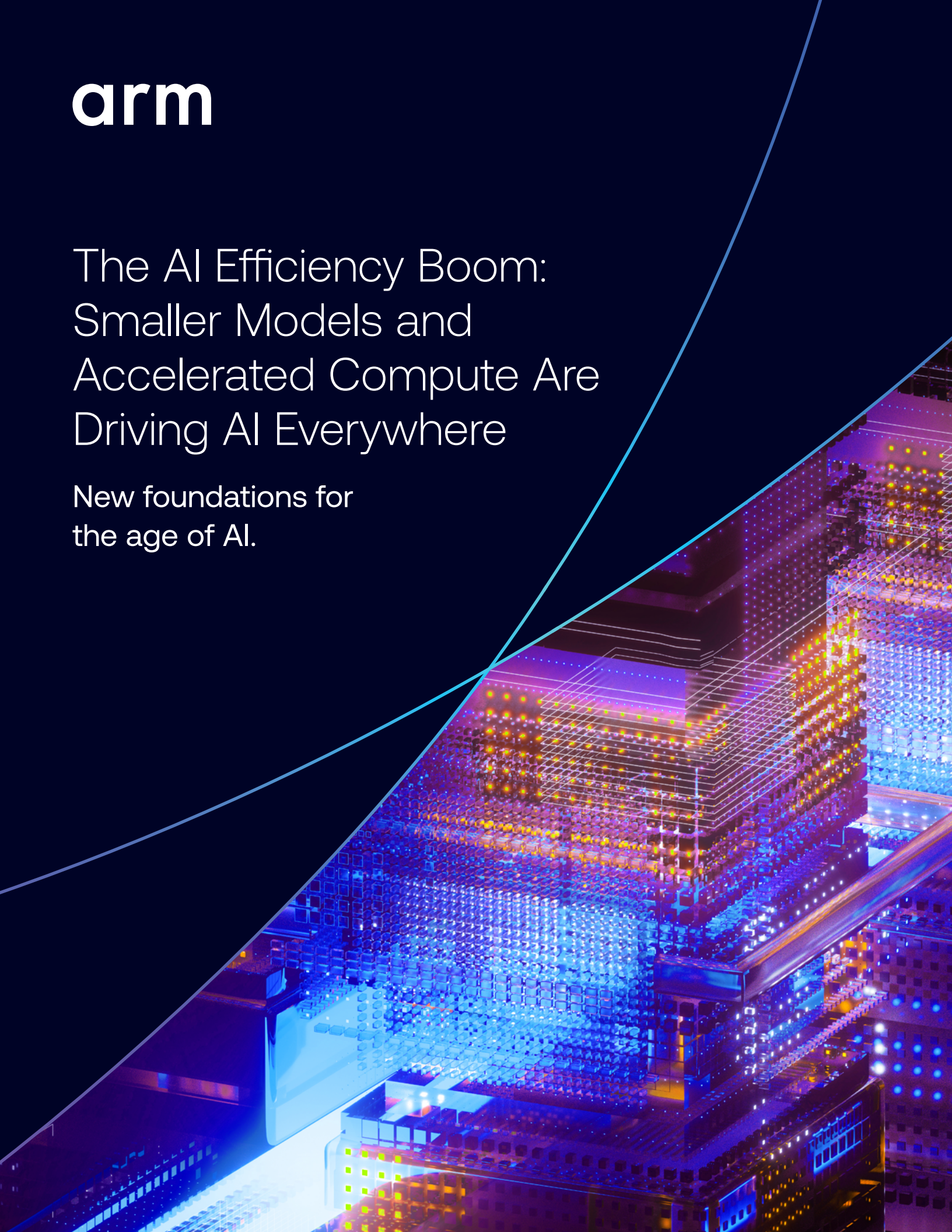
arm

The AI Efficiency Boom: Smaller Models and Accelerated Compute Are Driving AI Everywhere

New foundations for
the age of AI.



INTRODUCTION

THE ACCELERATION OF EDGE AI

Artificial Intelligence (AI) is undergoing a fundamental transformation. While early AI models were large, compute-heavy, and dependent on cloud processing, a new wave of efficiency-driven innovations is moving AI inference—the generation of model results—to the edge. Smaller models, improved memory and compute performance, and the need for privacy, low latency, and energy efficiency are driving AI adoption in mobile devices, wearables, robotics, and automotive applications.

However, this shift does not mean that demand for AI compute will decrease. Jevon's Paradox, the economics principle, tells us that when technological advancements improve efficiency, overall consumption increases rather than declines. The same is true for AI: as models become more efficient, AI adoption will become the norm across industries, with more intelligence embedded into billions of devices and systems worldwide to capture and analyze data.

This Executive Insights briefing paper explores how AI efficiency improvements—driven by model distillation, hardware acceleration, and emerging architectures—are fueling the rapid expansion of AI. We also analyze the impact of recent breakthroughs such as DeepSeek's ultra-efficient AI models and discuss the critical role of CPUs and accelerator compute subsystems in scaling AI inference at the edge.

THE ACCELERATION OF EDGE AI

THE EVOLUTION OF AI: FROM CLOUD TO AGENTIC AND PHYSICAL INTELLIGENCE

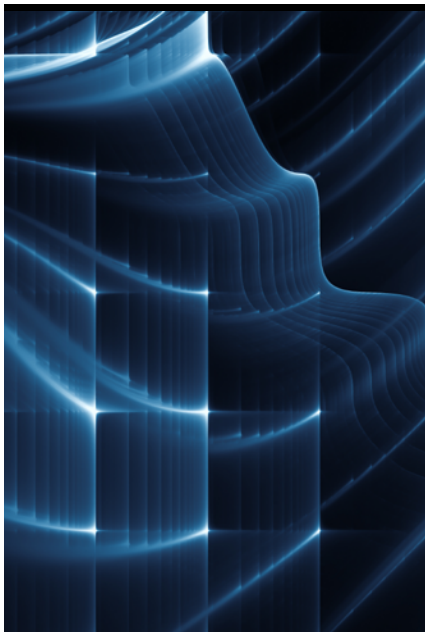
In just five years, AI has evolved rapidly in capability and in where and how it is deployed. Today, we've entered the **agentic AI phase**, where AI not only responds to inputs but proactively executes tasks, makes decisions, and adapts in real time. Think of personal assistants that summarize meetings, write code, adjust home settings, or automate customer service flows without constant human prompting.

Alongside this, **physical AI**—embedded intelligence in devices that sense, interpret, and act in the physical world—is emerging as the next frontier.

Together, agentic and physical AI mark a turning point: AI is no longer confined to digital interactions—it is becoming a ubiquitous, embedded force across every environment, industry, and form factor.

CHAPTER 1

WHY IS AI INFERENCE MOVING TO THE EDGE?



KEY DRIVERS OF EDGE AI ADOPTION

The shift of AI inference from centralized cloud environments to edge devices is a strategic response to both technological and market demands. Consumers increasingly expect faster, more private, and responsive experiences, while device manufacturers face growing pressure to deliver efficient, on-device intelligence. Indeed, generative AI inference is a major catalyst for edge computing, as enterprises prioritize low latency, enhanced security, and cost-effective processing of proprietary data in real time, according to Global management consulting firm Bain & Company. As further evidence of this, the consultancy McKinsey forecasts that by 2030 over 70% of AI inference will happen at the edge. Let's double click on some of the drivers.

Smaller AI models: Model distillation techniques compress large, complex models into efficient, compact versions capable of running on edge devices with minimal memory and compute requirements. This means devices with as little as a few megabytes of memory can now run inferencing tasks that previously required cloud connectivity and many gigabytes of RAM.

Improved compute capabilities: The rise of neural processing units (NPUs) and other specialized accelerators now enhance traditional CPUs by adding dedicated hardware optimized specifically for AI workloads – effectively supercharging overall AI performance.

Privacy and security: On-device AI inferencing reduces exposure to cloud-based security risks. With AI inferencing on the edge, sensitive data can stay securely on devices instead of being transmitted over networks.

Lower latency: Edge AI eliminates delays associated with cloud processing, ensuring real-time responsiveness for mission-critical applications. This is essential in scenarios such as when a self-driving car needs to identify a pedestrian, a surgical robot must respond to unexpected tissue movement, or an augmented reality system needs to overlay real-time information.

Energy efficiency: Cloud datacenters consume massive amounts of electricity—by some estimates, accounting for 1-2% of global energy usage—with cooling requirements further increasing their carbon footprint. Edge-based inference significantly reduces the power consumption associated with cloud computing, supporting sustainable AI adoption.

Economics: The financial burden of computation shifts depending on where inference occurs. In the cloud, companies absorb infrastructure costs and often subsidize usage to build customers and monetize data. At the edge, the consumer typically pays—indirectly through device costs or directly through app purchases—making efficiency critical for adoption. These systems-on-chip (SoCs) are highly integrated and incorporate all or most electronic components onto a single chip, including CPUs, GPUs, modems, image signal processors, memory, I/O interfaces and even AI accelerators, like NPUs. This enables a wide range of applications and services that are used on today's smartphones and other mobile devices.

INDUSTRIES RAPIDLY ADOPTING EDGE AI

The shift toward edge AI is not merely theoretical—it's already transforming multiple industries as organizations recognize the competitive advantages of deploying AI capabilities directly on devices. This revolution is happening across diverse sectors:

Mobile devices, including modern smartphones, now perform complex tasks like real-time language translation without internet connectivity, allowing travelers to communicate seamlessly in foreign countries even without cellular service. Generative AI capabilities enable on-device creation of images, text, and music, putting creative tools in everyone's pocket without privacy concerns. We're also seeing the rise of AI PCs—next-generation laptops and desktops designed with dedicated NPUs to accelerate local inference and deliver enhanced productivity, creativity, and security features.

IoT and smart home ecosystems represent fertile ground for edge AI implementation. Smart speakers, cameras, and sensors now process commands and detect anomalies locally, reducing response times and cloud dependency. AI-powered automation systems learn household patterns to optimize energy usage, security, and comfort without transmitting sensitive data to remote servers. Predictive analytics capabilities allow smart home systems to anticipate needs based on historical patterns—adjusting the temperature before residents return home or identifying potential appliance failures before they occur.

WHY IS AI INFERENCE MOVING TO THE EDGE?

Automotive edge AI enables features such as adaptive cruise control, personalized infotainment, and driver and passenger monitoring—powered by application-specific models running directly on the vehicle. Advanced driver assistance systems (ADAS) use on-vehicle neural networks to instantly recognize road hazards, pedestrians, and traffic signs, enabling split-second interventions that can help save lives. By shifting AI inference to the edge, automakers can meet growing consumer expectations for responsiveness, safety, and personalization—while reducing reliance on the cloud for those types of functions.

Wearables and healthcare devices benefit tremendously from edge AI capabilities. Smartwatches and fitness trackers now perform sophisticated health monitoring, detecting irregular heartbeats, sleep apnea, and even early signs of conditions like diabetes or atrial fibrillation—all while keeping sensitive health data on the device. Medical devices equipped with edge AI can provide AI-assisted diagnostics in remote locations without reliable internet connectivity, democratizing access to healthcare expertise.

Earlier we mentioned physical AI. In the industrial and robotics sectors, factory floors now employ visual inspection systems that identify defects with superhuman accuracy in milliseconds, improving quality while reducing waste. Predictive maintenance algorithms analyze equipment vibration, sound, and temperature patterns locally to forecast failures before they occur, dramatically reducing downtime. Autonomous robots and cobots (collaborative robots) use on-board AI to navigate dynamic environments and work safely alongside humans without requiring constant network connectivity.

CHAPTER 2

DEEPSEEK AND THE RISE OF ULTRA-EFFICIENT AI MODELS

DeepSeek's remarkable efficiency improvements, announced in January 2025, might initially suggest a future where less computational hardware is needed for AI. In fact, for a brief period, the announcement threw into question the future of datacenter investment levels. Enter Jevon's Paradox. The principle was named after William Stanley Jevons, who, in the 19th century, first observed that technological progress increases the efficiency with which a resource is used. Jevons determined that as coal-fired factories, machines, and vessels got more efficient, overall consumption of coal increased while costs declined. That prompted people to invent more ways to use coal. The same holds true for computing resources.

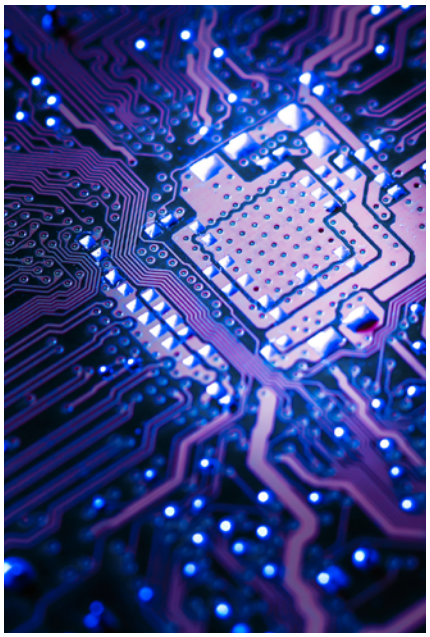
While DeepSeek demonstrated that models requiring 94% less computational cost per token can match the capabilities of more expensive alternatives, the demand for AI chips is paradoxically surging. In fact, OpenAI co-founder Sam Altman said that GPT-4.5 usage is being limited—not by demand, but by a shortage of available GPU capacity. In response to growing infrastructure constraints and competitive pressure, major cloud providers like Microsoft, Google, Meta, and Amazon have announced plans to invest hundreds of billions of dollars in AI chips and datacenters in 2025—nearly 50% more than their record-breaking 2024 spending.

This pattern mirrors what we've seen repeatedly in computing history. Moore's Law has driven a billionfold improvement in computing efficiency over seven decades, yet we don't satisfy all global computing needs with a single penny-sized processor. Instead, computation has proliferated throughout our economy, creating unprecedented demand for silicon. AI is following this same trajectory, but at an even more accelerated pace—with efficiency improvements estimated at 13.8x per year compared to Moore's Law's 2x every two years.

The more efficiently we can deliver AI capabilities, the more applications we discover and the more computational resources we ultimately need to fulfill humanity's appetite for artificial intelligence. Put it another way, do you want just one Einstein at your fingertips or countless Einsteins?

CHAPTER 3

THE FUTURE OF EDGE AI HARDWARE: CPU AND ACCELERATOR



As Jevon's Paradox drives demand for more AI, the question becomes not only how much compute is needed, but what kind. This expanding AI footprint requires specialized hardware architectures optimized for the unique computational patterns of machine learning. While the industry has made remarkable strides in algorithmic efficiency, the physical foundations for running these algorithms matter tremendously.

FINDING THE RIGHT BALANCE: CPUS AND ACCELERATORS AT THE EDGE

Today's AI revolution requires the right computing foundation. While CPUs have traditionally handled basic AI tasks adequately, modern AI demands more. The solution isn't choosing between CPUs or specialized AI accelerators—it's strategically combining them.

This balanced approach creates an optimal system where CPUs manage general operations while purpose-built accelerators tackle intensive AI workloads. This means devices that deliver both performance and efficiency—essential for competitive edge AI deployments that protect data privacy, reduce latency, and minimize power consumption.

High-performance matrix operations form the mathematical foundation of neural networks. These operations involve multiplying and transforming large arrays of numbers simultaneously—a task well-suited to GPUs and NPUs and at which these excel because of their parallel processing architectures. For example, a modern smartphone's NPU can process thousands of matrix calculations simultaneously, enabling real-time portrait mode photography by running sophisticated segmentation models that separate subjects from backgrounds.

Low-latency real-time processing is critical for applications where delay means failure. Consider autonomous vehicles, where milliseconds can mean the difference between collision and avoidance. Edge-specific AI chips can process multiple camera and sensor feeds with sub-millisecond response times, performing object detection and trajectory calculations locally without the latency of cloud-based solutions. A fully optimized heterogeneous platform offers an excellent opportunity to minimize latency in real-time processing—an essential requirement in use cases such as industrial automation.

THE FUTURE OF EDGE AI HARDWARE: CPU AND ACCELERATOR

Power-efficient AI execution carefully balances performance and energy constraints. This is necessary for battery-powered devices where every milliwatt matters. AI accelerators tend to achieve dramatically better performance-per-watt ratios compared to general-purpose processors. A smartwatch with a specialized NPU can continuously monitor heart rhythm for irregularities while consuming mere milliwatts, allowing the device to maintain multi-day battery life.

The integration of these accelerator capabilities with traditional CPU strengths creates a symbiotic relationship that's greater than the sum of its parts. CPUs excel at sequential processing, decision-making logic, and handling diverse workloads, while AI accelerators deliver the computational muscle for neural network operations.

Let's examine how this balanced approach translates to measurable business outcomes. Recent benchmarks demonstrate how Arm technology is delivering the performance, efficiency, and scalability needed for meaningful AI deployments—from smartphones to industrial equipment. These metrics illustrate what's possible when organizations implement the right blend of CPU processing and specialized acceleration.

KEY PERFORMANCE BENCHMARKS FOR LLM INFERENCE

Delivering seamless, on-device AI experiences requires the right mix of compute technologies, each optimized for different types of workloads. Here's how Arm CPUs, NPUs and KleidiAI provide a variety of performance and efficiency improvements across different AI tasks and workloads, from LLMs and generative AI to speech and vision.

CPUs: Efficient on-device LLMs

Arm CPUs are now powering real-time AI experiences on mobile devices, including running large language models like Llama 3.2 (1B tokens) with impressive efficiency:

- **5x faster performance** for key language tasks compared to previous methods
- **2-second response time** for summarizing multi-message prompts
- **Over 40% lower memory use**, enabling smoother performance on small devices
- **Models are over 50% smaller**, reducing download size and speeding up installation.

These improvements make it possible to run powerful AI apps on phones and PCs—without relying on a GPU or cloud connection.

NPUs: Maximum Performance at the Edge

When more demanding AI workloads are needed—like those used in advanced assistants, image recognition, or on-device translation—Arm dedicated neural processors (NPUs) deliver the acceleration:

- Runs complex transformer models (like BERT) fully on-device
- Powers small LLMs like Llama2-tiny with real-time performance
- Up to **78% more energy efficient** than previous AI hardware generations
- Can handle **over 4,000 inferences per second** on lightweight models
- Supports high-performance use cases like real-time video processing in wearables, vehicles, and smart cameras—all with low power draw.

KleidiAI: Scale and Speed Across the Stack

KleidiAI is the AI software engine from Arm that boosts performance across a wide range of AI workloads—from language models to speech recognition to computer vision—running on Arm-based cloud infrastructure and edge devices:

- Speeds up key language tasks like prompt response and token generation
- Delivers **up to 18× faster inference** for models like Llama and Gemma
- Provides **up to 3.5× better cost efficiency** than comparable x86 and GPU setups
- Enhances speech recognition and image processing with **multiple-fold performance gains**
- Integrated with industry-standard tools like PyTorch and TensorFlow, making it easy for developers to tap into this speed.

These performance indicators translate directly to competitive advantage—whether you're launching AI-powered consumer products or scaling enterprise solutions. By implementing the right balance of CPU and accelerator technologies, organizations can deliver more capable AI experiences while managing costs, energy consumption, and deployment complexity.

ACCELERATING DEVELOPMENT WITH THE ARM DEVELOPER HUB

To support developers and organizations building at the edge, Arm has launched the Arm Developer Hub—an optimized collection of open-source and proprietary models designed to run efficiently on Arm-based CPUs, GPUs, and NPUs. With models pre-tuned for edge performance, developers can quickly evaluate and deploy solutions across devices ranging from wearables to AI PCs and automotive platforms. The Developer Hub removes guesswork and accelerates time to value by enabling plug-and-play experimentation with top-tier AI models, right out of the box.

EMPOWERING DEVELOPERS TO BUILD THE NEXT GENERATION OF EDGE AI

Developers are central to the success of edge AI—and Arm is committed to making it easy to build, optimize, and deploy models across our ecosystem. Our Developer Hub provides pre-optimized models for rapid prototyping, while KleidiAI ensures seamless performance tuning for Arm-based hardware. Arm also offers AI performance libraries, compilers, and toolchains that integrate with popular ML frameworks like PyTorch, TensorFlow, and ONNX Runtime. Whether building in the cloud or at the edge, developers can count on broad support, best-in-class documentation, and unmatched flexibility to build AI apps that scale.

CHAPTER 4

CONCLUSION: THE ERA OF AI EXPANSION

Where AI Inference is Heading

During the coming decade, we'll witness a fundamental transformation away from primarily cloud-centric AI toward sophisticated hybrid architectures that balance centralized and distributed intelligence.

Models will continue their remarkable journey toward becoming smaller, faster, and more efficient. However, contrary to what intuition might suggest, these efficiency gains won't reduce AI deployment—they'll dramatically accelerate it. As barriers to implementation fall, previously impractical AI applications become viable, creating a virtuous cycle of innovation and adoption.

Why AI Hardware Matters

Despite the impressive advances in algorithmic efficiency, the physical substrate upon which AI pulses remains critically important. The future belongs to compute subsystems that seamlessly integrate CPUs with specialized accelerators like NPUs and GPUs, each handling the workloads for which they're optimized. CPUs will manage control flow, system operations, and sequential tasks, while accelerators tackle the most intensive matrix mathematics and pattern recognition that form the foundation of machine learning. These heterogeneous computing environments will become the norm across the entire computing spectrum, from massive datacenters to tiny, embedded devices.

Accelerating the AI Revolution

We're witnessing the early stages of an enthusiastic embrace of AI across every industry and sector. Intelligence is being embedded into processes, products, and services that previously operated without it—from agriculture to manufacturing, healthcare to transportation, education to entertainment. The scope of this transformation is difficult to overstate.

Companies that yesterday couldn't justify the cost of implementing AI will tomorrow consider it indispensable, while those already leveraging AI will dramatically expand their deployments.

The most profound implication of this pattern is that we're not approaching an end state for AI but rather standing at the beginning of a new technological frontier with immense possibilities. The efficiency breakthroughs demonstrated by companies in the AI ecosystem aren't signals of a maturing, stabilizing technology—they're nascent indicators of an intelligence revolution that will reshape our world in ways we're only beginning to understand.

CHAPTER 4

CONCLUSION: THE ERA OF AI EXPANSION

Why Arm Is Built for the Edge AI Era

Arm has been at the forefront of edge compute for decades—powering over 99% of smartphones and billions of IoT devices worldwide. With unmatched reach across the mobile ecosystem, a legacy of energy-efficient compute, and deep relationships with ecosystem partners, Arm is uniquely positioned to lead the AI transition at the edge. Our architecture is already embedded in the devices where the AI efficiency revolution is playing out—in phones, wearables, vehicles, smart cameras, and factory equipment. That scale provides a massive runway for edge AI adoption.

Learn more about how Arm is revolutionizing edge AI intelligence