PITON Global

EXECUTIVE WHITE PAPER · WP-51

AI DATA COLLECTION · PHILIPPINES

The corpus-fitness standard: the executive economics of AI data collection outsourcing to the Philippines.

An analysis of why samples collected is a volume vanity metric, how corpus fitness — coverage of the target distribution, label validity, and clean provenance — never collection throughput, decides the true cost of a training-data operation once off-distribution samples, mislabeled data, consent gaps, and re-collection are counted, and the vendor-selection discipline that builds a corpus a model can actually learn from.

AUTHORS

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

JUNE 2026
MANILA · ESTABLISHED 2001

Contents

ABOUT THIS SERIES — VOLUME 51

Every data-collection proposal is priced per sample, because volume is easy to count. This volume is about the only collection outcome a model's performance actually depends on: the sample that covers the target distribution, carries a valid label, and comes with clean, consented provenance — so it can be trained on. The gap between collecting a sample and collecting one that is fit for training is where model bias and re-collection 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 data-collection capacity for three decades, and learned that the fast operation that collects every sample and misses the distribution is the most expensive one an AI team can hire — because a corpus that over-samples the easy cases and skips the tail trains a model that fails exactly where it matters, and every consent-gap sample is a dataset that has to be thrown out. His standing question ignores the sample count: of the data you collected, how much was actually fit to train on — covered, labeled right, and cleanly sourced? Applied at PITON-Global, that question is what clients get on every data-collection 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 AI-data floors, and he judges a collection team by its corpus fitness — coverage, label validity, provenance — not its samples-per-day: whether the data spanned the target distribution and could be trained on, or piled up off-distribution and had to be re-collected. The collection contract in Section 3 codifies what he learned rebuilding teams that hit sample quotas and built unusable corpora. He now reads data-collection providers the way he once ran them — from the coverage audit and the consent 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 data-collection team's real corpus fitness and provenance discipline rather than its samples-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%
Fully-loaded cost per fit-for-training sample vs. US onshore

98%+
Label-validity rate on vetted teams, up from 84–92%

-64%
Reduction in re-collection on vetted teams

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

A data-collection engagement is bought on the price per sample and paid for on the samples that never should have counted. The vendor reports millions of samples collected at a fraction of the onshore cost; the AI team finds the corpus over-weighted on the easy, common cases and empty on the tail the model actually fails on, the labels that a spot-check shows are wrong, the samples with no consent record that legal makes them delete, and the collection round that has to be run again to fill the gaps. The operation hit its sample quota and the team inherited a biased corpus, a labeling-rework bill, and a compliance problem. The gap between collecting a sample and collecting one fit for training is where the real cost of cheap data collection hides, because an off-distribution, mislabeled, or unconsented sample is not a training example delivered — it is model bias and legal exposure acquired.

The Philippines — English-fluent, culturally broad, and with a mature AI-data talent pool — is where data collection is delivered with the most corpus-fitness discipline, because that pool can run the coverage design, label validation, and provenance documentation that fitness demands. Five findings:

1 **Samples collected is the industry's most flattering number.** Any floor can gather data at volume. Vetted teams are measured on outcome: distribution coverage, label validity, and clean provenance. Weak providers quote a low per-sample price the team repays in bias, rework, and deletion. Section 2 shows the gap.

2 **The fit-for-training sample, not the collected one, is the product.** Data a model can learn from comes from covering the target distribution, validating each label, and recording clean consent — not from hitting a quota. Teams that collect for fitness build a corpus that trains; teams that collect for volume build one that biases the model and gets re-run.

3 **A bad sample cuts three ways — off-distribution, mislabeled, or unconsented.** Off-distribution data biases the model; a wrong label teaches it the wrong thing; an unconsented sample is a legal liability the whole batch inherits. Vetted teams hold fitness high on all three because each failure mode alone can sink a training run.

4 **Re-collection and model bias, not the sample, are what cost.** A corpus built wrong costs the re-collection round, the labeling rework, and the model performance lost to bias — many times the per-sample saving. Teams that collect for fitness avoid the re-run and train a better model, and that avoided waste — not the offshore rate alone — is where the case study's return is built.

5 **Contract on the fit sample, not the collected one.** The contracting failure of data collection is paying per sample and hoping the corpus is usable. Vetted deals price against coverage targets, label-validity floors, and provenance standards, making a trainable corpus the team's problem, which is where it belongs.

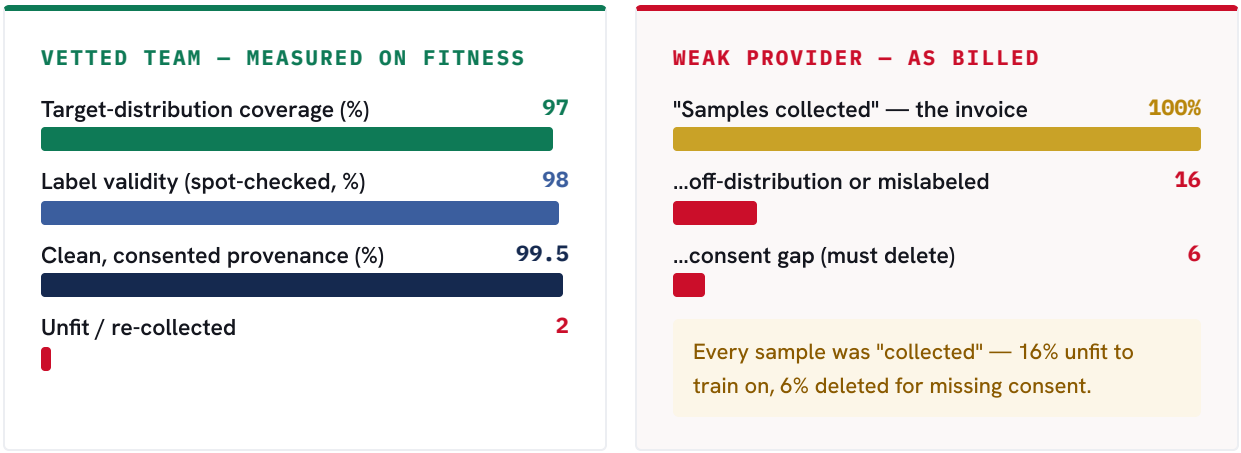
The intended reader is a head of ML data, VP of AI, or COO buying data collection per sample and paying for it in model bias, labeling rework, and compliance deletion. Sections 2–4 separate the teams that build fit corpora from the teams that hit quotas; Sections 5–7, the method and the proof.

SECTION 02

The volume mirage: samples collected versus fit-for-training data

Figure 1 shows one collection round two ways: as the samples-collected report a weak vendor bills, and as the corpus-fitness reality the coverage audit, the label spot-check, and the consent log experience. The samples-collected number is engineered to look like output while off-distribution data, mislabeling, and consent gaps accumulate underneath.

FIG. 1 — ONE COLLECTION ROUND: CORPUS-FITNESS QUALITY VS. REPORTED VOLUME



Left: PITON-Global engagement instrumentation, 2025–26, fitness audited. Right: composite of pre-engagement vendor reporting reviewed in diligence.

The economics follow the fitness, not the sample count. On the left, ninety-seven percent of the corpus covers the target distribution and provenance is clean, so the model trains on data it can learn from; on the right, a cheap per-sample price is off-distribution or mislabeled on sixteen percent and must delete six for missing consent, paying in model bias, labeling rework, and compliance deletion at once. A low per-sample price that yields an unfit corpus is not a saving; it is bias and liability, delayed and compounding through every model trained on it. The corpus-fitness audit — coverage, label validity, provenance — is the cheapest due-diligence instrument in this paper.

SECTION 03

The collection contract: cover the distribution, validate the label, document the provenance

Figure 2 lays out the three clauses of a working data-collection contract — the machinery a floor read verifies before any per-sample price is believed.

FIG. 2 — THE COLLECTION CONTRACT, AS OPERATED ON VETTED TEAMS

COVER THE DISTRIBUTION

Collection is designed against the target distribution — demographics, edge cases, rare classes, and the long tail the model will face — with quotas by stratum and a coverage audit that flags gaps before the round closes. The corpus is built to represent reality, not just the samples easiest to gather.

VALIDATE THE LABEL

Every label is applied to a written guideline, checked by a second annotator or gold-standard set, and inter-annotator agreement is tracked — with ambiguous cases escalated, not guessed. The label is made trustworthy at the source, so the model learns the right thing.

DOCUMENT THE PROVENANCE

Every sample carries its source, consent record, and licensing so the corpus is legally defensible and auditable — with PII handled to policy and a datasheet the AI team can read. Clean provenance is what keeps the whole dataset usable when legal or a regulator asks.

↓ yields ↓

THE FIT-FOR-TRAINING CORPUS

97%+ COVERAGE | 98%+ LABEL VALIDITY | CLEAN PROVENANCE

The corpus covers the distribution, carries valid labels, and comes with clean, consented provenance — so the model trains on data it can learn from, bias falls, and nothing has to be deleted or re-collected. The operation gets cheaper per fit sample and safer at the same time, which is the entire promise of outsourcing data collection to people who measure corpus fitness, not samples collected.

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

In diligence, the contract is tested from the corpus's side: pull a collected batch and audit it against the target distribution for coverage; re-check a label sample against the guideline; verify the consent records exist and are valid. 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 corpus fitness, not samples collected.

SECTION 04

Cost and performance: cost-per-fit-

sample economics and benchmarks

TABLE 1 — FULLY-LOADED COST PER FIT-FOR-TRAINING SAMPLE: THREE OPERATING MODELS (2026)

OPERATING MODEL	COST / FIT SAMPLE	VS. US BASELINE
US onshore in-house collection team	$1.72	baseline
Low-cost offshore data mill (per-sample)	$1.16	−33%
Vetted Manila team — covered, validated, sourced	$0.72	−58%

Basis: PITON-Global engagement pricing 2025–26; cost per fit-for-training sample = fully-loaded operating cost divided by samples that cover the distribution, carry valid labels, and have clean provenance (off-distribution, mislabeled, and deleted samples excluded); the data mill looks cheap per sample but re-collection, labeling rework, and deletion lift its true cost per fit sample; US comparator fully-loaded in-house cost.

TABLE 2 — DATA-COLLECTION BENCHMARKS, VETTED VS. BASELINE (2025–26)

METRIC	VETTED TOP TIER	BASELINE	EXECUTIVE READING
Target-distribution coverage	97%+	78–88%	The only collection number the model follows
Label validity	98%+	84–92%	Right lesson, or the wrong one
Re-collection rate vs. baseline	−64%	baseline	"Cover the distribution" — held or not
Provenance / consent completeness	99.5%+	88–95%	Defensible, or a deletion order
Inter-annotator agreement	0.90+	0.70–0.82	Consistent labels, or noise

Source: PITON-Global AI-data engagement data 2025–26; baseline = typical pre-engagement operations billed on samples collected. Agreement reported as Cohen's/Fleiss' κ.

The strategic read: the data mill's −33% per sample is a false economy — re-collection, labeling rework, and deletion lift its true cost per fit sample far above the per-sample gap, and a model biased by a bad corpus is a cost no rate recovers. The vetted Manila team's −58% is real because it is measured where the value is: the sample that covers the distribution, carries a valid label, and is cleanly sourced. Buy on samples collected and you buy the right column of Figure 1; buy on cost-per-fit-sample and you buy corpus fitness — 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 data collection

Data-collection diligence adds one hard requirement to every step: a per-sample price is real only when the corpus fitness behind it survives a coverage-and-provenance audit of a delivered batch. The funnel below is representative of a 40–90-seat data-collection search.

FIG. 3 — THE 7-STEP FRAMEWORK AS A DATA-COLLECTION FUNNEL

	STEP	WHAT IT ELIMINATES IN A DATA-COLLECTION SEARCH	FIELD REMAINING
1	**Requirements definition**	Volume-first scoping: fixes the coverage target, the label-validity floor, the provenance standard, and the agreement threshold the corpus must meet by data type and modality	1,000+
2	**Market mapping**	Data mills: operations that bill sample volume and treat off-distribution and consent gaps as the client's problem, and generalist floors with no coverage-design or annotation depth	~46
3	**Capability screening**	Teams without the machinery: coverage-design capability, a label-validation and inter-annotator-agreement workflow, consent/provenance tracking, and modality coverage across your data types	~13
4	**Compliance & security audit**	Weak governance: consent and licensing rigor, PII handling, data-residency and GDPR alignment, secure-collection controls, SOC 2 scope covering the data-collection stack	~8
5	**Operational diligence**	The batch audit: audit a delivered batch against the target distribution for coverage; re-check labels against the guideline; verify consent records; interview on how they fill a rare-class gap	4–6
6	**Competitive RFP & negotiation**	Per-sample pricing: finalists bid against coverage targets, label-validity floors, and provenance standards — all carrying credits for samples that come back unfit or unconsented	2–3
7	**Transition & launch governance**	Cold start: the team calibrates against your coverage spec, label guideline, and consent policy on a pilot batch before it scales, with the fitness floor certified before steady state, PITON-Global embedded buyer-side	1

Field-remaining counts represent a 40–90-seat data-collection search, PITON-Global engagements 2024–26.

The standing guarantees hold: **vendor neutrality** — PITON-Global operates no collection floors, gathers no data, and takes no placement economics, so the shortlist reflects audited corpus-fitness 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 corpus fitness: a 55-seat data-collection rebuild, reconstructed

Engagement DC-061 is a US AI company — collecting roughly 40,000,000 samples a year across text, speech, and image for model training — that had offshored data collection to a per-sample mill. The sample price was excellent; the corpus was not: distribution coverage sat near 82%, a label spot-check failed at 11%, consent records were incomplete on a slice legal forced them to delete, and the model was under-performing on tail cases. The head of ML data wanted the scale without the bias and the deletion. Identifying details are anonymized under NDA.

SELECTION (WEEKS 0–8, BATCH AUDITS)

Requirements fixed a 97% coverage target, a 98% label-validity floor, and a 99.5% provenance standard, with a mandatory inter-annotator-agreement and consent-tracking workflow. Mapping produced 46 claimants; the machinery screen — coverage design, label validation, provenance tracking, modality coverage — left 13; the consent-and-security audit left 8. Blind batch audits against the target distribution shortlisted 5. A Manila team won priced against coverage and provenance, with a batch audit written into the reporting spec.

TRANSITION (MONTHS 2–6, PILOT FIRST)

The team calibrated against the client's coverage spec, label guideline, and consent policy on a pilot batch before scaling; the fitness floor was certified in month 2. The coverage-and-agreement loop fed collection design continuously. By month 6 coverage reached 97.4%, label validity hit 98%, re-collection fell 64%, and provenance completeness reached 99.6%. PITON-Global chaired the corpus-fitness review board to steady state.

TABLE 3 — DC-061 FIRST-YEAR VALUE BUILD-UP (US$)

VALUE COMPONENT	YEAR 1	BASIS
Collection-cost arbitrage on the 55-seat operation	$1.98M	Onshore vs. vetted-offshore cost across the corpus
Re-collection eliminated (coverage held)	$0.72M	Repeat collection rounds avoided by fitness
Model-performance & bias value	$0.34M	Tail-case accuracy gain from better coverage
Compliance & deletion exposure avoided	$0.20M	Consent-gap deletion and rework removed
Gross value created	**$3.24M**	
Less: engagement & transition costs	($0.52M)	Pilot, coverage-spec build, tooling integration; advisory fee $0 (provider-paid model)
Net value ÷ cost = first-year ROI	**6.2×**	$3.24M ÷ $0.52M

Figures rounded; coverage, label, and re-collection effects audit-verified. Method in Methodology & notes.

Sixty-one percent arbitrage, thirty-nine percent from eliminated re-collection, better model performance, and avoided deletion — a typical split for the series, earned because the operation was re-based on corpus fitness rather than sample volume. Every line traces to a diligence step: the coverage target to Step 1, the fitness-based pricing to Step 6, the label-validation workflow to Step 3's contract. The head of ML data's line at the year-one review: "The sample price barely moved. What moved is that the model finally learned the tail — and legal stopped making us delete."

SECTION 07

The executive playbook: governance for the first 180 days

A data-collection transition succeeds when the team is measured and paid on corpus fitness and provenance — client in command throughout.

DAYS 0–30

Define fitness, contract on it

Set the coverage target, the label-validity floor, and the provenance and agreement thresholds with ML-data leadership before launch, and document the coverage spec, label guideline, and consent policy by data type. Contract against corpus fitness and provenance — never per sample alone — with targets and floors carrying credits. Stand up the batch audit as the program's arbiter of truth.

DAYS 30–90

Pilot, prove the coverage

Calibrate the team against your coverage spec, label guideline, and consent policy on a pilot batch before it scales; certify the fitness floor and label-validation loop. Switch on the coverage-and-agreement loop from day one. Publish the collection dashboard: coverage, label validity, re-collection rate, provenance completeness, inter-annotator agreement.

DAYS 90–180

Scale, certify steady state

Scale collection only after the team holds the coverage target and provenance standard. Review the fitness board monthly with the re-collection and consent-completeness trends on the table. Certify steady state on two rounds at floor including a rare-class fill; hand over on documented criteria; keep the batch audit standing.

THE FIVE FAILURE MODES WE ARE HIRED TO PREVENT

Paying per sample while the corpus misses the distribution · hitting quotas instead of covering the tail · labeling without validation or agreement tracking · collecting without consent and provenance records · scaling before fitness is proven. All five are settled at selection and contract.

The data-collection argument, in one sentence: a collected sample is volume and a fit-for-training one is data a model can actually learn from, cleanly sourced, and a vetted Philippine team delivers −58% cost per fit sample with 97%+ coverage, 98%+ label validity, and re-collection cut by nearly two-thirds — for buyers who audit the corpus fitness instead of counting samples collected, and compete the finalists through a fully transparent, fair, comprehensive, and highly competitive RFP. Anyone can collect a sample. We make sure you buy the team whose corpus a model can learn from.

METHODOLOGY & NOTES

How the numbers were built

Fitness and re-collection figures (Fig. 1, Table 2) derive from engagement instrumentation in PITON-Global AI-data engagements, 2025–26. Coverage and label validity are measured by auditing a delivered batch against the target distribution and re-checking labels against the guideline; re-collection rate, provenance completeness, and inter-

annotator agreement are measured against the client's pre-engagement baseline. The weak-provider column is a composite of pre-engagement vendor reporting reviewed in diligence, shown for contrast.

Cost-per-fit-sample figures (Table 1) are fully loaded — wages, benefits and statutory charges, facilities, collection and annotation tooling, management and QA overhead including coverage designers and provenance specialists, attrition-driven recruiting — divided by samples that cover the distribution, carry valid labels, and have clean provenance, with off-distribution, mislabeled, or deleted samples excluded from the denominator; US comparators reflect fully-loaded in-house cost on the same basis. A fit-for-training sample covers the target distribution, carries a valid label, and has consented, documented provenance.

The case study (DC-061) is anonymized under NDA; volumes and dates are generalized, figures rounded. Coverage, label, and re-collection effects are audit-verified; the model-performance line reflects tail-case accuracy gains, and the compliance 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; corpus fitness and provenance 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.