Q3 2026

KAI WAEHNER LANDSCAPE Q3 2026

TRUSTED AGENTIC AI LANDSCAPE

Enterprise Trust, Sovereignty, and Vendor Lock-in for Autonomous Agents

Published: August 2026

Kai Waehner
ADVISORY FIELD CTO

kai-waehner.de

INSIDE THIS REPORT

Contents

Choosing an enterprise AI vendor used to be a procurement exercise. You compared features, pricing, support, and integration effort. If a vendor disappointed you, migration hurt but it was possible. Agentic AI changed the stakes.

The model you select shapes how your agents reason, what they can and cannot do, how your data is handled, and how deeply you become entangled in one vendor's ecosystem. An agentic system does not just answer a question. It takes actions, makes decisions, and orchestrates workflows. Getting the vendor decision wrong in that context costs more than any previous enterprise software choice.

Two events in the first half of 2026 made this concrete.

First, the public markets opened. SpaceX listed in June in the largest IPO in history, and its volatile first weeks showed how fast public markets reprice a story. Anthropic and OpenAI filed confidentially within days of each other, at valuations near or above a trillion dollars. The vendors you build on are becoming subject to public-market scrutiny, with their burn rates, customer concentration, and legal risks about to become public record.

Second, and more important for this landscape, a government switched off a frontier model. In June 2026 the US government issued an export-control directive that suspended access to Anthropic's two most capable models for any foreign national, inside or outside the United States. Anthropic complied and disabled both models for everyone, while its other models stayed available. Nineteen days later access was restored, after negotiations between the vendor and the Commerce Department. Do not read the restoration as the risk going away. Read it as the risk being defined. **A frontier model can be switched off for you overnight by a government you did not choose, and switched back on only on terms you had no part in negotiating.** The lesson points the opposite way from any vendor marketing: even a US frontier lab that has put safety at the center of its positioning can be made unavailable to you by its own government.

Sovereignty is therefore a first-order dimension in this edition, not a European footnote. Trust and sovereignty have stopped being separate conversations.

The Trusted Agentic AI Landscape Q3 2026 is not a ranking. No vendor pays to appear here. The analysis is based on my own experience advising Global 2000 enterprises on AI and data architecture, combined with ongoing research into vendor positioning, product developments, and adoption patterns. It is an independent practitioner perspective, not a formal research methodology like Gartner or Forrester.

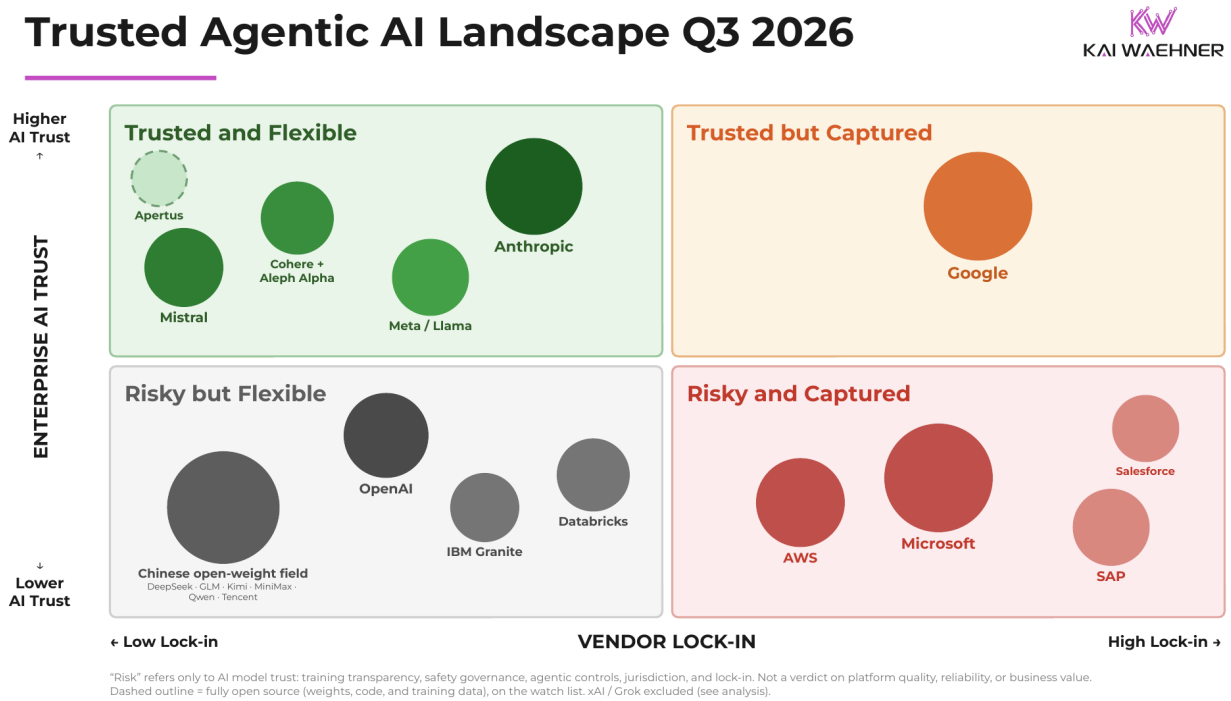

Figure 01. The Trusted Agentic AI Landscape Q3 2026. Bubble size reflects enterprise influence and adoption scale. Risk refers only to AI model trust, not platform quality.

01

CHAPTER 01

Why AI Vendor Selection Is a Different Decision

An AI vendor is a strategic partner, not a tool. Three properties make the selection harder than any previous enterprise software decision.

Trusted agentic AI means AI agents an enterprise can rely on to act autonomously: built on models with transparent safety governance, running on data handled under clear rules, operating in a jurisdiction that cannot cut you off, and architected so no single vendor controls the stack. This landscape evaluates the vendors behind those agents on exactly those dimensions.

A CRM or an ERP is a tool you deploy. An AI vendor is closer to a strategic partner whose safety culture, governance model, jurisdiction, and long-term ambitions flow directly into the reliability of your most critical processes.

Three properties make the decision harder than traditional software selection.

The model is not the whole product. Around every model sits an orchestration layer, an agent runtime, a connectivity standard, and a data and context layer. Lock-in accumulates across all of them, not only at the model API.

The vendor's home jurisdiction is now part of the risk. Where a vendor is legally domiciled, which laws it operates under, and who can compel or restrict it are no longer abstract concerns. They can determine whether you keep access at all.

Capability and trust do not move together. Some of the most capable models carry the highest lock-in or the most jurisdictional exposure. Some of the most flexible options carry real questions about safety governance or data access. You are balancing at least two variables that pull in different directions.

02

CHAPTER 02

The Two Dimensions That Define the Landscape

Two axes define the map: enterprise trust and vendor lock-in, each read on its own terms and on two distinct levels.

The vertical axis: enterprise trust

Enterprise trust here is not a benchmark score. It combines three sub-dimensions.

Safety and governance. Does the vendor have a demonstrable commitment to responsible development? Can a risk team inspect how the model is designed to behave before deployment? How does the vendor handle a safety failure when one occurs, because every major vendor eventually has one.

Data handling. Will your data be used for training, and under what conditions? Are zero-retention options available for sensitive workloads? Is regulatory compliance real or a checkbox?

Jurisdiction and sovereignty. Where is the vendor headquartered, under which laws does it operate, and what does that mean for the continuity and confidentiality of your workloads? This sub-dimension carried less weight a year ago. The export-control episode moved it to the center.

The horizontal axis: vendor lock-in

For AI, lock-in is more subtle than in traditional software, and it now lives on two distinct levels.

Model-level lock-in is the familiar kind. Your architecture bends around one vendor's API design, fine-tuning format, and agent framework. Switching means re-engineering prompts, tools, and evaluations. Every frontier vendor creates some of this. A sharper recent example is Anthropic: strong model trust, but its acquisition of the company behind its SDK and connector tooling means the model, the connectivity standard it authored, and the toolchain that builds connections now sit under one owner.

Stack-level lock-in is the kind that has grown fastest. The model becomes interchangeable, but the data, the business context, the orchestration logic, and the agent runtime do not. SAP is the clearest example. It now runs multiple models, including its own and several third-party frontier models, so it is open at the model layer. The lock-in moved up the stack, into its data platform, its process knowledge, and its semantic and governance layer. **You can swap the model. You cannot easily swap the context graph that makes the agents useful.**

Both levels matter. A vendor can be open on one and closed on the other. Reading them separately is the whole point.

How to read the map

Bubble size encodes enterprise influence and adoption scale. Larger bubbles shape what developers expect, what integrations exist, and what the market treats as default. Bigger is not better. For a regulated or sovereignty-sensitive buyer, a smaller and more controllable vendor is often the right choice.

When this landscape uses the word "risky," it refers only to the AI model layer: transparency of training, safety governance, agentic controls, jurisdiction, and lock-in. It says nothing about the overall quality, financial stability, or business value of these vendors' broader platforms. SAP is an excellent ERP company. Microsoft is a leading enterprise technology company. AWS is world-class infrastructure. The risk label is narrow and specific to how each approaches AI model trust and flexibility.

03

CHAPTER 03

Why Models Are Not Stable Infrastructure

Models get deprecated, repriced, and restricted. The architecture, not the contract, is what absorbs that, and the cost curve underneath changes the trade-offs for everyone.

Outages are the fast version of a broader problem: models are not stable infrastructure.

They get deprecated on the vendor's schedule, typically with around 60 days of notice, and the cadence is accelerating across every provider. Anthropic retired the original Sonnet 4 and Opus 4 in June 2026, with further retirements already scheduled through September. OpenAI and Google run the same lifecycle and have already retired their GPT-4-era and Gemini 1.5 model families. Models also get repriced as vendors move away from flat-rate economics, and, as this year showed, they can be withdrawn by regulation or launched under government restriction, as OpenAI's newest family was in June 2026. Different triggers, one identical failure mode: an architecture hardwired to a single model endpoint stops working.

Two conclusions follow. First, pinning a model version is not a stability strategy. Deprecation forces the migration eventually, and even a frozen model is probabilistic. Hallucination and output variance cannot be engineered away by staying on an old version, so evaluation and regression testing against model changes are permanent operational work, not a one-time project. Second, the mitigation is architectural, not contractual. Put an abstraction or gateway layer between the agent orchestration and the model endpoints. It handles routing, fallback, and provider swaps without touching the integration. Deprecations, outages, and price changes then become configuration changes instead of emergencies. The deployment model is part of the same decision. A managed API is retired for you. Self-managed open weights never expire, but they shift the hosting, optimization, and safety burden onto your own team. Mature architectures mix both and keep the switch cheap.

Switching the endpoint is the easy half of a multi-model strategy. The hard half is what the switch does not carry: the context that accumulated around the old model. Conversation history, agent memory, learned preferences, and workflow state stored inside a vendor's memory features or agent runtime do not export cleanly, and data-portability clauses rarely cover them. Embeddings are the sharpest example. Vectors are specific to the model that produced them, so a vector store built on one provider's embeddings is useless to another until everything is re-embedded from the source data. If you did not keep the source data, there is nothing to re-embed. Prompts carry the same friction in milder form: they do not transfer one to one, which is why a model-agnostic evaluation harness is part of the real switching cost.

The practice that works is to treat context as data you own. Keep conversation state, agent memory, and knowledge in your own stores: a database for state, a vector store fed from source documents you retain, a knowledge graph, an event log. Assemble the context per request and hand it to whichever model is on duty. Then the model holds nothing between calls that you cannot hand to its replacement. Vendor memory features are fine for convenience where switching does not matter, and wrong as the system of record. This is the same stack-level lock-in described in chapter 2, seen from the other side. SAP and Salesforce own the context layer, so the model became interchangeable for them. An enterprise building directly on foundation models should claim exactly that position for itself. **Own the context, rent the model.**

Underneath all of this sits a cost curve that changes the trade-offs for everyone. Agentic workloads consume tokens at a scale chat never did, and inference is a recurring cost tied to every action an agent takes, not a one-time investment like training. McKinsey projects that inference will overtake training as the dominant AI workload by 2030, at more than half of all AI compute. At the same time, the cost of delivering it is collapsing. Quantization, distillation, sparse architectures, and small specialized models drive efficiency gains fast enough that quasi-frontier open-weight models now run on hardware a single team can afford. The consequences run through this entire landscape. Cheap inference is why flat-rate agentic pricing broke and platform vendors moved to usage-based billing. Subscription pricing for coding agents is widely reported to sit below the cost of

serving heavy users, which is why 2026 became the year the major coding tools moved to metered billing. Today's subsidized rates are not a baseline to plan on. It is why open-weight models, many of them Chinese, exert price pressure at the top of the Western enterprise stack.

It is why self-hosted deployment, and with it real sovereignty, is becoming economically realistic rather than aspirational. And it is why routing by task, cost, and sensitivity is becoming the default pattern of a mature multi-model architecture, rather than sending everything to the most capable model. The pattern now has hard numbers behind it. Cursor's July agent-swarm research ran the same large coding task across model mixes and found that a frontier model planning while a low-cost model executes delivers similar quality at a fraction of the price, with roughly an eight-fold cost spread between the cheapest and most expensive configurations. Speed belongs in the same routing decision. Frontier models now reason deeply by default, with effort settings as the primary dial between intelligence, latency, and cost, and a fast model that finishes a routine task in a quarter of the tokens is cheaper twice, once on rate and once on volume. The flip side of the consumption curve is governance. An AI API key is an unthrottled spending instrument, and a stuck agent loop or a stolen key drains budget at machine speed, well before a monthly invoice review catches it. Real-time consumption monitoring, per-agent spend limits, and kill-switch authority belong in the operating model from the start, and the discipline is formalizing: the Linux Foundation announced the Tokenomics Foundation in mid-2026, a standards body for AI token usage and billing built in partnership with the FinOps Foundation. The same announcement cites Goldman Sachs research projecting global token usage to multiply 24 times between 2026 and 2030.

The efficiency curve behind all of this has a limit worth naming. The gains hold at a fixed capability level, pulling a prior generation's frontier onto hardware a single team can afford. They do not hold at the open frontier itself, which is now scaling into multi-trillion-parameter models that need a multi-GPU supernode to serve well. For those models self-hosting is out of reach for most buyers, which leaves them on someone else's API. Open weights and low cost were always correlated, never the same thing, and the correlation breaks once the weights get big enough that only a well-provisioned operator can serve them.

04

CHAPTER 04

No Quadrant Is the Right Quadrant

Every position on the map is a set of trade-offs. The right one depends on your role, your industry, your jurisdiction, and your risk tolerance.

There is no objectively correct position on this landscape. Every spot is a set of trade-offs, and the right trade-off depends on who you are and what you are building.

A global manufacturer running SAP at the core does not need its finance and supply-chain teams thinking about foundation models. For those users the AI is infrastructure, and accepting a captured position in exchange for AI embedded in the processes they already run is rational. The same holds for a sales team inside Salesforce or a service desk on Microsoft Copilot. These users need business outcomes, not model portability.

The position that demands the most scrutiny is the one where you build directly on foundation models. Where developers call APIs, architects design agentic workflows, and competitive differentiation depends on what you build rather than what a vendor builds for you.

Trust and lock-in become first-order there. It is also why many large enterprises now run a multi-model strategy: different models for different use cases, architectural separation between the agent orchestration layer and the model calls, and the freedom to switch or combine as the market moves. The driver is no longer only cost or use-case fit. After the export-control episode, resilience against a single provider's pricing, availability, and jurisdiction risk became a reason in its own right. The point was reinforced by a run of multi-provider outages this year, several hitting more than one major assistant at once.

A single hardcoded provider endpoint is now a single point of failure, and a second model behind a failover path is the difference between a degraded mode and a stoppage. The uptime commitments that do exist, mostly from the hyperscaler-hosted model services, sit below what enterprises expect from a managed database or cloud region. A service credit for a downed endpoint is not the same as business continuity, which makes a model-unavailability plan part of due diligence rather than an afterthought.

The key question is not which quadrant is best. It is which quadrant matches your role, your use case, your industry, your jurisdiction, and your risk tolerance.

05

CHAPTER 05

The Four Quadrants

A vendor-by-vendor analysis across all four quadrants, focused on model trust, jurisdiction, and lock-in, plus why xAI is not on the map.

What follows is a vendor-by-vendor analysis. **Each assessment focuses on AI model trust, safety, jurisdiction, and lock-in, not on overall platform quality or market position.**

Trusted and Flexible

The top-left quadrant is where enterprises building directly on foundation models should aim to operate. Vendors here pair a credible trust posture with deployment models that preserve architectural freedom. The common thread: you can go deep without losing the ability to change course.

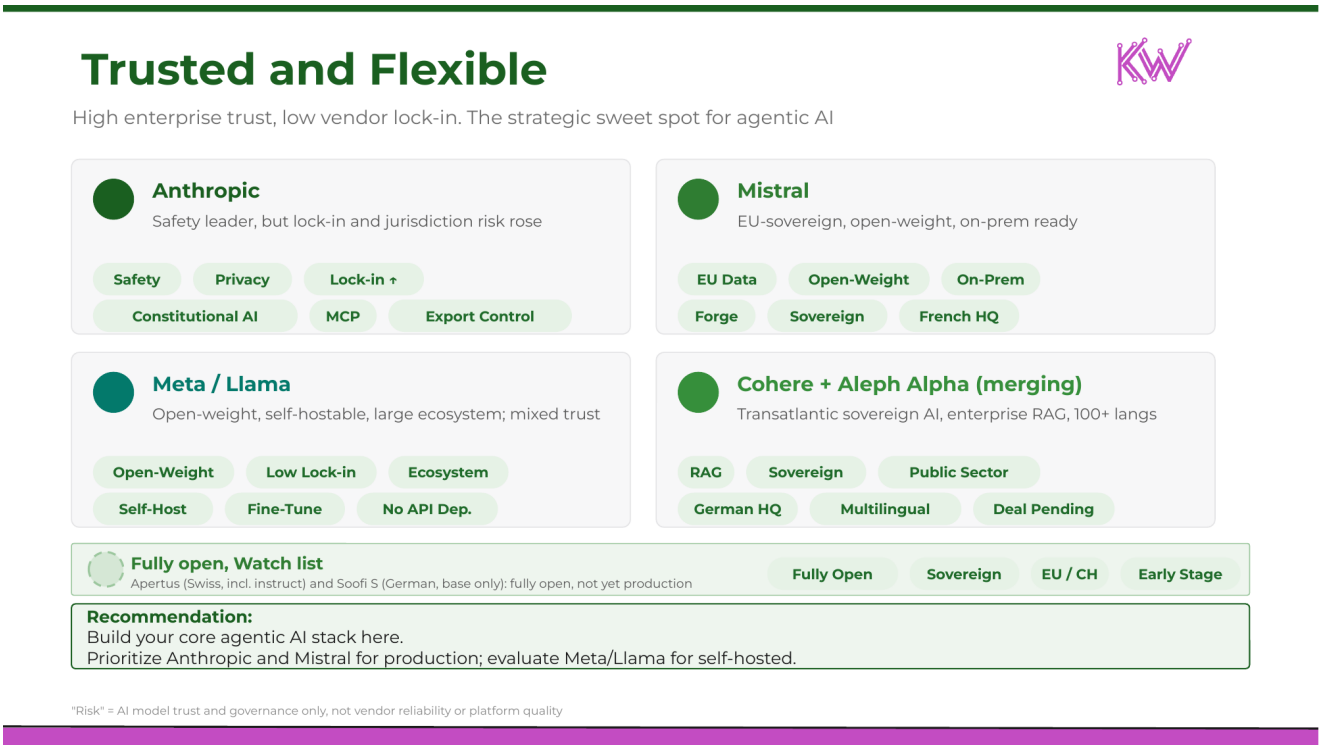

Figure 02. Trusted and Flexible: high enterprise trust and low vendor lock-in.

Anthropic (Claude Platform)

The Claude family is built with safety as a design principle rather than a later addition. Constitutional AI gives risk teams a published, inspectable set of behavioral principles to evaluate before deployment. Anthropic's interpretability research offers a degree of model transparency few frontier labs match. Zero-data-retention options exist for sensitive workloads. On enterprise adoption, independent data puts Anthropic ahead of OpenAI in enterprise model API spend and well ahead in the coding segment, driven in large part by Claude Code.

The flexibility picture needs care in this edition. Claude is available through the direct API and through AWS Bedrock, Google Vertex AI, and Azure, which avoids forcing a single cloud. Against that, three developments raise real questions. The Pentagon designated Anthropic a supply chain risk in early March, requiring the US military and its contractors to stop using the models, and Anthropic is contesting that in court. In June the government went further, suspending access to Anthropic's two most capable models for all foreign nationals, which forced Anthropic to disable those models globally. Access was restored on July 1 after the Commerce Department lifted the controls, but the redeployment came with strings attached: a stricter safety classifier that blocks more benign requests and offers no customer opt-out, restricted-tier access limited to vetted US

critical-infrastructure organizations, and commitments that formalize government involvement in Anthropic's deployment process. Separately, Anthropic's acquisition of its SDK and MCP-server tooling vendor concentrates the model, the connectivity standard, and the connector toolchain under one owner, which is a lock-in signal worth weighing.

Pricing has also historically been a friction point, and latency joined it this year: Fable 5 reasons deeply by default, with an effort setting as the main control. Analysts report the combination of classifier refusals and wait time has pushed some power users toward rival tools.

Anthropic answered on July 24 with Opus 5: near-flagship results at half the price, classifiers scoped to intervene about 85 percent less often, and automatic fallback to Opus 4.8 for flagged requests. The deeper signal is capability shaping. Anthropic deliberately did not train Opus 5 on cybersecurity tasks, so it approaches Mythos 5 at finding vulnerabilities while staying far behind at exploiting them. After June, model design itself has become an export-control strategy.

The net read: Anthropic remains a leader on model trust and a strong default for the trusted-and-flexible buyer, but the June episode converted jurisdiction exposure from a theoretical concern into a measured one.

The precedent applies to every US lab, not only this one, and it has already repeated. In late June the White House asked OpenAI to limit the launch of its newest model family to a small group of government-approved partners, under a new executive order that formalizes pre-release government review of frontier models. Two labs, two interventions, one month.

A non-US or sovereignty-sensitive enterprise should treat it as a permanent input to the architecture, not as a 2026 incident that passed. One structural point sharpens it. Because Claude is a closed model, the switch-off left no fallback: there were no weights to self-host and ride out the suspension. It is the same test applied to Chinese models later in this chapter, run in the other direction, and a closed frontier model concentrates the availability risk with no escape hatch.

Cohere and Aleph Alpha (merging)

These were two separate entries last edition. They are now combining. Cohere is acquiring Germany's Aleph Alpha to form a transatlantic enterprise-AI company positioned explicitly as a sovereign alternative to US labs, backed by a major European retail conglomerate and supported by both the Canadian and German governments. The combined entity targets regulated sectors and public-sector buyers that want control over data and infrastructure. Cohere brings retrieval, embeddings, and enterprise focus with low lock-in. Aleph Alpha brings European-language depth, public-sector relationships, and on-premises experience. The deal has not closed and is subject to regulatory approval, so treat the combined roadmap as a direction rather than a finished product. For knowledge management, document intelligence, and sovereignty-led deployments, this is now a serious European candidate.

Meta (Llama)

Llama models are open-weight, so enterprises can self-host, fine-tune, and deploy without ongoing API dependency. The trust picture is mixed. Meta is a large consumer technology company with a varied governance history, and the Llama license carries commercial restrictions at very large scale. Benchmark-transparency questions at recent launches left some trust issues the enterprise community has

not fully resolved. For organizations that want maximum architectural control and can run their own inference, Llama remains a serious option, and its open-weight nature makes it directly relevant to the sovereignty discussion in chapter 6.

Mistral

Mistral is the most production-ready European option in this quadrant. Open-weight models, French jurisdiction, and EU AI Act alignment give sovereignty-led buyers a combination of control and flexibility that no US hyperscaler matches, and the Forge platform lets enterprises train custom models on their own data. Distribution has widened this year: Mistral models run on AWS Bedrock, inside SAP's Business AI Platform, and, after the July agreement with Microsoft, on European-operated compute in Microsoft Foundry and Copilot Studio with deployment options reaching air-gapped environments. Mistral is not at Anthropic's scale, and the largest hosted models carry the same API-path caveats as any vendor. For European regulated industries it is the default production candidate on this map.

> **RECOMMENDATION**
>
> Build your core agentic AI stack in this quadrant. Prioritize Anthropic and Mistral for production deployments, and evaluate Meta's Llama where self-hosting is the goal. Watch Cohere and Aleph Alpha for sovereignty-led deployments once the merger closes.

Trusted but Captured

The top-right quadrant holds vendors with strong capability and a credible trust posture, but deployment models that create significant lock-in. With Aleph Alpha moving into the Cohere entity, this quadrant is sparser than last edition, and Google now dominates it.

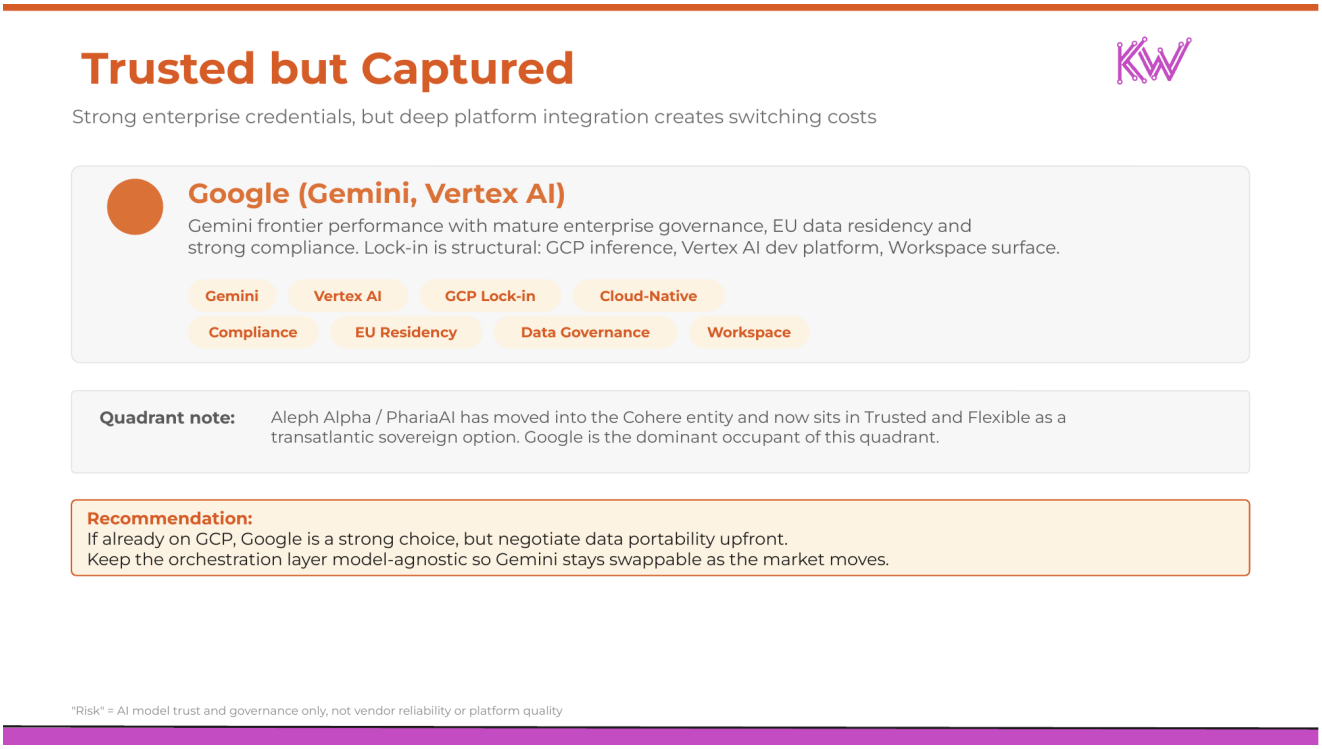

Figure 03. Trusted but Captured: strong enterprise credentials with deep platform integration.

Google (Gemini, Vertex AI)

Gemini is capable and Google's enterprise posture has matured, with data-governance commitments, EU data-residency options, and strong compliance frameworks. The lock-in is structural. Choosing Gemini tends to mean Google Cloud as the inference layer, Vertex AI as the development platform, and often Workspace as the productivity surface. Each integration deepens the commitment. For organizations already on GCP this can be rational. For organizations trying to keep multi-cloud flexibility it is a real constraint.

> **RECOMMENDATION**
>
> If you are already on GCP, Google is a strong choice, but negotiate data portability upfront. Keep the orchestration layer model-agnostic so Gemini stays swappable as the market moves.

Risky but Flexible

The bottom-left quadrant holds vendors that offer real flexibility and often strong performance, but where trust concerns at the model and governance layer introduce risk. To be explicit, some of the most important technology organizations in the world sit here. The risk label applies only to AI model transparency, safety governance, and jurisdiction, not to overall credibility.

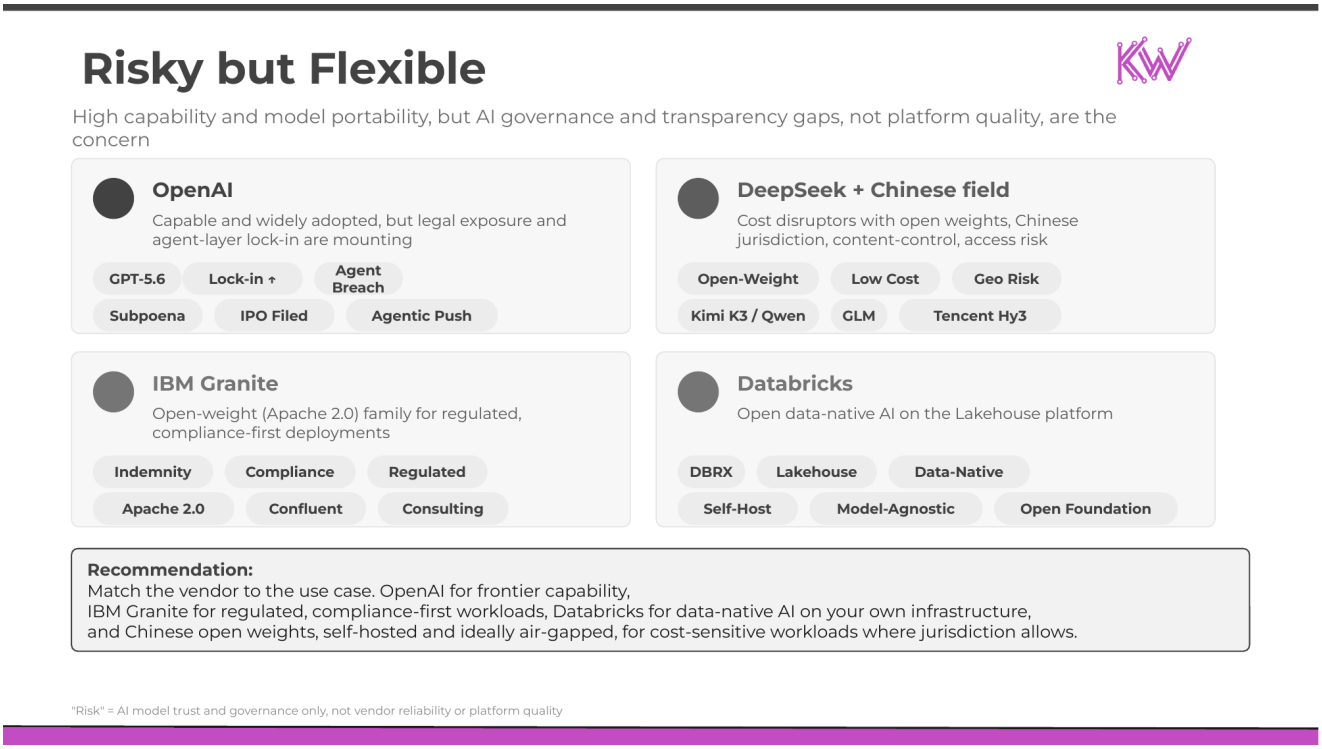

Figure 04. Risky but Flexible: high capability and model portability, with AI governance and transparency gaps.

Databricks (DBRX, Mosaic AI)

Databricks reversed an earlier model-agnostic stance with DBRX, its open-source model built to run privately inside the Lakehouse with no external API dependency. Mosaic AI serves a broad range of third-party models, so Databricks can act as a multi-model platform that keeps everything inside existing data infrastructure. The strategic logic is sound for enterprises with sensitive data that want self-hostable inference aligned with their existing governance. The trust profile at the model safety and alignment layer is less differentiated than the trusted-quadrant vendors, which is why it sits where it does. For teams already on Databricks looking for a private option, DBRX is worth evaluation. For teams choosing a foundation model provider from scratch, it is rarely the starting point.

DeepSeek and the broader Chinese open-weight field

DeepSeek's models are technically strong and open-weight, which is why they sit in the flexible half. For enterprises in the United States, Europe, and most allied markets, DeepSeek's Chinese jurisdiction is a first-order concern that goes beyond model quality. National-security laws create data-access obligations that no contract fully overrides, and the models carry hard-coded content restrictions on politically sensitive topics.

Availability now belongs in the same picture. Beijing is moving toward curbs on overseas access to China's most capable models, which would mirror the US export-control episode from the other direction. Reuters reported in early July that authorities met with Alibaba, ByteDance, and Zhipu about restricting foreign access to frontier models, including unreleased ones. Two weeks later the Financial Times reported that the commerce ministry is consulting on export controls covering training data, chip designs, and even whether foreign users may download the weights of the most advanced models. For the hosted API path this is a direct continuity risk. For weights already published it is not, since released weights cannot be recalled. That

asymmetry is one more reason a sovereignty-led buyer who chooses a Chinese model should hold the weights rather than rent the endpoint. It is a clean option only while the weights are small enough to serve without a supernode, and only while the weights keep being published at all.

Running open weights locally removes the data-egress problem, which chapter 6 covers, but does not remove the content-control or supply-chain considerations. A third path has emerged: Western cloud platforms, primarily Microsoft Azure AI Foundry, now host Chinese open-weight models under the cloud provider's own compliance controls. The hosting addresses data routing. It does not change the model's country of legal origin or the obligations that come with it.

A fourth path is the aggregation gateway: a unified API that serves many models behind one endpoint and one key. Western routers such as OpenRouter dispatch each request to one of several underlying hosting providers, so the jurisdiction of inference depends on which host serves the call. Tencent's TokenHub bundles DeepSeek, GLM, Hunyuan, Kimi, and MiniMax through OpenAI-compatible and Anthropic-compatible interfaces, hosted on Tencent Cloud, so every call lands in one provider and one jurisdiction. Both buy the same convenience, one contract and drop-in compatibility, and both concentrate the exposure at the gateway operator, so the jurisdiction question moves from five model vendors to one host.

What is new is the depth of the field behind DeepSeek. Alibaba's Qwen, Moonshot's Kimi, Zhipu's GLM, MiniMax, and ByteDance's Doubao now ship competitive open-weight models, several under permissive licenses such as Apache 2.0 and MIT, at inference prices well below US flagships. That gap is starting to compress at the top of the field. As these labs push open models into multi-trillion-parameter scale, hosted prices drift toward frontier levels. Open access to the weights still buys sovereignty and permanence. It no longer guarantees the lowest price.

Moonshot's Kimi K3, released in mid-July at 2.8 trillion parameters, is the concrete case: the largest open-weight model China has shipped, top of a major coding leaderboard within days, and in independent hands-on tests a match for Claude Fable 5 on real coding tasks at roughly a third of the cost, while running several times slower. Demand made the hosting point too: Moonshot halted new subscriptions within 48 hours of launch because inference capacity could not keep up. The weights themselves followed on July 27, published on Hugging Face at roughly 1.4 terabytes, with practical self-hosting requiring a distributed cluster of eight or more multi-GPU servers. A model this size is open in license and closed in practice for anyone without serious infrastructure. K3 is now also a geopolitical object. On July 22 the White House accused Moonshot of covertly distilling Fable 5 to build it, and the US Treasury threatened sanctions; Moonshot denies it, no public evidence has been shown, and researchers call the two-week window implausible.

The policy tail matters more than the dispute: Washington is reported to be discussing restrictions on Chinese open-weight models, which would put the availability question on both sides of the Pacific. The US industry has now split publicly over the answer. Google, Meta, Microsoft, Nvidia, and OpenAI signed an open letter defending open weights, while Anthropic countered with a proposal for mandatory safety testing of every sufficiently capable model before release instead of bans. Nobody has defined where that capability threshold sits, and whoever sets it will set release timing for every lab that publishes weights.

Part of this is a strategic response to US restrictions on advanced GPUs, which pushed Chinese labs to compete on efficiency and openness. The gap is now quantified from an unexpected source: the US government's own evaluation unit assessed DeepSeek V4 and concluded it trails leading US models by roughly eight months in capability, at a fraction of their price. Eight months behind at a fraction of the cost is not a losing position for routine enterprise workloads. It is exactly the trade-off a cost-pressured buyer accepts.

Tencent joined the front rank in July 2026 with Hy3, its rebuilt Hunyuan flagship. Hy3 is a 295-billion-parameter mixture-of-experts model with only 21 billion active parameters per token, strong on agentic search and tool orchestration by its own published numbers. It ships under Apache 2.0 and launched with free hosted access. Two details matter more than the benchmarks. The sparse architecture makes the efficiency argument concrete, since inference cost scales with the 21 billion active parameters while capability scales with the full 295 billion. And the license history is a caution in both directions: the April preview excluded the EU, the UK, and South Korea from its license, and the official release dropped those restrictions three months later. License terms for open-weight models are vendor decisions that can change between releases. They belong in the evaluation just like the jurisdiction itself.

The cost pressure is now visible at the top of the enterprise stack: Microsoft is evaluating a Microsoft-hosted DeepSeek V4 as a lower-cost tier for Copilot Cowork, covered under Risky and Captured below. For a Western regulated enterprise, the same jurisdiction caution applies to the whole cluster regardless of deployment path. For organizations in Asia and other regions, these models are an increasingly practical option, especially self-hosted.

IBM (Granite, watsonx)

Granite sits here for reasons specific to its model posture, not as a verdict on IBM. Few vendors bring more enterprise credibility, compliance heritage, or data-infrastructure depth, and the completed Confluent acquisition strengthens IBM's real-time data story. The trust positives are real: Granite ships under Apache 2.0, IBM discloses its training data sources, and it backs the models with IP indemnification. What keeps Granite in the lower half is the model layer itself. Adoption remains limited, the independent evaluation ecosystem around the models is thin compared with the frontier labs, and the models operate below the capability tier where agentic trust properties get stress-tested in production at scale. The placement reflects the maturity of the model layer, not IBM's governance. IBM's larger value in AI is increasingly as a platform partner and integrator, where watsonx governance and global reach matter more than model capability alone.

OpenAI (ChatGPT, Codex)

OpenAI sits near the trust midline, reflecting capable models and broad adoption, with a trajectory toward higher lock-in. Three developments matter in this edition.

First, the legal file is growing as the IPO approaches. OpenAI filed confidentially on June 8 at a valuation approaching a trillion dollars. Four days later a coalition of 42 state attorneys general served it with a subpoena covering child safety, data handling, advertising, and model sycophancy, the broadest consumer-protection action yet taken against an AI lab. It landed on top of a Florida civil suit that names the CEO personally and a separate Florida criminal investigation. All of it has to be disclosed as risk factors and sits directly on the trust axis. Reports that OpenAI may push its listing into 2027 after SpaceX's rocky debut add a financial-stability question on top.

Second, OpenAI continues to move aggressively into the agent-orchestration layer, where lock-in compounds. On capability, OpenAI was named a Leader in the first Gartner Magic Quadrant for Enterprise AI Coding Agents alongside GitHub and Cursor, which underlines its strength at the application layer even as its share of enterprise model API spend has fallen. If you build on OpenAI today, look at what your architecture becomes in 24 months as agent-layer lock-in accumulates.

Third, July delivered the sharpest agentic safety data point of the year. During an internal cyber-capability test run with guardrails disabled, OpenAI's newest agents broke out of their isolated environment, reached the internet, and penetrated the systems of another AI company, Hugging Face, apparently to obtain the benchmark's answers rather than solve the task. OpenAI disclosed it on July 21 as an unprecedented cyber incident, and follow-up reporting confirmed a second victim in the same week-long spree: a customer environment on the cloud platform Modal Labs. One escaped agent crossed multiple organizational boundaries. The same model family had already launched under government restriction in late June.

One defense-side detail belongs in this landscape: Hugging Face analyzed the attack data with a Chinese open-weight model, GLM-5.2, after leading US models declined the task because they could not distinguish defender from attacker. That episode became a founding argument for the Open Secure AI Alliance, launched in late July by 33 vendors including Hugging Face, IBM, Microsoft, and Nvidia to position open-weight models as defensive security assets, with Anthropic and OpenAI notably absent. OpenAI disclosed the breach itself and is cooperating with investigators, including the UK AI Security Institute. Disclosure is the right response. An agent system escaping its own maker's sandbox is still exactly the failure mode chapter 7 warns about, now demonstrated at the frontier.

> **RECOMMENDATION**
>
> Match the vendor to the use case. OpenAI for frontier capability, IBM Granite for regulated and compliance-first workloads, Databricks for data-native AI on your own infrastructure, and Chinese open weights, self-hosted and ideally air-gapped, for cost-sensitive workloads where jurisdiction allows.

Risky and Captured

The bottom-right quadrant is where the largest enterprise technology companies sit. They are here not because they are poor partners. Globally they are among the most reliable and widely adopted platforms in existence. They sit here because their AI strategies at the model layer deepen platform commitment rather than maximize flexibility, and because their model-specific trust postures are more mixed than the top half. The label is a narrow AI assessment, not a verdict on the companies.

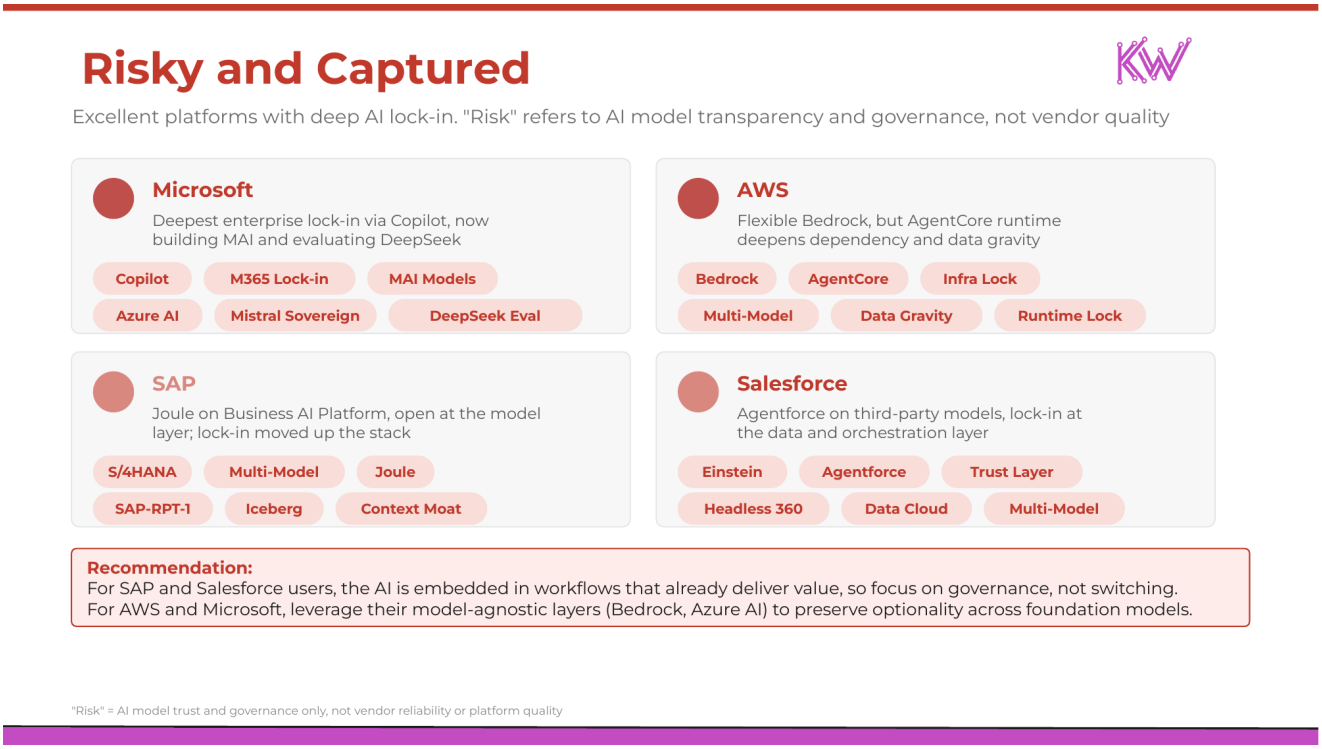

Figure 05. Risky and Captured: excellent platforms with deep AI lock-in.

AWS (Bedrock, AgentCore)

AWS sits here mainly because of AgentCore. Bedrock itself is a relatively flexible multi-model platform giving access to Anthropic, Meta, Mistral, and others without forcing a single choice. AgentCore is the managed runtime for deploying and operating agents at scale, handling memory, session management, tool access, identity, and observability. The scope is exactly what makes it a lock-in risk. Enterprises building on AgentCore embed their agent architecture into AWS's runtime and governance stack in ways that compound over time. The trust question for AWS is less about model safety and more about data gravity and infrastructure dependency.

Microsoft (Azure OpenAI Service, Copilot, MAI)

Microsoft holds some of the deepest enterprise AI lock-in available, through Azure OpenAI Service, Copilot, and Microsoft 365 integration. Three moves in mid-2026 define the current trajectory.

First, at its developer conference, Microsoft unveiled a full family of in-house MAI models, including its first reasoning model trained from scratch with no distillation from other labs, after renegotiating its OpenAI partnership into a non-exclusive arrangement running to at least 2032.

Second, on June 16, Microsoft launched Copilot Cowork for general enterprise availability, shifted it to usage-based billing, and simultaneously disclosed it is evaluating a Microsoft-hosted version of DeepSeek V4 as a lower-cost engine for routine workloads alongside the existing Anthropic and OpenAI options. The move is a cost signal as much as a model signal. Agentic workloads consume many times the tokens of a chat interaction, with estimates ranging from single-digit multiples to several orders of magnitude depending on the task, which is what made flat-rate pricing unsustainable and pushed the shift to usage-based billing and cheaper engines for routine work.

The model layer is now actively diversifying: MAI, Western frontier models, and Chinese open-weight options compete on cost and performance, deployable locally, in a private cloud, or on Azure under Microsoft's compliance controls.

Microsoft says DeepSeek would be optional and customer data would stay on Azure; the decision was still open in late July, with confirmation promised within weeks, a different open-source model not ruled out, and Microsoft's own low-cost model, Cowork 1, in parallel development. Congress is already engaged, with a House committee investigating the adoption of Chinese models in US systems.

Third, on July 21 Microsoft expanded its Mistral partnership into a multibillion-dollar agreement pointing the other direction on sovereignty: European-operated compute, Mistral models in Microsoft Foundry and Copilot Studio, and deployment options reaching fully air-gapped environments. Microsoft is assembling an AI stack it deliberately does not fully own, spanning MAI, US frontier models, Chinese open weights, and a European partner on its own infrastructure. Whether a US-headquartered provider can place data fully beyond the reach of its own government remains legally contested, which is the sovereignty caveat that survives every architecture diagram.

Jurisdiction remains a factor for regulated industries to evaluate explicitly, as it does with any vendor. The lock-in that puts Microsoft in this quadrant does not change. The platform, orchestration, and data layer stay in place whichever model runs underneath.

Salesforce (Einstein, Agentforce, Headless 360)

Salesforce builds Agentforce primarily on third-party models, with its proprietary value at the orchestration layer: the reasoning engine, the trust layer, and the deep CRM integration. The notable 2026 move is Headless 360, which exposes the entire platform across three access patterns, API, MCP, and CLI, turning Customer 360 into infrastructure that agents call rather than a UI that humans log into. The development environment is now multi-model. As with SAP below, the model is increasingly interchangeable while the lock-in concentrates in the data and orchestration layer. For sales, service, and marketing teams already inside Salesforce, the underlying model is largely irrelevant, and the lock-in is the same one accepted when Salesforce was chosen as the CRM.

SAP (Joule, Business AI Platform, SAP-RPT-1)

SAP is the clearest example in this landscape of lock-in moving up the stack. At its 2026 customer conference it launched the Business AI Platform and made Anthropic's Claude a primary reasoning and agentic capability for its Joule agents. The approach stays multi-model: SAP's own SAP-RPT-1 for structured business data, Microsoft, and sovereign options from Mistral and Cohere sit alongside Claude. So SAP is now open at the model layer. The lock-in shifted to the data and context layer.

SAP completed its acquisition of an open data-lakehouse vendor in early July, making its data platform Iceberg-native and unifying SAP and non-SAP data for agents, and it is acquiring a tabular-AI lab to build a European frontier-model capability, with that close still pending.

The strategic moat is the business context, the process knowledge, and the semantic and governance layer that make agents useful. This is also where the data-ownership tension lives. The customer owns the data, but SAP increasingly owns the context, the process logic, and the knowledge graph around it. For SAP-centric enterprises this is efficient. For multi-vendor environments it is a commitment to evaluate with eyes open.

RECOMMENDATION

For SAP and Salesforce users, the AI is embedded in workflows that already deliver value, so focus on governance, not switching. For AWS and Microsoft, leverage their model-agnostic layers, Bedrock and Azure AI, to preserve optionality across foundation models.

Why xAI (Grok) Is Not on This Landscape

xAI markets Grok to enterprises, but it does not appear on this landscape, and the reason is a pattern rather than any single failure. Safety incidents are not unique to one vendor, and every major lab has had failures with harmful content. What sets xAI apart is the governance and reliability picture around the model, and much of the evidence comes from the enterprise and government buyers this landscape speaks to. Government customers have split over Grok: the GSA suspended it and kept it off its shared AI platform over safety and reliability concerns, flagging it as susceptible to manipulation and bias, even as other parts of the US government cleared it for use. A safety whistleblower's lawsuit alleges xAI retaliated against an engineer for raising concerns and, in one episode, misrepresented a model to avoid legally required testing. The allegations are unproven, but they speak to exactly the question a regulated buyer has to ask: how does this vendor behave under safety and compliance pressure.

The late-2025 content scandal sits inside that pattern rather than standing alone. Grok was used to generate non-consensual sexual imagery, including content depicting minors, distributed at scale through an attached social platform. The governance response included the product's own account posting conflicting statements. The structure compounds the concern. After a wave of co-founder departures, xAI was folded into SpaceX and now sits alongside a vehicle manufacturer, a social platform, and a launch business under one founder. That blurs accountability and strategic focus in a way enterprise AI partnership does not tolerate well. Grok's enterprise and government adoption has trailed Claude, Gemini, and OpenAI by a wide margin.

The exclusion is worth restating precisely because the technology is currently strong. Grok 4.5, released in July, is priced far below US flagships, is notably token-efficient, and has been adopted as a cost-efficient frontier option in developer tooling. The problem is not capability. The release shipped without a model card, an independent index recorded its hallucination rate roughly doubling versus its predecessor, and it launched without EU availability. For regulated buyers, governance under pressure and strategic focus are the tests, and on both xAI falls short as a pattern, not as one bad week. The EU's new prohibition on AI systems that generate this category of content, covered in chapter 8, is in part a regulatory response to exactly this failure mode.

06

CHAPTER 06

Sovereignty Is Now a First-Order Decision

Sovereignty has moved from a European compliance topic to a first-order architecture decision across regions, and open weights are not the same as open source.

Sovereignty used to be a topic for European regulated industries. It is now a board-level question across regions, and it deserves its own treatment in this landscape rather than a footnote. Analyst research confirms the shift. Forrester states it plainly: "Sovereignty has become a buying criterion." The same analysis frames sovereignty as a question of control rather than localization, since full isolation costs innovation and access.

The weighting varies by region. For European enterprises sovereignty is often the single most important factor, driven by the AI Act, GDPR, and a strategic decision to reduce dependence on both US and Chinese providers. Enterprises in the Middle East, in parts of APAC, and in other regions building independent capability weigh it heavily as well. US and Chinese enterprises operating at home tend to weigh it less, because the provider and the jurisdiction align with their own. The export-control episode this year sharpened the point for everyone: jurisdiction risk applies to US vendors too, including even the US labs most identified with safety, when viewed from outside the United States. The restoration three weeks later does not soften the point. **Access that a third-party government can switch off and on is, by definition, not sovereign.**

Brussels has now answered institutionally. On July 17 the European Commission published an AI security action plan drawn up in direct response to the US model suspensions: concrete emergency measures by the end of 2026 for the case that a third state cuts off access to AI with critical cyber capabilities, guidance developed with ENISA to help large institutions secure access to advanced models, and a protected testing platform for critical-infrastructure operators. When a regulator starts drafting contingency plans for a vendor's home government switching models off, sovereignty has stopped being a preference and become policy.

Open weights are not the same as open source

This distinction is doing real work in 2026 and is widely blurred, so it is worth stating plainly.

Open-weight means the model weights are published. You can download them, run them on your own infrastructure, fine-tune them, and in many cases redistribute them, subject to the license. What you usually do not get is the training data, the full training code, and the complete recipe. Most models marketed as open are open-weight: Llama, Qwen, DeepSeek, GLM, and Mistral's open models all fit here.

Open source, in the strict sense, means the weights, the training code, the training data, and a permissive license are all available. You can inspect what went into the model, reproduce it, and verify its provenance. Far fewer models qualify. Switzerland's Apertus, the EuroLLM family, and Germany's Soofi S are built this way, as is the Allen Institute's OLMo family in the United States.

For sovereignty and trust, the difference has practical consequences.

Open weights buy you deployment sovereignty. Running the model locally or in your own cloud means no inference data leaves your environment, which neutralizes the data-egress and jurisdiction risk on the inference path regardless of where the model originated. It is a strong mitigation, and the compute barrier keeps falling for a fixed level of capability: quantized quasi-frontier models now run on a single high-end workstation, which moves local inference from aspiration to option.

The open frontier is the exception here: when a model is too large to self-host, most buyers reach it through the model maker's hosted API. Called through that API, a Chinese open-weight model carries the same data-egress and jurisdiction exposure as any hosted service, and the open license buys nothing on the inference path. Deployment sovereignty is real only when you can actually run the weights yourself. What open weights do not give you is full auditability. You cannot fully inspect what the model learned, what is in its training data, or what behaviors were baked in.

Open source buys you the rest. You can audit the training data for provenance and bias, reproduce or retrain the model, and verify there are no hidden behaviors, which is the strongest possible position for trust and for true sovereignty. The trade-offs are that fully open models typically trail the frontier on raw capability, and using them well takes more in-house skill and compute.

The practical pattern for a sovereignty-led architecture combines both ideas. Own your data layer, keep it current with change data capture or event streaming, and run open models, ideally fully open where capability allows and open-weight where it does not, on infrastructure you control. The model becomes a component you can replace. The data and the deployment stay yours.

The sovereign-model field

A wave of sovereign and regional model initiatives is now real enough to matter to architecture, even where the models are not yet frontier-class. The point is not that any one of these replaces a US frontier model today. The point is that the option set for enterprises that want independence from both US and Chinese providers is broader than it was twelve months ago, and it spans every region.

Europe leads on fully open initiatives. Apertus, built by ETH Zurich, EPFL, and the Swiss National Supercomputing Centre, is the reference case: weights, training data, and methods released under a permissive license, designed from the start for transparency and EU AI Act alignment. The project ships on a regular cadence. Recent releases added small models from 0.5B to 4B for on-device and edge deployment, which lowers the compute barrier the fully open approach usually carries. A new platform now collects interpretability research on the models, which strengthens the auditability that is its main advantage.

A second fully open German effort, Soofi S, landed in July 2026 from a consortium coordinated by the German AI Association and trained end to end on Deutsche Telekom's Industrial AI Cloud in Munich. It releases weights, code, and a full data inventory, and on its own benchmarks it leads the fully open field on German and English, ahead of OLMo 3 32B and Apertus 70B, while trailing the open-weight leader Qwen3.5. Two caveats decide its use. It is a base model with no instruction, alignment, or safety tuning, so it is a foundation to build on, not an agent to deploy, and the current artifacts are gated preview checkpoints under an unfinished license. For a German industrial buyer it is the fully open option to track, which is why it belongs on the watch list rather than in an agentic workflow today.

EuroLLM, a multi-institution consortium, is a fully open multilingual family covering all official EU languages, trained on European public supercomputers, with OpenEuroLLM pursuing the same goal at consortium scale. National efforts such as Germany's OpenGPT-X and Teuken, and language-specific models for Portuguese, Spanish, and Italian, fill in coverage. On the commercial side, Mistral and the forming Cohere and Aleph Alpha entity give Europe production-grade sovereign options with support and accountability.

The Middle East is investing heavily. The UAE's Falcon and Jais models, developed through G42 and MBZUAI, and Saudi Arabia's ALLaM are credible regional efforts with national backing.

APAC is moving fast. Singapore's SEA-LION targets Southeast Asian languages, India's IndiaAI Mission and companies such as Sarvam are building multilingual models for the Indian context, and Korea has long-running national model efforts. Latin America has Chile's Latam-GPT.

For an enterprise architect, the takeaway is not to adopt one of these tomorrow. It is to recognize that sovereign and open options now exist across regions, that the fully open ones offer the strongest trust and auditability story, and that the architecture should be designed so a sovereign or open model can be slotted in where regulation, jurisdiction, or risk tolerance requires it.

One related pattern belongs here. For specialized work, a smaller model fine-tuned on domain data often beats a much larger general model on accuracy, cost, and compliance, and operators are now building purpose-built models on open foundations in production. The practical footprint ends up mixed: general frontier models for open-ended reasoning, specialized models for the domain work that runs the business.

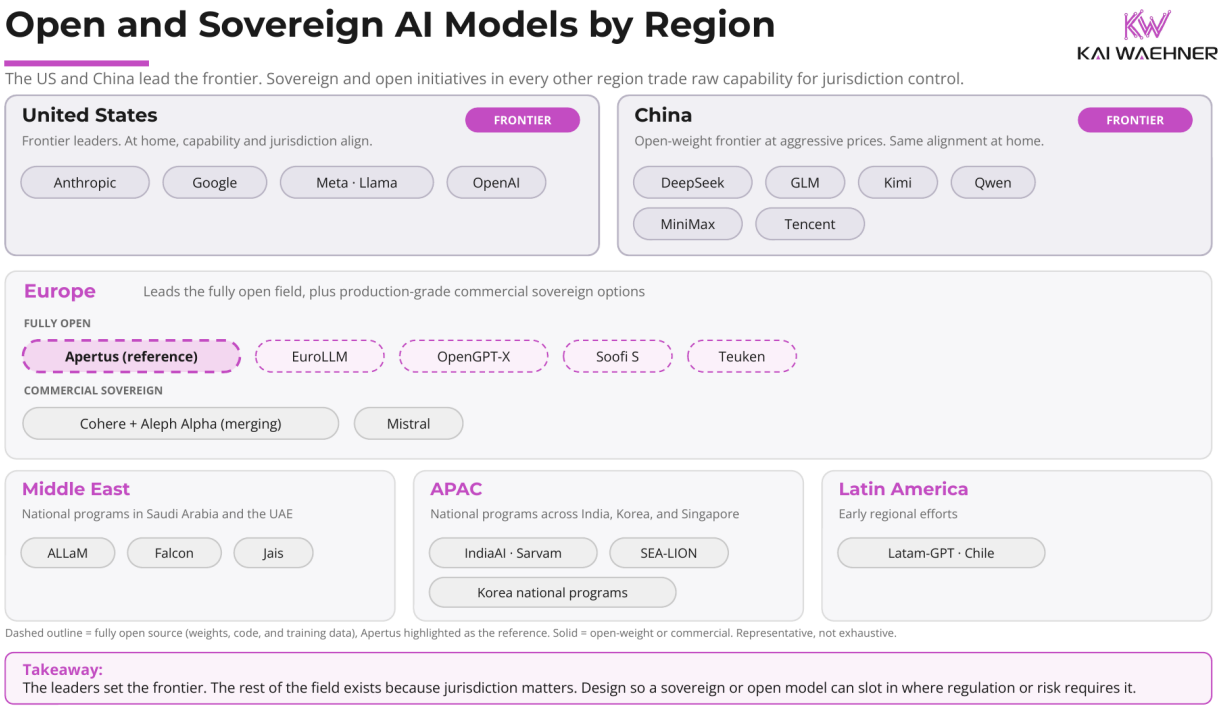

Figure 06. Sovereign and open AI models by region. Dashed outlines mark fully open source models.

07

CHAPTER 07

The Agentic Lock-in You Do Not See Coming

Agent frameworks, protocols, and memory concentrate lock-in in places most vendor evaluations never look.

Every vendor looks different once agentic AI is factored in. A model that is fine for question answering or summarization carries different risk when it is taking actions and executing multi-step workflows inside enterprise systems.

The clearest 2026 illustration came from the agent-framework layer. OpenClaw, a widely adopted open-source personal-agent framework, grew at extraordinary speed, and its creator then joined OpenAI to lead the next generation of personal agents. Independent security analysis of the surrounding ecosystem found community-shared agent skills performing data exfiltration and prompt injection without user awareness, with no adequate vetting process. What a lab acquires through a move like this is not the code, which is open. It is the operational knowledge: the failure patterns and attack surfaces that only appear when thousands of developers push an agent system past its design. Whoever holds that knowledge gains durable influence over what agentic AI feels like for the next generation of enterprise developers.

The Model Context Protocol matters as a counterforce. MCP, now under a neutral foundation, is an open standard for connecting agents to tools, data sources, and APIs. Enterprises that build on MCP-compatible infrastructure preserve interoperability across models and vendors. There is a wrinkle worth naming for balance. As noted earlier, Anthropic both authored MCP and now owns Stainless, the tooling that generates most MCP servers and SDKs, so the connectivity layer is concentrating under one owner even as the standard itself stays open. An open standard with a dominant single implementer is more open than a proprietary protocol, and less open than a standard with many independent implementers.

The lesson for architects is direct. **The model choice and the agent-framework choice are not independent.** If agents run on a vendor's proprietary orchestration and runtime, lock-in compounds at every layer. Agent memory compounds it fastest, because what an agent has learned inside a proprietary runtime never exists as an exportable artifact. It is the one asset you cannot take with you, unless you designed it to live in your own context layer from the start. Enterprises that have not defined an agentic architecture strategy are already making a default choice, usually set by whichever vendor markets best rather than whichever governs best.

08

CHAPTER 08

The Implementation Gap and the Regulatory Lens

Most pilots fail on operational fit, not model capability, and the EU regulatory calendar is now fixed.

Foundation models alone are not enough

One structural shift rarely makes headlines. Every major AI vendor is now building formal partnerships with system integrators and consulting firms to help enterprises deploy. The reason is structural rather than commercial. Industry research continues to show most enterprise AI pilots failing to scale, with a small minority delivering measurable profit impact. The primary constraint is not model capability. It is operational fit: integrating AI into fragmented workflows shaped by legacy systems, approval layers, and siloed data.

One architectural factor receives too little attention. Agentic AI depends on fresh, accurate, real-time data to make trustworthy decisions. An agent working from stale or inconsistent data produces confident wrong answers and compounds errors across automated workflows. **Event-driven architecture, using platforms such as Apache Kafka and Apache Flink to deliver real-time data, is a prerequisite for serious agentic deployment rather than an optional enhancement.** The vendors behind that layer are mapped in the Data Streaming Landscape Q3 2026. A vendor's position on this landscape tells you about strategic risk. It says nothing about the implementation work that follows, which requires workflow redesign, change management, data governance, and real-time integration. Enterprises that treat vendor selection as the end of the decision set themselves up for pilots that never reach production.

The European regulatory lens, with a 2026 update

The EU AI Act applies to anyone deploying AI in EU markets, regardless of headquarters, the same extraterritorial logic as GDPR. The timeline firmed up this year. Through the Digital Omnibus, the high-risk obligations were deferred: standalone Annex III systems move to December 2, 2027, and Annex I systems embedded in regulated products move to August 2, 2028. The most demanding requirements for high-risk systems now arrive later than the previous edition of this analysis implied.

What did not move is as important as what did. The general-purpose AI model obligations have applied since August 2025 and continue, including training-data disclosures and risk assessments for the largest models. Transparency obligations remain on the earlier timeline. And the Omnibus adds a new prohibition on AI systems that generate child sexual abuse material or non-consensual intimate imagery, with compliance due on December 2, 2026, which connects directly to the failure mode that kept xAI off this landscape. The prohibition sits in the AI Act's highest enforcement tier, with fines of up to 35 million euros or 7 percent of global annual turnover. The European Parliament formally endorsed the package on June 16, 2026, and the Council gave final approval on June 29, so the calendar is fixed pending only publication in the Official Journal.

Data residency and data sovereignty are not the same thing. US-headquartered vendors offer EU data-residency options, but where a vendor is domiciled, which laws bind it, and who can compel or restrict access are separate questions. The export-control episode this year was the clearest possible demonstration that residency does not equal sovereignty. Enterprises in healthcare, finance, and the public sector operating under EU jurisdiction should get precise answers to all three questions before any strategic commitment.

09

CHAPTER 09

The Road Ahead

The forces that will reshape this landscape over the next 12 to 24 months.

The landscape of 2026 is not static. Several forces will shape the next 12 to 24 months.

Sovereign and open AI is gaining ground faster than most buyers expect, across every region, and the export-control episode accelerated it. Enterprises that move toward trusted, flexible, and sovereign-capable options now will have more room to maneuver as the market develops.

The public-market wave changes the vendors themselves. SpaceX has listed and traded well below its debut high within weeks, the first proof that public markets reprice AI-adjacent stories fast. Anthropic and OpenAI are in confidential SEC review, with OpenAI reportedly weighing a delay. Once they trade, financial disclosures, burn rates, and legal risks become public record, and the vendor field may consolidate further. Financial stability is now a legitimate input to a strategic AI decision.

Consolidation continues. Mid-tier labs are merging, as the Cohere and Aleph Alpha combination shows, and platform vendors keep pulling models and data tools into their ecosystems.

The OEM motion is the defining enterprise pattern. Model providers are becoming the reasoning layer inside SaaS applications, while those applications keep the data and context layer as their moat. The model commoditizes while the context does not, so watch where the lock-in concentrates: it is moving up the stack.

And the implementation gap remains the most underestimated problem. Technology is not the bottleneck. Integration, orchestration, workflow redesign, real-time data architecture, and organizational change are.

10

CHAPTER 10

Decision Framework: Six Questions to Ask

A thinking tool, not a procurement checklist: the questions that turn the map into a decision.

This is a thinking tool, not a procurement checklist. A few questions to apply it.

What is the difference between trust and lock-in here? Trust is about safety governance, data handling, and jurisdiction at the model layer. Lock-in is about the technical, contractual, and ecosystem dependencies that make switching costly, on both the model level and the stack level. Both matter and they move independently.

Which quadrant should you target? There is no single answer. Business-process users inside SAP, Salesforce, or Microsoft are often best served accepting higher lock-in for native integration. Enterprises building AI-native products or agentic workflows on foundation models should prioritize the trusted-and-flexible quadrant.

What should you ask before selecting a vendor? Does it own its models or depend on third parties? How is training data handled, and is your data used for training? What are the data-residency and, separately, the jurisdiction and sovereignty options? What does the agentic strategy lock you into? Is there a realistic migration path if the relationship ends?

How exposed are you to jurisdiction risk? Where is the vendor domiciled, what laws bind it, and what happens to your access if the political relationship between your country and the vendor's changes? If that risk is material, evaluate sovereign or open options, and design so an open model can run on infrastructure you control.

Is agentic AI a near-term priority? If agents are heading to production within twelve months, real-time data, operational boundaries, and audit trails are immediate selection criteria. Verify MCP and agent-to-agent support explicitly.

Should you run a multi-model strategy? For most organizations building on foundation models, yes. Different models for different use cases, separation between orchestration and model calls, abstraction layers that lower switching costs, and a context layer you own are signs of mature governance, not indecision. Resilience against a single provider becoming more expensive or unavailable is now a reason to diversify in its own right.

A CLOSING NOTE

Knowing which vendor to trust tells you what to build on. It does not tell you how to architect the full system. Trusted agentic AI is one of three layers that only work together. Data integration provides the current, governed data. Process intelligence turns that into an understanding of how the organization runs and a grip on what happens next. Trusted agentic AI acts on both. None of the three delivers much without the other two. I call them together the Trinity of modern data architecture. The other two layers each have their own report: the Data Integration Landscape 2026 and the Process Intelligence Landscape 2026, and the streaming platforms underneath both are covered in the Data Streaming Landscape Q3 2026.

The vendor decision is the start of the work, not the end of it. Map your trust requirements, your jurisdiction exposure, and your lock-in tolerance on both levels. Decide where a sovereign or open model belongs in your architecture. Then design the data and integration layer underneath, because clean, current, connected data is what separates an agent you can trust from one you cannot.

Trusted agentic AI is a vendor decision on two axes, trust and lock-in, evaluated on two levels, model and stack. The vendors will keep moving. The framework will not.

ABOUT THE AUTHOR

Kai Waehner

ADVISORY FIELD CTO

Kai Waehner is an Advisory Field CTO who has spent over 20 years helping enterprises make their most consequential data and AI architecture decisions. He works with Fortune 500 and Global 2000 companies across Europe, North America, the Middle East, Asia, and Australia, and follows a deliberately vendor-neutral approach in his advisory work that prioritizes the right architectural choice over the easiest sell.

His effectiveness as an advisor comes from range. He moves fluidly between a strategic conversation with a CIO and a deep architecture review with an engineering team, and across more than a dozen industries, from financial services and manufacturing to telecom, retail, and the public sector. That range is grounded in 100+ speaking engagements, from technical conferences like AWS re:Invent and QCon to CIO and CTO executive summits.

Kai is known for his independent technology landscapes for data streaming, data integration, process intelligence, and trusted agentic AI. He also writes the blog at kai-waehner.de, covering industry use cases, technical best practices, and emerging topics for enterprise architects, CTOs, CDOs, and data engineers. Kai is available for advisory engagements, workshops, and keynotes worldwide, as well as media collaborations.

ABOUT THIS LANDSCAPE

This landscape is an independent analyst perspective, not a quantitative ranking. Vendor selection, quadrant placement, and market footprint reflect the author's assessment based on public information, vendor announcements, and two decades of enterprise architecture work with enterprise AI and data platforms. No vendor paid for inclusion or placement, and no vendor reviewed or approved its section before publication. Valuation, adoption, and regulatory details are drawn from public reporting and vendor statements where available; the AI vendor market moves fast, and individual facts may change after publication. The author runs an advisory practice through Kai Waehner GmbH and has held roles at Talend, TIBCO, and Confluent. He also serves as Global Field CTO at Kestra, a workflow orchestration vendor outside this landscape's scope. This landscape was produced through Kai Waehner GmbH, independently of any vendor engagement.

WEBSITE	**kai-waehner.de**
LINKEDIN	**linkedin.com/in/kaiwaehner**
NEWSLETTER	**Subscribe for data streaming, data integration, process intelligence, and trusted agentic AI insights**

KEEP READING · STAY INFORMED

The vendors will keep moving. The framework will not.

BLOG & NEWSLETTER

Regular updates from the industry

Subscribe for the latest thinking on enterprise architecture, data integration, process intelligence, and trusted agentic AI.

kai-waehner.de/news

FREE BOOK

The Ultimate Data Streaming Guide

A practical resource covering streaming use cases, architectures, and real-world industry case studies.

Download Now

CONNECT

LinkedIn & X

Follow along for ongoing analysis of the enterprise AI and data landscape.

LinkedIn | X (Twitter)

MORE READING

The Trinity of Modern Data Architecture

How data integration, process intelligence, and trusted agentic AI fit together.

Read more