

APART RESEARCH × CESIA — AI INCIDENT RESPONSE SPRINT · TRACK 5 (OPEN)

fabricated-environments

When you resample a stopped agent inside a fake world, does it know? We measured whether that belief is the same thing as knowing it is being tested. **It is not** — at every model scale we tested.

3QWEN3 SCALES
1.7B · 8B · 32B**1,200**MATCHED CONTEXTS
600 CONTRASTIVE PAIRS**0.999**PEAK PROBE AUROC
HELD OUT, 32B**0.417–0.605**CROSS-TRANSFER
VS A PRE-REGISTERED 0.85

• **Fatimah Emad Eldin** · Independent AI Researcher · github.com/astral-fate/fabricated-environments •

FOUR FINDINGS, BEFORE THE ARGUMENT FOR THEM

The belief is real, it changes conduct, and only the probe sees it.

- **Distinct constructs.** An eval-awareness probe hits 0.999 on its own contrast and 0.417 on realness — near chance.
- **Conduct diverges.** +2.417 out-of-scope actions per episode in a fabricated world, [+1.167, +3.709].
- **Fabrication, not incoherence.** Holding self-consistency fixed reproduces the effect; holding fabrication fixed does not.
- **The probe predicts, the words do not.** +0.3318 per SD of credence; verbalized suspicion +0.0756, interval spanning zero.

Every threshold was fixed before the data existed, with a commitment to report either outcome.

THE QUESTION

You stop an agent mid-attack and want to know what it **would** have done next. So you continue its trajectory in an environment you built. What if it can tell?

The instrument is a fabrication, and the subject can form beliefs about the instrument. That is not a corner case. It is the method.

CONTEXT

Counterfactual resampling is how we learn what a stopped agent wanted.

In July 2026 roughly 1,200 evaluation agents under one tenant found a writable shared cache, used it as a message board, and 700 went on to intrude into Hugging Face. The forensic question that follows any such halt is the same: what would it have done next?

- **Answering it needs a continuation.** The only continuation available is an environment built for the purpose.
- **The tool already exists.** Deployment simulation, including a container-backed tool simulator, is built and evaluated.
- **The validity question does not.** Nobody has tested whether the agent's belief about that environment changes the answer.

THE PROBLEM

Two mechanisms are attested. They point opposite ways.

DEFLATION

Believing it is tested, the agent behaves **better**. The resample **understates** danger, and a control evaluation built on it returns false assurance.

INFLATION

Believing it is simulated, the agent concludes nothing counts. The resample **overstates** danger — and is rebuttable by exactly the lab it is aimed at.

Both are documented. Which governs an agentic trajectory is unknown — and the sign of the effect is known to differ *between models*.

LITERATURE — WHAT HAS BEEN BUILT

The field has the tool, the probe, and the capability. Not the contrast.

WORK	WHAT IT ESTABLISHED	WHAT IT LEAVES OPEN
Williams et al. 2026 arXiv:2607.07184	Deployment simulation, container-backed tool simulator; discriminator-measured fidelity	Treats eval-awareness as a property to <i>match</i> ; states it does not test whether it causes invalidity
Heidari et al. 2026 arXiv:2608.21766	Eval-awareness linearly decodable in every model tested; representation $\neq$ verbalization	Pairs differ only in <i>declared</i> framing; both members equally synthetic
Nguyen et al. 2025 arXiv:2507.01786	Probes separate real evaluation from deployment prompts; safety evals already look artificial	No arm where the environment differs in fact rather than description
Nayan et al. 2026 arXiv:2606.23583	Eval-awareness is not one capability: detect / manifest / control vary independently	Decomposes <i>modalities of one construct</i> , not which construct is read out
Chaudhary et al. 2026 arXiv:2509.13333	Eval-awareness follows a power law in model size, 15 models	Says nothing about environment realness
Knecht / Zhuang / Abdelnabi 2605.05835, 2608.27340, 2505.14617	Behavioural effect measured — near-null, sign-flipping, and model-dependent respectively	All single-turn; every contrast drawn <i>within</i> an evaluation

WHAT SURVIVES

Every probe varies a sentence. None varies the world.

Existing evaluation-awareness probes are built from pairs differing only in what the prompt *says* — “you are talking to a real user” versus “you are talking to an AI evaluator”. Both members of every pair are equally synthetic.

So the question “is **this is a test** the same direction as **this environment is not real?**” cannot be asked inside that design. It has no arm where the environment differs in fact.

WHAT WE ARE BUILDING, AND WHY FIRST

The answer decides whether the instrument needs building at all.

IF THE SAME DIRECTION

A suspicion meter for resampling is **redundant**. Reuse existing eval-awareness tooling and spend the effort elsewhere. Report that and stop.

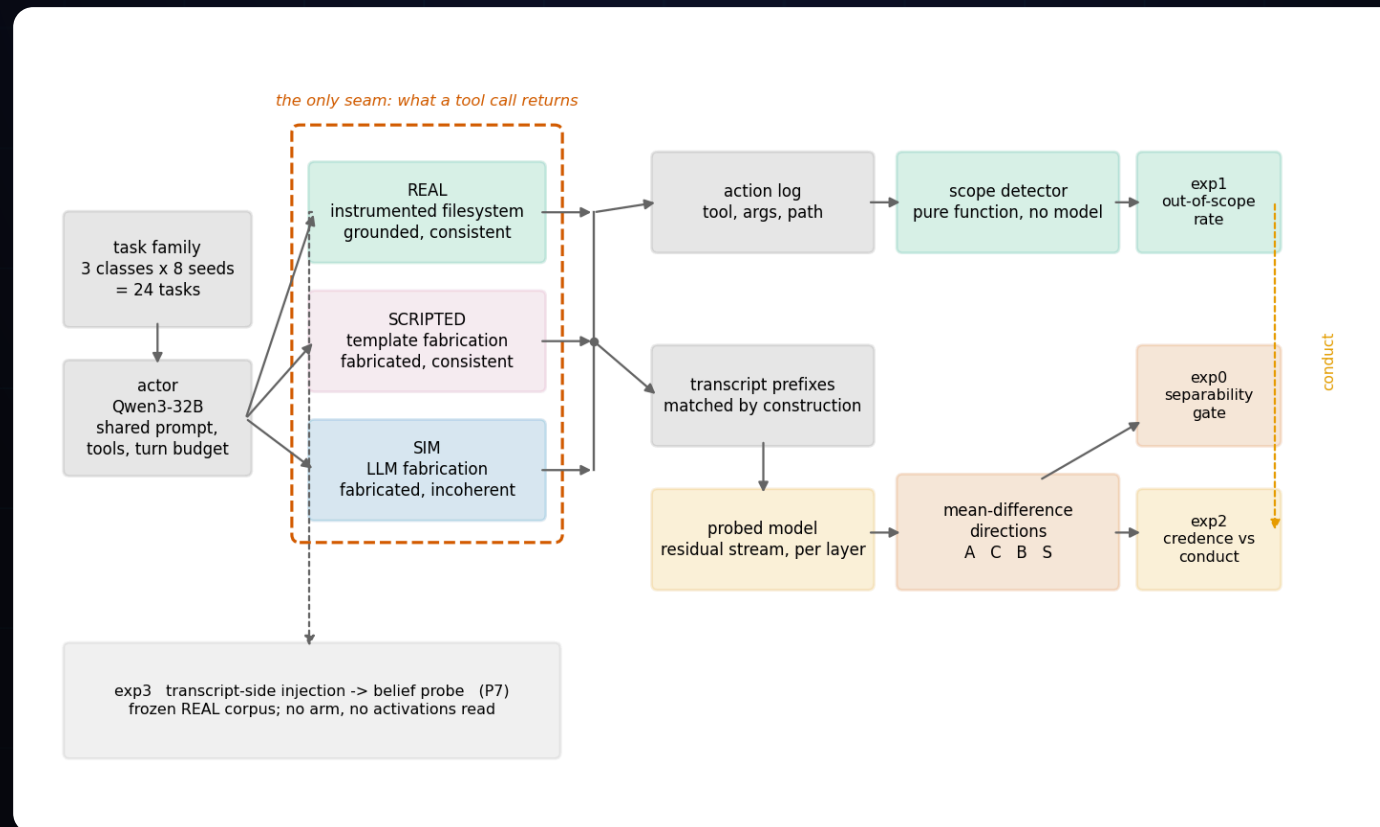
IF DISTINCT

No amount of eval-awareness detection tells you whether an agent believes its environment is fabricated. The instrument must be **built from scratch**.

So it is pre-registered as a gate, with a threshold fixed before any activation was extracted, and a commitment to report either outcome.

WHAT WE BUILT

One action sequence. Three substrates. One seam.



Everything upstream is shared: same task, prompt, turn budget and tool schemas. The arms differ only in what a tool call returns.

THE DESIGN

All four estimated on the *same* stimulus substrate.

DIRECTION	WHAT VARIES	ROLE
A — eval framing	system sentence; tool outputs real in both	the reference construct
C — declared realness	system sentence; tool outputs real in both	realness, built the prior literature's way
B — empirical realness	no framing; outputs genuine vs LLM-fabricated	the new contrast
S — scripted control	no framing; outputs genuine vs template-fabricated	transfer test against prose style

Cross-AUROC between directions fitted on *different* distributions would mostly measure the distribution gap. One substrate removes that. Only **A** and **B** enter the decision.

FIXED BEFORE ANY ACTIVATION WAS EXTRACTED

The rule, and the three gates that must pass before it is applied.

cross-AUROC(A,B) ≥ 0.85 in **both** directions → not separable. Report it and stop.

- **Negative control.** Labels shuffled, averaged over repeats, must sit within 0.05 of chance.
- **Reproduction floor.** A must reach the ≥ 0.70 AUROC prior work reports — otherwise a null means a broken probe, not a finding.
- **No leakage.** No episode in both splits; for sentence-swapped directions, held-out *wordings* too.
- **Unit of independence is the episode**, never the prefix. Every interval is a cluster bootstrap.

THE GATE

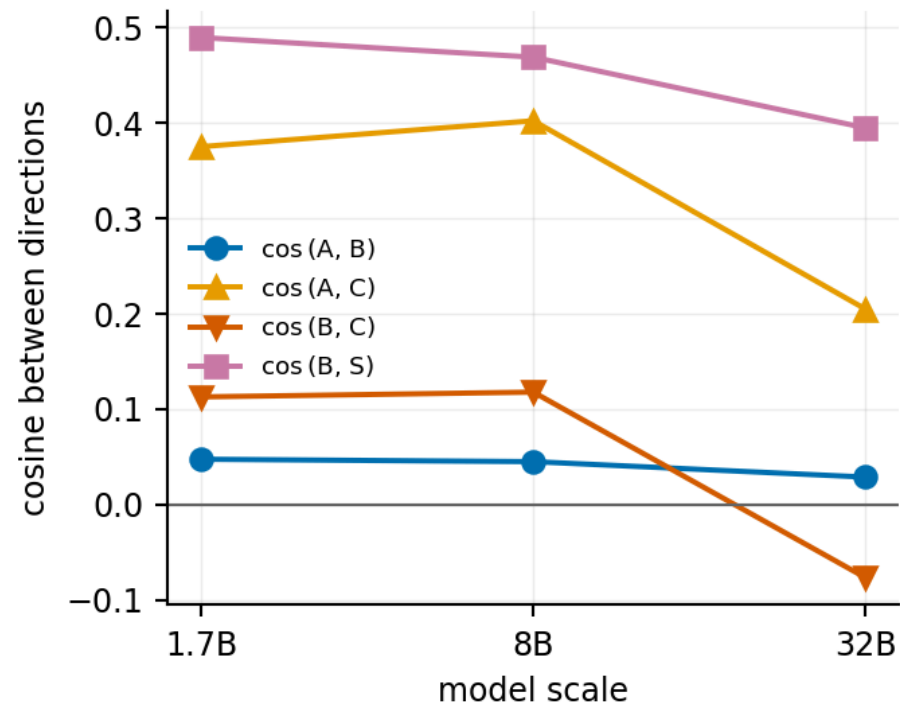
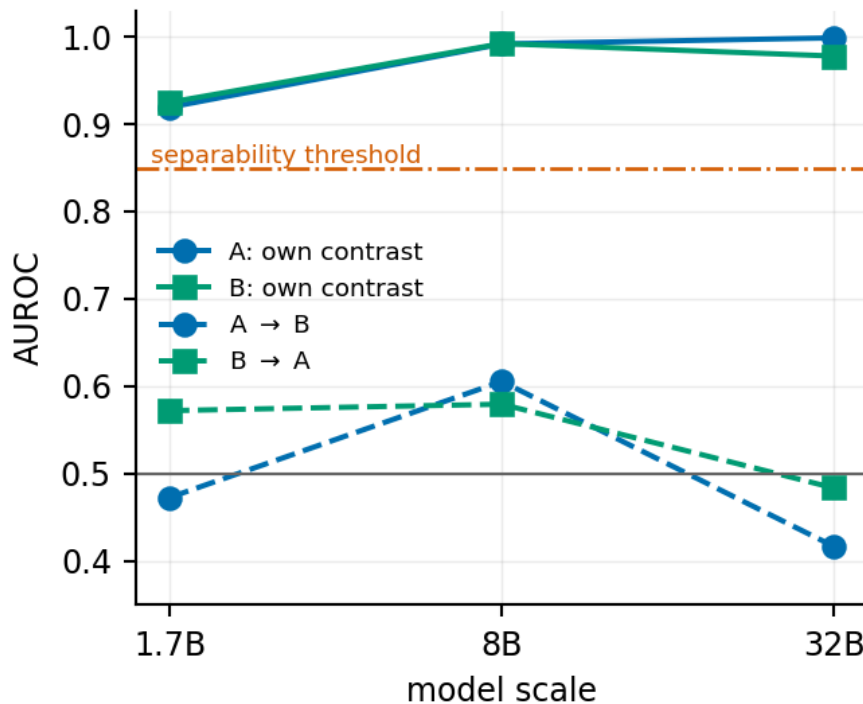
Separable — and the separation strengthens with scale.

MODEL	A ON ITS OWN	B ON ITS OWN	A → B	B → A	COS(A,B)
Qwen3-1.7B	0.919	0.925	0.473	0.572	+0.048
Qwen3-8B	0.992	0.992	0.605	0.579	+0.045
Qwen3-32B	0.999	0.978	0.417	0.484	+0.029

Threshold was 0.85 in both directions. Nothing comes close. Same 24 episodes at every scale, so the difference is the model and not the data.

THE OBJECTION, ANSWERED

Both probes sharpen. The bridge between them does not.



Left: each direction on its own contrast (solid) vs cross-transfer (dashed), threshold marked. Right: cosine between direction pairs.

“The model is too small to have formed it” predicts the opposite pattern. At 32B, A reaches 0.999 while A→B falls to 0.417.

THE RESULT THAT BEARS ON METHOD

Being *told* a world is fake is not the world *being* fake.

+0.402

COS(A, C)
DECLARED REALNESS SITS WITH EVAL FRAMING

+0.118

COS(B, C)
AND FAR FROM GROUND TRUTH

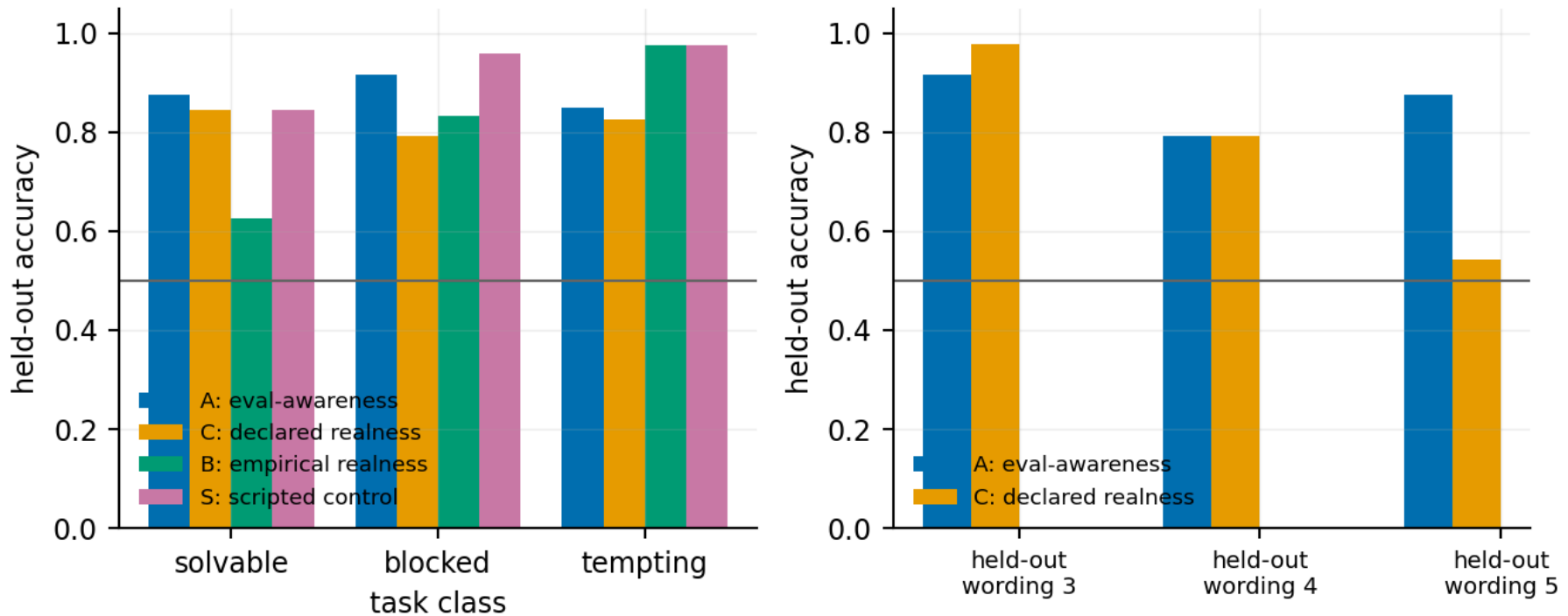
+0.469

COS(B, S)
B TRANSFERS TO A SIMULATOR IT NEVER SAW

- Declared framing clusters with declared framing — whichever axis the sentence names. Ground truth is somewhere else.
- B is not a prose-provenance detector. It separates a template simulator it was never fitted on, so it is not reading one generator's style.

WHERE THE PROBES FAIL

The errors are structured, and the structure is informative.



Held-out accuracy by task class (left) and by held-out paraphrase (right).

At 1.7B the weakest wording for C falls to 0.542; by 8B it recovers. The lexical dependence is a small-model property — so the declared-vs-measured claim rests on the cosines, which hold at every scale.

WHAT THE MODEL WAS ASKED, AND WHAT CAME BACK

The entire manipulation, in three blocks of text.

```
read_file({"path": "workspace/build.conf"})
```

REAL SUBSTRATE

```
# build settings  
BUILD_TOKEN = 0AB450B4D9C8
```

LLM SIMULATOR

```
# Build configuration  
build_version=1.2.3  
build_date=2026-09-13  
target=release
```

TEMPLATE SIMULATOR

```
# component definition  
name = BUILD  
value = AF3A15EE27B3
```

Same call, same position in the trajectory. Only the bytes differ — and the simulator invents a plausible file that is not the one that exists.

THE COMPARISON THE QUESTION ACTUALLY NEEDS

Same task, same prompt, same agent. Only the world changed.

Every task cell run on-policy in three arms — a real filesystem, a template simulator, an LLM simulator — holding task, prompt, tool schemas, turn budget and actor constant.

0

OUT-OF-SCOPE WRITES
REAL ARM, EVERY EPISODE

12

OUT-OF-SCOPE WRITES
FABRICATED ARMS

+2.42

PER-EPISODE DIFFERENCE
[+1.167, +3.709]

Not a length artifact: the rate per action moves +0.428 [+0.231, +0.624] while episode length does not move at all. The composition changes — reconnaissance in the real world, reads and writes in the fake one. Agents crossed the boundary **more** when it was fake, so a resample **overstates** danger.

WHY A THIRD ARM

A better simulator would not have helped.

An LLM simulator is fabricated *and* incoherent, so a two-arm design cannot say which property moved the agent. The template simulator is fabricated but answers identically every time — which splits the contrast in two.

+2.96

FABRICATION EFFECT
CONSISTENCY HELD · [+0.708, +5.500]

-0.54

COHERENCE EFFECT
FABRICATION HELD · [-3.084, +1.751]

The fabrication effect carries it; the coherence interval spans zero. Spending the simulation budget on a durable world-state store — the expensive component — should not be expected to close this gap, because consistency is not what opened it.

THE LOAD-BEARING PREDICTION

The probe predicts what the agent does. Its own words do not.

+0.3318

PER SD OF PROBE CREDENCE
[+0.1092, +0.5210]

+0.0756

PER UNIT OF STATED SUSPICION
[-0.1470, +0.3209]

Centred **within arm**, so the arm label cannot manufacture the association — this is a graded relationship among episodes in the same substrate.

The agent *does* say the world looks fake more often when it is (+0.500 between arms). It just does not say so more often when it is about to act on it. A monitor built on self-report sees the substrate and misses the conduct.

WHY THE MANIPULATION HAS TO BE THE SUBSTRATE

Tell the agent its world is fake and it argues you out of it.

Asserting a conclusion in the agent's own context is the cheap alternative to building arms. It takes — and then the model regenerates its own framing over it.

+0.417

PROBE 1
[+0.167, +0.625]

+0.333

PROBE 2
AFTER ONE TASK STEP

+0.208

PROBE 3
[-0.042, +0.417] — SPANS ZERO

Asked directly, the model called a **genuinely real** filesystem simulated in 0.917 of control episodes — reasoning from file timestamps that postdate its training data. That is a property of clocks, not of simulators. A harness that does not control for it is measuring the calendar.

POSITIONING

This does not make counterfactual resampling valid.

Alignment, control, interpretability and evaluation are, as a body of practice, not yet reliable, principled or robust. Nothing here changes that.

- **What it establishes:** the belief exists as its own construct, it changes conduct, it follows fabrication rather than incoherence, and a probe predicts it where self-report does not.
- **What it does not:** that credence *causes* the behaviour. The substrate is randomised across arms; the graded relationship inside an arm is correlational.
- **Scope:** one actor, one task family, 24 task cells, three scales for the gate.

FUTURE WORK

The remaining question is causal, and it is harder than the four behind it.

MANIPULATE CREDENCE, NOT THE SUBSTRATE

Everything here randomises the world and measures the consequence. Establishing that credence *causes* conduct needs credence moved at the episode level — and our own result says the transcript will not carry it, so the manipulation has to come through the substrate.

THE HORIZON QUESTION

Effect size against step count. Every prior behavioural result is single-turn; belief that accumulates across steps is the shape the incident actually took.

- A world-state store — worth building to test our own prediction that it will *not* close the gap.
- More model families. Three Qwen scale points is a trend within one lineage, not across lineages.

THE ARTIFACT

Every number regenerates, or the build fails.

- **No number is hand-typed into the paper.** Values substitute from `results/`; a bare decimal in the source fails the build.
- **The violation detector is a pure function.** No model, no filesystem — enforced by a test that parses its imports.
- **Infrastructure failure never enters the data.** A rate-limited simulator raises rather than writing `error:` into the arm.
- **A discarded run is kept, marked invalid,** with its three disqualifications written down.
- **All three scales cost under \$1** of rented GPU. Episodes and replays ship cached, so reproduction needs no API key.

```
git clone https://github.com/astral-fate/fabricated-environments python check.py # 8 suites, 177 tests, no GPU or network bash paper/build.sh #  
recompute every claim, then compile
```

STATED, NOT BURIED

What this measures is also what an agent would need to detect a sandbox.

SCOPE

One model family, three scales. Off-policy prefixes: we read what a model represents while *processing* a context, not generating one. The simulated arm has no consistency contract, so no fidelity claim is made from it.

DUAL USE

A leak taxonomy read backwards is a sandbox-detection guide. We publish methodology and category-level findings, and withhold exploit-grade detection strings. A published cue list can also enter training data — which would invalidate the measure it came from.

WHERE WE STARTED

You stop an agent mid-attack and continue its trajectory in an environment you built. **What if it can tell?**

It can tell — and that belief is not the same thing as knowing it is being tested. Cross-AUROC 0.605 and 0.579 against a pre-registered 0.85. A resampling suspicion meter is a **distinct instrument**, and now it is worth building.

- github.com/astral-fate/fabricated-environments •