LLM TRAINING & FINE-TUNING · PHILIPPINES

The capability-taught standard: the executive economics of LLM training and fine-tuning data outsourcing to the Philippines.

An analysis of why examples labeled is a volume vanity metric, how instruction quality and eval-verified taught capability — never labeling throughput — decide the true cost of a fine-tuning dataset once wrong demonstrations, off-policy answers, reward hacking, and eval regressions are counted, and the vendor-selection discipline that teaches the model what you actually wanted.

AUTHORS

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

JULY 2026
MANILA · ESTABLISHED 2001

Contents

ABOUT THIS SERIES — VOLUME 49

Every fine-tuning-data proposal is priced per example labeled, because volume is easy to count. This volume is about the only SFT outcome a model owner actually depends on: the demonstration that teaches the model the behavior you wanted, across the distribution it will see, verified on a held-out eval. The gap between labeling an example and teaching a capability is where eval regressions and wasted training runs 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 AI-data capacity for three decades of the outsourcing era, and learned that the fast vendor that labels every example and demonstrates the wrong behavior is the most expensive one a model owner can hire — because a fine-tuning set full of off-policy or sloppy demonstrations teaches the model exactly that, and the regression only shows up after an expensive training run. His standing question ignores the throughput dashboard: of the examples you paid for, how many actually moved the eval in the direction you wanted? Applied at PITON-Global, that question is what clients get on every SFT-data 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 data floors, and he judges an SFT-data team by its instruction quality and eval-verified lift, not its examples-per-day: whether the demonstrations taught the target behavior across the real distribution, or drifted off-policy and regressed the model. The SFT contract in Section 3 codifies what he learned rebuilding teams that labeled fast and taught the wrong thing. He now reads fine-tuning-data providers the way he once ran them — from the eval scorecard and the reward-hacking 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 SFT-data team's real instruction quality and eval-verified lift rather than its examples-per-day 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

-58%	**97%+**	**+2.4×**	**6.4×**
Fully-loaded cost per capability-grade example vs. US onshore	Instruction-quality (accept) rate on vetted teams, up from 76–86%	Eval-verified capability lift per labeling dollar vs. baseline	First-year ROI, reconstructed case (Section 6)

A fine-tuning-data engagement is bought on the price per example and paid for on the demonstrations that teach the model the wrong thing. The vendor reports a million examples labeled at a fraction of the onshore cost; the model owner finds the off-policy demonstrations that pull the model away from its target behavior, the sloppy chains of thought that teach shortcuts, the coverage gaps where the real prompt distribution was never demonstrated, and the eval regression that a costly training run only revealed at the end. The operation hit its examples-labeled number and the model team inherited a dataset that made the model worse in the ways that mattered. The gap between labeling an example and teaching a capability is where the real cost of cheap SFT data hides, because a bad demonstration is not throughput delivered — it is negative training signal you paid to inject.

The Philippines — English-fluent, degree-deep, and now several years into building AI-data delivery for frontier labs and applied-AI teams — is where fine-tuning data is produced with the most instruction-quality discipline, because that talent pool can write the correct demonstrations, cover the distribution, and hold the eval bar that taught capability demands. Five findings:

1. **Examples labeled is the industry's most flattering number.** Any vendor can generate examples at volume. Vetted teams are measured on outcome: instruction quality, distribution coverage, and eval-verified lift. Weak providers quote a low per-example price the model owner repays in regressions, re-labeling, and wasted training compute. Section 2 shows the gap.

2 **The correct demonstration, not the labeled row, is the product.** A dataset that improves a model comes from demonstrating the target behavior correctly, across the real distribution, checked against a rubric — not from filling a labeling quota. Teams that demonstrate correctly teach the capability once; teams that label loosely inject noise the model faithfully learns.

3 **A bad example cuts both ways — wrong answer and wrong process.** A demonstration with the right answer but a hacked or shortcut rationale teaches the model to game the objective; a wrong answer teaches it to be wrong. Vetted teams hold instruction quality above 97% precisely because both failure modes get trained in and only surface on eval.

4 **The wasted run and the regression, not the label, are what cost.** A dataset built wrong costs the training compute, the eval investigation, the re-labeling, and the shipped-capability delay — many times the per-example saving. Teams that demonstrate correctly convert labeling dollars into eval-verified lift, and that lift-per-dollar — not the offshore rate alone — is where the case study's return is built.

5 **Contract on the capability-grade example, not the labeled one.** The contracting failure of SFT data is paying per example and hoping it helps. Vetted deals price against instruction-quality floors, coverage targets, and eval-verified-lift minimums, making taught capability the team's problem, which is where it belongs.

The intended reader is a head of model training, applied-AI lead, or data-operations owner buying fine-tuning data per example and paying for it in regressions, re-labeling, and wasted compute. Sections 2–4 separate the teams that teach capability from the teams that label rows; Sections 5–7, the method and the proof.

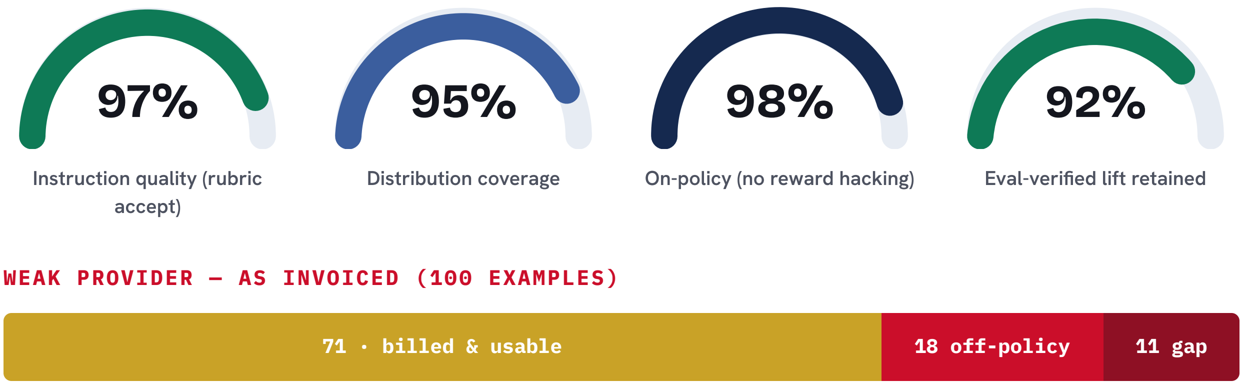
SECTION 02

The volume mirage: examples labeled versus capability taught

Figure 1 reads one fine-tuning batch two ways. The four dials are the vetted team's quality gauges — the outcomes eval actually rewards; the bar beneath is the same batch as a weak vendor invoices it, a full quota of "examples labeled" with the share that quietly regresses the model called out. The labeled count is identical in spirit to a chatbot's containment rate — engineered to look like output while off-policy demonstrations and coverage gaps accumulate underneath.

FIG. 1 — ONE SFT BATCH: TAUGHT-CAPABILITY QUALITY (DIALS) VS. REPORTED VOLUME (BAR)

VETTED TEAM — MEASURED ON TAUGHT CAPABILITY



Every example was "labeled" — but 18% demonstrated off-policy behavior and 11% fell in coverage gaps the eval punishes; nearly a third was negative or null training signal.

Left dials: PITON-Global engagement instrumentation, 2025–26, rubric- and eval-audited. Right bar: composite of pre-engagement vendor reporting reviewed in diligence.

The economics follow the taught capability, not the labeled count. On the dials, instruction quality and on-policy rate sit at the high-nineties and the eval-verified lift is retained, so labeling dollars become model capability; on the bar, a cheap per-example price ships eighteen percent off-policy and eleven percent in coverage gaps, and the training run faithfully learns all of it. A low per-example price that injects negative signal is not a saving; it is a regression you paid to train and will pay again to undo. The instruction-quality-and-eval audit — rubric accept rate, coverage, on-policy rate, held-out lift — is the cheapest due-diligence instrument in this paper.

SECTION 03

The SFT contract: demonstrate it correctly, cover the distribution, verify on eval

Figure 2 lays out the three clauses of a working fine-tuning-data contract as a build ladder — each rung must hold before the next means anything, and a per-example price is believable only when a floor read confirms all three.

FIG. 2 — THE SFT CONTRACT, AS A BUILD LADDER

1

DEMONSTRATE IT CORRECTLY

Every example shows the target behavior with a correct answer and a sound, on-policy rationale — checked against a written rubric, not a labeler's intuition — and an ambiguous prompt is escalated to a spec owner rather than guessed. The demonstration is made correct at the source, so the model learns the behavior you meant.

2

COVER THE DISTRIBUTION

The example mix mirrors the prompt distribution the model will actually face — the long tail, the hard slices, the adversarial edges — so no capability is taught only on the easy center. Coverage is tracked against a taxonomy, and gaps are commissioned deliberately rather than back-filled with whatever is cheap.

3

VERIFY ON EVAL

A held-out eval and a training probe confirm the batch moves the target metric before the full run, and reward-hacked or regression-inducing examples are caught and pulled — with an eval scorecard the model owner can read. Verification on eval is the only proof that labeling became capability, which is the only outcome the model actually feels.

THE CAPABILITY-GRADE EXAMPLE 97%+ INSTRUCTION QUALITY EVAL-VERIFIED

Demonstrations are correct, coverage matches the real distribution, and the batch is verified to lift the eval before the run — so labeling dollars convert to model capability and regressions are caught upstream. The dataset gets cheaper per capability-grade example and safer per training run at the same time.

Operating parameters from PITON-Global AI-data engagements, 2025–26.

In diligence, the contract is tested from the eval's side: commission a blind batch, re-grade a sample against the rubric for instruction quality, check coverage against the taxonomy, and run it through a training probe to measure real lift. Perhaps one team 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 taught capability, not examples labeled.

SECTION 04

Cost and performance: cost-per-capability-

grade-example economics and benchmarks

TABLE 1 — FULLY-LOADED COST PER CAPABILITY-GRADE EXAMPLE: THREE OPERATING MODELS (2026)		
OPERATING MODEL	COST / EXAMPLE	VS. US BASELINE
US onshore domain-expert labeling	$11.80	baseline
Low-cost offshore labeling mill (per-example)	$7.90	−33%
Vetted Manila team — rubric- & eval-verified	**$4.96**	**−58%**

Basis: PITON-Global engagement pricing 2025–26; cost per capability-grade example = fully-loaded operating cost divided by examples that pass rubric and retain eval-verified lift (off-policy, gap, and regressed examples excluded); the labeling mill looks cheap per example but re-labeling, wasted training runs, and eval investigations lift its true cost per usable example; US comparator fully-loaded onshore domain-expert cost.

TABLE 2 — SFT-DATA BENCHMARKS, VETTED VS. BASELINE (2025–26)			
METRIC	VETTED TOP TIER	BASELINE	EXECUTIVE READING
Instruction-quality (accept) rate	97%+	76–86%	The only label number capability follows
Distribution coverage vs. taxonomy	95%+	60–78%	Taught everywhere, or only the easy center
Re-label rate vs. baseline	−67%	baseline	"Demonstrate it correctly" — held or not
Eval-verified lift per $1k labeled	+2.4×	baseline	Capability bought, or compute burned
Wasted training runs vs. baseline	−71%	baseline	Ship, or debug a regression

Source: PITON-Global AI-data engagement data 2025–26; baseline = typical pre-engagement operations billed on examples labeled.

The strategic read: the labeling mill's −33% per example is a false economy — re-labeling, wasted runs, and eval investigations lift its true cost per usable example far above the per-example gap, and a shipped-capability delay is a cost no rate recovers. The vetted Manila team's −58% is real because it is measured where the value is: the example that passes rubric, covers the distribution, and moves the eval. Buy on examples labeled and you buy the bar in Figure 1; buy on cost-per-capability-grade-example and you buy taught capability — the same discipline every volume in this series keeps arriving at from a different door.

SECTION 05

Sourcing discipline: the 7-Step Vendor Vetting Framework, applied to SFT data

SFT-data diligence adds one hard requirement to every step: a per-example price is real only when the instruction quality behind it survives a rubric re-grade and a training-probe eval. The funnel below is representative of a 25–60-seat fine-tuning-data search.

FIG. 3 — THE 7-STEP FRAMEWORK AS AN SFT-DATA FUNNEL			
	STEP	WHAT IT ELIMINATES IN AN SFT-DATA SEARCH	FIELD REMAINING
1	**Requirements definition**	Throughput-first scoping: fixes the instruction-quality floor, the coverage target against the prompt taxonomy, the eval-verified-lift minimum, and the on-policy standard by capability	1,000+
2	**Market mapping**	Labeling mills: operations that bill example volume and treat off-policy demonstrations and coverage gaps as the client's problem, and generalist floors with no ML-literate labelers or eval capability	~40
3	**Capability screening**	Teams without the machinery: ML-literate and domain-expert labelers, a rubric-and-calibration workflow, coverage tracking against a taxonomy, and an eval/training-probe loop across your capabilities	~11
4	**Compliance & security audit**	Weak governance: data-provenance and licensing hygiene, PII and model-confidentiality controls, secure enclaves, IP assignment, SOC 2 scope covering the labeling stack	~7
5	**Operational diligence**	The blind batch: commission a batch, re-grade against the rubric for instruction quality, check coverage against the taxonomy, and run a training probe for real lift; interview on how they handle an ambiguous spec	4–5
6	**Competitive RFP & negotiation**	Per-example pricing: finalists bid against instruction-quality floors, coverage targets, and eval-verified-lift minimums — all carrying credits for examples that fail rubric or regress the eval	2–3
7	**Transition & launch governance**	Cold start: the team calibrates against your rubric, taxonomy, and eval harness on a pilot batch before it scales, with the instruction-quality floor certified before steady state, PITON-Global embedded buyer-side	1

Field-remaining counts represent a 25–60-seat SFT-data search, PITON-Global engagements 2024–26.

The standing guarantees hold: **vendor neutrality** — PITON-Global operates no labeling floors, produces no data, and takes no placement economics, so the shortlist reflects audited instruction-quality and eval performance, not inventory; **right-sizing** — teams for whom your operation 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 operation on taught capability: a 44-seat SFT-data rebuild, reconstructed

Engagement FT-063 is a US applied-AI company — commissioning roughly 900,000 fine-tuning examples a year across agentic, coding, and domain-reasoning capabilities — that had offshored SFT data to a per-example labeling mill. The example price was excellent; the taught capability was not: instruction quality sat near 82%, coverage skewed to the easy center, off-policy rationales were teaching shortcuts, and two training runs had regressed on the target eval. The head of model training wanted the labeling economics without the regressions. Identifying details are anonymized under NDA.

SELECTION (WEEKS 0–8, BLIND BATCHES)

Requirements fixed a 97% instruction-quality floor, a 95% coverage target against the prompt taxonomy, and an eval-verified-lift minimum, with a mandatory rubric-and-calibration and training-probe workflow. Mapping produced 40 claimants; the machinery screen — ML-literate labelers, rubric process, coverage tracking, eval loop — left 11; the provenance-and-security audit left 7. Blind batches re-graded and training-probed shortlisted 5. A Manila team won priced against instruction quality and eval-verified lift, with a training probe written into the reporting spec.

TRANSITION (MONTHS 2–6, CALIBRATED FIRST)

The team calibrated against the company's rubric, taxonomy, and eval harness on pilot batches before scaling; the instruction-quality floor was certified in month 2. The rubric-and-eval loop fed labeler feedback continuously. By month 6 instruction quality reached 97.3%, coverage hit 95%, re-labeling fell 67%, and eval-verified lift per labeling dollar rose 2.4×. PITON-Global chaired the instruction-quality-and-eval review board to steady state.

TABLE 3 — FT-063 FIRST-YEAR VALUE BUILD-UP (US$)		
VALUE COMPONENT	YEAR 1	BASIS
Labeling-cost arbitrage on the 44-seat operation	$1.58M	Onshore vs. vetted-offshore cost across 900k examples
Wasted training runs eliminated	$0.74M	Compute + engineer time on regressed runs removed
Re-labeling & rework removed	$0.30M	Lower re-label load vs. baseline
Shipped-capability acceleration value	$0.30M	Earlier capability release from clean lift
Gross value created	**$2.92M**	
Less: engagement & transition costs	($0.46M)	Calibration, rubric/eval-harness integration, pilot batches; advisory fee $0 (provider-paid model)
Net value ÷ cost = first-year ROI	6.4×	$2.92M ÷ $0.46M

Figures rounded; quality, coverage, and lift effects audit-verified. Method in Methodology & notes.

Fifty-four percent arbitrage, forty-six percent from eliminated wasted runs, removed re-labeling, and faster capability release — a typical split for the series, earned because the operation was re-based on taught capability rather than example volume. Every line traces to a diligence step: the instruction-quality floor to Step 1, the eval-verified-lift pricing to Step 6, the rubric-and-eval workflow to Step 3's contract. The head of model training's line at the year-one review: "The example price barely moved. What moved is that the runs stopped regressing — the data actually teaches the model now."

SECTION 07

The executive playbook: governance for the first 180 days

An SFT-data transition succeeds when the team is measured and paid on instruction quality and eval-verified lift — client in command throughout.

DAYS 0–30

Define quality & eval, contract on them

Set the instruction-quality floor, the coverage target against the prompt taxonomy, and the eval-verified-lift minimum with model-training leadership before launch, and document the rubric and escalation path by capability. Contract against instruction quality and eval-verified lift — never per example alone — with floors carrying credits. Stand up the rubric re-grade and training probe as the program's arbiters of truth.

DAYS 30–90

Calibrate, prove the floor on a pilot

Calibrate the team against your rubric, taxonomy, and eval harness on pilot batches before scaling; certify the instruction-quality floor and the training-probe loop. Switch on the rubric-and-eval feedback loop from day one. Publish the SFT dashboard: instruction quality, coverage, re-label rate, eval-verified lift per dollar, wasted-run count.

DAYS 90–180

Scale, certify steady state

Scale volume only after the team holds the instruction-quality floor and the eval-verified-lift minimum on the pilot. Review the eval board monthly with the coverage and wasted-run trends on the table. Certify steady state on two batches at floor across capabilities; hand over on documented criteria; keep the standing training-probe eval.

THE FIVE FAILURE MODES WE ARE HIRED TO PREVENT

Paying per example while demonstrations are off-policy · labeling to a quota instead of demonstrating the target behavior · teaching only the easy center of the distribution · shipping reward-hacked rationales that game the eval · scaling volume before instruction quality and lift are proven. All five are settled at selection and contract.

The fine-tuning-data argument, in one sentence: a labeled example is activity and a capability-grade example is a model that learned the behavior you wanted, verified on eval, and a vetted Philippine team delivers −58% cost per capability-grade example with 97%+ instruction quality and 2.4× more eval-verified lift per labeling dollar — for buyers who re-grade the quality and probe the eval instead of counting examples labeled, and compete the finalists through a fully transparent, fair, comprehensive, and highly competitive RFP. Anyone can label an example. We make sure you buy the team whose data teaches the model.

METHODOLOGY & NOTES

How the numbers were built

Quality and lift figures (Fig. 1, Table 2) derive from engagement instrumentation in PITON-Global AI-data engagements, 2025–26. Instruction quality and coverage are measured by re-grading a sample against the rubric and taxonomy; eval-verified lift, re-label rate, and wasted-run count are measured against the client's pre-engagement baseline via a held-out eval and training probe. The weak-provider figures are a composite of pre-engagement vendor reporting reviewed in diligence, shown for contrast.

Cost-per-capability-grade-example figures (Table 1) are fully loaded — wages, benefits and statutory charges, facilities, tooling and eval-harness compute, management and QA overhead including rubric reviewers and ML-literate leads, attrition-driven recruiting — divided by examples that pass rubric and retain eval-verified lift, with off-policy, gap, and regressed examples excluded from the denominator; US comparators reflect fully-loaded onshore domain-expert cost. A capability-grade example is demonstrated correctly, covers the distribution, and is verified to move the eval.

The case study (FT-063) is anonymized under NDA; volumes and dates are generalized, figures rounded. Quality, coverage, and lift effects are audit-verified; the wasted-run line reflects compute and engineer time removed, and the acceleration 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; instruction quality and eval-verified performance re-verified May 2026). 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 example price survive a rubric re-grade and a training probe?

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

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