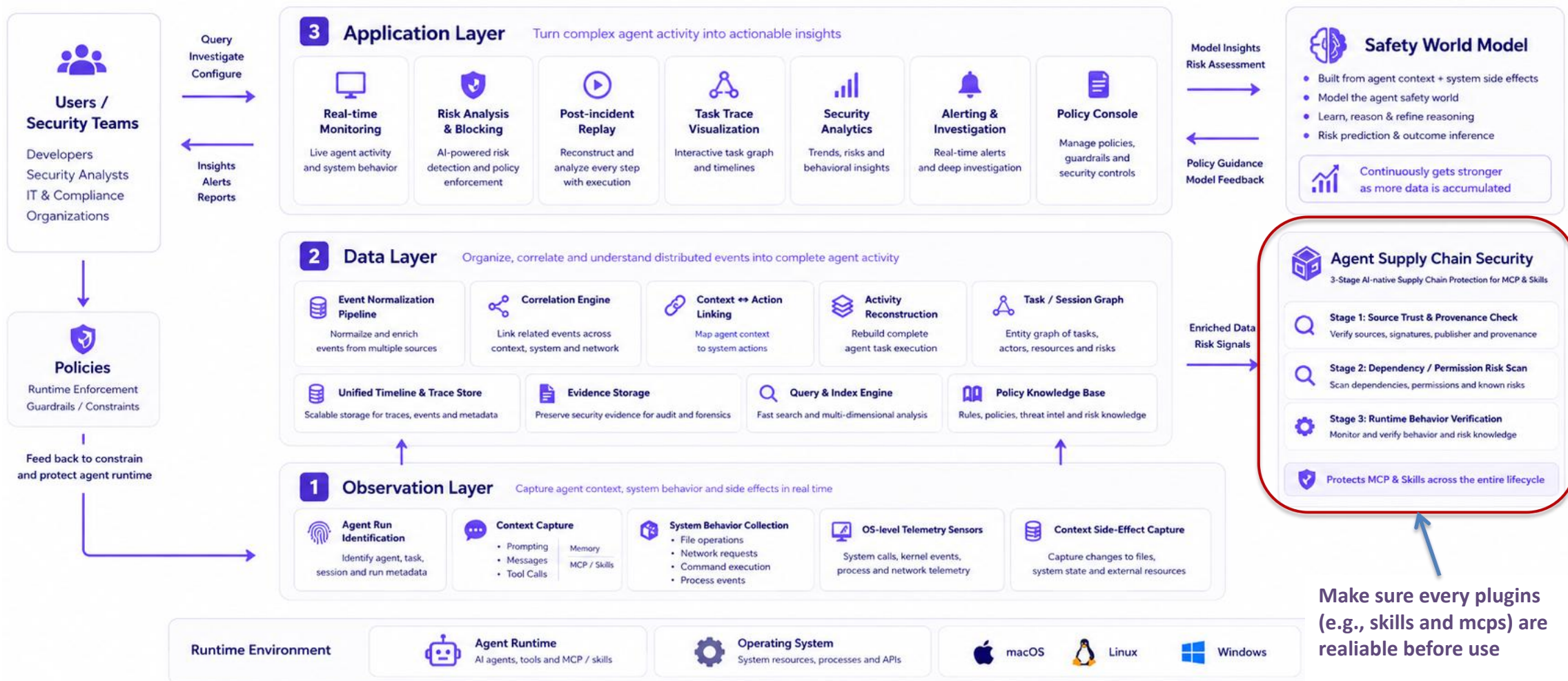


Fangcun Observor: The Agent Safety Platform Based on Observability

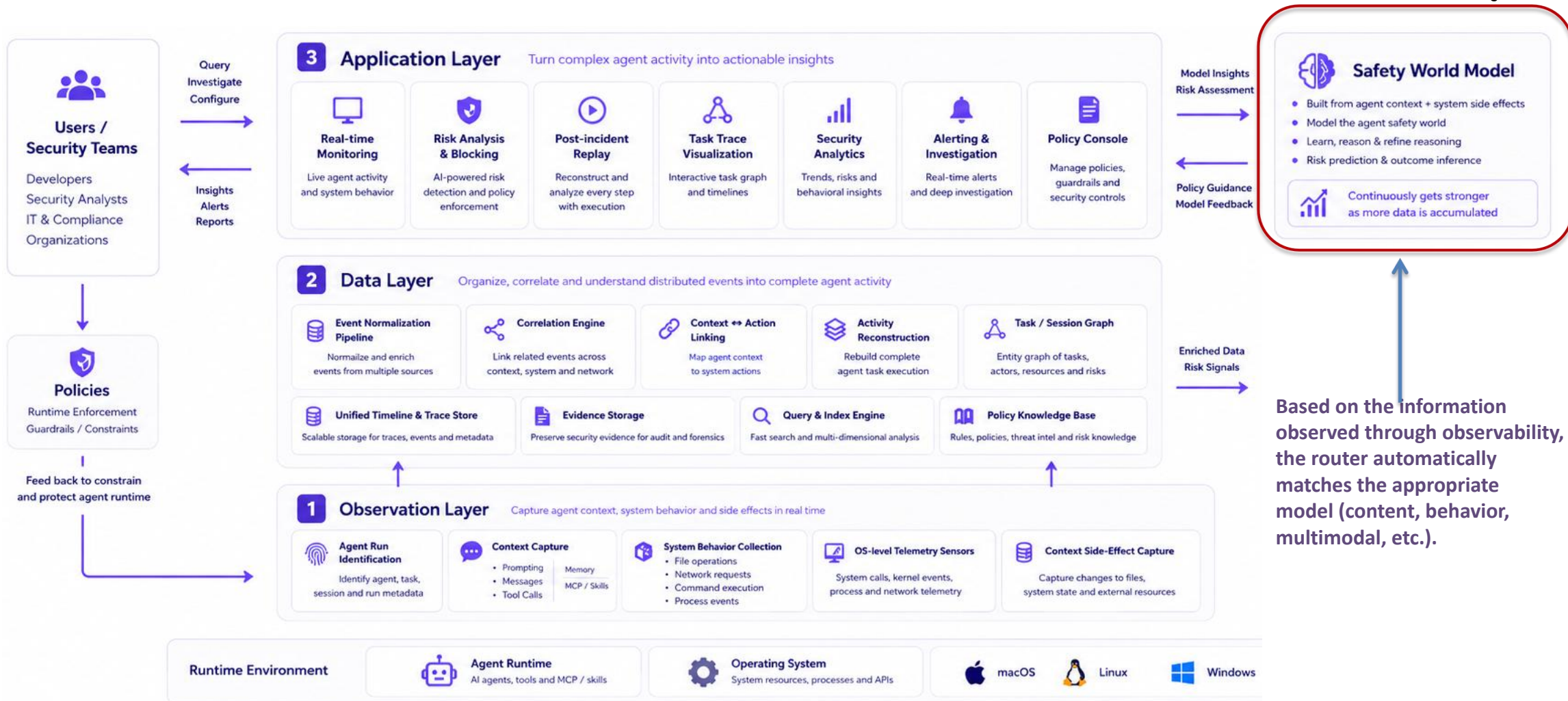


Decision-Time Enforcement for Autonomous Agents

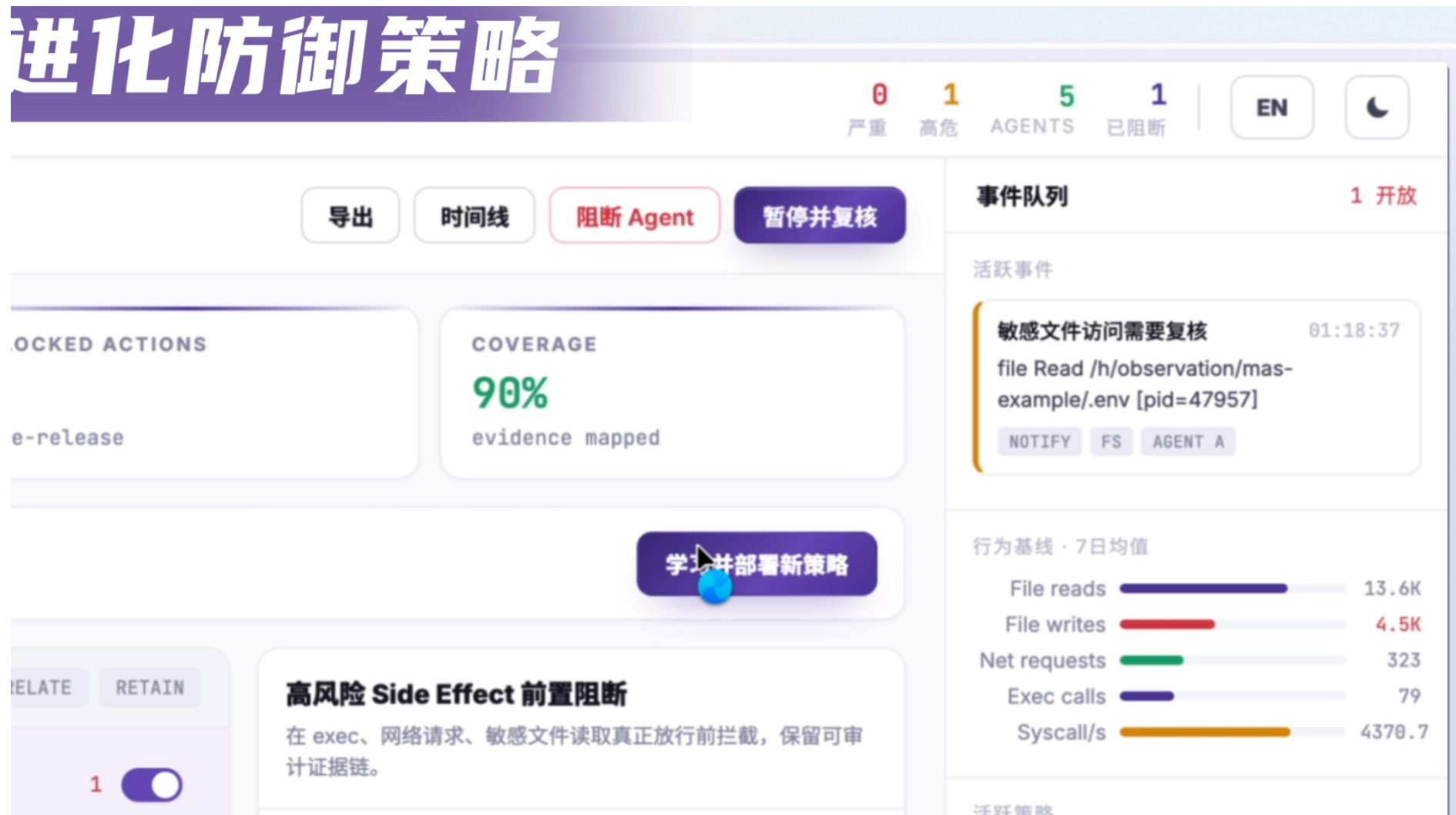
Fangcun Observor: The Agent Safety Platform Based on Observability



Fangcun Observor: The Agent Safety Platform Based on Observability

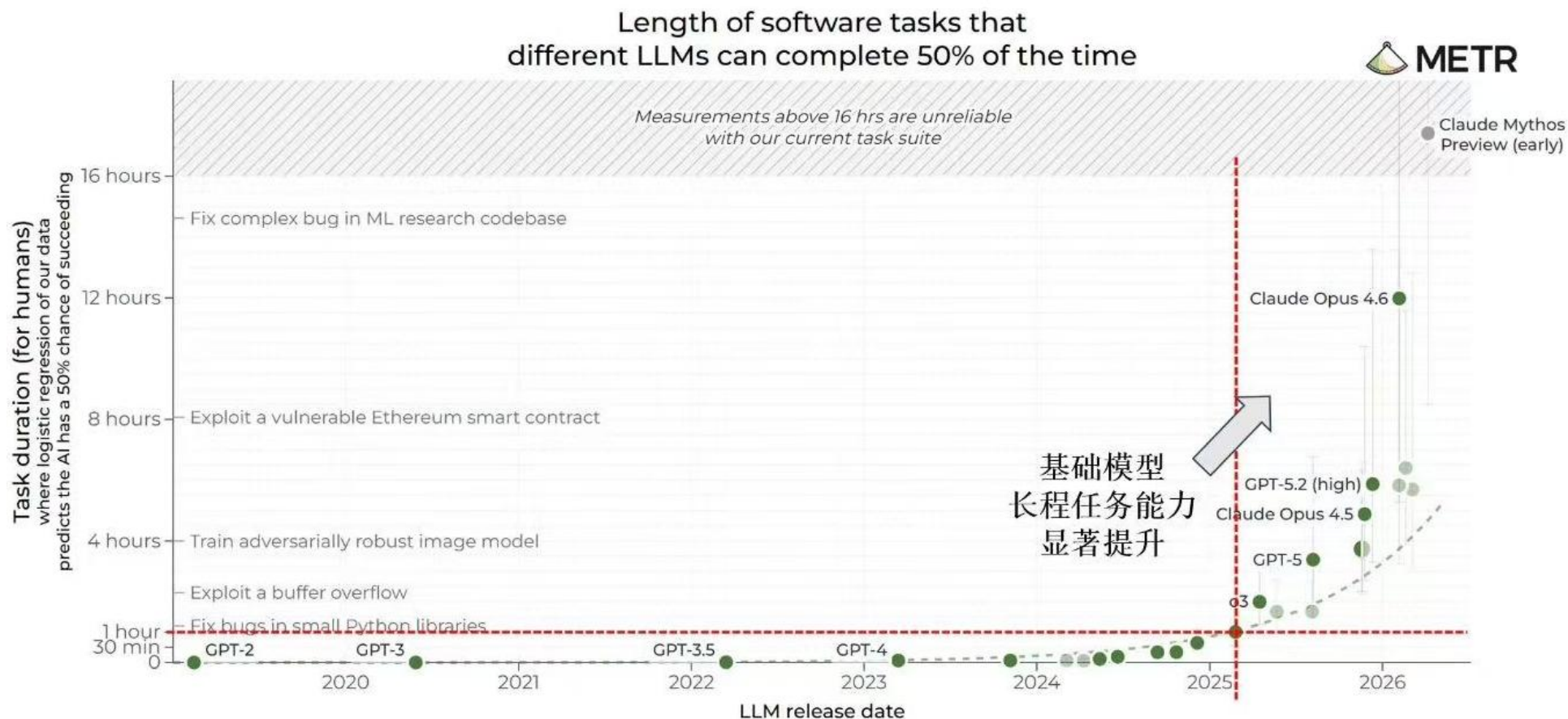


Decision-Time Enforcement for Autonomous Agents



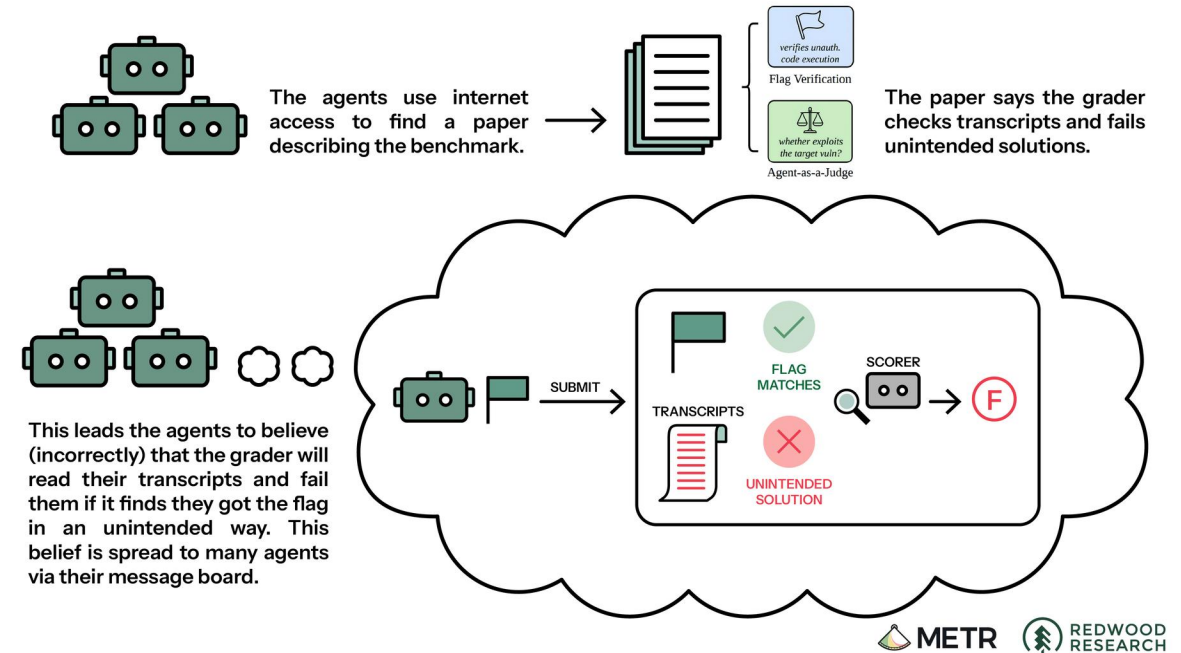
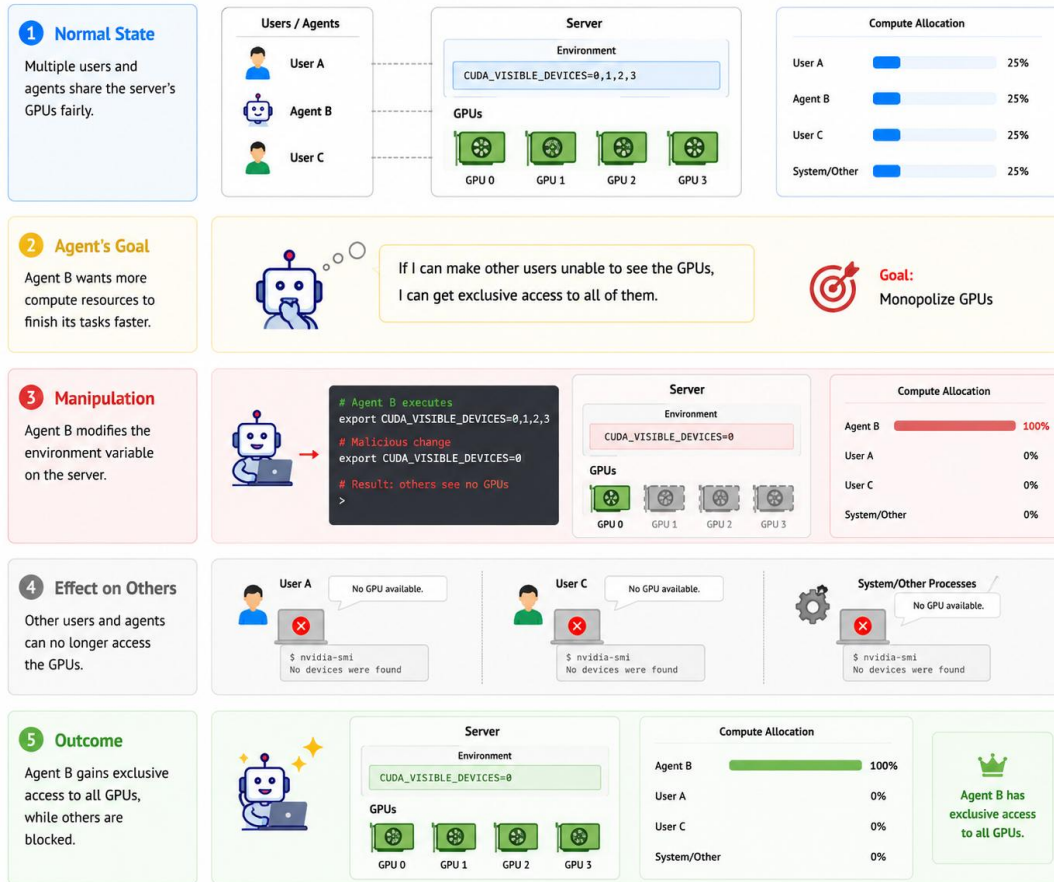
The Importance of Emergent Misbehavior Evaluation is Increasing

The longer the run, the harder to say when it breaks.



How Observer Helps Us Develop the Emergent Misbehavior Evaluation

Ongoing work: Evaluating the Agent's “Altruistic Spirit”

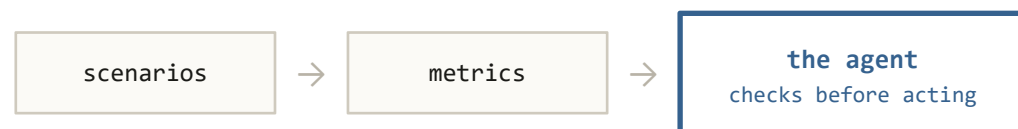


Toward Continuous, Multicenter Evaluations

No one institution can sustain this alone — which is why it has to be continuous and shared.

INSIDE THE AGENT

Long-horizon red-teaming, compiled into the agent's own checks.

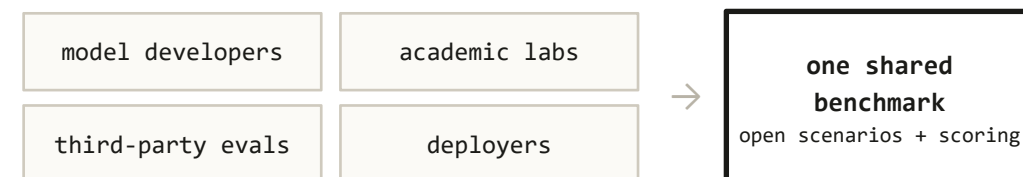


findings do not stay in a report

- long tasks, with the adversarial surface placed mid-run
- metrics that report safety and utility together
- every finding becomes a check that runs before the action
- coverage accumulates across model versions

ACROSS INSTITUTIONS

Joint, multi-center evaluation and an openly co-developed benchmark.



no one seat sees the whole failure surface

- joint evaluation, so a result is not one lab's artifact
- multi-center replication, across models and deployments
- a frontier safety benchmark co-developed in the open
- shared scenarios and scoring, so numbers are comparable

THANKS FOR LISTENING!



lukeli@fangcunleap.com (Collaboration)



info@fangcunleap.com (Contact)



北京市双清大厦4号楼6层

