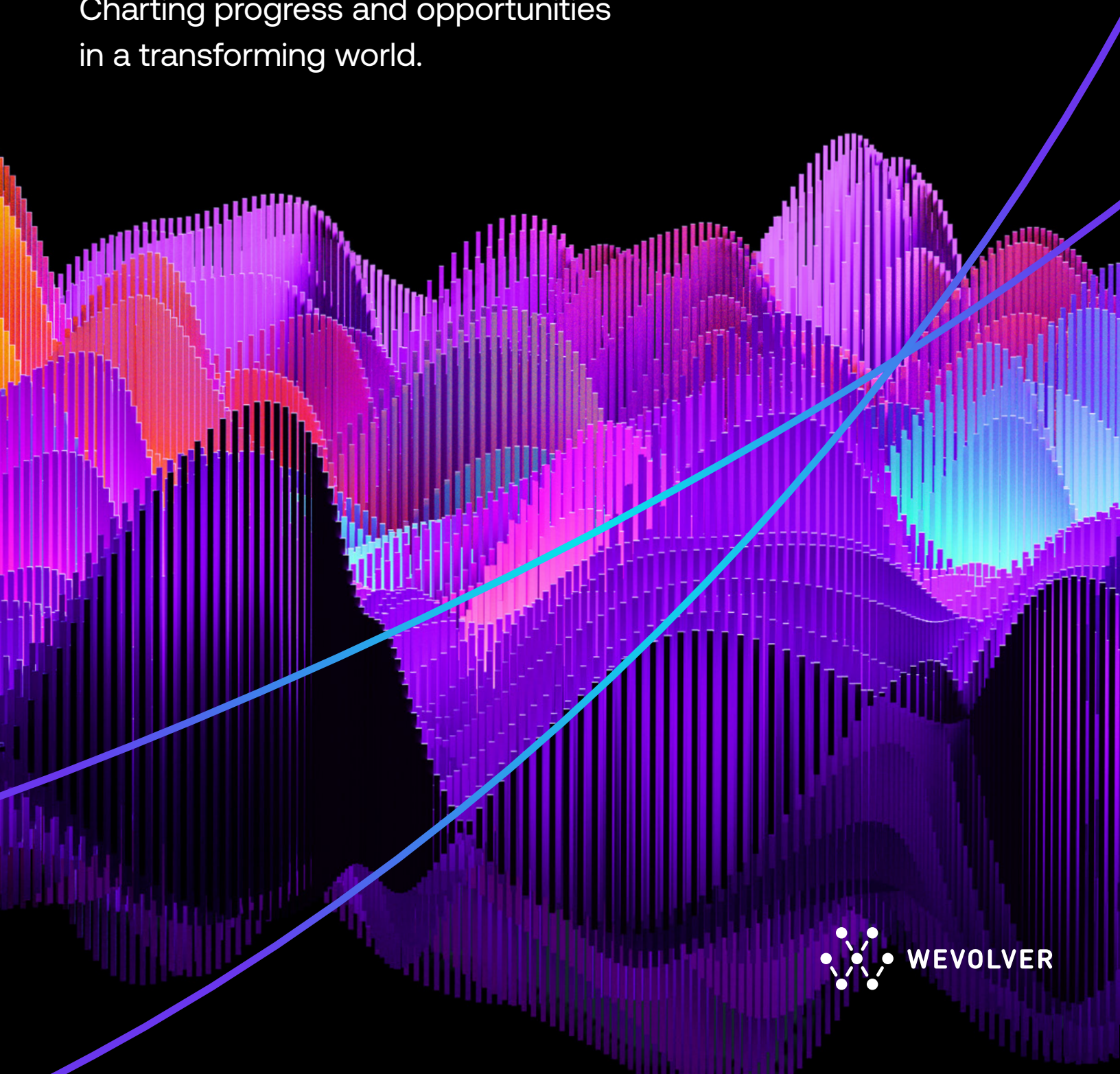
arm

Arm AI Readiness Index

Charting progress and opportunities
in a transforming world.



WEVOLVER

Foreword

WILL ABBEY,
EXECUTIVE VICE PRESIDENT AND CHIEF COMMERCIAL OFFICER, ARM

Throughout history, certain technological advancements have fundamentally altered our world. The printing press democratized knowledge. The industrial revolution transformed manufacturing. The internet connected humanity in ways previously unimaginable. Today, we stand at another such pivotal moment with artificial intelligence.

Like those historic inflection points, artificial intelligence (AI) is not merely an incremental improvement on what came before; it represents a paradigm shift that is redefining how we work, communicate, create, and solve problems. And like those earlier transformations, the full impact of AI will be determined not just by the technology itself, but by how we choose to develop, deploy, and govern it.

What was once considered the realm of science fiction has become an integral part of our daily lives and business operations. The AI Readiness Index report represents a comprehensive effort to understand where we stand today in a rapidly transforming world and chart a course for the future.

As we navigate this transformative era, one thing has become abundantly clear: AI is no longer a technology of tomorrow—it is the driving force of now. Our research reveals that 82% of global business leaders are already using AI applications, with overwhelming majorities expecting to increase their investments in the coming years. Yet despite this widespread adoption, only 39% report having a clearly defined, comprehensive strategy in place.

This gap between adoption and strategic implementation underscores the complex challenges organizations face. From building robust technical

foundations and addressing security concerns to navigating evolving regulatory landscapes and developing necessary talent, the path to AI success requires careful consideration of multiple interconnected factors.

What makes this moment particularly significant is the shifting nature of the conversation itself. We have moved beyond asking whether AI will transform our industries to asking how we can responsibly harness its potential to drive innovation, efficiency, and growth. The focus has rightfully expanded from purely technical considerations to encompass broader implications for security, sustainability, governance, and workforce development.

The AI Readiness Index aims to serve as both a mirror and a compass—reflecting the current state of AI readiness across industries and regions while providing direction for organizations at various stages of their AI journeys. By combining extensive survey data from 665 business leaders with expert analysis and case studies, we offer a nuanced view of the challenges and opportunities that lie ahead.

As you explore the findings in this report, I encourage you to consider not just where your organization stands today, but where it needs to be tomorrow. The organizations that will thrive in this new era will be those that approach AI with both ambition and responsibility—embracing its transformative potential while thoughtfully addressing the technical, ethical, and human dimensions of implementation.

The future of AI is being written today, by leaders willing to ask difficult questions, take bold risks and make strategic investments in the foundations of tomorrow's success. My hope is that this report will serve as a valuable resource as you write your organization's chapter in that future.

CHAPTER 1

The Global State of AI Readiness: A Data-Driven Analysis

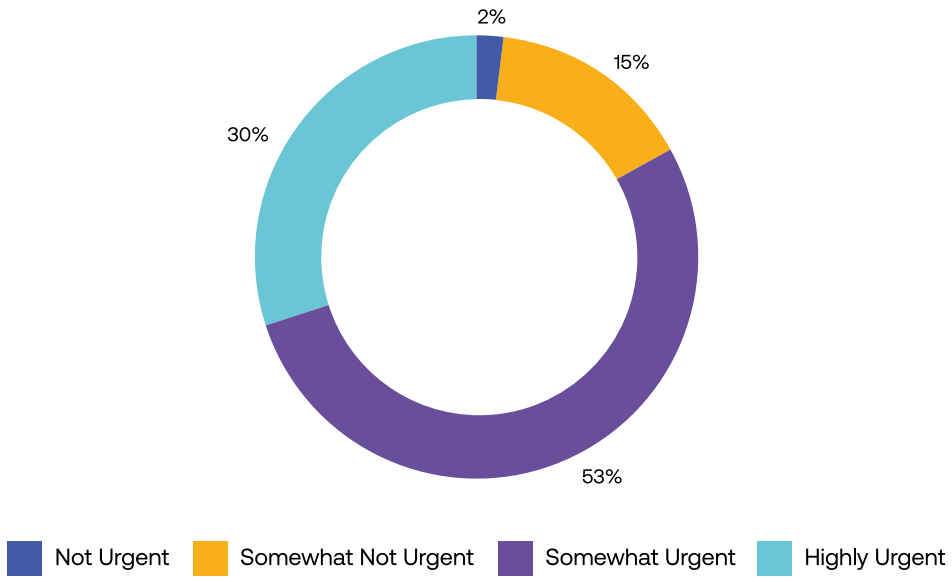
METHOD
COMMUNICATIONS

In boardrooms and strategy sessions across the globe, artificial intelligence has moved from a distant aspiration to an immediate priority. Our extensive survey of business leaders reveals a notable reality: AI adoption is accelerating rapidly, with 82% of global businesses already deploying AI applications in their day-to-day operations.. This broad adoption of artificial intelligence spans industries and regions, demonstrating that AI is no longer the exclusive domain of tech giants but has become an essential technology for enterprises of all types.

Customer service stands at the forefront of this adoption wave, with 63% of organizations deploying AI to enhance customer interactions. Document processing (54%), IT operations (51%), and security applications (51%) follow closely behind, showing how AI has rapidly infiltrated core business functions. This isn't merely experimentation—it's transformation at scale.

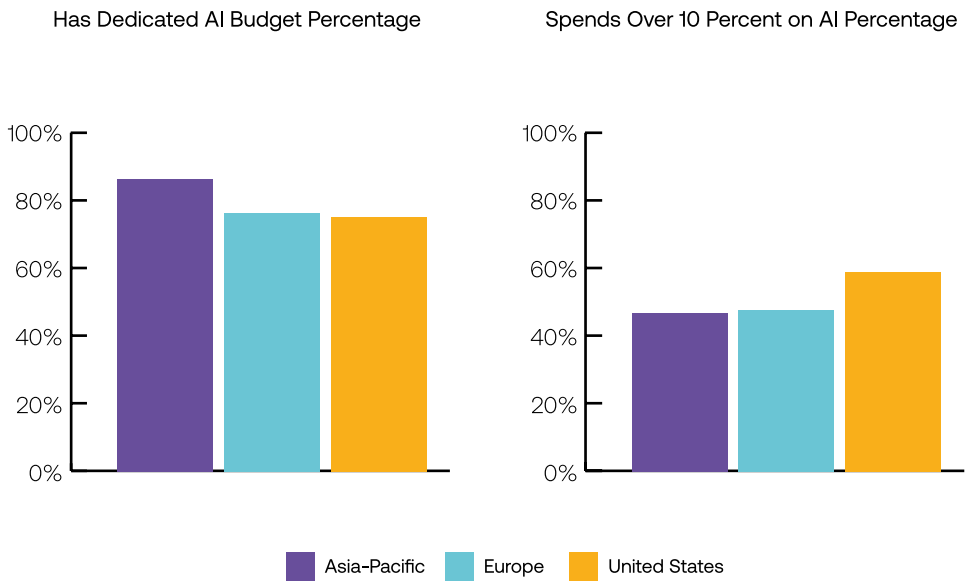
The enthusiasm for AI runs deep within organizational hierarchies. An overwhelming 90% of leaders report their organizations are receptive to AI-driven changes, with 83% considering AI adoption "urgent."

83% of Global Leaders Believe it is Urgent for their Organizations to Embrace AI; with 30% Saying it's Highly Urgent



Most tellingly, 82% of businesses have secured executive sponsorship at the highest levels, with CEOs and senior management personally championing AI initiatives. This top-down commitment signals that AI has transcended technical curiosity to become a strategic imperative.

Financial commitments reinforce this strategic shift. Eight in ten organizations have established dedicated AI budgets, with regional variations highlighting different approaches to investment. The Asia-Pacific region leads in budget allocation, with 86% of APAC businesses dedicating resources to AI initiatives, compared to 76% in Europe and 75% in the United States. However, American organizations are investing more aggressively, with 57% committing more than 10% of their IT budgets to AI, surpassing European (46%) and APAC (45%) counterparts.



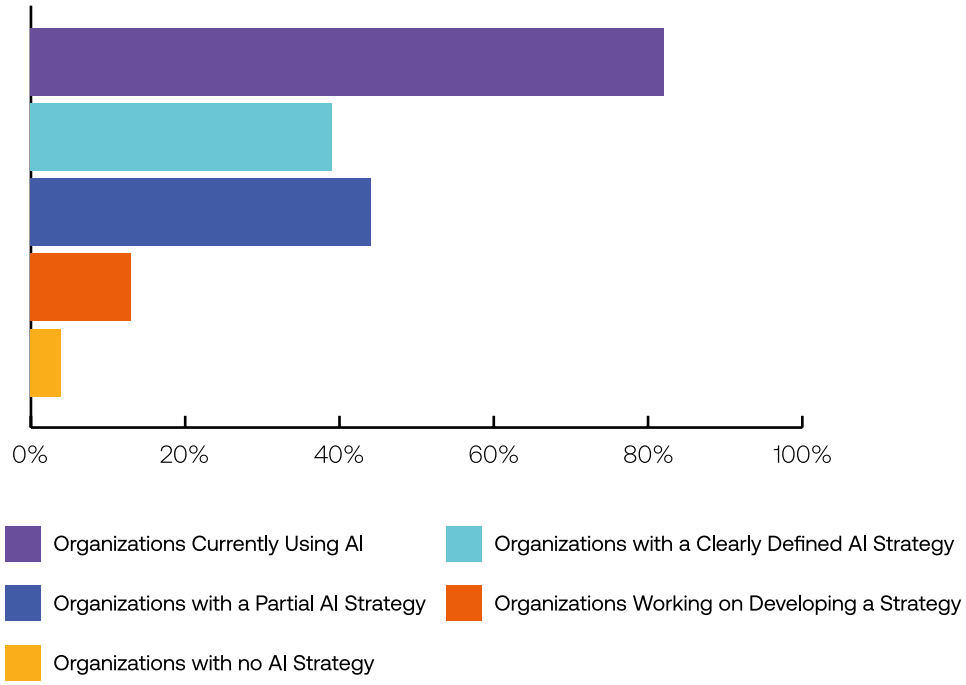
This financial commitment shows no signs of wavering, as 87% of business leaders anticipate increasing their AI budgets over the next three years.

"AI will eliminate repetitive work, allowing employees to engage in more strategic and creative tasks." - *Business leader survey respondent*

What's driving this rush toward AI? In a word, efficiency. A compelling 63% of global leaders identify operational efficiency as the primary benefit they expect from AI, with 80% making it the central focus of their 2025 AI strategy. As one business leader articulated, "AI will eliminate repetitive work, allowing employees to engage in more strategic and creative tasks." Another explained how "AI advancements will impact our organization by enhancing efficiency in decision-making and innovation across operations while requiring continuous adaptation to new technologies and regulatory changes."

Yet beneath this momentum lies a concerning paradox.

Adoption-Strategy Paradox



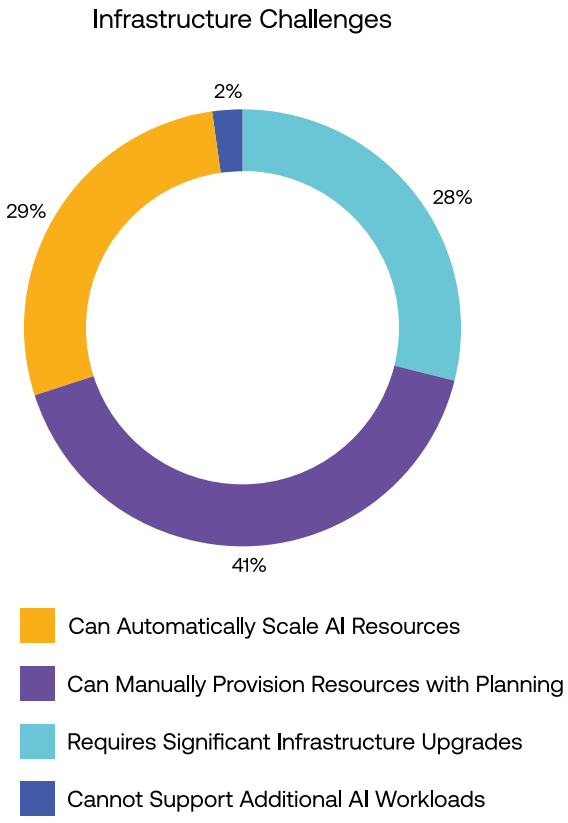
Despite widespread adoption and executive support, only 39% of organizations have developed a clearly defined, comprehensive AI strategy. Even fewer—just 37%—have implemented a robust change management plan to guide AI implementation across their businesses. This gap underscores that while AI is deployed, many organizations may not fully understand how to integrate it effectively or ready their workforce for the changes ahead.

One Vice President in the financial industry's IT department acknowledged: "We have our work cut out for us in formalizing these toolsets. Their use has grown organically within the business units, so we have duplication. I think we're getting there, and that's certainly part of my role—to figure that out."

This lack of a cohesive strategy raises critical questions about the sustainability of current AI initiatives and their long-term impact on business performance.

PROCEED WITH CAUTION

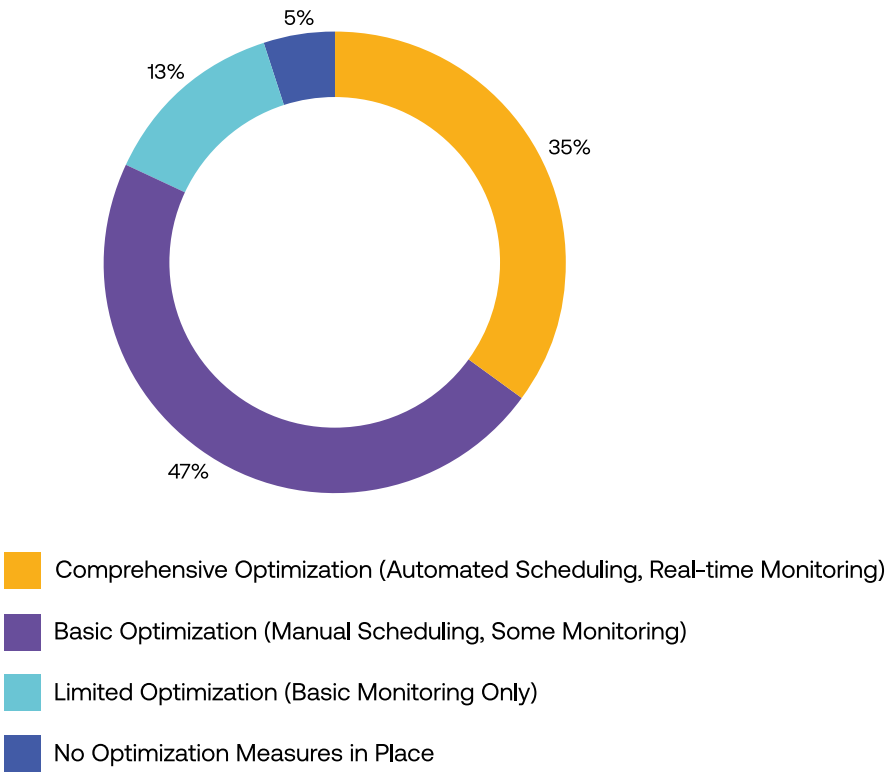
The enthusiasm for AI must be tempered with a sobering reality: many organizations remain ill-equipped to scale their AI initiatives effectively. Our research reveals significant shortcomings in three critical areas—**infrastructure readiness, talent availability**, and **data quality**—each representing a potential barrier to realizing AI's full potential.



Infrastructure limitations are particularly concerning. Only 29% of organizations have systems or storage resources to meet growing AI

demands. Even fewer—a mere 23%—have dedicated power infrastructure to handle the increased energy requirements of AI workloads. This means **77% of businesses have only begun upgrading their facilities**—if they have any power management strategy. Additionally, 65% of leaders acknowledge that their organizations lack comprehensive energy efficiency optimization measures for AI systems, raising questions about operational costs and environmental impact as these workloads grow.

AI Efficiency Optimization Level



The talent gap poses an equally significant challenge. A third of business leaders (34%) report their organizations are “significantly under-resourced” or “under-resourced” when it comes to AI expertise, while nearly half (49%) identify a lack of skilled talent as a primary barrier to successful AI implementation. As one C-level executive in manufacturing lamented, “It’s about manpower talent. It’s so hard to find people who know AI and can talk about AI.”

Despite recognizing the talent gap, organizations are taking inconsistent approaches to addressing it. While two-thirds (66%) of leaders express intentions to upskill existing employees to adapt to increased AI integration, 39% still don't have dedicated programs to develop AI skills among their workforce. This disconnect between recognized need and decisive action threatens to widen the talent gap rather than close it.

Perhaps most fundamentally, organizations struggle with data readiness—the essential foundation for effective AI applications. Nearly half of businesses leverage customer data (49%) and operational data (48%) for AI initiatives. However, data management practices remain largely rudimentary. Only about half (53%) of organizations have basic data automation processes for AI/ML models, while 18% rely on manual or ad hoc data cleaning procedures. Not surprisingly, 46% of leaders cite data quality and accessibility issues as a major barrier to successful AI implementation.

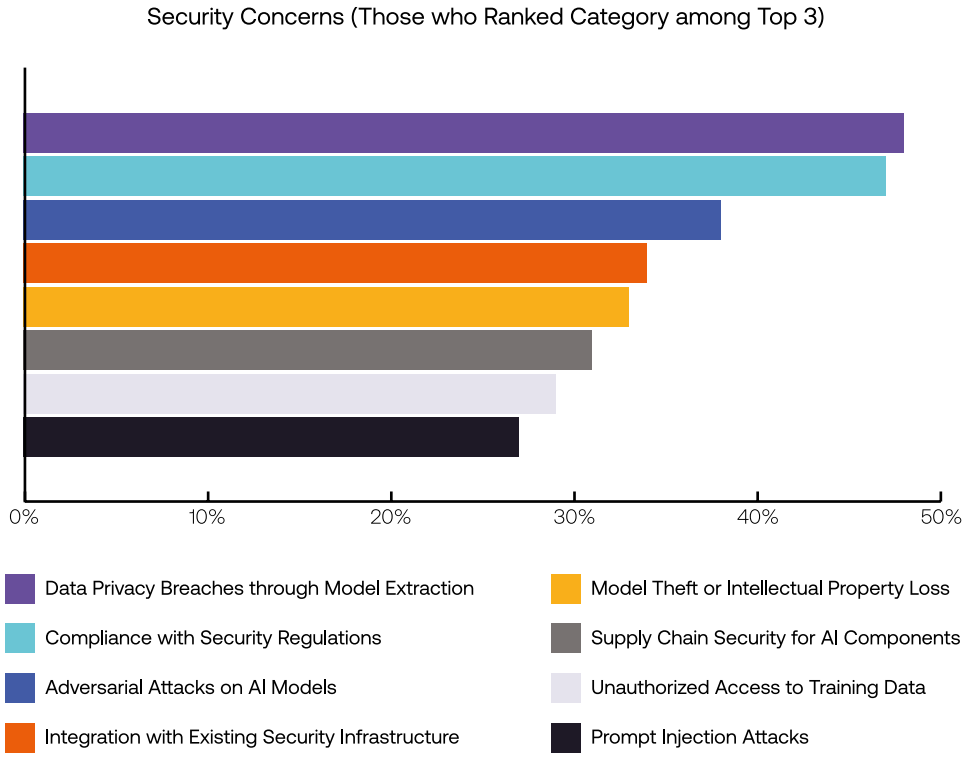
"We have our work cut out for us in formalizing these tool sets. Their use has largely grown organically within the business units, and that's why we have duplication."
- Vice President within the Information Technology department, working for an organization in the financial industry

Industry leaders recognize these challenges. One Vice President in the financial industry's IT department explained: "Our data quality and cataloging is so critical when it comes to AI...That is work in front of us." Another C-level IT executive in manufacturing emphasized the need for integrated data platforms: "We need to get the data into one platform so we can try to leverage the other functions to use AI."

These infrastructure, talent, and data challenges cannot be overlooked. Without addressing these fundamental capabilities, organizations risk their AI investments delivering diminishing returns as they attempt to scale beyond initial proof-of-concept deployments.

SECURITY SENSITIVITY

As organizations increasingly integrate AI into their operations, a new dimension of risk has emerged that demands urgent attention: data security and privacy. Security concerns are growing proportionally, with personally identifiable information (PII) rapidly becoming central to AI initiatives.



Nearly half (49%) of businesses already use customer data to power their AI initiatives, and the trend is accelerating. Looking ahead, 56% of leaders indicate their organizations plan to utilize personally identifiable information for future AI applications. This growing reliance on sensitive data introduces significant security considerations.

As one director in the life sciences and biotechnology industry explained: "I think when you get into some of these companies, ours in particular, we have to be a little bit more safeguarded because of the data sets we have.

Not all of it can be shared in public knowledge, but AI makes looking at that data, interpreting that data, analyzing that data 1000 times easier."

The expanded use of sensitive data in AI systems introduces security considerations and ethical concerns about bias and fairness. AI systems are only as objective as the data they're trained on, raising important questions about how organizations ensure their AI applications make fair and unbiased decisions.

"AI advancements will impact our organization by enhancing efficiency in decision-making and innovation across operations while requiring continuous adaptation to new technologies and regulatory changes."
- Business leader survey respondent

Our research reveals a significant gap in this area. Nearly half (47%) of business leaders acknowledge their organizations have limited bias detection and correction processes in their AI systems. Even more concerning, 17% reported having no formal bias correction process and relying instead on ad hoc bias checks. This inconsistent approach to bias mitigation poses significant risks, especially as AI becomes more deeply integrated into consequential business decisions.

Forward-thinking leaders recognize these challenges. As one Vice President in the financial industry observed: "We have to make sure that data is bias-free, and to do that, we have to come up with policies and standards controls that need to be implemented within the organization by support of data scientists and data stewards." This recognition drives 44% of leaders to identify AI ethics and data engineering as the most critical skills their organizations will need in the next five years.

On a more positive note, organizations are beginning to implement security measures for their AI systems. Only 5% of businesses report no specific AI security measures. The vast majority are taking at least some precautions, with 56% conducting regular security audits of AI models and infrastructure, 56% performing security testing of AI-powered applications, and 51% conducting vulnerability assessments for AI systems.

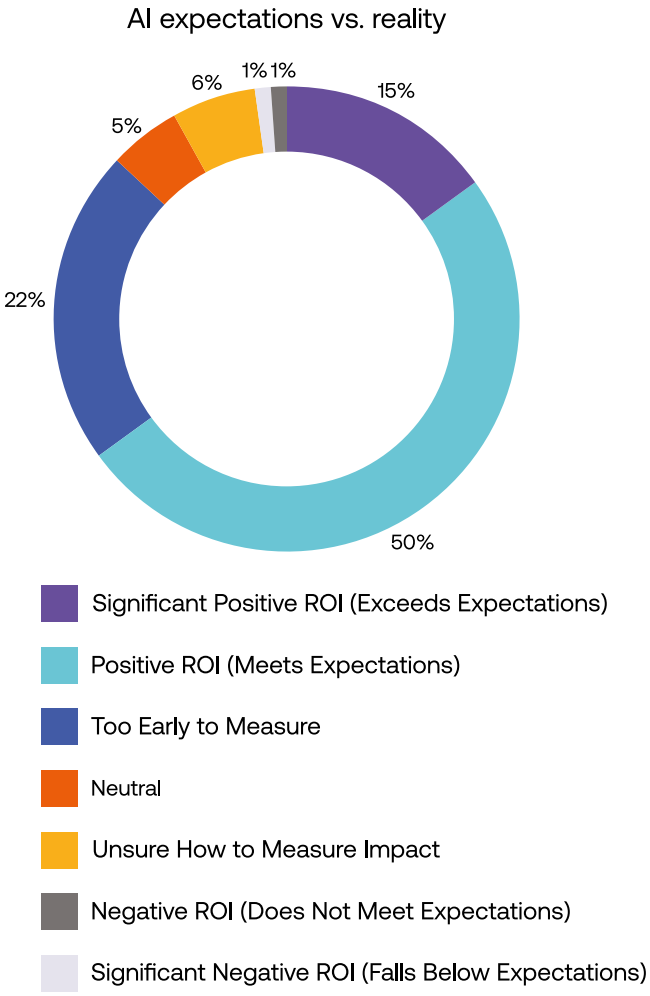
Despite these measures, significant concerns persist. Nearly half (48%) of business leaders identify data privacy breaches through model extraction as a top AI security concern. One Vice President of Marketing and Sales in retail observed: "I think there's a big open question on IT security that large companies are going to worry about."

As AI systems become more deeply integrated with sensitive business and customer data, organizations must evolve from basic security practices to comprehensive security strategies that address the unique vulnerabilities associated with AI technologies.

PREPARING FOR AN AI-POWERED FUTURE

One prediction seems certain as we look toward the horizon: AI adoption will continue to accelerate across all industries. Organizations that prepare effectively today will position themselves to capitalize on this technological revolution, while those that fail to address fundamental readiness gaps risk falling behind.

The business case for AI continues to strengthen among organizations using AI; 65% report that their AI applications are meeting or exceeding expectations for return on investment. This positive experience drives further adoption, with 92% of business leaders indicating plans to expand their AI usage in some capacity.



However, successfully scaling AI initiatives requires addressing several critical preparation gaps. A third (34%) of business leaders identify integration with existing security infrastructure as a top AI security concern. Meanwhile, 43% acknowledge their employees demonstrate only a basic understanding of AI at best, with additional training urgently needed. These challenges have led two-thirds (67%) of business leaders to identify AI security among the most critical AI-related skills their organizations will need to develop over the next five years.

Organizations must balance enthusiasm with strategic planning as they prepare for an AI-powered future. By systematically addressing

infrastructure limitations, closing talent gaps, enhancing data quality, and strengthening security measures, businesses can move beyond initial experimentation to realize AI's transformative potential at scale.

The path forward is clear, if challenging. Organizations must translate their AI enthusiasm into comprehensive strategies that address foundational readiness gaps. Those who successfully navigate this transition will be positioned to reap the efficiency gains and competitive advantages that AI promises. For the rest, the gap between AI ambition and AI readiness may prove increasingly difficult to bridge.

82%

of global business leaders are currently using AI applications.

39%

say their organization has a clearly defined, comprehensive AI strategy in place.

29%

of organizations can automatically scale compute or storage resources to meet AI demands.

23%

have dedicated power infrastructure to manage increased power demands for AI workloads.

95%

have implemented some kind of AI security measures.

arm

CHAPTER 2

The Technical Foundation: Technology Requirements for AI Success

DR JOHN SOLDATOS,
HONORARY RESEARCH FELLOW AT THE UNIVERSITY OF GLASGOW

2.1 OVERVIEW OF AI TECHNOLOGY REQUIREMENTS

2.1.1 The Need for a Robust Technology Infrastructure

Artificial Intelligence (AI) systems are advancing rapidly, becoming more flexible, scalable, and powerful. But to unlock their full potential, they need a solid technological foundation. Robust infrastructure is critical to managing the increasing complexity, scale, and sophistication of AI applications. It must support vast data processing, storage, and analysis while ensuring models run efficiently. As AI workloads grow, infrastructure must evolve to handle diverse tasks, reduce latency, and process massive data volumes—all while maintaining the speed and reliability modern applications demand.

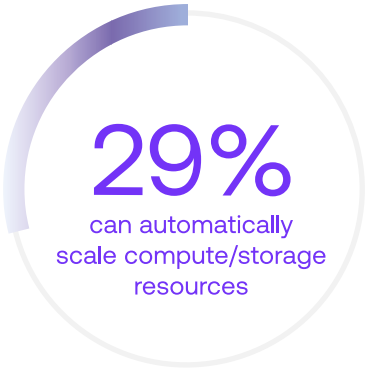
2.1.2 Technology Requirements: Main Challenges in Key Technological Areas

The development and deployment of robust AI technology infrastructures leverage various state-of-the-art developments, including high-performance computing architectures, power-efficient systems, scalable

infrastructures, and a proper balance between edge AI and cloud AI deployments. Recent advances in the above-listed areas help overcome considerable performance, scaling, and energy efficiency challenges. Specifically:

- **Evolution of computing architectures**: Over the past three decades we have witnessed a transition from simple, single-tier systems to complex multi-tier architectures that segment various aspects of an application (e.g., user interface, data access, business logic) into different functional layers. This transition has been driven by the need for more flexible, adaptable, and scalable systems. In line with this evolution, early AI systems were limited in their inability to efficiently process large datasets or perform complex computations. To address this limitation, there has been a shift towards multi-tier systems. This shift introduces multiple learning layers and service-oriented designs, enhancing system capabilities and performance.
- **Infrastructure scaling**: As AI models become larger and more sophisticated, there is a need for infrastructures that can scale in a cost-effective manner. As a prominent example, the size of Large Language Models (LLMs) and their training datasets are increasing at an unprecedented pace. This demands for infrastructures capable of scaling to accommodate both training data and data processing workloads. Recent advances in distributed computing offer solutions to these challenges, notably, solutions handle workloads at large scale without compromising speed or efficiency. Nevertheless, there is still a need for effective resource management strategies that optimize resource utilization while maintaining cost efficiency.
- **Edge vs. cloud computing considerations**: In recent years, organizations have managed workloads across the edge-cloud computing continuum. In this direction, the benefits of localized processing are confronted against those of centralized resources. In principle, edge computing reduces latency by processing data closer to its source, making it ideal for real-time applications. In contrast, cloud computing provides scalable

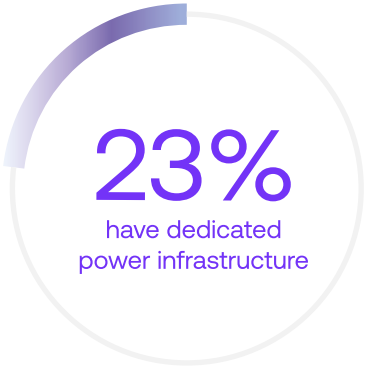
IT infrastructure readiness



access to virtually any number of computing resources, yet it also introduces latency issues for certain applications.

- **Power-performance trade-offs**: Today, energy consumption is one of the major concerns surrounding the deployment of AI systems at a very large scale. AI models are gradually growing in complexity, asking for more computational power and leads to increased energy consumption. Therefore, there is a need to balance high performance with energy efficiency, which is typically a significant challenge. At the same time, organizations must carefully consider the trade-offs between different processing architectures based on their specific workload requirements and constraints. Energy-efficient CPU architectures, like those based on Arm technology, have been crucial for enabling AI workloads at scale in data centers and edge computing environments. These processors combine general-purpose computing capabilities with optimizations for AI operations, providing an efficient foundation for inference workloads where versatility and power efficiency are paramount.
- **Specialized AI hardware**: Specialized AI hardware has also emerged to address specific computational challenges. Graphics Processing Units (GPUs) excel at parallel processing tasks common in AI training. Tensor Processing Units (TPUs) and Neural Processing Units (NPUs) are purpose-built for AI workloads, offering high performance per watt for specific types of neural network operations. Each of these technologies presents different trade-offs between computational power, energy efficiency, flexibility, and cost. Modern industrial enterprises are therefore offered a host of AI hardware options (e.g., AI-optimized CPUs, GPUs, TPUs, NPUs), which can be combined to meet their AI deployment requirements. In this context, companies must understand the capabilities of each option and how they can be integrated to serve their particular needs.

Power consumption management



2.1.3 User-driven technologies: Aligning infrastructure with end-user requirements

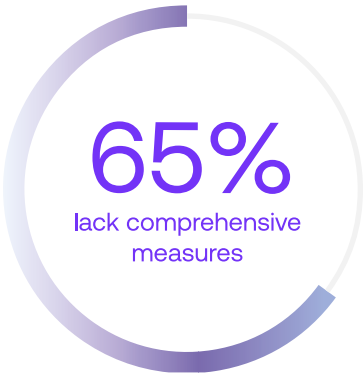
To ensure scalability, speed, and efficiency, it is also important to align technology infrastructure with end-user requirements. This involves understanding the specific needs of AI workloads, such as latency sensitivity and data volumes. In principle, AI infrastructures must provide tailored solutions for different AI application profiles in order to help ensure that performance, scalability, latency, energy efficiency, and other requirements are met. Emerging AI infrastructures must be driven by end-user requirements as much as they are driven by the evolution of AI technologies.

2.2 THE ARCHITECTURAL FOUNDATIONS OF MODERN AI

2.2.1 Edge-Cloud Integration: A New Paradigm for AI Infrastructure

The evolution of AI system architectures has played an essential role in enabling the development, deployment, and operation of state-of-the-art AI systems, such as edge-cloud AI systems and generative AI systems. For instance, large-scale generative AI systems (e.g., ChatGPT) and distributed learning systems (e.g., federated learning) could hardly be deployed based on old-fashioned, monolithic, single-tier architectures. Rather, they have become possible with the emergence of distributed, multi-tier, and hybrid architectures, which have opened new horizons for the scalability, adaptability, modularity, and performance of AI systems.

Energy efficiency optimization



In particular, AI architectures have undergone significant transformations over the decades. The main milestones in this evolution of AI architectures include:

- **Single-tier and multi-tiered systems**: As already outlined, AI systems were rarely monolithic. They consolidated user interface, logic, and data storage components into a single executable module. This offered

development and deployment simplicity yet limited scalability and adaptability. As such, single-tier architectures were unsuitable for non-trivial AI systems and tasks. To address these limitations, multi-tier (n-tier) architectures that segmented functionalities into specialized layers have emerged. Multi-tier systems enable modular development and distributed processing, which boost efficiency and scalability.

Cloud-based architectures: During the last two decades, the rise of cloud computing introduced on-demand access to computational resources, while also offering elasticity, capacity, and quality of services. The latter features have opened new horizons in the scalability and flexibility of enterprise workloads, including non-trivial AI applications.

- **Edge computing**: The advent of IoT devices, real-time applications, and mobile applications hosted in mobile devices like smartphones imposed new requirements that could hardly be met by cloud computing infrastructures. This led to the concept of edge computing, which is about processing data close to the data sources rather than to the cloud. Edge computing provides an effective solution for reducing latency, improving energy efficiency, and ensuring the confidentiality of sensitive datasets.

"Modern AI infrastructures typically employ a heterogeneous computing approach that leverages different processor types for different tasks."

Today, AI architectures are not purely cloud-based but instead support hybrid models. Hybrid architectures combine on-premises infrastructure with cloud resources to balance flexibility and control. For instance, in a hybrid deployment, on-premises systems can be used to handle sensitive data and/or latency-critical tasks, while cloud resources are used to provide scalability for training large models or storing vast datasets. Advanced hybrid solutions integrate cloud and edge computing to optimize performance. For instance, edge devices preprocess data locally before sending it to the cloud for deeper analysis. This helps leverage the advantage of both edge and cloud. In particular, edge-cloud deployments combine the scalability, cost flexibility (pay-as-you-go), and intelligent centralized resource management capabilities of the cloud with the reduced latency and enhanced privacy characteristics of the edge.

2.2.2 AI Models and Hardware

The evolution of AI system architectures has played an essential role in enabling the development and deployment of different AI models. Specifically, modern AI systems integrate a variety of models that feature different architectures, including:

- **Neural networks**: Neural networks are the foundational architecture for many AI systems. They consist of interconnected nodes (neurons) that mimic the human brain's structure. Prominent types of Neural Networks include

  (i) **Feedforward Neural Networks (FNNs)**, where data flows in one direction, from input to output. FNNs are used in basic classification and regression tasks;
- (ii) **Convolutional Neural Networks (CNNs)**, which are specialized for image processing and pattern recognition. As a prominent example, the AlexNet CNN, revolutionized image classification by using deep convolutional layers;
- (iii) **Recurrent Neural Networks (RNNs)**, which are designed for sequential data with feedback loops to retain memory. As a prominent example, Long Short-Term Memory (LSTM) networks are extensively used in time-series forecasting and natural language processing (NLP);
- (iv) **Residual Networks (ResNet)**, which introduced skip connections to address vanishing gradient issues in deep networks. As an example, ResNet50 is widely used in computer vision tasks like object detection. Overall, Neural networks are used in a wide variety of applications and tasks, including image recognition, speech-to-text systems, medical diagnosis, and more.
- **Transformer models**: Transformers have revolutionized AI by enabling parallel processing and self-attention mechanisms in order to handle sequential data in very efficient ways. From an architecture viewpoint, a transformer model consists of an encoder-decoder structure. The encoder processes input sequences into representations. At the same time, the decoder generates outputs based on these representations.

The core components of transformers also include multi-head attention layers and feedforward networks. One of the most prominent examples of transformer models is **BERT (Bidirectional Encoder Representations from Transformers)**, which is an encoder-only model for tasks like text classification and sentiment analysis. Another example is the **T5 (Text-to-Text Transfer Transformer)** model, which is an encoder-decoder model for translation, summarization, and other NLP tasks. Most importantly, the very popular GPT (Generative Pre-trained Transformer) a family of models, i.e., the models behind ChatGPT, are also decoder-only models for text generation and conversational AI. In general, transformers excel in NLP tasks such as machine translation, text summarization, and question answering. They are also applied in protein folding (e.g., AlphaFold) and computer vision tasks like object detection.

- **Generative AI Models**: Generative models create new data resembling the input data they were trained on. As a prominent example, Generative Adversarial Networks (GANs) comprise a generator that creates synthetic data and a discriminator that evaluates its authenticity. Notable examples of GANs are the DCGAN (Deep Convolutional Generative Adversarial Networks) for image generation and the CycleGAN for domain transfer, like turning photos into paintings. Another class of Generative AI Models is the Variational Autoencoders (VAEs), which learn latent representations to generate new data points. VAEs are used to generate realistic images or interpolate between data points. Overall, generative models are used in creative tasks like art generation, synthetic data creation, and drug discovery.
- **Multi-modal models**: These models process and integrate multiple types of data (e.g., text, images). For instance, **CLIP (Contrastive Language–Image Pre-training)** aligns visual inputs with textual descriptions for tasks like image captioning or zero-shot classification. As another example, **DALL-E**: Generates images based on textual prompts by combining NLP and computer vision capabilities.

Multi-modal models are usually applied in content creation, augmented reality, and cross-modal retrieval systems.

- **Other specialized architectures**: There are also other models tailored for specific domains or tasks. For instance, **Vision Transformers (ViTs)** adapt transformers for image processing by dividing images into patches. They are used in tasks like object detection and segmentation. As another example, **Deep Reinforcement Learning Models** combine neural networks with reinforcement learning techniques. The AlphaZero model, notorious for mastering board games like chess and GO, is a deep reinforcement learning model.

Overall, AI system architectures have evolved from basic neural networks to advanced designs like transformers and generative models. Each category serves a unique purpose. Neural networks provide foundational AI capabilities, while transformers dominate NLP and multi-modal applications. Generative AI models enable creative outputs, and multi-modal models bridge different data types.

Moreover, modern AI architectures and the above-listed AI models benefit from specialized hardware. Specifically:

- **CPUs** are fundamental to AI infrastructure, particularly for inference workloads where their versatility and power efficiency are crucial. Modern CPU architectures optimized for AI workloads combine sophisticated vector processing capabilities with energy efficiency, making them ideal for deploying AI at scale in datacenters and edge devices. Their ability to efficiently handle diverse workloads, from preprocessing to inference, makes CPUs essential for real-world AI deployments where flexibility and cost-effectiveness are paramount.
- **GPUs** excel at parallel processing. Their ability to handle matrix operations efficiently make them indispensable for training deep learning models. Today, AI companies are massively deploying GPUs to enable the provision of their products or services at scale.

– **NPUs** are purpose-built processors designed specifically for accelerating neural network computations. By incorporating dedicated circuitry for common AI operations like convolutions and matrix multiplication, NPUs achieve high performance per watt for both training and inference tasks. Their architecture makes them particularly effective for specialized edge AI applications where power constraints are critical, enabling sophisticated AI capabilities in mobile devices, IoT sensors, and other resource-constrained environments.
– **TPUs** are optimized for tensor-based operations, which provide high throughput for large-scale deep-learning tasks. While GPUs are versatile across various applications, TPUs tend to specialize in accelerating neural network computations.

AI-optimized hardware solutions are significantly reducing training and inferencing times and computational costs for complex AI models.

2.2.3 Trends in AI Systems and Architectures

Some of the most prominent AI systems trends and their AI hardware implications include:

– **Integration with machine learning frameworks**: Nowadays, distributed computing platforms (e.g., Apache Spark or Hadoop) are increasingly integrated with frameworks such as TensorFlow and PyTorch. This combination enhances scalability while simplifying model deployment across distributed environments.
– **Modular datacenters**: Modular designs allow datacenters to scale incrementally by adding or replacing components as needed. This approach caters to increased scalability and sustainability, as it reduces both energy consumption and construction costs.
– **Distributed training of AI models**: Distributed training techniques take advantage of multiple computing nodes to enable faster model training. This is particularly beneficial for large-scale transformer models like GPT

(Generative Pre-training Transformer) or BERT (Bidirectional Encoder Representations from Transformers).

- **Federated learning**: Federated learning allows decentralized devices to collaboratively train global, more accurate models without sharing raw data. This approach enhances privacy while reducing bandwidth usage.
- **Fine-tuning**: Fine-tuning adapts a pre-trained model to perform specialized tasks with greater accuracy. Instead of training a model from scratch, fine-tuning leverages the knowledge embedded in a pre-trained model, such as GPT or BERT, and refines it using task-specific data. This approach reduces computational costs and training time while improving performance on domain-specific tasks. Fine-tuning can range from full optimization of all model layers to partial fine-tuning, where only specific layers or parameters are updated. It is already widely used for tasks such as customizing conversational tones in chatbots, enhancing domain-specific knowledge in legal or medical applications, and addressing edge cases not covered during pre-training.
- **Retrieval-Augmented Generation (RAG)**: RAG enhances the capabilities of generative models by integrating real-time information retrieval. Unlike traditional large language models (LLMs) that rely solely on static training data, RAG incorporates external data sources (e.g., databases, APIs, document repositories) into its responses. This dynamic retrieval process allows RAG systems to provide more accurate, context-aware, and up-to-date outputs. The process involves indexing relevant data into vector databases, retrieving pertinent information based on user queries, augmenting the input with this data, and generating responses that combine both retrieved and pre-trained knowledge. RAG is already used in numerous applications, including advanced question-answering systems, content summarization, conversational agents, and personalized educational tools. Emerging trends in RAG include hybrid search techniques that combine structured and unstructured data retrieval, multimodal RAG for integrating text with images or audio, and on-device RAG for enhanced privacy and reduced latency.

"Recent studies by the International Energy Agency estimate that AI workloads now account for 10-15% of total electricity usage in data centers, with projections suggesting this could reach 25-30% by 2030."

RAG demands efficient data processing and inference based on the following key hardware components: (i) CPUs (e.g., multi-core processors like Intel Xeon or AMD EPYC) for managing data ingestion, preprocessing, and vector search operations; (ii) GPUs for embedding generation and LLM inference. These help ensure high throughput and low latency during real-time retrieval and response generation; (iii) Random Access Memory (RAM) (e.g., 64GB or more) to properly handle embeddings and vector databases; (iv) Fast NVMe SSDs for storing large datasets and vector indices in ways that help ensure quick access during retrieval steps; (v) Scalability solutions such as cloud-based solutions that can offload storage and search workloads in order to reduce the need for extensive on-premises hardware while maintaining scalability.

- **Agentic AI**: Agentic AI refers to AI systems designed to operate autonomously while exhibiting decision-making capabilities similar to human agents. These systems go beyond passive data processing to actively interact with their environment, adapt based on feedback, and pursue goals independently. Agentic AI incorporates elements like reinforcement learning and multi-agent collaboration to optimize its performance in dynamic settings. For instance, agentic AI systems can autonomously navigate supply chain logistics or manage financial portfolios based on real-time data. Agentic AI is closely tied to advancements in areas like reinforcement, fine-tuning and adaptive learning. These techniques enable AI agents to refine their decision-making processes over time while maintaining alignment with predefined objectives. Agentic AI systems are complex as they involve autonomous decision-making, learning, and action modules. Their hardware demands are driven by the need to process multimodal data, execute actions in real time, and adapt dynamically.

These hardware requirements include:

- (i) **High-performance CPUs** (e.g., Intel Xeon or AMD Zen 4) **paired with GPUs** like NVIDIA A100 or H100 in order to support the perception

module (e.g., computer vision) and decision-making engines that rely on reinforcement learning;

(ii) **Large memory capacities (128GB RAM or more)** are needed to manage simultaneous tasks across multiple agents in a modular architecture;

(iii) **NVMe SSDs** ensure fast access to knowledge graphs, training datasets, and historical data used by cognitive modules;

(iv) Networking & Power as distributed Agentic AI systems require robust networking infrastructure to facilitate communication between agents. Power-efficient designs are, therefore, critical to supporting long-term operations in autonomous environments.

2.3 POWER AND PERFORMANCE SOLUTIONS

2.3.1 Computational Demands

Large-scale AI models, such as large language models (LLMs) and generative AI systems, are growing exponentially in size and complexity. These models consist of hundreds of billions or even trillions of parameters, requiring large amounts of computational resources for efficient training and deployment.

Modern AI infrastructures typically employ a heterogeneous computing approach that leverages different processor types for different tasks. The foundation often begins with high-performance CPUs, which provide the versatile computing capabilities needed for data preprocessing, inference, and orchestrating complex AI workloads. Energy-efficient CPU architectures have proven particularly valuable for inference at scale, where power efficiency and consistent performance are crucial.

For training-intensive workloads, organizations often augment their CPU infrastructure with specialized hardware accelerators. GPUs excel at the parallel processing needed for model training, while NPUs offer optimized

performance for specific AI operations. This combination of general-purpose processors and specialized accelerators helps organizations balance performance, efficiency, and cost-effectiveness across their AI workflows.

The sheer scale of computational requirements continues to push the limits of existing infrastructures. Training state of the art LLMs like GPT-4 or Meta Llama 3 involves processing vast datasets using high-performance computing (HPC) setups that typically combine multiple types of processors working in concert. In this context, distributed computing frameworks and advanced hardware architectures that can effectively coordinate these various processing elements are required to meet these growing demands.

2.3.2 Computational Demands for Training and Inferencing

AI models have considerable computational demands at both their training and inferencing stages. These demands are significant given the complexity of models, the size of datasets, and the need for real-time performance. The hardware options previously presented play a considerable role in meeting these demands, providing specialized components optimized for different stages of AI workflows.

Hardware Support for AI Training

Training AI models is highly resource-intensive. It involves massive floating-point operations (FLOPs). For instance, training LLMs like GPT-4 requires tens of thousands of GPUs and consumes tens of gigawatt-hours of energy. Key factors influencing computational demand include:

- **Model complexity**, as larger models with billions or trillions of parameters require more compute power.
- **Dataset size,** as larger datasets increase training time and resource needs.
- **Parallel processing**, given that the efficient use of multiple processors is important when it comes to handling large-scale computations.

In this context, modern hardware for training AI includes:

- GPUs like the NVIDIA A100 or NVIDIA GeForce RTX™ 4090 are widely used due to their high parallel processing capabilities and ability to deliver trillions of operations per second (TOPS).
- Google's TPUs are optimized for deep learning tasks while offering high performance in datacenters.
- ASICs (Application-Specific Integrated Circuits), i.e., custom chips designed for specific AI tasks that provide energy efficiency and speed.
- GPUs with high memory bandwidth (e.g., 256 GB/s or more) are also used to help ensure efficient data transfer during training.

Hardware Support for AI Inferencing

Inferencing involves the application of trained models to new data in order to make predictions or decisions. While less computationally intense than training, they demand low latency and high throughput. The latter are very important in cases of real-time applications like autonomous vehicles or healthcare diagnostics. In this context, inferencing hardware focuses on speed and efficiency, including:

- Compact and power-efficient devices like NVIDIA Jetson Nano™ enable real-time inferencing at the edge.
- NPUs are typically specialized for inferencing tasks and can balance performance with energy efficiency.
- FPGAs (Field-Programmable Gate Arrays) offer flexible and efficient solutions for specific workloads yet are less powerful than GPUs.
- ASICs are also optimized for specific inference tasks in ways that offer low power consumption and high speeds.

There are also a variety of optimization techniques that are aimed to enhance inferencing performance, including:

- **Quantization**, which involves reducing model precision (e.g., from 32-bit to 8-bit) in order to improve speed and reduce memory usage.

– **Pruning**, which involves removing redundant model parameters to lower computational complexity without sacrificing accuracy.
– **Batching**, which processes multiple inputs simultaneously towards maximum GPU utilization.

The future of AI training and inferencing will be shaped by emerging trends in hardware development, including:

– **Distributed computing**, as inferencing is increasingly moving toward distributed systems to handle growing demands.
– **Energy efficiency**, given that advances in memory architectures and specialized chips aim to reduce energy consumption during both training and inferencing.
– **Real-time applications** based on edge computing which is becoming important for applications that require immediate responses (e.g., autonomous driving).

2.3.3 Energy Efficiency

AI workloads are notoriously energy-intensive, leveraging datacenters that consume significant power for computation and cooling. It is estimated that AI workloads account for 10–20 percent of total electricity usage in datacenters. This demand raises sustainability concerns due to the

associated carbon emissions. Overall, the main AI-related energy efficiency challenges are:

- **Datacenter power consumption**: The electricity required to train large AI models can be staggering. For instance, training a single LLM can consume energy equivalent to powering hundreds of homes for a year. To make matters worse, cooling systems are also highly energy-intensive.
- **Carbon footprint**: The environmental impact of AI is substantial and represents one of the main setbacks to its wider adoption. In particular, according to the International Energy Agency, modern datacenters contribute approximately 1 percent of global greenhouse gas emissions.

"Training a single LLM can produce carbon emissions equivalent to five cars over their lifetimes."

To address these power consumption challenges, AI vendors and integrators leverage the following solutions:

- **Energy-efficient chips**: Recent innovations in chip design are mitigating energy efficiency challenges. For example, the low-power Arm architecture optimizes performance per watt by minimizing energy consumption without sacrificing computational capability.
- **Sustainable datacenters**: Modular datacenters powered by renewable energy sources are emerging as a solution to reduce the carbon footprint of AI infrastructures. Furthermore, the use of advanced cooling technologies and virtualization techniques also improves operational efficiency.

2.3.4 Latency Sensitivity

Latency is another critical factor in AI applications, especially for the class of AI applications that incorporate real-time decision-making, such as autonomous vehicles, healthcare monitoring, and industrial automation applications. For instance, autonomous vehicles must process sensor data in milliseconds to make driving decisions, while wearable healthcare devices need low latency to monitor vital signs and alert users in near real-time.

To address latency challenges, AI vendors and integrators are employing local data processing at the edge of the AI deployment. Edge computing processes data closer to its source, which reduces round-trip times and improves responsiveness. The proliferation of AI applications with real-time constraints has given rise to the edge AI paradigm, which deploys powerful AI algorithms (e.g., deep neural networks) on edge devices like gateways or edge servers. One of the most prominent edge AI approaches is the TinyML paradigm, which deploys smaller AI models on ultra-low-power devices like microcontrollers towards enabling lightweight machine learning tasks with minimal latency. Edge AI systems commonly integrate hardware-level optimizations, which leverage advancements in memory bandwidth and dedicated AI accelerators to help minimize latency during AI computations. Apart from boosting real-time responsiveness, edge AI systems enhance privacy and improve bandwidth efficiency, as significantly less data is transmitted over the networking part of the AI infrastructure.

2.4 INFRASTRUCTURE SCALING REQUIREMENTS

2.4.1 Scalability Needs

AI infrastructures must scale in a cost-effective fashion in order to cope with growing data volumes and the increasing complexity of AI models. Large-scale AI workloads, such as training large language models (LLMs) and real-time inference systems, demand immense computational power and efficient resources. The exponential increase in data and compute demands leads to bottlenecks arising in storage, processing speed, and network bandwidth. Moreover, maintaining system performance while scaling infrastructure can be costly and operationally complex. To the rescue, distributed computing offers a scalable approach by dividing workloads across multiple nodes or machines. Parallel processing is also key for handling large-scale AI tasks more efficiently. In this direction, open-source platforms like Kubernetes enable the orchestration of distributed resources toward optimizing training times and inference processes.

2.4.2 Cost Management

Balancing performance with cost-effectiveness is another consideration when it comes to scaling AI infrastructures. Modern enterprises must navigate trade-offs between cloud-based elasticity and on-premises control to optimize resource utilization. Cloud computing provides dynamic scalability, which allows organizations to adjust resources based on workload demands without significant upfront investment. However, this flexibility comes at the cost of reduced control over infrastructure and potential long-term expenses for high usage. On the other hand, on-premises solutions offer greater control and customization but require substantial capital investments and ongoing maintenance costs.

At the same time, resource pooling is an effective strategy for optimizing computing utilization. Based on the aggregation of resources across multiple users or regions, organizations can maximize efficiency while minimizing idle time. For example, pooling workloads with non-overlapping peak demands allows for better utilization of shared resources. This approach reduces the need for over-provisioning while ensuring computational capacity during peak periods.

2.4.3 Modular Approaches

Modern modular infrastructure designs provide flexible solutions for scaling AI systems in an incremental fashion without requiring complete overhauls. Specifically, modular datacenters allow organizations to add or upgrade components as needed, enabling incremental scaling to meet growing demands. Modular datacenters are particularly valuable for edge computing environments where there are considerable space and power constraints.

2.5 EMERGING TRENDS IN AI INFRASTRUCTURE

2.5.1 Sustainability Initiatives: Shift Towards Greener Technologies

In an era where environmental performance and sustainability are at the very top of the political and enterprise agendas, there is a shift towards greener technologies to reduce the carbon footprint associated with AI infrastructure. This shift is powered by the following trends:

- **Renewable energy sources and green datacentres**: AI datacenters are transitioning to renewable energy sources such as solar and wind power. This shift not only reduces greenhouse gas emissions but also aligns with global sustainability goals. Companies like Google and Microsoft have committed to running their datacenters on 100% renewable energy, which set the bar high for other infrastructure providers in the industry. Moreover, Google has recently partnered with Kairos Power to deploy compact nuclear reactors for its AI-driven data centers. At the same time, Microsoft is funding the restart of a mothballed reactor at Three Mile Island to supply power for its operations by 2028.
- **Energy-efficient chips**: As already outlined, chip design innovations contribute to reducing energy consumption. Arm, for example, has developed low-power architectures that optimize performance per watt. These chips are designed to handle complex AI tasks while consuming less energy, which makes them ideal for both datacenters and edge devices.

2.5.2 Infrastructure Management Automation

Automation is becoming increasingly important in managing AI infrastructures. Generative AI (GenAI) tools have recently had a prominent role in realizing this automation. These tools enhance resource provisioning, workload optimization, and scaling decisions, which improve operational efficiency automation. More specifically, during the last couple of years, GenAI technologies have revolutionized resource allocation through auto-scaling features. Tools like K8sGPT enable GenAI-powered resource

management that adjusts resource provisioning dynamically based on real-time workload requirements. This helps to ensure that computational resources are allocated efficiently during peak usage while scaling down during low-demand periods. This level of automation reduces the need for manual intervention, which allows IT teams to focus on strategic initiatives rather than routine maintenance tasks. Overall, automation reduces operational costs without compromising performance.

Conclusion

"Edge computing processes data closer to its source, which reduces round-trip times and improves responsiveness."

As AI applications grow in size and complexity, there is a clear need for evolving the AI technology infrastructures. For over two decades, we have witnessed significant improvements in different aspects of AI infrastructures, including hardware, software, and architecture-related aspects. For instance, the transition from monolithic single-tier systems to distributed, multi-tier, and hybrid architectures has played an important role in addressing the scalability, adaptability, and performance challenges of AI systems. Nowadays, modern architectures integrate cutting-edge technologies such as CPUs, GPUs, TPUs, and microservices-based designs to enhance computational efficiency and flexibility.

Key infrastructure-related challenges such as balancing power-performance trade-offs, managing large-scale data processing, and minimizing latency have been effectively addressed through innovations in hardware and infrastructure design. In recent years, edge-cloud computing paradigms have emerged as a cornerstone for achieving low-latency processing while maintaining scalability and privacy. Furthermore, sustainability has taken center stage based on initiatives that reduce the environmental impact of AI operations. Prominent examples of such initiatives include renewable energy-powered datacenters and energy-efficient chip designs.

Moreover, the latest advances in automation are transforming the infrastructure management landscape. In the coming years, GenAI tools

that dynamically optimize resource allocation are expected to reduce operational costs while maintaining peak performance. Also, modular approaches to scaling infrastructure will provide flexibility for incremental growth without overhauling existing systems.

The above-listed advancements will enable businesses to overcome technical barriers and unlock the full potential of AI-driven solutions. Companies like Arm are playing a significant role in this transformation through innovative chip designs and ecosystem partnerships that address energy efficiency and computational demands. The current state and outlook of the AI technology infrastructure indicate that the future of AI will include not only technological breakthroughs but also a more efficient and environmentally responsible digital ecosystem.

ARM SIDEBAR

AI Everywhere: Arm's Vision for the Future of Compute

By Kevork Kechichian, EVP, Solutions Engineering, Arm

Artificial intelligence (AI) is reshaping computing, influencing everything from cloud-scale processing to real-time inference on personal devices. As AI adoption accelerates, the demand for high-performance, scalable, and energy-efficient compute platforms continues to grow. The challenge lies in enabling AI across a diverse range of environments—from cloud datacenters to power-constrained edge devices—without compromising efficiency or security.

The Arm compute platform plays a critical role in this transformation, providing the architectural foundation for AI workloads across industries. By enabling AI to run efficiently on CPUs, GPUs, and NPUs, Arm supports an ecosystem where AI can be deployed at scale. The approach is built on three fundamental pillars: accelerating AI from cloud to edge, delivering energy-efficient AI inference, and ensuring security and trust in AI workloads.

Accelerated AI Everywhere #onArm From Cloud to Edge

AI workloads span a vast spectrum of compute environments, ranging from cloud-based model training to on-device inference. The ability to run AI efficiently across this continuum is crucial for scaling AI applications and ensuring real-time performance. The Arm compute architecture is designed to support AI workloads at every level, enabling accelerated AI from cloud to edge.

In cloud environments, Arm-powered solutions (Nasdaq:ARM) provide scalable AI compute, helping datacenters optimize performance while maintaining energy efficiency. Arm CPUs are increasingly being adopted in cloud instances for AI inference, as they deliver high computational throughput at lower power consumption compared to traditional architectures. As AI models grow in complexity, cloud providers require efficient compute platforms that can handle large-scale workloads while managing operational costs.

Beyond the cloud, AI inference is rapidly shifting toward the edge, driven by the need for low-latency processing and enhanced privacy. Running AI on-device—whether on smartphones, industrial sensors, or embedded systems—reduces dependency on cloud resources and enables real-time decision-making. The Arm processor architecture facilitates this shift by providing a flexible foundation that supports diverse AI workloads, ensuring optimal performance across different hardware configurations.

The Armv9 architecture introduces key enhancements for AI acceleration, including Scalable Vector Extensions (SVE), which improve AI compute efficiency, and the Compute Matrix Engine (CME), which enhances vector processing capabilities. Additionally, Arm's heterogeneous computing

approach allows CPUs to work alongside GPUs and NPUs, providing a balanced AI execution framework tailored to specific workload demands.

As AI continues to scale, the ability to run workloads seamlessly from cloud to edge will become increasingly important. The Arm compute platform helps ensure that AI applications can be deployed across a broad range of devices while maintaining efficiency, flexibility, and scalability.

The Energy-Efficient Pervasive Compute Platform to Power GenAI

The exponential growth of AI workloads has raised concerns about power consumption and sustainability. Training large AI models requires significant computational resources, while AI inference at scale demands efficient processing to avoid excessive energy usage. The Arm compute architecture is designed to address this challenge, enabling AI workloads to run with high efficiency without compromising performance.

The power-efficient design philosophy of Arm is rooted in decades of innovation. The Armv9 architecture integrates advanced features that optimize AI compute while minimizing energy consumption. Technologies such as Helium for machine learning acceleration in IoT devices and SVE for improved AI inference performance allow AI applications to run with reduced power overhead. These innovations ensure that AI workloads can scale efficiently, whether in datacenters, edge devices, or battery-operated consumer electronics.

One of the key advantages of Arm-based solutions is their ability to provide high-performance AI inference while maintaining low power requirements. In the mobile industry, for example, over 70 percent of AI workloads in third-party applications already run on Arm CPUs. This trend is expanding

into PCs, industrial automation, and autonomous systems, where AI-driven applications require continuous processing within strict power constraints.

Sustainability is becoming a central focus in AI deployment, particularly in large-scale cloud operations. AI workloads contribute to increasing power demands in datacenters. By leveraging Arm-based compute platforms, cloud providers can optimize energy efficiency, reducing power consumption while maintaining high compute performance.

The Developer-Trusted Platform for Running AI Workloads Everywhere

As AI adoption grows, security has become a critical factor in ensuring trust and reliability in AI workloads. AI models process vast amounts of data, often handling sensitive or proprietary information, making protection against security threats, data breaches, and unauthorized access essential.

"Arm's compute platform plays a critical role in this transformation, providing the architectural foundation for AI workloads across industries."

Arm takes a security-first approach, integrating robust security features into its compute architectures to provide a trusted foundation for AI deployment across industries. The Armv9 architecture introduces Arm Confidential Compute Architecture (CCA) and Realm Management Extension (RME), ensuring that silicon protects AI data and code wherever computing happens. These security enhancements are crucial as AI inference shifts from cloud to edge, requiring devices to process sensitive data locally without exposing it to external threats.

By integrating hardware-based security features, the Arm architecture enables privacy-preserving AI, safeguarding data at rest and in transit. This is especially relevant in sectors such as healthcare, finance, and autonomous systems, where AI models handle confidential information. Additionally, in cloud environments, the Arm security framework provides

hardware-enforced protections that mitigate risks associated with multi-tenant infrastructures and prevent unauthorized model access.

As AI continues to scale, security must remain at the core of compute design. By embedding security at the architectural level, Arm ensures that AI applications operate in a trusted, resilient, and privacy-focused environment, empowering organizations to deploy AI workloads with confidence.

CHAPTER 8

Case studies

Streamlining ADAS Integration for Safer, Smarter Vehicles with LeddarTech

INTRODUCTION

As the automotive industry undergoes rapid transformation, vehicle OEMs face mounting pressure to accelerate development cycles and integrate advanced safety technologies like advanced driver-assistance aystems (ADAS). In the U.S. alone, ADAS is estimated to prevent up to 62% of total traffic deaths annually.

ADAS technologies rely on various sensing systems, such as cameras, LiDAR, radar, and ultrasonic sensors, combined with AI-driven algorithms to help detect potential hazards, assist drivers, and even take corrective action autonomously. These systems are essential for creating safer roads and reducing human error behind the wheel.

THE CHALLENGE

Despite the proven benefits of ADAS, OEMs and Tier 1 suppliers face several barriers to seamless integration. ADAS platforms must be compatible with a wide range of electronic control units (ECUs) and configurations from legacy vehicle development, requiring scalable solutions that work across both high-end and mass-market vehicles. Moreover, integrating ADAS into existing vehicle architectures demands complex software coordination capable of processing vast sensor inputs in real time.

In addition to hardware and software complexities, OEMs must comply with different global standards and regulations for ADAS implementation. Each region imposes varying requirements for compliance, making it challenging to ensure consistent safety and performance worldwide.

THE SOLUTION

To simplify ADAS adoption and improve efficiency, LeddarTech and Arm collaborated to optimize LeddarVision™, an AI-driven sensor fusion and perception software stack. This solution enables vehicles to create an accurate 3D environmental model using advanced AI and computer vision algorithms.

LeddarVision is built on Arm-based automotive platforms, which offer a standardized ADAS software framework that can be adapted across multiple platforms without extensive customization. This unified architecture integrates Arm advanced CPUs and accelerators to deliver optimized performance, while minimizing computational bottlenecks to result in improved system efficiency.

One of the key performance-enhancing features of LeddarVision is the use of Arm Cortex-A720AE CPUs, which provide a significant boost in

processing power for critical sensor fusion algorithms. This improvement reduces perception latency, enhancing the responsiveness of ADAS features and overall vehicle safety.

By leveraging the Arm pre-silicon port in the Armv9 architecture, LeddarTech is able to prototype and validate software enhancements before hardware deployment. This approach helps accelerate development cycles and ensure compatibility with future hardware advances, providing OEMs with an accelerated path to market.

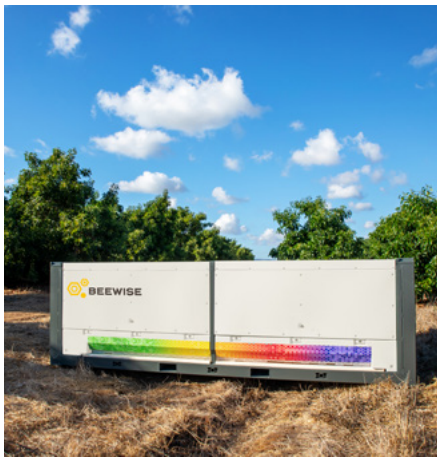
THE RESULT

LeddarVision, running on Arm-powered next-generation ECUs, delivers a range of benefits to OEMs and Tier 1 suppliers. The optimized software stack helps reduce CPU utilization, minimizing the need for high-power systems-on-chip (SoCs) while ensuring reliable ADAS performance. This improved computational efficiency leads to accelerated reaction times and reduced end-to-end latency, resulting in safer, more responsive ADAS functionality.

The standardized ADAS architecture simplifies the integration process across different vehicle models, helping to reduce complexity and lower development costs. In addition, LeddarVision undergoes rigorous real-world testing to ensure that its AI-driven perception models perform reliably in dynamic driving conditions.

By optimizing LeddarVision with Arm-based platforms, LeddarTech and Arm are setting a new standard for intelligent mobility. This collaboration helps to enable accelerated ADAS adoption, improved system efficiency, and enhanced vehicle safety, shaping the future of advanced automotive technology.

The First Autonomous Beehive: How Beewise is Revolutionizing Beekeeping with AI



INTRODUCTION

Bee populations are declining at an alarming rate, with 35% of bees dying annually due to climate change, pesticide exposure, and disease. Since bees pollinate one-third of the global food supply, their decline poses a major risk to agriculture and biodiversity.

Traditional beekeeping relies on manual inspections and wooden hives, making it slow, labor-intensive, and ineffective against modern threats. Beewise has developed Beehomes, the world's first AI-powered, robotic beehive that automates hive management and helps reduce bee mortality rates by 85%.

THE CHALLENGE

For centuries, beekeepers have used the same manual methods to manage their hives. They physically inspect colonies to check for diseases, pests, and environmental damage, but these visits happen only every few weeks. By the time an issue is detected, a colony may already be collapsing.

One of the biggest threats to honeybees is the Varroa mite, a deadly parasite that spreads viruses and weakens entire hives. Without early detection and treatment, an infestation can wipe out a colony within

weeks. Additionally, climate change has led to more extreme weather conditions, including rising temperatures and sudden cold snaps, further destabilizing bee populations.

Beekeepers also face economic and labor challenges. The manual nature of beekeeping is time-consuming, requiring significant resources and expertise. Honey harvesting disrupts bee colonies, affecting productivity and overall hive stability. A scalable, automated solution is needed to protect bees, reduce human intervention, and optimize hive management.

THE SOLUTION

Beewise's Beehomes are AI-driven, solar-powered hives that monitor, regulate, and protect bee colonies 24/7. Unlike traditional wooden hives, Beehomes use robotics, machine learning, and computer vision to provide real-time hive insights and automate critical processes.

These intelligent hives track temperature, humidity, and colony activity, automatically adjusting conditions for hive stability. The system detects early signs of disease and removes Varroa mites before infestations escalate. By using precision robotics, Beehomes automate honey harvesting, allowing for minimal disruption to the bees while maximizing production.

At the core of Beehomes is an Arm-based CPU, which enables real-time automation and AI-driven decision-making. NVIDIA Jetson processors analyze hive data, while Raspberry Pi modules provide cloud connectivity, allowing beekeepers to monitor their hives remotely. Solar-powered operation helps ensure that the system remains completely self-sufficient, reducing reliance on external energy sources.

Beehomes replace guesswork with data-driven precision, giving

beekeepers an advanced tool to manage their colonies effectively, while requiring far less manual intervention.

THE RESULT

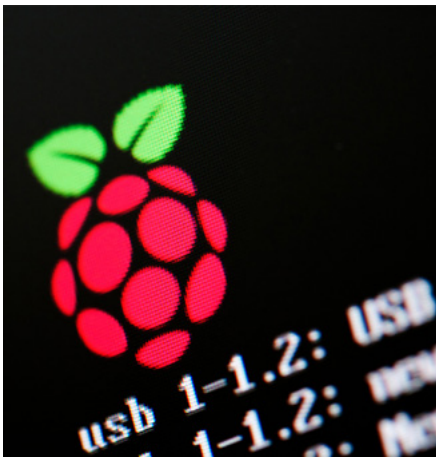
Since Beehomes were introduced, beekeepers have seen an 85% reduction in bee mortality rates, a game-changing improvement for colony survival. Healthier bees have led to higher honey yields, increasing productivity while maintaining hive stability.

By eliminating unnecessary manual interventions, Beewise has cut labor costs and made beekeeping more scalable and efficient. The ability to detect and address threats in real time has prevented colony collapses that would have been unavoidable under traditional methods.

Beehomes are also proving to be a sustainable solution for large-scale beekeeping operations. With their solar-powered design and energy-efficient architecture, they align with eco-friendly practices, while helping to ensure long-term hive stability. Beekeepers can now manage multiple hives remotely, freeing up time to focus on expansion rather than crisis management.

By merging AI, robotics, and automation, Beewise is helping to secure the future of beekeeping and global food production. As pollinators continue to face threats, Beewise innovation helps ensure that bees can thrive. Saving bees means saving ecosystems, agriculture, and the planet.

Accessible Computing Platform For Everyone: The Evolution of Raspberry Pi



INTRODUCTION

Fifteen years ago, the Raspberry Pi Foundation was established with a mission to democratize computing for young people. Founded by Eben Upton, the initiative aimed to create an affordable and accessible computing platform to empower students and makers worldwide. What began as an educational project soon evolved into a powerful computing tool, extending far beyond classrooms into industrial and embedded applications.

Today, Raspberry Pi is a critical platform in the industrial and embedded electronics sector, serving a diverse range of users across industries.

THE CHALLENGE

In the past, the computing industry struggled with accessibility and affordability. Traditional hardware solutions were expensive, proprietary, and primarily designed for large-scale enterprises, leaving students, hobbyists, and small businesses without viable options. The absence of an affordable, flexible computing platform stifled innovation, restricting opportunities for learning, experimentation, and custom development. At the same time, industries seeking embedded computing solutions needed compact, power-efficient, and scalable alternatives to costly proprietary systems. The market lacked a truly accessible and adaptable computing solution

that could serve as a versatile computing platform for education, hobbyists, and industry alike.

THE SOLUTION

Raspberry Pi, a compact, flexible, and cost-effective computing platform, emerged as the solution to this accessibility gap. Originally developed to support education, it quickly found adoption in industrial and embedded applications. Today, companies worldwide integrate Raspberry Pi into a variety of applications, including factory automation, smart manufacturing, medical devices, IoT systems, and autonomous monitoring solutions, leveraging its compact form factor, power efficiency, and adaptability to diverse computing needs.

A major enabler of Raspberry Pi's success is its partnership with Arm, which has played a massive role in delivering performance, energy efficiency, and scalability. From the Arm11 family of processors in the original Raspberry Pi to the Arm Cortex-A76, a high-performance processor optimized for modern computing tasks in Raspberry Pi 5, this collaboration has enabled Raspberry Pi to continuously advance in power and capability. The integration of Arm technology has ensured continuous innovation, making Raspberry Pi an essential platform for educators, developers, and businesses alike.

"Arm's technology allows us to offer cutting-edge AI capabilities in a way that's accessible to everyone." –Eben Upton, CEO, Raspberry Pi Foundation

Beyond education, companies worldwide integrate Raspberry Pi hardware into embedded products, industrial systems, and IoT solutions, leveraging its power, flexibility, and affordability. This synergy between Raspberry Pi and Arm continues to push the boundaries of accessible computing.

THE RESULT

Raspberry Pi has completely revolutionized computing accessibility, becoming a key player in industrial automation, AI development, and real-world problem-solving. Developers and engineers worldwide rely on Raspberry Pi to build autonomous systems, AI-driven analytics, edge computing frameworks, and connected devices. Its affordability has empowered researchers, small businesses, and students to experiment with robotics, environmental monitoring, and smart home solutions, proving that cost-effective computing can be powerful and scalable.

The partnership with Arm continues to drive Raspberry Pi's innovation, enabling it to deliver high-performance, low-power computing across industries. From supporting machine learning and AI deployments to fueling next-generation industrial control systems, Raspberry Pi remains a pioneer in advancing digital transformation. With its commitment to providing flexible, powerful computing platforms, the Raspberry Pi Foundation is shaping the future of embedded computing, helping to ensure technology remains accessible to all.

SpaceTech Smart City Infrastructure Powered by Arm's Edge AI

INTRODUCTION

The future of urban living depends on intelligent, data-driven solutions that help enhance safety, sustainability, and efficiency in real time.

SpaceTech, a subsidiary of China Vanke Co., Ltd., is a leading smart city solutions provider that manages more than 8 million residential units and 2,000 commercial buildings, integrating technologies to optimize urban living spaces. To streamline operations and improve the resident experience, SpaceTech has integrated Arm-based computing solutions into its infrastructure, enabling real-time decision making, automated building management, and energy efficiency enhancements.

THE CHALLENGE

With a vast portfolio that spans over 1 billion square meters, SpaceTech faced several key challenges in managing smart urban environments.

One of the major concerns was waste management, where inefficient manual oversight led to delays in service. AI-driven categorization and automated service requests were required to streamline operations. Another challenge lay in legacy building systems, where disparate devices, such as lighting, HVAC, and access control, operated in isolation. The lack

of interoperability between devices from different manufacturers created inefficiencies, making centralized management difficult.

Additionally, the growing volume of data traffic from smart devices demanded a high-performance, power-efficient solution to process AI workloads at the edge, while keeping energy consumption low.

THE SOLUTION

To address these challenges, SpaceTech partnered with Ampere Computing to transition from x86 edge servers to Arm-based Ampere Altra servers. This move provided significant advantages, including 2.5 times improved performance per rack, 2.8 times lower power consumption, and a 3 times smaller footprint, aligning with SpaceTech's sustainability goals.

A proof of concept demonstrated the efficiency of the Alibaba Qwen-VL Vision Language Model (VLM) on Ampere Arm-based servers, showing significant energy savings without compromising AI capabilities. Ampere also provided dedicated AI libraries with optimized quantization methods, improving performance by up to 2 times, while maintaining model accuracy.

In smart waste management, AI-driven categorization now automatically detects full bins and generates service tickets.

In building automation, real-time AI analytics integrate various systems, allowing lighting, HVAC, and security to function as a unified network. Motion sensors detect room occupancy, adjusting energy use dynamically, while surveillance AI enhances safety by monitoring elevators, preventing unauthorized scooter use, and managing cleanliness in public areas.

THE RESULT

The transition to Arm-powered edge AI servers has enabled SpaceTech to create a scalable, high-efficiency smart city infrastructure. The implementation of intelligent automation has led to a significant reduction in energy consumption, while improving security and operational efficiency.

With Arm-based edge computing, SpaceTech has streamlined multiple building systems, running various applications in virtual containers on edge server clusters. This unified data-lake approach has eliminated inefficiencies, allowing seamless integration across lighting, HVAC, air quality, and access control systems.

Additionally, the edge cloud-native platform developed by SpaceTech helps ensure that every device—from cameras to sensors to gateways—functions as a micro-cloud, enabling real-time AI decision making, while reducing power and cooling demands. The adoption of Arm-powered solutions positions SpaceTech to lead the smart city revolution, setting new benchmarks for intelligent urban management.

By embracing Arm technology and Ampere Arm-based edge servers, SpaceTech is redefining urban property management with scalability, AI-powered efficiency, and sustainability, offering a model for the future of connected smart cities worldwide.

Appendix

References

01 AI Risk Management Framework. NIST [Internet]. 2021 Jul 12 [cited 2025 Feb 20]; Available from: https://www.nist.gov/itl/ai-risk-management-framework

02 AI Verify Foundation [Internet]. [cited 2025 Feb 20]. What is AI Verify. Available from: https://aiverifyfoundation.sg/what-is-ai-verify/

03 AI Watch: Global regulatory tracker – Brazil | White & Case LLP [Internet]. 2024 [cited 2025 Feb 20]. Available from: https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-brazil

04 Artificial Intelligence and Data Act (AIDA) – Companion document [Internet]. Innovation, Science and Economic Development Canada; 2025 [cited 2025 Feb 20]. Available from: https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document

05 Arm. A leap forward in auto safety with LeddarTech and Arm [Internet]. Available from: https://armkeil.blob.core.windows.net/developer/Files/pdf/case-study/a-leap-forward-in-auto-safety-with-leddartech-and-arm.pdf

06 Arm. First autonomous beehive: Beewise uses AI to protect pollinators [Internet]. Available from: https://www.arm.com/company/success-library/made-possible/first-autonomous-beehive
Arm. Raspberry Pi: From education tool to AI-enabled edge computing [Internet]. Available from: https://www.arm.com/company/success-library/made-possible/raspberry-pi

07 Arm Edge AI technology powers SpaceTech's next-generation smart cities. Arm Blog [Internet]. 2025 Jan 17. Available from: https://newsroom.arm.com/blog/arm-edge-ai-powers-spacetech-smart-cities

08 Arm Newsroom Blog. AWS Graviton Decarbonize Compute. Arm Blog [Internet]. 2023 November 21. Available from: https://newsroom.arm.com/blog/aws-graviton-decarbonize-compute

09 Arab News/AFP. The software behind the self-driving Uber crash did not recognize jaywalkers [Internet]. 2019. Available from: https://www.arabnews.com/node/1579826/science-technology

10 BBC News. Google partners with Kairos Power to deploy small nuclear reactors for AI data centers [Internet]. 2024 Sep. Available from: https://www.bbc.com/news/articles/c748gn94k95o

11 Bellamy RKE, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, et al. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM J Res Dev. 2019;63(4/5):4:1–4:15. doi: 10.1147/JRD.2019.2942287

12 Biggio B, Corona I, Nelson B, Rubinstein BIP, Maiorca D, Fumera G, et al. Evasion attacks against machine learning at test time. Machine Learning and Knowledge Discovery in Databases. 2013;8190:387–402.

13 Biggio B, Nelson B, Laskov P. Poisoning attacks against support vector machines. In: Proceedings of the 29th International Conference on Machine Learning (ICML). 2012;1807–14.

14 Bradford A. The Brussels Effect: How the European Union Rules the World. Oxford University Press; 2020.

15 Brookings Institution. Explainability won't save AI [Internet]. 2020. Available from: https://www.brookings.edu/articles/explainability-wont-save-ai

16 Brookings Institution. Global networked AI governance [Internet]. (No date). Available from: https://www.brookings.edu/articles/network-architecture-for-global-ai-policy/

17 California Consumer Privacy Act (CCPA). California Civ. Code § 1798.100 et seq.

18 Campbell G. Arm and industry leaders launch Semiconductor Education Alliance to address the skills shortage [Internet]. Arm Newsroom; 2023 Aug 1. Available from: https://newsroom.arm.com/news/semiconductor-education-alliance

19 China. Data Security Law. 2021.

20 China. Ethical Guidelines for AI. 2022.

21 China. Personal Information Protection Law. 2021.

22 China Privacy Law | Office of Ethics, Risk and Compliance Services [Internet]. Available from: https://oercs.berkeley.edu/privacy/international-privacy-laws/china-privacy-law

23 China State Council. New Generation Artificial Intelligence Development Plan. 2017.
'New Generation Artificial Intelligence Development Plan' (2017)" in List 1)

24 CIODive. California governor vetoed SB 1047. What's next for AI regulation? [Internet]. 2024 Sept. 30. Available from: https://www.ciodive.com/news/california-senate-bill-SB1047-AI-regulation-landscape/728481

25 Clark L. GenAI's dirty secret: It's set to create a mountainous increase in e-waste. The Register. 2024 Oct 28. Available from: https://www.theregister.com/2024/10/28/genai_dirty_secret/

26 CNN Business. Microsoft to restart reactor at Three Mile Island for AI power by 2028 [Internet]. 2024 Sep 20. Available from: https://edition.cnn.com/2024/09/20/energy/three-mile-island-microsoft-ai/index.html

27 Cornell University (INFO 2040 course blog). The 2010 Flash Crash: information cascades [Internet]. 2020. Available from: https://blogs.cornell.edu/info2040/2020/11/13/the-2010-flash-crash-how-information-cascades-shape-our-world/

28 Cyberspace Administration of China. Internet Information Service Algorithmic Regulation Provisions. 2022.

29 Dastin J, Nellis S. Focus: For tech giants, AI like Bing and Bard poses billion-dollar search problem [Internet]. Reuters; 2023 Feb 23. Available from: https://www.reuters.com/technology/tech-giants-ai-like-bing-bard-poses-billion-dollar-search-problem-2023-02-22

30 Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT). 2019;4171–4186.

31 Deloitte. Powering artificial intelligence: A study of AI's environmental footprint—today and tomorrow. November 2024. Available from: https://www.deloitte.com/global/en/issues/climate/powering-ai.html

32 Dung L. Current Cases of AI Misalignment and Their Implications for Future Risks. Synthese. 2023;202:138. doi:10.1007/s11229-023-04367-0

33 European Commission. EU AI Act – Regulatory framework [Internet]. 2024. Available from: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

34 European Commission. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance) [Internet]. Jun 13, 2024.

35 European Union. AI Act. Official Journal of the European Union; August 2024.

36 Falade PV. Decoding the threat landscape: ChatGPT, FraudGPT, and WormGPT in social engineering attacks. Int J Sci Res Comput Sci Eng Inf Technol. 2023;8(5):185–200. doi: 10.32628/CSEIT2390533

37 Garcia-Tobin C, Knight M. Elevating security with Arm CCA. Commun ACM [Internet]. 2024 Oct;67(10):34–39. Available from: https://cacm.acm.org/practice/elevating-security-with-arm-cca

38 Gallup. Strategy Will Fail Without a Culture That Supports It [Internet]. 2024. Available from: https://www.gallup.com/workplace/652727/strategy-fail-without-culture-supports.aspx

39 Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative Adversarial Nets. In: Advances in Neural Information Processing Systems (NeurIPS). 2014;27:2672–2680.

40 Government of Canada. Bill C-27: The Artificial Intelligence and Data Act. 2022.

41 Hadshar R. A Review of the Evidence for Existential Risk from AI via Misaligned Power-Seeking [Preprint]. arXiv; 2023 Oct 27. Available from: https://arxiv.org/abs/2310.18244

42 He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV; 2016. pp. 770–778. doi: 10.1109/CVPR.2016.90

43 Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Comput. 1997;9(8):1735–1780. doi: 10.1162/neco.1997.9.8.1735

44 ICO (UK). Google DeepMind and NHS patient data [Internet]. ICO. Available from: https://ico.org.uk/for-the-public/ico-40/google-deepmind-and-class-action-lawsuit/

45 Illinois Biometric Information Privacy Act (BIPA). 740 Ill. Comp. Stat. 14/1 et seq.

46 Infocomm Media Development Authority (Singapore). Model AI Governance Framework. 2019; updated 2020.

47 International Energy Agency. Data centres and data transmission networks [Internet]. Available from: https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks

48 ISO & IEC. ISO/IEC 42001: AI Management Systems – Requirements. 2023.

49 K8sGPT. AI-Powered resource and cluster orchestration [Internet]. 2024. Available from: https://k8sgpt.ai/

50 Kabashkin I. End-to-end service availability in heterogeneous multi-tier cloud-fog-edge networks. Future Internet. 2023;15:329.

51 Kilian KA. Beyond Accidents and Misuse: Decoding the Structural Risk Dynamics of Artificial Intelligence [Preprint]. arXiv; 2024 Jun 21. Available from: https://arxiv.org/abs/2406.14873

52 KPMG. ISO/IEC 42001 AI Management Systems [Internet]. 2023. Available from: https://kpmg.com/ch/en/insights/artificial-intelligence/iso-iec-42001.html

53 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS'12). Curran Associates Inc.; 2012. pp. 1097–1105.

54 Laux J, Wachter S, Mittelstadt B. Trustworthy artificial intelligence and the European Union AI Act: On the conflation of trustworthiness and acceptability of risk. Regul Gov. 2024;18(1):3–32. doi: 10.1111/rego.12512.

55 Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems. 2017;30:4765–4774.

56 Mayer Brown (Bloomberg Law). Conducting an AI Risk Assessment [Internet]. 2024. Available from: https://www.mayerbrown.com/en/insights/publications/2024/01/conducting-an-ai-risk-assessment

57 McMahan HB, Moore E, Ramage D, Hampson S, Arcas BA. Communication-efficient learning of deep networks from decentralized data. arXiv [Preprint]. 2017 Feb 17. Available from: https://arxiv.org/abs/1602.05629

58 Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Comput Surv. 2021;54(6):Article 115. doi: 10.1145/3457607

59 Microsoft. Tay chatbot incident [Internet]. 2016. Available from: https://en.wikipedia.org/wiki/Tay_(chatbot)

60 Millière R. The Alignment Problem in Context [Preprint]. arXiv; 2023 Nov 3. Available from: https://arxiv.org/abs/2311.02147

61 Nemko. Ensuring AI Safety and Robustness [Internet]. (No date). Available from: https://www.nemko.com/blog/ai-safety-and-robustness

62 New America OTI. Global cooperation – International Network of AI Safety Institutes [Internet]. 2024. Available from: https://www.newamerica.org/oti/blog/to-achieve-global-ai-cooperation-we-need-multidisciplinary-dialogues-in-addition-to-the-international-network-of-ai-safety-institutes

63 Personal Data Protection Commission (Singapore). AI Verify Toolkit. 2022.

64 Radford A, Narasimhan K, Salimans T, Sutskever I. Improving Language Understanding by Generative Pre-Training. OpenAI; 2018.

65 Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. J Mach Learn Res. 2020;21(140):1–67.

66 Raffin E, Hamidouche W, Nogues E, Pelcat M, Menard D, Tomperi S. Energy efficiency of a parallel HEVC software decoder for embedded devices. In: Proceedings of the 12th ACM International Conference on Computing Frontiers (CF '15). New York: Association for Computing Machinery; 2015. pp. 1–6. doi: 10.1145/2742854.2747286

67 Reuters. Amazon scrapes secret AI recruiting tools that show bias against women [Internet]. 2018. Available from: https://www.ml.cmu.edu/news/news-archive/2016-2020/2018/october/amazon-scraps-secret-artificial-intelligence-recruiting-engine-that-showed-biases-against-women.html

68 Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining. 2016:1135–1144.

69 Silver D, Schrittwieser K, Simonyan K, Antonoglou I, Huang A, Guez A, et al. Mastering the game of Go without human knowledge. Nature. 2017;550(7676):354–359. doi: 10.1038/nature24270

70 Singh RK, Mishra S. TinyML meets IoBT against sensor hacking. In: Proceedings of the 2024 Workshop on Security and Privacy in Standardized IoT. 2024. doi: 10.14722/sdiotsec.2024.23010

71 Sofia RC, et al. A framework for cognitive, decentralized container orchestration. IEEE Access. 2024;12:79978–80008. doi: 10.1109/ACCESS.2024.3406861

72 Securiti.ai. Overview of Australia's AI Safety Standard [Internet]. 2023. Available from: https://securiti.ai/australia-voluntary-ai-safety-standard/

73 Srivatsa M, Abdelzaher T, He T. Artificial intelligence for edge computing. Springer; 2023. 365 p. Available from: https://doi.org/10.1007/978-3-031-40787-1

74 TechMonitor. Frontier Model Forum launched for safe AI [Internet]. 2023. (OpenAI, Google, Microsoft, Anthropic partnership).

75 TIME. Exclusive: New Research Shows AI Strategically Lying [Internet]. 2024 Dec 18. Available from: https://time.com/7202784/ai-research-strategic-lying

76 Time.com. AI's growing demand for data-center electricity [Internet]. 2024. Available from: https://time.com/6987773/ai-data-centers-energy-usage-climate-change/

77 UNFCCC. Paris Agreement. Bonn, Germany: UNFCCC; 2015. Available from: https://unfccc.int/sites/default/files/english_paris_agreement.pdf

78 University of Cambridge. Arm donates £3.5 million for Cambridge PhD students to study computer architecture and semiconductor design [Internet]. 2024. Available from: https://www.cam.ac.uk/news/arm-donates-ps3-5-million-for-cambridge-phd-students-to-study-computer-architecture-and

79 Voigt P, Von dem Bussche A. The EU General Data Protection Regulation (GDPR): A Practical Guide. Springer International Publishing; 2017.

80 Wikipedia. AI Alignment [Internet]. (No date). Available from: https://en.wikipedia.org/wiki/AI_alignment

81 World Economic Forum. Why upskilling your existing workforce can be far more cost-effective than hiring new employees [Internet]. 2024. Available from: https://www.weforum.org/stories/2024/01/ai-training-workforce/

82 Zhou I, Tofigh F, Piccardi M, Abolhasan M, Franklin DR, Lipman J. Secure multi-party computation for machine learning: A survey. IEEE Access. 2024;12:53881–53899.

83 Zhu J-Y, Park T, Isola P, Efros AA. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy; 2017. pp. 2242–2251. doi: 10.1109/ICCV.2017.244

84 Translation: Internet Information Service Algorithmic Recommendation Management Provisions – Effective March 1, 2022 [Internet]. DigiChina. [cited 2025 Feb 20]. Available from: https://digichina.stanford.edu/work/translation-internet-information-service-algorithmic-recommendation-management-provisions-effective-march-1-2022

85 Translate CL. Data Security Law of the PRC [Internet]. China Law Translate. 2021 [cited 2025 Feb 20]. Available from: https://www.chinalawtranslate.com/datasecuritylaw/

86 Position Paper of the People's Republic of China on Strengthening Ethical Governance of Artificial Intelligence (AI) [Internet]. [cited 2025 Feb 20]. Available from: https://www.fmprc.gov.cn/eng/zy/wjzc/202405/t20240531_11367525.html

87 Model AI Governance Framework 2024 - Press Release | IMDA [Internet]. [cited 2025 Feb 20]. Available from: https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2024/public-consult-model-ai-governance-framework-genai

Survey methodology

METHOD

Arm conducted this research using an online survey prepared by Method Research and distributed by RepData and iTracks among n=655 adults (age 18+) who are business leaders across various industries. All respondents are from companies with more than 1,000 employees and influence purchase decision-making within their organizations. The sample was from the following countries: U.S., U.K., Germany, France, Hungary, China, Taiwan and Japan. Data was collected from January 16 to February 5, 2025.

Research led by Svetlana Gershman, Vice President and the head of Method Communications' specialized marketing research division, and Fatima Aslam, Research Analyst at Method Communications.

SAMPLE DEMOGRAPHIC

COUNTRY

US - 15%

UK - 15%

China - 15%

France - 8%

Germany - 8%

Japan - 16%

Taiwan - 15%

AGE

Gen Z (18–27) - 13%

Millennial (28-43) - 43 %

Gen X (44-59) - 35%

Boomer (60-78) - 10%

GENDER

Man - 60%

Woman - 40%

DEPARTMENT

Finance -13%

Marketing/Sales - 14%

Accounting- 10%

Operations- 12%

Human Resources- 14%

Information Technology- 24%

Legal services - 13%

FUNCTIONAL ROLE

C-level executive- 17%

Vice President - 13%

Director - 36%

Manager - 34%

COMPANY SIZE

SMB - 44%

Mid Market - 28%

Enterprise - 28%

COMPANY ANNUAL REVENUE (US)

Less than $1 Million - 9%

$1 million - under $40 Million - 20%

$40 million - $1 billion - 37%

More than $1 billion - 24%

Prefer not to say - 12%

INDUSTRY

Financial Services - 18%

Manufacturing - 19%

Healthcare - 4%

Retail - 9%

Technology - 17%

Energy - 4%

Other - 29%

About the Contributors

EXTERNAL AUTHORS

Deval Shah

Senior Machine Learning Engineer at the Australian Institute for Machine Learning (AIML)

Deval Shah is a Senior Machine Learning Engineer at the Australian Institute for Machine Learning (AIML), where he specializes in designing Retrieval Augmented Generation (RAG) systems for enterprise search and policymaking applications. His focus on scalability and secure data handling underscores his expertise in production-ready AI solutions. Prior to AIML, Deval advanced from Machine Learning Engineer to Senior Software Engineer at Uncanny Vision, spearheading projects that enhanced license plate recognition models and vehicle re-identification microservices. Deval has collaborated with over 15 AI companies, published more than 100 articles, and successfully built substantial organic inbound channels for AI startups through his content-driven initiatives. His skill set spans full-stack development (Next. js, FastAPI), containerization (Docker), and cloud infrastructure (AWS, Azure DevOps).

Dr John Soldatos

Honorary Research Fellow at the University of Glasgow

John Soldatos holds a Ph.D. in Electrical & Computer Engineering from the National Technical University of Athens (2000) and is currently an Honorary Research Fellow at the University of Glasgow, UK (2014-present). He was

Associate Professor and Head of the Internet of Things (IoT) Group at the Athens Information Technology (AIT), Greece (2006–2019), and Adjunct Professor at the Carnegie Mellon University, Pittsburgh, PA (2007–2010). He has significant experience working closely with large multi-national industries (e.g., IBM, INTRACOM, INTRASOFT International) as an R&D consultant and delivery specialist while being a scientific advisor to various high-tech startup enterprises. Dr. Soldatos is an expert in Internet-of-Things (IoT) and Artificial Intelligence (AI) technologies and applications, including IoT/AI applications in smart cities, finance (Finance 4.0), and industry (Industry 4.0).

Mark Hinkle

CEO and Co-Founder Peripety Labs

Mark Hinkle has over 30 years of experience in technology, open source, and enterprise software. He served in executive roles at Cloud.com, Citrix and The Linux Foundation, spearheading innovative programs in cloud computing, DevOps, and emerging technology adoption. Over the past two years, he has trained thousands of professionals on how to use AI for real-world impact—ranging from beginners to C-suite executives. Mark publishes TheAIE.net, a newsletter network with more than 250,000 subscribers dedicated to working smarter with AI. He is also the co-founder of All Things AI, an organization focused on advancing enterprise AI through events, community, and industry collaboration.

Dr Nicole Höher

Project Manager for Sustainability & Digitalisation at juS.TECH GmbH

Dr Höher is a natural scientist with many years of experience in data analysis and visualisation. She specialises in the application of data science and artificial intelligence, and has experience in the development of AI

software solutions. Her expertise combines technical know-how with an understanding of environmental, economic and social contexts to develop innovative and sustainable solutions.

Dr Nora von Ingersleben Seip

Postdoctoral Researcher at the Political Science Department of the University of Amsterdam, RegulAite project

Nora von Ingersleben-Seip is a Postdoctoral Researcher at the University of Amsterdam, where she investigates international cooperation on AI governance. Previously, she was the Managing Director of a venture capital-funded tech startup in Southeast Asia and held other executive roles in the region's startup ecosystem. She also led political advocacy efforts for major global tech firms in North America and Europe. Her research on AI governance has been published in leading peer-reviewed journals and cited by The Wall Street Journal and Politico, among others. She is a member of several expert groups, including the AI Adoption Initiative and the Forum for Cooperation on Artificial Intelligence at the Brookings Institution and the Center for European Policy Studies. Nora holds a PhD (with Distinction) from the Technical University of Munich and an MBA from INSEAD.

Dr Vanessa Just

Founder and CEO of juS.TECH GmbH

Dr Just is an entrepreneur and expert in business, industrial engineering and artificial intelligence. As a leading voice on sustainability, digitalisation and AI, she speaks on numerous panels and events and is a lecturer at the FOM University of Applied Sciences. Dr Just is also an author and editor for renowned publishers such as Springer: "Digitalisation and Sustainability" and "Sustainability through Innovation, Digitalisation and Technologies". Dr Just is also a board member of the German AI Association.

Method Communications

Research for the AI Readiness Index survey and narrative for Chapter 1 led by Svetlana Gershman, Vice President and the head of Method Communications' specialized marketing research division, and Fatima Aslam, Research Analyst at Method Communications.

ARM AUTHORS

Kevork Kechichian

Executive Vice President, Solutions Engineering

Kevork is leading the team that works closely with Arm's world-class partner ecosystem to develop and accelerate our technologies as we keep ahead of changing market needs.

His career spans 30 years in various semiconductor engineering and leadership roles. Before joining Arm, he was the executive vice president of MCU/MPU Engineering at NXP Semiconductors, where he led the company's global architecture and IP & SoC design. Previously, he was the senior vice president of engineering at Qualcomm where he managed the Snapdragon SoC and Technology teams and was responsible for delivering more than 100 SoCs in leading-edge technologies, enabling breakthrough products in mobile and consumer markets. Kevork holds a BEng degree in electrical engineering from American University of Beirut and a M.S. in electrical engineering from Concordia University.

Khaled Benkrid

Senior Director, Education and Research

With a background that spans both academia and industry, Benkrid has been instrumental in addressing the skills gap in the semiconductor

industry. In his current position, he leads Arm's Education and Academic Engagements Team, leveraging his experience to bridge the divide between academic education and industry needs. He also holds the position of Visiting Professor in Computing at Anglia Ruskin University (ARU) in Cambridge, England.

Benkrid has been actively involved in several key initiatives, including playing a formative role in the creation of the Semiconductor Education Alliance and working with organizations such as the Semiconductor Research Corporation (SRC) in the U.S., the Taiwan Semiconductor Research Institute (TSRI), and the All India Council for Technical Education (AICTE) to develop necessary semiconductor workforce, especially in light of recent initiatives like the US CHIPS Act.

Benkrid is also a prolific writer and thought leader in the field of education and technology. He has authored articles on topics such as "The Evolving Nature of Work" and "Talent Multiplication: The Holy Grail of Modern Organizations," discussing the impact of AI on the job market and the importance of continuous learning. Benkrid holds a Ph.D. and an MBA and is a Chartered Engineer (CEng).

Maureen McDonagh

Head of Sustainability

Maureen is passionate about driving Arm's sustainability journey in this transformational era of AI, balancing the opportunities and challenges for Arm to contribute to a sustainable, equitable and resilient future. Energized by Arm's incredible talent of global problem solvers and the role we can all play at the intersection of technology, innovation, people, and planet. Connecting ideas, people, resources to catalyse the new thinking we need to solve our environmental and social challenges to shape a more sustainable and inclusive world.

Vince Jesaitis

Senior Director, Government Affairs

Vince is a seasoned government affairs professional with over 20 years of experience in the high-tech sector. He currently serves as the Senior Director of Government Affairs at Arm, the world's leading semiconductor and software design company.

Vince joined Arm in October 2021 and is responsible for developing and executing advocacy campaigns, building relationships with key policymakers and influencers, and representing the company's vision and values in various forums.

Prior to his role at Arm, Vince held senior positions at the Information Technology Industry Council and in the U.S. Congress, where he worked on several legislative initiatives benefiting the high-tech sector. Vince holds a Bachelor of Arts degree in International Relations, Philosophy, and Political Science from William Jewell College, which he earned between 1997 and 2001.

Will Abbey

Executive Vice President and Chief Commercial Officer

Will leads sales, field engineering, and partner enablement at Arm, helping innovative organizations harness the latest technologies to prepare for the next wave of AI-driven digital transformation. From IP to AI, his insight has guided technology leaders in transforming their products and operations.

Since joining Arm in 2004, Will has held several leadership roles, including General Manager of the Physical Design group. Will has held product management roles at Celoxica, Infineon Technologies and Loughborough Sound Images, and serves on the board of EnPro Industries.

Additional Arm Contributors

Arm Content

Brian Fuller

Jack Melling

Omkar Patwardhan

Kurt Wilson

Arm Branding

Julie Jervis

Sarah Nguyen

Arm Creative

Chris Salvador

Hoa Tong

Arm Sustainability Team

Mohammed Chunara

Fiona Maudslay

Fiona Riggall

Wevolver Editorial team

Editor: Rebecka Durén

Operations: Jessica Miley

We would like to extend our sincere gratitude for the dedication and hard work of everyone who contributed to the creation of this report. Your expertise, collaboration, and commitment made this possible—thank you!

About Arm

Arm is the industry's highest-performing and most power-efficient compute platform with unmatched scale that touches 100 percent of the connected global population. To meet the insatiable demand for compute, Arm delivers advanced solutions that allow the world's leading technology companies to unleash the unprecedented experiences and capabilities of AI. Together with the world's largest computing ecosystem and 20 million software developers, we are building the future of AI on Arm.

www.arm.com

arm

About Wevolver

Wevolver is a global platform and community that provides engineers with the knowledge and connections to develop better technology.

We bring a professional audience of engineers informative and inspiring content, such as articles, videos, podcasts, and reports, about state-of-the-art technologies.

The knowledge on Wevolver is published by various sources: universities, tech companies, and individual community members. Next to that, we manage a network of over 50 technical writers who create content for our customers and publish that on Wevolver.com

Millions of engineers leverage Wevolver to stay up to date, find knowledge when they are developing products, and leverage the platform to make meaningful connections.

Wevolver has won the SXSW Innovation Award, the Accenture Innovation Award, and the Top Most Innovative Web Platforms by Fast Company. Wevolver is how today's engineers stay cutting edge.

www.wevolver.com