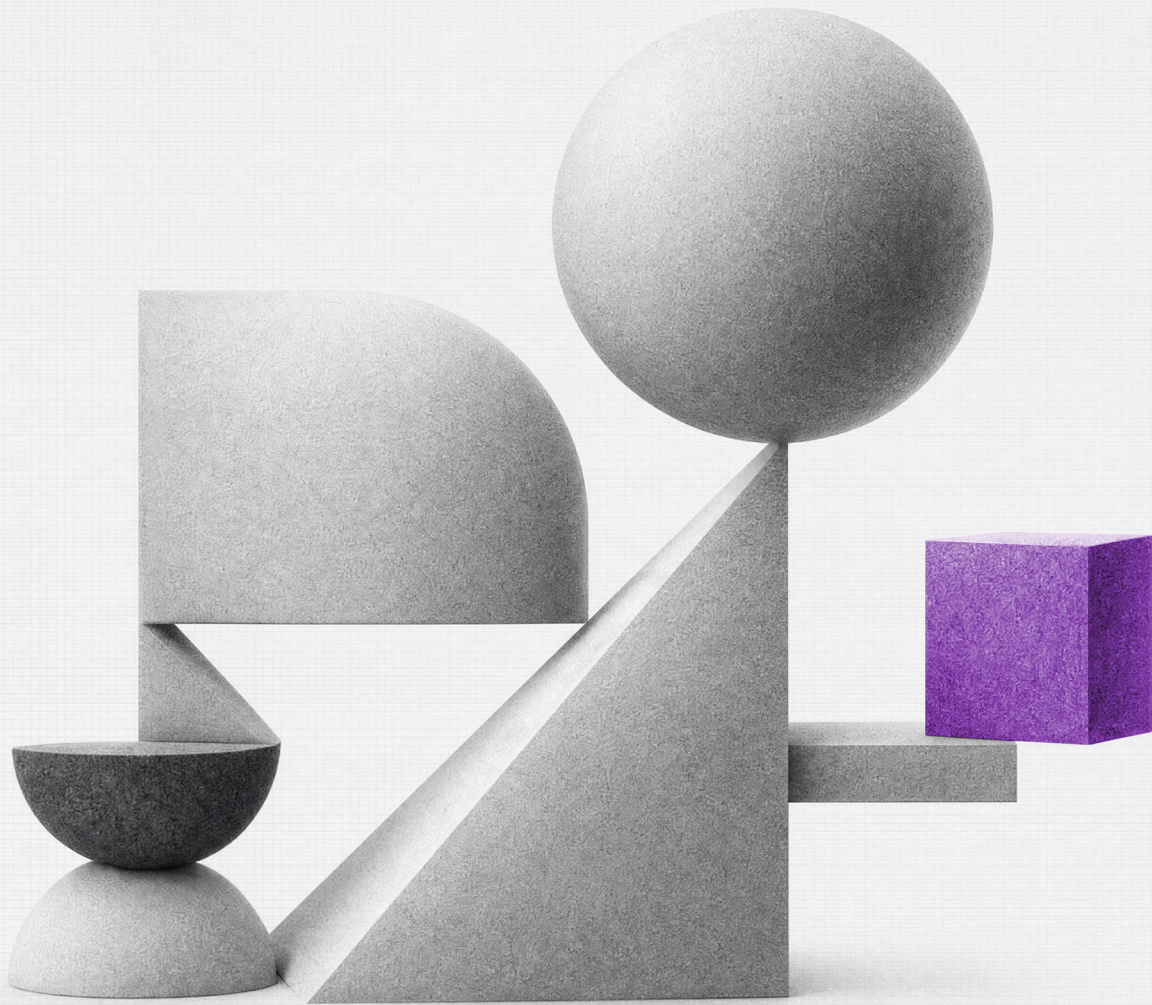
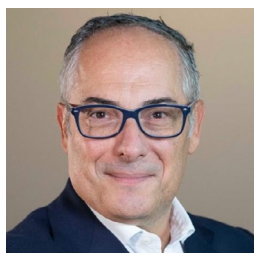




AI designed with intention

How Europe can win with Specialist AI

Authors



**Mauro
Macchi**

CEO – Europe,
Middle East and
Africa (EMEA)

[LinkedIn](#)



**Ashish
Goel**

Senior Managing
Director, Digital
Core Lead – Europe,
Middle East and
Africa (EMEA)

[LinkedIn](#)



**Mauro
Capo**

Managing Director,
Digital Sovereignty
Lead – Europe,
Middle East and
Africa (EMEA)

[LinkedIn](#)



**Surya
Mukherjee**

Principal Director –
Accenture Research

[LinkedIn](#)

1	Executive summary	Page 4-6
2	Europe is spending billions on AI, but who controls it?	Page 7-11
3	Depth beats scale: Where Europe leads in AI	Page 12-13
4	Four forces breaking the scale premise	Page 14-20
5	Agentic AI needs the right models, not bigger ones	Page 21-22
6	The efficiency dividend: What smart models actually deliver	Page 23-24
7	Europe's economy is built for precisely this	Page 25-27
8	Building the ecosystem, fast	Page 28-30
9	Six sectors where Specialist AI wins and a framework for others	Page 31-35
10	What governments can do	Page 36-38
11	Four moves European enterprises should make now	Page 39-41
12	Specialization: Europe's edge	Page 42
13	Calculation note: Frontier model training energy	Page 43-44

Europe holds a stronger hand in AI than the investment headlines suggest.

The US and China are setting the pace in hyperscale spending, where capital intensity determines who leads. Europe's advantage sits elsewhere: in production AI, where value depends on applying the right model to the right task, on the right data, with the right controls.

That advantage reflects the structure of Europe's economy. Precision manufacturing, regulated financial services, life sciences, energy infrastructure and the public sector all generate proprietary operational data that outside providers cannot easily replicate. Europe also has mature governance requirements, deep industrial expertise and research institutions that helped shape the foundations of modern AI. Together, those assets matter most where AI moves from experimentation into embedded, accountable use.

The investment numbers put that strategic position in perspective. The US has committed \$500 billion through Stargate and attracted

\$109.1

billion in private AI investment in 2024 alone, exceeding the combined generative AI total of China, the EU and the UK by \$25.4 billion.¹ Europe does not need a smaller version of the US scale strategy. It needs an AI strategy built around the assets it already controls.

That strategy is becoming more credible as the scale premise weakens. Since 2022, the dominant assumption has been that whoever accumulates the most compute captures the most enterprise value.

The top 15 frontier models now sit within three percentage points of each other on benchmarks that once clearly separated them. Leading labs have missed their own gross margin targets.

Enterprise AI bills keep rising even as token prices fall, because consumption expands to fill every efficiency gain. Europe's grid connection queues in major hubs run seven to ten years.

These constraints point to a specialist approach to AI deployment: an operating discipline that deliberately optimises model size, deployment location, training data, inference pathway and governance controls for a specific operational task, with measurable targets for accuracy, latency, cost, energy consumption, auditability and sovereignty. This specialist approach determines when each model is the right instrument, reserving frontier capability for the roughly 10% of enterprise workloads that genuinely require it.

The strongest enterprise systems already follow this logic. They combine many specialised models with a smaller number of frontier models used selectively for complex reasoning, exception handling and strategic synthesis. The European Patent Office processes 400,000 pages a day on a one-billion-parameter specialist model.² Aleph Alpha's Pharia-1 serves 80,000 German civil servants on fully sovereign on-premises infrastructure.³ BMW runs domain-tuned models directly on vehicle hardware, where cloud Large language Models (LLMs) cannot reach.⁴ Task-specific ai models combined with prompt optimisation can cut energy consumption by up to 90% with no accuracy loss.⁵ At high utilisation, on-premises deployment can reach breakeven against hyperscale cloud within three months.

The pattern extends beyond isolated deployments. Gartner predicts that by 2027, organizations will use small, task-specific ai models three times more than general-purpose large language models.⁶ The economics strengthen as usage grows: each bounded, high-volume workload routed to a right-sized model reduces inference cost, latency, energy demand and exposure to cross-border data risk at the same time.

Europe's regulatory and procurement environment reinforces this momentum. The EU AI Act, GDPR, DORA and emerging EU cloud policy all reward architectures that enterprises can audit, govern, localise and control. Sixty-two percent of European organisations actively seek sovereign AI solutions. Among banks, that figure rises to 76%. Procurement criteria increasingly focus on whether an architecture meets a workload's sovereignty, resilience and auditability requirements. That shift favours Specialist AI directly because it aligns technical fit with regulatory fit.

62%

of European organisations actively seek sovereign AI solutions

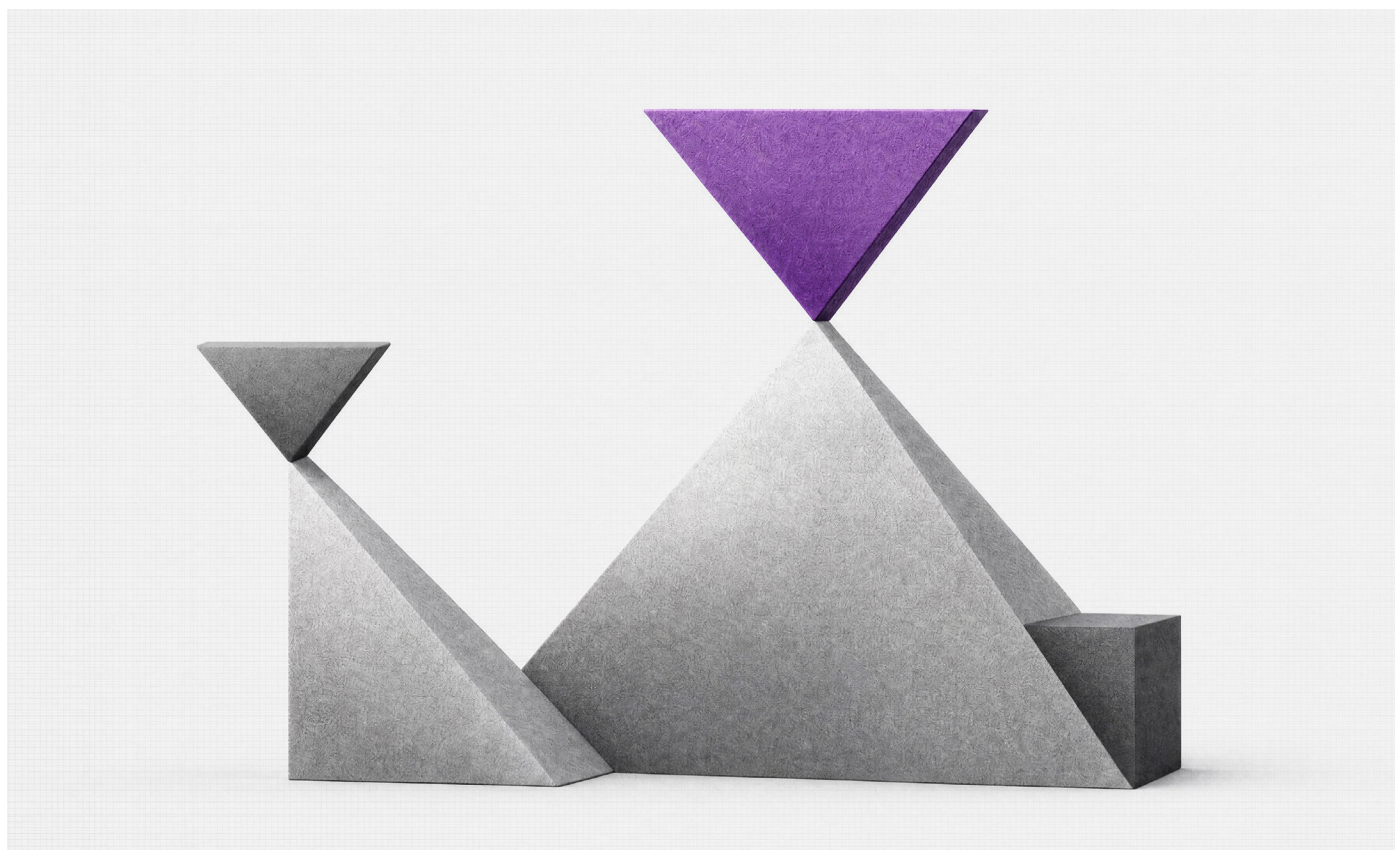
For companies, the strategic logic reduces to four moves: **reserve** frontier models for the 10% of workloads that genuinely need them such as advanced cybersecurity, strategic synthesis, exception handling; **deploy** small but smarter models on high-volume operational tasks in manufacturing, healthcare, financial services, and the public sector; **build** sovereign layers where the data and risk profile demand it, using federated learning to reduce GDPR exposure without sacrificing participation in global ecosystems; and **invest** in domain data now because, as foundation model performance converges, proprietary industrial data becomes the moat, and European enterprises hold datasets no global provider can replicate.

Europe can build an AI advantage around the assets it already controls: proprietary industrial data, regulated-sector expertise, sovereign infrastructure, mature governance and deep operational knowledge. Companies that establish domain-data and fine-tuning relationships now will shape the market as Specialist AI becomes a European standard for production AI.

Research Snapshot

Our December 2025 survey of 1,928 executives across 28 countries, coupled with our client experience, show that undifferentiated scale is not the only path to growth today when Europe's alternative is proving itself in production. The survey captures enterprise sovereignty priorities, adoption stage, and the layers where sovereignty matters most (data, operations, and

models), providing the demand-side evidence for the Specialist AI approach developed in this paper. We use generative AI in our research production process. Our research experts review and validate the generative AI outputs with traditional research methods where possible, applying Accenture's Responsible AI standards



Europe is spending billions on AI, but who controls it?

Europe is spending billions on AI to strengthen control over its technological future.

The EU AI Continent Action Plan, the AI gigafactory programme and the proposed EU Technological Sovereignty Package signal real strategic ambition, with commitments of €200 billion (~US\$229 billion) for data centres, €100 billion (~US\$114 billion) for cloud and AI, and €2 billion (~US\$2.28 billion) for open source.^{7,8,9} The direction is right, but scale alone will not secure Europe's AI future.

In 2024, the US attracted US\$109.1 billion in private AI investment.

Europe trailed far behind. China's private US\$9.3 billion understates its true effort; much of its AI investment flows through state-backed programs and government guidance funds that fall outside private-investment tallies, making the real total difficult to quantify. In generative AI specifically, US investment exceeded the combined total of China, EU and UK by US\$25.4 billion.¹

On strategic infrastructure, the US has announced the US\$500 billion Stargate commitment, while Europe's response is taking shape through the Tech Sovereignty Package.¹⁰

Behind the investment gap sits a sharper question: How much of Europe's AI future depends on systems where infrastructure, data and terms of access are set elsewhere?

61%

of business leaders say AI sovereignty has become a strategic priority.

When we surveyed business leaders on the importance of AI sovereignty in mid-2025, 61% said it had become a strategic priority. By early 2026, that figure exceeded 80%. Executives increasingly understand that AI supply chain concentration carries its own category of risk, distinct from data governance or compliance. The recent suspension of access to Anthropic's Fable 5 and Mythos 5 models, following a US export-control directive on foreign-national access, made that risk concrete. A capability available on Monday was gone on Tuesday, for reasons no European company controlled.¹¹

Europe should respond, but matching hyperscale at the speed and coordination required may not offer the most realistic or highest-return path.

The case for simply matching hyperscale is weakening, as the assumption that compute-scale translates directly into enterprise value fractures. Frontier model performance is converging at a shared point of diminishing marginal returns. On benchmarks that once clearly separated the leading models, the top 15 systems are now separated by as little as three percentage points—more compute is producing less differentiation at the frontier.

The economics are under equal pressure. Enterprise AI bills are rising even as unit prices fall, partly because demand has gone up but token prices don't yet reflect the full cost of running these systems. Leading companies and labs have missed their own gross margin targets, with the cost of running models at scale coming in significantly higher than they had planned. This is a familiar move in technology markets: Price low to win customers first and fix the economics later. The worry for enterprise buyers is what their contracts will look like when pricing normalizes to cover cost and recover margins. Europe's energy grids cannot support this race regardless.¹²

A complimentary architecture is emerging. Specialist AI — sovereign, smarter and specialist by design — is an architecture in which model size, deployment location, training or fine-tuning data, inference pathway and governance controls are deliberately optimised for a specific operational task, with measurable targets for task accuracy, latency, cost, energy consumption, auditability and sovereignty.

Specialist AI does not displace frontier models. It determines when each is the right instrument. The strongest AI systems combine many specialised models with a smaller number of frontier models, used selectively.

Specialist AI is typically built on open-weight foundations that can be fine-tuned on sovereign data and audited at the model level, making open source a practical enabler of EU's technological sovereignty agenda. This is supported by open-source-first procurement preferences and a dedicated

€2B
(~US\$2.28 billion)
funding envelope.⁹

Specialist AI is also the architecture most naturally compatible with Europe's regulatory direction. Emerging EU policy direction would regulate AI at the service layer, not the model itself wherein sovereignty is determined by infrastructure, control and data governance, not model architecture.¹³ High-sovereignty deployment requirements such as data localisation, auditability, EU-only infrastructure and limited third-country access structurally favour the deployment characteristics of Specialist AI. Smaller models are easier to localise, audit, fine-tune on sovereign data and deploy within constrained environments. Specialist AI is not mandated by any single framework.

The case for Specialist AI is not purely technical or economic. It is increasingly structural, shaped by the regulatory and procurement environment in which European enterprises operate. For high-frequency, well-bounded enterprise tasks, fine-tuned smaller models are smarter and can match or outperform larger, general-purpose models when evaluated on task-specific accuracy, latency, cost, auditability, energy consumption and deployment control.

Analyst forecasts already point to task-specific models becoming the dominant enterprise AI deployment model within two years.⁶

Europe does not need to out-invest the US or China. It needs to build better models for specific jobs, sovereign by design, energy-efficient by necessity, rooted in the industrial expertise no other region can replicate.

The advantage is what a small model is trained on and where it runs:

proprietary industrial and regulated-sector data held in Europe, deep domain expertise and deployment inside European jurisdiction.

A German government agency is running a sovereign AI assistant for

80K

civil servants, hosted entirely within German infrastructure and processed in compliance with GDPR.²

The European Patent Office processes 400,000 pages a day on a 1-billion parameter specialist model.³ BMW embeds domain-tuned models directly in vehicle hardware, where cloud LLMs simply cannot reach.⁴

Together, these deployments show a repeatable pattern that the US or China cannot replicate simply by shrinking their own models.

Depth beats scale: Where Europe leads in AI

In 1973, Japanese and American competitors were flooding global markets with mass-produced watches. Switzerland faced a choice: match their volume or redefine the game. It chose precision. The Swiss watch industry established

the Contrôle Officiel Suisse des Chronomètres (COSC), an independent organisation dedicated to certifying chronometer precision, doubling down on accuracy and craftsmanship rather than scale.

Over time, “Swiss Made” became a globally recognised mark of precision and quality.¹⁴ Today, Switzerland produces only around 2% of global watch volumes yet remains the world’s largest exporter of watches by value, generating CHF 25.6 billion (~US\$31.7 billion) in exports in 2025.¹⁵

2%

of global watch volumes yet remains the world’s largest exporter of watches by value

Europe’s position in AI follows the same logic. Many of the foundational ideas that underpin today’s frontier models—including advances in transformers, attention mechanisms and modern deep learning—draw heavily on European research talent and institutions. DeepMind operates out of London, Mistral out of Paris and Interuniversity Microelectronics Centre (IMEC) out of Leuven. They represent world-class AI research capacity.

Europe’s industrial base further strengthens that position. Precision manufacturing, regulated financial services, life sciences and energy infrastructure are the most data-rich, domain-intensive sectors in the global economy. Enterprises with long-standing operations and extensive historical data hold a significant advantage in training and deploying AI systems, and Europe’s industrial base is rich in both.



Europe's industrial base further strengthens that position. Precision manufacturing, regulated financial services, life sciences and energy infrastructure are the most data-rich, domain-intensive sectors in the global economy. Enterprises with long-standing operations and extensive historical data hold a significant advantage in training and deploying AI systems, and Europe's industrial base is rich in both.

Europe has also advanced the policy agenda. The AI Continent Action Plan, national compute investments from member states such as France, Germany and the Nordics and an AI Act with influence beyond EU borders all signal strategic intent, even as the policy conversation shifts toward regulatory simplification and competitiveness. Europe's relative weakness remains hyperscale AI. Its advantage lies where technical depth, responsible AI and sector expertise converge such as industrial AI and regulated-market applications.¹⁶

The US and China still lead in scaling and industrialising foundation models. China has also built a regulatory mechanism less often noted: a mandatory generative AI registry, widely considered the world's only public registry of its kind.¹⁷ It reached

302 approved services by end of 2024	346 filed services by April 2025	250–300 new registrations per month. ^{18,19}
---	---	--

That creates a different kind of competitive advantage, and Europe has no direct equivalent yet.





Four forces breaking the scale premise

Since 2022, the dominant AI narrative has run on a single assumption: bigger models, more Graphics Processing Units (GPUs), more data, more value. Whoever accumulates the most compute captures the most enterprise value. That premise is fracturing, and four forces are driving the break.

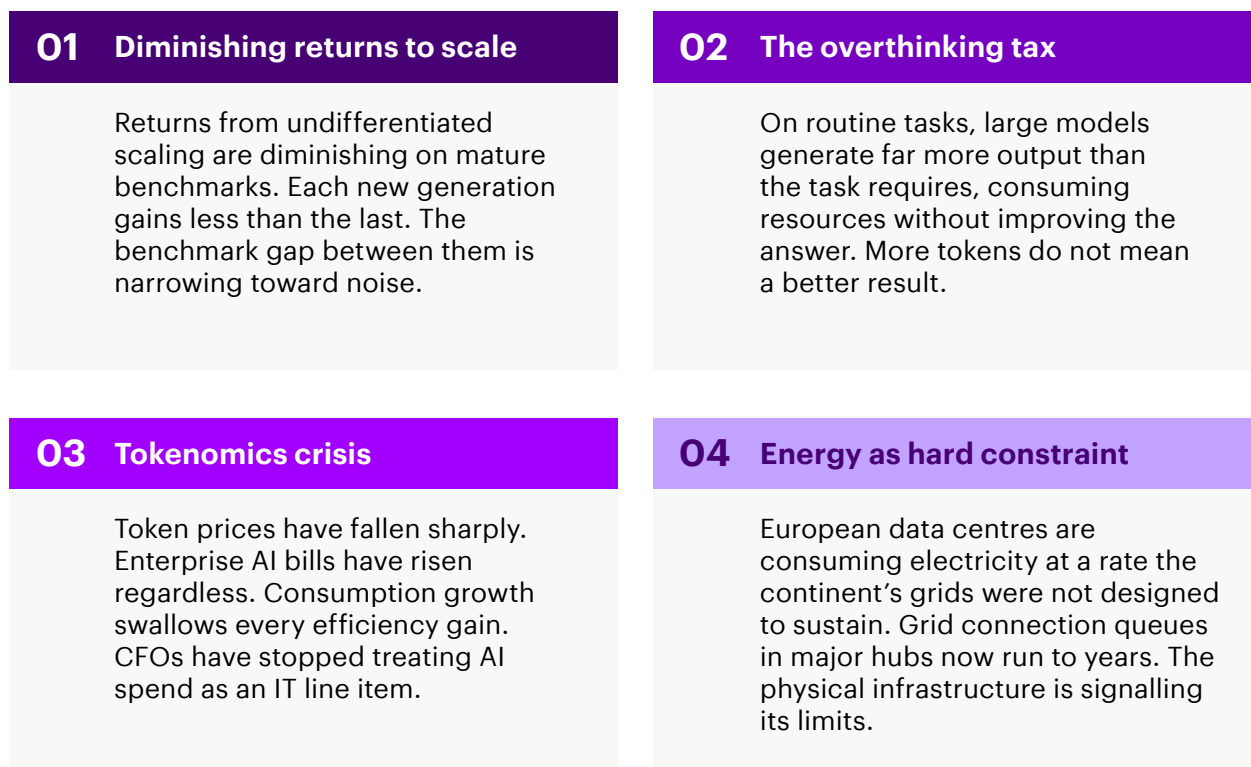
Four forces are weakening the scale premise. First, **returns to scale are diminishing** as benchmark gains saturate, costs rise, and deployment realities become harder to ignore.

Second, the **overthinking tax is growing**: larger models often generate more noise than signal on bounded enterprise tasks. Third, the **tokenomics crisis** is intensifying as token bills outpace productivity gains. Fourth, **energy has become a hard constraint**, making large-scale compute structurally difficult across much of Europe (**Figure 1**).

Increasingly, value depends not simply on generic capability, but on how effectively AI is integrated into workflows, grounded in domain data and evaluated against specific operational tasks.

Beyond these four forces lies a separate strategic risk: access. Export controls, national-security rules and provider decisions can change the terms on which frontier models are available across jurisdictions. Meaningful provider choice and credible domestic alternatives for critical workloads reduce concentration risk and position Europe as a confident partner rather than a dependent customer.

Figure 1: Four forces breaking the AI scale premise



01 Diminishing returns to scale

Empirical analysis of hundreds of AI models²⁰ suggests that return from additional scale become less predictable as benchmarks mature and gains narrow. Leading researchers have increasingly questioned whether parameter scaling alone can deliver the same magnitude and predictability of gains as before, particularly as high-quality training data becomes more constrained (Epoch AI).²¹ At NeurIPS 2024, OpenAI co-founder Ilya Sutskever declared that pre-training scaling is "essentially tapped out,"²² with the internet's finite data supply setting a hard ceiling on the traditional approach. For enterprises, model gains matter only when they translate into useful task performance in real workflows, with domain data and operating requirements as the test.

02 The overthinking tax

Research published in 2025 formally characterised what practitioners had been noticing for months: Large reasoning models generate unnecessarily long outputs on simple tasks without improving the answer.²³ Researchers call this phenomenon overthinking. Consuming more tokens does not necessarily produce a better result. Sometimes it means the opposite.

03 Tokenomics crisis

Enterprise CFOs have given the cost crisis its own name: tokenomics. Token prices have fallen sharply, but bills have not, because consumption rises to fill every efficiency gain and makes the value of any individual token hard to forecast.²⁴ The arithmetic is unforgiving: a 20% fall in per-token price is exactly cancelled by a 25% rise in consumption. Beyond that, the bill grows even as every token gets cheaper.²⁵

Many organisations are building general-purpose AI architectures for specialised tasks. That mismatch is structural, and expensive.

In enterprise settings, output matters only when it improves performance on a defined task inside a real workflow.

04 Energy as hard constraint

Two forces make compute a resource to spend carefully, and only one of them is about location. The first is global and unavoidable: Electricity supply is not keeping pace with AI demand, and power prices are rising across every major market, so inefficient compute costs more everywhere. A data centre cannot escape this simply by relocating. Global data centre electricity consumption grew 17% in 2025, more than five times the global electricity demand growth rate of 3%.²⁶ The International Energy Agency projects data centre demand reaching approximately 950 Terawatt-hours by 2030, nearly 3% of all global electricity.²⁷ The second force is regional: In Europe specifically, grid capacity and connection timelines add a further, harder ceiling.

22%

of Ireland's electricity consumed
by data centres in 2024²⁸

7-10yrs

grid connection queue for new data
centres in European/FLAP-D hubs
(Frankfurt, London, Amsterdam,
Paris and Dublin²⁹)

Locational strategy can ease the regional constraint, but it cannot answer the global one: wherever the compute runs, spending it on oversized models for bounded tasks is a waste that rising energy prices make increasingly expensive.

Training frontier AI models can consume substantial electricity, but exact figures are rarely disclosed. One independent analysis estimates GPT-4's training energy at 51,773–62,319 Megawatt-hour.³⁰ OpenAI announced **Stargate UK** in September 2025 with NVIDIA and Nscale but reports in April 2026 said the project had been paused, citing energy costs and regulatory uncertainty.^{31,32}

The EU's InvestAI initiative
aims to mobilise

€200B (~US\$229 billion), including a
€20 billion (~US\$22.9 billion)
fund for AI gigafactories.

These facilities are themselves major electricity consumers, so the programme's success depends as much on grid capacity as on funding; public descriptions indicate a financing model with substantial private co-investment.³³

If scale is not the whole story, the question becomes what is. A fair challenge to any small/smart-model argument is whether it confers a durable edge at all. The largest providers already build excellent compact models, and query costs have fallen more than a hundredfold in two years. If Specialist AI were only about size, it would be a commodity.

But size is not the moat. What distinguishes deployment is where and how a model runs. The advantage comes from jurisdiction, proprietary domain data and procurement fit, including model-level auditability and fit with emerging EU procurement expectations that reward European supply-chain contribution. Those factors do not depend on a bigger model; they depend on where the model runs, what data it uses and how it fits the buying criteria.

The opportunity is real, time-bound and collaborative. Major international providers, including US firms, are building EU-controlled offerings that can serve much of the market at lower and mid sovereignty tiers under emerging EU sovereignty rules. The high-sovereignty core is where domain-tuned European deployments fit most naturally.

The advantage lies in establishing those domain-data and fine-tuning relationships now, building on the global ecosystem rather than apart from it.

Agentic AI needs the right models, not bigger ones

Agentic systems turn complex workflows into sequences of small, repeatable tasks. Most of those tasks reward precision, auditability, latency and cost control more than frontier-scale capability. As workflows break into discrete steps, Specialist AI can take on more of the work.

Recent research by Belcak and colleagues at NVIDIA Research supports this view.³⁴ Specialist AI fits classification, extraction, compliance checks and domain summarisation, where accuracy and auditability matter most. Frontier models remain valuable for exception handling and strategic synthesis, where broad context and open-ended reasoning justify the cost. The strongest systems use many specialised models and reserve frontier models for the moments that need them.

Crystalline: How agentic architecture can turn model capability into domain performance

Crystalline, built by Accenture Italy's cybersecurity team, shows how that architecture can turn model capability into domain performance. On CyberGym, the UC Berkeley benchmark of 1,507 real-world vulnerabilities, Crystalline scored 89.6%, compared with 66.6% for the same frontier model without it. The 23-point gain came without

a bigger model or fine-tuning. Crystalline improves performance by reusing principles and patterns from past tasks, so knowledge compounds instead of starting cold each time. That is the fuller claim of Specialist AI: domain specialisation, workflow memory, and selective use of frontier capability can matter more than model size alone.

Cybersecurity also clarifies the boundary. It remains a brute-force domain where frontier power still matters, but Crystalline narrows that requirement through design. Accenture Italy is now extending

the same pattern to finance through Accenture Financial Advanced Solutions & Technology (AFASST), an early template for domain-specific agentic systems in regulated industries.

Specialist AI on the Factory Floor

Aerospace and industrial engineers spend more time searching and documenting than engineering. Legacy systems, technical drawings, and regulatory records fragment the work — and every hour lost to retrieval is an hour not spent on design, validation, or innovation.

Accenture's Engineering Orchestrator addresses this by coordinating specialized AI agents across requirements, design, simulation, and technical documentation within a single, traceable execution layer. It's deployed on air-gapped, on-premises hardware that keeps sensitive engineering data under full organizational control. Each agent handles a narrow domain such as requirements extraction, systems modelling, design validation, FMEA, drawing interpretation, and PLM integration.

The architecture is toolchain agnostic, integrating with existing engineering tools such as DOORS, CATIA, and Teamcenter rather than replacing them. Smaller, fine-tuned models handle these bounded tasks and outperform larger general-purpose models. Frontier capability enters only where genuine complexity demands it: cross-domain synthesis and exception handling.

For aerospace and defense firms, proprietary engineering data is a strategic asset. Routing it through external cloud infrastructure is not a compliance question—it is a competitive risk. The Engineering Orchestrator eliminates that risk by design. The result: engineers spend less time retrieving and more time engineering, on a system that is faster, cheaper to run, and sovereign by architecture.

The efficiency dividend: What smart models actually deliver

The efficiency case for Small Language Models (SLMs), the technical core that makes Specialist AI smart, is a performance argument backed by numbers.

Start with energy. Research shows the energy gap between an SLM on edge hardware and a frontier LLM on a GPU cluster typically spans one to two orders of magnitude per token, although the exact figure depends on deployment architecture, not model size alone.^{35,36} For bounded, well-defined enterprise tasks, a well-trained SLM is not a compromise, it's often the more rational choice.³⁷

90%

**energy reduction using
task-specific models and
prompt optimization, no
accuracy loss**

The performance picture reinforces this. Capable reasoning models now available can run on a high-end laptop and outperform models from two years ago that required a data centre, though the absolute cutting edge still demands one. Recent academic work suggests that returns to scale can decelerate, and that well-tuned smaller models can match or outperform larger general-purpose systems on specific tasks.³⁸ In enterprise settings, smaller models can be a form of precision.

The cost case follows directly. Fine-tuned specialised models can deliver domain-task inference at a fraction of frontier-model cost, sometimes matching or exceeding larger models on task-specific accuracy.

A smaller model run on edge hardware typically consumes one to two orders of magnitude less energy per token than a frontier model on a GPU cluster. Independent modelling puts the breakeven for on-premises deployment against hyperscale cloud at within three months for high-volume workloads. Organisations switching narrow, high-volume workloads to smaller fine-tuned or

distilled models report substantial cost and latency gains.³⁹

Falling frontier prices do not erase this advantage. A right-sized model reduces consumption at the source, improving the economics of each task rather than relying on unit-price declines that usage can quickly outpace.

The Specialist AI architecture that Europe is particularly well-placed to lead reframes smartness as a strength where the workload justifies it:

Targeted, energy-efficient models are often the right choice for most industrial workloads, a practical alternative to the race toward ever-larger parameters. It combines technical optimisation, including quantization, model distillation, mixture of experts, local deployment and a focus on reducing inference consumption as well as training demand. This aligns closely with the operating realities European enterprises face, including budget constraints, existing infrastructure limitations, available internal capabilities, GDPR and AI Act compliance requirements, and growing demands for data sovereignty.

Mistral AI has quantified what this looks like in practice. For a 400-token response in Le Chat, excluding users' terminals, it estimates marginal inference impacts of 45 millilitres of water consumption, 1.14 grams Carbon Dioxide Equivalent (CO₂e), and 0.16 milligrams antimony-equivalent resource depletion. Mistral describes this as the first comprehensive lifecycle assessment of an AI model, peer-reviewed, and compliant with ISO 14040/44.⁴⁰

By 2027, Gartner forecasts that organizations will use small, task-specific AI models at least three times more often than general-purpose LLMs.⁶

Europe's economy is built for precisely this

Specialist AI maps naturally onto the structure of the European economy. The fit is structural, not aspirational. This does not eliminate the importance of frontier models; it changes where and how enterprises derive the greatest operational value from them, decomposing industrial workflows into governable task units and providing more control.

A black box delivers automation, but it conceals where each action occurs. A workflow broken into agents responsible for discrete steps makes every action explicit, so operators can log, measure, govern and audit each one. Transparency and accountability become achievable by design, not aspirations bolted on afterward. Each model sits where its data and risk profile require: edge, on-premises, sovereign cloud, or air-gapped. Sovereignty becomes a property of the operating model, shaped by workload-level risk assessments and procurement requirements.

European SMEs represent around 99% of all businesses in the EU,⁴¹ but they lag large firms in AI adoption, blocked by cost, talent and data-sensitivity. Specialist AI removes all three barriers

directly as it is cheaper to run, deployable on existing infrastructure and able to keep sensitive data in-house. The continent's industrial base—Mittelstand manufacturing, specialised life sciences, regulated financial services, public sector institutions—does not need a model that can discuss philosophy. It needs a model that can process a German-language patent filing in under a second, flag a GDPR-relevant data field without routing it to an American cloud, or process

400K

pages daily with over 99% accuracy.² Those are not edge cases, but the workload.

The demand signal confirms this: 62% of European organisations actively seek sovereign AI solutions. Among European banks, that figure rises to 76%. Gartner surveys reveal that geopolitics will drive 61% of CIOs and IT leaders in Western Europe to increase reliance on local cloud providers. Fifty-three percent of Western European CIOs and IT leaders said geopolitics will restrict their organizations' future use of global cloud providers.⁴² This is not a compliance posture. It is a procurement strategy.

~60-80% of European organisations seeking sovereign AI solutions

76% of European banks specifically seeking sovereign AI

Procurement makes this architectural shift operational. Emerging EU expectations increasingly reward European supply-chain contribution, model-level auditability, open-source options, sovereignty and control. That favors systems that minimize data transfer, limit extraterritorial exposure and offer verifiable governance. Specialist AI fits these criteria directly. Smaller, task-specific models on sovereign or hybrid infrastructure are easier to audit, cheaper to run and better aligned with procurement expectations in regulated sectors such as energy, banking, healthcare, transport and telecoms.

Specialist AI also reduces strategic risk. Emerging European sovereignty frameworks assess providers across four dimensions: control and ownership, exposure to foreign laws with extraterritorial effect, service continuity ('kill-switch') risk and operational autonomy from foreign-controlled subcontractors and support chains. Smaller and smarter models deployed on sovereign or on-premises infrastructure reduce reliance on third-country-controlled subcontractors, telemetry and support chains. They do not eliminate risk, but they reduce exposure across each dimension. This creates a tiered market rather than a binary choice. Under the emerging EU approach, trusted non-EU providers are expected to remain eligible for most workloads, while only the most sensitive, sovereignty-critical workloads move toward a no-third-country-control standard. The strategic goal is fit-for-risk placement, not a foreign-free technology stack.

The economic case reinforces this structural advantage. For high-volume workloads, the economics increasingly favour on-premises deployment. Independent academic modelling puts breakeven against hyperscale cloud at within three months at high utilisation, with a token-cost advantage that compounds with scale.

| Sovereignty and cost efficiency, on this architecture, point in the same direction.



Building the ecosystem, fast

Europe's sovereign AI ecosystem accelerated sharply through 2025 and into 2026.

This is no longer only a story of pilots. It is increasingly visible in live deployments, substantial capital commitments and strategic partnerships.

The ecosystem is best read in three layers: the task-specific and edge deployments that are Specialist AI in action, the open-weight and efficiency foundations they build on and the sovereign infrastructure that hosts them.

Task-specific and edge deployments are already moving into production.

Aleph Alpha's Pharia-1-LLM-7B serves 80,000 German civil servants on fully sovereign, on-premises infrastructure compliant with GDPR and EU AI Act requirements.³

The European Patent Office and Mistral AI use a one-billion-parameter OCR model, fine-tuned on 50,000 patents, to process 400,000 pages per day at an error rate below 1%, cutting lead times from five days to minutes.² BMW Group, in a proof of concept with

Google Cloud, is developing domain-specific SLMs that run directly on vehicle hardware.⁴ The important distinction is deployment: Inference runs at the edge, with data remaining on-device, avoiding the latency and connectivity constraints of cloud-based LLMs in vehicles.

Open-weight and efficiency foundations are also strengthening. Mistral's compact edge models show how smart architectures can outperform much larger predecessors on constrained hardware.

In early 2026, the company raised

€830M

or about US\$950 million, in institutional debt without US venture capital, making it the first European AI company to finance infrastructure at that level independently.

Its April 2026 white paper also called for European preference in public AI procurement, linking open-weight capability to sovereignty and market access.⁴³

Cohere fits this layer through its deployment model and efficiency logic. It deploys on customers' sovereign infrastructure rather than serving them from global servers, and its large partnership in Germany reinforces the European relevance of that approach. Cohere also describes a capital-efficient model that scales compute in proportion to client demand rather than ahead of it.⁴⁴ Its Aya Vision 32B outperforms models up to 2.8 times its size on multilingual vision-language benchmarks, and its AI Energy Score initiative with Hugging Face, Salesforce, and Carnegie Mellon University aims to standardise inference-efficiency measurement.^{45,46}

Sovereign infrastructure and delivery are also moving from concept to capability. SAP's EU AI Cloud, launched on 27 November 2025, gives customers a choice of SAP data centres, trusted European infrastructure, or fully managed on-site deployment, all with full EU data residency.⁴⁷

Deutsche Telekom's T Cloud Public, significantly expanded in February 2026, has committed to full core feature parity with US providers by the end of 2026 while guaranteeing EU-compliant data processing protected against third-country access.⁴⁸ The Accenture and Mistral AI partnership, announced in February 2026, adds an enterprise delivery layer, combining open-weight models with sovereign deployment options, architecture expertise and compliance support.⁴⁹

Open-weight foundations are becoming a structural feature of Europe's AI ecosystem. They can be fine-tuned on sovereign data, deployed across sovereign nodes without licensing friction and audited at the model level. EU policy strengthens that advantage through open-source procurement preferences, co-equal investment in open-source cloud and frontier AI and EuroCloud⁵⁰'s interoperable architecture. Together, these moves create policy-backed market access for open-weight systems and give European AI providers a competitive advantage grounded in both technology and regulation.

The CEOs of Airbus, ASML, Ericsson, Mistral AI, Nokia, SAP, and Siemens published a joint May 2026 op-ed warning that Europe is "losing global competitiveness every day" and demanding regulatory simplification and coordinated industrial policy.⁵¹ They stressed that "long-term sovereignty requires us to fiercely support and scale homegrown European innovation, alongside fostering the skills and mobility of our workforce." In direct response to that call, a sovereign-by-design, precision-over-scale model is emerging as the clear direction of travel for European industry.



Six Sectors Where Specialist AI Wins

Our analysis points to six sectors where the structural conditions for Specialist AI are strongest.

In these sectors, narrow task definition, regulatory pressure and deep European data advantages combine to make domain-specific small models more effective, not merely cheaper, than frontier alternatives.

Industrial manufacturing leads because Mittelstand firms run high-volume, repetitive shop-floor tasks where local inference can reduce latency, improve resilience and keep sensitive production data inside the factory.

Automotive is a priority case because domain-tuned models running directly on vehicle hardware, as in the BMW example above, can handle latency- and safety-critical tasks that cloud LLMs cannot reach. Regulated financial services follow a similar logic. GDPR, the AI Act, DORA and audit requirements do not universally mandate sovereign architectures, but they strongly favour auditable, controllable, resilient systems.

| In healthcare and life sciences, strict protections for health data often constrain centralised inference, making controlled data movement, precision and auditability hard requirements.

The public sector faces a parallel constraint as digital sovereignty becomes a procurement priority and institutions seek reduced dependency, resilience and control over strategic data. Legal and compliance services are a natural fit because the work is document-intensive, jurisdiction-specific, multilingual and audit-sensitive. Energy and utilities are a frontier case. As grids digitalise and distributed assets expand, reliable local inference becomes increasingly important for remote, constrained, latency-sensitive or safety-critical infrastructure.

Where specialist AI has the strongest potential

Sector	Why specialist AI fits	European strength
 Industrial manufacturing	Narrow, high-volume tasks. Latency-critical. Data stays on-premises. Automotive edge/in-vehicle inference	Mittelstand ecosystem, world-leading automotive base, deep OT/IT integration
 Regulated financial services	GDPR, DORA and AI Act compliance by design. Auditability required.	Strong regulatory frameworks, risk culture
 Healthcare and life sciences	Sensitive patient data. Domain specificity essential.	World-class research institutions, clinical data
 Public sector and defence	Sovereignty scales with sensitivity; maximal for defence.	National data infrastructure, EU procurement shift
 Legal and compliance	Document-intensive, language-specific, jurisdiction-bound.	Civil law tradition, multilingual complexity
 Energy and utilities	Edge deployment on constrained grid infrastructure.	Renewables transition, grid modernisation

The fourth lens: which sovereign architecture is right?

[In our previous Sovereign AI research](#), we found that roughly one third of AI workloads require sovereign treatment, with three lenses framework calibrating that decision: industry risk, national context and use-case criticality.

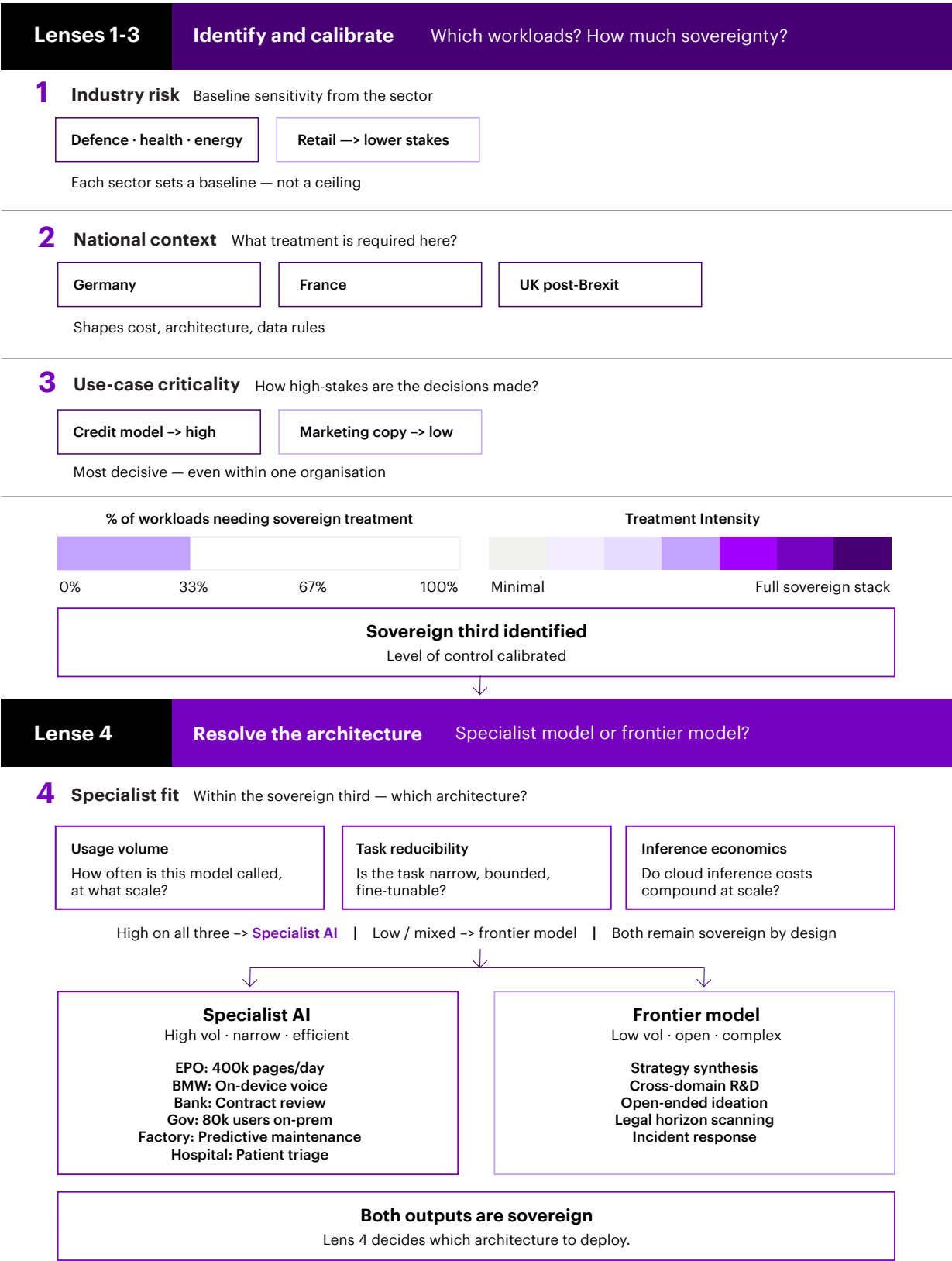
This paper adds a fourth lens: Which architecture is right?

Three signals determine the architecture choice: usage volume, inference economics and task reducibility. Specialist AI fits workloads that run often, cost meaningfully at inference and break into narrow, repeatable steps. Smaller, fine-tuned models can often serve those workloads more efficiently within the required sovereign controls.

Frontier models should be reserved for irreducible tasks: work that depends on non-obvious judgment, long chains of reasoning, synthesis across extensive context or novel combinations of multiple domains. These workloads may occur less frequently, but they require capabilities that smaller models cannot reliably reproduce.

Both architectures can be sovereign.
The fourth lens determines which sovereign architecture best fits the workload.

The diagram below maps the full decision logic:



What governments can do

The policy architecture for Specialist AI is still forming, and governments can shape it now through three moves: adding an efficiency track to AI Factory funding, using procurement to reward sovereign and auditable architectures, and creating a cost and energy transparency standard. These are starting points for discussion, not prescriptions.

01 Build an efficiency track into AI factories

The EU's AI gigafactory programme,⁵² part of the package's roughly €100 billion cloud and AI envelope and expected to include up to five facilities, is strategically important but oriented toward frontier scale. A parallel track within Horizon Europe and the EIC could focus on energy-efficient edge AI and domain fine-tuning, redirecting capital toward Europe's natural advantage. That direction is already visible in the Strategic Roadmap for Digitalisation and AI in the Energy Sector, part of the June 2026 Tech Sovereignty Package. Its 'AI for Energy' pillar backs energy-specific AI foundation models built on dedicated data spaces with Horizon Europe funding, while its data-centre rating scheme rewards energy efficiency, clean-energy use and waste-heat reuse rather than raw scale. Emerging EU cloud and AI policy, including the proposed Cloud and AI Development Act,⁵³ is turning sovereignty from an idea into a practical deployment question. Policy is moving toward a risk-tiered model in which workload sensitivity determines the required level of control and infrastructure, from standard cloud to fully sovereign EU environments. That direction aligns with Specialist AI because high-sensitivity use cases favour smarter, task-specific models on sovereign or hybrid infrastructure over general-purpose, frontier-scale systems. The final scope and obligations of any legislation remain unsettled, so the relevant signal is policy direction rather than any specific article or tier.

02 Where Specialist AI fits: A sovereignty-driven workload framework

The tiering below maps most directly onto public-sector and regulated procurement, where higher sovereignty requirements are most likely to become explicit expectations over time. Private enterprises can apply the same logic voluntarily, and many are expected to adopt it as a de-facto standard. In practice, lower-sensitivity tiers are typically self-assessed, while higher-sensitivity tiers call for independent audit. Where the sector table earlier asked which industries fit Specialist AI, this table applies a sovereignty lens: how the sensitivity tier of a workload, not the sector it sits in, drives the level of control and the architecture choice.

Sovereignty Level*	Workload Type	Typical Use Cases	Architecture Implication	Role of Specialist AI
1 Level 1* (Baseline)	Low sensitivity General workloads	Customer service copilots Content generation General enterprise assistants	Hyperscale cloud frontier models Global infrastructure	Limited role: dominated by frontier models
2 Level 2* (Enhanced)	Moderated sensitivity Regulated workflows	Financial reporting Legal summarisation Enterprise knowledge copilots	Hybrid: cloud and regional control Stronger data governance	Selective use: Smaller fine-tuned models improve cost, latency and control
3 Level 3* (High Sovereignty)	High sensitivity Sector-critical Workloads	Banking risk models Healthcare diagnostics Critical infrastructure AI	Sovereign or partner-controlled infrastructure Strong localisation	Strong fit: Domain-specific models with higher auditability, efficiency, and control
4 Level 4* (Maximum Sovereignty)	Critical national/public sector workloads	Defence Law enforcement Classified public-sector AI	Fully sovereign, EU-controlled, on-prem, edge, air-gapped	Dominant architecture: Specialist AI often required to meet sovereignty, auditability, latency constraints

*Based on proposed CADA regulation, which has not been adopted and is included for illustrative purposes only.

The implication is a structured mix of frontier and smaller but smart models. As sovereignty requirements rise, architectures naturally shift toward Specialist AI, with task-fit deployment closer to data and stronger control and accountability.



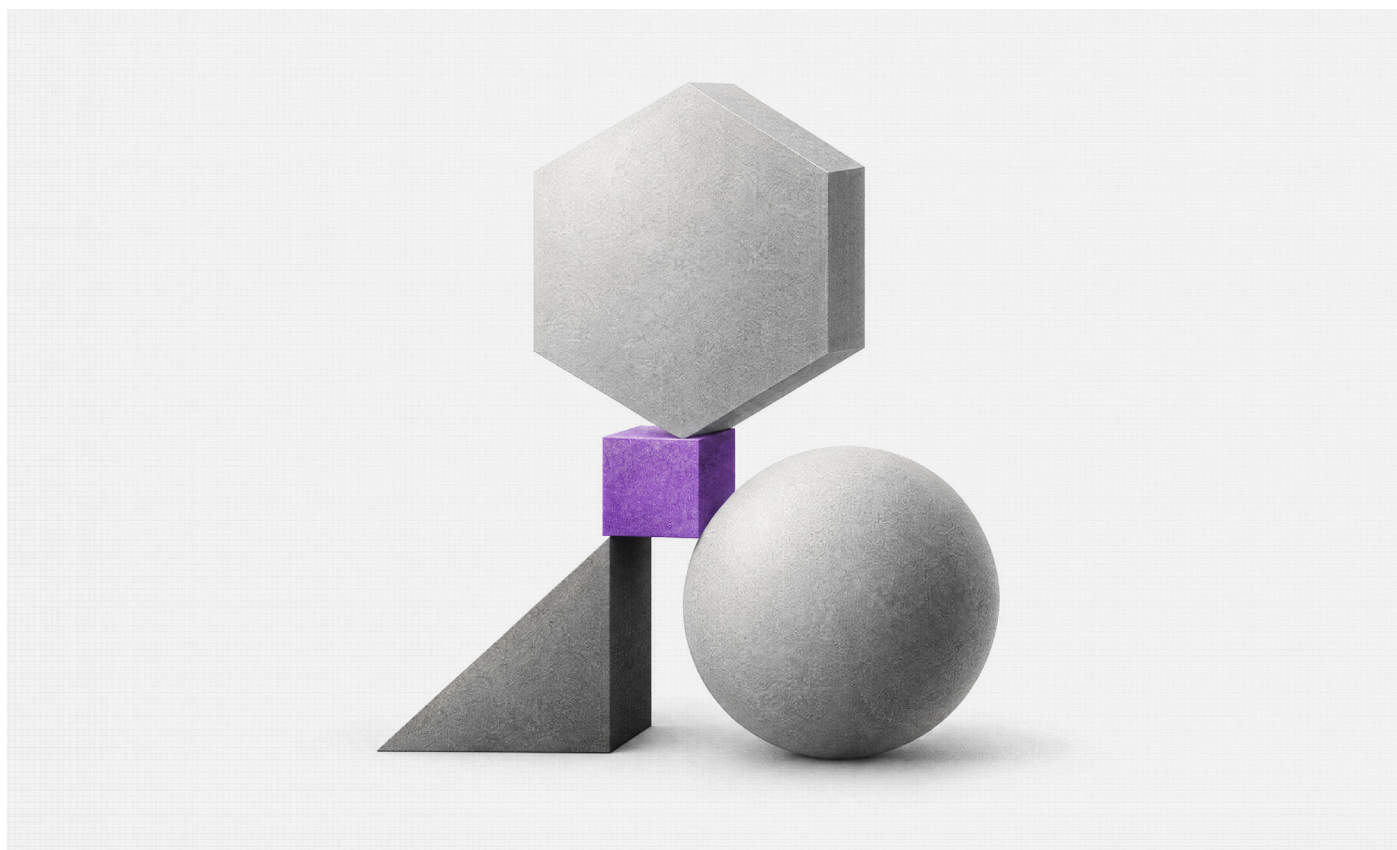
03 Procure for workload risk, not model prestige

Procurement can set the architecture. Public-sector and regulated-industry buyers are shifting from asking, "Which model is best?" to asking, **"What level of sovereignty does this workload require, and can Specialist AI meet it?"** This reframing puts workload sensitivity, auditability, data control and resilience alongside price as procurement criteria. It also favours providers that can demonstrate European technological contribution and open-source options. Some member states are already moving this direction. France's February 2026 State Digital Purchasing Doctrine,⁵⁴ for example, includes exposure to extraterritorial laws as one factor in digital purchasing decisions.

This emerging model centres on verifiable control and risk-aligned architecture within an open framework. Enterprises building sovereign SLM deployments today will be better positioned as this shift develops, with the AI Act increasingly functioning as a competitive feature rather than only a constraint.

04 Create a cost and energy transparency standard

The EU AI Act's data governance, transparency and human oversight requirements are easier to satisfy with sovereign, domain-specific edge models than with large centralised systems. **Governments can strengthen that advantage by creating a cost and energy transparency standard.** Today, no standardised disclosure captures the full cost of AI deployment across energy use, compliance overhead and inference cost. A voluntary standard, similar to appliance energy labels, would help enterprises and public authorities compare deployment models on equal terms. It could build on the data centre rating scheme proposed under the Sovereignty Package's Energy Sector Roadmap, which covers energy and water efficiency, clean energy use, waste-heat reuse and flexibility. It could also draw on the Commission's recent consultation on measuring AI energy use and emissions, which closed in May 2026.⁵⁵



Four moves European enterprises should make now

European enterprises can act now, even as policy continues to evolve. The priority is to allocate AI deliberately: reserve frontier capability for the work that truly requires it, move high-volume operational tasks to smarter models, build sovereign layers around the data and applications that matter most and turn domain data into a durable advantage.

Enterprises should reserve frontier models for genuinely complex, high-stakes or open-ended tasks such as advanced cybersecurity, frontier research, strategic synthesis, cross-domain discovery, exception handling and incident response.

These tasks may represent only

10%

of most enterprise AI workloads. The rest often gets over-served at unnecessary cost.

They should deploy SLMs on operational tasks where narrow scope, high volume, latency sensitivity and data sensitivity define the work. Manufacturing, healthcare, financial services, legal and public-sector environments all have those characteristics. In these settings, smaller models can improve performance and economics by reducing unnecessary inference at the source.

Enterprises should also build sovereign layers where the data, model components or regulated applications demand them. Data, sensitive model layers and regulated application environments should run on sovereign or hybrid sovereign infrastructure. Federated learning, which shares model updates rather than underlying data, can reduce Article 9 GDPR risk by design, according to the European Data Protection Supervisor, while still allowing enterprises to participate in global ecosystems.⁵⁶

The next priority is domain data. As foundation model performance converges, proprietary industrial data becomes the differentiating asset. European enterprises hold domain-specific datasets that global training corpora do not fully capture.^{57,58} That advantage will compound only if companies curate, govern and use those datasets to fine-tune models for specific workflows.

Finally, enterprises should instrument their AI before a formal standard arrives. COSC gave watchmakers a shared language for precision before standards became part of the industry's value proposition. European enterprises face the same moment in AI. The four-lens framework above offers a starting point: map workloads against industry risk, national context and use-case criticality to identify the sovereign third, then apply the precision-fit lens to determine where a smart model is the right instrument. For each workload, track domain task accuracy at a defined parameter ceiling, energy per inference at deployment scale and verified data sovereignty of the deployment architecture. Review these metrics quarterly. Enterprises that build this discipline now will help shape the criteria for a future European Specialist AI standard.

The next enterprise AI winners will pair strong models with workflow redesign, controls and production scale. Specialist AI puts industry expertise back at the centre of AI transformation.

Specialization: Europe's edge

The Swiss watch industry did not try to out-manufacture Japan. It went deeper into the parts of the value chain where precision, trust and craft created defensible, compounding advantage. The result was

2%
of global volume and a disproportionate share of global value.

Specialist AI applies that same strategy to artificial intelligence.

Europe will not win by replicating Stargate on a fifth of the budget. It can win by building best-in-class AI for the industrial, regulated, data-sensitive, energy-constrained environments that define its economy. Specialist AI gives Europe a credible path to compete by combining smaller, smarter, task-specific models with sovereign infrastructure and deep domain expertise, rather than trying to match the hyperscale spending of the US and China.

Specialist AI compliments frontier AI. It gives enterprises an operating discipline for deciding when frontier capability is necessary, when small but smart models perform better, and how to orchestrate both inside sovereign, observable, industry-specific workflows. Sovereignty means controlling the workloads, data, infrastructure and model layers that matter most while staying open to global technology. Cybersecurity may offer the clearest exception, because the most advanced defensive work still requires trusted access to frontier LLM capability. For most industrial, regulated and high-volume enterprise tasks, Europe's advantage lies in deploying smaller, task-specific models closer to the data, under European control, with measurable gains in cost, energy, auditability and resilience.

Calculation note: Frontier model training energy

Published estimates for training energy of frontier models vary widely depending on model size, hardware generation and training duration:

- GPT-3 (175B parameters): approximately 1,287 MWh, per [Patterson et al. \(2021\), "Carbon Emissions and Large Neural Network Training"](#) (Google/UC Berkeley). (Note: this figure is frequently but incorrectly attributed to the [UMass Amherst study \(Strubell et al., 2019\)](#), which examined earlier, pre-GPT-3 models).
- GPT-4 (estimated 1.7 trillion parameters, mixture-of-experts): [third-party analysis](#) estimates training energy at 51,773–62,319 MWh (~40–48× GPT-3), based on reported hardware and training duration. The upper bound (~62.3 GWh) has been cited by [Balkan Green Energy News \(2024\)](#) and [Extreme Networks \(2024\)](#). (estimate as OpenAI has not officially disclosed GPT-4's parameter count or training energy).

Real-world equivalents for the central estimate of ~60 GWh:

European households: At the EU average household electricity consumption of approximately 3.6 MWh per dwelling per year, as reported by ODYSSEE-MURE, 60 GWh is equivalent to the annual electricity consumption of roughly 16,700 average European households.

Small-town equivalent: Using Denmark's per-capita electricity consumption of roughly 5,900 kWh per year, broadly consistent with the World Bank electricity-use-per-capita indicator for Denmark, 60 GWh corresponds to the annual electricity consumption associated with about 10,000 Danish residents.

Sources: [ODYSSEE-MURE](#), [World Bank](#)

This final table sets the two model types side by side on a technical lens—deployment, cost, latency, sovereignty exposure and energy—independent of sector or sovereignty tier.

Frontier Models vs Specialist AI

Dimension	Frontier LLM	Specialist AI / SLM
Typical deployment	Centralised cloud	Edge / sovereign / hybrid
Strength	General reasoning	Domain-specific execution
Cost profile	High inference cost	Lower operational cost
Latency	Higher	Lower
Sovereignty exposure	Cross-border risk	Local control
Energy consumption	High	Significantly lower

References

1. Economy | The 2025 AI Index Report, [Stanford Institute for Human-Centered Artificial Intelligence](#), 2025.
2. From PDFs to structured output: automated patent transcription at scale, [Mistral AI](#), accessed September 1, 2026.
3. Welcome to the Pharia-1-LLM-7B documentation, [Aleph Alpha Documentation](#), accessed September 1, 2026.
4. Michael Menzel and Jens Kohl, Small models, high quality: Inside BMW Group's experiments evaluating domain-specific language models, [Google Cloud Blog](#), March 4, 2026.
5. Practical changes could reduce AI energy demand by up to 90%, [UCL News](#), July 8, 2025.
6. Gartner, "Predicts 2025: AI-Powered Analytics Will Revolutionize Decision Making," Peter Krensky, Georgia O'Callaghan, Sumit Agarwal, Pieter den Hamer, David Pidsley, Tong Zhang, 26 February 2025
7. Commission launches InvestAI initiative to mobilise €200 billion of investment in artificial intelligence, [European Commission press release](#), February 11, 2025.
8. Proposal for the Chips Act 2.0, [European Commission](#), June 3, 2026.
9. OSI welcomes the European Union's "Tech Sovereignty" package, [Open Source Initiative](#), June 5, 2026.
10. Announcing the Stargate Project, [OpenAI](#), January 21, 2025.
11. Redeploying Fable 5, [Anthropic](#), accessed September 1, 2026.
12. Grids for data centres: Ambitious grid planning can win Europe's AI race, [Ember](#), accessed September 1, 2026.
13. Proposal for a Cloud and AI Development Act (CADA), [European Commission](#), accessed September 1, 2026.
14. Discover COSC, [Contrôle Officiel Suisse des Chronomètres](#), accessed September 4, 2026.
15. Swiss watch exports in 2025, [Federation of the Swiss Watch Industry](#), January 2026.
16. Apply AI Strategy, [European Commission](#), accessed September 4, 2026.
17. Inside China's AI registry, [Gradient Flow](#), accessed September 1, 2026.
18. Chinese Cyberspace Administration registration of generative AI services, [Digital Policy Alert](#), accessed September 1, 2026.
19. 346 generative AI services registered with China's cyber authority, [CGTN](#), April 9, 2025.
20. Mubashara Akhtar, Anka Reuel, Prajna Soni et al., When AI benchmarks plateau: A systematic study of benchmark saturation, [arXiv](#), February 2026.
21. Will we run out of data? Limits of LLM scaling based on human-generated data, [Epoch AI](#), accessed September 1, 2026.
22. AI reasoning power will be less predictable, Sutskever says, [Reuters](#), December 13, 2024.
23. Jinyan Su, Jennifer Healey, Preslav Nakov and Claire Cardie, Between underthinking and overthinking: An empirical study of reasoning length and correctness in LLMs, [arXiv](#), April 30, 2025.
24. Token economics: The atomic unit of AI value, [FinOps Foundation](#), accessed September 1, 2026.
25. AI cost explosion, [Xpert.Digital](#), accessed September 1, 2026.
26. Data centre electricity use surged in 2025, even with tightening bottlenecks driving a scramble for solutions, [International Energy Agency](#), accessed September 1, 2026.
27. Energy demand from AI, [International Energy Agency](#), accessed September 1, 2026.
28. Data centres metered electricity consumption 2024: Key findings, [Central Statistics Office Ireland](#), 2025.
29. Overcoming energy constraints is key to delivering on Europe's data centre goals, [International Energy Agency](#), accessed September 1, 2026.
30. Confronting AI's growing energy appetite, part 1, [Extreme Networks](#), accessed September 1, 2026.



31. Liv McMahon and Zoe Kleinman, OpenAI pauses UK data centre deal over energy costs and regulation, [BBC News](#), April 9, 2026.
32. Introducing Stargate UK, [OpenAI](#), accessed September 1, 2026.
33. EU launches InvestAI initiative to mobilise €200 billion investment in artificial intelligence, [European Commission](#), February 11, 2025.
34. Peter Belcak et al., Small language models are the future of agentic AI, [arXiv](#), June 2, 2025.
35. Erik Johannes Husom et al., The price of prompting: Profiling energy use in large language models inference, [arXiv](#), July 2024.
36. Harri Renney, Fouad Trad, Michael Mattarock et al., Cloud to edge: Benchmarking LLM inference on hardware-accelerated single-board computers, [arXiv](#), April 2026.
37. Patrick Wilhelm, Thorsten Wittkopp and Odej Kao, Beyond test-time compute strategies: Advocating energy-per-token in LLM inference, [arXiv](#) / [EuroMLSys](#), March 4, 2026.
38. Zhengyu Chen et al., Revisiting scaling laws for language models: The role of data quality and training strategies, [Proceedings of ACL 2025](#), July 2025.
39. Martin Juan José Bucher and Marco Martini, Fine-tuned “small” LLMs (still) significantly outperform zero-shot generative AI models in text classification, [arXiv](#), June 12, 2024.
40. Our contribution to a global environmental standard for AI, [Mistral AI](#), accessed September 1, 2026.
41. SMEs, European Commission, Internal Market, Industry, [Entrepreneurship and SMEs](#), accessed September 1, 2026.
42. Gartner Press Release, “[Gartner Survey Reveals Geopolitics Will Drive 61% of CIOs and IT Leaders in Western Europe to Increase Reliance on Local Cloud Providers](#),” 12 November 2025
43. Europe’s AI imperative, [Mistral AI](#), 2026.
44. AI startup Cohere’s revenue and IPO plans, [CNBC](#), February 13, 2026.
45. Aya Vision: A massively multilingual vision-language model, [Hugging Face](#), accessed September 1, 2026.
46. AI Energy Score, [Salesforce](#), accessed September 1, 2026.
47. SAP EU AI Cloud: A unified vision for Europe’s sovereign AI and cloud future, [SAP News](#), November 2025.
48. T Cloud Public: Sovereign power, [Deutsche Telekom](#), accessed September 1, 2026.
49. Accenture and Mistral AI accelerate enterprise reinvention with scalable AI that delivers strategic autonomy for customers, [Accenture Newsroom](#), 2026.
50. European Union Tech Sovereignty package, ST 10220/2026 ADD 1, [Council of the European Union](#), 2026.
51. Europe’s tech creators’ wake-up call: Let’s act as one Europe now, [Nokia Newsroom](#), May 2026.
52. AI Gigafactories, [European Commission](#), accessed September 4, 2026.
53. Commission proposes Cloud and AI Development Act, [European Commission press release](#), 2026.
54. Achats publics numériques: l’État précise sa doctrine, [Direction interministérielle du numérique](#), February 2026.
55. Targeted consultation on measuring the energy consumption and emissions of AI models and systems, [European Commission](#), 2026.
56. TechDispatch 1/2025: Federated learning, [European Data Protection Supervisor](#), June 10, 2025.
57. Salim Fares, Utilizing LLMs for industrial process automation, [arXiv](#), February 26, 2026.
58. Moe Kayali, Fabian Wenz, Nesime Tatbul and Çağatay Demiralp, Mind the data gap: Bridging LLMs to enterprise data integration, [CIDR 2025](#) / [arXiv](#), December 29, 2024.

About the report

Based on a December 2025 survey of 1,928 executives across 28 countries, combined with Accenture's client experience, this research explores how organizations can balance performance, cost, governance and sovereignty through more deliberate AI deployment choices.

Accenture Research

Amal Sebastian,
Chiara Addis,
Gargi Chakrabarty,
Meredith Trimble

About Accenture

Accenture helps the world's leading enterprises reinvent by building their digital core and unleashing the power of AI to create value at speed for organizations across industries. Our strategy is to be the reinvention partner of choice for our clients and lead in the safe, widespread adoption of AI, and to be the most client-focused, AI-enabled, great place to work in the world. We bring together the talent of our approximately 786,000 people with proprietary assets and platforms, deep process and industry expertise, and leading ecosystem relationships to deliver end-to-end solutions and measurable outcomes at scale. Through our Reinvention Services, we offer broad expertise across Cybersecurity, Digital Core, Finance, Industry and Enterprise, Song, Supply Chain and Engineering, and Talent, with advanced capabilities in AI and Data, Industry and Process, and Technology. We serve approximately 9,000 clients and generated approximately \$70 billion in FY25 revenue.

[Visit us at **accenture.com**.](https://www.accenture.com)

Accenture Research

Accenture Research creates thought leadership about the most pressing business issues organizations face. Combining innovative research techniques, such as data-science-led analysis, with a deep understanding of industry and technology, our team of 300 researchers in 20 countries publishes hundreds of reports, articles and points of view every year. We use generative AI in our research production process. Our research experts review and validate the generative AI outputs with traditional research methods where possible, applying Accenture's Responsible AI standards. Our thought-provoking research developed with world-leading organizations helps our clients embrace change, create value and deliver on the power of technology and human ingenuity.

[Visit us at **accenture.com/research**.](https://www.accenture.com/research)

Disclaimer: The material in this document reflects information available at the point in time at which this document was prepared as indicated by the date provided on the front page, however the global situation is rapidly evolving and the position may change. This content is provided for general information purposes only, does not take into account the reader's specific circumstances, and is not intended to be used in place of consultation with our professional advisors. Accenture disclaims, to the fullest extent permitted by applicable law, any and all liability for the accuracy and completeness of the information in this document and for any acts or omissions made based on such information.

Accenture does not provide legal, regulatory, audit, or tax advice. Readers are responsible for obtaining such advice from their own legal counsel or other licensed professionals. This document refers to marks owned by third parties. All such third-party marks are the property of their respective owners. No sponsorship, endorsement or approval of this content by the owners of such marks is intended, expressed or implied.

Copyright © 2026 Accenture. All rights reserved.
Accenture and its logo are registered trademarks of Accenture.