
PERSPECTIVE

Model Access Is Not a Services Moat

Whitepaper

Bart de Graaff · BdG Advisory · October 2026

For two years, a frontier-model partnership was a reason to pick a services partner. Now the same models run on every major cloud, partner badges are open to any firm with a certified bench, and labs and clouds have committed billions to putting their own engineers inside clients. Follow what each party is paid for, and the buyer turns out to be the only one paid by the outcome.

The badge everyone has

Every large integrator now has a frontier-model partnership. That is exactly why none of them can sell one.

For two years, access to the best models was a credible reason to choose a services partner. That reason has expired. The same frontier models now run on every major cloud. The labs' partner programmes are open to any firm that certifies enough staff. And the labs and the clouds have started hiring thousands of engineers to sit inside clients themselves. Access has become a utility, and a utility cannot carry a services premium. What stays scarce is the capacity to redesign how an organisation makes decisions, to run AI in production under real controls, and to carry accountability for the result, while staying independent of whichever stack is funding the deal. That shifts the buyer's test. A model relationship tells you very little. What matters is whose income rises when your outcome improves, and whose rises when your consumption does.

This paper traces how the balance of power between systems integrators, model labs and hyperscale clouds is shifting, names what each party is actually paid for, and sets out what an enterprise should conclude about whom to trust with its AI estate.

Access became a utility

Start with the mechanism, because it explains why exclusivity collapsed so quickly.

A frontier lab has two problems that only scale solves. It needs enormous amounts of compute, and it needs distribution into enterprises that buy through the clouds they already use. A hyperscaler has the mirror-image problem. It needs the models its customers ask for, or those workloads, and the data gravity that follows them, move elsewhere. Each side is rational to want the other as widely as possible. The result is that every serious model ends up on every serious cloud.

The public record shows how fast this happened.

In November 2025, Microsoft, NVIDIA and Anthropic announced a partnership to bring Claude to Microsoft Foundry, which would make it, in Anthropic's words, the only frontier model available on all three of the world's most prominent clouds. Anthropic committed to buy \$30 billion of Azure compute in the same announcement.

A month earlier, Microsoft and OpenAI rewrote their agreement. Microsoft gave up its right of first refusal to be OpenAI's compute provider, and OpenAI gained the right to release open-weight models. In February 2026, Amazon announced a \$50 billion investment in OpenAI, with a stateful runtime built on OpenAI models to be offered through Amazon Bedrock. The lab once synonymous with one cloud now reaches customers through a second.

The models themselves are converging. Stanford's 2025 AI Index found that the cost of running a system at GPT-3.5 level fell more than 280-fold between November 2022 and October 2024. The gap between open-weight and closed models on some benchmarks shrank from 8% to 1.7% in a year. The score difference between the top-ranked and tenth-ranked models fell from 11.9% to 5.4%, and the top two were separated by 0.7%. The 2026 edition reports that US and Chinese models have traded the lead several times since early 2025. When the frontier is that crowded, the identity of the model matters less every quarter.

Then look at what a partner badge now certifies.

Anthropic launched its Claude Partner Network in March 2026 with \$100 million of support, free to join for organisations bringing Claude to market. By June, more than 40,000 firms had applied and more than 10,000 consultants had earned a Claude certification. The top services tier requires at least 1,000 certified professionals and 100 customers across three or more regions. That is a sensible quality bar for a lab to set. It is also a bar defined by bench size and exam passes, which any well-capitalised firm can clear with a training budget and time.

The same firms appear on every list. Capgemini, to take one public example, is named in OpenAI's Frontier Alliances, among the firms alongside which Google Cloud will embed its engineers, and as a founding partner of OpenAI's new deployment company. That is a rational hedge for an integrator, and it is the norm among the large ones. It also means the badge no longer distinguishes one shortlisted firm from another.

When every serious bidder holds the same credential, the credential becomes a hygiene factor. It qualifies a firm for the shortlist. It does not tell you who should win.

The labs and the clouds are walking towards the client

If access is a utility, where does the value go? It moves to the point where the model meets the organisation: the data, the tools, the controls and the decisions. Whoever builds that

connection learns where value sits, shapes the architecture, and sets the consumption path for years. In a commoditising model market, that is the most valuable position in the chain. So the labs and the clouds are moving towards it.

Look at what was announced between February and July 2026.

- **February.** OpenAI launched Frontier, a platform for building and managing enterprise agents, and Frontier Alliances with two strategy consultancies and two integrators, with OpenAI's own forward-deployed engineers working alongside them.
- **April.** Google Cloud committed a \$750 million fund to its 120,000-member partner ecosystem, including embedded Google forward-deployed engineers and usage incentives.
- **May.** Anthropic, with Blackstone, Hellman & Friedman and Goldman Sachs, announced a new enterprise AI services company that pairs Anthropic's applied AI engineers with its own team to serve mid-sized companies.
- **May.** OpenAI launched the OpenAI Deployment Company with more than \$4 billion from 19 partners, and agreed to acquire the consultancy Tomoro to bring about 150 forward-deployed engineers in from day one. The company is majority-owned and controlled by OpenAI.
- **June.** AWS committed \$1 billion to a new unit of embedded AI engineers, planned to number in the thousands and working in pods placed with customers for 45-day periods, according to Reuters.
- **July.** Microsoft announced Microsoft Frontier Company, a \$2.5 billion investment embedding 6,000 industry and engineering experts at customers to deploy and improve AI systems based on measurable business outcomes.

Two labs and three clouds committed capital and people to delivery in six months. None of these announcements says integrators are finished. Most of them explicitly include partners, and the labs say openly that demand exceeds what they can serve alone. But the direction is unmistakable. The parties that used to supply the integrator are now also standing next to it at the client.

The reasons are rational, and they are worth stating plainly.

A lab whose model is one of several near-equivalents needs to be embedded in the workflow before a competitor is. A forward-deployed engineer who builds an agent on a lab's platform makes that lab the default for the next ten agents. A cloud that has just committed tens of billions to data centres needs those data centres consumed, and nothing drives consumption like its own engineers designing the workload. Delivery is no longer a low-margin distraction for these firms. It is how they defend their position.

None of this makes their engineers bad advisers. Many of them will be the best platform engineers a client can get. It does make them conflicted ones. OpenAI says its deployment company will operate as an extension of OpenAI, keeping customers close to its research and product teams. That is an honest description, and a buyer should take it at face value. An

engineer employed by a lab will design for that lab's models. An engineer employed by a cloud will design for that cloud. Nobody should expect otherwise, and nobody should mistake it for neutral advice.

Notice also what Microsoft chose to promise. Its Frontier Company announcement states that a customer's data, IP and competitive advantage will not be used to train models in ways that commoditise what differentiates them. A vendor does not make that promise unless buyers are asking. The question underneath is the right one: when a vendor's engineers spend months inside your operations, who keeps what they learn?

Every party is paid by something other than your outcome

To decide whom to trust, follow what each party is paid for. The answer is uncomfortable for everyone, including the buyer.

The clouds are paid by consumption. Alphabet raised its 2026 capital expenditure guidance to \$195–205 billion in July 2026, and attributed the increase to the data centre capacity needed for AI workloads. Capital on that scale is recovered one inference call at a time. Every incentive inside a cloud, from account teams to partner funds, points towards more workload on that cloud. Google's partner fund lists usage incentives among its components. That is not hidden and not improper. It is the business model.

The labs are paid by tokens, and they owe the clouds for compute. OpenAI has contracted to buy an incremental \$250 billion of Azure services and has expanded its AWS agreement by \$100 billion over eight years on top of an existing \$38 billion. Anthropic has committed \$30 billion to Azure, announced access to up to one million Google TPUs in a deal worth tens of billions of dollars, and names AWS its primary training partner after \$8 billion of Amazon investment. Microsoft and NVIDIA have committed to invest up to \$5 billion and \$10 billion respectively in Anthropic.

Read those together and the structure is circular. Clouds invest in labs. Labs commit to buy cloud capacity. Clouds resell lab models to enterprises and fund partners to drive usage. Each link is rational. Each party is paid by volume moving through the chain.

The integrators are paid by hours, and the hours are deflating. The classic services model bills people against time. AI attacks exactly that unit. In August 2026, Reuters reported that clients were demanding the same work for 25% to 30% less, quoting the chief executive of one Indian services firm. Tata Consultancy Services said about 80% of contracts in its business services segment were now tied to outcome measures, double the share before AI went mainstream in late 2023. Its chief executive framed growth as a race to stay ahead of revenue deflation. The Nifty IT index had fallen by a fifth in 2026. A year earlier, the same firm had announced it would cut about 2% of its workforce, more than 12,000 roles.

That is not a story about one firm or one country. It is what happens to any business priced on labour when the labour content of the work falls. And it creates a quieter structural risk. As hourly economics weaken, the funding that flows from labs and clouds to their partners,

in co-investment, usage incentives, certifications and joint go-to-market, becomes a larger share of what makes an AI engagement attractive to the integrator. Nobody has to act in bad faith for that to tilt the recommendation. It only has to be how the deal is funded.

The pattern is easy to recognise. A regional bank wants to automate parts of its commercial lending review. Its preferred integrator arrives with a pilot that is largely funded: the cloud contributes credits, the lab contributes engineers, the integrator contributes a discounted team. The bank pays little and the demonstration is impressive. Nine months later, the bank discovers that the agent definitions, the retrieval layer and the monitoring all assume one model on one cloud, and that the business case for production was written on that stack's list prices. Nobody misled anyone. The architecture was simply decided by whoever paid for the pilot.

In the whole chain, only the buyer is paid by the outcome.

The cloud is paid when you consume. The lab is paid when you call its model. The integrator is paid when you buy hours, or when its partner funding is triggered by your consumption. The enterprise is the one party whose return depends on whether the decision got better, the claim was settled correctly, or the customer stayed.

That is the real structure of the AI services market. A buyer that ignores it will be advised, competently and in good faith, towards somebody else's volume.

What stays scarce

If access is abundant and every vendor wants to be in the room, what is a services position actually worth paying for? Five things, and none of them comes with a badge.

First: knowledge of how this organisation decides. A model does not know what a correct claims decision looks like in a particular insurer, which exceptions a credit committee tolerates, or which supplier deviations a plant manager will accept. That knowledge lives in people, policies and history. To make AI useful, someone has to turn it into data, evaluation sets and explicit controls. The evaluation set, the agreed test of what "right" means for a given decision, is the most underrated asset in enterprise AI. It cannot be brought in. It has to be built with the client, and it compounds.

Consider a pattern. A mid-sized insurer's claims team runs a promising pilot on a frontier model. Accuracy looks strong on the vendor's demonstration data. In production, the system mishandles the cases the claims handlers know are hard: partial liability, policy endorsements, repeat claimants. The fix is not a better model. It is a few hundred adjudicated cases, written down by senior handlers, used as the standing test any model has to pass before it touches a live claim. Whoever builds that test owns the insurer's definition of quality. That should be the insurer.

Second: the redesign of decision rights. AI fails at the boundary between technology and organisation, as the series' paper on transformation failure argues. Integrating a model with a

core system is hard engineering. Deciding who owns an automated decision, who can override it, and what happens to the roles around it is harder. That work is political, slow and specific to each organisation. A 45-day engineering pod is not designed to do it, and a lab's platform cannot do it.

Third: running it. The value of AI is captured in production, over years, not in the pilot. Someone has to monitor drift, own incidents at 2am, retest when a model version changes, retire agents that no longer earn their place, and swap the model when a better or cheaper one arrives. This is the least glamorous part of the estate and the most durable services position in it. It is also where the governance-first ratio in the 7:1 paper is either honoured or abandoned.

Fourth: accountability. Clients do not ultimately buy effort. They buy someone who will stand behind the result. When a deliverable produced with AI turns out to contain fabricated references or invented quotations, the model that produced the text is not the party that answers for it, and neither is the cloud that hosted it. The firm that signed the work does. Model access was never the scarce input. Review, judgement and the willingness to bear the consequences were. The contract mechanism that holds is a fee tied to an acceptable deliverable, not to the hours or tokens consumed.

Fifth: independence. The rarest asset in this market is a partner whose economics do not depend on which model or which cloud wins. Every force described above pushes against it. Labs want their models designed in. Clouds want their capacity consumed. Partner programmes reward concentration. A services firm that can genuinely swap a model, move a workload, and show the buyer the incentives it receives is offering something none of the platform owners can offer by construction.

Put those five together and the shape of a durable services position becomes clear. It rests on assets that are specific to the client, accumulate over time, and cannot be bought by certification: the client's decision logic, its operating model, its production estate, its liability and its freedom to switch. Access to models is an input to all five. It is no longer a reason to choose anyone.

This also answers the question of who will win among integrators, labs and clouds. There is no settled winner, and there will not be a single one. The labs and clouds will win the platform engineering they are best placed to do. Integrators that keep selling access and hours will be squeezed from both sides. Firms of any kind that organise around the five scarce assets, and price against outcomes rather than effort, will hold a position the platforms cannot easily take, because the platforms' own incentives prevent them from offering independence.

The strongest objection: the badge still buys something

The best counter-argument deserves a fair hearing, because parts of it are true.

It runs like this. Access is not yet a utility where it counts. Frontier capacity is constrained, and the labs ration their best engineers and their earliest models. Google's partner fund explicitly includes early model access for select partners. Microsoft's agreement with OpenAI still keeps API products developed with third parties exclusive to Azure. The labs themselves say they need partners: Anthropic's chief financial officer said enterprise demand for Claude is significantly outpacing any single delivery model. Partner programmes come with real money, \$100 million from Anthropic and \$750 million from Google, that subsidises pilots and training the buyer would otherwise fund. So the rational buyer should choose the best-connected integrator, because that firm brings priority, co-investment and a direct line to the lab.

All of that is true. Here is why it does not change the conclusion.

Priority decays fast. Early access to a model is a lead measured in weeks, in a market where the top two models are separated by less than a percentage point. A programme designed around a temporary head start is designed around the wrong asset. Priority matters most to a buyer who has already decided which stack to build on, which is exactly the decision a buyer should not delegate.

Funding comes with direction. Co-investment and usage incentives are not charity. They are paid by a specific lab or cloud, against adoption of that lab's or cloud's products. That is a fair commercial arrangement. It is also vendor financing, and vendor financing is priced, conditional and never neutral. A buyer should take it the way a sensible treasurer takes vendor financing: gladly, with full disclosure, and never as the reason for the purchase.

"We need partners" is a capacity statement, not a commitment. The same lab that said demand outpaces any single delivery model was, in the same announcement, launching a new delivery company. Its competitor launched a majority-owned one a week later. The labs need partners today because they cannot hire fast enough. Their capital allocation shows where they intend to be when they can. A buyer signing a multi-year programme should plan for the market the investments point to, not the one the press releases describe.

The best-connected firm is the most concentrated one. The deeper an integrator's alliance with one lab or cloud, the more of its economics depend on that party's volume. That can make it an excellent implementer of a stack the buyer has already chosen. It makes it a poor adviser on which stack to choose.

So the objection lands as a correction, not a rebuttal. Partner depth is useful for execution once the architecture is decided. It is a poor basis for deciding the architecture.

What buyers should change

The structural conclusion for an enterprise buyer is simple to state and harder to practise. Trust with the AI estate should go to the party whose economics are tied to the buyer's outcome and whose architecture keeps the buyer free to switch. In practice that is rarely one party. It is a deliberate division of roles. Here is the order in which to build it.

1. Own the architecture and the evaluation set before choosing anyone. Decide internally which assets the organisation will always own: the evaluation sets, the agent definitions and prompts, the decision logs, the integration layer, the data contracts. Design the model layer to be swappable from the first day. If a partner's proposal cannot work without a specific model or cloud, that is a finding, not a feature.

2. Separate advice from consumption. The party that designs the target architecture should not be paid by the consumption that architecture generates. Require every bidder to disclose its partner tiers, co-investment, usage incentives and resale margins tied to the proposal. This is ordinary procurement hygiene in other markets. Apply it here.

3. Contract for outcomes and for exit. Tie a meaningful share of fees to measured operational outcomes, defined by the buyer. Write exit in at the start: data export, portability of agents and evaluation sets, and transition support. Regulation is moving the buyer's way. The EU Data Act requires cloud providers to remove switching charges, including data egress fees, from 12 January 2027. The UK competition regulator's cloud market investigation found that competition was not working well and identified egress fees as a barrier to switching. Bring both into the negotiation now; do not wait for enforcement.

4. Buy the run, not the pilot. Select partners on their willingness to own production: monitoring, incident response, model retesting and replacement, over years. A vendor's embedded engineering pod is a valuable way to start. It is not a service model. Ask who will be accountable in month nineteen.

5. Keep a small internal team that can judge. Not a large AI department. A handful of people who can read an evaluation, challenge a vendor's benchmark, and run a bake-off between models on the organisation's own cases. Without that, the buyer cannot tell whether advice is good or merely confident.

6. Use each party for what its incentives make it good at. Use lab and cloud engineers for deep platform work on a stack you have chosen. Use integrators for integration, operating-model redesign and production operations, under outcome terms. Use independent advice for architecture and vendor choice. Make the division explicit, so no party is quietly doing a job its incentives conflict with.

Just as important is what to stop doing.

- **Stop selecting partners by badge or tier.** It tells you how many people passed an exam, not whether your claims will be settled correctly.
- **Stop accepting "free" pilots without an architecture review.** A funded pilot on a single stack is the first instalment of a dependency.
- **Stop letting a vendor's account team write the AI roadmap.** It will be a good roadmap for that vendor.
- **Stop signing multi-year commitments to one model family.** The frontier has changed hands repeatedly in two years. Your contracts should assume it will again.

Labs, clouds and integrators are each acting rationally. Each is protecting the thing it is paid for. The buyer's job is to arrange the work so that at least one party in the room is paid for the same thing the buyer is.

A partner badge can open a conversation. It cannot carry your transformation.

What to do about it

- **Own the architecture and the evaluation set.** Decide which assets you always own, from evaluation sets and agent definitions to decision logs and integration, and keep the model layer swappable from day one.
- **Separate advice from consumption.** Whoever designs the architecture should not be paid by the consumption it generates. Require disclosure of partner tiers, co-investment and usage incentives.
- **Contract for outcomes and for exit.** Tie a meaningful share of fees to outcomes the buyer defines, and write portability and transition support into the contract at the start.
- **Buy the run, not the pilot.** Select partners on their willingness to own monitoring, incidents, retesting and model replacement for years, not on the quality of a 45-day sprint.
- **Keep a small team that can judge.** A handful of people who can read an evaluation and challenge a benchmark is what lets you tell good advice from confident advice.
- **Stop buying badges.** Stop selecting by tier, accepting funded pilots without an architecture review, and signing multi-year commitments to one model family.

Sources

- Anthropic, Microsoft, NVIDIA and Anthropic announce strategic partnerships, 2025: <https://www.anthropic.com/news/microsoft-nvidia-anthropic-announce-strategic-partnerships>
- Microsoft, The next chapter of the Microsoft–OpenAI partnership, 2025: <https://blogs.microsoft.com/blog/2025/10/28/the-next-chapter-of-the-microsoft-openai-partnership/>
- Amazon, OpenAI and Amazon announce strategic partnership, 2026: <https://press.aboutamazon.com/2026/2/openai-and-amazon-announce-strategic-partnership>
- Stanford HAI, AI Index Report 2025: <https://hai.stanford.edu/ai-index/2025-ai-index-report>
- Stanford HAI, AI Index Report 2026: <https://hai.stanford.edu/ai-index/2026-ai-index-report>
- Anthropic, \$100 million for the Claude Partner Network, 2026: <https://www.anthropic.com/news/claude-partner-network>

- Anthropic, Claude Partner Network Services Track and Partner Hub, 2026: <https://www.anthropic.com/news/services-track-partner-hub>
- TechCrunch, OpenAI calls in the consultants for its enterprise push, 2026: <https://techcrunch.com/2026/02/23/openai-calls-in-the-consultants-for-its-enterprise-push>
- Google Cloud, Google Cloud commits \$750 million to accelerate partners' agentic AI development, 2026: <https://www.googlecloudpresscorner.com/2026-04-22-Google-Cloud-Commits-750-Million-to-Accelerate-Partners-Agentic-AI-Development>
- Anthropic, Building a new enterprise AI services company, 2026: <https://www.anthropic.com/news/enterprise-ai-services-company>
- The Next Web, OpenAI launches \$4bn Deployment Company, 2026: <https://thenextweb.com/news/openai-deployment-company-4bn-tpg-tomoro>
- Pulse 2.0, OpenAI confirms Deployment Company launch and Tomoro acquisition, 2026: <https://pulse2.com/openai-confirms-deployment-company-launch-with-more-than-4-billion-backing-acquires-ai-consulting-firm-tomoro/>
- Reuters via The Star, Amazon's AWS commits \$1 billion toward new unit for embedded AI engineers, 2026: <https://www.thestar.com.my/tech/tech-news/2026/06/30/amazons-aws-commits-1-billion-toward-new-unit-for-embedded-ai-engineers>
- Microsoft, Microsoft Frontier Company: AI engineering that amplifies and protects your intelligence, 2026: <https://blogs.microsoft.com/blog/2026/07/02/microsoft-frontier-company-ai-engineering-that-amplifies-and-protects-your-intelligence/>
- interactive investor, Alphabet raises 2026 capital expenditure guidance, 2026: <https://www.ii.co.uk/analysis-commentary/ii-view-google-firm-alphabet-tests-investor-nerves-ii539766>
- Anthropic, Expanding our use of Google Cloud TPUs and services, 2025: <https://anthropic.com/news/expanding-our-use-of-google-cloud-tpus-and-services>
- Amazon, Amazon to invest additional \$4B in Anthropic, 2024: <https://www.aboutamazon.com/news/aws/amazon-invests-additional-4-billion-anthropic-ai>
- Reuters via The Star, AI reshapes India's IT services sector contracts as clients demand more for less, 2026: <https://www.thestar.com.my/tech/tech-news/2026/08/21/ai-reshapes-india039s-it-services-sector-contracts-as-clients-demand-more-for-less>
- Reuters via Yahoo Finance, TCS to cut more than 12,000 jobs, 2025: <https://finance.yahoo.com/news/tcs-cut-more-12-000-091131053.html>
- European Commission, Data Act explained: <https://digital-strategy.ec.europa.eu/en/factpages/data-act-explained>
- UK Competition and Markets Authority, Cloud services market investigation, 2025: <https://www.gov.uk/cma-cases/cloud-services-market-investigation>

Read online: bdgadvisory.com/perspectives/si-model-cloud-dynamics