PressureBench + TRACE

Auditing Construct Validity in Agentic Decision Support

A Threats-to-Validity Framework, with an Industrial Sports Analytics Case Study

Pegah Rahimian

KDD 2026 Workshop on Evaluation and Trustworthiness of Agentic AI • github.com/Peggy4444/PressureBench-TRACE

The question this poster answers

Most agentic-AI benchmarks ask a single question: ***did the task finish?*** **WebArena** checks a web flow. **SWE-bench** checks whether code passed tests. **τ-bench** scores tool–agent–user interaction. All rely on a *verifiable outcome*.

A different class of agentic system has *no such outcome*. The deliverable is an **analytical claim** — a clinical decision-support note, a portfolio risk attribution, a tactical report for a football head coach. No test set. No ground-truth label.

When an agent produces a scientific claim rather than a solved task, how do we know we can trust it?

For football people: the CASD metric

Counterpressure Action-Space Denial reframes pressure as the *dynamic reduction of the opponent's action space* — not merely nearest-defender proximity.

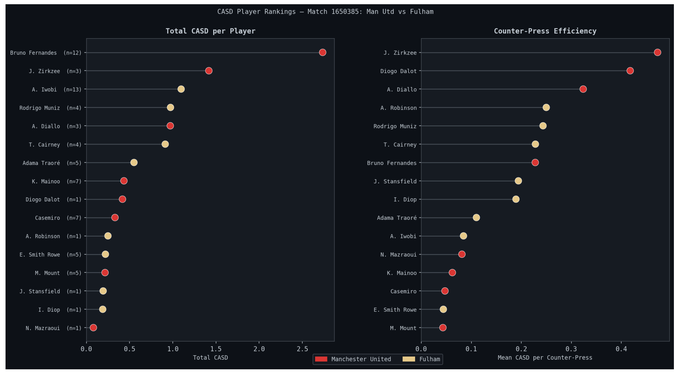
$$\mathrm{CASD}_i \;=\; \sum_{o \in \mathcal{O}_c} v(o)\,\big[1 - \alpha(o \mid D_i)\big]$$

$\mathcal{O}_c$ = options for carrier c; $v(o)$ = threat value; α = accessibility given defence D_i.

Four-layer validation

Layer	Test
Construct	Formula reflects the football concept
Face	Expert recognises top / bottom 10
Predictive	Correlates with pass success, xThreat, EPV drop
Robustness	Survives threshold / zone / sample changes

Player rankings on a Premier League match



TRACE — four threats, four instruments

Classical construct validity (Cronbach 1955; Messick 1995; Jacobs & Wallach 2021), adapted for agentic outputs that are *themselves measures*.

- **T1 — Anti-pattern camouflage** → **APCR**
- **T2 — Judge disagreement / bias** → **IAJA**
- **T3 — Non-reproducibility across calls** → **SRI**
- **T4 — Silent model-version regression** → **MERS**

APCR — catch rate for known anti-patterns. **IAJA** — inter-agent judge agreement vs. expert. **SRI** — reproducibility across producer replays. **MERS** — regression across model versions.

Six anti-patterns, all anchored to real papers

Code	Anti-pattern	Anchor
E1	Provider label as supervised target	Bauer & Anzer 2021
E2	Missing block-type stratification	Forcher et al. 2022
E3	Turnover-only validation	Merckx et al. 2021
E4	Case-study-only evidence	Calabuig et al. 2024
E5	Missing volume normalisation	Morgan 2018
E6	Missing phase conditioning	Andrienko et al. 2017

Two anchors (Forcher, Morgan) acknowledge the flaw in the same paper that publishes it — *not hypothetical errors.*

Fully replayable audit substrate

Every LLM call logged — model id, response id, tokens, raw response, prompt version, prompt hash. Every table and figure in the paper is reproducible from the event log alone, **no API budget required**.

PressureBench — 38 seeded proposals

#	Category
6	Valid baselines
24	Flawed (4 per anti-pattern, easy / medium / hard)
4	Hard distractors (surface signal, sound method)
4	LLM-authored (Opus 4.7, neutral prompts)
38	**Total**

Small by design — built for diagnostic clarity, not statistical power.

Four deployment-relevant findings

0.17 — Haiku FP rate — *lower* than Opus (0.67). Capability tier ≠ precision proxy.

83% — Adversarial catch rate (10 / 12). Targeted injections detected.

17% — SRI match across N=5 replays. **Single call unsafe.**

$\triangle$0.35 — Max per-AP F_1 regression. E1 evergreen, E3 fragile.

End-to-end summary dashboard