ב"ה

THE AI *EITHER/OR* THAT LOSES BY 2027

Established firms are reading the AI threat through the wrong frame. The response window is rapidly closing.

A WHITEPAPER BY

Dr. Yosef Wolf | April 2026

Executive Summary

Established services firms in knowledge-work categories are reading the AI-era competitive threat through the lens of displacement by small AI-native startups. The reading is not wrong in its particulars, rather it is wrong in its conclusion. The firms that built defensive moats over a decade of operation still stand. What the moats buy is runway rather than permanent protection, and the threat that runway leaves unaddressed is not the startup pitching against the incumbent at the enterprise tier. The actual threat is that AI-native production economics are becoming the cost structure of the category. Firms whose economics were designed for the prior era are being priced out of the work, with buyer confidence as the binding constraint.

Of three structural responses, two are insufficient. Layering AI tools onto an unchanged organization captures bounded productivity gains and stops there. The Copilot and Klarna post-mortems make the ceiling visible: Copilot at 3.3 percent penetration of Microsoft's installed base, and Klarna's reversal to human agents after an aggressive automation push degraded customer satisfaction. Building AI capability into the product was the right move in 2024-25, when services firms still competed on human-labor cost structures. *In 2026 the cost basis has shifted*. A firm adding a platform on top of an unchanged services operation now runs two different cost structures simultaneously: the lean platform economics pulling up, and the legacy services economics dragging down. **The two cannot both win.**

The third response restructures the production architecture itself, internalizing the AI-native cost structure instead of competing against it. *Small autonomous pods directing agent fleets, with the human role shifting from executional to directive and evaluative*. The documented cases inside adjacent categories, including Accenture's embedded technologists and Salesforce's agentic squads, show the pattern at scale, though it is not yet documented at scale inside B2B demand generation specifically. **The window for execution is measured in months, not the 18-24 months traditional engagements assume.** *The closing edge is AI-native enterprise-tier credibility, and that inflection is arriving faster than incumbent planning cycles account for*.

The case is built from the freshest Q1 2026 data, named firms on each side of the transition, and the structural-limit reasoning that determines what any of the three responses can and cannot produce. The firms positioned inside the compressed window are a smaller cohort than most services firms realize.

1. What Established Firms Are Reading Wrong

Established services firms in knowledge-work categories are reading the AI-era threat as a competitive question about which firms take which accounts, rather than as a cost-structure question about what it costs to produce category output under different production architectures. The reasoning is familiar enough. Established firms hold defensive moats that ten-person entrants do not reproduce on any short timeline, including first-party data accumulated across a decade or more of operation, partner relationships built one renewal at a time, compliance infrastructure executed across dozens of jurisdictions, and the enterprise client trust that long-cycle relationships produce. Enterprise procurement does not switch core service providers to firms with no track record. The moats are real.

The competitive framing is correct in its particulars and wrong in what it concludes. The moats buy runway, not permanent protection. The threat the framing misses is not a startup arriving at the enterprise tier with a competing pitch, but rather that the production cost structure those startups operate from is becoming the cost structure of the category. Firms whose production economics were designed for the prior era risk being priced out of the work buyers are now demanding.

Exhibit 1. Two readings of the AI-era threat to established services firms

DIMENSION	Common reading	Cost-structure reading
Nature of the threat	Competitors taking specific accounts	Category cost structure resetting under architectural pressure
What moats protect against	Direct displacement by competing firms	Buy runway, not permanent protection
Timeline that matters	When competitors reach enterprise scale and credibility	When category pricing resets under the new cost floor
Binding constraint	Feature parity in client-facing product	Production cost per unit of category output
Required response	Platform investment, agentic features	Production architecture restructured
Failure mode if wrong	Underestimate margin compression	Priced out before platform matures

Source: HLA analysis based on public corporate disclosures and trade press coverage, 2024-2026.

Exhibit 1. Two readings of the AI-era threat to established services firms. The common reading drives platform-side investment, while the cost-structure reading drives operating-model response.

The error in the dismissal is treating AI as a tool and a feature rather than as the substrate that determines what production economics are now possible. A tool-framing produces deployment programs, productivity measurement, and incremental optimization. A feature-framing produces platform investment, agentic capability in the client-facing product, and competitive positioning around proprietary engines. Both framings leave the production base where it was, and neither is wrong inside its own logic. Both are insufficient against a competitive pressure that operates on the cost structure underneath the firm, not on the tools its workforce uses or products the firm sells.

2. The Compression That Already Happened

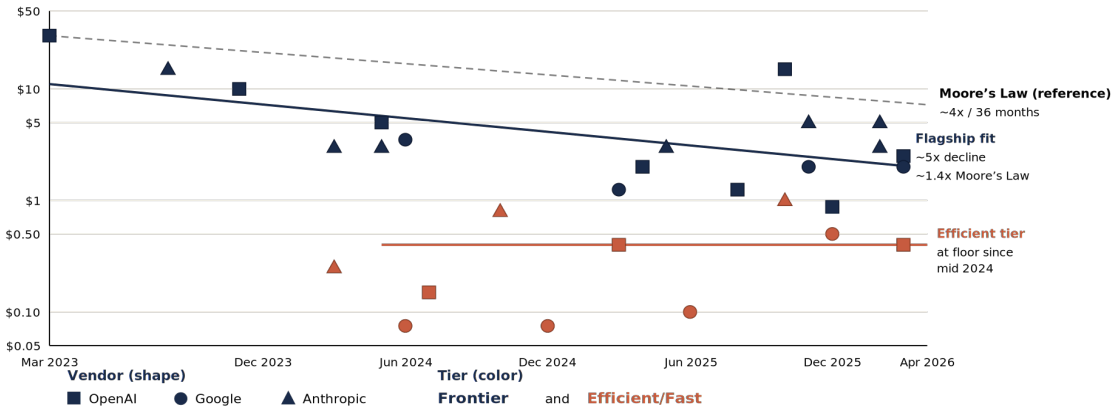
The infrastructure underneath every AI-augmented production workflow is restructuring at compounding rates in multiple independent vectors at once.

Inference cost compression was the 2024-2025 story. Flagship-tier GPT-4-class pricing fell from approximately $30 per million input tokens at GPT-4 launch to a stable band of $1 to $5 by late 2024. The efficient tier fell over the same window and has held its floor near $0.10 to $1.00 since Gemini 1.5 Flash established it in mid-2024. Pricing stability through 2025 and 2026 reflects hardware supply constraints colliding with demand, not the end of underlying algorithmic compression. The question that now drives 2026 planning is not where token prices go next, as token cost is no longer the binding constraint. The fundamental question is what happens to labor costs when the compression that already landed is fully internalized into services production architecture. That internalization is the work of the next 18 months.

Memory compression at the algorithmic layer is collapsing the cost of self-hosted inference in parallel, with research-stage breakthroughs landing in production-grade implementations on cycles substantially shorter than either hardware refresh or organizational adoption. The detail of that compression frontier, and why its timescale determines what ought to concentrate planning attention, sits in the subsection that follows.

Exhibit 2. GPT-4-class inference pricing by vendor and tier

Public API pricing milestones, March 2023 through April 2026. Shape encodes vendor; color encodes tier. Compression landed; tiers have stabilized.



Flagship band stabilized at $1-$5 by late 2024. Efficient tier holds floor at $0.10-$1.00.
2025-2026 pricing stability reflects hardware supply constraints, not end of compression.

Source: HLA analysis from public API pricing pages (OpenAI, Anthropic, Google) at model release dates. Prices are $/M input tokens. Flagship fit is log-linear least-squares across the full window; efficient tier shown as floor line from Gemini 1.5 Flash launch. Moore's Law reference at 2x per 18 months.
Implication: the token-cost compression is the precondition for the labor-cost restructuring question that Sections 4 through 8 address. The compression has already landed.

Exhibit 2. GPT-4-class inference pricing by vendor and tier, March 2023 through April 2026. Flagship-tier pricing declined roughly 5x over the window, at approximately 1.4x the Moore's Law pace, and has stabilized in a $1 to $5 band since late 2024. The efficient tier reached the pricing floor near $0.10 to $1.00 in mid-2024 and has held there. Pricing stability through 2025-2026 reflects hardware supply constraints rather than end of compression. The charted compression is a precondition for the upcoming labor-cost consideration.

Hardware throughput improvements multiply the cost-structure effect of the inference pricing decline and the memory compression breakthroughs. NVIDIA's Vera Rubin platform, announced at GTC in March 2026, delivers an order-of-magnitude increase in token throughput against the prior generation, with revenue guidance from the company indicating sustained demand at scale through 2027.

Managed agent products have reached general availability in Q1 2026, with Anthropic's Cowork and Claude Code leading the category of reliable and secure agentic development. Cowork's early 2026 launch reportedly triggered a $285 billion selloff in enterprise software stocks as investors repriced SaaS incumbents whose products overlapped with agent capabilities. Microsoft has moved from positioning Copilot as its internal standard to actively encouraging adoption of Anthropic's products in multiple teams, with reported spending approaching $500 million annually. Claude Code, a product category that did not exist at credible enterprise scale at the start of 2025, is at $2.5 billion in annualized revenue by Q1 2026. The velocity signal is not about any single vendor or product. It is the rate at which agent categories that did not exist in early 2025 are reaching enterprise availability and substantial revenue inside the timeline conventional planning cycles are scoped against. What the velocity signal forces into view is the question of

whether a firm can deliver buyer confidence at the cycle time the market is moving to, and the cost basis underneath the work is what determines whether the answer is yes.

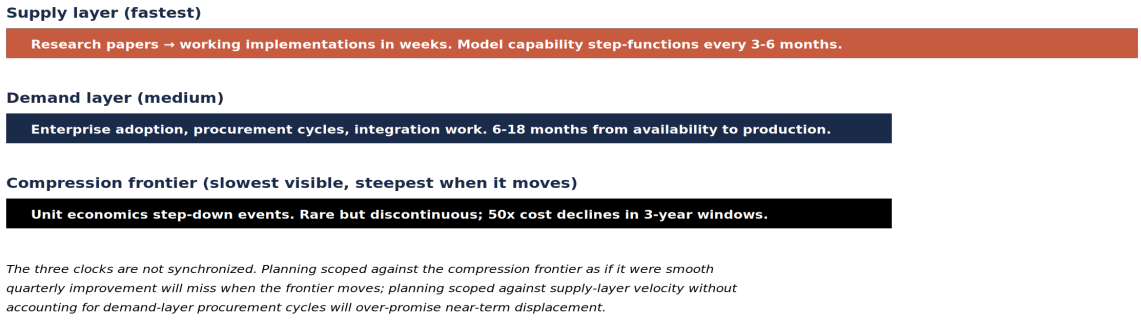
These three forces (stabilized-but-compressed inference cost, hardware pressure on the new band, and the arrival of managed agent platforms) interact multiplicatively, not additively. Their combined effect is that the cost pressure has migrated from the infrastructure layer to the labor layer. The managed-agent launch in early 2026 is what completed that migration, because it is what made the labor restructuring operationally possible at scale. The question of what production architecture a services firm should be operating from twelve months out is now a labor question, not a token question.

The asymmetry that planning is most likely to underweight.

The three forces in this layer do not move on the same clock. Fab supply moves on cycles measured in years, with new DRAM facilities taking three to five years from groundbreaking to volume production and multi-year hyperscaler supply contracts already locked in. Demand moves on adoption curves measured in quarters, bounded by procurement cycles and the friction of organizational change. ***Compression moves at the speed of research papers***, with no new hardware, no new manufacturing capacity, and no organizational change required to deploy what a research team published last week. The asymmetry between these three timescales matters more than any single compression ratio. The fastest-moving variable in a system whose other variables are structurally slow is the variable most likely to produce the discontinuous step that the rest of the planning has not been built to absorb.

Exhibit 2b. Three timescales of change in the AI infrastructure stack

Supply, demand, and compression frontier operate on different clocks. Planning cycles scoped against the slowest are structurally mispriced.



The three clocks are not synchronized. Planning scoped against the compression frontier as if it were smooth quarterly improvement will miss when the frontier moves; planning scoped against supply-layer velocity without accounting for demand-layer procurement cycles will over-promise near-term displacement.

Source: HLA analysis. Supply-layer pace from arXiv-to-implementation tracking and public model release cadence. Demand-layer estimates from enterprise procurement benchmarks. Compressi

Exhibit 2b. Three timescales of change in the AI infrastructure stack. Supply, demand, and compression each advance on different clocks. The fastest-moving force is the most underweighted in conventional planning.

The pattern that ought to concentrate attention is the algorithmic compression frontier. In March 2026, Google Research published TurboQuant, a memory-compression algorithm that compresses the working memory transformers use during inference by approximately 6x with no measurable accuracy loss and no retraining required. Five independent community implementations shipped within two weeks of the paper, with integration into the standard inference frameworks underway. Concurrently, work demonstrating that the transformer can host arbitrary deterministic computation inside its own forward pass, including a published implementation of a WebAssembly interpreter compiled directly into model weights, has begun to suggest that the working memory of the model is not just a cost line item but the constraint on what classes of computation become possible. ***The combination is a step-function rather than another quarterly increment***: the same GPU that previously served nine concurrent users now serves roughly fifty, and effective context windows of approximately 100,000 tokens are extending to roughly 600,000 without new hardware purchase.

The implication for incumbent planning is that the cost basis competitors will operate against on the far side of a step-function landing inside the planning window is not the same cost basis the planning was scoped against. Continuous-curve assumptions about quarterly compression of margin understate the discontinuity that lands when a compression breakthrough reaches production scale. ***The window between the step landing at production scale and AI-native firms pricing against the post-step economics is the window the structural response has to fit inside***.

> *"The dividing line in 2026 will be between B2B marketing organizations that are AI-enhanced and those that are truly AI-native."*
>
> **— Al Lalani, Omnibound AI, in Demand Gen Report, December 2025**

. . .

3. AI-Native Cost Structure: Actual Competitive Pressure

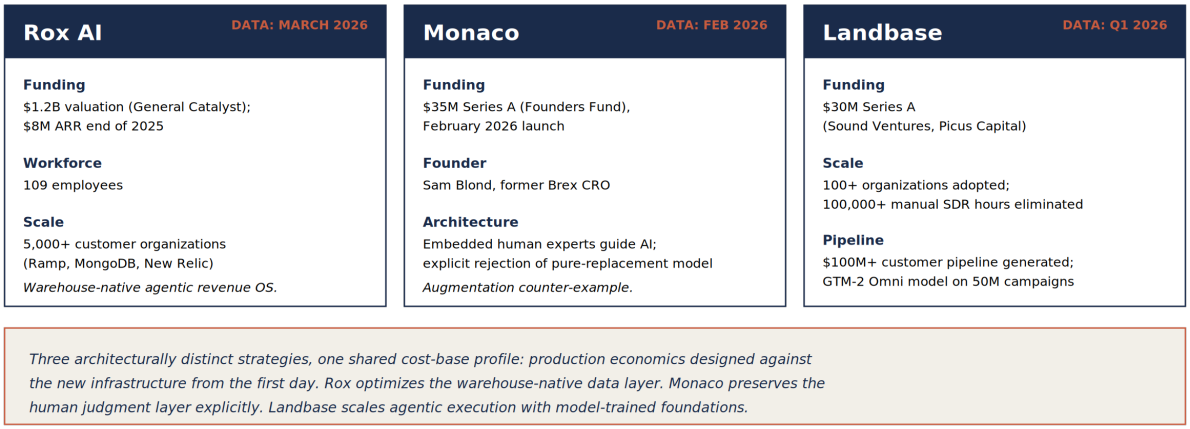
The AI-native entrants in B2B demand generation are not winning on the strength of being smarter or faster. They are winning because their production architecture was designed against the new economics from the first day, while the established firms' production architecture was designed against the old economics over the last fifteen years.

The gap between a traditional services architecture and an AI-native services architecture is structural in the way a Carnot efficiency limit is structural. That is, a ceiling set by the architecture itself, not a margin that better execution can close. *A firm that optimizes relentlessly inside a traditional production architecture will approach that ceiling but cannot exceed it*, and the ceiling sits below the point where AI-native cost structures operate.

The pattern in the freshest Q1 2026 cohort is consistent. Rox AI raised a $1.2 billion valuation round in March 2026 (General Catalyst-led) with 109 employees serving more than 5,000 customer organizations including Ramp, MongoDB, and New Relic, positioned as a warehouse-native agentic revenue OS. The $8 million ARR against the $1.2 billion valuation is itself a signal worth reading carefully: the 150x revenue multiple reflects capital-market repricing of the category rather than proven operating economics, and the repricing affects what multiples traditional services firms can command for their own equity even before the operating proof arrives.

Monaco, launched in public beta in February 2026 with $35 million Series A funding led by Founders Fund and founded by former Brex CRO Sam Blond, has built an explicit counter-example to the pure-replacement narrative: embedded human experts guide AI actions, not agents pretending to be humans. Landbase closed a $30 million Series A from Sound Ventures and Picus Capital in Q1 2026, with more than 100 organizations adopted, over 100,000 hours of manual SDR work eliminated, and more than $100 million in customer pipeline generated by its GTM-2 Omni action model trained on 50 million B2B campaigns and 175 million sales conversations. Daydream, which closed a $15 million Series A from WndrCo in April 2026, has built an AI-native SEO agency combining agent fleets with senior human experts, reinforcing the pattern that the strongest of the new entrants do not eliminate human judgment but redesign the role around it.

Exhibit 3. AI-native B2B revenue and demand-gen entrants: Q1 2026 profile



Source: HLA analysis from public funding announcements, founder communications, and customer-reported metrics, Q1 2026.

Exhibit 3. Funding, scale, and compression claims across three named AI-native entrants. The pattern is the architecture, not any individual firm.

The defensive read on these firms is that customer-reported figures are not independently audited and that the enterprise tier remains protected by procurement complexity. Both observations are correct, though neither removes the structural pressure. The compression of the timeline between mid-market disruption and enterprise disruption is itself a function of the velocity in the unit-economics layer described in the previous section. The mid-market positions where these entrants operate today are the positions established firms expect to defend last, and the cost-structure asymmetry that makes those positions tenable for the entrants is the same asymmetry that will reach enterprise margins on a shorter timeline than incumbent moats imply.

4. Three Structural Responses (and Why Two Are Insufficient)

Established services firms facing this transition have three structural responses available. Two are common, rational, and inadequate. The third is uncommon, harder, and the only one that addresses the actual threat.

Option 1. Use AI to find efficiency, keep the same organization. Productivity tools layered onto the existing workforce, measured against existing workflows, optimized where the tools work and accepted as inadequate where they do not. Industry analysis in Q1 2026 has named the failure mode the Improvement Trap: firms automating broken or obsolete processes, producing zero measurable productivity impact despite AI adoption rates reaching 70 percent in some categories. The failure mode is not that AI does not work. It is that layering AI onto unchanged processes compounds the inefficiencies of those processes instead of eliminating them.

The Microsoft 365 Copilot deployment is the most visible case in the public record. As of the January 2026 earnings call, Copilot reached 15 million paid seats against an installed base of 450 million Microsoft 365 commercial users, a 3.3 percent penetration rate. Recon Analytics survey data from January 2026 shows Copilot's share of paid AI subscribers declined from 18.8 percent in July 2025 to 11.5 percent in January 2026, a 39 percent decline. Of employees with Copilot access, only 35.8 percent use it weekly; the remaining 64 percent of licenses are effectively shelfware. Citi and J.P. Morgan analyst reports document 40 to 60 percent discounting on large deployments. SemiAnalysis recently observed that "Claude for Excel effectively is what Copilot for Excel should have been," capturing in one line the gap that Q1 2026 enterprise data has made legible.

The ceiling these programs hit is not a matter of implementation quality. It is Amdahl's Law expressing itself in organizational form. Amdahl's Law states that the maximum speedup achievable by accelerating one component of a system is bounded by the fraction of the system that does not use that component. A program that speeds up the automatable 30 percent of a knowledge-work process by 20x, while leaving the remaining 70 percent human-serial and unchanged, produces a total system speedup of roughly 1.40x regardless of how much the AI component continues to improve. The 30 percent figure is illustrative; the structural point holds under any reasonable parameter choice, because any fraction of the work left human-serial bounds the total speedup below the component speedup. Efficiency programs that layer AI onto unchanged workflows run into this ceiling by construction, not by misexecution.

Klarna executed an aggressive version of the same logic and is now the best-documented Option 1 case in the public record. Headcount reduced from 7,000 to approximately 3,000, a 49 percent reduction. Revenue per employee climbed to $1.24 million by Q1 2026, a 3.6x increase since 2022. Revenue grew 104 percent since Q4 2022 while operating expenses declined 8 percent. The

AI assistant handled 2.3 million customer conversations in its first month with $60 million in projected annual cost savings. In January 2026, Klarna reinstated human agents for escalations after customer satisfaction declined for complex cases, and CEO Sebastian Siemiatkowski admitted that cost had become too predominant an evaluation factor. The structural lesson is that aggressive Option 1 execution hits the ceiling exactly where the human judgment layer was removed.

A distinction worth naming directly. The Amdahl ceiling binds on execution work, where the human-serial fraction is the bottleneck in client-facing production. Coordination work (meetings, translation artifacts, state synchronization) has a different ceiling and a different structural response. Firms currently running coordination-compression experiments at scale (Meta's defragmentation push, Nvidia's wide-span model, the broader middle-management compression pattern Gartner projects for 2026) are working on a different problem than the one Klarna addressed. Coordination compression has its own failure modes, including retention strain and cultural drift, and none of these experiments has yet produced a clean terminal outcome. What the distinction establishes is that Option 1 is not a single strategy. Applied to coordination work, it produces the compression pattern now being tested across multiple firms. Applied to execution work, it hits the Amdahl ceiling documented in the Klarna case. Execution compression at scale requires the architectural change Option 3 describes; coordination compression does not, which is why the two patterns produce different trajectories under the same Option 1 label.

> *"Cost had become too predominant an evaluation factor."*
>
> **— Sebastian Siemiatkowski, CEO, Klarna (January 2026)**

Option 2. Build AI capability into the client-facing product. Proprietary engines, unified intelligence layers, agentic features in the platform. This is the dominant move among the most visible firms in the category. Informa TechTarget formed in the 2024 Informa-TechTarget combination and signed a strategic platform integration with Demandbase in April 2025 to embed intent data directly into Demandbase One. Demandbase has expanded its ABM tooling into a unified GTM platform. Typeface launched a Marketing Orchestration Engine in 2025 positioned as a coordination layer among people, agents, and systems. The strategic logic was correct in 2024, but is no longer sufficient in 2026 because a firm executing this response on a traditional services cost base ends up running two cost structures against each other inside the same company. Platform revenue grows, but the production cost base does not meaningfully shrink. The platform's pricing power is then eroded by AI-native competitors whose production cost structure was designed for the era the platform is being sold into.

Option 3. Restructure the production architecture so the AI-native cost structure is internalized, not competed against. Small autonomous pods of three to five humans each, with an agent fleet performing the volume that previously required a substantially larger team and meaningful P&L visibility at the pod level. The freshest public articulation of this response sits in Block's April 2026 essay co-authored by Jack Dorsey and Roelof Botha, which lays out an explicit decomposition: information routing handled by an AI world model querying machine-readable artifact state throughout the firm, sensemaking assigned to Directly Responsible Individuals with time-bounded authority over specific cross-cutting problems (one named example holds full authority over merchant churn for ninety days, with the ability to pull resources from multiple teams), and accountability held by player-coaches who continue to write code and design interfaces while coaching the practitioners around them.

Block is architecturally closer to what a mid-sized services firm would execute than either Accenture or Salesforce because it is founder-led, explicitly decomposed, and structured to prevent middle-management accumulation by design. Accenture's 2025 restructuring program of $865 million moved technology practitioners out of a centralized function and embedded them within business units, with company-reported revenue per employee approximately 15 percent higher through the 2024 to 2026 window. Salesforce restructured internal sales and service into agentic squads, with company-reported reductions of 30 to 50 percent in manual campaign management time and 25 percent in delivery timelines. Both are company-reported, not independently audited, and the pattern is not yet documented at scale inside B2B demand generation specifically.

The tier split visible in Exhibit 2 has direct operational consequence inside the pod. Routing volume through the efficient tier captures the cost-structure advantage the exhibit describes, but the efficient tier does not handle every task well. Work that requires frontier capability. Nuanced judgment, long-horizon reasoning, high-context synthesis still belongs on the flagship tier, and work that can run at the efficient tier needs to be routed there deliberately. Matching the model to the task is the operational skill the pod member learns first. A pod that runs everything on the flagship tier pays flagship-tier costs and loses the compression that justified the operating-model move; a pod that runs everything on the efficient tier produces quality failures on work the efficient tier cannot support. The selection is a learned judgment, not a configuration setting.

Exhibit 4. Three structural responses to AI-era cost-structure pressure

DIMENSION	Option 1: Efficiency overlay	Option 2: AI as product feature	Option 3: Production restructured
What it addresses	Bounded productivity gains	Pricing power via differentiation	Cost basis under the firm
What it leaves exposed	Org structure unchanged; Improvement Trap	Production cost base unchanged	Cultural and selection risk
Documented ceiling	~10-15% efficiency, then plateau	Margin compression continues	15%+ revenue/FTE reported
Public Option 1 cases	Microsoft 365 Copilot (15M seats, 3.3% penetration, Q1 2026) Klarna (5,500 to 3,400; partial reversal)		
Public Option 2 cases		Informa TechTarget + Demandbase Typeface Marketing Orchestration Engine	
Public Option 3 cases			Block (Dorsey/Botha decomposition) Accenture (technologists embedded) Salesforce Agentforce squads

Source: HLA analysis. Block decomposition per Dorsey/Botha essay (April 2026). Accenture (Sept 2025), Salesforce (2025-26), Klarna (2024-26). Outcome figures company-reported.

Exhibit 4. Three structural responses to AI-era cost-structure pressure, with publicly executing firms named under each. Option 3 is what makes Option 1's gains durable and Option 2's pricing power defensible.

5. Why the Third Response Is More Than Efficiency

The rest of this analysis concerns what happens to the labor line of the P&L when the cost compression documented in Section 2 is fully internalized into production. The question moves from token economics to the economics of the human team that directs the tokens. This is where the argument lives.

The reason the third response amounts to more than an efficiency play sits in the nature of the human role inside the pod. The pod member is not a worker assisted by an agent. The pod member is a person whose primary output is the specification of outcomes, the auditing of agent work against those outcomes, the detection of drift and hallucination in the work the agents produce, and the redesign of workflows when the agent fleet hits the boundary of its competence.

Executor mode: one human, one output, no compounding

The legacy unit of production. Output is bounded by the human hours that go in.

Executor mode. The legacy unit of production: one human, one output unit, no compounding. Output is bounded by the human hours that go in.

Exhibit 5a. The legacy production unit: a 30-person account team, sequential

Specialized functions, sequential handoffs, executional human roles

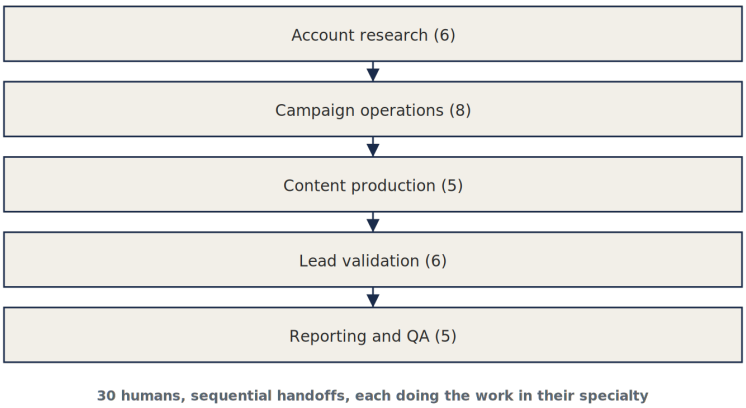
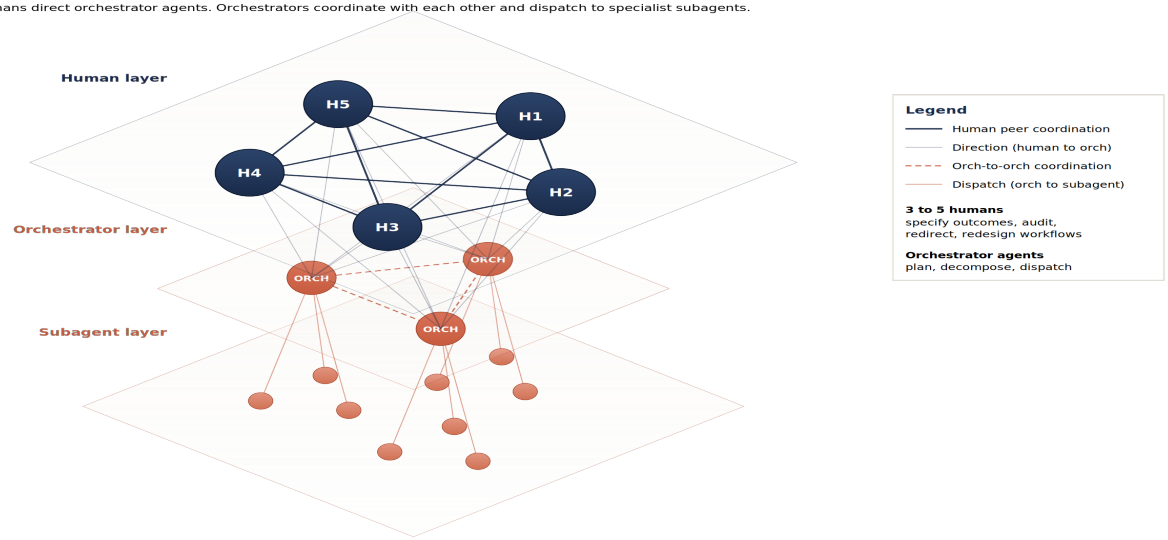


Exhibit 5a. The legacy production unit. A 30-person account team organized by specialty, with sequential handoffs and executional human roles.

Exhibit 5b. The pod production unit: three tiers of coordination

Humans direct orchestrator agents. Orchestrators coordinate with each other and dispatch to specialist subagents.



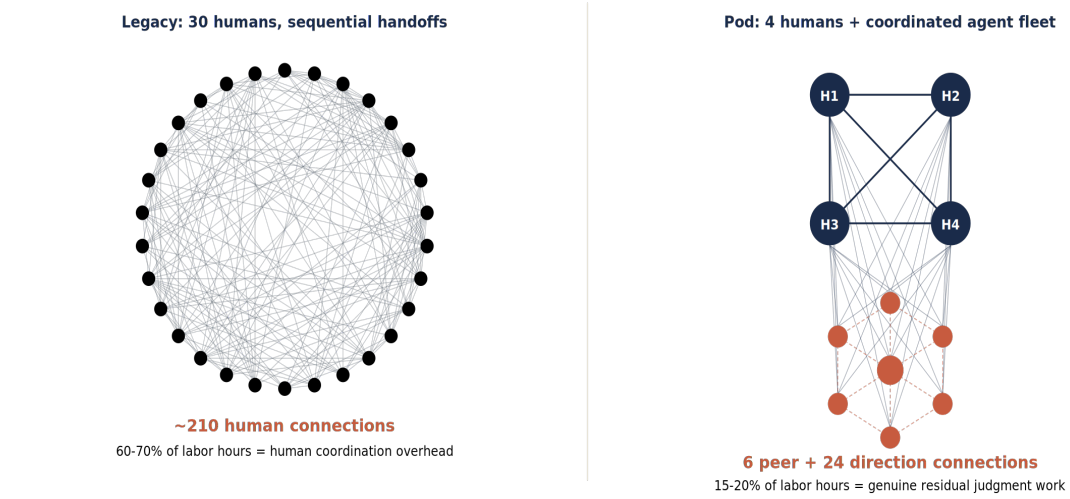
Direction flows down through tiers; agent coordination runs at machine speed.

Source: HLA pod design framework. Multi-tier agent topology per production orchestration systems (Anthropic Claude Code, Cowork, Q1 2026).

Exhibit 5b. The pod production unit: three tiers of coordination. Humans direct orchestrator agents, which coordinate with each other and dispatch to specialist subagents. The human role is directive and evaluative; throughput is set by the agent fleet, not by human hours.

Exhibit 5c. Human coordination overhead, two structures compared

The axis is human coordination hours. Pod agent fleet shown at orchestrator layer only; subagent coordination runs below the line.



Source: HLA pod design framework. Multi-tier agent topology shown in detail in Exhibit 5b.

Exhibit 5c. Coordination overhead, two structures compared. The legacy 30-person account team carries ~210 communication connections; the 4-person pod with a single-layered shared agent fleet carries 30. The structural innovation is the elimination of coordination overhead, not just the automation of tasks.

> *The **pod** is neither a team in the conventional sense nor an automation project. **It is a new unit of production in which the human role is primarily directive and evaluative rather than executional.***

The job description is different enough from manual execution that the transition cannot be accomplished through training alone. It requires a reframing of what the contribution is, how it is measured, and what makes the person valuable. ***Workers trained for decades to derive professional identity from being the person who does the work do not instinctively become the person who directs a system that does the work***. That conversion is a cognitive and identity-level shift before it is a workflow change.

Organizations that treated the conversion as a tooling rollout consistently captured a fraction of the upside that the tools themselves had made available. Firms running the second response on an unchanged production base are increasing the pricing power they project to clients while leaving the cost structure underneath that pricing power exposed. Firms running the first response are capturing the bounded gains that productivity tools deliver and stopping there. Firms running the third response are rebuilding the production layer so the platform's pricing power is defensible and the efficiency gains compound instead of plateauing.

Decomposing the management role

The directive-and-evaluative framing sharpens further when management work itself is decomposed. Knowledge-work managers perform three bundled functions: information routing, sensemaking, and accountability. Routing is fully automatable; agent fleets handle it. Sensemaking is human-centered work that translates noise into signal, buffers the team from irrelevant information, and makes sharper decisions where tacit knowledge matters. Accountability is fundamentally human; ownership of outcomes across long time horizons, honest feedback delivery, and identity-level care for the work do not transfer to systems that do not have skin in the game. The pod member holds sensemaking and accountability. The agent fleet holds routing and execution.

The ARR-per-FTE benchmark

The cost-structure asymmetry is now apparent: AI-native firms are averaging $1.13 million in ARR per FTE. Traditional SaaS benchmarks sit substantially lower. Atlassian at $0.46 million per FTE, GitLab at $0.54 million. High performers in the AI-native cohort run materially higher: Cursor between $6.1 million and $25 million per FTE, Gamma at $2.0 million, OpenAI at $1.5 million. Burn multiple, defined as free cash flow divided by net new ARR, runs at 0.4x for AI-native firms under $100 million ARR against 2.0x for the median non-AI SaaS comparable. AI-native firms scale with roughly eighty percent less capital than prior-era SaaS. BCG analysis indicates that seat-based pricing for AI products produces 40% lower gross margins than outcome-based pricing, suggesting that even the pricing model of the legacy SaaS business is structurally exposed.

The coordination collapse

The compression compounds because the pod model eliminates coordination overhead, not just because it automates tasks. In a typical 200-person knowledge-work firm, 60-70% of labor hours go to coordination: meetings, translation artifacts, state synchronization, handoff management. The remaining 30-40% constitutes genuine creation and judgment work. Roughly half of that is verifiable execution that agent harnesses can already perform. The other half is genuinely hard residual work: zero-to-one insight, political navigation, moral judgment, aesthetic taste at the frontier. **The pod compresses by collapsing the coordination layer, not by speeding up individual contributors.**

Operator evidence

Operator evidence has begun to confirm the compression ratio the pod model assumes. Wade Foster, CEO of Zapier, has stated publicly that AI agents are already performing more than ninety percent of the work, with only the last mile going to human account representatives. SaaStr, a B2B media and events firm, reports running an eight-figure business with single-digit headcount by deploying more than twenty agents that handle 139,000 conversations autonomously. The agent layer that the pod architecture depends on is no longer hypothetical infrastructure. Managed agent products available in production handle the execution volume the pod architecture requires.

Middle-management compression

Middle-management compression data confirms the structural shift. Gartner projects that by the end of 2026, twenty percent of organizations will use AI to eliminate more than half of current middle management roles. LinkedIn data for Q1 2026 shows manager job postings down twelve percent year over year while lead and principal roles are up eighteen percent. Harvard Business School research published in March 2026, analyzing nearly all U.S. job postings from 2019 through March 2025, found that automation-prone roles saw a thirteen percent decrease in postings since the launch of ChatGPT while analytical, technical, and creative roles saw a twenty percent increase. Job postings for automation-prone occupations also required seven percent fewer skills, suggesting employers are not just posting fewer of those jobs but posting simpler ones, consistent with AI tools absorbing the complexity that once required dedicated human labor.

When the judgment layer is removed: A cautionary case

The pod model's structural innovation is the human quality gate sitting between the agent fleet's output and the client. Pure-automation approaches that omit this gate produce predictable failure. Artisan, the AI BDR firm that marketed Ava as a set-it-and-forget-it autonomous outbound system, entered active retrenchment in Q1 2026 with reported customer churn, layoffs, and a G2 rating of 3.9 driven primarily by data-accuracy flags. Users reported that managing Ava consumed up to 30% of the Chief AI Officer's time, contradicting the marketing. The failure mode is not the agent's capability. It is the architectural decision to remove the judgment layer that the pod model places at the center.

The selection problem

Traditional hiring and performance assessment methods were designed to identify executional competence. The directive-and-evaluative role inside a pod is structurally different work, and the constructs that predict performance (tolerance for ambiguity, calibrated trust in automated systems, cross-domain synthesis, mindset plasticity when the output of the work is no longer the output of one's hands) are not reliably captured by existing assessment instruments. A rigorous selection methodology designed specifically for the transition is a prerequisite for executing the structural move at scale**. Off-the-shelf tools do not exist for this problem.**

. . .

6. Structural Conditions That Determine Readiness

The trust-architecture buildout that buyer-side research has now identified as the binding constraint on B2B purchase decisions operates at the production layer. Buyer confidence is not manufactured at the platform layer, where most strategic attention has been concentrated; it is manufactured at the precision and timeliness and contextual fit of the work that lands in front of the buying group. The cost-structure response described in the previous sections is what defends the trust-architecture position. The conditions that follow determine which firms are positioned to execute it.

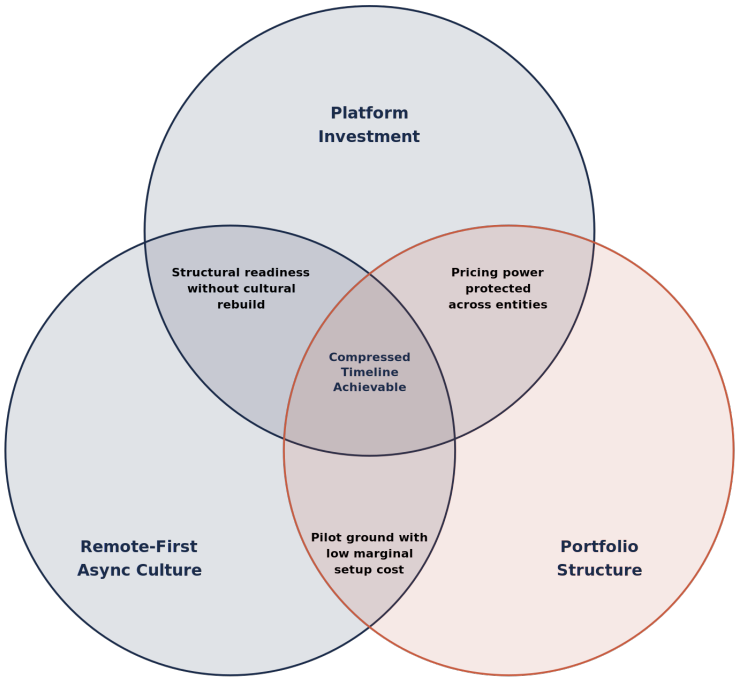
The pod model in the documented cases has typically been built inside organizations that were previously co-located and synchronous. Most of the friction in those transitions was cultural rather than technical. The work involved restructuring communication patterns, decision rights, management cadence, and tolerance for operating without real-time coordination.

Three conditions distinguish firms that are structurally closer to the third response than they appear. Remote-first asynchronous operating culture eliminates the cultural rebuild. Distributed teams in different time zones, asynchronous handoffs as the default, documented workflows, written decision logs, and an organizational tolerance for operating without standing meetings already constitute the cultural substrate the pod model needs.

Portfolio structure provides natural pilot grounds. A firm operating across related entities with shared data and infrastructure can run the pod model in one entity before scaling the structural change throughout the others, and the marginal cost of standing up agent fleets in additional entities is materially lower than starting from scratch in each.

Existing platform investment signals strategic readiness for AI-driven differentiation, and the production-side restructuring is what defends the platform's pricing power instead of leaving it eroded. Firms holding all three conditions simultaneously sit at the intersection where the operating-model move is both most urgent and most achievable.

Exhibit 6. The three structural conditions that determine readiness



Source: HLA framework.

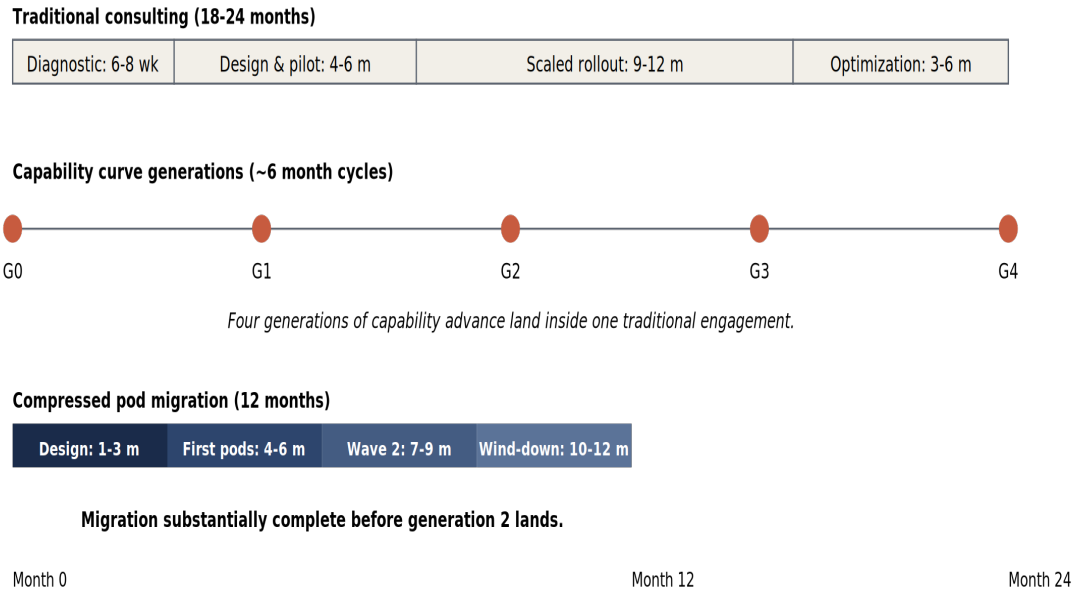
Exhibit 6. The three structural conditions that determine readiness for the operating-model move. Firms holding all three simultaneously occupy the center intersection, where the compressed timeline in Section 7 is achievable without a cultural rebuild.

7. The Timeline Question

Strategy consulting firms approaching transformations of this kind typically propose engagements of eighteen to twenty-four months: a diagnostic phase of six to eight weeks, a design and pilot phase of four to six months, a scaled rollout of nine to twelve months, and an optimization handover of three to six months. The shape is built for categories where the underlying technology and competitive environment are stable enough that a twenty-four-month program design remains valid through to execution. That condition does not hold in the current environment.

Exhibit 7. Capability generations land inside the traditional engagement window

Traditional 18-24 month consulting engagement vs. compressed 12-month pod migration



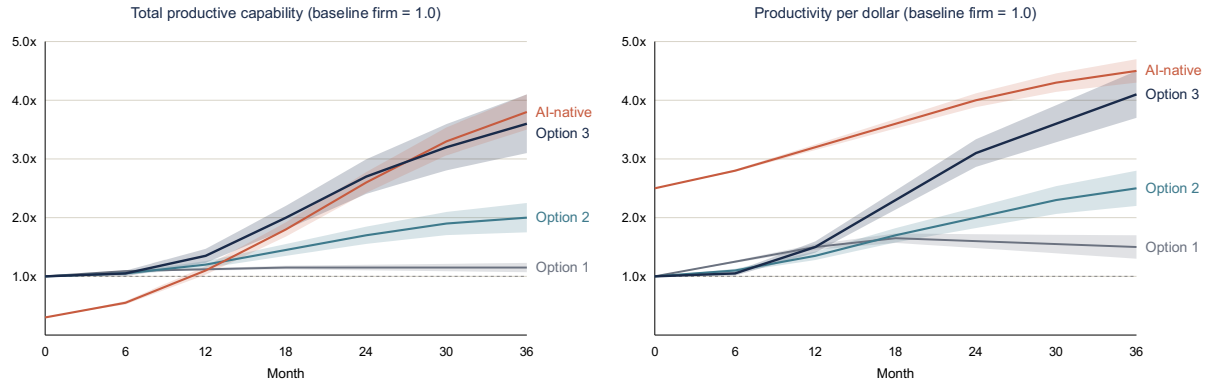
Source: HLA analysis. Traditional engagement structure from public McKinsey, BCG, Bain, Deloitte program descriptions.

Exhibit 7. Capability advance lands inside the traditional consulting engagement window, with intervals that may be irregular: a step-function from the compression frontier can land as a single discontinuous event rather than as a smooth quarterly increment. The compressed pod migration completes before the second generation lands.

The CEO executing Option 1 in the first twelve months captures visible margin expansion. Board reports look strong. What Exhibit 7b shows is that this short-term victory is bounded by the Amdahl ceiling on output and the subsequent margin erosion as quality reversals force rehiring. The firm running Option 3 does not trade a visible trough for the long-term position. **Executed competently, first-wave pods deploy on contained slices of work with AI-savvy employees, absorb the learning curve at small scale, and fund the transition costs of platform and agent infrastructure out of first-wave efficiency gains.** Net productivity holds near baseline through the first six months. Then the compounding starts. The payoff through months 12 to 36 is not recovery from self-inflicted damage; it is the output of an architecture that was designed against the new economics from the first wave. Every dollar of short-term margin gain that Option 1 captures is a dollar that never funded the restructuring capital that Option 3 deploys during the first-wave window. Firms that optimize for short-term margin in 2026 arrive at month 36 trapped below the Amdahl ceiling while competitors who executed the architectural shift reach the new structural frontier.

Exhibit 7b. Modeled productivity and margin trajectories under four strategic responses

Left: total firm output indexed to baseline. Right: productivity per dollar. Band widths reflect execution-risk uncertainty specific to each response.



Source: HLA model. Bands reflect execution-risk uncertainty. Option 1: workforce absorption and quality-reversal pace. Option 2: platform pricing power, adoption velocity, engineering allocation, revenue vs. cost-erosion race. AI-native: enterprise-tier timing and cost-advantage persistence. Option 3: compound risk across pod selection, learning transfer, workforce transition.

Exhibit 7b. Modeled productivity and margin trajectories under three strategic responses. Left panel: total firm output indexed to baseline traditional firm. Right panel: productivity per dollar (output divided by cost), a proxy for operating margin at fixed pricing. Option 1 delivers short-term margin expansion through cost reduction, peaking around month 18 and eroding as quality reversals force rehiring; output caps at 1.15x per Amdahl ceiling calculation in Section 4. AI-native runs high margin from day one with compounding through scale, but total output remains bounded by small headcount. A competent pod restructuring shows baseline-steady performance through the first six months as first-wave pods are deployed, followed by accelerating compounding as subsequent waves benefit from learning captured by the first cohort. The chart shows no visible transition dip because real transition costs (severance, platform investment, agent infrastructure, training) are absorbed by first-wave efficiency gains, not because those costs are hidden. AI-native firms reach enterprise-tier credibility before their total output approaches incumbent scale; see Exhibit 8 for the execution-timing implication of that asymmetry. All projections are modeled under stated assumptions, not observed.

The capability curve is advancing on roughly six-month cycles, while the cost basis has already completed its compression and now sits at the new, lower band that enterprise volume is deploying against. AI-native entrants are reaching credible scale on timelines that do not accommodate a twenty-four-month incumbent restructuring as a defense. The combination is a step-function, not another quarterly increment: the same firm that could credibly plan a three-year transformation against the old cost basis cannot credibly plan the same transformation against the internalized new one.

The continuous-curve framing of the capability advance is itself an underestimate of what the timeline window has to absorb. As the discussion in Section 2 indicates, the algorithmic compression frontier is moving on a timescale that produces step-function change, not incremental, with research-stage breakthroughs reaching production-grade implementation in weeks, not quarters. A step-function in effective inference cost or context capacity that lands inside a twenty-four-month incumbent program is a step the program was not scoped against, and the AI-native firms operating against the post-step economics are pricing into the same client base the incumbent is restructuring to defend. The compressed twelve-month migration window is the timeline that fits both the continuous compounding and the discontinuous step. The traditional eighteen-to-twenty-four-month engagement shape fits neither.

8. Implications and the Diagnostic Conversation

A caveat worth naming: the trajectories in Exhibit 7b are modeled, not observed. The frontier is new enough that no firm has yet executed any of these three responses at scale across a full three-year window with outcomes that are fully measurable. Analyses of this kind proceed under stated assumptions about rates of change that are themselves moving targets. Anyone claiming certainty about the terminal outcome of any of these strategic responses is overclaiming**. What the analysis does support is a question reframe. For leadership at a firm positioned at the intersection of the three conditions in Section 6, the load-bearing decision is no longer which response is provably optimal. The load-bearing decision is whether the risk-reward balance of inaction remains defensible given the direction and pace of the pressure that has now become legible.**

A firm that has already invested in a unified platform layer, that already operates a remote-first asynchronous workforce in many jurisdictions, and that holds adjacent portfolio entities in the AI-native pressure zone is the firm for which the operating-model move is both most urgent and most achievable. The platform investment is real and is increasing the pricing power the firm projects to its clients. The production cost base underneath that platform is what determines whether the pricing power survives when AI-native competitors reach enterprise-tier credibility. The remote-first asynchronous culture is what makes the migration runnable on the compressed timeline. The portfolio structure is what makes the migration testable in one entity before scaling to the others.

The case for executing the response is robust across scenarios, not only the specific future Section 2's velocity argument most vividly describes. Even if AI-native enterprise credibility arrives later than the compressed timeline projects, the pod architecture produces meaningful coordination compression and cost-structure improvement on its own terms. The response is not a single-scenario bet; it is the move that produces superior structural position under the range of futures the current analysis supports.

The load-bearing decision is whether the risk-reward balance of inaction remains defensible.

This is the question the analysis leaves for leadership.

The factors that determine the specific shape of the migration include the production workforce headcount subject to restructuring, the sequence in which workflows are migrated from the legacy structure to the pod structure, the selection approach for identifying which current workers can operate in the new mode, the management cadence that supports pods without recreating matrix overhead, the integration between the platform investment and the pod buildout for engineering capacity allocation, and the strategic question of which portfolio entity hosts the first pilot. None of these can be specified from outside the firm. They are the substance of a diagnostic conversation, not of a document like this one.

Two dimensions deserve explicit naming even at this altitude. The migration shape must preserve customer experience during the transition window. Account continuity, service-level consistency, and client perception of a firm in restructuring determine whether an internal architectural change produces external damage. The timing and sequencing of pod migration against client commitments is a first-order constraint, not an operational detail.

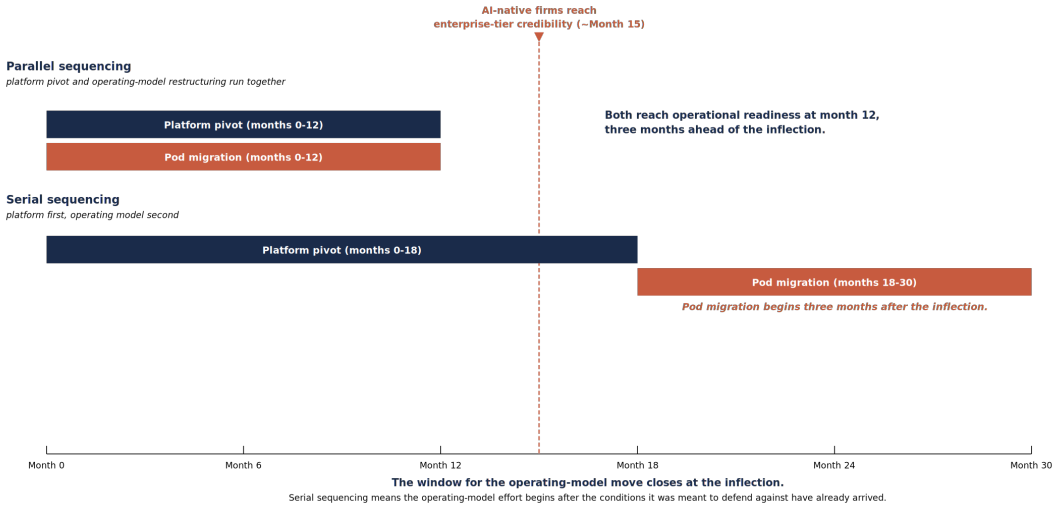
The migration must also engage the workforce dimension honestly. Pod selection defines which workers make the transition into the new architecture. The workers whose roles do not transition require an equally deliberate design: severance aligned with tenure and jurisdiction, internal mobility across portfolio entities where the fit exists, retraining investment for workers whose aptitude spans the transition even if their current role does not. Neither the client dimension nor the workforce dimension is developed in this document. Both are central to the conversation that follows.

The window for executing the production-side restructuring in parallel with the platform pivot, such that both reach maturity together before AI-native entrants reach enterprise-tier credibility, is measured in months, not years. Firms that sequence these moves, completing the platform first and then turning to the operating model, are likely to discover that the platform's unit economics have been priced out of the work the platform was built to capture by the time the operating-model effort begins.

The trust gap that buyers now name as the binding constraint operates at the production layer. The firms whose cost structure permits them to deliver buyer confidence at the cycle time the market now demands will hold the discoverability-to-revenue work, while the firms whose cost structure does not will lose it to firms whose cost structure does. The competitive question is which side of that line a firm finishes on, and the structural question is whether the production architecture the firm is operating from in twelve months is the architecture that side requires.

Exhibit 8. Parallel sequencing closes the window; serial sequencing misses it

Platform pivot and operating-model restructuring, two execution sequences, plotted against the AI-native enterprise-readiness inflection.



Source: HLA analysis. Inflection timing reasoned from AI-native firm scale trajectories, 2024-2026; specific timing of enterprise-tier credibility varies by category.

Exhibit 8. Two execution sequences for the platform pivot and the operating-model restructuring, plotted against the AI-native enterprise-readiness inflection. In the parallel sequence, both reach maturity at month 12, three months ahead of the inflection. In the serial sequence, the operating-model effort begins at month 18, three months after the inflection has already arrived and the platform's unit economics have been exposed to AI-native pricing pressure. The window for the operating-model move is the window between today and the inflection.

Appendix A. Sources and Methodology

Token cost decline. Epoch AI inference price trend data, synthesized in Introl, 'Inference Unit Economics' (Dec 2025) and TLDL, 'LLM API Pricing 2026' (Apr 2026). Flagship-tier decline of approximately 5x over 36 months; efficient tier reached a floor in mid-2024 and has stabilized. Per-milestone pricing milestones for GPT-4 through GPT-5.4, Claude 2 through Claude Opus 4.6, and Gemini 1.5 Pro through Gemini 3.1 Pro drawn from vendor API pricing pages at model release dates.

TurboQuant memory compression. Google Research, 'TurboQuant' (March 25, 2026); five independent community implementations within two weeks of publication. Reported compression of the transformer KV cache to approximately 3 bits per value with no measurable accuracy loss; concurrency gains of approximately 5x revenue per GPU at production scale.

Three-timescale framing. Synthesis of fab-supply, demand-adoption, and algorithmic-compression cycles drawing on Nate Jones, 'The Third Body: Compression as the Fastest-Moving Force in AI Infrastructure' (April 2026), and the underlying TurboQuant primary source above. The asymmetry of the three timescales is the load-bearing argument.

Step-function framing. TurboQuant production-grade implementation timeline (community implementations shipped within 14 days of arxiv publication; vLLM and llama.cpp integration underway; Google official release expected Q2 2026) supports the characterization of the compression frontier as discontinuous rather than incremental at the production layer.

NVIDIA Vera Rubin throughput and revenue guidance. NVIDIA GTC keynote, March 2026.

Landbase growth and cost-reduction figures. Company-reported 2025 metrics; customer-reported cost reductions, not independently audited.

Informa TechTarget and Demandbase integration. Demand Gen Report, 'Informa TechTarget, Demandbase Announce Strategic Partnership' (April 2025).

Typeface Marketing Orchestration Engine. Typeface product launch, 2025.

Lalani quote. Demand Gen Report, 'AI Agents Revolutionized B2B Marketing in 2025' (December 2025), citing Al Lalani of Omnibound AI.

Accenture restructuring. Public corporate announcement, September 2025; $865M program; revenue per employee figures company-reported, not independently audited.

Salesforce Agentforce. Salesforce internal metrics published 2025 to 2026; not independently audited.

Microsoft Copilot adoption and ceiling. Synthesized from 2024 to 2025 enterprise post-mortem coverage and analyst reporting; commercial adoption near 2 percent in select published analyses.

Klarna workforce reversal. Public corporate communications, 2024 to 2025 cycle; 5,500 to 3,400 reduction followed by partial reversal.

Managed agent category Q1 2026. Anthropic public product communications, Q1 2026; SaaS market reaction reporting; enterprise procurement disclosures.

Improvement Trap framing. Industry analysis Q1 2026 covering AI adoption ceilings in enterprise rollouts.

Microsoft Copilot Q1 2026. January 2026 earnings call (15M paid seats); Recon Analytics survey of 150,000 enterprise users (January 2026); Citi and J.P. Morgan analyst reports on enterprise discounting; SemiAnalysis commentary.

Klarna Q1 2026 figures. Public corporate communications, 2024-2026 cycle; CEO Sebastian Siemiatkowski public statements; revenue per employee, headcount, and reversal data per January 2026 reporting.

Rox AI. Series funding announcement, March 12, 2026 (General Catalyst-led); customer disclosures.

Monaco. Public beta launch, February 2026; Series A announcement (Founders Fund); founder communications, Sam Blond.

Landbase Q1 2026 refresh. Series A announcement (Sound Ventures, Picus Capital), Q1 2026; customer-reported metrics.

Daydream. Public Series A announcement, WndrCo lead, April 2026.

Methodology. Analysis combines public corporate disclosures, primary product announcements, trade-press coverage, peer-reviewed and preprint research releases, and HLA-internal pattern synthesis across the 2024 to 2026 enterprise AI cohort. Tier flags applied internally: VERIFIED for claims confirmable from primary sources, REASONED for inferences from adjacent verified material, UNKNOWN for load-bearing claims requiring further verification. Company-reported figures for restructuring outcomes are flagged as such throughout.

Appendix B. About the Author

Dr. Yosef Wolf is the founder of High Leverage Analytics. His work focuses on the operating-model implications of agentic AI for established services firms, with particular attention to the assessment and selection problems that arise when the unit of production shifts from the individual contributor to the human-directed agent fleet. His background combines doctoral work in astrophysics with applied analytical practice in multiple knowledge-work categories.

High Leverage Analytics works with founder-led services firms on the structural questions that determine which firms remain competitive through the AI-native transition. The firm's posture is analytical at the diagnostic stage, with implementation work available when the analysis points at a structural move the client desires to execute. Clients receive the diagnostic frame, the documented evidence base, and the design constraints the structural change has to satisfy. Engagements are sized to the diagnostic conversation first, with downstream scope following the answers that conversation surfaces.