Rethinking critical AI infrastructure

How on-device infrastructure is emerging as a platform for enterprise AI development and deployment

RESEARCHED BY

Omdia by informa techtarget

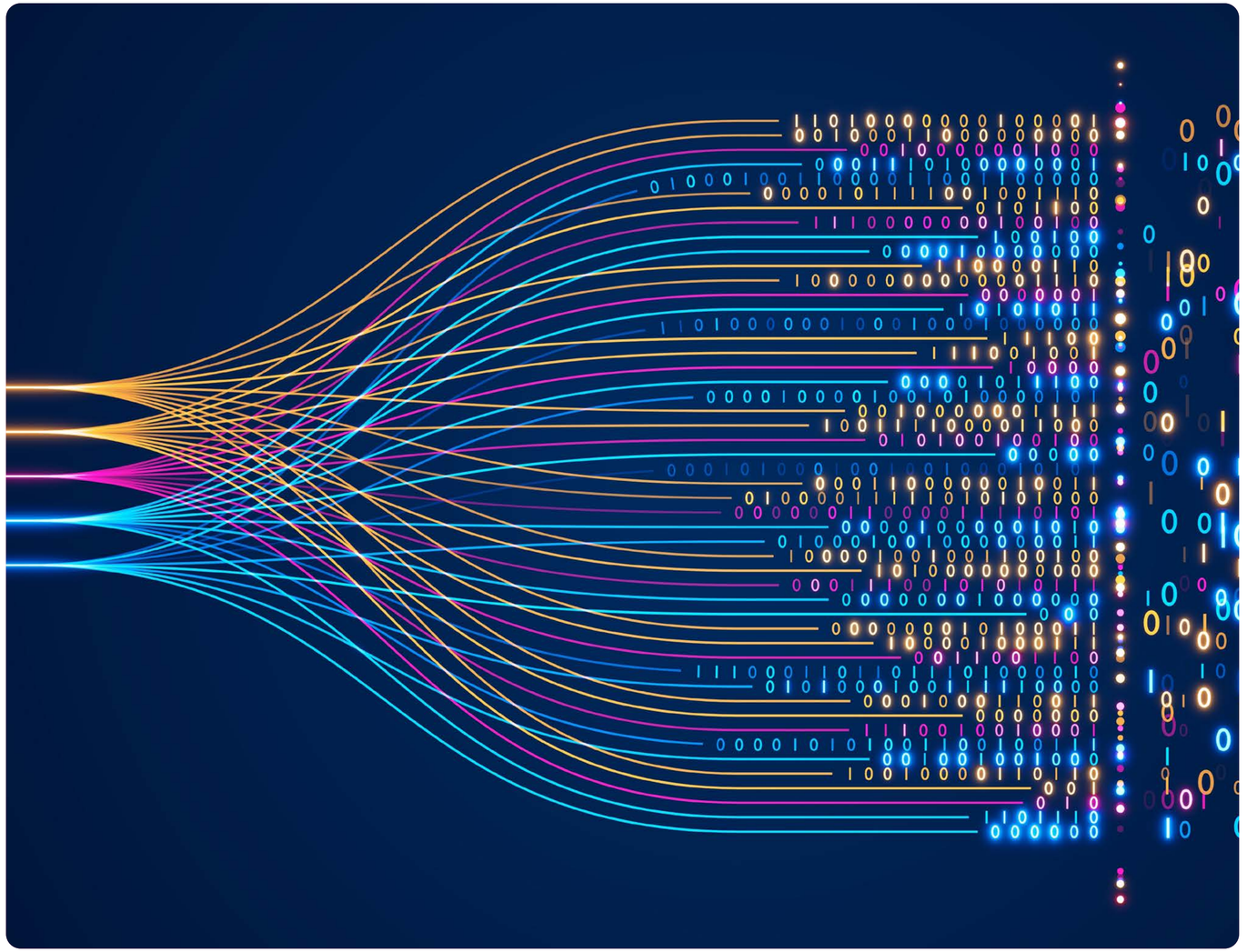
COMMISSIONED BY

Rethinking AI with on-device infrastructure

In today's rapidly evolving AI landscape, enterprises have leaned heavily on cloud and on-premises infrastructure to fuel AI development and deployment. Many organizations adopt a hybrid approach, utilizing the most appropriate infrastructure for the task at hand. However, many haven't fully evaluated where on-device AI delivers measurable advantages.

Insights from Omdia's exclusive study of over 1,500 enterprise technology leaders and practitioners — including CIOs, CTOs, software engineers, and AI specialists — reveal organizations are increasingly questioning whether their current infrastructure truly serves three critical dimensions: their security requirements, unpredictable and ongoing costs, and the need to match infrastructure to workload requirements.

These pillars define the on-device advantage, providing an architectural solution to the security, economic, and workload challenges facing the modern enterprise.



Security

Architectural privacy over procedural workarounds

Three-quarters of enterprises say they are concerned about potential data leakage through cloud services; on-device AI has a clear advantage in that data that never transmits cannot leak in transmission. With 99% of enterprises handling proprietary data in their AI workflows, the ability to process locally is becoming a critical security requirement — eliminating transmission risks and securing AI-assisted workflows without relying on procedural controls.

Economics

Unlimited experimentation and predictable production costs

On-device infrastructure has near-zero marginal cost after initial investment, enabling unlimited experimentation without budget constraints. For production inference, fixed hardware costs replace cloud's unpredictable operational expenses that scale with success. And unlike cloud seats that become stranded assets if user adoption disappoints, on-device hardware retains full productivity value regardless of AI usage patterns.

Workload viability

Matching infrastructure with models

Organizations don't need massive foundation models to unlock value — 57% of enterprise models are under 10 billion parameters, well within the capabilities of modern devices like MacBook Air or the entry-level MacBook Pro. Surprisingly, the shift towards on device is even more pronounced at the larger end: enterprises who primarily use on-device infrastructure are nearly twice as likely to deploy models over 100 billion parameters.

The enterprise AI landscape today

Mature organizations are moving to hybrid by leveraging more on-device

Organizations at advanced stages of AI maturity show a higher adoption of hybrid deployment approaches. This indicates that each infrastructure type serves distinct purposes for companies that have or plan to scale AI across multiple business functions.

On-device infrastructure handles development workflows, sensitive data processing, distributed deployment of smaller models across teams, and production inference for workloads requiring local processing. While on-premises GPU clusters may be preferred for batch training and multi-user shared environments, cloud infrastructure provides elastic scaling and production APIs requiring high availability.

Notably, organizations don't choose exclusively, they combine infrastructure types strategically. Among organizations using Mac for on-device AI workloads, 56% also deploy locally with NVIDIA GPUs, demonstrating that on-device platforms complement rather than replace specialized compute resources.

The pattern reflects workload economics: security-sensitive development happens on controlled workstations, large-scale training uses dedicated clusters when needed, and production deployment matches infrastructure to scale requirements.

Organizations deploy infrastructure where it delivers the best combination of performance, cost, and control for specific tasks. A financial services firm might prototype fraud detection models on workstations, train production versions on GPU clusters when scale requires it, deploy high-volume inference to cloud, and run branch-level models on local systems, using each architecture where it provides optimal value.

Platform diversity provides strategic flexibility. Organizations can match infrastructure to requirements rather than forcing all workloads onto a single architecture. As AI capabilities evolve and workload patterns shift, infrastructure portfolios adapt more readily than single-platform commitments.

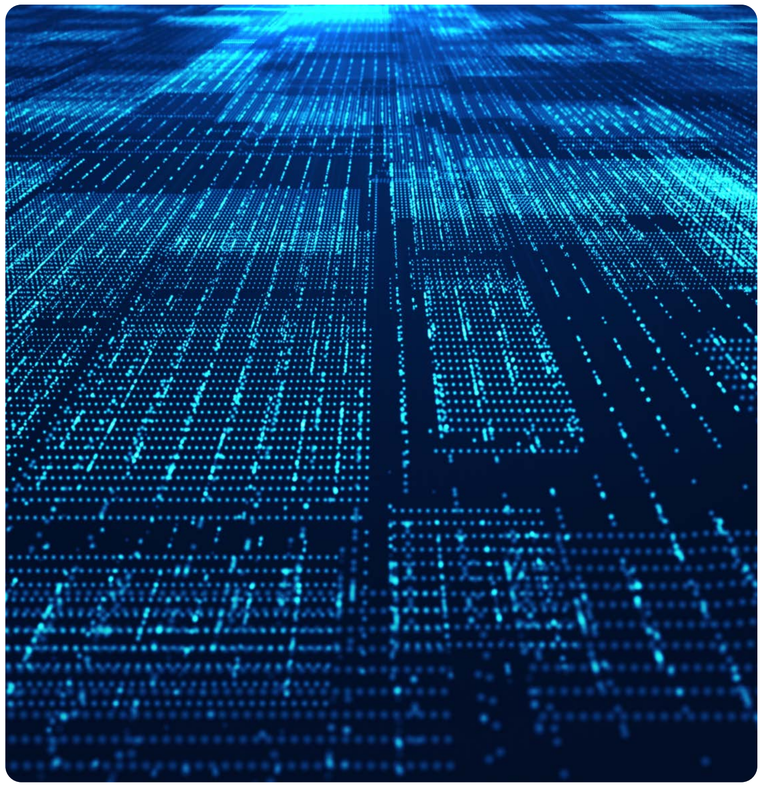


Figure 1: 41% actively using AI in business processes

What is your organization's current AI maturity?

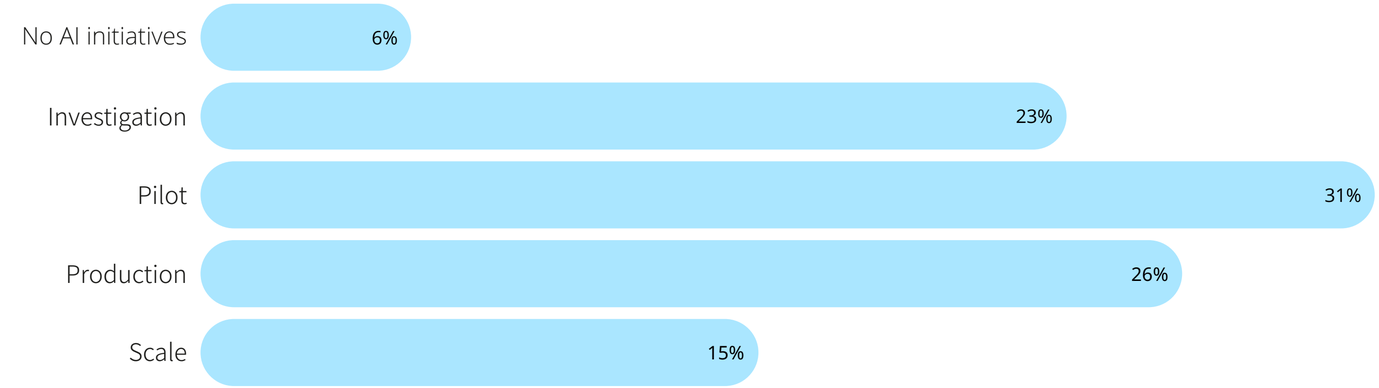
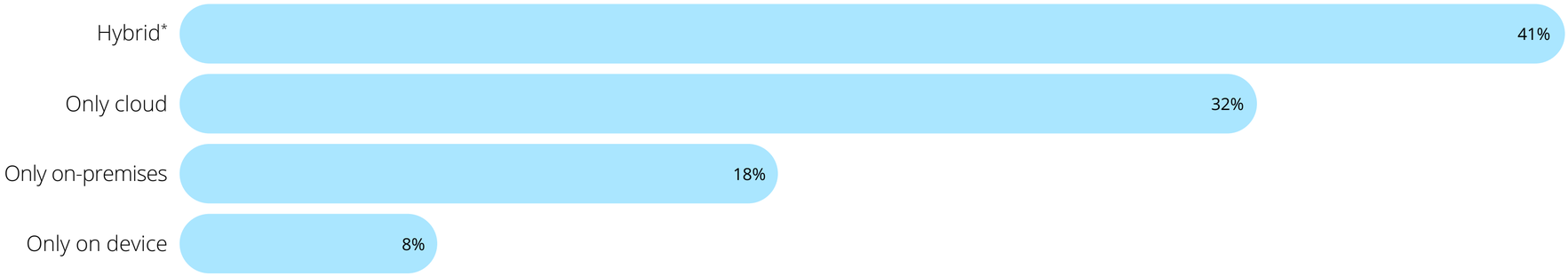
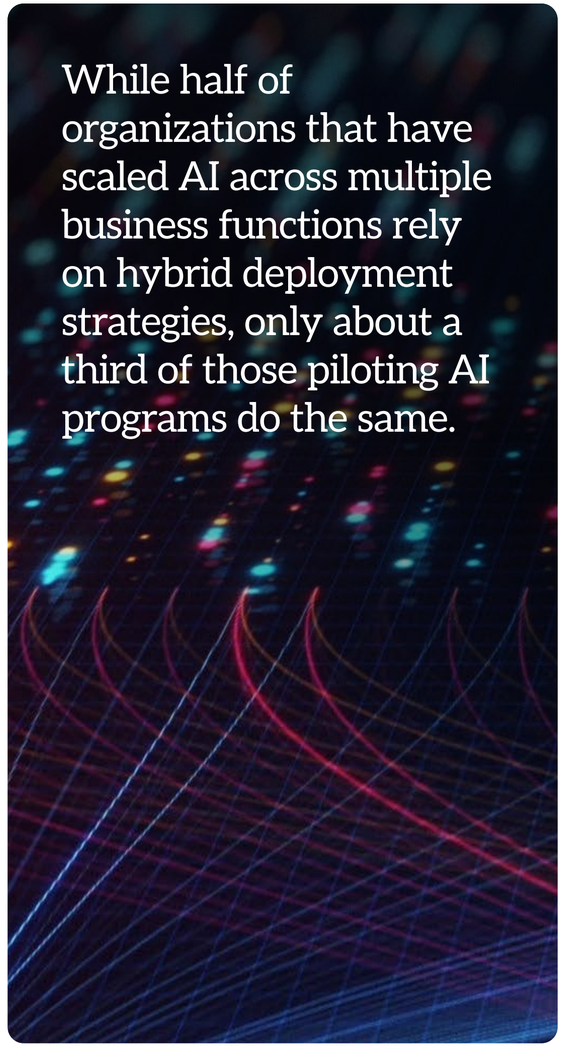


Figure 2: 73% use cloud or hybrid infrastructure to deploy

Where do you deploy AI workloads?



SOURCE: OMDIA
NOTES: N=1584, *HYBRID STRATEGIES ARE WHEN AI DEPLOYMENTS MIX A COMBINATION OF CLOUD, ON-PREMISES, AND ON-DEVICE COMPUTE. PREDOMINANTLY IT MIXES CLOUD AND ON-PREMISES SERVER.



Organizations are moving some AI workloads on-device

Over a third of organizations plan to shift more AI workloads on device in the next 12 months. These organizations have identified specific workloads where on-device infrastructure addresses gaps in their current approaches.

For security-sensitive workloads, procedural controls create ongoing overhead. Each project involving proprietary information, customer data, or regulated content triggers documentation requirements, audit cycles, and transmission risk assessments. On-device processing reduces this overhead by keeping data local. What doesn't transmit doesn't require transmission security reviews or third-party access approvals.

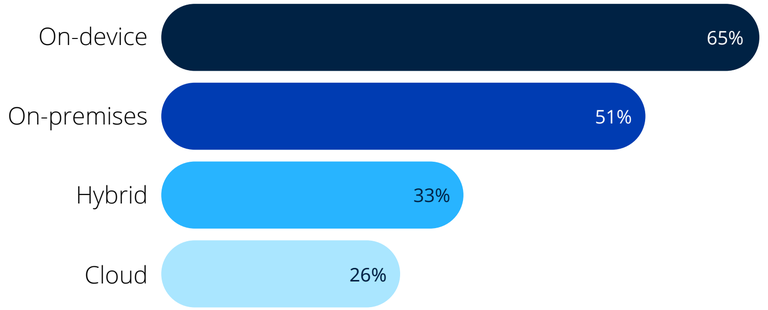
For development workflows, iteration speed determines how quickly teams can refine models and applications. Teams need to test hypotheses, optimize parameters, and validate approaches without waiting for cloud resource allocation or monitoring usage quotas. On-device infrastructure removes these constraints: development teams work directly on local datasets with no transmission delays, no token limits, and no marginal costs per iteration.

Organizations aren't abandoning cloud infrastructure, they are following the same impulse that led many of them to adopt hybrid strategies in the first place — they need the right tool for the right job. Each architecture serves distinct purposes. The pattern among AI-mature organizations is strategic diversification rather than single-platform dependency.

Among organizations using on-premises infrastructure, 51% plan to add more on-device capabilities. For firms already using on-device infrastructure, 65% plan to shift more workloads on-device.

Figure 3: Share of organizations planning to shift more AI workloads on-device, split by current deployment strategy

How do you expect your on-device versus on-premises and cloud AI split to change in the next 12 months?*



SOURCE: OMDIA
NOTES: N=1568, *ONLY "MORE ON-DEVICE" RESPONSES SHOWN PER CURRENT PRIMARY INFRASTRUCTURE

On device supplements AI vendors by filling their gaps

Only 9% report no significant gaps with their strategic AI partner

Strategic AI partnerships deliver capabilities many enterprises couldn't build alone: pre-trained models, deployment infrastructure, and specialized tools. But partnerships also introduce constraints. They require data access, create pricing complexity, and generate integration dependencies. Organizations discover these frictions at different stages: security-conscious firms encounter them when deploying on sensitive data, while cost-conscious organizations discover them when bills arrive.

Omdia's research reveals widespread dissatisfaction. Insufficient data privacy controls, unpredictable costs, and poor integration with existing systems emerge as consistent pain points across organizations working with major cloud providers, traditional enterprise software vendors, or AI-native companies. Only 9% of enterprises report their partnerships fully meet their needs.

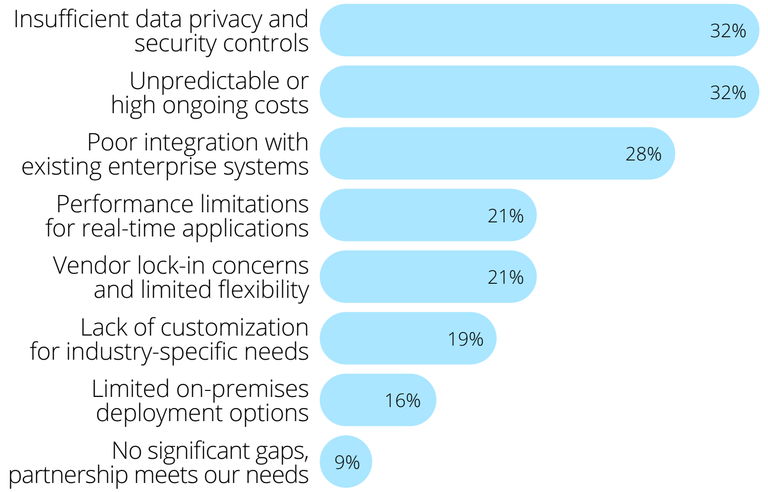
This dissatisfaction creates an opening for a hybrid approach. Organizations aren't rejecting AI partnerships, they are seeking the flexibility to solve the specific operational challenges that cloud-only strategies cannot fully address. By developing locally on controlled infrastructure, organizations gain a high-speed environment for sensitive data and real-time tasks, providing both predictable costs and direct data access.

Supplementing infrastructure capacity with on-device computing closes the functional gaps inherent with cloud-only partnerships. It streamlines operations by enabling AI at every stage of a workflow — from annotation to cleaning of sensitive data — where privacy controls or egress costs previously acted as a barrier. Using on-device infrastructure as a foundation ensures that strategic cloud resources are prioritized for large-scale distribution and centralized coordination rather than being consumed by tasks better suited for on device.

On-device infrastructure gives you a secure foundation from which to manage your entire AI ecosystem.

Figure 4: Security, cost, and poor integration are consistent partnership pain points

Top two limitations with current AI partnership? (Select up to two)



SOURCE: OMDIA
NOTES: N=1352

Enterprises track visible costs, missing the key drivers

Over two-thirds track cloud costs but less than half track security-related costs

Cloud compute, security, and software licensing investments emerge as top concerns for AI infrastructure's Total Cost of Ownership (TCO). But concern doesn't automatically translate to accurate tracking and attribution.

Among those ranking cloud compute as a TCO concern, 72% report systematically, tracking it as a distinct line item in their AI infrastructure budget. This makes sense since cloud providers deliver frequent invoices with clear per-service breakdowns.

Security tells a different story. Despite being the second largest TCO concern, fewer than half track security costs as a discrete component of AI infrastructure TCO. Security expenses are real - compliance overhead, audit cycles, incident response, specialized tooling - but they're distributed across existing budgets and organizational silos, making attribution far less straightforward.

This visibility gap creates strategic blind spots. Organizations optimize what they can measure. When cloud costs arrive in an itemized bill while security investments remain fragmented across procurement, personnel, and operational budgets, decision-makers naturally focus on the visible number. Yet the invisible costs may be larger and more critical to long-term success.

The gap also obscures possible cost advantages for deployment strategies beyond exclusively cloud. Organizations may systematically undervalue hybrid approaches because the potential cost savings aren't immediately visible, while the upfront investments can dominate cost evaluations and overshadow long-term value. This visibility asymmetry can bias decision-making toward cloud-centric strategies even when total cost of ownership favors a more balanced approach.

Attribution gaps and budget structures can hide TCO advantages of hybrid deployment strategies

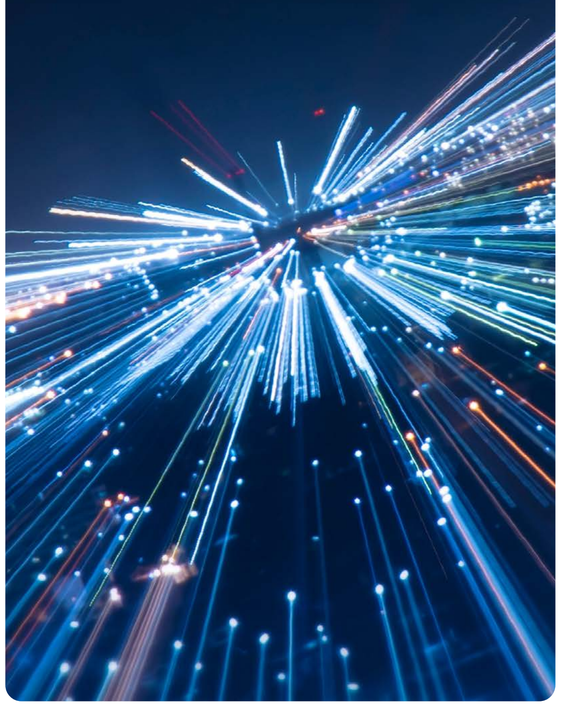
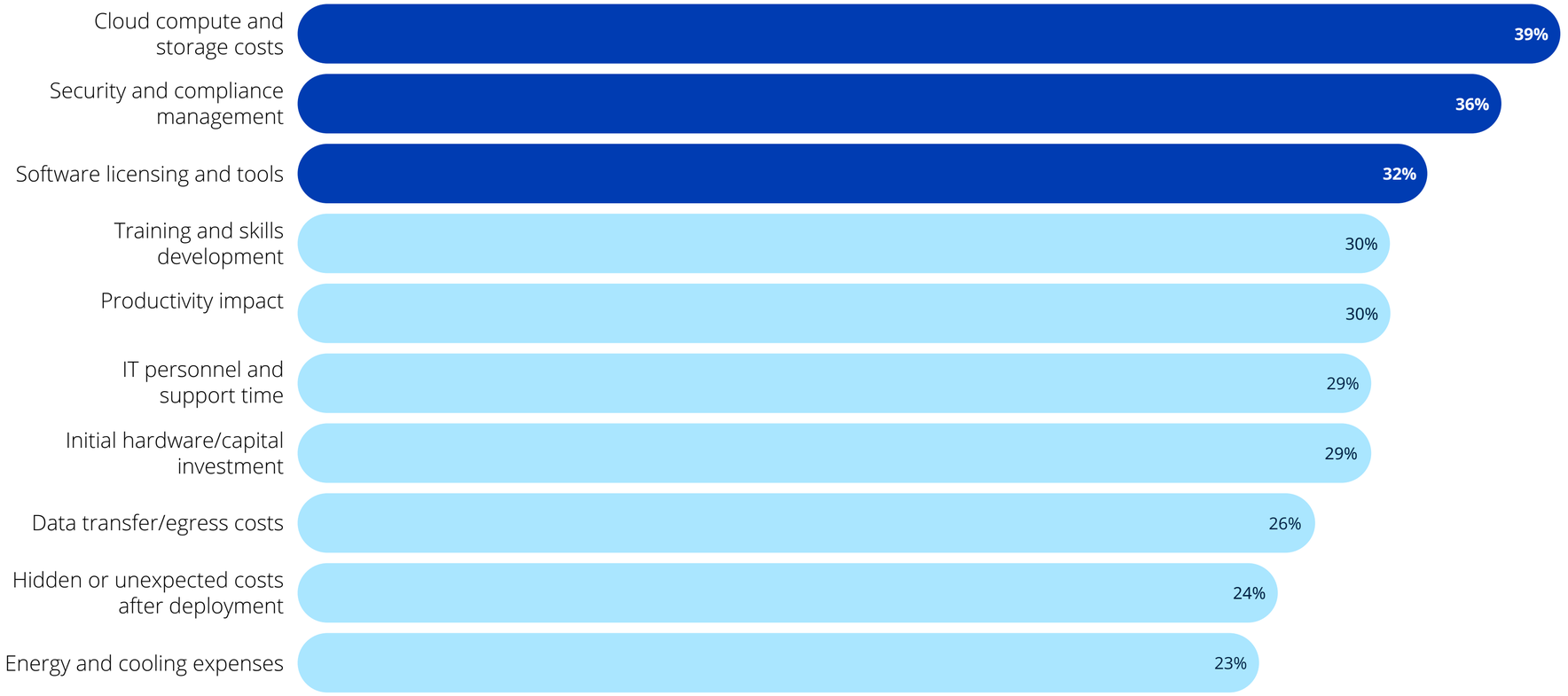


Figure 5: Cloud, security, and software licenses lead as top cost concerns for AI infrastructure's total cost of ownership

What are your top concerns regarding the TCO of AI infrastructure (Select three)



SOURCE: OMDIA
NOTES: N=1584

The security model is right

Security as friction vs. security as architecture

Minimizing data exposure by design

Current cloud-centric approaches create an inherent tension: to leverage cloud AI capabilities, data must leave the organization and be transmitted to third-party infrastructures. This transmission creates exposure that 76% of enterprises identify as a concern. The concern is not necessarily from intentional attacks, but from inadvertent transmission or configuration errors that are far more common.

This concern intensifies when organizations fine-tune or train models using their most valuable information: customer data, internal knowledge bases, and proprietary research. While 87% of organizations recognize that on-premises or on-device is critical or important, cloud-first architectures can only offer risk mitigation, not risk elimination.

The fundamental challenge: innovation requires data access, while data protection requires restricting transmission. Development teams need sensitive, proprietary data to build effective models. Each new project triggers compliance reviews cycles, creating planning overhead that limits experimentation velocity on new use cases.

Across regulated industries, this tension is particularly acute. Healthcare organizations training models on patient health information face mandated controls. Financial services using transaction histories and customer financial data must navigate data residency restrictions. And professional services firms working with confidential client data face contractual prohibitions on third-party data access.

Security requirements are one of the top reasons enterprises say they are using on-device infrastructure. Concerns drive preference for architectural solutions over procedural controls.

⚠ CONCERN

76%

Of enterprises are concerned or very concerned about potential data leakage through cloud services

✓ PREFERENCE

87%

Of enterprises say keeping sensitive data on-premises or on-device is important

On-device AI development strengthens security

Greatly reducing risks through architectural privacy over procedural workarounds

On-device development resolves the innovation-security tension architecturally. Data never leaves the secure environment, eliminating entire categories of security review and compliance overhead. This intensifies with data sovereignty requirements — when organizations operate across jurisdictions they face country-specific mandates requiring data to remain within geographic boundaries. Cloud architectures can complicate compliance; on-device processing makes it easier for compliance.

The security advantage for on device is straightforward: data that never transmits cannot leak in transmission. Development teams can work directly on local datasets without requiring third-party infrastructure access. Models train where data resides, resulting in no data transfer costs to manage, no transmission encryption to verify, no cloud provider controls to audit, and no cross-border data transfer documentation.

On-device infrastructure creates previously impossible workflow advantages: using AI to prepare data for AI. Organizations deploying to cloud environments must clean and prepare sensitive data before transmission, scrubbing personally identifiable information or redacting confidential details. Without on-device AI capabilities, this data preparation happens using traditional tools, creating significant bottlenecks.

With on-device AI, development teams can use AI tools locally to accelerate the cleaning and preparation process, using models to identify sensitive fields, automate redaction, validate data quality, and transform datasets, all before any data leaves the secure environment.

Organizations can innovate at full speed on sensitive data without compromise. Development cycles that previously required days of security review overhead compress dramatically. Engineering teams regain velocity and security teams gain confidence through architectural guarantees rather than procedural controls.

This doesn't eliminate all security requirements. Physical device security, access controls, and audit logging remain essential. But it protects against the transmission-related security concerns that cloud architectures endure by design. On-device infrastructure removes the architectural risk, and also removes the procedural friction.


99%
of enterprises use proprietary data when training or fine-tuning AI models, making on-device AI a critical workflow advantage in eliminating transmission risks, reducing procedural overhead, and enabling AI-assisted data preparation workflows

The economics are right

On-device AI development eliminates the experimentation tax

AI development thrives on iteration. The best solutions emerge when teams can experiment freely. This allows them to test hypotheses, refine models, and explore creative approaches without constraints. For the 54% of organizations still investigating or piloting AI solutions, establishing the right infrastructure balance early translates directly to cost savings as experimentation intensifies and workloads scale to production.

On-device infrastructure changes the economics of experimentation. A fixed hardware investment unlocks unlimited iterations. Once hardware is deployed, every iteration, test run, and validation cycle costs nothing additional.

Whereas cloud-centric strategies accumulate costs for each exploration. Every experimental run accumulates charges from compute time, data transfer, and storage. When a promising approach fails after 20 iterations, those 20 iterations still appear on the invoice. Learning from unsuccessful approaches costs just as much as successful ones.

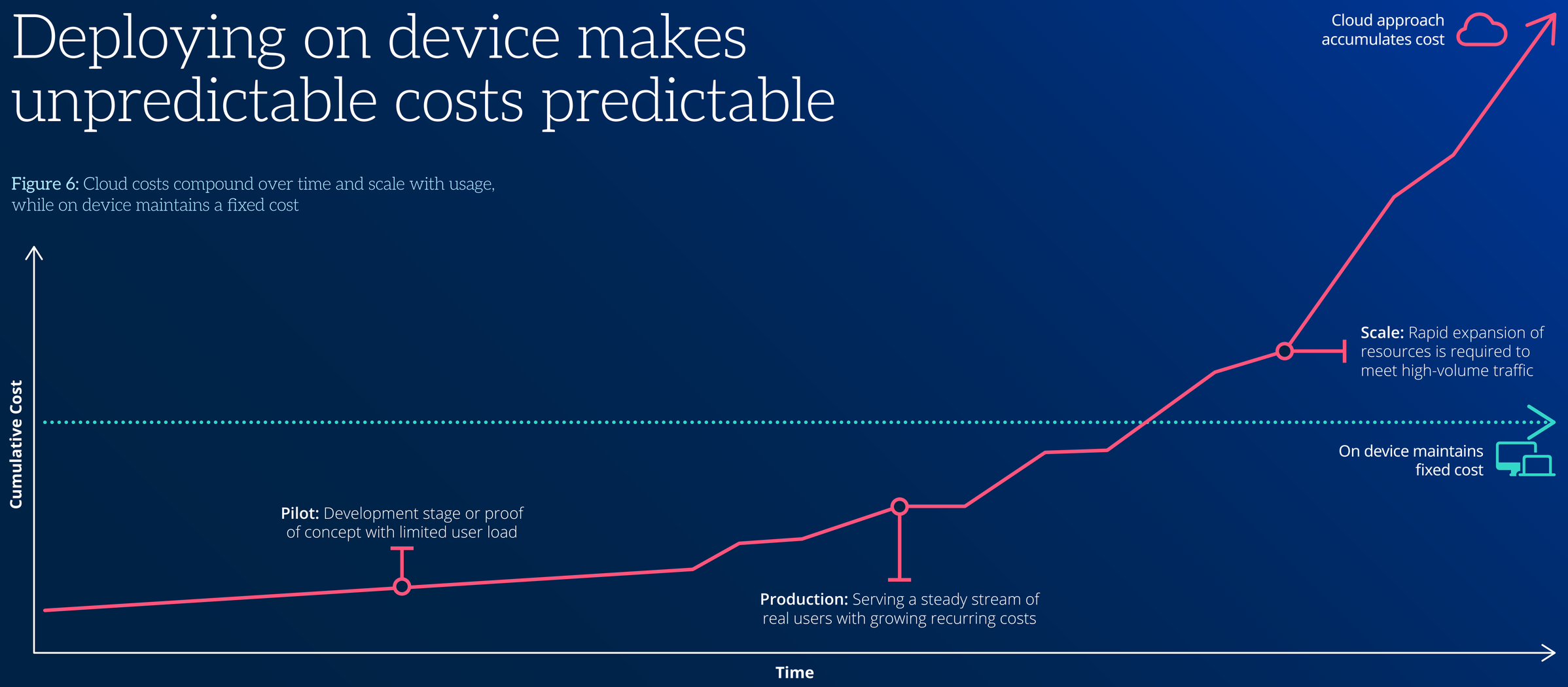
For longer running projects, organizations face materially different total costs between cloud and on device. Cloud charges compound with each version, while hardware that can run AI fully on device delivers unlimited experimentation at zero marginal cost. And the same hardware investment serves multiple projects over its lifecycle.

The strategic advantage compounds over time. Development teams can validate thoroughly, explore dead ends, and iterate rapidly without budget anxiety. The constraint becomes human attention and engineering time rather than infrastructure economics. This removes the financial friction that forces teams to limit experimentation, skip validation runs, or avoid exploring alternative approaches.

On-device infrastructure doesn't just reduce costs, it enables the rapid iteration cycles that produce better AI. Teams ship better products and solutions because they can afford to test more and learn from more failures.

Deploying on device makes unpredictable costs predictable

Figure 6: Cloud costs compound over time and scale with usage, while on device maintains a fixed cost



Deploying AI on device protects organizations against runaway inference costs and unused cloud licenses

Beyond development, infrastructure choice has a significant impact on the costs of production inference. Cloud-based AI follows two primary pricing models: per-API-call charges that accumulate with every request, or seat-based subscriptions that remain fixed regardless of usage intensity. Both create economic challenges that compound costs over time.

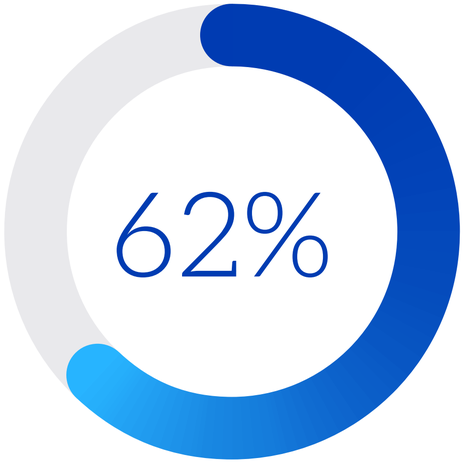
On-device infrastructure fundamentally changes the economics of AI inference. After the initial hardware investment, every inference request costs nothing additional. Generating 1,000 or 100,000 tokens on device maintains the same total cost of ownership. This predictability proves particularly valuable for organizations scaling AI across their workforce, where departmental silos often make it difficult to track cloud costs accurately and in real-time across the enterprise.

Seat-based licensing introduces a different challenge. Enterprises purchasing cloud subscriptions, expecting broad adoption, often see actual usage fall short of expectations. 32% of organizations said they were not satisfied with getting end users to adopt AI tools, a further 30% were only somewhat satisfied. These unused licenses represent ongoing costs with no productivity return.

On-device infrastructure avoids this risk entirely. Whether used for AI workloads or standard computing tasks, the hardware continues to deliver value over its lifecycle. Organizations can eliminate the waste inherent to unused subscriptions while maintaining flexibility as requirements evolve.

Figure 7: Per-seat cloud licenses cost the same regardless of usage

Assume an organization purchases 1,000 licenses of cloud-based AI at $30-per-user-per-month and just 10% go unused, that is $3,000 per-month gone to waste.



of organizations say they are somewhat satisfied or not satisfied with getting end users to adopt AI tools.

The workloads are right

Enterprise AI models serve specific purposes

The models making headlines with trillions of parameters in pre-training have created a perception that all meaningful AI work requires hyperscale data centers. This does not reflect the reality of enterprise deployment.

95% of organizations are not building models from scratch; they are fine-tuning and running inference on existing models. These tasks are significantly less compute-intensive.

Further still, 57% of models enterprises are typically deploying are less than 10 billion parameters. Even for the largest models with hundreds of billions of parameters, there are options available to develop and deploy locally on clustered desktop devices. The infrastructure requirements for enterprise AI development, in other words, frequently match the capabilities of modern devices.

Figure 8: 95% of organizations fine-tune and run inference

What is the typical size of AI models your organization deploys? Do you run inference, fine-tune, or train your own models?

A majority of enterprise models are under 10 billion parameters, well within the capabilities of modern devices like MacBook Air or an entry-level MacBook Pro

SOURCE: OMDIA
NOTES: N=1529

Organizations already deploying on device are twice as likely to run models over 100 billion parameters

A 10 billion parameter model requires approximately 12GB of memory when quantized. This is comfortably handled by devices with 24GB memory, which provide ample headroom for simultaneous applications and workflows. Even base configurations with 16GB can support smaller 1B parameter models.

Unified memory architecture enables on-device deployment of even the largest models

A common assumption about AI infrastructure is that larger models automatically necessitate cloud deployment. Omdia's analysis found no significant relationship between the size of models an organization deploys and their infrastructure deployment preferences.

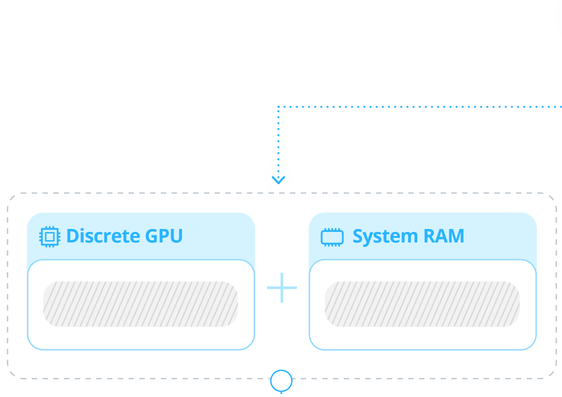
Organizations running 100B+ parameter models showed no greater likelihood of preferring cloud infrastructure than those running 10B models. The factors that actually predict deployment choice are security requirements, cost considerations, and development workflow needs.

This pattern reflects a technical reality: the constraint isn't model size per se, but rather available memory capacity and architecture. The determining factor is whether infrastructure can provide sufficient memory for the workload, not whether the model crosses some arbitrary size threshold.

For the 20% of enterprise deployments using models over 100B parameters, memory availability becomes the critical consideration. Infrastructure that can accommodate 100GB+ memory requirements, whether through single-system configurations with large unified memory pools or clustered multi-device deployments, enables local processing of these larger models.

Memory architecture determines viability: unified memory systems enable 100GB+ configurations through a single-device rather than multi-GPU infrastructure.

Why memory architecture matters



High-end GPUs provide 24-48GB of high memory bandwidth. When models exceed the GPUs memory capacity, performance degrades as processing falls back to slower system RAM.

Unified Memory

Pools memory and eliminates the bottleneck, providing access to larger memory configurations at consistent bandwidth. For example, a 128GB MacBook Pro with M4 Max can run 70B parameter models at Q8 precision locally, while a 512GB Mac Studio with M3 Ultra supports 300B+ parameter models — all without the performance degradation typical of discrete GPU systems.

Figure 9: Model memory requirements for model inference, by precision

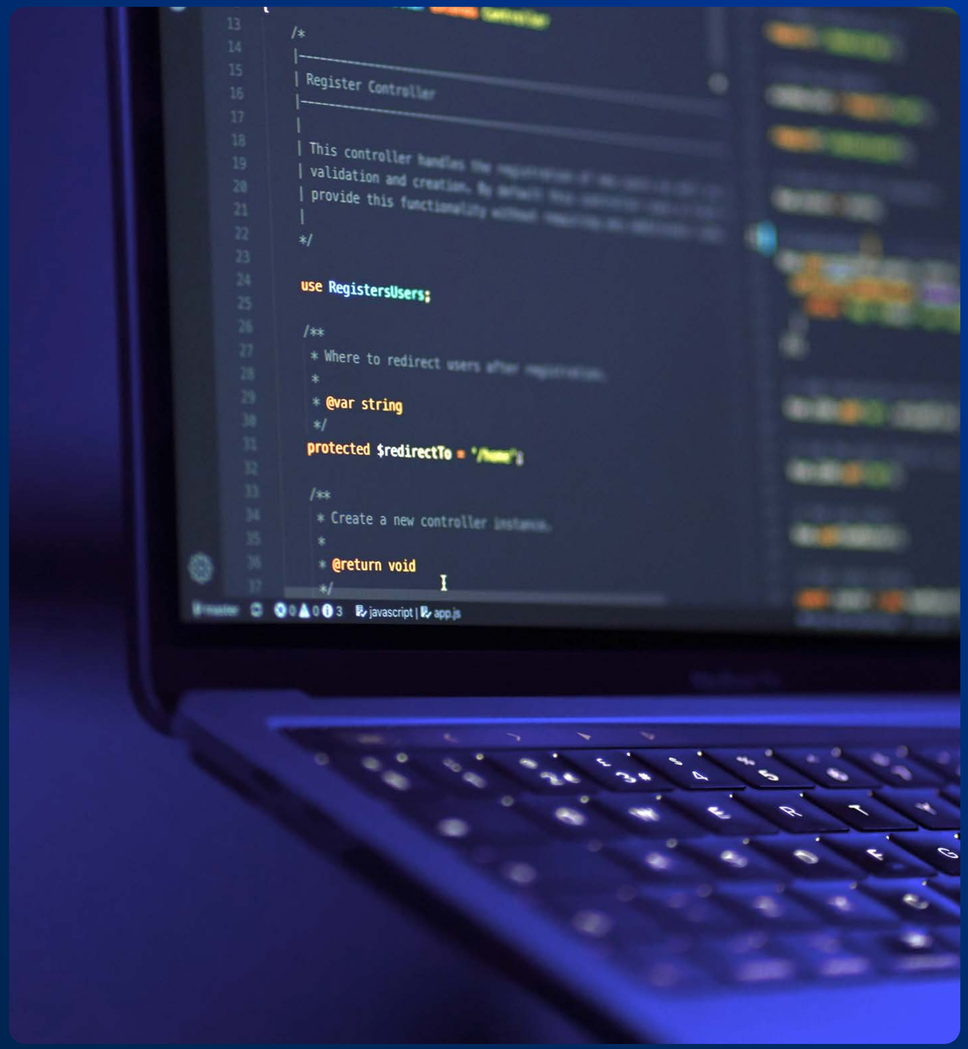
Includes model weights + 20% inference overhead

NUMBER OF PARAMETERS AND INDICATIVE CAPABILITIES	Q8	FP16	FP32
1 billion: Text classification, sentiment analysis, simple extraction	1.2GB	2.4GB	4.8GB
10 billion: Conversational AI, Q&A, summarization, basic code assistance	12GB	24GB	48GB
15 billion: Document analysis, translation, structured content generation	18GB	36GB	72GB
70 billion: Advanced code generation, multi-step tasks	84GB	168GB	336GB
100 billion: Multi-modal analysis, research synthesis	120GB	240GB	480GB
300 billion: Frontier class, complex reasoning, autonomous agent workflows	360GB	720GB	1440GB

INDICATIVE CAPABILITIES BY MODEL SIZE (ACTUAL PERFORMANCE VARIES BY ARCHITECTURE AND FINE-TUNING).

Q8 (8-BIT QUANTIZATION), FP16 (16-BIT FLOATING POINT), AND FP32 (32-BIT FLOATING POINT) REFER TO THE NUMERICAL PRECISION USED TO STORE MODEL WEIGHTS. LOWER PRECISION REDUCES MEMORY REQUIREMENTS WHILE PRESERVING OUTPUT QUALITY — 8-BIT QUANTIZATION IS WIDELY REGARDED AS NEAR-LOSSLESS FOR MOST INFERENCE TASKS. Q8 IS THE MOST MEMORY-EFFICIENT OPTION SHOWN HERE; IN PRACTICE, 4-BIT QUANTIZATION IS ALSO WIDELY USED FOR ON-DEVICE DEPLOYMENT WHERE MEMORY IS MORE CONSTRAINED.

SOURCE: OMDIA



Companies heavily investing in AI workloads choose Mac

The relationship between AI strategy and platform choice reveals a significant pattern. Organizations that build AI solutions in-house adopt Mac for AI workloads at nearly double the rate of organizations buying commercial solutions.

This is a clear case of revealed preference. In-house development teams have the deepest technical requirements. They need iteration flexibility without marginal cost, absolute data security for proprietary information, and performance they can directly observe, control, and optimize against.

These requirements align with on-device infrastructure strengths, and the developers doing the most intensive local development choose Mac most frequently.

The organizations who find it imperative that their infrastructure platforms serve development workflows need iteration speed, local data access, reliable performance — find their strongest adoption among organizations with sophisticated AI capabilities. The teams that know what they need are choosing Mac.

When technical practitioners control platform decisions, they optimize for the tools that make them productive. Mac's adoption advantage among in-house developers suggests that something in the platform serves their workflows that alternatives don't provide.

Organizations developing in-house choose Mac primarily for its performance and developer familiarity.

...at nearly double the rate of commercial solution buyers. When technical practitioners control platform decisions, they optimize for the tools that make them productive — **with Mac adopted at 1.8x the rate for in-house AI development.**

Key takeaways

Building balanced AI infrastructure

This research reveals a gap between current infrastructure approaches and enterprise AI requirements. Organizations face security barriers that add overhead, cost structures that hide operational expenses, and workflow constraints that slow development velocity.

The path forward involves completing infrastructure portfolios with on-device capability, enabling organizations to match workload requirements to infrastructure, rather than routing all AI initiatives through cloud-centric approaches. Organizations that build this flexibility establish advantages that compound as AI capabilities evolve and workload patterns shift.

Match AI infrastructure to workloads

Mandate workload-infrastructure alignment based on model size, data sensitivity, usage patterns, and user requirements. The majority of enterprises running models under 10B parameters have a different optimal infrastructure than the AI researchers building frontier models. Development workloads have different requirements than production serving. Resist one-size-fits-all infrastructure defaults that create misalignment.

Solve security architecturally, where possible

For projects involving proprietary or sensitive information, evaluate infrastructure options that eliminate transmission risks through architecture rather than managing them through process. On-device processing isn't the only approach, but for development workflows and inference on smaller models, it removes entire categories of compliance overhead and security review delays.

Make unpredictable costs predictable

Expand TCO measurement beyond easily tracked line items like cloud compute bills and hardware purchases. Quantify security review delays, compliance overhead, productivity impact from resource constraints, and the marginal cost of experimentation under usage-based pricing. Organizations optimize what they measure, incomplete cost visibility drives suboptimal infrastructure decisions.

Enable in-house technical teams to build hybrid solutions

Empower in-house development teams to incorporate diverse infrastructure types where each serves specific workflows optimally. The teams doing the most sophisticated AI work are already building hybrid portfolios that match the infrastructure to workloads. Create procurement frameworks and technology policies that enable rather than restrict this strategic diversification.

The AI advantage of Mac is intentional, not incidental

The preference for Mac among organizations is not arbitrary. It stems from an architecture that directly addresses the concerns and gaps of cloud-centric or on-premises-centric approaches.

Apple silicon's unified memory architecture in Mac provides ample capacity for typical enterprise AI workloads, while also offering unparalleled scale in memory capacity for a given price point, allowing larger models to run on device. This addresses the workload mismatch, providing ample performance without requiring ongoing cloud costs or expensive on-premises server options.

Data security

On-device processing eliminates transmission risks by design. For the vast majority of organizations that need to keep sensitive data local, Mac provides architectural privacy rather than procedural workarounds.

Unified memory for AI inference

20% of enterprise deployments run models over 100B parameters — too large, or costly, for discrete GPU VRAM, but well suited to Apple silicon’s unified memory, which offers up to 128GB on MacBook Pro and 512GB on Mac Studio.

Predictable cost model

Fixed hardware investment replaces usage-based cloud costs that scale over time. Development teams gain unlimited experimentation without marginal costs, while production deployments avoid costs that scale with usage.

Developer preference

Organizations building AI in-house choose Mac 1.8x more frequently than those buying commercial AI solutions. When technical teams control platform decisions, they optimize for performance, security, and developer familiarity.

Appendix

About

Omdia

Expertise that listens. Advice that leads.

Omdia is a global technology research powerhouse, established following the merger of the research division of Informa TechTarget (Ovum, Heavy Reading, Tractica, and Canalys) and the acquired IHS Markit technology research portfolio. We combine the expertise of more than 300 analysts across the entire technology spectrum, covering over 200 markets. We publish over 3,000 research reports annually, and cover thousands of technology, media, and telecommunications companies. Our exhaustive intelligence and deep technology expertise enable us to uncover actionable insights that help our customers connect the dots in today's constantly evolving technology environment and empower them to improve their businesses – today and tomorrow.

About this research – methodology and sample details

The Critical AI Infrastructure Survey was fielded and analyzed in Q4 2025 by Omdia, covering enterprise IT decision-makers across five global markets: United States, United Kingdom, Germany, China, and India.

A total of 1,584 responses were gathered, representing organizations with 500+ employees in Healthcare, Financial Services, and Professional Services. All participants held either direct budget authority for technology purchases or significant influence over their organization's AI strategy. Respondents included, but were not limited to, CIOs, CTOs, software engineers, and AI specialists.

The research examined enterprise attitudes toward AI infrastructure decisions, security and privacy requirements, total cost of ownership considerations, and platform preferences for AI workloads.

In which country are you located?

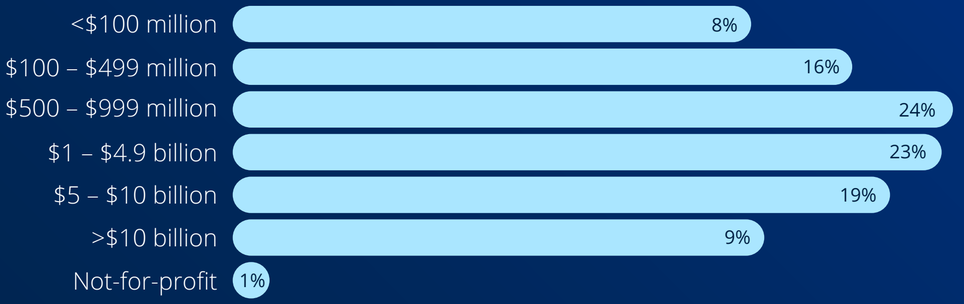
COUNTRY	
USA	20%
UK	20%
Germany	20%
China	20%
India	20%

Which is Your Primary Industry?

INDUSTRY	
Healthcare	33%
Financial services	34%
Professional services	33%

How many employees globally?

What is your organization's annual revenue?



NOTE: PERCENTAGES MAY NOT ADD UP TO 100% DUE TO ROUNDING

Omdia
by informa techtarget

The Omdia team of 400+ analysts and consultants are located across the globe

Americas

Argentina
Brazil
Canada
United States

Asia-Pacific

Australia
China
India
Japan
Malaysia
Singapore
South Korea
Taiwan

Europe, Middle East, Africa

Denmark
France
Germany
Italy
Kenya
Netherlands
South Africa
Spain
Sweden
United Arab Emirates
United Kingdom

Omdia

E insights@omdia.com
E consulting@omdia.com
W omdia.tech.informa.com

OmdiaHQ
Omdia

Citation Policy

Request external citation and usage of Omdia research and data via citations@omdia.com