INDIUM

Harnessing the Power of LLMs

Without Losing on Privacy and Security

A practical, three-tier framework for enterprise LLM deployment Risk landscape, mitigation strategies, and capability–cost–control tradeoffs

Whitepaper

Written by Ashish Kumar
Chief AI Architect

Executive Summary

Large Language Models have moved from pilots to mission-critical production workloads, shifting the key question from "does this work?" to "can we run this safely on regulated data without ceding control?" The answer is yes, but only with a deliberately chosen deployment architecture. Three patterns dominate, each trading off capability, cost, control, and operational burden: Tier 1, public LLM APIs with guardrails, offers the fastest path to capability through contractual controls and a policy/redaction layer, ideal for non-sensitive workloads and prototyping; Tier 2, cloud-managed models (Bedrock, Azure OpenAI, Vertex AI) inside a private VPC via PrivateLink/VNet, is the default choice for regulated industries; and Tier 3, self-hosted open-weight models (DeepSeek V4, Llama 4, Qwen 3.5, GLM-5) on owned or leased GPUs, delivers the highest control and lowest per-token cost at scale but carries substantial fixed costs and operational overhead. In practice, the recommended approach is a hybrid model that routes workloads by data sensitivity—public APIs for non-sensitive bulk traffic, VPC-resident managed models for the regulated middle, and a small self-hosted footprint for the most sensitive or highest-volume work—and this paper provides the benchmarks,

1. The Privacy and Security Challenges of LLM APIs

LLM APIs introduce a new trust boundary in the enterprise stack. Unlike traditional APIs, where input is structured and validated, LLM endpoints accept arbitrary natural-language text that is often packaged with internal context – system prompts, retrieved documents, user PII, tool outputs. That combination of open-ended input and contextual access creates a wider attack surface than most legacy integrations, and one that organizations are still learning to govern.

1.1 The Risk Landscape

The OWASP Top 10 for LLM Applications (2026 update) and the MITRE ATLAS knowledge base catalog over 130 distinct attack techniques against LLM systems. For most enterprises, six risk classes account for the majority of real-world exposure:

Risk class	What it looks like	Impact
Prompt injection (direct & indirect)	User input or embedded content overrides system instructions; persistent injection across connected agents.	Jan 2025: embedded instructions in a public document exfiltrated proprietary data via a major enterprise RAG system.
Sensitive information disclosure	Outputs leak PII, credentials, prior context, or fine-tuning data – often via RAG with weak scoping.	1 in 12 employee prompts to public LLMs contain confidential information; RAG can surface restricted records to unauthorized users.
Training-data leakage & memorization	Weights regurgitate verbatim secrets from training corporate	Truffle Security: ~12,000 live API keys and passwords in a single training dataset; 63% on multiple pages.
Supply-chain compromise	Third-party model, library, plug-in, or vector-DB component compromised; sub-processor chain spans jurisdictions.	Foundation-model providers depend on layered sub-processors often outside enterprise contracts.
Insecure output & excessive agency	Generated code, SQL, shell, or API calls executed without verification; agents with broad permissions.	~45% of AI-generated code contains security flaws (Veracode 2025); over-scoped agents can move money or modify identity.
Cross-border transfer & residency	Prompts and embeddings routed across jurisdictions; CLOUD Act exposure on EU data held by US providers.	Cumulative GDPR fines €5.88B by 2026; EDPB Opinion 28/2024 confirms LLM training is in-scope; EU AI Act overlays.

Sources: OWASP LLM Top 10 (2026), MITRE ATLAS, EDPB Opinion 28/2024, Lasso Security, Veracode 2025, Truffle Security 2025.

1.2 The Shadow-AI Multiplier

Public consumer chatbots (ChatGPT free tier, personal Claude accounts, browser plug-ins) bypass formal access logs, RBAC, data-masking, and audit trails. Studies estimate that 1 in 12 employee prompts to public LLMs contains confidential information, and recent reporting (December 2025) documented a Chrome extension intercepting millions of users' AI chats. This is a workforce-side problem that no contract can fix. The first line of defense is replacing consumer tools with sanctioned enterprise tiers and giving employees a fast, capable, approved alternative so they do not work around the policy

1.3 What Happens at the API Boundary

Even on enterprise-tier APIs from reputable providers, several layers of exposure remain. Understanding them is the basis for the mitigations in Section 2.

- **Retention By Default**

OpenAI retains API data for 30 days for abuse monitoring; Anthropic for 7 days (as o Sept 2025); Bedrock does not log prompts or completions by default. None of these are training data, but retention windows matter for breach exposure and subpoena risk.

- **Sub-Processors**

Foundation-model providers depend on cloud infrastructure from other vendors. Documenting and contracting across that chain is the customer's problem.

- **Inference-Time Leakage**

Outputs can echo prior context (RAG), fine-tuning data, or training data. Standard output filters frequently fail to detect this, while behavioral anomaly detection and token-level redaction are emerging defenses.

- **Logs, Transcripts, and Observability Tools**

LangSmith, Langfuse, and similar telemetry stacks can ship full prompts and completions to third-party SaaS – a shadow data store that is easy to miss in a vendor inventory.

2. A Three-Tier Mitigation Framework

There is no single "secure LLM deployment." There is a spectrum of architectures, each with different blast radius, cost structure, and operational demands. The framework below maps the three patterns most enterprises converge on, and the rest of this section drills into each one.

Dimension	Tier 1: Public API + guardrails	Tier 2: Managed LLM in private VPC	Tier 3: Self-hosted open-source	Driver of the difference
Accuracy ceiling	Highest – frontier proprietary	Same as Tier 1 for proprietary; ~5–10% lower for cost-optimized	Strong open models close most gaps; ~8 months behind frontier on agentic	Frontier proprietary still leads on agentic, reasoning, SWE-bench Pro
Throughput	Provider rate limits; cross-region adds latency	Predictable with provisioned throughput; latency-optimized options	Bounded by GPU count; ~793 tok/s on H100 for 70B FP8	Self-host = own the queue; managed = rent SLAs
Privacy & data control	Contractual (ZDR, BAA, DPA); data leaves perimeter	Data stays in your cloud account & region; CMK keys	Data never leaves network; air-gap possible	Trust shifts: contract → infrastructure → physical
Security posture	Provider-managed; you control prompts and authentication /authorization	Inherits cloud IAM, KMS, audit; you set model policies	You own patching, network, hardening, GPU drivers	Control follows responsibility
CapEx	$0	$0 – pay-as-you-go or commit	$80K–$500K+ on-prem (8× H100 ≈ $200–240K)	Self-host is front-loaded
OpEx (per 1M tokens)	$0.20–$75 by model tier	Same per-token; +PT hourly when reserved	~$0.10–$1 marginal once amortized; power & cooling	Variable vs fixed cost
Maintenance overhead	Low – prompts + evals	Medium – IAM, networking, RAG, version upgrades	High – GPU ops, serving tuning, observability	Cost shifts to headcount
Time to first production	Days	Weeks	Months	Complexity grows with control

Sources: OWASP LLM Top 10 (2026), MITRE ATLAS, EDPB Opinion 28/2024, Lasso Security, Veracode 2025, Truffle Security 2025.

Tier 1 is the right starting point for any workload that does not handle regulated data, identifiable customer information, or material non-public information. The economics are unbeatable for low-to-medium volumes (well under 10M tokens/day on a single model), and the capability ceiling is the highest of any tier. The risk-reduction work happens around the API, not inside it.

Guidelines for Safer API Usage

- **Sign the Right Paperwork**

  Use enterprise-tier accounts only. Execute a Zero Data Retention (ZDR) addendum where available. For PHI workloads, get a signed BAA before any data flows; verify which specific services are in scope (a BAA is not a blanket covenant).

- **Put a Stateless Trust Layer in Front of the API**

  A proxy that handles PII redaction, secret-scanning (regex + ML), prompt-injection detection, output policy enforcement, and per-user / per-tenant rate limits. This is the single highest-leverage control in Tier 1 – it works regardless of what the upstream vendor changes.

- **Separate System Prompts from User Input**

  Treat system instructions as a locked vault; never concatenate raw user text into them. Validate the context layer before generation.

- **Scope RAG Aggressively**

  Per-user/per-tenant document filters at the vector-DB layer, not just in the prompt. Encrypt embeddings; log access. Embeddings are reversible, protect them like the source documents.

- **Never Auto-Execute Model Output**

  Generated code, SQL, shell commands, or tool calls must pass policy checks. For agents, scope tool permissions tightly and require human approval for any action that touches money, identity, or external systems.

- **Log Everything in Your Own Infrastructure**

  Audit trails must live in customer-owned storage (CloudWatch, Splunk, Elastic), not the vendor's observability tool. So, retention and access controls are yours to set.

- **Red-Team Continuously**

  Map MITRE ATLAS techniques to test cases; run injection corpora and jailbreak suites on each model upgrade. Providers update models without announcement and behavior drifts.

Provider / Endpoint	Default retention	Training on customer data	ZDR available	BAA / HIPAA
Anthropic (Claude API)	7 days (Sep 2025)	No	Yes (ZDR addendum)	Yes, enterprise tier
OpenAI API	30 days, abuse monitoring	No	Yes – org or project level	Yes, with BAA
Azure OpenAI	30 days (overridable)	No	Yes, configurable	Yes (Online Services DPA)
AWS Bedrock	No prompt/ completion logging	No	N/A – no logging by default	Yes, AWS BAA
Google Vertex AI	Stateless; configurable	No	Yes	Yes
DeepSeek (China-hosted)	Up to 30 days; profiling permitted	Yes per ToS	No	No

Bedrock, Anthropic, OpenAI, Azure OpenAI, and Vertex AI support enterprise contracts that disable retention and provide BAAs. Self-hosted Tier 3 is the only path for fully air-gapped workloads.

Data & AI Governance for Tier 1

Tier 1's deepest risk is not the API itself. It is the absence of an enterprise governance wrapper around it. Without that wrapper, shadow AI proliferates, sensitive data leaks through prompts, and there is no audit trail when something goes wrong.

The table below summarizes a minimum-viable control set covering inventory, classification, contracts, identity, policy enforcement, evaluation, audit, and workforce enablement. Most of these controls carry forward into Tier 2 and Tier 3 with minimal change, and the governance layer is the durable part of any LLM platform.

Control area	What it looks like in practice	Why it matters
AI inventory & catalog	Registry of approved models, tools, agents and owners; auto-discovery via egress scanning to surface shadow AI.	You can't govern what you can't see. Shadow AI is the largest leakage source in Tier 1.
DLP at the prompt boundary	PII / PHI / secrets / IP scanners on every outbound prompt; classified data auto-blocked or redacted; per-class routing.	Closes the shadow-AI leakage gap technically, not just by policy
Contracts: ZDR + BAA + DPA + SCCs	Zero Data Retention addendum; BAA for PHI; DPA for personal data; SCCs for EU↔US transfers.	Default API terms retain data 7–30 days and span unexpected jurisdictions.
Identity & access on the gateway	OIDC/SAML SSO; per-user / per-team token budgets; tool and model allowlists by role.	Limits credential-compromise blast radius; gives Finance cost attribution.
Governance bodies	AI Risk Governance Committee (AIRGC) owns policy; AI Risk Management Office (AIRMO) runs day-to-day approvals, inventory, incidents.	Without named owners, AI risk falls between Security, Legal, and the business.
Model & tool approval workflow	Recorded approval for each use case, capturing data classes, provider terms, region, evaluation results, business owner, and approved tools published in the catalog.	Closes the gap between intent ("we have a policy") and enforcement ("only approved tools work").
Third-party AI vendor due diligence	Vendor questionnaire on training-data provenance, sub-processors, certifications (SOC 2, ISO 42001), incident SLAs, right-to-audit.	AI risk is now third-party risk – the contract is part of the architecture.
Prompt / output policy enforcement	System-prompt isolation; jailbreak detection; output classifier; refusal logging.	Most production incidents are policy bypass, not novel attacks.
Continuous evaluation & red-teaming	Versioned evaluation suite (accuracy, safety, bias, injection corpora mapped to MITRE ATLAS); runs on every model update.	Providers update models silently; behavior drift is detectable only with regression evaluations.
Tamper-evident audit logging	All prompts, completions, tool calls, and policy decisions to customer-owned WORM storage with Legal-set retention.	Required for regulator investigations, subpoena response, post-incident forensics.
AI risk register & review	Documented risk assessment per use case, refreshed periodically; AIRGC reviews material risks quarterly.	Satisfies GDPR DPIA and EU AI Act high-risk-system documentation.
Workforce enablement	Approved tools easier than the shadow alternative; role-based training; clear escalation for edge cases.	Governance fails when the approved path is harder than the unapproved one.

These controls are Tier-1-specific but most carry into Tier 2 and Tier 3 unchanged.

White-Labelled Enterprise Tools as a Managed Alternative

A pragmatic middle path between rolling your own Tier 1 platform and writing direct API integrations is to procure the foundation labs' own enterprise products. Anthropic Claude Enterprise, OpenAI ChatGPT Enterprise, Google Gemini Enterprise, Microsoft 365 Copilot, and Cursor/Cody/Windsurf Teams editions include SSO, audit logs, admin policy controls, DLP integration, zero data retention by contract, and BAA eligibility. These capabilities provide a built-in governance framework for enterprise deployments.

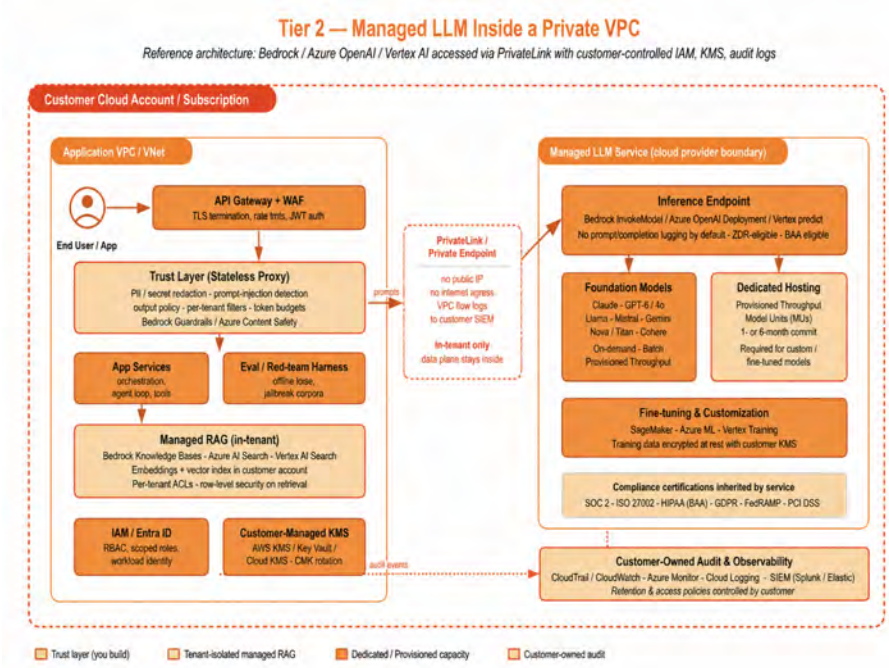
Per-seat pricing makes the cost obvious, and the procurement risk is lower than building. The tradeoff is reduced customization (a fixed UX), tenant data still living inside the vendor's perimeter (Tier 1 by definition), and less leverage when prices change. These products are an excellent fit for general-purpose knowledge work and developer productivity. They are not a substitute for the trust layer and gateway that Tier 1 still needs around bespoke applications.

When Tier 1 is The Right Answer

Non-sensitive workloads (marketing copy, public-document Q&A, internal knowledge search on already-public material), prototyping, low-volume agentic experimentation, and any case where rapid iteration on the latest model matters more than data sovereignty. The moment regulated data or material non-public information enters the picture, the conversation moves to Tier 2.

2b. Tier 2 – Cloud-Managed LLMs in a Private VPC

Tier 2 is the default enterprise choice for regulated industries. The model still lives on the cloud provider's infrastructure, but inference happens inside the customer's cloud account: VPC endpoints, IAM-controlled access, customer-managed encryption keys, customer-owned audit logs, and (optionally) dedicated capacity via provisioned throughput. Data does not traverse the public internet and models do not see traffic outside the tenant boundary.



Reference Architecture – Key Elements

- **Network**

  AWS PrivateLink (Bedrock VPC endpoints), Azure Private Link with VNet integration, or Google Private Service Connect. No public IPs on the inference path.

- **Identity**

  IAM roles (AWS), Microsoft Entra ID with RBAC (Azure), or Workload Identity Federation (GCP). Customer Lockbox on Azure for break-glass auditing.

- **Encryption & Keys**

  Customer-managed KMS keys (AWS KMS, Azure Key Vault, Cloud KMS). Storage encrypted at rest; TLS 1.3 in transit.

- **Throughput Model**

  On-demand for variable load, provisioned throughput for predictable workloads where consistent latency matter and Batch (50% discount, async) for non-realtime processing.

- **Managed RAG**

  Bedrock Knowledge Bases, Azure AI Search + Cognitive Search, Vertex AI Search – all keep retrieval inside the tenant boundary, at the cost of a per-query floor (e.g., OpenSearch Serverless ~$700/month minimum on Bedrock).

Feature	AWS Bedrock	Azure OpenAI / AI Foundry	Google Vertex AI
Model catalog	Claude, GPT-5 (since Apr 2026), Llama, Mistral, Nova/Titan,	GPT-5/4o, o-series, Llama, Mistral, Phi-4	Gemini (1M+ context), Claude, Llama, Mistral via Model Garden
Private network	PrivateLink VPC endpoints	Private Link with VNet integration	Private Service Connect; VPC Service
Default retention	No prompt/completion logging	30 days (overridable)	Stateless;
Compliance	ISO 27001, SOC 1/2/3, HIPAA, GDPR, FedRAMP Moderate, PCI DSS	HIPAA, FedRAMP High, SOC 1/2/3, GDPR, EU Data Boundary	ISO 27001/17/18, HIPAA, GDPR, PCI DSS, CSA
Dedicated capacity	Provisioned throughput (1- or 6-mo); ~15–40% off	PTUs; month / year commits	Provisioned throughput;
Custom-model hosting	PT required for fine-tuned models	Fine-tuned hosting via deployment quotas	Dedicated endpoints for custom models
Best fit	AWS-first enterprises wanting model breadth + IAM depth	Microsoft-first shops; FedRAMP High needs	Google Cloud-native, BigQuery-heavy estates

Claude on Bedrock matches Anthropic's direct API pricing; Llama on Bedrock runs 22–201% above dedicated Meta partners (Together AI, Fireworks, Groq). Cross-region inference adds ~10%.

On-Demand vs Provisioned Throughput: The Break-Even

Provisioned throughput reserves dedicated model capacity at a fixed hourly rate, eliminates throttling, and offers 15–40% off on-demand pricing in exchange for a 1- or 6-month commitment.

The break-even rule of thumb tracked across enterprise deployments is roughly $30–$40/day of consistent on-demand spend on a single model. Below that, on-demand flexibility wins and above it, reserve capacity.

Model class	On-demand ($/M tokens, in / out)	Provisioned Throughput	Break-even (daily on-demand)
Claude Opus 4.6	$15 / $75	Hourly per Model Unit; commitment discount	≈ $40/day sustained
Claude Sonnet 4.6	$15 / $75	Hourly per Model Unit	≈ $30/day sustained
Claude Haiku 4.5	$15 / $75	Hourly per Model Unit	≈ $30/day sustained
Llama 3.3 70B (Bedrock)	$0.72 / $0.72	Hourly per Model Unit	Often cheaper to use together AI/Fireworks for Llama at any volume v
Amazon Nova Pro	Competitive; Bedrock-exclusive	Hourly per Model Unit	Cost-quality tradeoff vs Claude/GPT
Batch inference (any model)	50% off on-demand	—	Use for non-realtime backlog processing

Claude pricing on Bedrock matches Anthropic's direct API. Cross-region inference adds ~10%. Knowledge Bases (managed RAG) add a per-query fee plus the OpenSearch Serverless floor (~$700/mo).

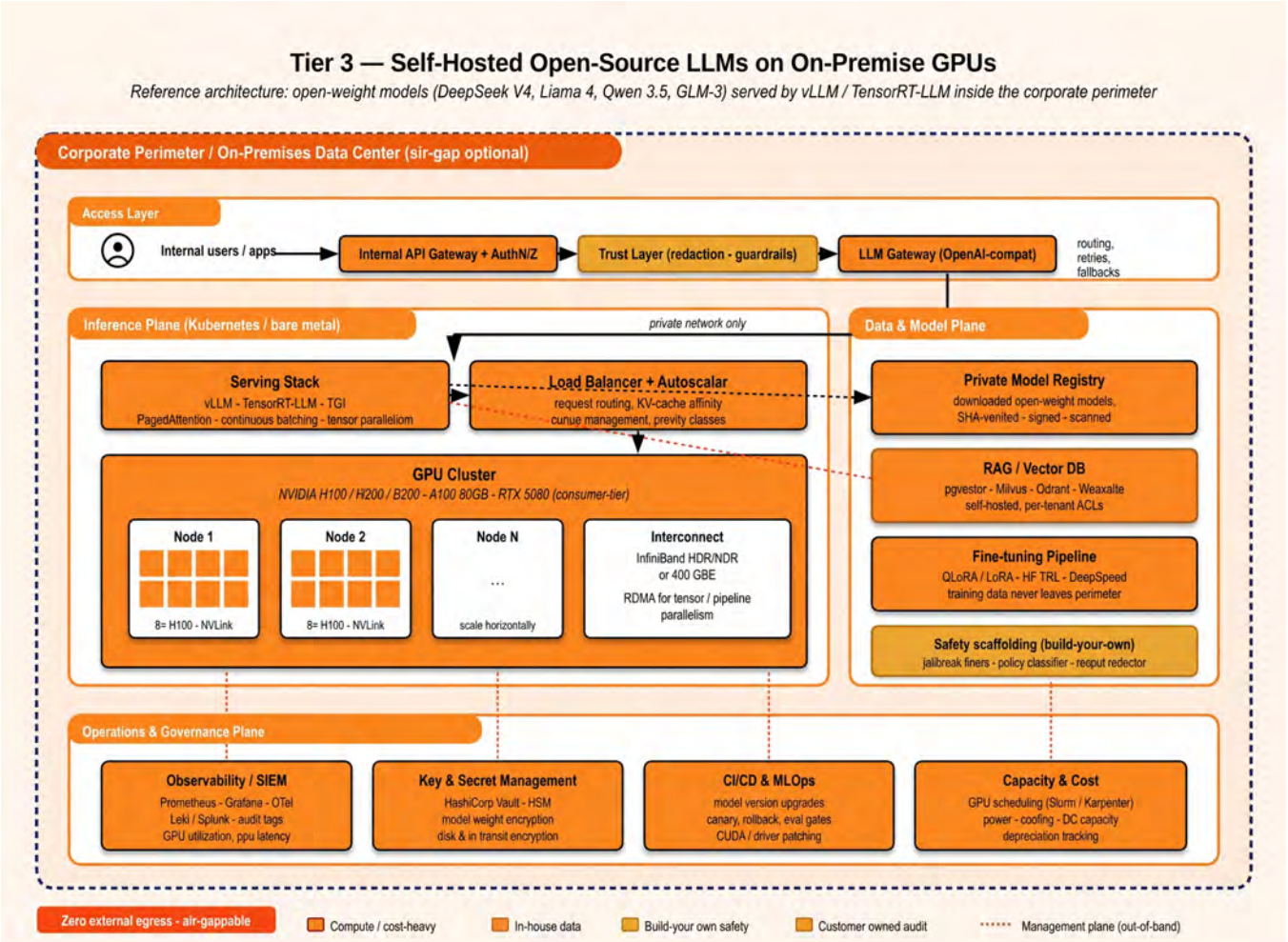
Use for non-realtime backlog processing

Almost any regulated enterprise workload: healthcare (HIPAA), financial services (PCI-DSS, SOX), public sector (FedRAMP), anything subject to GDPR data-residency where the cloud provider has in-region capacity. Tier 2 buys data-perimeter guarantees and audit-friendly architecture without the operational tax of running GPUs. The tradeoff is per-token cost. At very high volumes, the economics eventually favor Tier 3

2c. Tier 3 – Self-Hosted Open-Source LLMs on On-Premise GPUs

Tier 3 is the maximum-control option. Open-weight models (DeepSeek V4, Llama 4, Qwen 3.5, GLM-5, Mistral) are downloaded, served on customer-owned or leased GPUs, and exposed through a private endpoint.

Data never leaves the customer's network and air-gapped deployments are feasible. The 2026 landscape has changed the math here significantly. Open-weight models are now within striking distance of frontier proprietary models on most tasks and serving stacks have matured.



Serving Stack

- **vLLM**

  Production-grade, with PagedAttention and 2–4× higher throughput than standard Hugging Face Transformers implementations. Achieves approximately 793 tokens per second sustained on a single H100 for a 70B FP8 model.

- **TensorRT-LLM**
  NVIDIA-optimized, maximum throughput on Hopper/Blackwell.

- **TGI (Hugging Face)**
  Production serving with rolling batches and streaming.

- **Ollama / llama.cpp**
  Easy local serving, CPU-friendly quantized formats; not for production scale.

- **Managed RAG**
  GGUF Q4_K_M, AWQ, GPTQ, FP8: 2–4× VRAM reduction with modest quality loss. FP8 / FP4 also cuts power consumption by 30–50%.

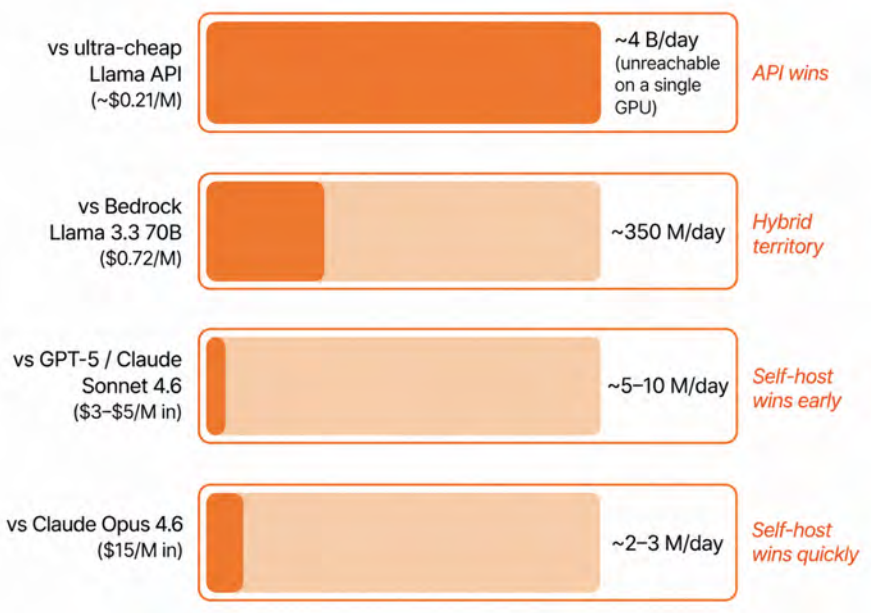
Model size class	VRAM needed (Q4)	Single-GPU option	Multi-GPU option	Indicative throughput
7B–13B (e.g., Gemma 3, Phi-4)	~6–10 GB	RTX 4090 / 5090, L4, A10G	—	150–400 tok/s on a single GPU
~32B (e.g., Qwen 2.5 32B, DeepSeek R1 distill 32B)	~20 GB	Single A100 80GB, dual RTX 4090	—	100–150 tok/s
70B (e.g., Llama 3.3 70B, Mistral Large 3)	~40 GB (Q4)	Single A100 80GB or H100	2× RTX 5090 ≈ H100 perf at ~25% cost	vLLM: ~793 tok/s on H100 FP8; ~1,500–3,000 tok/s aggregate with batching
~200B (e.g., Qwen3-235B, dense)~200B (e.g., Qwen3-235B, dense)	~120 GB	—	4× A100 80GB or 2× H100 with TP	Multi-GPU NVLink required for usable latency
400B+ MoE (DeepSeek V4, Kimi K2.6, Llama 4 Maverick	~136 GB (Q4, MoE active params)	—	8× H100 (~$200K hardware) or rented 8× H100 cluster	Cloud GPU compute ~$2–5K/month if rented full-time

Sources: BenchLM 2026, Introl hardware pricing guide, DeepSeek V4 model card, vLLM published benchmarks.

Self-Host vs. API: Where the Math Actually Tips

Self-hosting only beats an API on cost when the comparison is against an expensive frontier model, and volume is high. Against ultra-cheap API-hosted Llama (~$0.12/M tokens on DeepInfra; $0.21/M blended on OpenRouter), self-hosting is rarely competitive.

The GPU is the floor, and a single A100 at $1.79/hr on Lambda Labs costs ~$43/day, whereas the equivalent API output costs roughly $0.21–$0.71/day at the same 1M-token volume. The break-even shifts dramatically when the alternative is GPT-5 or Claude Opus at $5–$15/M input tokens.



Hidden Costs of Self-Hosting

GPU rental or capex is the floor, not the ceiling. Realistic TCO must include:

a. SRE and MLOps headcount for driver and CUDA upgrades, vLLM tuning, observability and on-call support.

b. High availability. A single GPU is a single point of failure, while production environments require redundancy.

c. Model upgrade cycles every 3–6 months as the open-weight landscape continues to evolve.

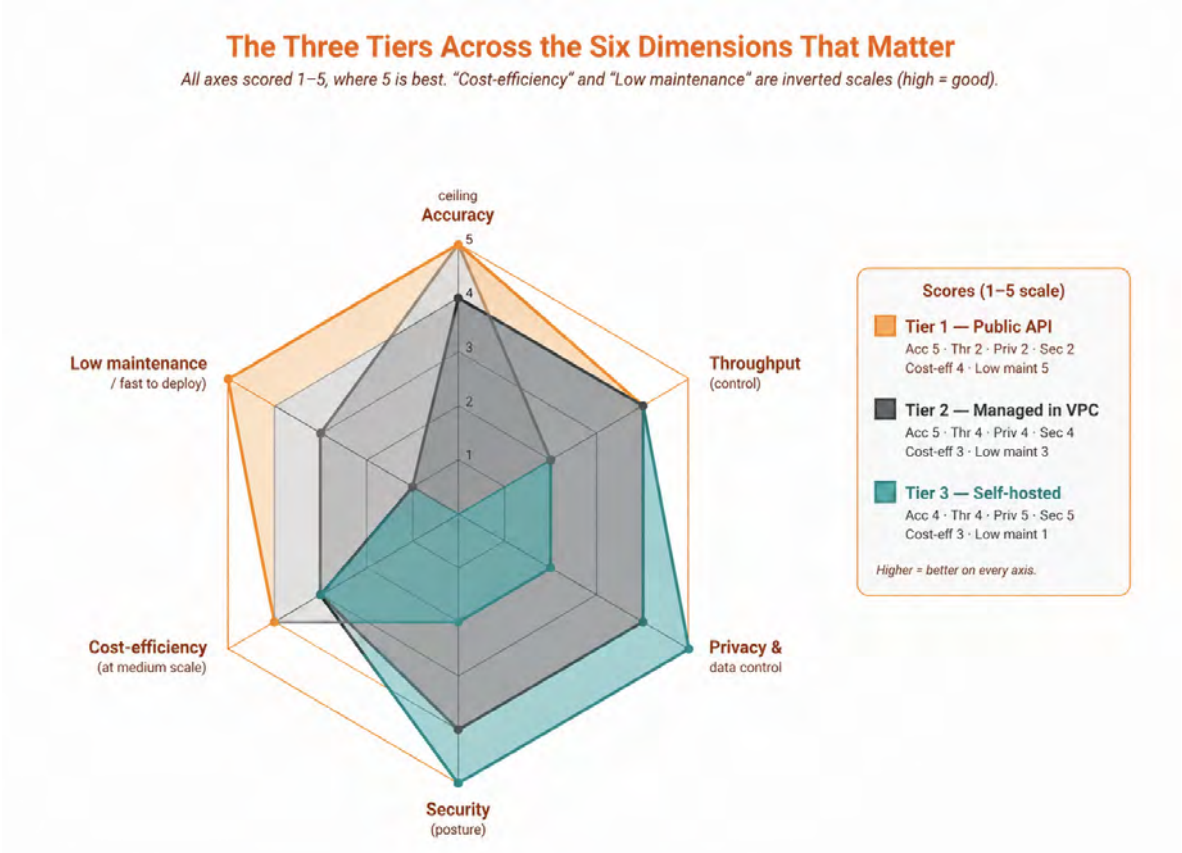
d. Safety scaffolding, including guardrails and evaluation pipelines, that managed services typically provide.

e. Power, cooling, and data center space if the deployment is truly on-premises.

Industry consensus suggests that an on-premises GPU cluster rarely outperforms a well-negotiated cloud API contract on total cost of ownership below approximately 50 million tokens per day of consistent traffic against frontier models.

3. Tradeoff Analysis – Accuracy, Throughput, Privacy, Cost

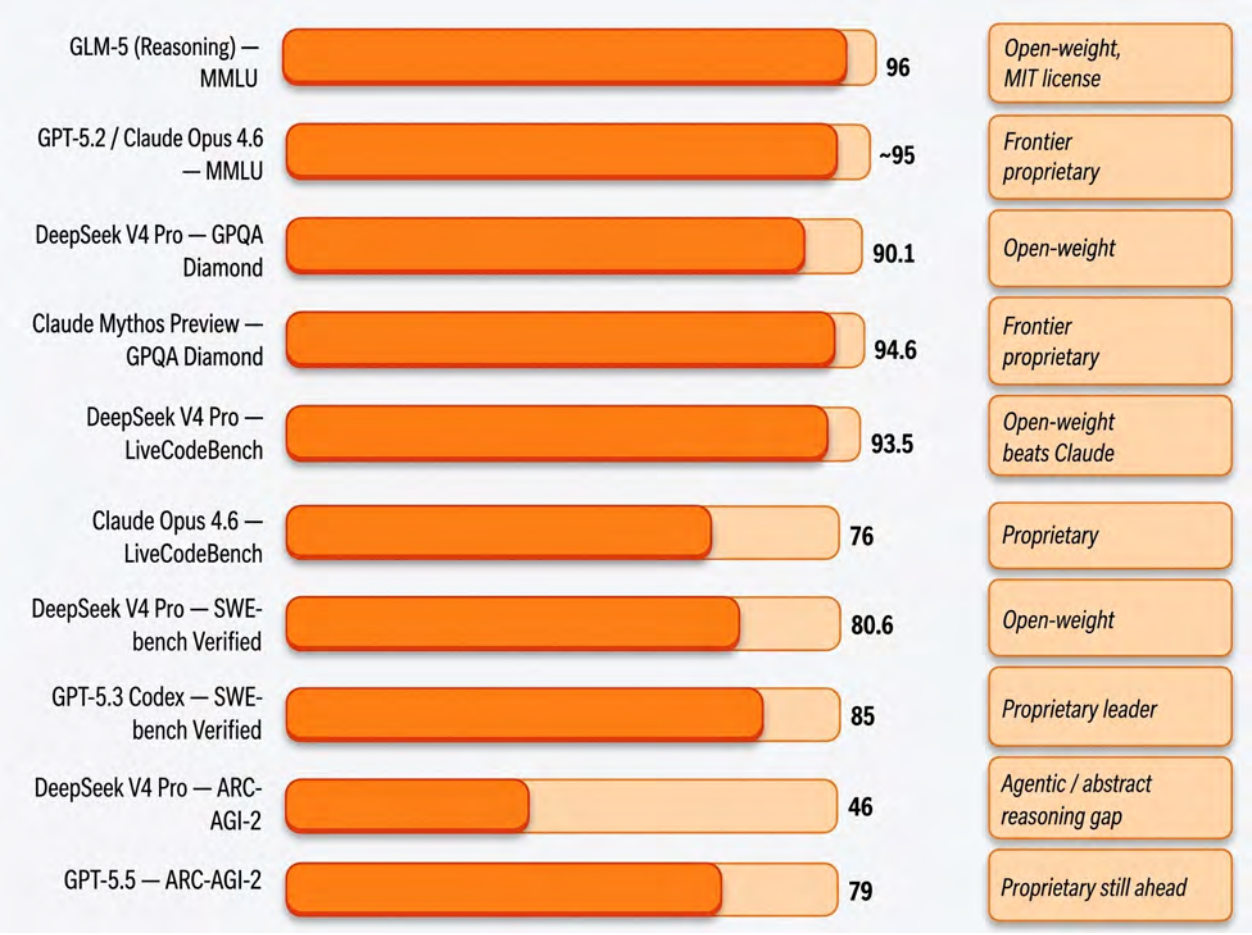
The choice between tiers is rarely about a single dimension. The radar below summarizes how the three tiers score against the six dimensions introduced in Section 2 – Accuracy ceiling, throughput control, privacy & data control, security posture, cost-efficiency, and low maintenance/time-to-deploy. The subsequent views drill down into the three dimensions with the largest spread: accuracy benchmarks, per-token cost, and total cost of ownership.



3.1 Accuracy: How Close Have Open-weight Models Gotten?

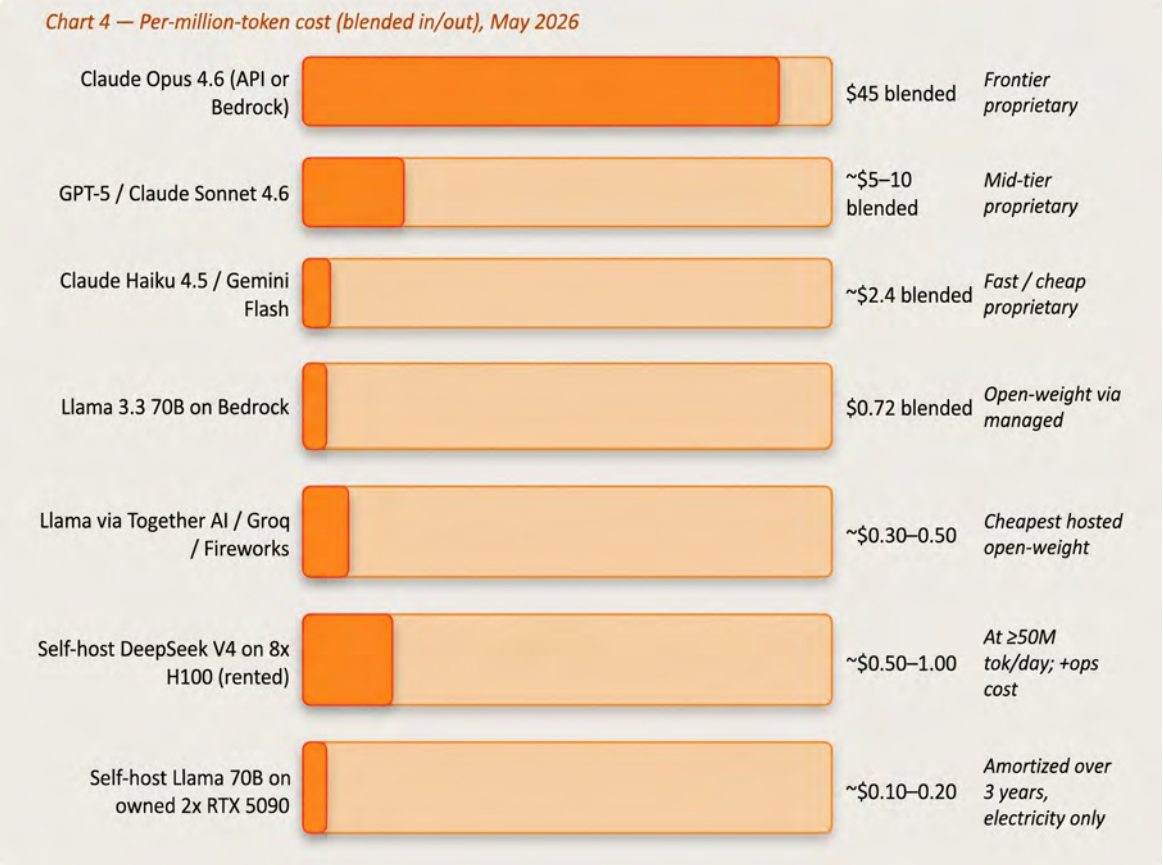
Two years ago, the gap between the best closed model and the best open-weight alternative on MMLU was ~17 percentage points. In Q2 2026, that gap on knowledge and reasoning benchmarks has narrowed to a few points; on coding (LiveCodeBench) and graduate-level science (GPQA), open-weight models hold their own.

The remaining differentials are on agentic benchmarks (BrowseComp, TerminalBench, OSWorld, SWE-bench Pro) and on multi-benchmark consistency. NIST CAISI's April 2026 evaluation pegs DeepSeek V4 Pro at roughly 8 months behind frontier closed models on aggregated capability.



Sources: BenchLM 2026 leaderboard, NIST CAISI April 2026 evaluation, model card disclosures. Benchmarks 0–100. The pattern: knowledge and coding gaps are small; agentic and abstract-reasoning gaps remain meaningful.

3.2 Per-token cost across tiers



3.3 Total Cost of Ownership at Three Workload Sizes

Per-token cost flatters Tier 3 and Tier 1 in different directions. TCO at realistic workload sizes is the cleaner comparison.

Workload	Tier 1: API + guardrails	Tier 2: Managed in VPC	Tier 3: Self-hosted
Small (1M tok/day)	$10–25K compute + $50K ops = $60–75K	$15–30K compute + $75K ops = $90–105K	$80K capex + $200K ops = $280K (not justified)
Medium (50M tok/day)	$200–500K + $100K ops = $300–600K	$250–600K + $150K ops = $400–750K	$200–250K rented GPU + $300K ops = $500–550K
Large (500M tok/day)	$2–5M + $200K ops = $2.2–5.2M (rate-limit risk)	$2.5–4M with PT + $250K ops = $2.75–4.25M	$240K capex + $400K facilities + $500K ops = $1.1–1.4M
Tier 3 dominates when...	—	—	≥50–100M tok/day on a single frontier model, or air-gap mandated

Illustrative ranges; exact numbers depend on model mix, batch profile, redundancy targets, and locale.

4. A Decision Framework

The right deployment for an organization is rarely the right deployment for every workload. Treating the choice as a per-workload routing decision rather than a one-time platform commitment is the pattern that scales.

Workload characteristics	Recommended tier	Why	Watch-outs
Public-facing chatbot on non-sensitive content; rapid model iteration	Tier 1	Lowest friction; capability ceiling matters more than data control	PII redaction at ingress; output filtering; jailbreak monitoring
Customer support with PII (orders, addresses)	Tier 2	BAA / DPA needed; data should stay in-region	VPC endpoints; customer-managed KMS; tenant isolation in RAG
Healthcare clinical assistance (PHI)	Tier 2 (or Tier 3 if mandated)	HIPAA BAA; data residency	Verify BAA scope; ZDR; output policy filters
Proprietary fine-tuned models (legal, medical, IP-heavy)	Tier 3	Weights are the moat	Capex justification; safety scaffolding must be built
High-volume bulk processing (>50M tok/day, latency-tolerant)	Tier 3 or Tier 2 Batch	Per-token economics dominate	Batch APIs at 50% discount can close the gap with no ops burden
Agentic systems with tool use & external actions	Tier 1 or 2 (frontier models)	Agentic gap to open models still meaningful	Tight tool scopes; human-in-loop for high-stakes actions; injection defense
Developer coding agents (Claude Code, Cursor, Cline, Windsurf, Zed)	Any tier – including local Ollama for source-code privacy	Source code is IP; default SaaS tools send it to the vendor; coding agents support gateway or Ollama endpoints	Streaming + tool-calling end-to-end; model pinning; per-developer IAM; SIEM logging
Decentralized / laptop-side workloads (offline drafting, local code review, prototyping)	Local via Ollama / LM Studio	Zero data egress, zero per-token cost; modern laptops run 7B–32B models	Approved catalog via MDM; pin versions; escalate to gateway when capability exceeded

Claude on Bedrock matches Anthropic's direct API pricing; Llama on Bedrock runs 22–201% above dedicated Meta partners (Together AI, Fireworks, Groq). Cross-region inference adds ~10%.

Recommended Path

- **Start With a Data Classification Scheme**

  Public/Internal/Confidential/Regulated. Map each workload to the lowest tier that satisfies its constraint.

- **Deploy an AI Gateway as the Single Ingress to All Tiers**
  Centralizes auth, budgets, fallbacks, redaction, audit, and routing. Details in Section 5.

- **Invest in the Trust Layer Regardless of Tier**
  PII redaction, injection detection, output policy, per-tenant filters, customer-owned audit logging – enforced at the gateway.

- **Treat Tier 3 as Additive, not Replacement**
  Justified by air-gap requirements, fine-tuned competitive moats,or sustained high volume – not by aspiration.

- **Push Selected Workloads onto Local Laptop/CPU Models via Ollama**
  Decentralized variant of Tier 3 for developer agents, offline scenarios, and prototyping. Details in Section 5.

- **Re-Evaluate Annually**
  Open-weight models improve fast; API pricing falls fast; GPU prices fall slowly. The gateway makes re-evaluation cheap.

5. Other Useful Patterns

Two cross-cutting patterns deserve their own treatment because they apply to more than one tier and are easy to get wrong if treated as afterthoughts. A decentralized small-model layer running on developer hardware, and the AI Gateway that sits in front of everything else.

Decentralized Hosting of Smaller LLMs on Laptops and CPUs (Ollama/LM Studio)

Modern open-weight models in the 3B–35B range run comfortably on developer hardware – Apple Silicon (M-series), single RTX 4090 / 5090 workstations, and even CPU-only machines with 32GB+ RAM via 4-bit quantization. Ollama and LM Studio wrap llama.cpp or vLLM in an OpenAI-compatible local endpoint (localhost:11434 for Ollama). Any tool that supports a custom base URL, including Claude Code, Cursor, Cline, Windsurf, Zed, and OpenCode, can point to it with a one-line configuration. The key facts that make this pattern serious in 2026:

- **Quality Has Crossed the Threshold**

  Qwen 3.6-35B-A3B (MoE, 3B active params) scores 73% on SWE-bench Verified, which is close to cloud frontier models.Phi-4, Gemma 4, Llama 4 8B,DeepSeek-Coder, Mistral Small handle most non-frontier tasks.

- **Throughput is Usable**
  7B–14B at 20–50 tok/s on consumer hardware; 32B at 8–15 tok/s;CPU-only 7B at 8–15 tok/s with quantization.

- **Zero Data Egress, Zero Per-token Cost**

  Source code, local documents, and personal data never leave the machine. Offline/travel/air-gapped scenarios work natively..

- **Governance Still Applies**

Distribute an approved model catalog via MDM (Jamf, Intune), pin model versions, push the same DLP and audit hooks the gateway uses, and route anything beyond local-model capability back to the gateway.

Good fits: Developer coding agents on local source code, sensitive document drafting and summarization, prototype iteration without burning API budget, classification and extraction on personal or HR data, edge/disconnected use.

Poor fits: Latency-critical multi-user services, large-context analysis beyond 32K–128K tokens, agentic workflows requiring frontier reasoning.

AI Gateway – The Single Ingress to All Tiers

Modern open-weight models in the 3B–35B range run comfortably on developer hardware – Apple Silicon (M-series), single RTX 4090 / 5090 workstations, and eveAn OpenAI-compatible (or Anthropic-Messages-compatible) gateway sits between every application and every LLM endpoint – Tier 1 APIs, Tier 2 managed services, Tier 3 self-hosted clusters, and local Ollama instances.

Common implementations:

LiteLLM (open source), Bifrost, Portkey, TrueFoundry, Kong AI Gateway, or AWS's published Multi-Provider Generative AI Gateway reference architecture. The gateway is the single highest-leverage control in an enterprise LLM platform because it centralizes every cross-cutting concern in one place: n CPU-only machines with 32GB+ RAM via 4-bit quantization. Ollama and LM Studio wrap llama.cpp or vLLM in an OpenAI-compatible local endpoint (localhost:11434 for Ollama). Any tool that supports a custom base URL, including Claude Code, Cursor, Cline, Windsurf, Zed, and OpenCode, can point to it with a one-line configuration. The key facts that make this pattern serious in 2026:

Capability	What it does
Identity & budgets	Per-developer/per-team virtual keys with token budgets and rate limits; OIDC federation from Okta/Entra ID.
Routing & fallback	Tier or model selection by workload class, automatic fallback when a provider throttles or fails.
Trust layer	PII redaction, prompt-injection detection, output policy enforcement, tenant-scoped allowlists – applied once, not per app.
Caching	Prompt caching and semantic caching across providers for cost reduction.
Audit & observability	Unified logging of prompts, completions, tool calls, and policy decisions into customer-owned SIEM with user attribution.
Vendor decoupling	Stable internal endpoint; switching Claude on Bedrock to Gemini on Vertex or vLLM is a config change, not a code change.

Conclusion

The privacy and security challenges of using LLMs are real, but they are architectural problems with well-understood solutions. The right answer is rarely "don't use LLMs" and is even more rarely "use them everywhere on a single tier."

It is to match each workload to the deployment pattern whose tradeoffs fit its data, its volume, and its risk surface, supported by a consistent trust layer and a flexible routing fabric. Public APIs give frontier capability with contractual safeguards; VPC-resident managed LLMs give the same models behind your own perimeter; self-hosted open-source gives total control where it is justified.

Used together, deliberately, these three tiers let an enterprise harness what LLMs do best – without surrendering the privacy, security, or control that its customers and regulators expect.

References

Standards & frameworks: OWASP Top 10 for LLM Applications (2026); MITRE ATLAS; EDPB Opinion 28/2024; NIST CAISI evaluation of DeepSeek V4 (April 2026); ISO/IEC 42001.

Benchmarks: BenchLM.ai (April 2026); Artificial Analysis 2026; LMSYS Chatbot Arena; HF Open LLM Leaderboard; vLLM / TensorRT-LLM.

Provider docs: AWS Bedrock pricing, PT, HIPAA BAA (2026); Anthropic, OpenAI, Microsoft, Google enterprise data-retention, ZDR, BAA (Q2 2026); Ollama documentation.

Industry analysis: Lasso Security, Wiz, Duality Technologies, TrueFoundry (2025–2026); Veracode 2025 GenAI Code Security Report; Truffle Security 2025.

INDIUM

About Indium

We are an AI services company specializing in Agentic AI, Data & Analytics, Application Engineering, and Quality Engineering.

By combining autonomous AI systems, modern engineering, and data intelligence, we build innovative AI solutions that drive measurable business outcomes and long-term value.

With 5,000+ associates worldwide, Indium partners with Fortune 500 companies, Global 2000 enterprises, and leading technology firms in Financial Services, Healthcare, Manufacturing, Retail, and Technology. Our teams operate in North America, India, the UK, Singapore, Australia, and Japan, helping businesses stay ahead in an AI-first world.

USA	INDIA	UK	SINGAPORE
Cupertino | Princeton Toll-free: +1-888-207-5969	Chennai | Bengaluru | Mumbai Hyderabad | Pune Toll-free: 1800-123-1191	London Ph: +44 1420 300014	Singapore Ph: +65 6812 7888

www.indium.tech

For Sales Inquiries
sales@indium.tech

For General Inquiries
info@indium.tech