

AI beyond data centers:

On-device, on-demand,
everywhere

Our firm

We make innovation work for the world

We work with businesses, governments, and multilateral institutions to unlock new markets, drive commercial success, and create solutions that advance humanity.

Our multidisciplinary team includes geopolitical analysts, regulatory specialists, and communications strategists, as well as economists, engineers, physicians, and scientists.

We combine on-the-ground expertise across every region with deep specialism, letting us go beyond conventional consulting to design and execute bold, end-to-end strategies that help you scale sustainably and deliver meaningful impact for society.

We help ambitious organisations turn ideas into real-world outcomes quickly and decisively. Whether you're expanding into new markets, launching new solutions, or tackling global challenges, we combine commercial insight, public-policy and regulatory analysis, and deep networks to help you move with confidence and clarity.

Find out more here: accesspartnership.com⁷

Copyright © 2025. The information contained herein is the property of AA Access Partnership Limited and is provided on condition that it will not be reproduced, copied, lent, or disclosed, directly or indirectly, nor used for any purpose other than that for which it was specifically furnished.

AA Access Partnership Pte Ltd

Contents

- Executive Summary
- 1. AI’s next gold rush
- 2. The data center crunch.....
- 3. The breakthrough
- 4. Beyond efficiency
- Ecosystem-wide benefits
- Industry-level benefits
- Consumers and end-users
- 5. The path forward
- 6. Appendix.....
- Methodology for estimating future AI compute gap
- Methodology for the regional AI compute gapEmergency services
- Additional trends in the AI ecosystem today
- 7. Endnotes.....

AI beyond data centers: On-device, on-demand, everywhere

AI is the defining technology of our time; Adoption is set to be widespread

\$7 trillion to global GDP
The total benefits that generative AI alone could add over the next decade, on top of increasing productivity growth by 1.5 percentage points





Three-quarters
of global consumers use gen AI tools today



9 in 10
executives say accelerating AI adoption is a top priority in 2025

Demand for AI compute is expected to accelerate

125x
Projected growth in compute demand for Gen AI applications between 2024 and 2030



18.7 QFLOPs
Estimated global shortfall in AI inferencing capacity by 2030



Equivalent to adding 130% of all planned data center capacity or 1,000 hyperscale facilities


Relying only on data centers to host AI models is unsustainable

- Requires additional investments worth **\$2.8 trillion**
- Energy and water requirements **exceed local grid capacities**


A complementary path is needed: on-device AI

- On-device AI reduces power use by up to **90%**
- And improves water efficiency by **96%**


On-device AI also facilitates broader benefits




It enables **real-time processing** for latency-sensitive tasks



On-device AI can **lower** enterprise deployment **costs**



It **protects user privacy** by keeping sensitive data local



It **spreads AI access** to regions underserved by cloud infrastructure

The future of AI architecture is hybrid

The AI ecosystem today... ... could be complemented with on-device AI



Executive Summary

Artificial intelligence (AI) is the defining general-purpose technology of our time. Generative AI alone could add \$7 trillion to global GDP and increase productivity growth by 1.5 percentage points over the next decade.¹

Adoption is widespread: three-quarters of global consumers already use generative AI tools,² while nearly nine in ten executives say accelerating AI adoption is a top business priority in 2025.³

This opportunity presents significant potential, along with important considerations. Compute demand for generative AI applications is projected to grow **125-fold** between 2024 and 2030. Access Partnership estimates a global shortfall of **18.7 QFLOPs** (quettaFLOPs) of inferencing capacity by 2030⁴ — equal to 130% of all data center capacity currently planned, or the equivalent of approximately 1,000 additional hyperscale facilities. In some regions, particularly Sub-Saharan Africa and Latin America, up to 98% of compute demand for inference could go unmet.

Scaling the exclusively data center-based model is unsustainable. Additional investments of **\$2.8 trillion** would be required, while energy and water requirements are set to exceed local grid and utility capacities in many markets.

A complementary path is needed: **on-device AI**. By distributing inference workloads across billions of devices — smartphones, PCs, IoT systems, and robotics — AI compute can scale more efficiently, securely, and sustainably. On-device AI reduces **power use by up to 90%**, **improves water efficiency by 96%**, and cuts network bandwidth needs. In practice, distributing inference workloads across devices can save as much electricity as South Korea generates annually (507 TWh) and enough water annually to fill 126,000 Olympic swimming pools.

The benefits go well beyond efficiency. On-device AI enables real-time processing for latency-sensitive tasks, protects user privacy by keeping sensitive data local, lowers enterprise deployment costs, and spreads AI access to regions underserved by cloud infrastructure.



The future of AI is hybrid: data centers for training, complemented by distributed inference on billions of edge devices. This balance will allow AI to scale sustainably, equitably, and securely — making intelligence truly on-demand, on-device, everywhere.

“

It's an easy prediction of where things are headed. Devices will just be edge nodes for AI inference, as bandwidth limitations prevent everything being done server-side.

Elon Musk, CEO, Tesla, SpaceX and xAI

1. AI's next gold rush — the demand for AI compute is expected to be massive

AI has emerged as a pivotal general-purpose technology. Generative AI alone is expected to facilitate a 7% increase in global GDP, equivalent to around \$7 trillion, while lifting productivity growth by 1.5 percentage points over the next decade.⁵

Consumers and businesses are well aware of the benefits of AI, and most are already using it. Since becoming mainstream in late 2023, generative AI has been assisting users across a wide range of applications, including speeding up writing and research, instantly generating images and videos for creative work, offering personalized scheduling and writing assistance on smartphones as well as producing new permutations of chemical compounds for pharmaceuticals. At least three-quarters of consumers across the globe are already using generative AI tools, primarily for translation or text and image generation.⁶

Meanwhile, 88% of C-suite executives surveyed globally say that helping their businesses speed up AI adoption is a top priority in 2025.⁷ Industrial AI applications promise to transform sectors such as manufacturing, where leading factories are already using AI-enabled analytics to deploy digital twins and forecast performance metrics to improve productivity. In the finance sector, AI algorithms are used to analyze large amounts of data to assess creditworthiness.⁸ These examples illustrate the technology's broad impact, precipitating a rapid acceleration of AI-related investment.⁹



88% of C-suite executives say speeding up AI adoption in their business is a top priority in 2025

Supporting this wave of technology-driven innovation is the AI ecosystem, built on layers of infrastructure and technology that enable its development, deployment, and use. Each layer plays a vital role in the broader AI economy, driving innovation and



advancing the capabilities of AI technology. It encompasses a wide range of stakeholders, including those involved in manufacturing and designing chips, data centers, model developers, application developers, and end-users, all of whom rely on essential infrastructure such as electricity grids and internet connectivity.

It is projected that the total compute required to run generative AI applications alone could grow **125 times** between 2024 and 2030, mainly due to the rising number of use cases and users.¹⁰ How best to meet this projected surge in AI demand and assign resources is still up for debate among policymakers and businesses.¹¹



The amount of computation necessary to do [newer] reasoning process(es) is 100 times more than what we used to do.

Jensen Huang, CEO, NVIDIA, on the possibilities of next-generation AI

2. The data center crunch — the data center-based AI ecosystem can't meet this demand alone

The characteristics of the AI ecosystem discussed today largely reflect a specific type of AI architecture — one that mainly relies on data center infrastructure for operations and deployment. Scaling this ecosystem to meet current AI demand will be a significant undertaking.

This ecosystem centralizes large volumes of data and AI models for training and inference in public data centers or private on-premises/co-located servers, while end-users interact with the models remotely. This is largely driven by the dominance of AI models, particularly large-language models (LLMs), which require significant volumes of data and energy to train and refine. This leads to them being hosted centrally on a remote server for scalability and cost-effectiveness.

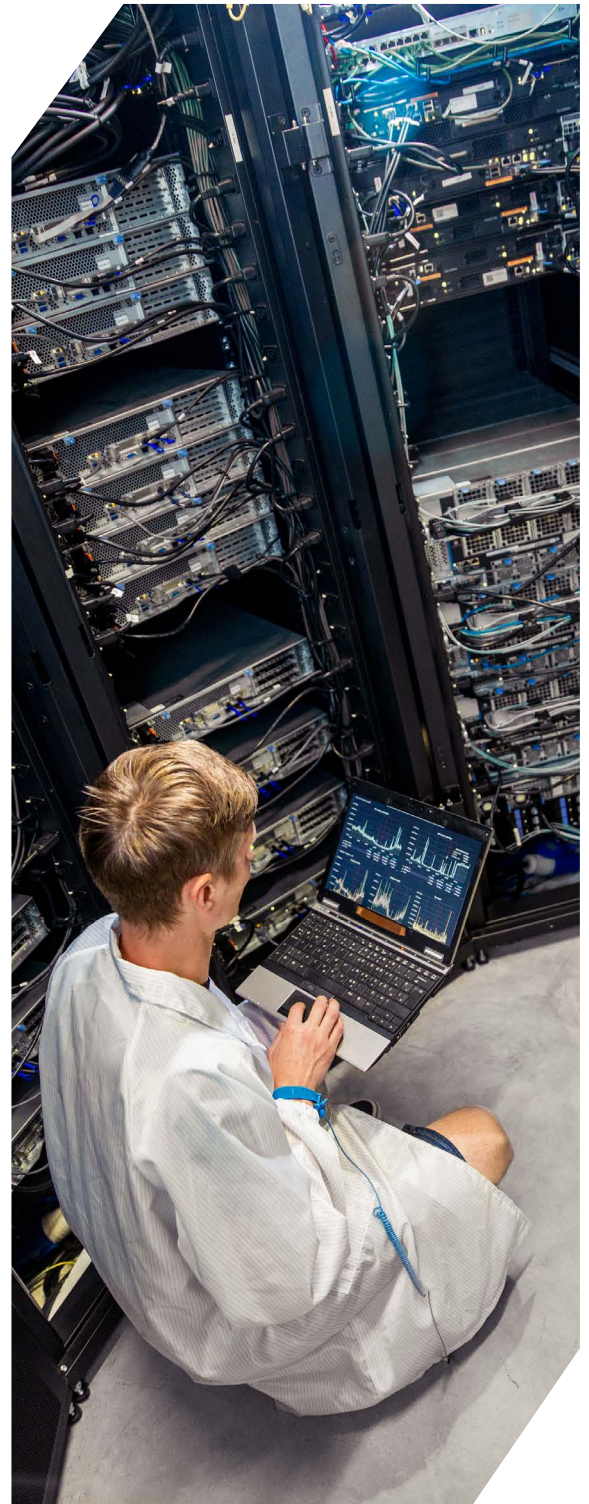
Meeting the AI demand through this ecosystem will be a challenge. Access Partnership estimates the total supply of AI compute resources for inferencing based on public commitments made by data center operators up to 2030 (see *Exhibit 1*).

The analysis highlights a potential shortfall of approximately **18.7 QFLOPs** of AI inference compute capacity¹² by 2030, which is around 130% more than the projected inference compute that committed data center investments could provide.

How do we measure AI compute?

The wide variation in task duration and complexity across AI use cases means there is a wide variety of methods for assessing compute needs. FLOP (floating-point operations per second), among others, remains one of the most accessible metric, serving as a proxy for the volume of calculations — and thus resources — AI models require. For the purposes of this study, both the demand for AI compute and the supply of data center capacity were converted into QFLOPs (quettaFLOPs).

How large is a QFLOP? One QFLOP = 10^{30} FLOPs. This is equivalent to processing over 800 million copies of English Wikipedia or 5 trillion copies of the King James Bible.



It is crucial to note that AI compute requirements differ based on the task at hand: either for AI **training** or **inference**. **Training** refers to the process of building, running, and refining AI models using significantly large datasets over a length of time, often done iteratively. These phases typically require large amounts of compute resources for a certain length of time. Inference refers to the task an AI model carries out to process an input to generate outputs, from answering queries to creating images or videos. Inference tasks typically require a marginal volume of compute resources compared to training. Still, when done for a significantly large consumer base regularly, the total compute resources dedicated to inferencing begin to dwarf the amount used in training.¹³ As such, this analysis focuses on compute workloads related to **inference**, given how AI inferencing will require significant compute capacity. Even if all the projected training workloads in data centers were shifted to cater to inference, total data center capacity will still leave around **12 QFLOPs** worth of compute demand unfulfilled.

How many data centers would it take to plug the 18.7 QFLOPs gap?

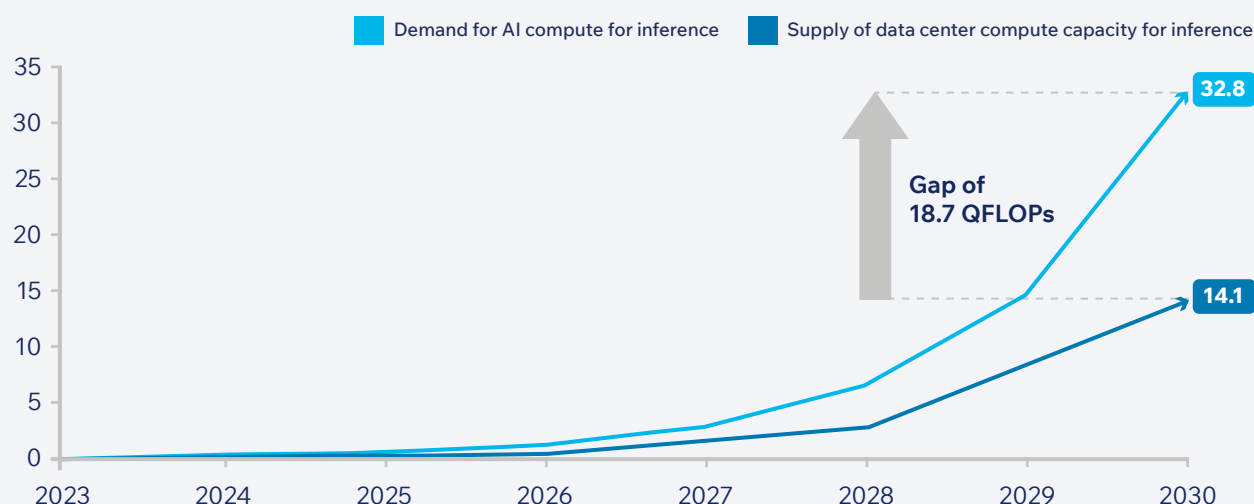
Standard data center capacity is conventionally measured in terms of wattage (or gigawatts). As such, determining the number of data centers needed to run a specific volume of compute also requires projections on a data center's power usage effectiveness (or PUE).

Assuming PUE does meaningfully improve by 2030, it could mean installing at least another **1,000 hyperscale data centers** across the globe.

Exhibit 1

The gap between the supply of data center compute capacity for inferencing and projected demand will widen from 2023 to 2030

Projected annual demand vs supply of AI computing capacity (inference only)



Source: Access Partnership analysis
Note: QFLOPs (quettaFLOPs) = 10^{30} FLOPs

The projected growth in demand for compute is expected to outpace installed data center capacity in every region by 2030

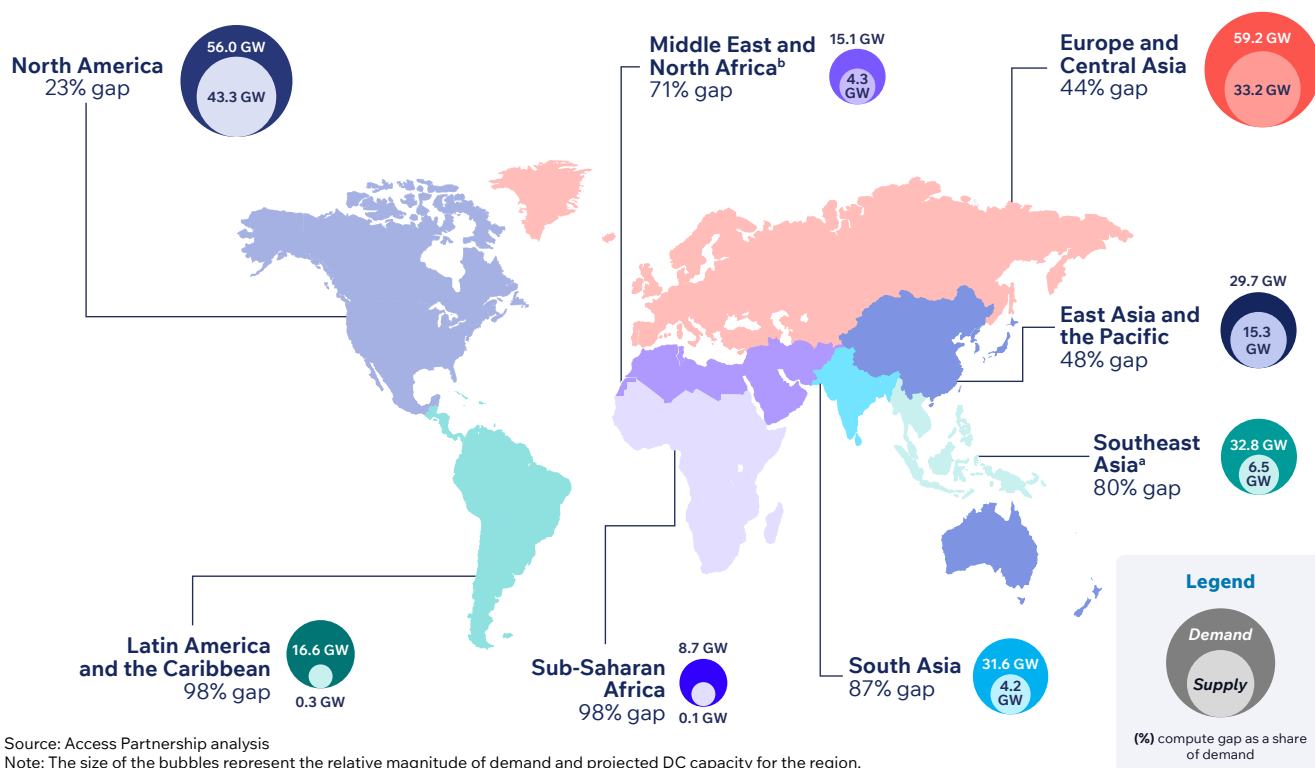
As consumers become more open to using AI and businesses pour more investments into AI tools, demand for AI compute is expected to dramatically accelerate across the globe and filter down at the regional level as well. These regions reflect clusters of data center infrastructure that serve groups of customers within regional boundaries, owing to how they traditionally concentrate within specific areas due to the technical, power and data storage requirements they have.¹⁴

When accounting for wide regional variations in (a) projected data volume growth, (b) growth in AI investments and (c) each region's state of AI infrastructure, the level of compute resources that major data center regions can provide is expected to be significantly outpaced by rapid demand growth in every region — as illustrated in *Exhibit 2*. Gaps are most prominent in regions with limited data center capacity today (and estimated capacity tomorrow) such as Sub-Saharan Africa and Latin America and the Caribbean, where the compute gap could be at least 98% of projected total demand. Even developed regions such as North America, Europe, and East Asia and the Pacific will experience substantial compute gaps of between 23% and 44% of projected demand.

Exhibit 2

The volume of compute resources needed to meet demand for AI inferencing is expected to outstrip supply in every region

Equivalent demand and projected data center capacity for AI inferencing in 2030 by region, gigawatts (GW)



Limitations in infrastructure and cost prevent data centers from becoming a scalable compute solution

Building up data centers to host this 18.7 QFLOPs gap would face structural limits.

- **Investment gap:** Meeting demand purely through data centers would require at least \$2.8 trillion in additional infrastructure investment — on top of \$3–8 trillion already committed.¹⁵ This is equivalent to over **five Stargate-level investment commitments**¹⁶ over the next five years. Firms will find it difficult to justify this upfront investment, as the cost of acquiring further capital for infrastructure buildout places enormous debt burdens on data center operators, making it difficult for them to raise funds without novel financing structures.

596 TWh worth of electricity would be required to address the compute gap in data centers — **as much as Brazil's total power generation in 2022**

- **Power bottlenecks:** Data centers already draw up to 10% of a city's electricity.¹⁷ In the US, this number could go up to 12% by 2028, triple of its usage in 2023.¹⁸ Meeting the 18.7 QFLOPs gap would require **596 TWh** of additional electricity — equivalent to the total amount of electricity generated by Brazil in 2022 or the output of approximately 180 Hoover dams. Already in the Asia Pacific, demand from data centers is expected to exceed available power supply well into 2028, with the energy gap projected to be 25% to 42% of total forecasted demand.¹⁹
- **Water stress:** Cooling hyperscale data centers consumes vast quantities of water, raising sustainability concerns in water-scarce regions.
- **Regional inequities:** Compute deserts persist. By 2030, Sub-Saharan Africa and Latin America could face compute gaps of 98% relative to demand, locking out entire regions.



It is worth noting that these figures reflect the constraints in building data centers to address compute demand related to **inference**. It is assumed that **training-related workloads** would remain hosted by data centers for the foreseeable future, given their capacity to provide large volumes of computational resources at scale and over long durations. Beyond centralizing AI compute resources into data centers, stakeholders across the AI ecosystem must consider envisioning a new future — one where AI compute resources are distributed across devices and infrastructure.

“

[Agentic AI] eat tokens for breakfast, lunch and dinner... puts more pressure on tokens... They're looking at a... re-architecture because the demand for tokens is putting pressure on more large-scale systems

Matt Garman, CEO, AWS on new demands reshaping enterprise infrastructure

3. The breakthrough — the on-device AI future

A future where a robust on-device AI ecosystem complements data centers will lead to compute resources being well distributed across the ecosystem — and save precious resources.

While data centers will continue to host AI training workloads, on-device AI is quickly emerging as a critical complement — helping to shift part of the inference workload off the cloud. This involves deploying (typically lightweight²⁰) AI models onto a device — including smartphones, personal computers, smart portable devices, and drones — and running these algorithms to generate outputs and outcomes all within that device rather than in centralized host servers. Several emerging technological and user trends are creating an environment for on-device AI to play a larger role, owing to its inherent characteristics, which include:

- **The growing shift in data gravity toward end-users is making on-device AI increasingly effective for processing data locally.** Given the sheer volume of data produced by digital applications, it is becoming increasingly efficient to process data on-device rather than in the cloud. Internet of Things (IoT) devices are expected to generate nearly 80 zettabytes of data in 2025 alone, equivalent to storing 250 billion DVDs.²¹ This shift is particularly pronounced in applications that require real-time video processing, such as autonomous vehicles or smart home systems, where latency-sensitive operations would benefit from immediate, local computation offered by on-device AI.
- **A rise in hyper-personalization, which requires sensitive data that users may prefer keeping on-device.** Consumers are demanding more personalization in the products and services they engage with, with 71% of consumers in the US stating so.²² Such services would require AI applications to have access to highly sensitive and personal data. The tension lies in ensuring

users also have control over their data and minimizing risks of data breaches. Over half of smartphone consumers are concerned about their data being accessed by AI²³ and 40% of businesses are exploring ways to mitigate cybersecurity risks of using generative AI.²⁴ Localizing AI processing to be on-device could effectively circumvent these concerns.

- **Challenges in accessing data center-based AI compute, with on-device AI enabling inclusive access to AI compute.** The economics of using AI are reaching a critical juncture, with costs estimated to increase by 89% between 2023 and 2025.²⁵ In the US, 59% of IT decision makers report bandwidth shortages when using large AI applications within their organizations, creating a bottleneck when leveraging data center-based AI at the industry level.²⁶ Additionally, the heavy concentration of data centers in a few geographical clusters has inadvertently created “compute deserts”, where there is no public cloud AI infrastructure at all. To make AI widely accessible, a robust on-device AI ecosystem is needed — one built around lightweight models that work with minimal connectivity.

“

By 2030, we aim to make Samsung an AI-powered firm. We aim to use AI in over 90% of its workflows, with the technology intended to support, not replace, its human workforce. We expect to embed AI features in more than 400 million Galaxy devices by the end of this year.

Roh Tae-moon, Acting Head of Samsung's Device Experience Division

3. The breakthrough

As on-device AI processing requires low latency and is available in a broader range of devices, on-device AI would enable computational workloads to be easily distributed across the ecosystem regardless of connectivity conditions and locations.

Shifting the workload gap to on-device AI could also bring about significant energy and resource savings. Running AI on devices instead of centralizing compute workloads on data centers is estimated to reduce total **power consumption by around 90% and increase water efficiency by 96%.**²⁷ Estimates suggest that running AI locally can reduce energy consumption by at least 100 times per task.²⁸ Preliminary experiments by researchers also estimate that shifting inferences from remote servers to edge AI applications can reduce the power consumption of IoT devices by 21%.²⁹

Running AI on devices could lead to 315 billion liters of water being saved, **equivalent to 126,000 Olympic-sized swimming pools**

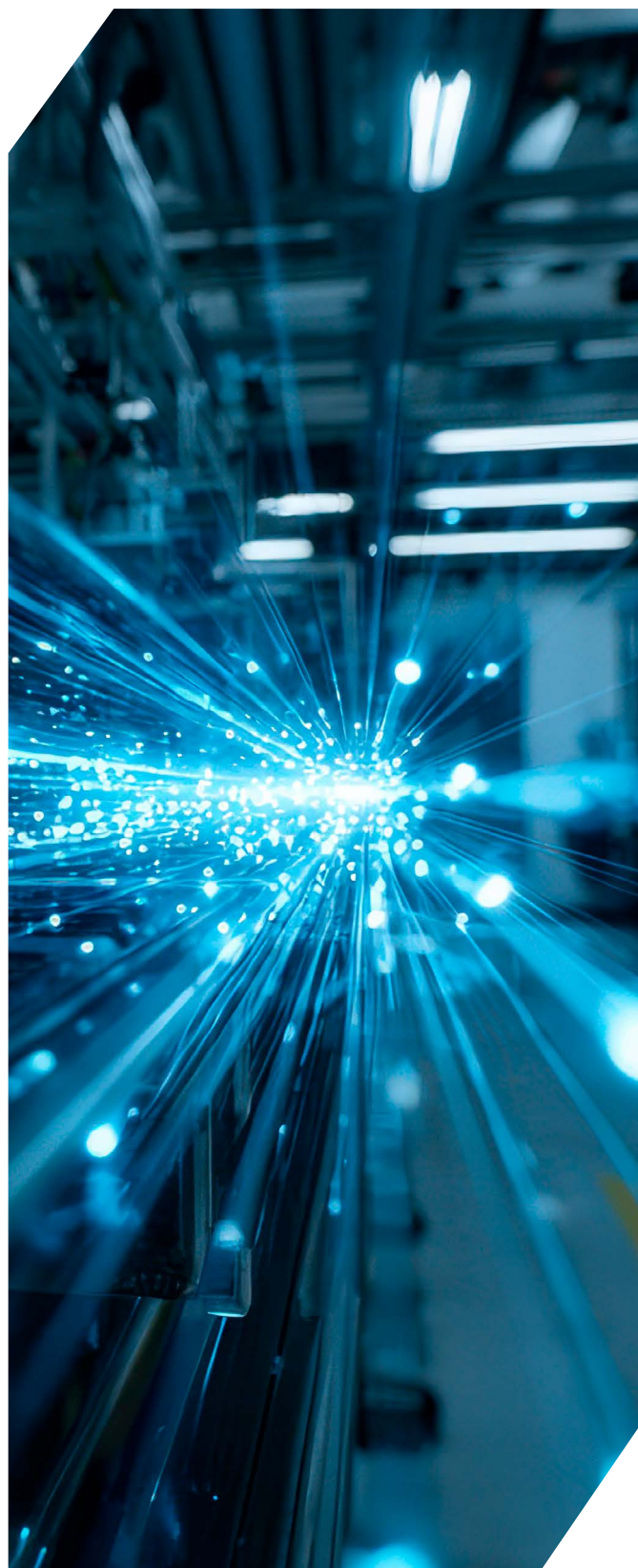
Compared with relying solely on data center AI to address the compute gap, incorporating on-device AI is expected to:



Reduce grid load by 507 TWh, **equivalent to the total amount of electricity generated by South Korea in 2022.**



Lead to around 315 billion liters of water being saved, equivalent to redistributing **126,000 Olympic-sized swimming pools** back to consumers.



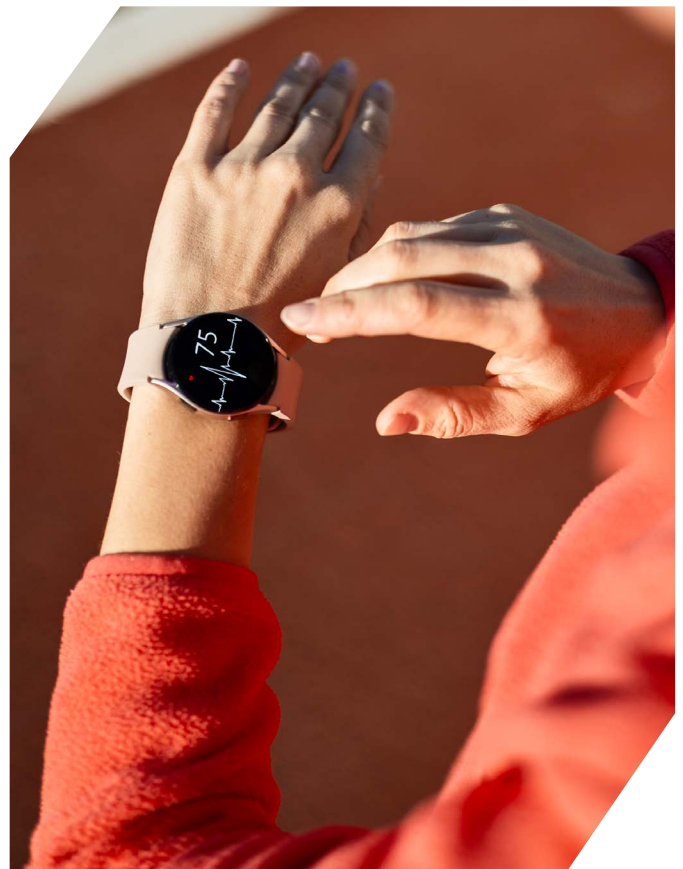
4. Beyond efficiency — additional benefits of on-device AI

The emerging trends and unique characteristics of on-device AI open up a wide range of use-cases — particularly ones that are constrained by the current data center infrastructure. On-device AI applications are expected to facilitate benefits such as:

- **Enabling greater customization and functionality of consumer products.** Smartphones, PCs, consumer wearables, IoT devices, and camera-based hardware would be supercharged as on-device AI enables devices to process user data in real-time without relying on internet connectivity. Consumers would be able to use AI for real-time translation, photo editing, text generation, or even to solve specific problems instantly without the worry of data transmission vulnerabilities or network downtimes.
- **Improving the efficiency of industrial products.** Manufacturing plants that integrate robotics and sensor-based hardware would become more efficient as on-device AI processors circumvent network outages and reduce the latency of large data transmissions. Aramco leverages on-device AI processing to assist workers in forecasting and preventing essential equipment downtime, revolutionizing maintenance practices.³⁰
- **On-device AI facilitates more effective and widespread deployment of robotics.** Equipping robots with on-device AI capabilities reduces reliance on constant cloud connectivity and allows for faster, low-latency decision-making. This capability has measurable productivity benefits: research shows that every 1% increase in industrial robot density can improve overall productivity by 0.8%.³¹ In addition, on-device AI enhances operational safety and security by allowing robots to respond in real time to hazards, monitor sensitive environments, and execute critical processes autonomously, reducing human exposure to risk and improving process reliability.

- **Transforming industries, including by creating safer and more secure processes for the financial services sector.** For sensitive sectors like financial services, on-device AI can process sensitive data and tasks locally, ensuring close alignment with data privacy laws. This allows them to verify customer identities as part of Know Your Customer (KYC) requirements or even detect fraudulent transactions and unauthorized log-in attempts at greater speed and with greater security.

Beyond offloading power and resource use compared to data centers, on-device AI also offers other benefits to the broader ecosystem, industry, and consumers — as summarized in the following pages.





Ecosystem-wide benefits:

- **Promotes data privacy and protection by bringing data closer to the end-user.** By removing the need for transmitting user data to central servers for processing, on-device AI promotes data protection and enables companies to comply with regulations easily, and provides users with the confidence that their data will not be tampered with or stolen.
- **Reduces bandwidth usage.** Processing data locally minimizes the need for large data transmissions across networks, reducing bandwidth usage and the load on telecommunication networks.
- **Promotes economic growth and supports jobs.** By enabling more use cases, on-device AI is expected to create productivity gains that lead to economic growth. The proliferation of more AI-capable devices will also help to drive sustainable growth and innovation by unlocking efficiencies, increasing productivity, and enabling new business models.³² Additionally, promoting the on-device model allows more companies to participate in AI development and deployment, while also creating new jobs focused on how AI interacts with end-users at the device level.



Industry-level benefits:

- **Optimizes usage of compute, memory, and storage for AI model training.** Shifting growing model inference workloads to on-device AI is expected to free up bandwidth and capacity for cloud data centers to focus on model training, especially for advanced AI models that drive cutting-edge breakthroughs.³³
- **Reduces data transfers and lowers latency, opening possibilities for real-time processing.** Some AI application providers have observed that edge-based video analytics result in lower latency — with kinara.ai, an edge AI solutions developer, estimating a latency of 5 to 50 milliseconds when using edge AI, compared to 50 to 250 milliseconds using cloud-based analytics.³⁴
- **Mitigates the impact of network vulnerabilities and downtimes.** By enabling AI applications to process data locally, the effects of abrupt downtimes or data breaches would be minimized. These downtimes can be costly — a report by IBM indicates that 33% of surveyed enterprises estimated that an hour of downtime can cost between \$1 million and \$5 million.³⁵ Data breaches are also an expensive affair, costing firms an average of \$4.9 million in 2024.³⁶

- **Reduces costs in AI deployment.** For AI application developers, on-device AI also eliminates costs associated with sending and processing tokens to cloud servers. Typical costs for sending long-form or image-based prompts range from \$0.02 to \$0.10 per prompt, excluding the cost of integrating applications with existing remote models via APIs.³⁷ On-device AI also reduces dependencies on data centers. This includes the expense of hosting data in a centralized server, maintaining data center networks, and moving data out of cloud servers during transitions.
- **Increases productivity for businesses.** On-device AI use cases also have the potential to improve productivity. Analysis by Accenture estimates that companies adopting a more integrated approach to using edge AI are four times more likely to achieve accelerated innovation, nine times more likely to increase efficiency, and seven times more likely to reduce costs.³⁸ Specific use cases also display positive benefits. For example, on-device AI enables more extensive deployment of robotics through swarm-based coordination algorithms, which can support robots to act independently while facilitating collective behavior. Creating more impactful use cases can encourage broader economic gains, with a 1% rise in industrial robot density improving business productivity by 0.8%.³⁹



Consumers and end-users:

- **Aligns with end-user preferences.** On-device AI appears to fulfil a broader range of consumer preferences, including the sustainability, safety, and effectiveness of AI products. For example, younger and digitally savvy consumers are becoming more particular about the impact of companies, with 27% of them more likely to purchase products when they believe a brand cares about its impact on people and the planet.⁴⁰ AI products that are perceived as more carbon-intensive due to their reliance on data center-based AI may therefore become perceived as harmful and less desirable.
- **Facilitates hyper-personalization.** Through direct user interaction and observation, on-device AI can better understand the context of a user's actions, allowing for more accurate and relevant personalization. These features have already become prevalent, with instances of generative AI applications adjusting their text outputs to match the way users type and the tone they use.

5. The path forward

The future of AI will be hybrid. Data centers will continue to play a central role in training and hosting the largest and most complex models. However, inference — the day-to-day workload of billions of users and enterprises — will increasingly shift to devices and edge infrastructure. The resulting hybrid AI ecosystem of the future is illustrated in *Exhibit 3*.

How policymakers can promote a hybrid AI architecture

Policymakers should recognize a hybrid AI ecosystem, a model where cloud data centers handle training and the heaviest inference workloads, while edge devices (phones, PCs, vehicles, and IoT) manage lighter inference tasks. This approach reduces costs, latency, water and energy use, and dependence on hyperscale data center buildouts. Hybrid AI should be positioned as the **default design pattern** for AI infrastructure planning. Governments can integrate this into national digital strategies, infrastructure roadmaps, and procurement guidelines, signaling clearly that scaling AI sustainably requires distributing workloads intelligently across the ecosystem.

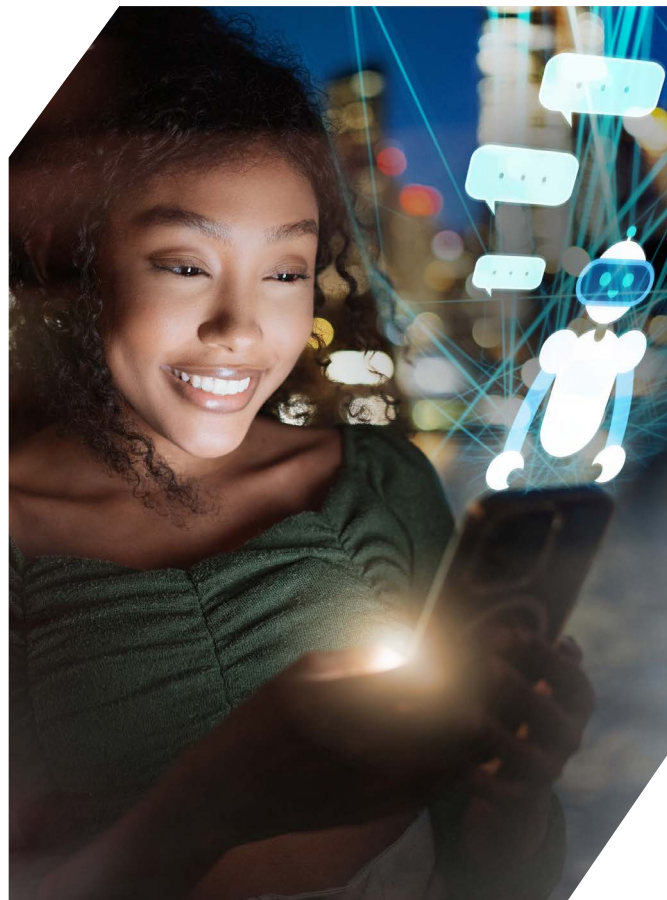
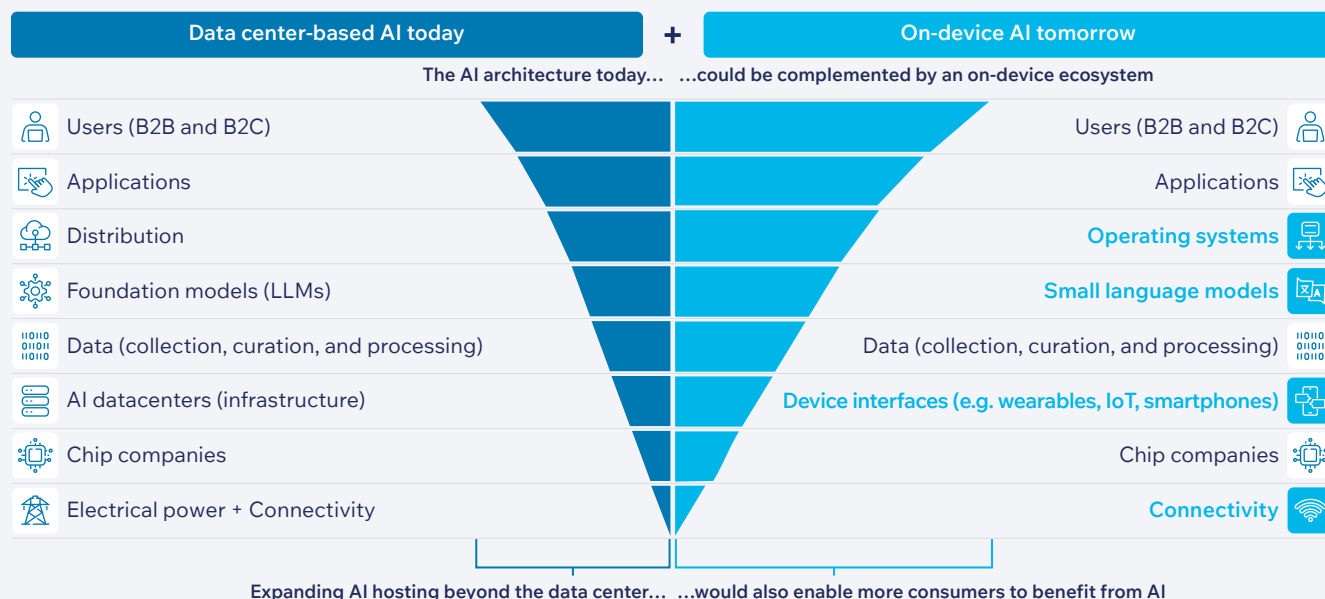


Exhibit 3

The AI ecosystem could be expanded from enabling on-device AI



In a hybrid model, connectivity becomes the bridge between large-scale AI models hosted in data centers and the diverse ecosystem of edge devices that now possess AI processing capability. Smartphones, PCs, tablets, wearables, and a growing class of connected systems, from vehicles and drones to industrial robots and IoT sensors, will become active participants in AI computation (see *Exhibit 4*)

1. Support research and innovation in hybrid workload optimization

To operationalize hybrid AI, countries need research programs that model how to best distribute workloads between the cloud and devices. These programs should be designed to create open knowledge for governments, academia, and industry.

- **Fund foundational research on workload allocation frameworks:** Governments can sponsor national labs or consortia to model performance, cost, and energy trade-offs between cloud, edge, and device AI. This helps identify “optimal split” scenarios.
- **Support testbeds and simulation platforms:** Establish large-scale testing environments (public and private) where models can be trialed across hybrid configurations under real-world conditions.
- **Create open modelling benchmarks:** Publish transparent datasets and benchmarking tools that allow comparison of cloud vs. on-device performance. This avoids vendor lock-in and encourages fair competition.
- **Develop skills and expertise pipelines:** Invest in training programs for engineers and researchers specializing in edge AI optimization, ensuring the workforce can implement hybrid architectures at scale.

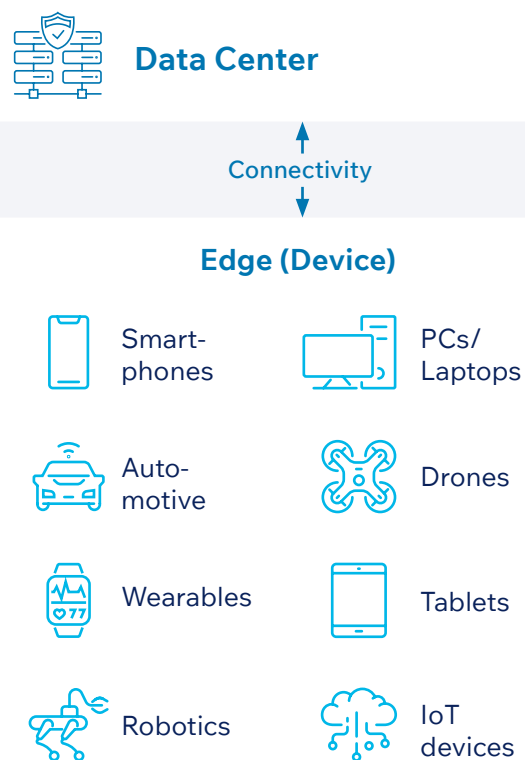
2. Build a robust on-device AI ecosystem

On-device AI requires an enabling ecosystem of innovation, adoption, and scale. Policymakers can accelerate this by creating the right incentives, reducing barriers, and encouraging uptake in strategic sectors.



Exhibit 4

AI will extend to the edge through multiple interfaces



- **Provide targeted incentives for lightweight models and chip R&D:** Offer grants and tax credits for companies developing energy-efficient chips, operating systems, and lightweight models optimized for on-device use.
- **Launch regulatory sandboxes in strategic sectors:** Create controlled environments for piloting on-device AI in healthcare, finance, mobility, and manufacturing, where privacy, safety, and performance requirements are critical.
- **Adopt on-device AI in public services:** Set adoption targets for areas such as smart cities, education, and environmental monitoring, signaling demand and providing reference use cases.
- **Support small, medium-sized enterprise (SME) participation in the AI economy:** Establish funding streams, accelerator programs, or public procurement set-asides for startups working on on-device AI solutions. This ensures innovation does not remain confined to large incumbents.
- **Facilitate industry-led technical standards:** Support collaborative bodies to set benchmarks for model portability, energy efficiency, and device-level performance evaluation.
- **Establish safety and provenance frameworks:** Develop mechanisms for content provenance, safety grading of device-level AI, and protocols for handling sensitive data locally.
- **Create certification schemes for AI-ready devices:** Introduce voluntary certifications or “trust labels” that signal compliance with privacy, sustainability, and safety requirements, enabling consumer confidence.
- **Promote international coordination:** Engage in global standard-setting (OECD, ISO, ITU, etc.) to ensure interoperability across borders and avoid fragmented regulatory regimes.

3. Develop standards, safety, and trust frameworks for on-device AI

For hybrid AI to scale sustainably, governments must ensure that trust and interoperability are built into the system from the start. Policymakers should work with industry to co-develop standards that increase safety,

By 2030, the global AI economy will depend on a rebalanced compute model. Success will hinge not on building ever more data centers, but on distributing intelligence across billions of connected devices as well. This shift will allow AI to scale sustainably, equitably, and securely — positioning on-device AI as the cornerstone of an everywhere, always-on AI future.



6. Appendix

1. Methodology for estimating future AI compute gap

This analysis consists of two components: (a) compute capacity for AI; and (b) compute demanded by AI applications.

• Calculating compute capacity for AI

To describe total AI compute resources available in the world, Access Partnership modelled the total compute capacity available in data centers dedicated to AI (public and on-premise) up to 2030 as a proxy (including compute used for training and inference). This estimate relies on publicly available data. Specific assumptions to estimate AI capacity include:

- The total data center capacity can be represented by total power capacity in gigawatts (GW)
- Of total cloud workloads, the share dedicated to AI would grow from 14% in 2025 to 27% in 2027, based on external estimates from firms like Goldman Sachs and McKinsey & Company
- The ratio of total AI compute to GW (i.e., a compute-per-GW ratio) changes over time to reflect increasing GPU efficiency
- Around 40% of workloads are dedicated to training-related compute, based on analysis from the Center for the Governance of AI

To estimate total AI capacity, component (a) was multiplied by (b) and (c) to reflect total compute capacity toward AI and further disaggregated with (d) to reflect compute capacity for inference.

• Calculating compute demanded

To measure the potential compute capacity that on-device AI can serve, Access Partnership also estimated the overall AI compute capacity demanded. The “gap” between demand and current AI capacity represents the total compute that



could be offloaded from data centers. Specific assumptions to estimate demand for AI compute include:

- Annual demand from the use of generative AI is estimated to grow from 0.2 QFLOPs in 2024 to 25 QFLOPs in 2030, as outlined in Chapter 1
- By 2030, 40% of total AI compute demanded by businesses and consumers will be toward generative AI applications
- Of total AI demand by businesses and consumers, demand for training applications could range between 10% and 25% of total compute

To estimate total demand, component (a) was extrapolated using (b) to reflect total compute demand for AI applications in general, and further disaggregated using (c) to reflect compute demand for AI inference applications.

2. Methodology for the regional AI compute gap

This analysis consists of two components: (a) regional compute capacity for AI; and (b) regional compute demanded by AI applications.

- **Calculating regional compute capacity for AI**

To describe AI compute resources available by region, Access Partnership disaggregated existing modelled figures on global compute capacity for AI into region-specific figures. This illustrates the relative levels of installed DC capacity for each region up to 2030. This estimate relies on publicly available data. Specific assumptions to estimate regional AI capacity include:

- a. Count of the number of data centers (split by region) from the Data Center Map
- b. Data center energy usage (by region) from the International Energy Agency (IEA)
- c. Growth of data center capacity (by region) from KPMG

Components (a) and (b) served as proxies to disaggregate existing modelled figures into region-specific figures, whilst component (c) served to inform region-specific growth rates.

- **Calculating the compute demanded by the region**

Access Partnership also estimated the AI compute capacity required by each region, with the “gap” representing the total compute for each region as a percentage of the region’s demand. Specific assumptions to estimate demand for AI compute include:

- a. The number of average visits to AI tools, applications, and platforms from web traffic analysis companies
- b. The CAGR of AI spending for each region from market research firms and consultancies

Component (a) served as a proxy to disaggregate existing modelled figures into region-specific figures, whilst component (b) was used to contextualize the global growth rate into region-specific growth rates.



7. Endnotes

- 1 \$ refers to US\$ in this instance, and in the rest of the report (unless otherwise stated). Goldman Sachs (2023), “Generative AI could raise global GDP by 7%.” Available at: <https://www.goldmansachs.com/insights/articles/generative-ai-could-raise-global-gdp-by-7-percent>
- 2 SwissRe (2024), Trust in the era of GenAI: the Swiss Re Global AI Perception Survey. Available at: <https://www.swissre.com/dam/jcr:0d2816b7-9300-4151-816a-b575286bf007/global-ai-perception-survey.pdf>
- 3 LinkedIn and Access Partnership (2025), Work Change Report: AI is Coming to Work. Available at: <https://economicgraph.linkedin.com/content/dam/me/economicgraph/en-us/PDF/Work-Change-Report.pdf>
- 4 AI compute needs vary between training and inference. Training involves building and refining models on large datasets over time and requires intensive compute resources. Inference is the application of trained models to generate outputs (e.g. answers, images, video), and typically requires far less compute per task.
- 5 Goldman Sachs (2023), “Generative AI could raise global GDP by 7%.” Available at: <https://www.goldmansachs.com/insights/articles/generative-ai-could-raise-global-gdp-by-7-percent>
- 6 SwissRe (2024), Trust in the era of GenAI: the Swiss Re Global AI Perception Survey. Available at: <https://www.swissre.com/dam/jcr:0d2816b7-9300-4151-816a-b575286bf007/global-ai-perception-survey.pdf>
- 7 LinkedIn and Access Partnership (2025), Work Change Report: AI is Coming to Work. Available at: <https://economicgraph.linkedin.com/content/dam/me/economicgraph/en-us/PDF/Work-Change-Report.pdf>
- 8 EY (2024), “How artificial intelligence is reshaping the financial services industry.” Available at: https://www.ey.com/en_gr/insights/financial-services/how-artificial-intelligence-is-reshaping-the-financial-services-industry
- 9 BCG (2025), “From Potential to Profit: Closing the AI Impact Gap.” Available at: <https://www.bcg.com/publications/2025/closing-the-ai-impact-gap>
- 10 McKinsey & Company (2023), “Generative AI: The next S-curve for the semiconductor industry?” Available at: https://www.mckinsey.com/~media/mckinsey/industries/semiconductors/our_insights/mckinsey_on_semiconductors_2024/mck_semiconductors_2024_webpdf.pdf
- 11 OECD (2023), A blueprint for building national compute capacity for artificial intelligence. Available at: https://www.oecd.org/en/publications/a-blueprint-for-building-national-compute-capacity-for-artificial-intelligence_876367e3-en.html
- 12 Data center capacity includes both public cloud and private on-premise or co-location servers.
- 13 It is estimated that out of the total AI compute workloads happening in data centers today, around 80-90% are for inference-related tasks. MIT Technology Review (2025), “We did the math on AI’s energy footprint. Here’s the story you haven’t heard.” Available at: <https://www.technologyreview.com/2025/05/20/1116327/ai-energy-usage-climate-footprint-big-tech/>
- 14 Sources include: International Energy Agency (IEA) (2025), Energy and AI. Available at: <https://www.iea.org/reports/energy-and-ai> and Lehdonvirta, V.; Wu, B.; Hawkins, Z. (2024), Compute North vs. Compute South: The Uneven Possibilities of Compute-based AI Governance Around the Globe. Available at: <https://ojs.aaai.org/index.php/AIES/article/view/31683/33850>
- 15 McKinsey & Company (2025), “The cost of compute: A \$7 trillion race to scale data centers.” Available at: <https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/the-cost-of-compute-a-7-trillion-dollar-race-to-scale-data-centers>
- 16 OpenAI (2025), “Announcing The Stargate Project” Available at: <https://openai.com/index/announcing-the-stargate-project/>
- 17 International Energy Agency (IEA) (2025), Energy and AI. Available at: <https://www.iea.org/reports/energy-and-ai>
- 18 Broadband Breakfast (2025), Dateline Ashburn: Data Centers Drive New Energy Disputes in Northern Virginia. Available at: <https://broadbandbreakfast.com/dateline-ashburn-data-centers-drive-new-energy-disputes-in-northern-virginia/>
- 19 NetworkWorld (2025), “AI boom exposes infrastructure gaps: APAC’s data center demand to outstrip supply by 42%.” Available at: <https://www.networkworld.com/article/4000184/ai-boom-exposes-infrastructure-gaps-apacs-data-center-demand-to-outstrip-supply-by-42.html>
- 20 Lightweight models are smaller AI systems that are optimized to run efficiently on consumer devices without needing data center-level hardware.
- 21 Sources include: IDC (2021), “Future of Industry Ecosystems: Shared Data and Insights.” Available at: <https://blogs.idc.com/2021/01/06/future-of-industry-ecosystems-shared-data-and-insights/> and Rivery (2025), “Big data statistics: How much data is there in the world?” Available at: <https://rivery.io/blog/big-data-statistics-how-much-data-is-there-in-the-world/>
- 22 McKinsey & Company (2021), “The value of getting personalization right—or wrong—is multiplying.” Available at: <https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/the-value-of-getting-personalization-right-or-wrong-is-multiplying>

7. Endnotes

- 23 Canalys (2024), Now and Next for AI-Capable Smartphones 2024. Available at: <https://www.canalys.com/insights/consumer-ai-inclination-index>
- 24 McKinsey & Company (2025), The state of AI: How organizations are rewiring to capture value. Available at: <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai>
- 25 IBM (2024), Use Gen AI Economics to lap the Competition. Available at: <https://www.ibm.com/thought-leadership/institute-business-value/report/ceo-generative-ai/ceo-ai-cost-of-compute>
- 26 Flexential (2025), State of AI Infrastructure report. Available at: <https://www.flexential.com/system/files/file/2025-05/flexential-2025-state-of-ai-infrastructure-report-hvc.pdf>
- 27 Axios (2025), "Moving AI compute to phones massively reduces power use, study finds." Available at: <https://www.axios.com/2025/06/26/ai-compute-phones-qualcomm>
- 28 World Economic Forum (2025), "How on-device AI can help us cut AI's energy demand." Available at: <https://www.weforum.org/stories/2025/03/on-device-ai-energy-system-chatgpt-grok-deepx/>
- 29 Internet of Things (2023), Power consumption reduction for IoT devices thanks to Edge-AI: Application to human activity recognition. Available at: <https://www.sciencedirect.com/science/article/pii/S2542660523002536>
- 30 Qualcomm (2024), "Qualcomm and Aramco Lead Industrial Innovation with Transformative Generative AI IoT Solutions at the Edge." Available at: <https://www.qualcomm.com/news/releases/2024/09/qualcomm-and-aramco-lead-industrial-innovation-with-transformati>
- 31 SelectUSA (2024), Robots and the Economy: The Role of Automation in Driving Productivity Growth. Available at: <https://www.trade.gov/sites/default/files/2022-08/SelectUSAAutomationReport2020.pdf>
- 32 World Economic Forum (2024), "On-device AI would democratise GenAI and ensure inclusive participation in the economy." Available at: <https://www.weforum.org/stories/2024/01/the-case-for-on-device-ai-for-society-and-the-economy/>
- 33 GSMA (2025), "Distributed inference: AI adds a new dimension at the edge." Available at: <https://www.gsma.com/newsroom/article/distributed-inference-ai-adds-a-new-dimension-at-the-edge/>
- 34 Kinara.ai (2023), Optimizing latency for edge AI deployments. Available at: <https://kinara.ai/wp-content/themes/kinara/files/Latency-on-the-Edge-WP-v2.pdf>
- 35 IBM (2024), "How much does an hour of downtime cost the average business?" Available at: <https://www.ibm.com/support/pages/system/files/inline-files/2024-01-Hybrid-Cloud-on-IBM-Power.pdf>
- 36 IBM (2024), Cost of a Data Breach Report 2024. Available at: <https://www.ibm.com/reports/data-breach>
- 37 BytePlus (2025), "How much does one ChatGPT prompt cost? A comprehensive guide." Available at: <https://www.byteplus.com/en/topic/408663?title=how-much-does-one-chatgpt-prompt-cost-a-comprehensive-guide>
- 38 Accenture (2023), Leading with Edge Computing: How to reinvent with data and AI. Available at: <https://www.accenture.com/content/dam/accenture/final/accenture-com/document-2/Accenture-Leading-With-Edge-Computing.pdf#zoom=40>
- 39 SelectUSA (2020), Robots and the economy: The role of automation in driving productivity growth. Available at: <https://www.trade.gov/sites/default/files/2022-08/SelectUSAAutomationReport2020.pdf>
- 40 Harvard Business Review (2023), "Research: Consumers' Sustainability Demands Are Rising." Available at: <https://hbr.org/2023/09/research-consumers-sustainability-demands-are-rising>

Follow us



Our offices

Europe

London

The Tower, Buckingham Green
Buckingham Gate
London, SW1E 6AS
United Kingdom

+44 20 3143 4900
london@accesspartnership.com

Brussels

8th Floor, Silversquare Europe
Square de Meeûs 35
B-1000 Brussels
Belgium

brussels@accesspartnership.com

North America

Washington DC

1300 Connecticut Avenue NW,
Suite 250
Washington, DC 20036
USA

+1 202 503 1570
washingtondc@accesspartnership.com

Asia

Singapore

Asia Square, Tower 2
#11-20
12 Marina View
Singapore 018961

+65 8323 7855
singapore@accesspartnership.com

Jakarta

Revenue Tower 21st Floor
Unit 104 SCBD Lot 13, Jl. Jend. Sudirman
Kav. 52-53
Provinsi DKI Jakarta, 12190
Jakarta, Indonesia

+62 21 5020 0949

Kuala Lumpur

Common Ground Q Sentral
Level 39, Unit 39-02 (East Wing), 2A,
Jalan Stesen Sentral 2, Kuala Lumpur
Sentral, 50470
Kuala Lumpur, Malaysia

Bangkok

188 Spring Tower
11th Floor, Unit 106, Phayathai Road
Thung Phayathai, Ratchathewi,
10400 Bangkok, Thailand

+ 66 (2)-8216148

Hanoi

19th floor, Tower 1
Capital Place Building
No 29 Lieu Giai Street
Ngoc Khanh Ward, Ba Dinh District
Hanoi, Vietnam

Manila

28F & Penthouse
World Plaza
5th Ave
Bonifacio Global City
Manila, 1634
Philippines

Middle East and Africa

Abu Dhabi

Al Wahda City Tower, 20th Floor
Hazaa Bin Zayed The First Street
PO Box 127432
Abu Dhabi, UAE

abudhabi@accesspartnership.com

Johannesburg

119 Witch-Hazel Avenue
Highveld Technopark
Johannesburg
Gauteng, South Africa