

EXECUTIVE PAPER · AI

Choose Your AI Platform by What You Must Control

Executives are choosing between enterprise AI platforms, custom software on model APIs, and open-weight models run in-house. The right answer follows from five things the business must keep under its own control, and vendor demos reveal none of them.

AUTHOR

Chris Watkins

SERVICE

AI

FOR

CEO, CTO, CFO, CIO · 12 min read

AT A GLANCE

The problem. Companies pick an AI platform on demos and incumbent relationships, then find out later what they gave up. In 2025 and 2026 model providers retired models on 60 days' to six months' notice, OpenAI shut down an entire API and closed its fine-tuning platform to new users, and a flaw in a hosted assistant could have leaked internal files without a single click.

Why it persists. The three options look alike in a pilot. Differences appear only at volume, during an audit, or when a provider changes its product, and by then prompts, workflows and data are tied to one vendor.

The recommendation. Before comparing vendors, write down what the business must control on five points: where data may go, cost at expected volume, the ability to change models, a record of what the AI did, and the skills you can staff. Choose the lightest option that meets those requirements, and build a thin layer of your own for prompts, evaluations and logs so the choice can change later.

The Decision in Front of You

There are three ways to put generative AI into a business. An enterprise platform, from a large software or cloud vendor, bundles models, connectors to your data, agent tooling, security settings and billing. Custom orchestration means your own software calls one or more model providers through their APIs, so you own the workflow while they own the model. A private deployment runs open-weight models, meaning models whose trained parameters are published for download, on hardware you own or lease.

Andreessen Horowitz's 2025 survey of 100 CIOs across 15 industries found 37% running five or more models in production, up from 29% a year earlier [1]. The same survey found switching getting harder, because prompts are tuned to one provider's model [1]. A choice made for a pilot quickly becomes a multi-year commitment.

What's Happening

Providers change the product under you

Hosted models have a published shelf life. OpenAI commits to at least six months' notice before retiring a generally available model, three months for specialized variants, and as little as two weeks for preview models [2]. Anthropic commits to at least 60 days' notice for public models [3]. Both retired whole model families in 2025 and 2026: OpenAI's gpt-4.5-preview was announced for shutdown in April 2025 and turned off in July 2025 [2], and Anthropic retired its Claude 3 Opus model on January 5, 2026 [3]. Each retirement means retesting every prompt built on the old model.

The interfaces change too. OpenAI's Assistants API, which stored conversation threads and assistant settings on OpenAI's side, was deprecated on August 26, 2025 and shut down on August 26, 2026, with the Responses and Conversations APIs as replacements [2]. OpenAI's own guidance now states that it is winding down its fine-tuning platform, which is closed to new users; fine-tuned models stay available only until their base models are retired [4]. A company that tuned a model there keeps its training data, and the tuned model lasts only as long as its base model.

Open-weight models became a practical option

OpenAI released gpt-oss-120b in August 2025 under the Apache 2.0 licence; Google's documentation says it runs on a single 80GB GPU [5]. Other licences carry conditions. Meta's Llama 4 Community License requires a separate licence from Meta, granted at Meta's discretion, for any licensee whose products had more than 700 million monthly active users and requires "Built with Llama" attribution and compliance with Meta's acceptable use policy [6]. Each licence needs legal review.

Gartner predicted in April 2025 that by 2027 organizations will use small, task-specific models three times more than general-purpose large language models [7]. Adoption lags that forecast: coverage of Menlo Ventures' mid-2025 update put open-source models at 13% of enterprise workloads, down from 19% at the start of the year, partly because they are harder to deploy [8].

Regulators want records of what the AI did

The EU AI Act requires deployers of high-risk AI systems to keep the logs the system generates automatically, where those logs are under their control, for at least six months [9]. Under Regulation (EU) 2026/1744, which entered into force on July 27, 2026, those obligations apply from December 2, 2027 for standalone high-risk systems and from August 2, 2028 for AI embedded in regulated products [10]. A deployer that cannot export logs from a hosted tool has a gap it cannot close alone.

NIST's AI Risk Management Framework (AI 100-1, 2023) and its Generative AI Profile (AI 600-1, 2024) organize AI risk work under Govern, Map, Measure and Manage [11]. Neither prescribes a platform, and both assume the organization can see, measure and change its AI systems.

Four cases from 2025 and 2026

EchoLeak, Microsoft 365 Copilot, June 2025. Aim Security found a flaw, CVE-2025-32711, rated 9.3 out of 10: hidden instructions in an attacker's email, pulled in when a user asked Copilot a routine question, could send the user's internal data out through links with no click required [12]. Microsoft fixed it on the server side, customers needed to take no action, and no exploitation was reported [12]. The hosted platform patched every customer at once. Customers could not inspect or log the step that failed; their only control was what the assistant could read.

Mixpanel and OpenAI API customers, November 2025. Mixpanel, the web analytics vendor OpenAI used on its developer platform, detected a breach on November 9, 2025 [13]. Names, email addresses, approximate locations and organization IDs of API users were exposed; passwords and keys were not

[13]. OpenAI ended its use of Mixpanel [13]. A hosted AI path includes the provider's own suppliers, and your data residency review has to reach them.

Assistants API shutdown, 2025 to 2026. Teams whose conversation threads and assistant settings lived in OpenAI's Assistants API had twelve months to rebuild on a different API [2]. No cost figure has been published. The work fell on everyone who had let the provider hold application state.

HSBC and Mistral, December 2025. HSBC announced a multi-year partnership with Mistral AI to run Mistral's models self-hosted on HSBC's internal technology systems, starting with client communications, lending file analysis and translation [14]. No results or costs have been disclosed. A bank with US$3,234 billion in assets [14] chose to buy models from a vendor and run them inside its own walls, separating who builds the model from who controls the data.

Why the Usual Approach Fails

Most platform decisions start from the vendor list. A team scores demos and picks the vendor the company already pays. A demo on sample data cannot show cost at a million requests a month or whether logs can be exported for an auditor.

The second mistake is choosing one platform for every use case. A regulated lending workflow and a marketing copy tool have requirements that differ more than the products do.

The third mistake is letting the provider hold the parts that are expensive to rebuild. Model-specific prompts, conversation state in a vendor API and fine-tuned weights on a closed platform all become switching costs [1][2][4], usually chosen by default in the first month.

The Framework: The Control Test

The Control Test asks five questions in order: data, cost at volume, portability, audit and skills. Each answer is a written requirement with an owner, set before any vendor is compared. The option that meets every requirement with the least engineering is the right one.

Data. State where each use case's data may be processed and stored. Hosted options sell regional processing; Anthropic charges a 1.1x multiplier on all tokens for US-only inference [15]. If the data may not leave your infrastructure at all, the answer is private deployment, whatever the other four say.

Cost at volume. Price the use case at its expected production volume, with your own token counts. Salesforce launched Agentforce Flex Credits in May 2025 at $0.10 per agent action [16]. Model APIs charge per token; Anthropic lists its Sonnet 5.5 model at $2 per million input tokens and $10 per million output tokens, half that through its batch interface [15]. Private deployment turns most cost into fixed hardware and staff; the paper on the real cost of AI in production covers the full stack.

Portability. Decide how fast you must be able to change models. Retirement notice periods of 60 days to six months set the floor [2][3]. Keeping prompts, test sets and conversation state in your own systems keeps that move to weeks.

Audit. Decide what record you must keep of each AI action: inputs, retrieved documents, model version, output, and any action taken in another system. For high-risk uses in the EU, deployers must retain logs for at least six months [9]. Confirm before signing that you can export them.

Skills. Custom orchestration needs software engineers who can build evaluation and logging. Private deployment adds people who can run GPU infrastructure, patch model servers and plan capacity. Without them, the platform is safer.

Exhibit 1. How each requirement points to an option

Control requirement	Enterprise platform	Custom orchestration on model APIs	Private open-weight deployment
Data must stay in your infrastructure	Fails unless the vendor offers a private or on-premises edition	Fails; regional processing narrows but does not remove the gap	Meets
Data may go to a named region with a contract	Meets if the vendor offers the region	Meets; regional pricing applies [15]	Meets
Cost at high, steady volume	Per-action or per-seat pricing grows with use [16]	Per-token pricing; batch and caching reduce it [15]	Mostly fixed cost for hardware and staff, spread across every request
Change models within weeks	Limited to models the vendor supports	Meets if prompts, tests and state live in your code	Meets; you choose when to upgrade
Export a full record of every action	Depends on the vendor's logging; verify in contract	Meets; you write the log	Meets; you write the log
Skills on staff	Administrators and business analysts	Software engineers with AI evaluation skills	Engineers plus GPU and infrastructure operators

Source: author's analysis; pricing from [15] and [16].

SIDEBAR: PROMPT, RETRIEVAL OR FINE-TUNE

Inside any option there is a smaller choice: how to make a general model good at your task. Start with the prompt. OpenAI's own guidance says prompt engineering may be all a use case needs [4]. Add retrieval-augmented generation (RAG), which looks up your documents and hands the relevant passages to the model, when the model needs facts it was not trained on or must answer only from approved content. Fine-tune, meaning retrain the model on your examples, only when format, tone or cost at scale cannot be fixed any other way; OpenAI notes it can let a smaller, cheaper model do a task where a larger one is not cost-effective [4]. Fine-tuning is the least portable: a hosted fine-tune disappears with its base model [4], and prompts move between providers more easily [1]. If you must fine-tune, do it on open weights you hold.

Costs, Risks, and Tradeoffs

The author's illustrative estimate below shows why the pricing model matters. Assume a customer service use case with 2 million requests a month, each using 3,000 input tokens (instructions plus retrieved documents) and 500 output tokens.

Exhibit 2. Pricing model sets cost at volume more than the model does (author's illustrative estimate)

Option, public list price	Cost per request	Monthly cost at 2 million requests	Annual cost
Per-action platform pricing at $0.10, one action per request [16]	$0.10	$200,000	$2,400,000
Model API, Sonnet 5.5 at $2 input and $10 output per million tokens [15]	$0.011	$22,000	$264,000
Same, with US-only processing at 1.1x [15]	$0.0121	$24,200	$290,400
Model API, Haiku 5.5 at $0.10 input and $0.50 output per million tokens [15]	$0.00055	$1,100	$13,200

Source: author's illustrative estimate. Assumptions: 2,000,000 requests a month, 3,000 input and 500 output tokens per request, list prices without negotiated discounts, caching or batch. Private deployment omitted because its cost depends on your hardware and how busy it is.

The platform price includes hosting, connectors and administration. The API rows leave out the engineers, evaluation and logging you would build around them. At 2 million requests a month the gap is large enough to fund that team. At 20,000 requests a month the platform row falls to $2,000, less than the cost of building your own.

Exhibit 3. Tradeoffs of each option

Option	What it costs you	Main risk	Mitigation
Enterprise platform	Per-action or per-seat fees; roadmap set by the vendor	Lock-in of workflows and data inside the vendor; logs you cannot export	Contract for log export and exit assistance; keep test sets and prompts in your own repository
Custom orchestration on APIs	Engineering team; you build evaluation and logging	Model retirements on 60 days' to six months' notice force retesting [2][3]	A model-neutral layer, a standing evaluation suite, and pinned model versions
Private open-weight deployment	Hardware, power, GPU operators, security patching	Idle capacity; falling behind newer hosted models; licence conditions [6]	Start with one high-volume, high-sensitivity use; legal review of each licence
When this is wrong	The Control Test assumes your requirements are knowable	A first pilot with no volume, sensitivity or audit history to measure	Run the pilot on the fastest option, record token counts and data types, then apply the test before production

Source: author's analysis.

How to Act

Exhibit 4. A one-year roadmap

Window	Actions	Owner	Milestone
First 30 days	List every AI use case with its data class, monthly volume, audit needs, model and provider.	CIO with data owners	Inventory complete; every use case has a written answer to the five control questions
First 30 days	Check contracts for log export, processing locations, subprocessors and retirement notice.	General counsel with procurement	Gaps listed per vendor; renewals flagged
First 90 days	Move prompts, test sets and conversation state into your own repository; build an evaluation suite per use case.	CTO	Every production use case can be rerun against a second model within one week
First 90 days	Price the top three use cases at production volume under each option with real token counts.	CFO with CTO	Cost per request and annual cost per option, signed off by finance
First 365 days	Log inputs, sources, model version, output and actions, kept six months or more; map controls to the NIST AI RMF.	CISO with risk and compliance	Logs exportable on request for every use case; audit evidence on file ahead of December 2, 2027
First 365 days	Rehearse a forced model change on one production use case. Decide whether any use case meets the bar for private deployment.	CTO	Model swap completed within the notice period; private deployment decision recorded with its reasons

Source: author's recommendation.

Questions to Ask Your Team

1. For each AI use case, where is the data processed and stored, including the provider's own subprocessors?
2. What does this use case cost per request at the volume we expect in a year, and under which pricing model?
3. If our main model were retired with 60 days' notice, how long would the change take, and who would do it?
4. Where do our prompts, test sets and conversation history live today, and could we export them this week?
5. Can we produce a record of every action an AI system took last month, with the model version that took it?

6. Which open-weight licences are we using, and has legal reviewed their conditions?

Closing

Choosing between an enterprise platform, custom orchestration and private open-weight models is a choice about control. Write down what the business must control on data, cost at volume, portability, audit and skills, then pick the lightest option that meets every requirement. Keep prompts, tests and logs in your own systems whatever you choose, so that a retired model or a changed contract means weeks of work.

Sources

[1] Andreessen Horowitz. "How 100 Enterprise CIOs Are Building and Buying Gen AI in 2025." Survey of 100 CIOs across 15 industries. 2025. https://a16z.com/ai-enterprise-2025/.

[2] OpenAI. "Deprecations." OpenAI API documentation. 2026. https://developers.openai.com/api/docs/deprecations.

[3] Anthropic. "Model deprecations." Claude Platform documentation. 2026. https://platform.claude.com/docs/en/about-claude/model-deprecations.

[4] OpenAI. "Model optimization." OpenAI API documentation. 2026. https://developers.openai.com/api/docs/guides/model-optimization.

[5] Google Cloud. "OpenAI gpt-oss 120B." Vertex AI model documentation. 2025. https://docs.cloud.google.com/vertex-ai/generative-ai/docs/maas/openai/gpt-oss-120b.

[6] Meta. "Llama 4 Community License Agreement." Effective April 5, 2025. Copy published by Wind River. 2025. https://open.windriver.com/info/uni-license-list/licenses/llama-4-cla-2025.html.

[7] Gartner. "Gartner Predicts by 2027, Organizations Will Use Small, Task-Specific AI Models Three Times More Than General-Purpose Large Language Models." Press release. 2025. https://www.gartner.com/en/newsroom/press-releases/2025-04-09-gartner-predicts-by-2027-organizations-will-use-small-task-specific-ai-models-three-times-more-than-general-purpose-large-language-models.

[8] TechCrunch. "Enterprises prefer Anthropic's AI models over anyone else's, including OpenAI's." Coverage of Menlo Ventures' 2025 mid-year LLM market update. 2025. https://techcrunch.com/2025/07/31/enterprises-prefer-anthropics-ai-models-over-anyone-elses-including-openais/.

[9] European Union. "Article 26: Obligations of Deployers of High-Risk AI Systems." Regulation (EU) 2024/1689 (AI Act), as published by the AI Act Explorer. https://artificialintelligenceact.eu/article/26/.

[10] National Law Review. "EU Digital Omnibus on AI Enters Into Force." Regulation (EU) 2026/1744. 2026. https://natlawreview.com/article/eu-digital-omnibus-ai-enters-force.

[11] National Institute of Standards and Technology. "AI Risk Management Framework." NIST AI 100-1 (2023) and NIST AI 600-1 Generative AI Profile (2024). https://www.nist.gov/itl/ai-risk-management-framework.

[12] The Hacker News. "Zero-Click AI Vulnerability Exposes Microsoft 365 Copilot Data Without User Interaction." CVE-2025-32711. 2025. https://thehackernews.com/2025/06/zero-click-ai-vulnerability-exposes.html.

[13] The Register. "OpenAI cuts off Mixpanel after analytics leak exposes API users." 2025. https://www.theregister.com/2025/11/27/openai_mixpanel_api/.

[14] HSBC Holdings plc. "HSBC and Mistral AI join forces to accelerate AI adoption across global bank." Media release. 2025. https://www.hsbc.com/news-and-views/news/media-releases/2025/hsbc-and-mistral-ai-join-forces-to-accelerate-ai-adoption-across-global-bank.

[15] Anthropic. "Pricing." Claude Platform documentation. 2026. https://platform.claude.com/docs/en/about-claude/pricing.

[16] Salesforce. "Salesforce Introduces New Flexible Agentforce Pricing to Accelerate the Digital Labor Revolution." Press release. 2025. https://www.salesforce.com/news/press-releases/2025/05/15/agentforce-flexible-pricing-news/.

The views in this paper are the author's own and do not represent any employer. Scenarios are hypothetical. This paper is not legal, financial, or compliance advice.