PITON Global

EXECUTIVE WHITE PAPER · WP-69

AI SUPPORT OUTSOURCING · PHILIPPINES

The containment standard: the executive economics of AI support outsourcing to the Philippines.

An analysis of why deflection dashboards flatter while reopen curves tell the truth, the small human layer that decides whether support AI is an asset or a liability, and vendor-selection discipline for the operation that governs what the machine says to your customers.

AUTHORS

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

AUGUST 2026
MANILA · ESTABLISHED 2001

Contents

ABOUT THIS SERIES — VOLUME 69

Every support AI vendor sells the same slide: contacts in, headcount out. What the slide never shows is where the deflected contacts went. This volume follows them — through reopen curves, silent churn, and late escalations — and prices the small Philippine human layer that turns automation theater into resolution you can audit. The full series lives at piton-global.com.

THE PITON-GLOBAL LEADERSHIP TEAM

About the authors

John Maczynski

Chief Executive Officer

Three times in John's thirty-plus years running Fortune-500 CX budgets, a technology promised to end support headcount — IVR, scripted chatbots, and now LLMs. Each time the pitch was deflection; each time the invoice arrived as reopens. The discipline in this paper is the one he learned to

Ralf Ellspermann

Chief Strategy Officer

Ralf has spent twenty-five years building Manila operations, including some of the first floors where humans supervised machine answers rather than typing their own. What that era taught him: automation tells the truth in its reopen curve and nowhere else. When he walks an AI support floor today, he

demand in procurement rooms: show me the reopen curve, not the deflection dashboard. At PITON-Global he applies that demand for buyers directly, from first scoping call to certified launch.

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

ignores the demo and pulls three artifacts — the error taxonomy, the QA sampling plan, and the route from a caught mistake back into the knowledge base. Floors that have all three are rare. They are the shortlist.

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

PITON Global

ABOUT PITON-GLOBAL

PITON-Global has advised on Philippine outsourcing from Manila since 2001, and only ever from the buyer's side of the table: no delivery operations of its own, no placement commissions, no consulting revenue from providers. The firm maintains a continuously re-verified map of more than 1,000 Philippine delivery operations, distills it to the six to ten providers genuinely qualified for a given program, and runs a fully transparent, competitive RFP among them — free to the buyer. Tripadvisor, TSYS, Garmin, HealthPartners, Yetipay, Workrise, and Norman Disney & Young have all bought through the practice.

SECTION 01

Executive summary

−60%	55–65%	<0.5%	5.7×
Fully-loaded cost of the human layer vs. US onshore	True containment on governed programs — reopen-checked	Hallucination escape rate with a full output-QA layer	First-year ROI, reconstructed case (Section 6)

Support AI has a measurement problem before it has a technology problem. The industry's favorite metric — deflection — counts conversations the bot ended, not problems the bot solved, and the difference between those two numbers is invisible until you follow the customer for two weeks. Followed honestly, a large share of "deflected" contacts reopen in another channel, escalate late and angrier, or go silent in the way that precedes churn. The fix is not a better model. It is an operating layer — small, disciplined, and overwhelmingly built in the Philippines — that audits what the machine says, catches what it gets wrong, and feeds every error back into the system so the same mistake is not made twice.

This paper prices that layer, benchmarks what it achieves, and shows how to buy it. Five findings:

1. **Containment is the only honest metric.** A contact counts as contained when it does not come back — any channel, fourteen days. Governed programs sustain 55–65% true containment; ungoverned bots claiming 70% deflection routinely measure out at 30–45% once reopens are netted. Section 2 follows the missing contacts.

2. **The human layer is small but decisive.** Roughly one governed human for every eight to twelve bot-handled conversations' worth of volume — sampling outputs, taking escalations with full context, and running the closed loop. It is the difference between an asset and a liability, and it costs −60% to run from Manila.

3 **The unit economics compound.** The bot absorbs the volume; the Philippine layer governs it at $12.10 a blended hour against $30.10 onshore. Net of the bot tier's own cost, per-resolved-contact economics land around −74% versus a pre-AI onshore baseline — but only when containment is real. Deflection theater delivers negative savings once reopens are re-worked.

4 **What reaches humans is harder now.** As the bot absorbs the FAQ tier, every conversation that escalates is by definition one the machine could not handle — higher stakes, higher emotion, often carrying a machine mistake to undo. The hiring bar for the human layer is therefore rising exactly as its headcount falls, which makes floor selection worth more per seat, not less.

5 **Governance is permanent, not transitional.** Models change under you — vendor updates, knowledge drift, new products. The QA layer is not scaffolding to remove once the bot "settles"; it is the standing function that notices when last month's accurate answer became this month's hallucination. Buyers who budget it as temporary rebuy it within a year.

The intended reader owns support, CX, or digital operations where an AI tier is live or imminent. Sections 2–4 price the problem; sections 5–7, the method and the proof.

SECTION 02

The deflection mirage: what bots actually resolve, and where the rest goes

Figure 1 is the audit most AI support programs have never run: take a month of "deflected" conversations, follow each customer across every channel for fourteen days, and classify what actually happened. The data comes from instrumented programs audited during PITON-Global engagements.

FIG. 1 — WHERE "DEFLECTED" CONTACTS ACTUALLY WENT, UNGOVERNED BOT TIER (14-DAY FOLLOW, 2025–26)

Cross-channel identity-joined follow of bot-closed conversations, pre-engagement audits across PITON-Global AI support programs, 2025–26. A 72% "deflection rate" measured out at 41% true containment.

The arithmetic behind the mirage: every reopened contact is handled twice and costs roughly 1.7× a first-touch resolution, because the second agent inherits a frustrated customer and a wrong answer to unwind. At the audited mix above, a bot tier "deflecting" 72% of volume was saving less than a third of what its dashboard implied — and the silent-19% was quietly the most expensive line on the chart, because those customers do not complain before they leave.

Governed programs run the same audit and land differently: 55–65% of bot-handled contacts genuinely contained, reopens under 5%, and — critically — a defined set of intents the bot is forbidden to touch, because the error cost dwarfs the handling saving. The

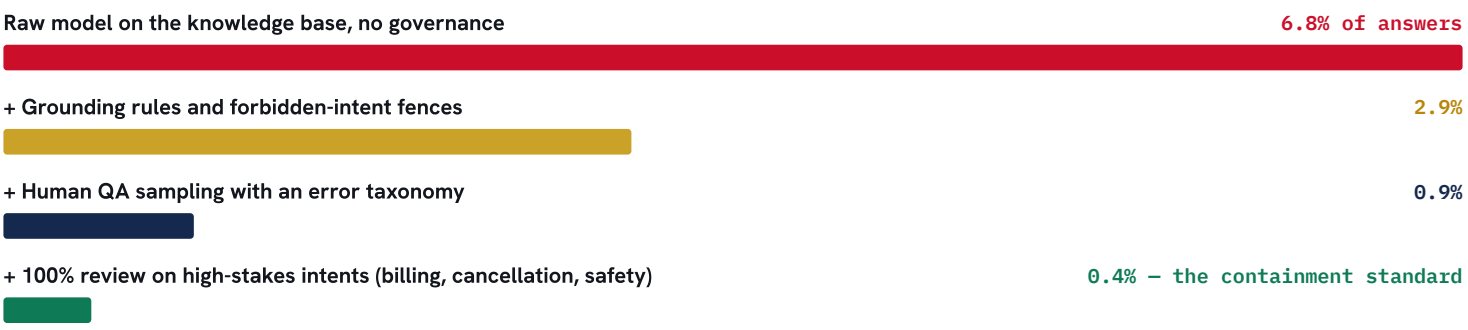
instrument that separates the two outcomes is not the model, which is often literally the same vendor product. It is the human layer of Section 3, and the discipline of counting resolution the way a CFO would: a contact is closed when the customer stays gone.

SECTION 03

The human layer: output QA, escalation with context, and the closed loop

The layer that turns Figure 1's mirage into Figure 2's floor has three functions, staffed in the Philippines for the same reason support itself moved there: deep written English, process discipline, and a labor market that can hire judgment at scale. Figure 2 shows the first function — output QA — doing its work on the metric that carries legal and brand risk.

FIG. 2 — HALLUCINATION ESCAPE RATE BY GOVERNANCE LAYER (MATERIAL ERRORS REACHING CUSTOMERS, 2025–26)



Material error = factually wrong commitment, price, policy, or instruction reaching a customer; measured by stratified transcript sampling, PITON-Global engagement QA data, 2025–26.

The second function is **escalation with context**. On weak programs the bot fails, apologizes, and dumps the customer into a queue where a human asks them to start over — the single most CSAT-destructive moment in modern support. On governed floors the escalation arrives as a case file: what the customer asked, what the bot answered, what it was uncertain about, and why the handoff fired. Escalation specialists trained to pick up mid-conversation hold 88–92 CSAT on contacts that arrive pre-frustrated, and their handle time runs 30% shorter because the diagnosis travelled with the transfer.

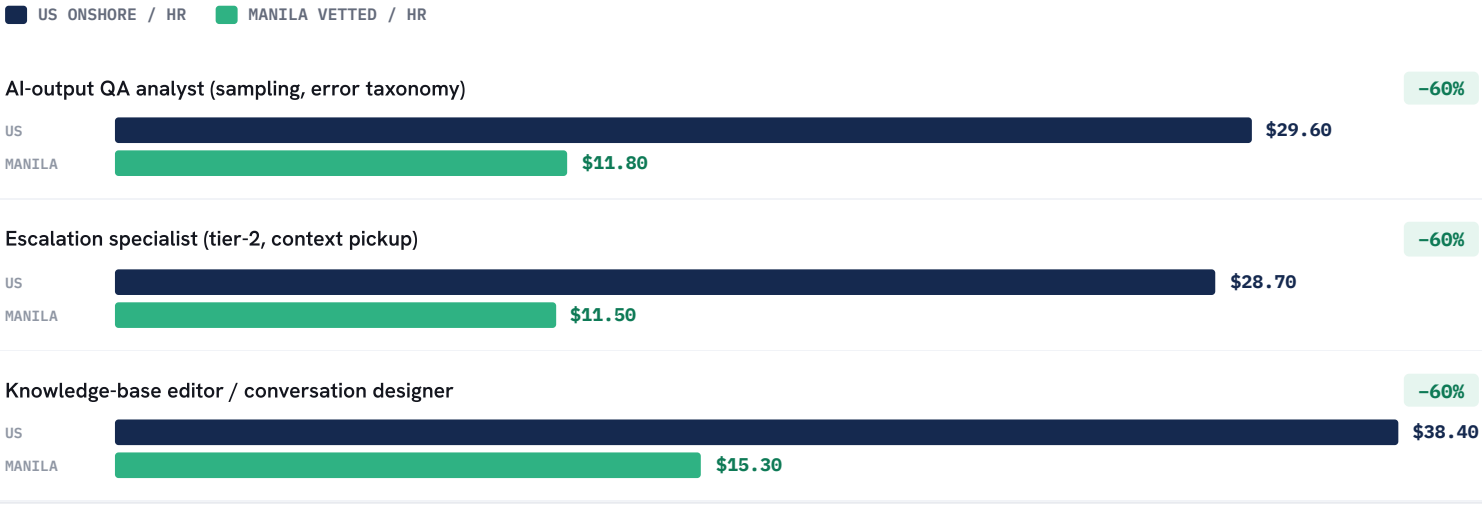
The third function is the **closed loop** — the reason the first two compound instead of merely coping. Every caught error lands in a taxonomy; every taxonomy cell routes to a fix — a knowledge-base correction, a prompt adjustment, a new fence, a bot-scope change — with a 48-hour service level on the highest classes. Floors running the loop cut their error inflow roughly in half every quarter for the first year. Floors without it employ QA as a fire watch: the same fires, every week, indefinitely. In diligence, one question separates them: *show me the last ten KB changes your QA team caused.* The good floors pull the log; the rest describe a meeting.

SECTION 04

Cost and performance: role-band economics and top-tier

benchmarks

TABLE 1 — FULLY-LOADED COST PER PRODUCTIVE HOUR BY ROLE BAND (US$, 2026)



Basis: PITON-Global AI support engagement pricing 2025–26 vs. US metro contact-center labor, fully loaded; per-resolved-contact figures include bot-tier platform costs and are netted for reopens.

TABLE 2 — AI SUPPORT OPERATING BENCHMARKS, GOVERNED VS. BASELINE (2025–26)

METRIC	GOVERNED TOP TIER	BASELINE	EXECUTIVE READING
True containment (14-day, cross-channel)	**55–65%**	30–45%	The number the deflection dashboard hides
Reopen rate on bot-closed contacts	**<5%**	18–31%	Every reopen is handled twice at 1.7× cost
Hallucination escape rate	**<0.5%**	3–7%	Legal and brand exposure, measured weekly
CSAT on escalated contacts	**88–92**	71–80	Context pickup vs. "please start over"
Error-to-fix cycle (highest severity classes)	**<48h**	Quarterly, if ever	The closed loop, or the same fires forever

Source: PITON-Global AI support engagement data 2025–26; baseline = typical pre-engagement operations, bot tier live without a governance layer.

Read together: the wage delta gets the human layer to −60%; real containment takes the per-resolution economics to −74%; and the benchmark gaps decide whether the AI tier is compounding value or manufacturing rework. A twenty-point containment gap at enterprise volume is worth more than the entire human-layer contract — which is why Section 5 treats the governance artifacts, not the model demo, as the decision documents.

SECTION 05

Sourcing discipline: the 7-Step Vendor Vetting Framework, applied to AI support

Every provider in the market now claims an AI story. Diligence is the art of finding the dozen with an AI operation — taxonomies, sampling plans, fix logs — behind the story. The funnel below is representative of a 20–50-seat human-layer search.

FIG. 3 — THE 7-STEP FRAMEWORK AS AN AI-SUPPORT FUNNEL

	STEP	WHAT IT ELIMINATES IN AN AI-SUPPORT SEARCH	FIELD REMAINING
1	**Requirements definition**	Vendor-led scoping: fixes the intent taxonomy first — what the bot owns, what it must never touch, where 100% review applies — and defines containment with a reopen window before anyone quotes a price	1,000+
2	**Market mapping**	The AI-washing tier: floors whose "AI practice" is a reseller license and a slide, and giants for whom a 30-seat governance layer ranks nowhere	~40
3	**Capability screening**	Floors without the machinery: living error taxonomies, stratified QA sampling, escalation-with-context tooling, KB-fix pipelines with service levels, agents hired for judgment on machine output	~12
4	**Compliance & security audit**	Weak AI governance: model-access controls, customer-data boundaries in prompts and logs, retention rules on transcripts, SOC 2 posture, ISO 42001 readiness where the program warrants it	~9
5	**Operational diligence**	The floor read: pull the reopen curve, read twenty escalation transcripts cold, open the error taxonomy and the last month's fix log, watch QA sampling happen live, ask what the bot is forbidden to answer and why	6–8
6	**Competitive RFP & negotiation**	Deflection-priced proposals: finalists bid against true containment with a 14-day reopen window, hallucination-escape ceilings, escalation-CSAT floors and 48-hour fix SLAs — all carrying credits	2–3
7	**Transition & launch governance**	Scope creep and silent model drift at cutover: bot boundaries configured as fences not guidelines, golden-set certification for escalation agents, weekly containment audits, PITON-Global embedded buyer-side to steady state	1

Field-remaining counts represent a 20–50-seat AI-support human-layer search, PITON-Global engagements 2024–26.

The standing guarantees hold: **vendor neutrality** — PITON-Global runs no AI support floors, resells no platforms, and takes no placement economics, so the shortlist reflects fix logs and reopen curves, not inventory; **right-sizing** — providers for whom your governance layer is a flagship program. 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

From deflection theater to audited containment: a 30-seat

rebuild, reconstructed

Engagement AS-026 is a US consumer-subscription software company — roughly $180 million in revenue, 1.4 million support contacts a year — that had deployed a leading bot platform and celebrated a 52% deflection rate for two quarters, until finance noticed support costs had fallen 9%, not 40%. A PITON-Global audit found the reopen rate on bot-closed contacts was 29%, hallucinated billing answers were escaping at 4% and driving refund leakage, and the onshore escalation team was re-diagnosing every handoff from scratch. Identifying details are anonymized under NDA.

SELECTION (WEEKS 0–7, INCL. GOVERNANCE FLOOR READS)

Requirements rebuilt the program around a 23-intent taxonomy: twelve intents the bot owned outright, seven it could draft but a human approved, four — cancellation, billing disputes, safety, legal threats — it was forbidden to close. Mapping produced 38 claimants of AI-operations capability; the machinery screen left 11; the model-governance and SOC 2 audit left 8. Floor reads — fix logs opened, reopen curves pulled, twenty escalation transcripts read cold — shortlisted 6. A Manila floor already running output QA for two SaaS brands won at $12.10 blended, bidding a 60% true-containment target with credits.

TRANSITION (MONTHS 3–6, FENCES BEFORE VOLUME)

The four forbidden intents were fenced in the platform before the new floor took a single conversation. Escalation specialists certified against a 50-case golden set of context pickups; the error taxonomy went live in week one, the 48-hour fix SLA in week three. Deflection was retired as a KPI and replaced by weekly containment audits with the 14-day reopen window. The bot's scope then expanded intent by intent — each promotion earned by four clean audit weeks. PITON-Global chaired buyer-side governance to steady state, then quarterly.

TABLE 3 — AS-026 FIRST-YEAR VALUE BUILD-UP (US$)

VALUE COMPONENT	YEAR 1	BASIS
Labor arbitrage on the 30-seat human layer	$1.01M	56.1k productive hrs × ($30.10 – $12.10)
Reopen collapse (29% → 4.6%) — contacts no longer handled twice	$0.33M	Avoided second-touch volume at 1.7× blended rate
Refund leakage stopped (hallucinated billing commitments, 4% → 0.3%)	$0.19M	Erroneous credits and make-goods vs. prior-year run rate
Churn saves on the silent cohort, cancellation intents human-handled	$0.12M	Retained subscriptions vs. cohort baseline, margin basis
Gross value created	**$1.65M**	
Less: engagement & transition costs	($0.29M)	Golden-set certification, dual-running, platform fencing, taxonomy build, governance overlay; advisory fee $0 (provider-paid model)
Net value ÷ cost = first-year ROI	**5.7×**	$1.65M ÷ $0.29M

Figures rounded; reopen, leakage and churn effects cohort-verified and taken at margin. Method in Methodology & notes.

Sixty-one percent arbitrage, thirty-nine percent operational value — and true containment ended the year at 61%, nine points *above* the old deflection number, measured honestly. The support VP's line at the year-one review: "We stopped asking how many tickets the bot closed and started asking how many customers never came back. Different question — and it turned out, a different vendor."

SECTION 07

The executive playbook: governance for the first 180 days

An AI support transition succeeds when the boundaries are enforced by the platform and the metrics are enforced by the contract — client in command throughout.

DAYS 0–30

Contract containment, fence the bot

Retire deflection from the contract. Define containment with a 14-day cross-channel reopen window, set hallucination-escape ceilings and escalation-CSAT floors, and attach credits to all three. Fix the intent taxonomy — owned, human-approved, forbidden — and configure the forbidden tier as platform fences before the first conversation routes. Policy-only fences erode with the first model update.

DAYS 30–90

Stand up the loop, certify the pickup

Launch the error taxonomy and stratified QA sampling in week one; the 48-hour fix SLA by week three. Certify escalation agents on a golden set of context pickups — the customer must never repeat themselves — before volume scales. Publish the weekly dashboard: true containment, reopen rate, escape rate, escalation CSAT, fixes shipped. If the fix log is empty in month two, the loop is not running.

DAYS 90–180

Expand scope on evidence, certify steady state

Promote intents into bot scope one at a time, each earning promotion with four clean audit weeks — and demote without ceremony when a model update degrades one. Re-run the full containment audit after every major model or KB release. Verify the program through one real surge before certifying steady state on two clean months; keep the quarterly transcript read and the model-change review standing.

THE FIVE FAILURE MODES WE ARE HIRED TO PREVENT

Buying deflection instead of containment · letting the vendor set the bot's scope · QA that samples only what the bot flags as uncertain · escalations that arrive without context · a governance layer budgeted as temporary scaffolding. All five are settled at selection and contract.

The AI support argument, in one sentence: the machine can absorb most of your volume, but only a governed human layer decides whether that absorption is resolution or theater — and a vetted Philippine floor runs that layer at a 60% cost advantage with the fix logs to prove it, for buyers who audit reopen curves instead of deflection dashboards and compete the finalists through a fully transparent, fair, comprehensive, and highly competitive RFP. The bot answers. Someone must own what it says. We make sure you know exactly who, and exactly how well.

METHODOLOGY & NOTES

How the numbers were built

Containment and reopen figures (Fig. 1, Table 2) derive from identity-joined, cross-channel contact histories on programs audited during PITON-Global engagements, 2025–26: a bot-closed contact counts as contained only if the same customer raises no related contact in any channel for fourteen days. Deflection comparators are the platforms' own dashboard figures over the same periods.

Hallucination escape rates (Fig. 2) are measured by stratified transcript sampling scored against a material-error rubric — a factually wrong commitment, price, policy, or instruction reaching a customer. Cost figures are fully loaded — wages, benefits and statutory charges, facilities, technology, management and QA overhead, attrition-driven recruiting — net of shrinkage; US comparators reflect metro contact-center labor on the same basis. Per-resolved-contact figures include bot-tier platform costs and are netted for reopens.

The case study (AS-026) is anonymized under NDA; volumes and dates are generalized, figures rounded. Reopen, refund-leakage and churn effects are verified against prior-year cohorts and taken at margin, not revenue. 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; AI-operations capabilities re-verified July 2026). Estimates and projections are PITON-Global analysis and should be cited as such. The full series is available at piton-global.com.

What is your bot's real containment rate?

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

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