CONTENT MODERATION · PHILIPPINES

The decision-accuracy standard: the executive economics of content moderation outsourcing to the Philippines.

An analysis of why items actioned is a volume vanity metric, how decision accuracy and moderator wellness — never moderation throughput — decide the true cost of a trust-and-safety operation once wrong takedowns, missed violations, policy inconsistency, appeals, and moderator attrition are counted, and the vendor-selection discipline that gets the decision right and keeps the moderator whole.

AUTHORS

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

AUGUST 2026
MANILA · ESTABLISHED 2001

Contents

ABOUT THIS SERIES — VOLUME 52

Every content-moderation proposal is priced per item actioned, because volume is easy to count. This volume is about the two outcomes a platform's integrity actually depends on and a cheap contract quietly trades away: the decision made accurately and consistently against policy, and the moderator kept well enough to keep making it. The gap between actioning an item and deciding it accurately — sustainably — is where wrong takedowns, missed harms, and burnout live. The full series lives at piton-global.com.

About the authors

John Maczynski

Chief Executive Officer

John has bought trust-and-safety capacity for three decades, and learned that the fast operation that actions every item and burns out its moderators is the most expensive one a platform can hire — because a wrong takedown is a user harmed, a missed violation is a harm shipped, and a floor churning at high attrition never holds decision quality. His standing question ignores the queue-cleared report: of the items you actioned, how many were decided right, consistently, by moderators who were properly supported? Applied at PITON-Global, that question is what clients get on every moderation 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 trust-and-safety floors, and he judges a moderation team by its decision accuracy and its wellness and attrition record, not its items-per-hour: whether calls were right and consistent, and whether the people making them were supported enough to stay. The moderation contract in Section 3 codifies what he learned rebuilding teams that cleared queues fast and lost their moderators. He now reads moderation providers the way he once ran them — from the QA-accuracy score and the wellness-and-attrition 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 moderation team's real decision accuracy and wellness-and-attrition record rather than its items-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%	**98%+**	**<25%**	**6.1×**
Fully-loaded cost per accurate, sustainable decision vs. US onshore	Decision-accuracy (QA) rate on vetted teams, up from 82–90%	Annualized moderator attrition, vetted teams, vs. 60%+ typical	First-year ROI, reconstructed case (Section 6)

A content-moderation engagement is bought on the price per item and paid for on the decisions it gets wrong and the moderators it burns through. The vendor reports millions of items actioned at a fraction of the onshore cost; the platform finds the wrong takedowns that anger users and draw press, the violations missed and shipped to the feed, the inconsistent calls that make policy look arbitrary on appeal, and a floor churning at sixty-percent attrition that can never hold quality because it is always training new hires. The operation cleared its queues and the platform inherited a trust problem, an appeals backlog, and a workforce it was burning out. The gap between actioning an item and deciding it accurately — sustainably — is where the real cost of cheap moderation hides, because a wrong or unsustainable decision is not throughput delivered; it is user harm, regulatory exposure, and a human cost the contract externalized.

The Philippines — English-fluent, culturally fluent with Western content norms, and with a mature trust-and-safety talent pool — is where moderation is delivered with the most decision-accuracy and wellness discipline, because that pool, properly supported, can run the accurate policy application, calibration, and moderator care that a sustainable operation demands. Five findings:

1 **Items actioned is the industry's most flattering number.** Any floor can clear a queue at speed. Vetted teams are measured on outcome: decision accuracy, policy consistency, and a sustainable, low-attrition workforce. Weak providers quote a low per-item price the platform repays in wrong calls, appeals, and burnout. Section 2 shows the gap.

2 **The accurate, sustainable decision, not the fast action, is the product.** A decision that protects users and the platform comes from applying policy correctly and consistently, made by a moderator who is supported and calibrated — not from clearing a queue. Teams that decide accurately and sustain their

people hold quality; teams that action loosely and burn out their floor produce wrong calls and a workforce that never stabilizes.

3 **A wrong decision cuts both ways — over-removal and under-removal.** Take down legitimate content and you silence users and invite backlash; leave a real violation up and you ship harm and court regulatory action. Vetted teams hold decision accuracy above 98% precisely because both tails cost — the speech wrongly removed and the harm wrongly allowed.

4 **Attrition, not the per-item rate, is the hidden cost driver.** A floor that churns at high attrition pays endlessly to recruit and retrain, and never accumulates the tenure that accuracy needs — while the wellness program that lowers attrition is exactly what a cheap contract cuts first. Teams that sustain moderators hold accuracy and cost less per accurate decision, and that stability — not the offshore rate alone — is where the case study's return is built.

5 **Contract on the accurate, sustainable decision, not the actioned item.** The contracting failure of moderation is paying per item and hoping the calls are right and the floor is cared for. Vetted deals price against decision-accuracy floors, policy-consistency targets, and wellness-and-attrition standards, making accurate, sustainable decisions the team's problem, which is where it belongs.

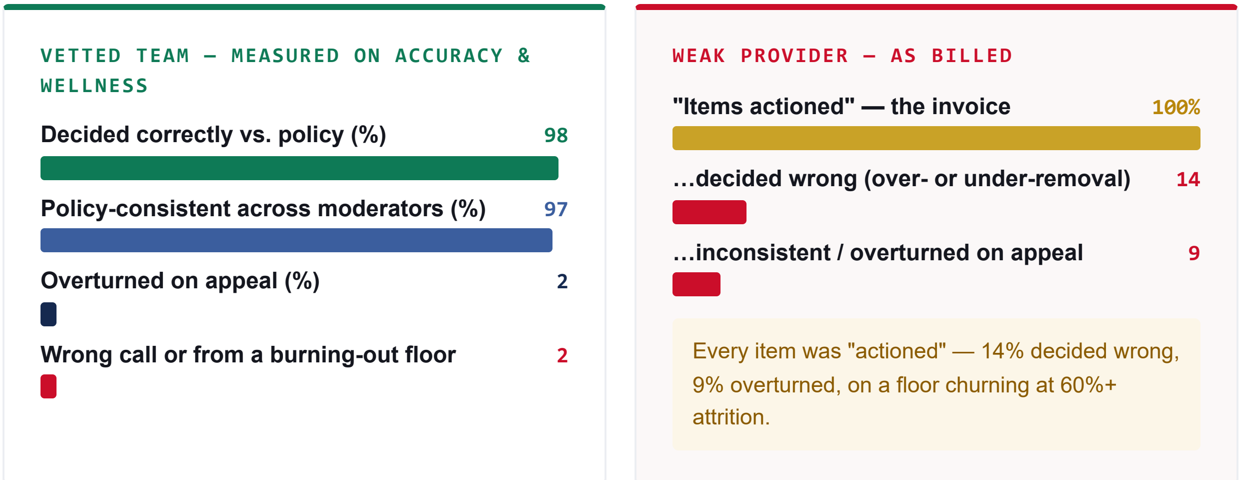
The intended reader is a head of trust and safety, VP of operations, or COO buying content moderation per item and paying for it in wrong decisions, appeals, and moderator burnout. Sections 2–4 separate the teams that decide accurately and sustainably from the teams that clear queues; Sections 5–7, the method and the proof.

SECTION 02

The volume mirage: items actioned versus accurate, sustainable decisions

Figure 1 shows one week of a platform's moderation queue two ways: as the items-actioned report a weak vendor bills, and as the decision-accuracy reality the QA sample, the appeals queue, and the wellness log experience. The items-actioned number is identical in spirit to a chatbot's containment rate — a figure engineered to look like output while wrong calls, appeals, and moderator burnout accumulate underneath.

FIG. 1 — ONE WEEK OF THE QUEUE: DECISION-ACCURACY QUALITY VS. REPORTED VOLUME



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

The economics follow the decision accuracy and the workforce, not the items count. On the left, ninety-eight percent of decisions are correct and the floor is sustained at low attrition, so trust holds and appeals stay rare; on the right, a cheap per-item price decides fourteen percent wrong and overturns nine on appeal, on a floor churning so fast it can never build the tenure accuracy needs — paying in user harm, appeals, and perpetual retraining at once. A low per-item price that produces wrong calls and burnout is not a saving; it is trust erosion and a human cost, delayed and compounding. The decision-accuracy-and-wellness audit — QA accuracy, policy consistency, attrition, wellness program — is the cheapest due-diligence instrument in this paper.

SECTION 03

The moderation contract: decide it accurately, apply policy consistently, sustain the moderator

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

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

DECIDE IT ACCURATELY

Every item is decided against the current policy on the evidence in front of the moderator, with the context and language understood, and an ambiguous or edge case escalated rather than guessed to clear the queue. The decision is made right at the source, so users are neither wrongly silenced nor wrongly exposed.

APPLY POLICY CONSISTENTLY

The same content gets the same call across moderators and shifts, held there by calibration sessions, a QA sample, and fast policy-update propagation — so the platform's rules are applied uniformly and hold up on appeal. Consistency is what makes moderation look like policy rather than caprice.

SUSTAIN THE MODERATOR

Moderators get wellness and resilience support, on-staff psychological care, exposure limits and rotation, and a career path — so the people making the calls stay well and stay, and attrition stays low. A sustained moderator is both an ethical obligation and the only way decision accuracy holds over time.

↓ yields ↓

THE ACCURATE, SUSTAINABLE DECISION

98%+ DECISION ACCURACY | POLICY-CONSISTENT | SUSTAINABLE FLOOR

Items are decided accurately, policy is applied consistently, and moderators are sustained — so wrong takedowns and missed harms fall, appeals drop, and the workforce stays well and stable enough to keep quality high. The operation gets cheaper per accurate decision and more humane at the same time, which is the entire promise of outsourcing moderation to people who measure decision accuracy and wellness, not items actioned.

Operating parameters from PITON-Global trust-and-safety engagements, 2025–26. Wellness provisions reflect on-staff psychological support, exposure limits, and rotation.

In diligence, the contract is tested from the QA sample's and the floor's side: pull actioned items and re-decide a sample against policy to measure true accuracy and consistency; read the appeal-overturn log; and audit the wellness program, exposure limits, and attrition record — because a floor churning at high attrition cannot hold quality, whatever the QA snapshot says. 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 accurate, sustainable decisions, not items actioned.

SECTION 04

Cost and performance: cost-per-accurate-

decision economics and benchmarks

TABLE 1 — FULLY-LOADED COST PER ACCURATE, SUSTAINABLE DECISION: THREE OPERATING MODELS (2026)		
OPERATING MODEL	COST / DECISION	VS. US BASELINE
US onshore in-house moderation team	**$0.71**	baseline
Low-cost offshore moderation mill (per-item)	**$0.49**	**−31%**
Vetted Manila team — accurate, sustainable	**$0.31**	**−57%**

Basis: PITON-Global engagement pricing 2025–26; cost per accurate, sustainable decision = fully-loaded operating cost (including wellness program and stable-tenure staffing) divided by items decided correctly and upheld (wrong and overturned decisions excluded); the moderation mill looks cheap per item but wrong calls, appeals, and 60%+ attrition retraining lift its true cost per accurate decision; US comparator fully-loaded in-house cost.

TABLE 2 — CONTENT-MODERATION BENCHMARKS, VETTED VS. BASELINE (2025–26)			
METRIC	VETTED TOP TIER	BASELINE	EXECUTIVE READING
Decision-accuracy (QA) rate	**98%+**	82–90%	The only moderation number trust follows
Policy consistency (inter-rater)	**97%+**	80–88%	Policy, or caprice on appeal
Appeal-overturn rate vs. baseline	**−63%**	baseline	"Decide it accurately" — held or not
Annualized moderator attrition	**<25%**	60%+	A sustained floor, or a churning one
Wellness program & on-staff support	**Full**	Minimal	People cared for, or externalized

Source: PITON-Global trust-and-safety engagement data 2025–26; baseline = typical pre-engagement operations billed on items actioned.

The strategic read: the moderation mill's −31% per item is a false economy — wrong calls, appeals, and the retraining treadmill of 60%+ attrition lift its true cost per accurate decision far above the per-item gap, and trust lost to a viral wrong takedown is a cost no rate recovers. The vetted Manila team's −57% is real because it is measured where the value is: the decision made accurately and consistently, by a moderator who is supported and stays. Buy on items actioned and you buy the right column of Figure 1; buy on cost-per-accurate-decision and you buy decision accuracy and a sustainable floor — 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 content moderation

Moderation diligence adds two hard requirements to every step: a per-item price is real only when the decision accuracy behind it survives a QA re-decision against policy, and only when the wellness program and attrition record behind it prove the floor is sustainable. The funnel below is representative of an 80–200-seat trust-and-safety search.

FIG. 3 — THE 7-STEP FRAMEWORK AS A CONTENT-MODERATION FUNNEL

	STEP	WHAT IT ELIMINATES IN A MODERATION SEARCH	FIELD REMAINING
1	**Requirements definition**	Throughput-first scoping: fixes the decision-accuracy floor, the policy-consistency target, the appeal-overturn ceiling, and the wellness-and-attrition standard the operation must meet by content type	1,000+
2	**Market mapping**	Moderation mills: operations that bill actioned volume, run high attrition, and treat wrong calls and burnout as the client's problem, and generalist floors with no trust-and-safety or wellness depth	~48
3	**Capability screening**	Teams without the machinery: audited QA-accuracy history, calibration workflow, a real wellness and resilience program with on-staff psychological support, and coverage across your policy areas and languages	~14
4	**Compliance & security audit**	Weak governance: user-data protection, secure-facility and content-handling controls, regulatory alignment (DSA and peers), duty-of-care and labor standards, SOC 2 scope covering the moderation stack	~8
5	**Operational diligence**	The QA-and-wellness read: re-decide a sample against policy for true accuracy and consistency; read the appeal-overturn log; audit the wellness program, exposure limits, and attrition; interview on how they handle an edge-case policy call	4–6
6	**Competitive RFP & negotiation**	Per-item pricing: finalists bid against decision-accuracy floors, policy-consistency targets, appeal-overturn ceilings, and wellness-and-attrition standards — all carrying credits for wrong or overturned calls	2–3
7	**Transition & launch governance**	Cold start: the team calibrates against your policy and content types on a sample before it owns the live queue, with the accuracy floor and wellness program certified before steady state, PITON-Global embedded buyer-side	1

Field-remaining counts represent an 80–200-seat trust-and-safety search, PITON-Global engagements 2024–26.

The standing guarantees hold: **vendor neutrality** — PITON-Global operates no moderation floors, actions no content, and takes no placement economics, so the shortlist reflects audited decision-accuracy and wellness 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 decision accuracy: a 120-seat trust-and-safety rebuild, reconstructed

Engagement CM-053 is a US digital platform — moderating roughly 46,000,000 items a year across text, image, and video — that had offshored content moderation to a per-item mill. The item price was excellent; the decisions and the floor were not: decision accuracy sat near 87%, appeal overturns were high and drawing press, policy was applied inconsistently across shifts, and the floor was churning at over sixty percent attrition with a thin wellness program. The head of trust and safety wanted the scale without the wrong calls and the human cost. Identifying details are anonymized under NDA.

SELECTION (WEEKS 0–8, QA + WELLNESS READS)

Requirements fixed a 98% decision-accuracy floor, a 97% policy-consistency target, an appeal-overturn ceiling, and a wellness-and-attrition standard (on-staff psychological support, exposure limits, <25% attrition). Mapping produced 48 claimants; the machinery screen — QA history, calibration, real wellness program, language coverage — left 14; the compliance-and-duty-of-care audit left 8. Blind QA re-decisions plus wellness-and-attrition audits shortlisted 5. A Manila team won priced against decision accuracy and wellness, with both written into the reporting spec.

TRANSITION (MONTHS 2–7, CALIBRATED FIRST)

The team calibrated against the platform's policy and content types on sample queues before owning the live queue; the accuracy floor was certified in month 3, and the wellness program and exposure limits stood up from day one. The calibration-and-QA loop fed moderator feedback continuously. By month 7 decision accuracy reached 98.2%, policy consistency hit 97%, appeal overturns fell 63%, and attrition dropped below 25%. PITON-Global chaired the decision-accuracy-and-wellness review board to steady state.

TABLE 3 — CM-053 FIRST-YEAR VALUE BUILD-UP (US$)		
VALUE COMPONENT	YEAR 1	BASIS
Labor arbitrage on the 120-seat operation	$3.06M	225.0k productive hrs × ($23.10 − $9.50)
Attrition & retraining cost removed	$0.82M	Recruiting/retraining avoided at <25% vs. 60%+
Trust, appeals & regulatory exposure avoided	$0.46M	Appeal-handling and wrong-call exposure removed
Rework & re-moderation reduction	$0.20M	Re-moderation of wrong calls removed
Gross value created	**$4.54M**	
Less: engagement & transition costs	($0.74M)	Calibration, wellness-program build, tooling integration; advisory fee $0 (provider-paid model)
Net value ÷ cost = first-year ROI	6.1×	$4.54M ÷ $0.74M

Figures rounded; accuracy, appeal, and attrition effects audit-verified. Method in Methodology & notes.

Sixty-seven percent arbitrage, thirty-three percent from removed attrition/retraining, avoided appeals and exposure, and less re-moderation — a typical split for the series, earned because the operation was re-based on decision accuracy and a sustainable floor rather than item volume. Every line traces to a diligence step: the accuracy floor and wellness standard to Step 1, the accuracy-and-wellness-based pricing to Step 6, the calibration workflow to Step 3's contract. The head of trust and safety's line at the year-one review: "The item price barely moved. What moved is that the decisions are right, the appeals dried up — and my moderators are staying."

SECTION 07

The executive playbook: governance for the first 180 days

A moderation transition succeeds when the team is measured and paid on decision accuracy and moderator wellness — client in command throughout.

DAYS 0–30

Define accuracy & wellness, contract on them

Set the decision-accuracy floor, the policy-consistency target, and the appeal-overturn and wellness-and-attrition thresholds with trust-and-safety leadership before launch, and document the calibration and escalation workflow by policy area. Contract against decision accuracy and wellness — never per item alone — with floors and standards carrying credits. Stand up the QA re-decision and the wellness/attrition audit as the program's arbiters of truth.

DAYS 30–90

Calibrate, stand up wellness

Calibrate the team against your policy and content types on sample queues before it owns the live queue; certify the accuracy floor, the calibration loop, and the wellness program with exposure limits. Switch on the QA-and-feedback loop from day one. Publish the moderation dashboard: decision accuracy, policy consistency, appeal-overturn rate, attrition, wellness-program utilization.

DAYS 90–180

Own the queue, certify steady state

Hand the live queue to the team only after it holds the accuracy floor and the wellness-and-attrition standard. Review the board monthly with the appeal-overturn and attrition trends on the table. Certify steady state on two months at floor including a policy-update cycle; hand over on documented criteria; keep the quarterly QA re-decision and wellness audit standing.

THE FIVE FAILURE MODES WE ARE HIRED TO PREVENT

Paying per item while decisions are wrong · clearing queues instead of deciding accurately · applying policy inconsistently across shifts · running a floor at 60%+ attrition with a thin wellness program · owning the queue before accuracy and wellness are proven. All five are settled at selection and contract.

The content-moderation argument, in one sentence: an actioned item is activity and an accurate, sustainable decision is a user protected, a policy upheld, and a moderator kept whole, and a vetted Philippine team delivers −57% cost per accurate decision with 98%+ decision accuracy, appeal overturns cut by nearly two-thirds, and attrition held under 25% on a real wellness program — for buyers who audit the accuracy and the wellness instead of counting items actioned, and compete the finalists through a fully transparent, fair, comprehensive, and highly competitive RFP. Anyone can action an item. We make sure you buy the team that decides it right and keeps its people whole.

METHODOLOGY & NOTES

How the numbers were built

Accuracy and wellness figures (Fig. 1, Table 2) derive from engagement instrumentation in PITON-Global trust-and-safety engagements, 2025–26. Decision accuracy and policy consistency are measured by QA re-deciding a sample against policy; appeal-overturn rate, attrition, and wellness-program provision 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-accurate-decision figures (Table 1) are fully loaded — wages, benefits and statutory charges, facilities, technology and tooling, management and QA overhead including calibration leads and on-staff wellness/psychological support, and attrition-driven recruiting — divided by items decided correctly and upheld, with wrong or overturned decisions excluded from the denominator; US comparators reflect fully-loaded in-house cost on the same basis. An accurate, sustainable decision is decided correctly against policy, upheld on appeal, and made by a supported, stable workforce.

The case study (CM-053) is anonymized under NDA; volumes and dates are generalized, figures rounded. Accuracy, appeal, and attrition effects are audit-verified; the attrition-removed line reflects avoided recruiting and retraining, and the exposure-avoided 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; decision accuracy and wellness/attrition performance re-verified May 2026). Moderator wellbeing is treated here as both an ethical obligation and an operating fundamental; wellness provisions reflect on-staff psychological support, exposure limits, and rotation. 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 item price survive a QA-and-wellness audit?

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

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