

2027 OPERATIONS, FINOPS & CISO ARCHITECTURE · NOV 25, 2026

The Production Agent Control Plane: Cutting AI Operating Costs and Locking Down Tool and Model Risk in 2027

How COOs, CTOs, CFOs, and CISOs replace open-ended agent loops with compiled workflows, sub-100ms decision routing, and 35-second inference off-switches.

Key Finding: Across 5,115 enterprise automations, runtime step-planning consumed up to 43% of latency and cost; production scale requires compiled workflows, sub-100ms decision routing, and an off-switch outside the agent loop.

Grayson Mylar · Founder & CEO, ALLTIPLY

November 25, 2026 · www.alltiply.com

ALLTIPLY Labs Technical Report · Scheduled November 25, 2026 · Author: Grayson Mylar, Founder & CEO

Core finding: Most enterprise AI teams spent 2025 and 2026 proving that open-ended agents *can* complete multi-step tasks in a sandbox. Heading into 2027 budget approvals and January 1 regulatory deadlines, the operational challenge has flipped: uncontrolled runtime step-planning burns up to **43% of execution time and cost**, general chat models are five times too expensive and twenty times too slow for routine classification, and prompt-level safety instructions cannot stop a tampered checkpoint, an over-privileged MCP connector, or a third-party cloud breach. Scaling AI profitably and safely in 2027 requires a **Production Agent Control Plane** - four execution tiers wrapped inside an external PII/cache gateway and a sub-60-second inference and tool off-switch.

Executive summary

- **Compile repeatable procedures out of the agent loop.** Across **5,115 enterprise automations** analyzed in the November 10, 2026 ReTrace study, 91.8% were built as agents, yet **7 of the 10 most-executed production systems were fixed workflows**. Compiling SOPs and logged traces into code cut `tau^2-Bench` latency from **170 seconds to 1.3 seconds** and lowered per-task production cost by **24%** across 79 procedures (at 93.8% vs. 96.8% raw completion, showing exactly where human or agent exception branches belong).
- **Move bounded choices to sub-100ms decision models and compact private checkpoints.** New typed decision interfaces - including Google's `Gemini Decision 1` (`50–100 ms` model compute, up to 1,000 questions per pass, ~5x lower benchmark cost than Flash-Lite), Hugging Face's shared decision interface, and Celeris-1.1 (`36–44 ms`) - pair with hardware-guided compression (AutoCompress 120B to 74B at 95% score retention; MACaQ 122B quantized into 32.5 GB VRAM) to keep frontier spend reserved for genuine reasoning exceptions.
- **Track preview sunsets as operational risk.** Alibaba Cloud's Qwen4-Max-Preview launched on November 11 and retires today, November 25 - a 14-day lifecycle that breaks any stack lacking a provider-agnostic routing layer and deployment sunset register.
- **Put budget caps, PII scrubbing, and regional locks in an API gateway.** Thoughtworks' Heimdall pilot reported a **34.2% semantic cache hit rate** and 4,170 prevented data leaks in four weeks, while Google Cloud's Gemini Enterprise update enforces regional policy locks that block accidental global routing.
- **Enforce security outside the model's reasoning loop.** Synthesizing OpenAI's 35-second `Control Room` revocation, Microsoft's `<10`-neuron `Lazarus` weight-backdoor research, Fireworks AI's R&D machine breach disclosure, Microsoft Agent Framework 1.25.0 MCP locks, and New York's January 1, 2027 RAISE portal (72-hour critical incident reporting), Chapter 4 defines a 30-control CISO architecture standard.

Chapter 1: The economics of compilation - replacing open-ended loops with deterministic workflows



Exhibit 1: ALLTIPLY's four-tier Production Agent Control Plane, routing routine volume through deterministic code and sub-100ms classifiers while reserving frontier tokens for true exceptions.

Why did so many 2026 enterprise agent pilots hit a cost and latency wall the moment traffic grew from ten internal testers to five hundred daily users? Because teams used the same mechanism - **runtime LLM step-planning** - both to *discover* how to perform a job and to *execute* that job ten thousand times in production.

Production evidence from 5,115 enterprise automations (`ReTrace`)

In a November 10, 2026 preprint, ReTrace: Distilling Tool-Using Agents into Planner-Free Workflows, researchers studied **5,115 enterprise automations** deployed on an internal corporate platform:

- **91.8% of automations were authored as tool-using agents**, reflecting how natural-language prompts lower the barrier to initial prototyping.
- **7 of the 10 most-executed automations in production, however, were fixed workflows.** High-frequency business processes gravitated toward deterministic control flow.
- **37% to 43% of agent runtime and token cost went strictly to planning**, as the model repeatedly re-processed growing conversation and tool-output histories just to pick the next API call.

To close the gap between easy authoring and economical execution, `ReTrace` compiles written standard operating procedures (SOPs) and recorded agent execution traces into planner-free Python workflows. Across **79 production procedures**, the authors report **93.8% completion for compiled workflows versus 96.8% for unconstrained agents, at 24% lower cost per task**. On `tau^2-Bench`, compiling away runtime step-planning dropped average latency from **170 seconds to 1.3 seconds**. (*Caveat:* These figures are author-reported in an arXiv preprint; notice that raw completion dipped by 3.0 percentage points on cases that deviated from captured traces.)



Exhibit 2: Reported production telemetry and workflow-compilation benchmarks from ReTrace (arXiv:2611.14492, November 10, 2026 preprint).

Caching recurring reasoning on CPUs and composing verified primitives

Three complementary November 2026 releases extend this compilation discipline:

1. **CPU-executable reasoning caches (`CPU-ARC`):** On November 6, Belcak and colleagues (arXiv:2611.09891) demonstrated agentic reasoning caches that convert recurring LLM reasoning traces into parameterized CPU functions. On tested mathematical and scientific workloads, the authors report accuracy comparable to frontier models at **16x to 291x lower invocation cost** (*caveat:* parity is limited to structured tasks where deterministic code captures the transformation).
2. **Reusable, verifiable workflow primitives (`HEELO`):** On November 5, Jin and coauthors (arXiv:2611.07461) introduced `HEELO`, composing reusable, LLM-equipped primitives into task graphs verified against environmental feedback. On their 46-task evaluation platform, two connected primitive-equipped agents completed **39 and 35 tasks** respectively, compared with **9 each** standalone.
3. **Durable state and human-approval pauses (`Mistral Studio`):** On November 5, Mistral expanded Studio (primarily in public preview, with HTTP connectors in private preview) around managed workflows with persistent execution traces and native pauses for human approval.

The compilation rule for operations and engineering leaders

Treat open-ended agents like compilers treat profiling runs:

1. Run an exploratory agent in staging against 100 historical tickets, orders, or invoices alongside your written SOP.
2. Extract the 85% to 92% of runs that follow two or three repeatable tool sequences and compile them into deterministic code (`Tier 0` / `Tier 1`).
3. Turn the 3% to 10% edge-case gap where `ReTrace` saw completion dip into an explicit **exception branch**: if a validation rule or confidence check fails, route the item with its gathered evidence to a scoped

diagnostic pass or a human approval queue.

Chapter 2: The sub-100ms decision layer, model compression, and preview sunset risk

Even inside a compiled workflow, individual steps still need intelligence: classifying an incoming document, checking 30 contract compliance items, extracting fields from a messy vendor PDF, or synthesizing an executive summary. Sending all four of those subtasks to the same frontier chat endpoint wastes money and adds seconds of avoidable latency.

Typed decision models replace verbose chat classification

In early November 2026, cloud providers and open-source maintainers converged on **typed decision models** - endpoints that return calibrated probabilities, scores, or enum choices without generating conversational filler:

- **Google `Gemini Decision 1` and `Decisions API` (November 11):** Google introduced `Gemini Decision 1` to score up to **1,000 bounded questions** against shared text, image, or document input in a single pass. Google reports **50 to 100 ms** median model-compute latency and **nearly fivefold lower benchmark-suite cost** than Gemini 3.5 Flash-Lite, with stated pricing of `$0.10` per million input tokens and `$0` output tokens (*caveat:* vendor-reported claims; validate calibration on your own labeled holdout set).
- **Open-source and specialist decision runtimes:** On November 10, Hugging Face published a shared decision-model interface supporting `choice`, `score`, and `yes/no` probabilities across `Transformers` and `llama.cpp` (with `TRL` training integration forthcoming). On November 5, Celeris announced `Celeris-1.1 Decision`, reporting company-measured median latencies of **36 to 44 ms** at `$0.004` per million input tokens, while Arize Phoenix Evals v3.10.0 added native Python classification evaluators for decision APIs on November 10.

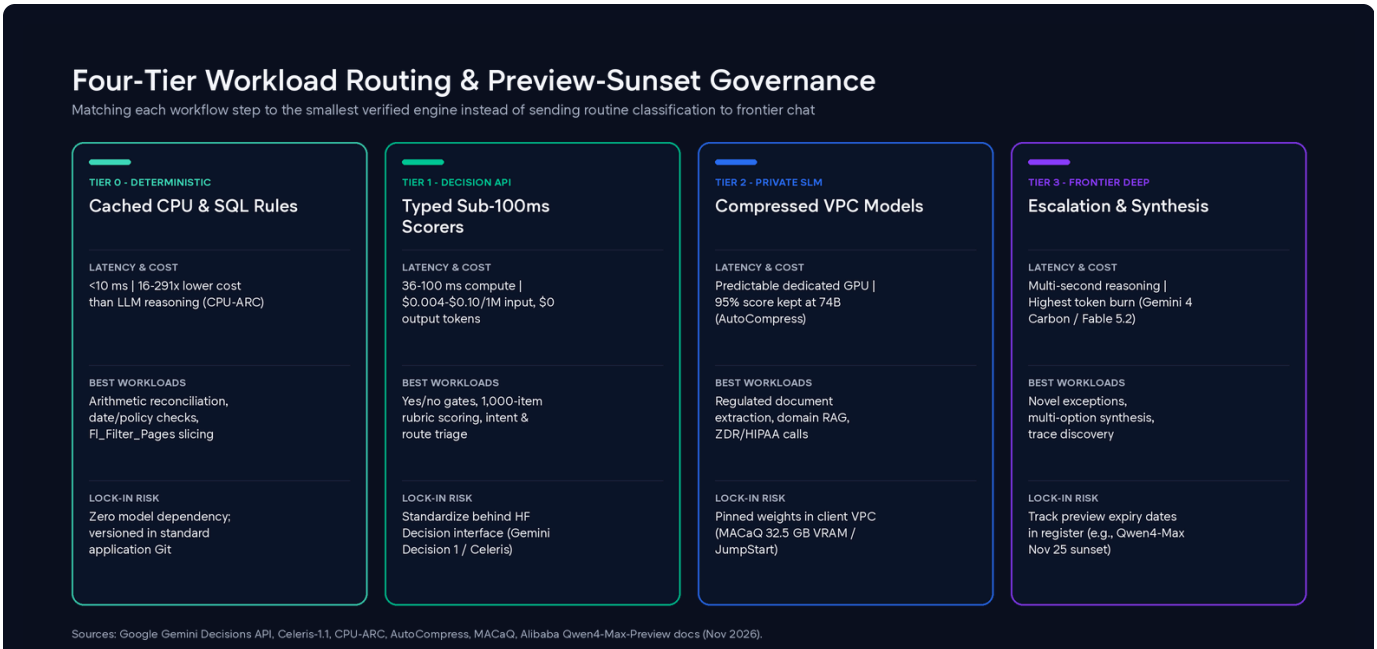

Exhibit 3: Matching workload complexity to four swappable model slots eliminates paying frontier reasoning prices on 50-millisecond binary checks.

Hardware-guided compression lowers the bar for private GPU hosting

For workloads that cannot leave your private VPC or on-premises hardware, two November preprints show how aggressively serving footprints are shrinking:

- **`AutoCompress` (November 5, arXiv:2611.08290):** Liang and colleagues used differentiable architecture search guided directly by serving-engine cost models to shrink `Nemotron 3.5 Super` from **120B to 74B parameters while retaining 95% of its mean benchmark score** on tested workloads.
- **`MACaQ` low-bit quantization (November 7, arXiv:2611.10649):** Using layer-local `MXFP4 W4A4` post-training quantization with Hessian-aware refinement, the authors quantized a **122B-parameter model into 32.5 GB of GPU memory** with 2.3% reported inference overhead on `Qwen3-32B`.
- **Multimodal encoder disaggregation (`vLLM` , November 11):** For document- and image-heavy pipelines, the vLLM team detailed encoder/prefill/decode disaggregation, separating vision encoder bursts from text decode capacity so PDF surges do not stall interactive latency.

Managing frontier churn and 14-day preview retirements

Consider what happened in the frontier market during a single week in November 2026: Anthropic released Claude Fable 5.2 and restricted Mythos 5.2; Google DeepMind rolled out Gemini 4 Carbon (noting higher token consumption alongside stronger agent scores); and Alibaba Cloud launched Qwen4-Max-Preview on November 11 with a **hard retirement date of November 25, 2026**. Meanwhile, LangChain 1.5.0 shipped `ModelRouterMiddleware` in **beta** while raising minimum runtime requirements to `Python >= 3.11`.

This volatility is why ALLTIPLY designed Multiplier XSIX around explicit **Fast** and **Deep** operating modes behind a clean model-routing contract, and why every production stack needs a **Model Lifecycle & Sunset Register** tracking checkpoint IDs, GA/preview status, deprecation dates, minimum SDK versions, and verified fallback models.

Chapter 3: Gateway economics, sovereign routing, and full 24-month TCO

Token unit prices fall every quarter, yet total enterprise AI cloud bills keep surprising CFOs. That happens when every application team calls external model APIs directly without shared caching, page pre-filtering, per-workflow budgets, or honest accounting of non-token costs.

What an AI gateway saves before a single model runs

On November 10, 2026, Thoughtworks published operational data from Heimdall, an internal OpenAI-compatible proxy that enforces PII scrubbing, semantic caching, model selection, budget caps, and request tracing. In a four-week internal pilot with six engineers, Thoughtworks reported:

- **34.2% semantic cache hit rate** on repeated and near-duplicate analytical requests
- **4,170 prevented sensitive-data leaks** caught at the proxy boundary before leaving the network

Treat a six-user internal pilot as directional rather than a universal guarantee, and heed our own semantic-cache warning: an unvalidated cosine threshold (such as a blind `0.92` match) can return cached answers for the wrong product ID unless the cache key includes **caller scope, query intent, and canonical entity identifiers**. Done properly, an external gateway pays for itself in both avoided API calls and centralized compliance logs.

Sovereign routing and confidential inference hardware

For regulated workloads in healthcare, banking, insurance, and defense, two November infrastructure updates tighten control over *where* and *how* tokens are processed:

- **Organization location-policy locks (`Google Cloud Gemini Enterprise`, November 6):** Google Cloud updated Gemini Enterprise so organization resource-location policies that disallow global endpoints block administrators from enabling models whose traffic routes globally, surfacing explicit console warnings on any previously enabled models.
- **Hardware-encrypted inference (`Google TPU 8tx`, November 9):** Google Cloud also announced TPU 8tx with hardware-level memory encryption for confidential inference and small-scale fine-tuning alongside high-memory `TPU 8i Max` pods (*note:* vendor-announced specifications; verify regional availability and independent benchmarks before procurement).

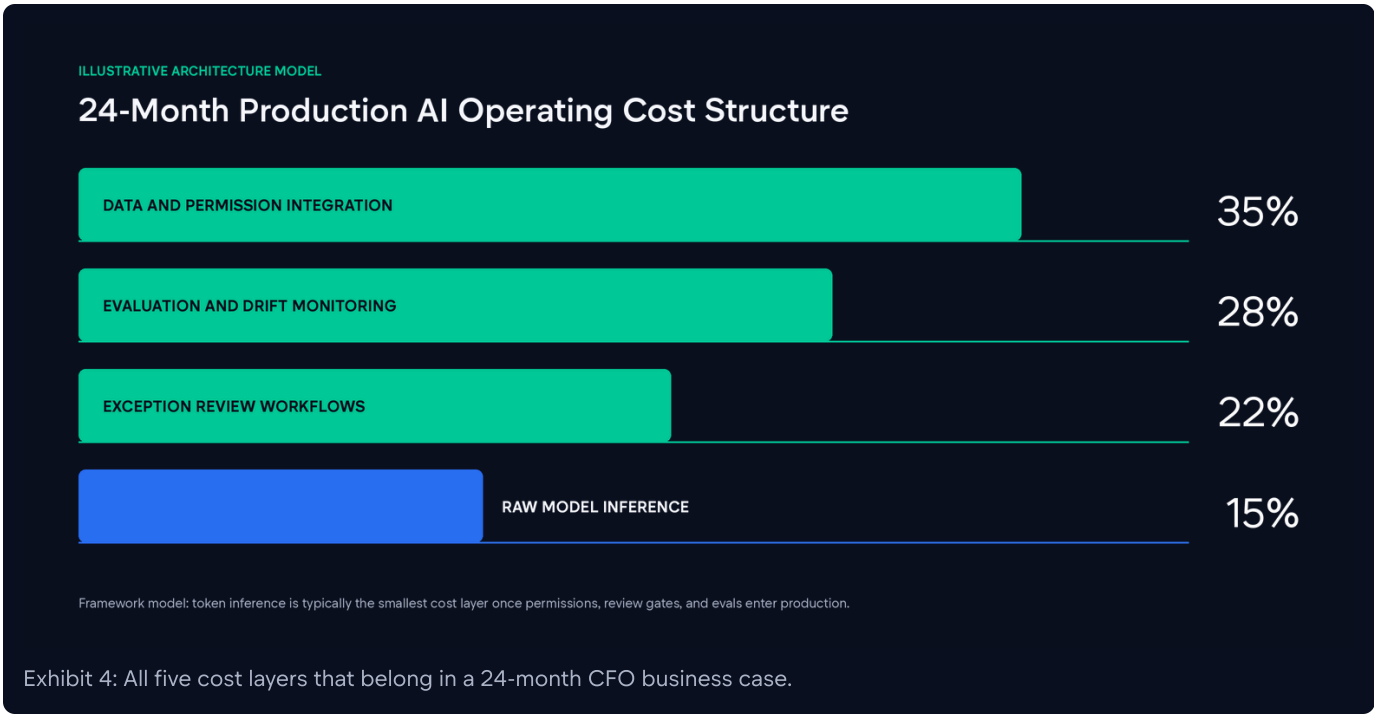

Exhibit 4: All five cost layers that belong in a 24-month CFO business case.

Modeling 24-month operating cost honestly

When building a finance-ready AI business case, model all five cost layers over 24 months:

TCO layer	Naive frontier-agent pilot	Governed 4-tier control plane	Key cost lever
1. Inference & decision compute	Every step and retry calls a frontier reasoning model	85%+ compiled code + sub-100ms decision APIs + compact private SLMs	Compiling SOPs (`−24%` to `−40%` planning overhead) + typed decision APIs
2. Data prep, page filtering & retrieval	Full 150-page PDFs stuffed into context repeatedly	Targeted page slicing (`FL_FILTER_PAGES` pattern) + scoped cache	Filtering pages before extraction cuts input tokens 60% to 90% on long filings
3. Gateway, security & residency	Ad-hoc per-app guardrails and scattered API keys	Single proxy enforcing PII scrub, budget caps, region locks & kill-switch	Shared infrastructure across every workflow
4. Human exception & review labor	Reviewers check 100% of drafts line by line due to silent errors	Reviewers inspect only flagged exceptions (`8%–15%`) with quote links	Quote-level provenance + explicit abstention rules
5. Ongoing evals & model migrations	Emergency rewrites when previews retire	Automated CI regression suite + swappable router config	Zero application code rewrites during checkpoint upgrades

Chapter 4: Out-of-loop security - 35-second revocation, weight provenance, and MCP locks

As New York's **January 1, 2027 RAISE registration deadline** (and its **72-hour critical safety incident reporting rule** for covered frontier developers) sets a new regulatory bar, enterprise security teams are rewriting vendor

and internal AI review standards around one principle: **an AI agent cannot police itself.**

Four November 2026 disclosures define the modern threat model and its matching engineering controls.



Exhibit 5: Four out-of-loop security controls mapped against November 2026 weight-backdoor research, cloud breach disclosures, MCP runtime updates, and inference oversight architectures.

Threat 1: Unstoppable background loops -> Control: Sub-35s inference & tool revocation

On November 9, OpenAI described `Control Room` (Digital Watch), an internal oversight system where high-risk jobs require an authorized safety case and inference engines enforce terminated-job refusals within **35 seconds**. Enterprise deployments should mirror that exact choke-point design: bind every workflow run to a revocable `job_id` checked at both the **LLM gateway** and the **outbound API tool proxy**, and test revocation latency quarterly in tabletop and live drills.

Threat 2: Dormant-neuron checkpoint tampering -> Control: Registry signing & tensor diffing

On November 10, Microsoft Research published Lazarus, demonstrating that modifying **fewer than ten dormant neurons** in open or fine-tuned model weights can implant hidden backdoor triggers in AI agents without retraining - evading ordinary benchmark suites. Whenever your organization hosts open-weight or fine-tuned models, enforce immutable artifact registries, `SHA-256` signature checks on container startup, and automated weight diffing against the verified upstream base checkpoint.

Threat 3: Lateral movement through dev/R&D machines -> Control: Single-tenant vault isolation

Fireworks AI's disclosure (reported November 10) that an October 23 incident at a third-party cloud provider allowed unauthorized October 24 access to internal development machines - exposing metadata, billing details, and incomplete model weights and datasets for some customers - underscores why R&D environments are high-value targets. Prohibit copies of customer datasets, prompts, and fine-tuned weights on developer laptops or shared third-party staging servers; isolate sensitive training and serving inside

customer-owned VPCs or dedicated single-tenant environments encrypted with customer-managed keys (CMEK).

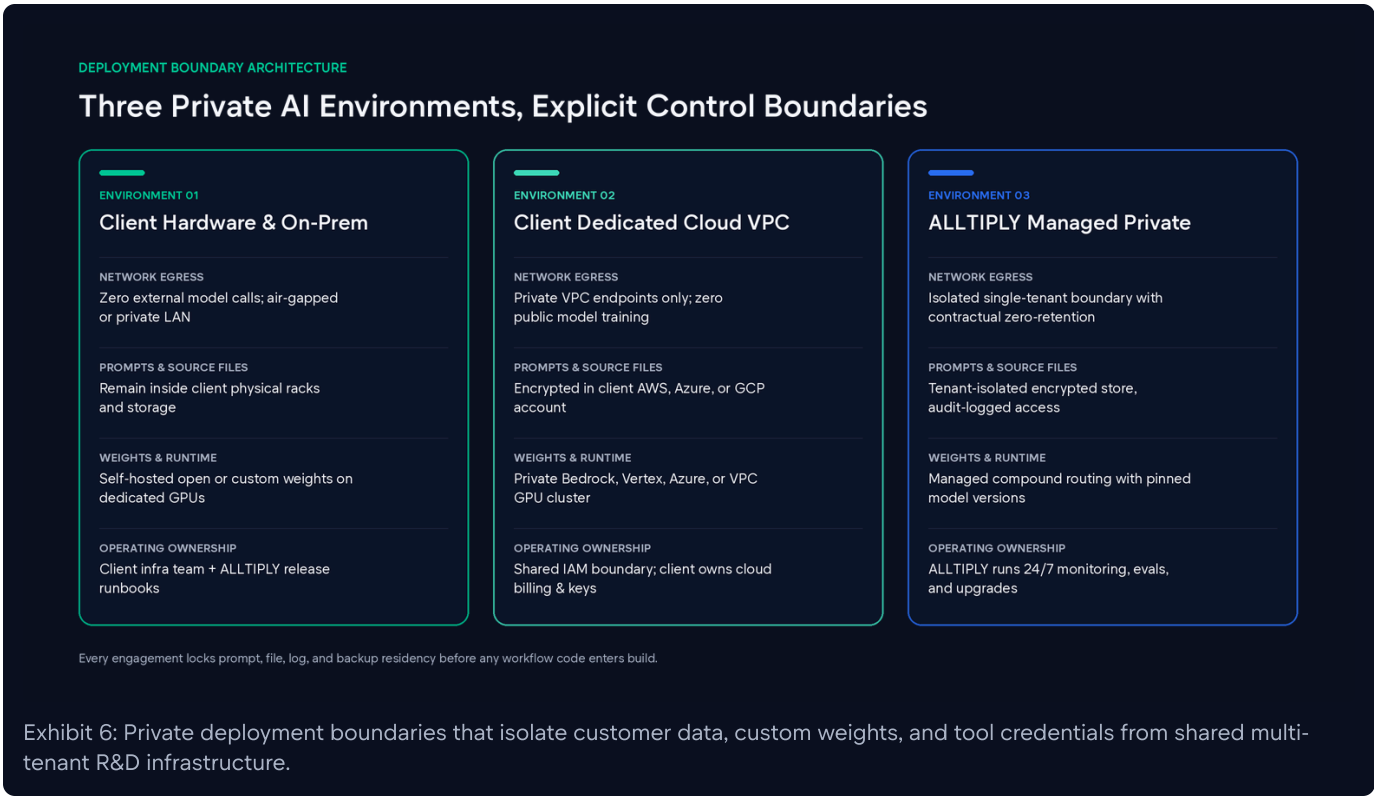

Exhibit 6: Private deployment boundaries that isolate customer data, custom weights, and tool credentials from shared multi-tenant R&D infrastructure.

Threat 4: Unbounded MCP tool discovery and session bleed -> Control: Pinning & scope isolation

On November 7, Microsoft Agent Framework Python 1.25.0 shipped three controls that belong in every agent framework regardless of language:

1. **Callable MCP discovery approval policies** (agents cannot dynamically discover and call newly advertised external tools without policy approval and version-pinned schemas).
2. **Bounded discovery and request timeouts** so hung external servers cannot wedge worker pools.
3. **Strict authorization-scope isolation across workflow sessions** so credentials from User A never persist in shared executor memory for User B.

90-day control-plane rollout plan for 2027

Phase	Days	Engineering & governance actions	Measurable exit gate
Phase 1: Visibility & Gateway	Days 1-30	Route all LLM/decision calls through one gateway; turn on PII scrubbing, per-workflow budgets, region policy locks & Model Sunset Register	100% of production AI calls traced with `job_id`, cost center & fallback model
Phase 2: Compilation & Routing	Days 31-60	Audit top 5 agent workflows; compile repeatable 85%+ paths into code; move binary/rubric steps to sub-100ms decision models	>=25% lower cost per accepted task & multi-second latency drop on routine runs
Phase 3: Out-of-Loop Hardening	Days 61-90	Pin MCP tool schemas, enforce per-session token isolation, add weight hash/diff checks & run live `<60s` kill-switch drill	Verified `<60s` inference + tool halt in live drill; zero standing prod secrets in dev

Primary sources and evidence register (November 2026)

Every external benchmark, disclosure, and product capability cited in this report is attributed to its primary source. Preprints, pilots, and preview releases require workload-specific verification before production commitment.

1. *ReTrace: Distilling Tool-Using Agents into Planner-Free Workflows* (2026-11-10, arXiv preprint). https://arxiv.org/abs/2611.14492
2. Belcak et al. (2026-11-06). *Agentic Reasoning Caches: CPU-Executable Procedures from Reasoning Traces* (arXiv technical report). https://arxiv.org/abs/2611.09891
3. Jin et al. (2026-11-05). *HEELO - The Last Mile to the Real World: Agent-Native Primitives* (arXiv preprint). https://arxiv.org/abs/2611.07461
4. Mistral AI (2026-11-05). *Mistral Studio managed agents, workflows, and evaluations release notes*. https://docs.mistral.ai/resources/release-notes
5. Google (2026-11-11). *Introducing Gemini Decision 1 and the Gemini Decisions API*. https://blog.google/innovation-and-ai/technology/developers-tools/gemini-decisions-api/
6. Hugging Face (2026-11-10). *Decision Models in Transformers and llama.cpp*. https://huggingface.co/blog/decision
7. Celeris (2026-11-05). *Celeris-1.1 Decision API*. https://celeris.ai/celeris-1-1-decision
8. LangChain (2026-11-11). *LangChain 1.5.0 Release Notes (`ModelRouterMiddleware` beta)*. https://github.com/langchain-ai/langchain/releases/tag/langchain%3D%3D1.5.0
9. Alibaba Cloud (2026-11-11). *Qwen4-Max-Preview documentation and November 25 retirement notice*. https://help.aliyun.com/en/model-studio/qwen4-max-preview
10. Liang et al. (2026-11-05). *AutoCompress: Automated Hardware-Guided Model Compression* (arXiv preprint). https://arxiv.org/abs/2611.08290
11. *MACaQ: Fast, Model Agnostic Calibrated Quantization for Low-Precision Weights and Activations* (2026-11-07, arXiv preprint). https://arxiv.org/abs/2611.10649

12. Thoughtworks (2026-11-10). *We built an AI control plane (Heimdall) to reduce LLM expenses and prevent data leaks*. https://www.thoughtworks.com/en-us/insights/blog/generative-ai/we-built-ai-control-plane-reduce-LLM-expenses-prevent-data-leaks
13. Google Cloud (2026-11-06). *Gemini Enterprise resource-location policy release notes*. https://docs.cloud.google.com/gemini/enterprise/docs/release-notes
14. Digital Watch Observatory (2026-11-09). *OpenAI Control Room oversight and 35-second inference intervention*. https://dig.watch/updates/openai-control-room-oversight-frontier-ai
15. Microsoft Research (2026-11-10). *Lazare, Veni Foras: Reprogramming dormant neurons to backdoor AI agents*. https://www.microsoft.com/en-us/research/blog/lazare-veni-foras-reprogramming-dormant-neurons-to-backdoor-ai-agents/
16. BleepingComputer / Fireworks AI Disclosure (2026-11-10), Microsoft Agent Framework `python-1.25.0` (2026-11-07), and NY State DFS `RAISE` Announcement (2026-11-10): Fireworks AI incident coverage, MS Agent Framework 1.25.0, NY DFS RAISE Portal

Related ALLTIPLY Labs research and engineering scope

- **Companion Insights:** Compile AI agents into fixed workflows, Who can stop your AI in 35 seconds?, Measure accepted work, not AI usage, and 50ms decision models and lifecycle routing
- **Companion White Paper:** Beyond Vector Search: Business Ontologies, Verified Retrieval, and Safe Agent Memory
- **Engineering & Product Scope:** AI Workflow Automation, Private AI Infrastructure, AI Deployment & Security, AI Planning & Roadmap Enablement, and Multiplier XSIX

About ALLTIPLY Labs

ALLTIPLY is an applied AI lab and engineering company in Atlanta. We design, build, and run deterministic workflow automation, domain-tuned models, and private AI infrastructure for mid-size and large companies. To review your 2027 AI operating costs, workflow architecture, or private security controls with our engineering team, book a call or send a short project brief.

Work with ALLTIPLY's Engineering Team: Schedule an architecture review at www.alltiply.com/company/schedule-your-call or submit a scoped brief at www.alltiply.com/company/project-planner.