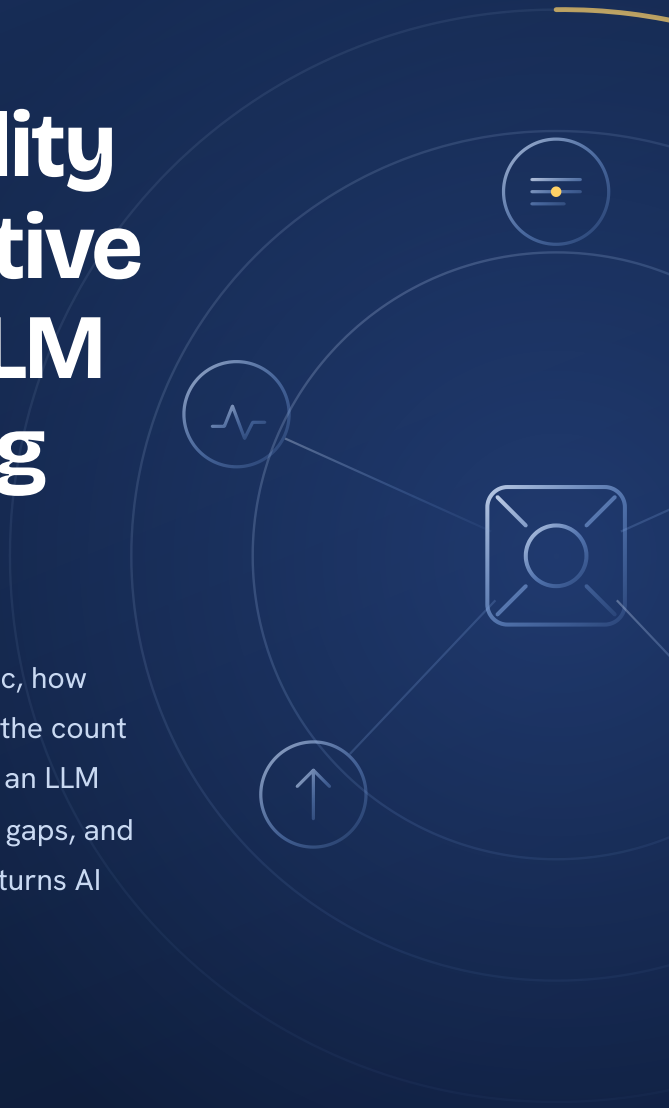
PITON Global

EXECUTIVE WHITE PAPER · WP-47

AI & LLM SERVICES · PHILIPPINES

The shipped-capability standard: the executive economics of AI & LLM services outsourcing to the Philippines.



An analysis of why services engaged is a breadth vanity metric, how shipped model capability and operational reliability — never the count of AI workstreams a vendor claims — decide the true cost of an LLM program once wasted compute, unshipped experiments, eval gaps, and rework are counted, and the vendor-selection discipline that turns AI spend into capability that ships.

AUTHORS

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

AUGUST 2026
MANILA · ESTABLISHED 2001

Contents

ABOUT THIS SERIES — VOLUME 47

This is the umbrella volume for PITON-Global's AI & LLM sub-series. Every AI-services proposal is sold on the breadth of workstreams a vendor can staff — data, training, eval, deployment, monitoring — because a long capability list is easy to print. This paper is about the only AI outcome a program's budget actually depends on: capability that ships — a model measurably better on the tasks that matter, running reliably in production. The gap between engaging AI services and shipping capability is where wasted compute and stalled experiments 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-program capacity since the first offshore data teams, and learned that the vendor with the longest capability deck and the most workstreams engaged is the most expensive one a program can hire — because breadth without shipped capability is compute burned on experiments that never reach production. His standing question ignores the services matrix: of the AI work you funded, how much shipped as a measurable capability gain running reliably? Applied at PITON-Global, that question is what clients get on every LLM-ops sourcing decision — personally, with no account layer in between.

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

Ralf Ellspermann

Chief Strategy Officer

Ralf has spent twenty-five years running Philippine delivery floors, latterly the AI-ops pods that support model programs end to end, and he judges a pod by its shipped-capability rate and production reliability, not its breadth of engaged services: whether experiments reached production as measurable gains, or stalled as sunk compute. The LLM-ops contract in Section 3 codifies what he learned rebuilding pods that ran every workstream and shipped none. He now reads AI providers the way he once ran them — from the ship log and the eval record 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 AI pod's real shipped-capability rate and production reliability rather than its services-engaged 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

-56%
Fully-loaded cost per shipped capability gain vs. US onshore

3.1×
Experiment-to-production ship rate on vetted pods vs. mills

-49%
Compute wasted on unshipped experiments, vetted pods

6.3×
First-year ROI, reconstructed case (Section 6)

An AI-services engagement is bought on the breadth of the capability deck and paid for on how little of it ships. The vendor staffs a dozen workstreams across the LLM lifecycle at a fraction of onshore cost; the program finds the fine-tunes that never beat the baseline, the evals that measured the wrong thing, the deployment that stalled in staging, the compute burned on experiments nobody shipped, and a capability roadmap that looked busy while the model in production barely moved. The pod ran every service and the program inherited sunk compute, a stalled roadmap, and a model no better than last quarter. The gap between engaging AI services and shipping capability is where the real cost of cheap AI programs hides, because a workstream that runs but never ships is not capability delivered — it is compute and quarters spent for nothing.

The Philippines — with a deep, English-fluent, technically-trained talent pool that has matured across the full data-and-model lifecycle — is where AI & LLM services are delivered with the most shipped-capability discipline, because that pool can run the lifecycle as one accountable pod rather than a menu of disconnected workstreams. Five findings:

1. **Services engaged is the industry's most flattering number.** Any vendor can staff a workstream and bill the hours. Vetted pods are measured on outcome: shipped-capability rate, production reliability, and compute efficiency. Weak providers quote a broad, cheap capability menu the program repays in sunk compute and stalled roadmaps. Section 2 shows the gap.

2 **Shipped capability, not engaged breadth, is the product.** A model that helps the business comes from running the lifecycle end to end against a capability target and shipping it evaluated — not from staffing every box on a services matrix. Pods that ship run fewer workstreams better; menus that engage everything ship little and cost more per gain that reaches production.

3 **Unshipped work cuts both ways — sunk compute and opportunity lost.** The experiment that never ships burns the GPU spend and the quarter; the capability that could have shipped but was crowded out by busywork is the win a competitor banks. Vetted pods hold ship rate high precisely because both tails cost — the compute wasted and the capability foregone.

4 **Compute and rework, not the hourly rate, are what cost.** A program that ships little pays for the GPUs regardless, re-runs experiments that a proper eval would have killed early, and re-does deployments that were never production-ready — many times the labor saving. Pods that ship efficiently cut wasted compute and rework, and that reclaimed spend — not the offshore rate alone — is where the case study's return is built.

5 **Contract on shipped capability, not engaged services.** The contracting failure of AI programs is paying for a capability menu and hoping some of it ships. Vetted deals price against shipped-capability targets, production-reliability floors, and compute-efficiency standards, making capability that reaches production the pod's problem, which is where it belongs.

The intended reader is a VP of AI/ML, head of data, or CTO buying AI & LLM services on breadth and paying for it in sunk compute, stalled roadmaps, and a model that will not move. Sections 2–4 separate the pods that ship capability from the vendors that engage services; Sections 5–7, the method and the proof. The sub-series volumes (WP-49 training, WP-50 evaluation, and onward) go deep on each lifecycle stage this umbrella frames.

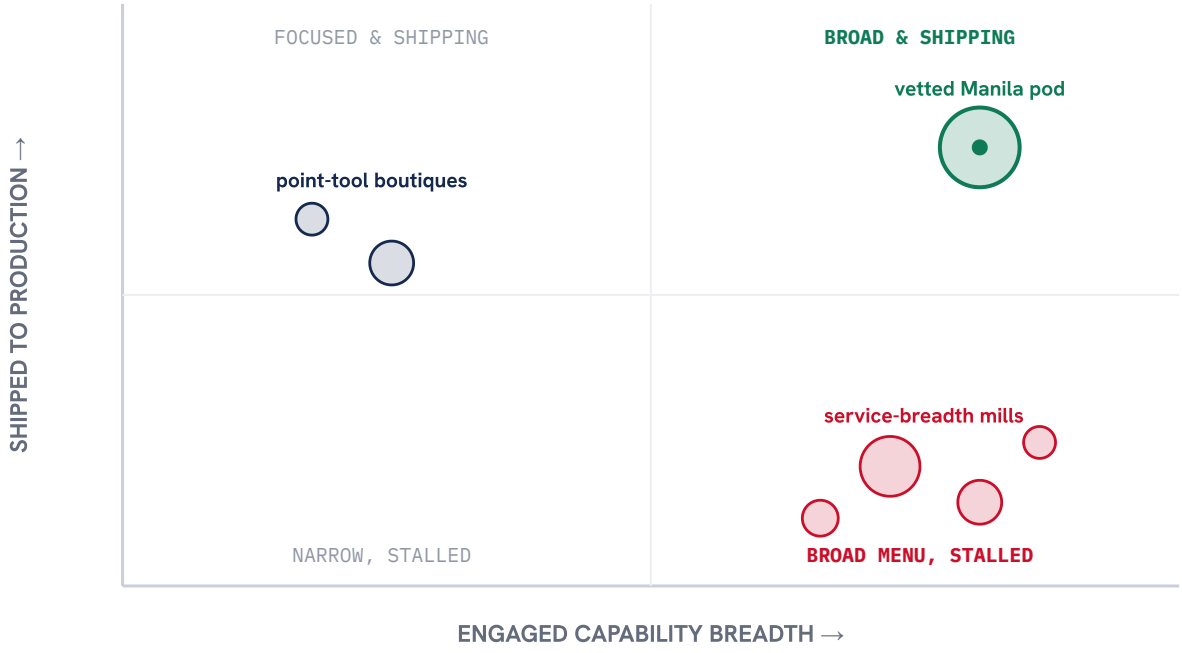
SECTION 02

The breadth mirage: services engaged versus shipped capability

Figure 1 maps AI providers on the two axes that decide the outcome: how broad a capability menu they engage (horizontal) against how much of it actually ships to production (vertical). Breadth is easy to sell and clusters providers to the right; ship rate is hard to sustain and lifts only the disciplined into the top band.

The engaged-services count is a chatbot's containment rate for AI programs — a figure engineered to look like progress while sunk compute accumulates underneath.

FIG. 1 — ENGAGED BREADTH VS. SHIPPED CAPABILITY: THE PROVIDER MAP



Positions are illustrative of PITON-Global diligence 2025–26: bubble size ≈ program spend. The market crowds bottom-right — broad menus that stall; disciplined pods reach top-right — broad and shipping.

The economics follow the vertical axis, not the horizontal. A provider parked bottom-right engages every workstream and ships almost none — the program pays for the whole menu and banks a fraction; a pod in the top-right runs a deliberate breadth and ships most of it, so spend converts to capability. Breadth that does not ship is not a saving on a broad rate card; it is compute and quarters spent to sit still. The shipped-capability read — ship rate, production reliability, compute per shipped gain — is the cheapest due-diligence instrument in this paper.

SECTION 03

The LLM-ops contract: scope to capability, run it reliably, ship it evaluated

Figure 2 lays out the three clauses of a working LLM-ops contract — the machinery a floor read verifies before any broad capability deck is believed.

FIG. 2 — THE LLM-OPS CONTRACT, AS OPERATED ON VETTED PODS

SCOPE TO CAPABILITY

Work is scoped backward from a named capability target the business needs, not forward from a menu of services on offer — so the pod runs the workstreams that move that target and declines the ones that just fill the roadmap. Capability is the unit of work, so compute is spent on what ships.

RUN IT RELIABLY

The lifecycle runs as one instrumented pipeline — reproducible experiments, tracked runs, compute budgets, and staged rollout — so failures are caught early and cheap rather than in a stalled production push. Reliability is what keeps compute from leaking into re-runs and dead ends.

SHIP IT EVALUATED

Nothing ships without an eval that proves the capability gain against a held-out baseline, and nothing counts as shipped until it runs reliably in production — with a ship-and-eval report the program can read. An evaluated ship is the only outcome the roadmap actually feels.

↓ yields ↓

SHIPPED CAPABILITY

HIGH SHIP RATE | RELIABLE IN PROD | EVAL-PROVEN

Work is scoped to capability, run reliably, and shipped only when an eval proves the gain — so compute converts to production capability, experiments stop stalling, and the roadmap moves. The program gets cheaper per shipped gain and faster at the same time, which is the entire promise of outsourcing AI to a pod that measures shipped capability, not services engaged.

Operating parameters from PITON-Global AI-ops engagements, 2025–26. Client owns model IP, data governance, and release decisions throughout.

In diligence, the contract is tested from the ship log's side: pull the last two quarters of experiments and trace how many shipped as evaluated capability gains running reliably; read the compute-utilization and re-run records; interview on how they kill a losing experiment early. Perhaps one pod in ten survives it cleanly. Steps 3 and 5 of the framework exist to find that one before the contract does — because everything in the value bridge of Section 6 depends on shipped capability, not services engaged.

SECTION 04

Cost and performance: cost-per-shipped-capability economics and benchmarks

Table 1 reads the LLM lifecycle as a capability-coverage matrix rather than a price list: which stages each operating model genuinely owns end to end (●), partially staffs (◐), or leaves to the client (○). Breadth of ● marks is not the point — depth that ships is; a mill's row of ◐ is exactly the menu that engages everything and ships little.

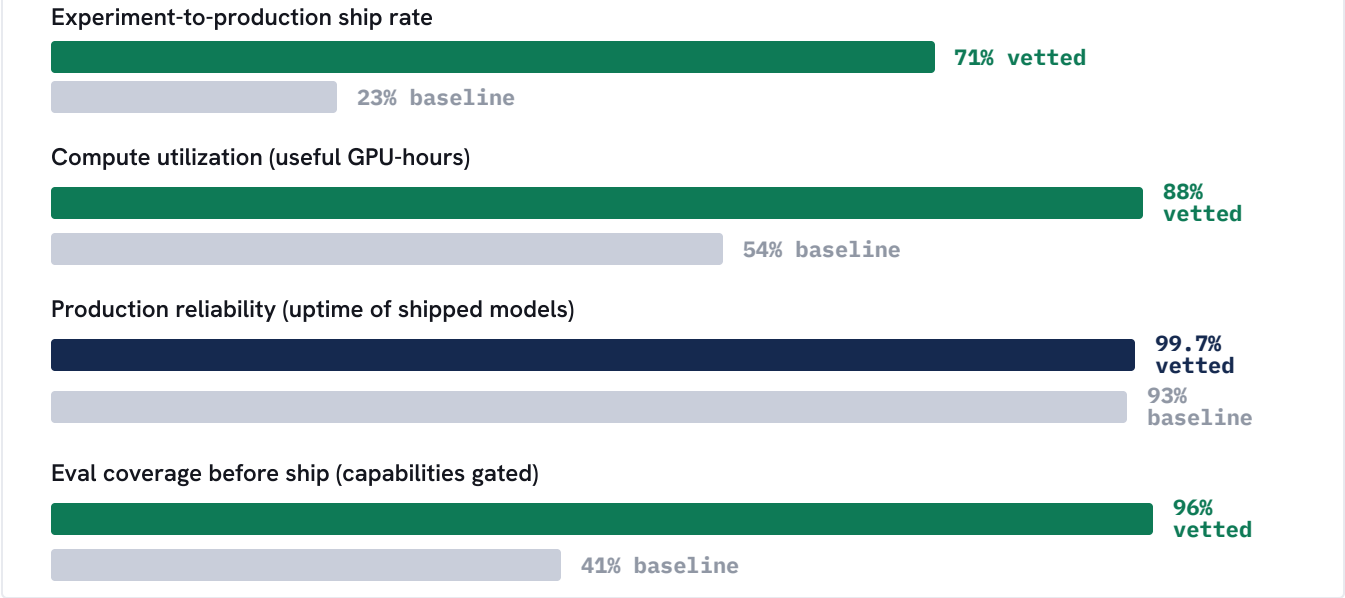
TABLE 1 — LLM-LIFECYCLE CAPABILITY COVERAGE, BY OPERATING MODEL (2026)

LIFECYCLE STAGE	US ONSHORE	OFFSHORE MILL	VETTED POD
Data collection & curation	●	◐	●
Training & fine-tuning (SFT)	●	◐	●
Evaluation & RLHF	◐	○	●
Deployment & serving	●	○	●
Monitoring & drift response	◐	○	●
Safety & red-teaming	◐	○	●

● owns end to end · ◐ partial / hand-off · ○ left to client. Source: PITON-Global AI-ops diligence 2025–26. The mill's ◐/○ rows are the breadth that engages without shipping; the vetted pod owns the stages that convert compute to production capability.

Table 2 shows the operating benchmarks as paired bars — vetted pod against the market baseline — on the four measures that decide cost per shipped capability.

TABLE 2 — LLM-OPS BENCHMARKS, VETTED POD VS. BASELINE (2025–26)



Source: PITON-Global AI-ops engagement data 2025–26; baseline = typical pre-engagement operations billed on services engaged. Bars scaled to 100%.

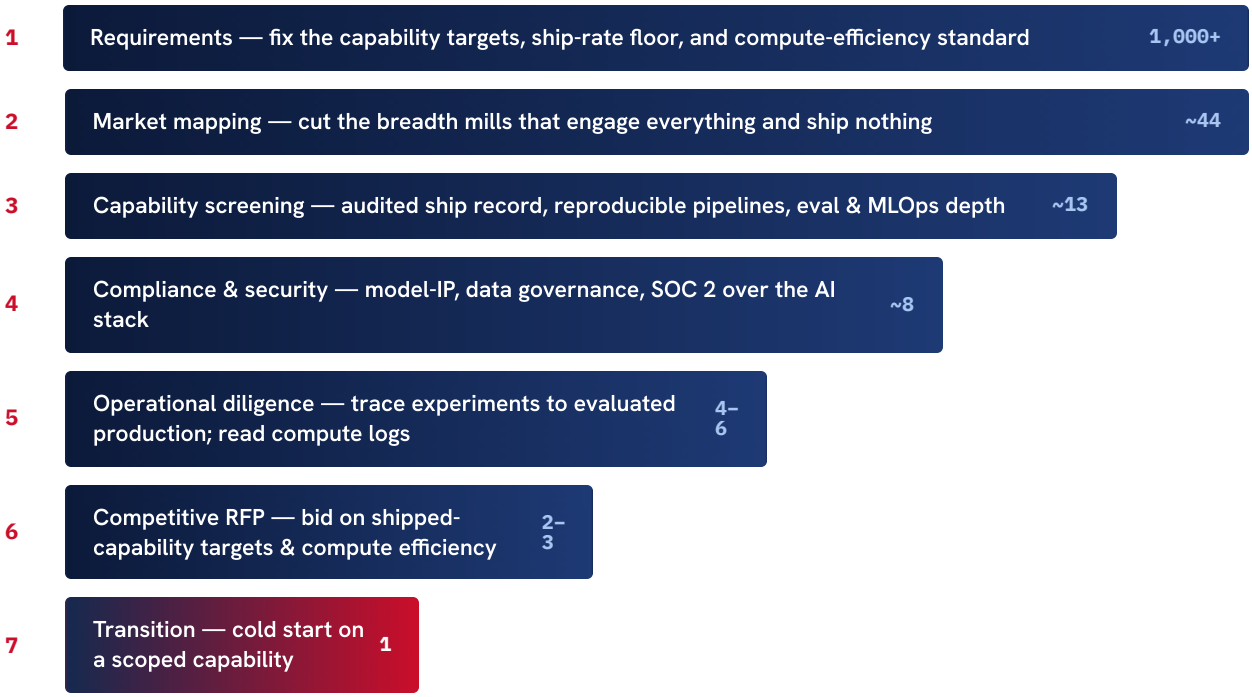
The strategic read: a broad, cheap capability menu is a false economy — a 23% ship rate and 54% compute utilization mean roughly half the GPU spend produces nothing that reaches production, so the true cost per shipped gain runs far above the labor gap. The vetted pod's −56% is real because it is measured where the value is: the capability that ships, evaluated and reliable, at high compute efficiency. Buy on services engaged and you buy the bottom-right of Figure 1; buy on cost-per-shipped-capability and you buy a model that moves — 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 LLM ops

LLM-ops diligence adds one hard requirement to every step: a broad capability deck is real only when the shipped-capability record behind it survives a trace of experiments through to evaluated production. Figure 3 reads the framework as a descending ladder — each step narrows the field, shown by the shrinking bar, for a 40–90-seat AI-ops search.

FIG. 3 — THE 7-STEP FRAMEWORK AS AN LLM-OPS SELECTION LADDER



Bar width ≈ field remaining. Counts represent a 40–90-seat AI-ops search, PITON-Global engagements 2024–26.

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

SECTION 06 · ANONYMIZED CASE STUDY

Re-basing the program on shipped capability: a 50-seat LLM-ops pod, reconstructed

Engagement AL-062 is a US software company — running a product-LLM program across data, fine-tuning, eval, and deployment — that had offshored AI services to a broad-menu vendor. The capability deck was impressive; the shipping was not: ship rate sat near 22%, half the compute budget went to experiments that never reached production, evals were run on a minority of releases, and the model in production had barely moved in three quarters. The VP of AI wanted the lifecycle covered without the sunk compute and the stalled roadmap. Identifying details are anonymized under NDA.

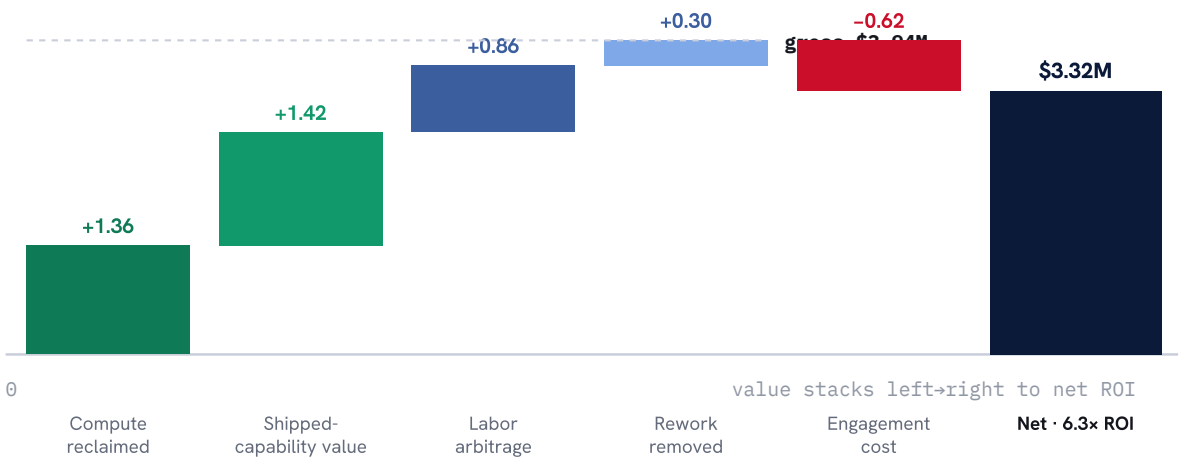
SELECTION (WEEKS 0–8, SHIP-LOG TRACE)

Requirements fixed named capability targets, a 65% ship-rate floor, a compute-efficiency standard, and mandatory eval-before-ship. Mapping produced 44 claimants; the machinery screen — audited ship record, reproducible pipelines, eval and MLOps depth — left 13; the model-IP-and-security audit left 8. A trace of each finalist's experiments through to evaluated production shortlisted 5. A Manila pod won priced against shipped capability and compute efficiency, with the ship-and-eval report written into the spec.

TRANSITION (MONTHS 2–7, SCOPED FIRST)

The pod scoped to a single capability target and shipped it evaluated before widening; the ship-rate floor was certified in month 3. The eval-gate and compute-budget loop ran from day one. By month 7 ship rate reached 70%, compute utilization rose to 88%, eval coverage hit 96%, and three capability gains had shipped reliably to production. PITON-Global chaired the shipped-capability review board to steady state.

TABLE 3 — AL-062 FIRST-YEAR VALUE BRIDGE (US$M)



Value bridge: reclaimed compute + shipped-capability value + labor arbitrage + rework removed = $3.94M gross; less $0.62M engagement cost (advisory fee $0, provider-paid) = $3.32M net ÷ cost ≈ 6.3× first-year ROI. Figures rounded; ship, compute, and eval effects audit-verified.

Reclaimed compute and shipped-capability value together outweigh the labor arbitrage roughly two to one — the signature of a program re-based on shipping rather than breadth. Every bar traces to a diligence step: the capability targets to Step 1, the shipped-capability pricing to Step 6, the eval-gate to Step 3's contract. The VP of AI's line at the year-one review: "The capability deck barely changed. What changed is that the model ships now — and we stopped paying for GPUs that produced nothing."

SECTION 07

The executive playbook: governance for the first 180 days

An LLM-ops transition succeeds when the pod is measured and paid on shipped capability and compute efficiency — client in command of model IP, data, and release decisions throughout.

DAYS 0–30

Define capability, contract on shipping

Name the capability targets, the ship-rate floor, and the compute-efficiency and eval-coverage thresholds with AI leadership before launch, and document the eval-gate by capability. Contract against shipped capability and compute efficiency — never engaged breadth — with targets and floors carrying credits. Stand up the ship-log trace as the program's arbiter of truth.

DAYS 30–90

Scope narrow, ship one thing evaluated

Scope the pod to a single capability target and ship it evaluated before widening; certify the ship-rate floor and the eval-gate. Switch on compute-budget tracking from day one. Publish the LLM-ops dashboard: ship rate, compute utilization, production reliability, eval coverage, capability gains shipped.

DAYS 90–180

Widen the lifecycle, certify steady state

Widen to the full lifecycle only after the pod holds the ship-rate floor and compute standard on the first capability. Review the shipped-capability board monthly with the compute-utilization and eval-coverage trends on the table. Certify steady state on two quarters at floor; hand over on documented criteria; keep the quarterly ship-log trace standing.

THE FIVE FAILURE MODES WE ARE HIRED TO PREVENT

Paying for a broad menu while little ships · staffing every workstream instead of scoping to capability · burning compute on experiments no eval would pass · shipping without an eval that proves the gain · widening the lifecycle before the first capability ships. All five are settled at selection and contract.

The AI-services argument, in one sentence: an engaged service is activity and a shipped capability is a model measurably better in production, and a vetted Philippine pod delivers −56% cost per shipped capability gain with a 3.1× ship rate and compute waste cut by half — for buyers who trace the ship log instead of counting services engaged, and compete the finalists through a fully transparent, fair, comprehensive, and highly competitive RFP. Anyone can engage an AI service. We make sure you buy the pod whose work ships.

METHODOLOGY & NOTES

How the numbers were built

Ship-rate and reliability figures (Fig. 1, Table 2) derive from engagement instrumentation in PITON-Global AI-ops engagements, 2025–26. Shipped capability is measured by tracing experiments to evaluated production; compute utilization, production reliability, and eval coverage are measured against the client's pre-engagement baseline. The baseline column is a composite of pre-engagement vendor reporting reviewed in diligence, shown for contrast.

Cost-per-shipped-capability figures (Fig. 1 read, Table 2) are fully loaded — wages, benefits and statutory charges, facilities, compute and tooling, management and eval/MLOps overhead, attrition-driven recruiting — divided by capability gains shipped to evaluated, reliable production, with unshipped experiments excluded from the denominator; US comparators reflect fully-loaded in-house cost on the same basis. A shipped capability is a measurable gain against a held-out baseline running reliably in production.

The case study (AL-062) is anonymized under NDA; volumes and dates are generalized, figures rounded. Ship, compute, and eval effects are audit-verified; the shipped-capability-value line reflects business value attributed to shipped gains, and the reclaimed-compute 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; shipped-capability and reliability performance re-verified May 2026). Client owns model IP, training data, and release decisions throughout; the pod operates under client governance. Estimates and projections are PITON-Global analysis and should be cited as such. This umbrella volume frames the AI/LLM sub-series; the full series is available at piton-global.com.