PITON Global

EXECUTIVE WHITE PAPER · WP-50

MODEL EVALUATION & RLHF · PHILIPPINES

The eval-discipline standard: the executive economics of model evaluation and RLHF outsourcing to the Philippines.

An analysis of why evals run is a volume vanity metric, how eval discipline and preference-data quality — never judgment throughput — decide the true cost of a model program once missed regressions, false alarms, grader drift, reward hacking, and shipped defects are counted, and the vendor-selection discipline that produces a verdict you can ship on.

AUTHORS

John Maczynski — Chief Executive Officer
Ralf Ellspermann — Chief Strategy Officer

JUNE 2026
MANILA · ESTABLISHED 2001

Contents

ABOUT THIS SERIES — VOLUME 50

Every eval-and-RLHF proposal is priced per judgment — per rating, per comparison, per prompt graded — because volume is easy to count. This volume is about the only evaluation outcome a model program actually depends on: the verdict you can trust enough to ship on, and the preference data clean enough to move the model the right way. The gap between running an eval and running one that catches the regression is where shipped defects and reward hacking live. The full series lives at piton-global.com.

THE PITON-GLOBAL LEADERSHIP TEAM

About the authors

John Maczynski

Chief Executive Officer

John has bought evaluation capacity for three decades of quality work and the last several for AI programs, and learned that the fast grader who rates every prompt and misses the regression is the most expensive one a model team can hire — because an eval that passes a broken model ships the defect to users, and preference data collected loosely teaches the model to game the reward. His standing question ignores the judgments-per-hour report: of the evals you ran, how many would you actually ship on? Applied at PITON-Global, that question is what clients get on every evaluation sourcing decision — personally, with no account layer in between.

j.maczynski@piton-global.com
+1 402-598-8740

Ralf Ellspermann

Chief Strategy Officer

Ralf has spent twenty-five years running Philippine delivery floors and the last several building the eval-and-RLHF pods now central to AI programs, and he judges an evaluation team by its grader calibration and regression-catch rate, not its judgments logged: whether a blind re-grade agrees with the panel, and whether the preference data survives a reward-model audit. The eval contract in Section 3 codifies what he learned rebuilding teams that rated fast and drifted. He now reads evaluation providers the way he once ran them — from the inter-grader agreement and the caught-regression log backward.

r.ellspermann@piton-global.com
Manila, Philippines

PITON Global

ABOUT PITON-GLOBAL

PITON-Global is a buyer-side advisory practice, Manila-based since 2001, that has never operated a delivery floor or taken a placement commission. Its stock-in-trade is knowing the Philippine provider market at the operating level — 1,000+ mapped operations, re-verified continuously, including each evaluation team's real grader calibration and regression-catch rate rather than its judgments-per-hour headline — and converting that knowledge into a shortlist of six to ten genuinely qualified partners per engagement. Buyers from Tripadvisor, Garmin, TSYS, HealthPartners, Yetipay, Workrise, and Norman Disney & Young have run their searches through the practice.

SECTION 01

Executive summary

−57%
Fully-loaded cost per trustworthy eval judgment vs. US onshore

0.87+
Inter-grader agreement (κ) on vetted pods, up from 0.55–0.70

−71%
Reduction in regressions shipped to production, vetted pods

6.2×
First-year ROI, reconstructed case (Section 6)

An evaluation-and-RLHF engagement is bought on the price per judgment and paid for on the verdicts it gets wrong. The vendor reports hundreds of thousands of prompts graded and comparisons ranked at a fraction of the onshore cost; the model team finds the regression an eval passed and shipped to users, the false alarms that blocked a good release and burned a week of debugging, the grader drift that made this month's scores incomparable to last month's, and the preference data so noisy the reward model learned to game it. The pod hit its judgments-per-hour number and the program inherited a shipped defect, a trust problem with its own metrics, and a reward signal pointing the wrong way. The gap between running an eval and running one you can ship on is where the real cost of cheap evaluation hides, because an untrustworthy judgment is not throughput delivered — it is a defect shipped or a model aligned in the wrong direction.

The Philippines — English-fluent, analytically trained, and now with a mature pool of calibrated AI graders and preference annotators — is where evaluation and RLHF are delivered with the most eval discipline, because that pool, properly calibrated, can run the blind judgment, agreement measurement, and reward-data hygiene that trustworthy evaluation demands. Five findings:

1 **Evals run is the industry's most flattering number.** Any pod can rate prompts at speed. Vetted pods are measured on outcome: grader calibration, regression-catch rate, and preference-data quality. Weak providers quote a low per-judgment price the program repays in shipped defects, false alarms, and a reward signal it cannot trust. Section 2 shows the gap.

2 **The trustworthy verdict, not the fast rating, is the product.** A judgment you can ship on comes from a defined rubric, a blind protocol, and calibrated graders whose agreement is measured — not from clearing a grading queue. Pods that evaluate with discipline catch the regression before release; pods that rate loosely pass it through and cost far more than the judgments saved.

3 **A bad eval cuts both ways — missed regression and false alarm.** Miss a real regression and you ship a worse model; raise a false alarm and you block a good release and burn engineering days chasing a phantom. Vetted pods hold agreement above 0.85 κ precisely because both errors cost — the defect that gets through and the good release that gets stopped.

4 **Reward hacking, not the per-judgment rate, is the hidden cost.** Noisy or inconsistent preference data teaches the model to satisfy the grader rather than the user, and that damage surfaces downstream where it is expensive to trace. Pods that keep preference data clean and calibrated align the model the right way, and that trustworthy reward signal — not the offshore rate alone — is where the case study's return is built.

5 **Contract on the trustworthy judgment, not the logged one.** The contracting failure of evaluation is paying per judgment and hoping the graders agree and the data is clean. Vetted deals price against agreement floors, regression-catch targets, and preference-data-quality standards, making trustworthy verdicts the pod's problem, which is where it belongs.

The intended reader is a head of AI, ML platform lead, or research operations director buying evaluation and RLHF per judgment and paying for it in shipped defects, false alarms, and reward hacking. Sections 2–4 separate the pods that evaluate with discipline from the pods that log judgments; Sections 5–7, the method and the proof.

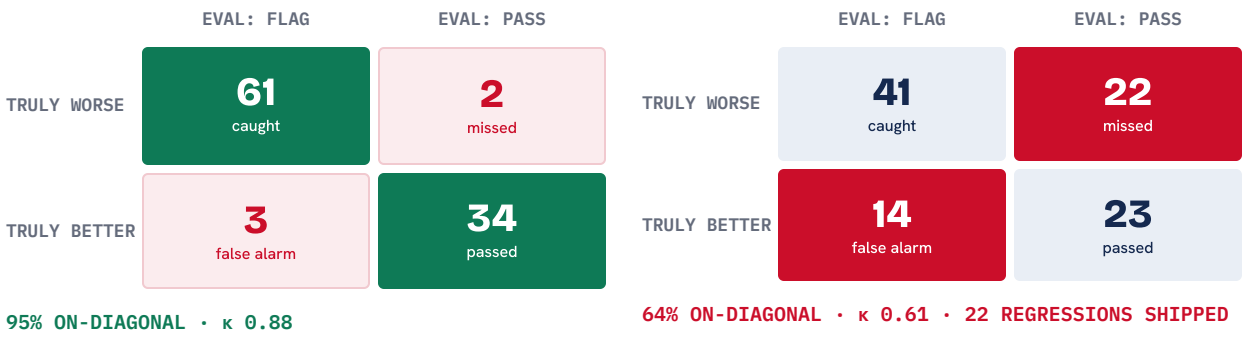
SECTION 02

The volume mirage: evals run

versus regressions caught

Figure 1 reads one release-gate eval as a pair of confusion matrices — what the eval said against what was actually true of the model. The vetted pod's judgments land on the diagonal (caught the regression, passed the good release); the weak vendor's scatter off it (missed regressions shipped, false alarms raised). Both pods "ran the eval." Only one produced a verdict you could ship on.

FIG. 1 — ONE RELEASE-GATE EVAL: VERDICT TRUTH TABLE, VETTED POD VS. WEAK VENDOR



Same 100 model-change judgments, scored against a held-out ground-truth panel. Left: PITON-Global engagement instrumentation, 2025–26. Right: composite of pre-engagement vendor reporting reviewed in diligence.

The economics follow the diagonal, not the judgment count. On the left, ninety-five percent of verdicts match ground truth and only two regressions slip — the eval is a gate you can ship on; on the right, twenty-two regressions ship and fourteen good releases get blocked, so the program pays in shipped defects, wasted debugging, and a metric nobody trusts. A low per-judgment price that lands off-diagonal is not a saving; it is a defect pipeline and a false-alarm tax. The eval-discipline audit — grader agreement, regression-catch rate, false-alarm rate — is the cheapest due-diligence instrument in this paper.

SECTION 03

The eval contract: define the criteria, judge it blind, close the loop

Figure 2 lays out the three clauses of a working eval-and-RLHF contract as a closed loop — the machinery a floor read verifies before any per-judgment price is believed. Unlike a one-shot rating, a trustworthy eval program is a cycle: criteria feed blind judgment, judgment feeds the reward loop, and measured drift feeds re-calibration back into the criteria.

FIG. 2 — THE EVAL CONTRACT, AS A CLOSED LOOP

1 · DEFINE THE CRITERIA

A written rubric fixes what "better" means for this model and task — helpfulness, safety, factuality, format — with anchored examples and a tie-break rule, so graders judge the same thing the same way rather than by private taste.

→

2 · JUDGE IT BLIND

Candidates are graded blind to model identity, order-randomized to kill position bias, with overlap samples measuring inter-grader agreement live — so the verdict reflects the response, not which system produced it or which came first.

→

3 · CLOSE THE LOOP

Clean preference data feeds the reward model, disagreements are adjudicated and fed back as rubric fixes, and grader drift is monitored and re-calibrated — so the eval improves the model in the right direction and stays comparable over time.

↺ MEASURED DRIFT & DISAGREEMENT RE-CALIBRATE THE CRITERIA

THE TRUSTWORTHY VERDICT — κ 0.87+ · CATCHES REGRESSIONS · SHIP-GRADE REWARD DATA

Criteria are defined, judgment is blind and agreement-measured, and the loop closes with clean reward data — so regressions are caught before release, false alarms fall, and the model aligns the right way. The program gets cheaper per trustworthy judgment and safer to ship at the same time, which is the entire promise of outsourcing evaluation to people who measure eval discipline, not judgments logged.

Operating parameters from PITON-Global eval-and-RLHF engagements, 2025–26. Client owns rubrics, preference data, and reward models; graders operate under client research direction.

In diligence, the contract is tested from the agreement log's side: seed a known-regression set and measure the catch rate; re-grade an overlap sample blind and compute κ; audit a slice of preference data for consistency and reward-hacking signatures. Perhaps one pod in ten survives it cleanly. Steps 3 and 5 of the framework exist to find that one before the contract does — because everything in Table 3 of Section 6 depends on eval discipline, not judgments logged.

SECTION 04

Cost and performance: cost-per-trustworthy-

eval economics and benchmarks

TABLE 1 — FULLY-LOADED COST PER TRUSTWORTHY EVAL JUDGMENT: THREE OPERATING MODELS (2026)

OPERATING MODEL	COST / JUDGMENT	VS. US BASELINE
US onshore in-house eval / research staff	$4.10	baseline
Low-cost offshore rating mill (per-judgment)	$2.70	−34%
Vetted Manila pod — calibrated, blind, looped	$1.76	−57%

Basis: PITON-Global engagement pricing 2025–26; cost per trustworthy eval judgment = fully-loaded operating cost (including calibration, overlap grading, and QA) divided by judgments that agree with a ground-truth re-grade and survive preference-data audit (off-diagonal and drifted judgments excluded); the rating mill looks cheap per judgment but missed regressions, false-alarm debugging, and reward-hacking cleanup lift its true cost per trustworthy judgment; US comparator fully-loaded in-house cost.

TABLE 2 — EVAL & RLHF BENCHMARKS, VETTED VS. BASELINE (2025–26)

METRIC	VETTED TOP TIER	BASELINE	EXECUTIVE READING
Inter-grader agreement (κ)	0.87+	0.55–0.70	The only eval number ship-decisions follow
Regression-catch rate	97%+	65–80%	Caught at the gate, or shipped to users
False-alarm rate vs. baseline	−64%	baseline	"Judge it blind" — held or not
Preference-data reward-model win-rate lift	+2.6×	baseline	Aligns the model, or teaches it to game
Grader drift (month-over-month κ loss)	<0.03	0.10–0.18	Comparable over time, or not

Source: PITON-Global eval-and-RLHF engagement data 2025–26; baseline = typical pre-engagement operations billed on judgments run. κ = Cohen's/Fleiss' kappa on overlap samples.

The strategic read: the rating mill's −34% per judgment is a false economy — missed regressions, false-alarm debugging, and reward-hacking cleanup lift its true cost per trustworthy judgment far above the per-judgment gap, and a defect shipped to users on a passed eval is a cost no rate recovers. The vetted Manila pod's −57% is real because it is measured where the value is: the verdict that agrees with ground truth and

the preference data that aligns the model. Buy on judgments run and you buy the right matrix of Figure 1; buy on cost-per-trustworthy-eval and you buy eval discipline — the same standard every volume in this series keeps arriving at from a different door.

SECTION 05

Sourcing discipline: the 7-Step Vendor Vetting Framework, applied to evaluation and RLHF

Eval diligence adds one hard requirement to every step: a per-judgment price is real only when the grader agreement and regression-catch rate behind it survive a blind re-grade against a ground-truth panel. The funnel below is representative of a 40–90-seat eval-and-RLHF search.

FIG. 3 — THE 7-STEP FRAMEWORK AS AN EVAL-AND-RLHF FUNNEL

	STEP	WHAT IT ELIMINATES IN AN EVAL SEARCH	FIELD REMAINING
1	**Requirements definition**	Throughput-first scoping: fixes the agreement floor (κ), the regression-catch target, the false-alarm ceiling, and the preference-data-quality standard the pod must meet by task family	1,000+
2	**Market mapping**	Rating mills: operations that bill judgments and treat missed regressions and drift as the client's problem, and generalist floors with no calibration protocol or research-grade domain graders	~44
3	**Capability screening**	Pods without the machinery: measured inter-grader agreement, a rubric-and-calibration workflow, blind/randomized protocols, reward-data QA, and domain depth across your task families	~13
4	**Compliance & security audit**	Weak governance: model-output and prompt confidentiality, IP and data-handling controls, secure grading environments, red-team/safety-content protocols, SOC 2 scope covering the eval stack	~8
5	**Operational diligence**	The blind re-grade: seed a known-regression set and measure catch; re-grade an overlap sample for κ; audit preference data for reward-hacking; interview on how they adjudicate a rubric edge case	4–6
6	**Competitive RFP & negotiation**	Per-judgment pricing: finalists bid against agreement floors, regression-catch targets, and preference-data-quality standards — all carrying credits for judgments that fail a ground-truth re-grade	2–3
7	**Transition & launch governance**	Cold start: the pod calibrates against your rubric and a gold set before it grades live release gates, with the agreement floor certified before steady state, PITON-Global embedded buyer-side	1

Field-remaining counts represent a 40–90-seat eval-and-RLHF search, PITON-Global engagements 2024–26.

The standing guarantees hold: **vendor neutrality** — PITON-Global operates no eval floors, grades no prompts, and takes no placement economics, so the shortlist reflects audited agreement and regression-catch performance, not inventory; **right-sizing** — pods for whom your program is a flagship account. Six to ten genuinely qualified providers, delivered within 48–72 hours, free to the buyer, competing through a fully transparent, fair, comprehensive, and highly competitive RFP.

SECTION 06 · ANONYMIZED CASE STUDY

Re-basing the program on eval discipline: a 55-seat eval-and-RLHF rebuild, reconstructed

Engagement EV-064 is a US foundation-model and applied-AI company — running release-gate evals, red-team scoring, and RLHF preference collection across several model lines — that had offshored evaluation to a per-judgment rating mill. The judgment price was excellent; the discipline was not: inter-grader agreement sat near 0.62 κ, a regression had shipped on a passed eval, false alarms were burning engineering weeks, and a reward-model audit found preference data noisy enough to be teaching the model to game the grader. The head of AI wanted the throughput without the shipped defects and the reward hacking. Identifying details are anonymized under NDA.

SELECTION (WEEKS 0–8, BLIND RE-GRADE)

Requirements fixed a 0.85 κ agreement floor, a 97% regression-catch target, a false-alarm ceiling, and a preference-data-quality standard, with mandatory calibration and overlap grading. Mapping produced 44 claimants; the machinery screen — measured agreement, rubric/calibration workflow, blind protocols, reward-data QA — left 13; the confidentiality-and-safety audit left 8. Blind re-grades against a seeded ground-truth panel shortlisted 5. A Manila pod won priced against agreement and regression-catch, with a standing blind re-grade written into the reporting spec.

TRANSITION (MONTHS 2–7, CALIBRATED FIRST)

The pod calibrated against the company's rubric and a gold set before grading live release gates; the agreement floor was certified in month 3, and blind/randomized protocols plus overlap grading ran from day one. The disagreement-adjudication loop fed rubric fixes weekly. By month 7 agreement reached 0.88 κ, regression-catch hit 97%, false alarms fell 64%, and a re-audit found the preference data clean, lifting reward-model win-rate 2.6×. PITON-Global chaired the eval-discipline review board to steady state.

TABLE 3 — EV-064 FIRST-YEAR VALUE BUILD-UP (US$)		
VALUE COMPONENT	YEAR 1	BASIS
Labor arbitrage on the 55-seat eval-and-RLHF pod	$1.58M	103.1k productive hrs × ($24.80 − $9.50)
Shipped regressions prevented	$0.94M	Incident, rollback, and trust cost avoided at the gate
Engineering time reclaimed (false alarms cut)	$0.42M	Researcher/engineer days off phantom-regression chases
Reward-hacking cleanup avoided	$0.24M	Re-collection and re-training avoided on clean data
Gross value created	**$3.18M**	
Less: engagement & transition costs	($0.51M)	Calibration, gold-set build, protocol/tooling integration; advisory fee $0 (provider-paid model)
Net value ÷ cost = first-year ROI	**6.2×**	$3.18M ÷ $0.51M

Figures rounded; agreement, regression-catch, and reward-data effects audit-verified. Method in Methodology & notes.

Fifty percent arbitrage, fifty percent from prevented regressions, reclaimed engineering time, and avoided reward-hacking cleanup — a typical split for the series, earned because the program was re-based on eval discipline rather than judgment volume. Every line traces to a diligence step: the agreement floor to Step 1, the agreement-based pricing to Step 6, the calibration-and-loop workflow to Step 3's contract. The head of AI's line at the year-one review: "The judgment price barely moved. What moved is that I trust the eval now — we stopped shipping regressions, and the reward model finally points the right way."

SECTION 07

The executive playbook: governance for the first 180 days

An eval-and-RLHF transition succeeds when the pod is measured and paid on grader agreement and regression-catch — client in command throughout.

DAYS 0–30

Define discipline, contract on it

Set the agreement floor (κ), the regression-catch target, and the false-alarm and preference-data-quality thresholds with AI leadership before launch, and document the rubric, gold set, and blind/randomized protocol by task family. Contract against agreement and regression-catch — never per judgment alone — with floors and ceilings carrying credits. Stand up the blind re-grade as the program's arbiter of truth.

DAYS 30–90

Calibrate, prove the floor

Calibrate the pod against your rubric and gold set before it grades live release gates; certify the agreement floor and the overlap-grading loop. Switch on the disagreement-adjudication loop from day one. Publish the eval dashboard: inter-grader κ, regression-catch rate, false-alarm rate, preference-data quality, month-over-month drift.

DAYS 90–180

Grade live gates, certify steady state

Route live release-gate evals to the pod only after it holds the agreement floor and regression-catch target, always under client research direction. Review the eval board monthly with the caught-regression and drift trends on the table. Certify steady state on two months at floor including a model-family change; hand over on documented criteria; keep the standing blind re-grade.

THE FIVE FAILURE MODES WE ARE HIRED TO PREVENT

Paying per judgment while regressions ship on passed evals · rating fast instead of judging blind against a rubric · letting graders drift until scores are incomparable · collecting noisy preference data that teaches reward hacking · grading live gates before agreement is proven. All five are settled at selection and contract.

The evaluation argument, in one sentence: a logged judgment is activity and a trustworthy verdict is a regression caught before release and a model aligned the right way, and a vetted Philippine pod delivers −57% cost per trustworthy eval judgment with 0.87+ grader agreement, regressions-shipped cut by over two-thirds, and a 2.6× reward-model lift — for buyers who blind-re-grade the agreement instead of counting judgments run, and compete the finalists through a fully transparent, fair, comprehensive, and highly competitive RFP. Anyone can rate a prompt. We make sure you buy the pod whose verdict you can ship on.

METHODOLOGY & NOTES

How the numbers were built

Agreement and regression figures (Fig. 1, Table 2) derive from engagement instrumentation in PITON-Global eval-and-RLHF engagements, 2025–26. Inter-grader agreement is measured as Cohen's/Fleiss' kappa on overlap samples; regression-catch and false-alarm rates are measured against a seeded ground-truth panel; preference-data quality is measured by reward-model win-rate lift and a reward-hacking audit. The weak-vendor column is a composite of pre-engagement vendor reporting reviewed in diligence, shown for contrast.

Cost-per-trustworthy-eval figures (Table 1) are fully loaded — wages, benefits and statutory charges, facilities, tooling, management and QA overhead including calibration leads and overlap grading, attrition-driven recruiting — divided by judgments that agree with a ground-truth re-grade and survive preference-data audit, with off-diagonal or drifted judgments excluded from the denominator; US comparators reflect fully-loaded in-house cost on the same basis. A trustworthy eval judgment agrees with ground truth and is safe to ship on.

The case study (EV-064) is anonymized under NDA; volumes and dates are generalized, figures rounded. Agreement, regression-catch, and reward-data effects are audit-verified; the prevented-regression line reflects avoided incident and rollback cost, and the reward-hacking line is a conservative estimate. ROI = net first-year value ÷ all engagement-related client costs. PITON-Global's advisory fee is provider-paid and appears at $0 in the client cost base; neutrality safeguards are documented at piton-global.com.

Market context draws on IBPAP reporting and PITON-Global's proprietary provider mapping (1,000+ operations; grader agreement and regression-catch performance re-verified May 2026). Evaluation and RLHF pods operate under client research direction and never substitute for the client's own release judgment. Estimates and projections are PITON-Global analysis and should be cited as such. The full series is available at piton-global.com.

Would your provider's judgment price survive a blind re-grade?

Thirty minutes with John will tell you. Free to buyers, strictly vendor-neutral — an agreement-verified shortlist within two to three business days.

j.maczynski@piton-global.com
+1 402-598-8740