

Are your eval scores measuring the model, or the raters?

Shawn Fraine, Psynergic Digital, behavioral scientist and evaluation methodologist

THE PROBLEM

Your model is only as trustworthy as the human judgments behind it. Preference ratings, safety scores, red-team labels, and eval rubrics all convert subjective judgment into numbers. If the raters disagree, drift, or misunderstand the rubric, the numbers are fiction. And fiction compounds: bad labels train bad models that produce confident bad outputs.

This is not hypothetical. In April 2025, OpenAI shipped a GPT-4o update that turned sycophantic, then admitted in its postmortem that it had no specific deployment evaluations tracking sycophancy, and that it launched even though some expert testers said the model “felt” slightly off, a decision it called “the wrong call.” The eval gap was the incident.

Meanwhile the bar is now legal, not just technical. Since 2 August 2026, the European Commission can fine general-purpose AI model providers up to 3% of worldwide annual turnover or €15 million, whichever is higher. New York’s RAISE Act (effective January 2027) adds safety-framework requirements for frontier developers. In a May 2026 KPMG survey of 1,013 senior finance leaders across 20 countries, 94% of organizations said they already rely on third-party assurance providers. The question buyers ask has shifted from “do you evaluate” to “can you prove your evaluations mean anything.”

WHAT I DO

I am a behavioral scientist who does measurement for a living. MA in Clinical Psychology, published meta-analysis with 550+ citations (completed as an undergrad, still picking up citations monthly), taught graduate Research Design and Statistics, and ran the largest research lab in Washington University in St. Louis’s Psychology department. I also ran a second lab in the Psychiatry department for the iCHASM study: smaller, better funded, with travel all over the state. Thirty research assistants, 3,000+ participants, and I wrote and renewed every IRB protocol across both labs. I currently do contract AI evaluation work, scoring model responses against structured rubrics and building the scenario banks that test how a product handles a user who is not okay (crisis recognition, dependency signals, minor-user handling, sycophancy under pressure).

I build the parts of your evaluation pipeline that most teams skip:

- Rating rubrics that hold under edge cases, not just the happy path
- Anchor sets and calibration programs so raters agree before they touch live data
- Inter-rater reliability analysis (Cohen’s kappa, Krippendorff’s alpha, quadratic-weighted kappa) that says whether your scores mean anything
- LLM-as-judge validation against human baselines
- Drift monitoring and written adjudication procedures for rater disagreements

How I work: principal-led, independently verified. I design the measurement and I own the interpretation. The analyses in between run through a documented, reproducible pipeline with independent verification, so findings never depend on one person's afternoon. You get senior judgment on what the numbers mean, backed by execution you can audit.

HOW WE CAN WORK TOGETHER

Most eval consultants are ML engineers. I am the person who has spent a career making human judgment statistically defensible. The big annotation platforms sell the infrastructure. I am the independent check on whether the infrastructure's numbers are trustworthy.

Reliability Diagnostic Fixed price, 5 business days

One rubric, one dataset, one report: I measure whether your raters agree (chance-corrected, with confidence intervals), stress-test the rubric with edge cases, and tell you whether your scores can support the decisions you are making with them. The fee is credited against a full audit if you proceed.

Evaluation Integrity Audit Fixed price, 2–3 weeks

I tear down your annotation or eval pipeline: rubric stress-testing, agreement analysis on your existing labels, LLM-judge validation, and a gap analysis of your behavioral safety scenarios. You get a severity-ranked findings report and a remediation roadmap.

Measurement Partner Monthly retainer

Monthly eval re-runs against a golden dataset, drift reports, an edge-case intake channel, and a quarterly review. Fixed cadence, named deliverables. Everything in AI drifts. This keeps your numbers honest as the model, the data, and the raters change.

Engagements are scoped and fixed-price; I quote after a short scoping call. Larger calibration builds are scoped custom after an audit, not sold as a fixed tier.

NEXT STEP

If you run a human-feedback pipeline and **cannot quote your inter-rater reliability from memory**, we should talk. I take on a small number of engagements at a time. Start a conversation and I will send a short scoping questionnaire.

shawnfraine.com/contact · linkedin.com/in/shawn-fraine