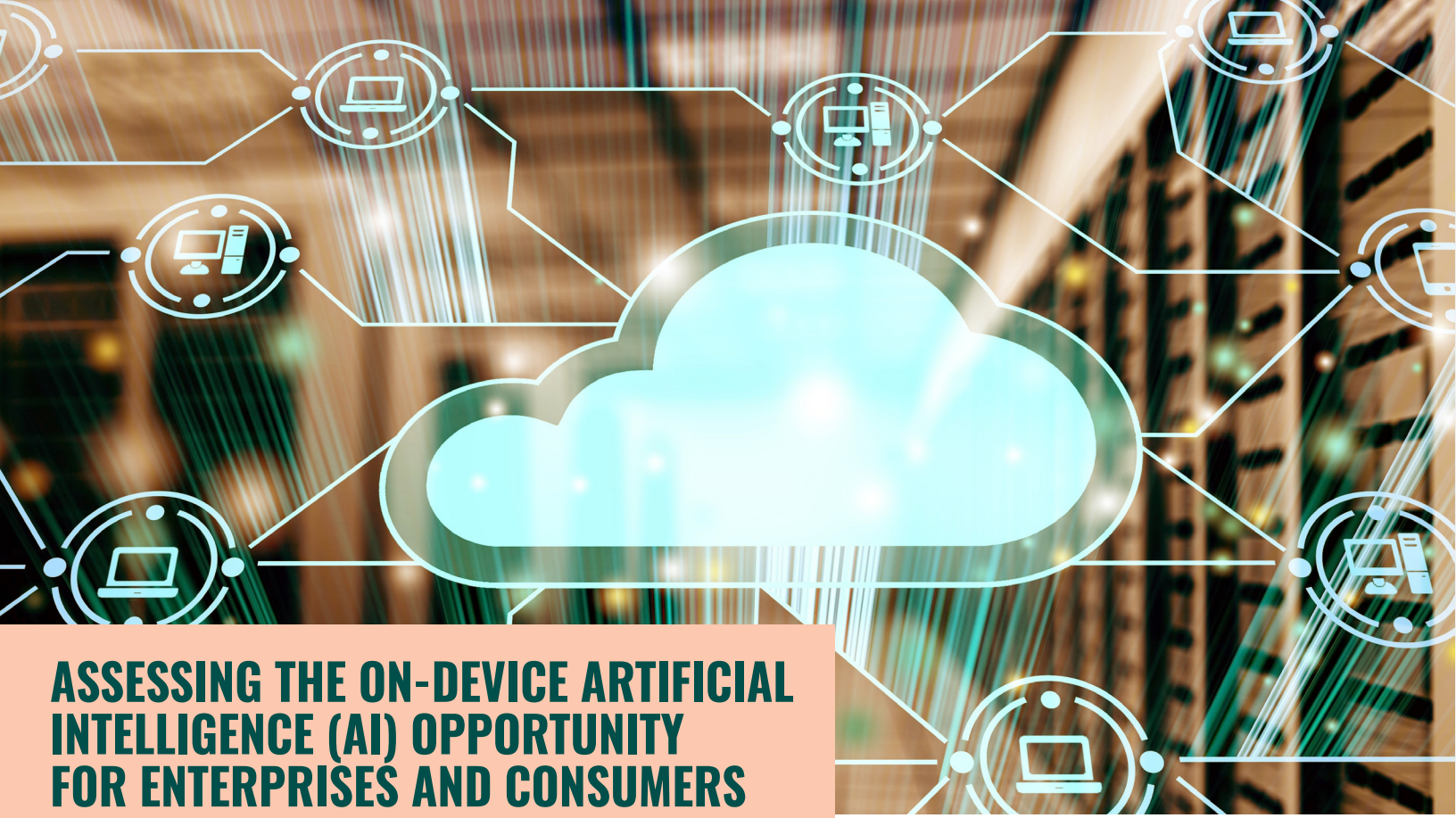


ASSESSING THE ON-DEVICE ARTIFICIAL INTELLIGENCE (AI) OPPORTUNITY FOR ENTERPRISES AND CONSUMERS

Lead Analyst: Reece Hayden
Analysts: Paul Schell, Malik Saadi

ABIresearch
THE TECH INTELLIGENCE EXPERTS™

TABLE OF CONTENTS

INTRODUCTION

The Artificial Intelligence (AI) market continues to develop rapidly, spurred by the accessibility of generative AI, which has expanded opportunities for enterprise and consumer deployments. Until now, generative, and other AI models have been mostly deployed in the cloud, but as AI becomes more accessible and the market looks to scale across use cases, this cloud-centric framework presents significant commercial and technical challenges. On-device AI capabilities—and eventually hybrid AI—in which inferencing workloads for large complex models are performed locally, are emerging to better support the case for deploying AI at scale. This white-paper explores the opportunity of on-device AI for enterprises and consumers, and the role that "productive AI" applications will play in building a strong Return on Investment (ROI)-based value proposition for end users.

KEY TAKEAWAYS

- **Moving inference workloads to the device will help enterprises and consumers unlock AI at scale by solving foundational commercial and technical challenges** like data privacy, network and server latency, and networking and cloud infrastructure subscriptions and/or costs, while enabling users to enjoy a variety of generative AI applications tailored to their personal needs. However, on-device generative AI will bring potential challenges around workload management, power consumption, and memory burden, which will require Neural Processing Units (NPUs) and model shrinking already happening today.

- **On-device generative AI will drive enterprise deployment of smart devices and new AI models into existing processes, enabling users to be more productive and efficient, thanks to potential cost and time savings.** For example, manufacturers have resisted deploying Augmented Reality (AR) on factory floors; however, local generative AI processing enables high-value applications that will offer tangible ROI.
- **On-device AI will be the first step toward hybrid AI systems,** which will help optimize resource usage, application performance, and data privacy, enabling joint processing from cloud to device. Making the leap to hybrid AI will require: intelligent systems able to distribute AI workloads across multiple cloud resources and multiple devices and Operating Systems (OSs); and integrated model architectures and clearly defined AI workload rules.
- **The biggest barrier to on-device generative AI's commercial success will be software immaturity and applications targeting experience, not productivity.** Encouraging end users to buy new devices remains challenging; this can only be achieved by developing a strong ROI-driven business case supported by either time or money savings. Achieving this requires the market to go beyond on-device hardware and to develop a killer productivity AI application built on a small, optimized model. Market leaders like Qualcomm have been quick to recognize the importance of developing full-stack products in collaboration with cloud, software, and AI vendors such as Google, Meta, and Microsoft.

STATE OF ARTIFICIAL INTELLIGENCE

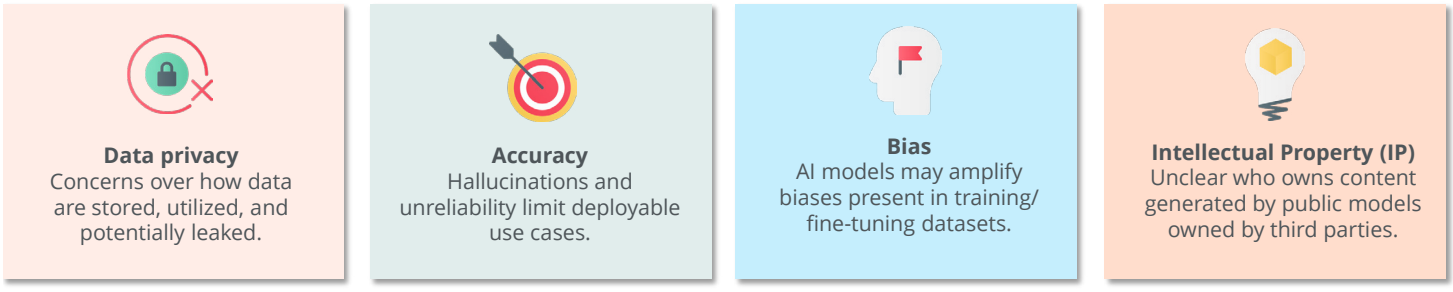
OVERVIEW

AI is not new to consumers or enterprises, but the introduction of ChatGPT has created public awareness of the capabilities of generative AI and enabled wider usage of models, tools, and applications. Generative AI applications are now being used to augment consumer and enterprise processes. These basic chatbot or content generation applications use Large Language Models (LLMs), or Large Visual Models (LVMs), trained on billions or even trillions of data points to determine parameters (weights and biases) within their neural network structure that are capable of generating content (i.e., text, image, sound) based on inputs or prompts.

Although using generative AI at scale brings opportunities, it also creates significant risks and anxieties that everyone must be aware of prior to usage, as explored in Figure 1.

Figure 1: Generative AI Risks

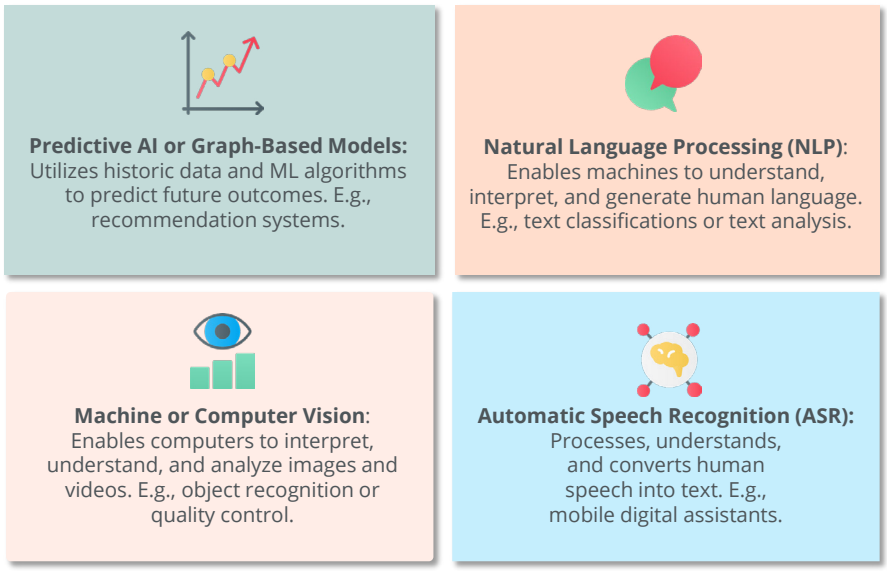
(Source: ABI Research)



But generative AI is not the only framework that offers consumers and enterprises value. Figure 2 explains the other four traditional AI frameworks. Generative AI is capable of handling many of the same applications and use cases, but often due to other considerations like price, memory, training data, and risk, other "traditional" AI frameworks are used. Expect most consumer and enterprise use cases to leverage a combination of traditional and generative AI models. Each has core competencies, risks, and commercial factors that suit different use cases and must be assessed prior to deployment.

Figure 2: "Traditional" AI Frameworks

(Source: ABI Research)



IMPACTFUL TRENDS

This section explores the key trends supporting the development of the on-device AI market.

- **Availability of Neural Processing Units (NPUs) Will Complement Graphics Processing Units (GPUs) and Central Processing Units (CPUs):** Different processors excel at different AI tasks. CPUs are used for sequential control, GPUs for parallel data streaming, and the NPU for core AI workloads. Combining these architectures within a System-on-Chip (SoC) will enable more efficient AI workload processing with improved application performance, thermal efficiency, and battery life to enable new and enhanced generative AI experiences. One example is Qualcomm's Snapdragon 8 Gen 3, a heterogeneous computing platform combining the company's flagship Kryo CPU, Adreno GPU, and Hexagon NPU. This combination enables the platform to support inference of a variety of AI workloads from specific use cases like Machine Vision (MV) or voice control to more sophisticated models like LLMs with up to 10 billion in smartphones and 13 billion parameters on Personal Computers (PCs).

- **Chip Vendors, Original Equipment Manufacturers (OEMs), and Independent Software Vendors (ISVs) Align:** Partners are coming together to build "productivity-focused" applications optimized for on-device AI hardware that aligns with consumer and enterprise pain points. This will create ROI-driven demand for new devices based on time/money savings.

- **Open-Source Market Garners Further Investment:** Vendors look to support speed of innovation, lower barriers to entry, and further democratized access to market-leading generative AI models. This is particularly the case for building bullet-proof development tools that developers can easily use to develop innovative AI applications.

- **Strong Support for Compressed and Optimized Models with Fewer than 15 Billion Parameters:** These models, developed through optimization techniques like parameter pruning, reduce power consumption, reduce inferencing time, and limit memory burden, while demonstrating behaviors and accuracy similar to "giant" models. These are suited to on-device deployments with lower resource burdens. Google, Meta, Mistral, Baichuan, and Microsoft have invested strongly in these models, aiming to replicate the performance of "giant" models.

- **Chip Vendors Build Software Development Kits (SDKs):** Building an on-device generative AI ecosystem requires lower developer barriers to entry. SDKs support efficient optimization of applications in underlying hardware. Qualcomm's AI Stack provides a Neural Processing Engine SDK, which ISVs (e.g., Facebook) have already leveraged to deploy new applications. This is supported by Qualcomm AI Hub, which offers a library of optimized models for supported platforms and specific use cases.
- **No/Low-Code Platforms Will Reduce Barriers to Development, Accelerating a Productive AI Ecosystem:** Vendors are investing in visual software development environments to improve accessibility and enable "anyone" to build applications. The next step in the market will leverage natural language to build applications, as exemplified by OpenAI's release of Generative Pre-Trained Transformers (GPTs).

BUILDING THE CASE FOR ON-DEVICE AI

AI models have mostly run training and inferencing workloads in the cloud, as it offers scalable computing resource capacity, high-speed networking, and high memory and storage reserves. However, moving forward, centralized cloud deployments will inhibit use case scalability for numerous commercial and technical reasons. ABI Research believes that the answer is moving inferencing closer to the end user with on-device AI capable of handling larger models (e.g., LLMs for generative AI).

On-device AI deployment moves inferencing workloads from the cloud to the device. This brings an improved commercial and technical case for AI deployment. On-device AI is not an entirely new phenomenon, as basic AI workloads for audio processing or visual enhancement are already running locally; hardware (and software) innovation has now made it possible to run large generative, Automatic Speech Recognition (ASR), Natural Language Processing (NLP), or image generation workloads on device. This offers both consumers and enterprises a strong value proposition compared to the "traditional" cloud-centric models, as explored in Section 4.1 and Section 4.2. But providing hardware capable of running large models will not offer sufficient commercial value to create growth in a largely stagnant market (especially smartphones and PCs). Subsequently, ABI Research believes that this must be complemented with targeted "productivity" applications that either save time or money, as this will build a strong ROI-driven case for buying new devices with on-device generative AI capabilities.

CONSUMER MARKET

A cloud-centric AI model deployment provides immense compute and power resources to power AI workloads, but as AI scales across consumer applications, on-device AI will become increasingly necessary to keep networking and server costs under control and mitigate data privacy challenges. Figure 3 explains on-device AI's experience and technical value proposition.

Figure 3: On-Device AI Value Proposition

(Source: ABI Research)

Experience

Enhance experience by eliminating network latency: improve application performance and enable offline applications. This is especially important for real-time interactive use cases like digital assistants by ensuring more natural and productive interaction.

Develop a personal AI twin without compromising data privacy: device sensors and memory can leverage personal data to fine-tune models locally enabling models to adapt over time and provide personalized insights to the user.

Benefit from a seamless multi-device experience: AI workloads can move across ecosystem of devices and locations to automatically optimize performance, power consumption, and experience for end users

Technical

Data privacy and personal AI models: user prompts, personal data, ingested sensor data all remains on the device which eliminates data privacy challenges.

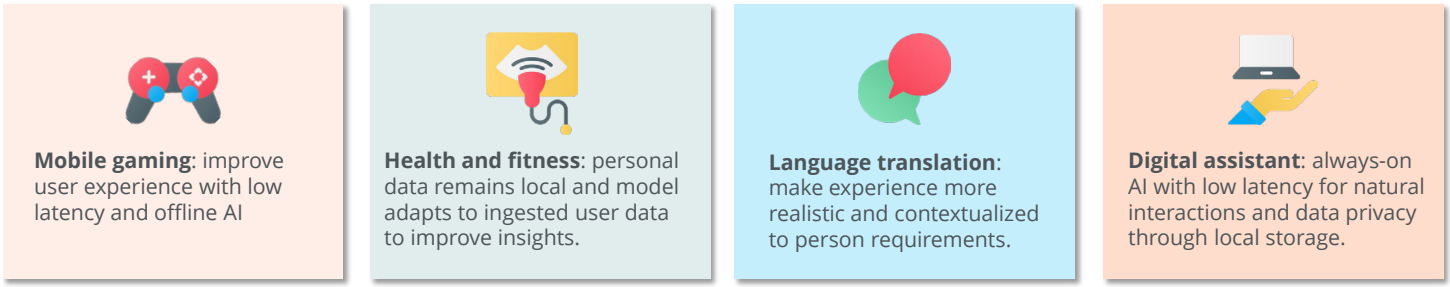
Lower networking, cloud subscription, and processing costs: as AI models scale, networking and compute costs will grow exponentially. On-device eliminates many of these costs and reduces network burden, as well as reducing energy usage.

Embed AI-based security: AI models can provide enhanced threat detection and other cybersecurity functions with very low latency. E.g., voice/facial recognition and malware detection.

Data privacy, cost optimization, and more life-like human-machine communication will be key value drivers for consumer on-device AI. Many AI use cases will benefit from local processing, as it brings lower latency, which will drive performance and accessibility for consumers and enterprises. Some of the use cases that will benefit more are explored in Figure 4.

Figure 4: Examples of On-Device AI Consumer Use Cases

(Source: ABI Research)



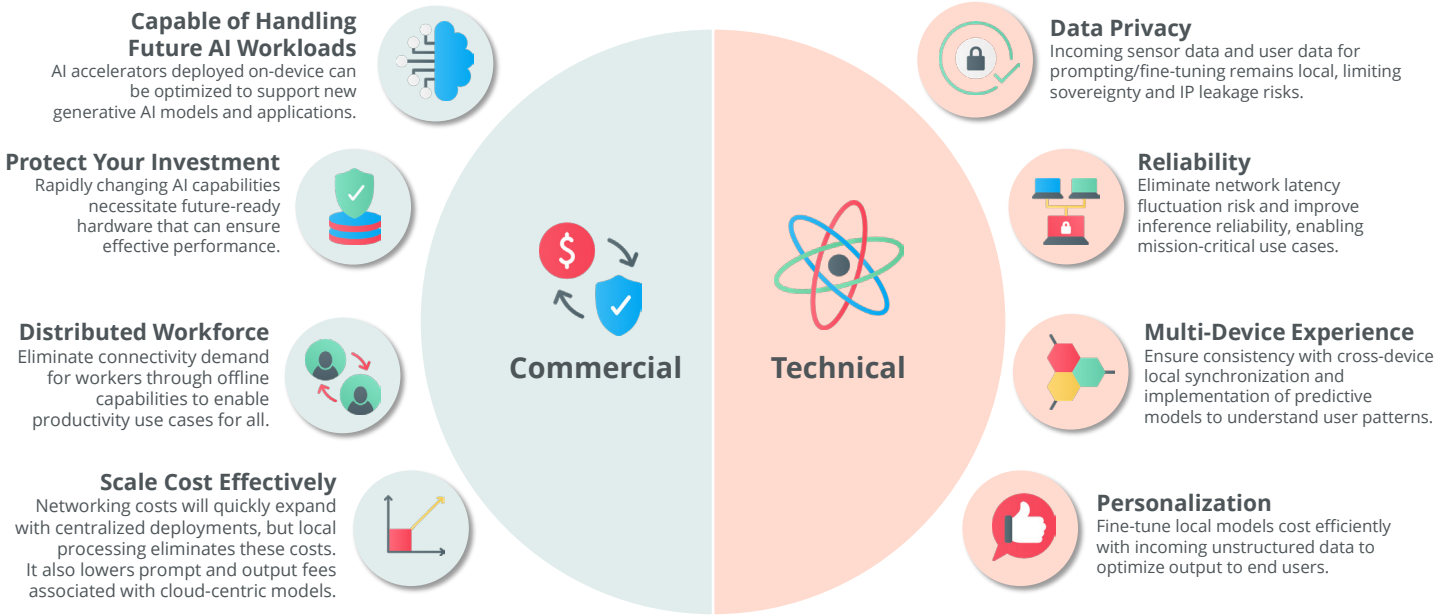
But on-device hardware must be complemented with "productivity" AI applications that go beyond device experience, and build a tangible ROI, necessitating device upgrades. For example, generative AI applications that automatically complete consumer tax returns could save money, as users would no longer need an external accountant and time. This builds a strong ROI that can justify buying a new device. ABI Research expects that this will be important for both the consumer and enterprise markets.

ENTERPRISE MARKET

AI has already been deployed across enterprises, enabling hundreds of use cases. But as enterprises look to scale models, cloud deployment will create commercial and technical barriers such as networking costs, bandwidth congestion, and soaring cloud hosting costs. On-device AI can solve these challenges. Its value proposition is highlighted in Figure 5.

Figure 5: On-Device AI Enterprise Value Proposition

(Source: ABI Research)



Although on-device AI has cross-vertical appeal, early adopters will have common features that lend themselves to on-device AI deployment: a highly distributed workforce, latency-sensitive use cases, low connectivity deployment environments, strict data requirements, and large upstream data flows. Table 1 explores six early deployments of on-device AI within the enterprise.

Table 1: Early Deployments of On-Device AI within the Enterprise

(Source: ABI Research)

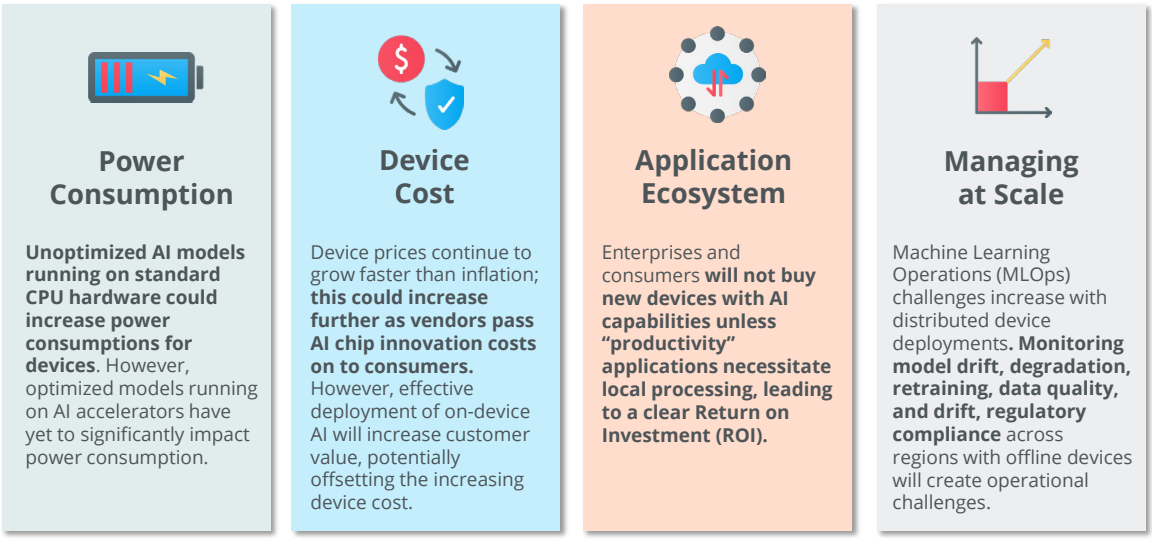
Vertical	Manufacturing	Healthcare	Logistics & Transportation	Telecommunications	Back End Operations & Offices	Professional Services
AI deployment pain points	Reliability for autonomous Internet of Things (IoT) Low connectivity environments with distributed workforce	Patient data security Reliability and latency sensitivity	Distributed workforce with poor connectivity Global distribution with asymmetric data regulation Network reliability	Distributed workforce High associated costs Customer data privacy concerns High Operational Expenditure (OPEX) due to workforce deployment	Productivity cost and lack of intuitiveness of current tools Shortage of AI talent Poor data and AI governance frameworks	Distributed and moving workforce Data sovereignty
On-device AI use case(s)	IoT automated workflows, AI-generated workflow instructions, remote assistance for repair; staff training, machine predictive monitoring	Patient monitoring and care, patient history summarization, remote patient assistance and remote diagnostics, patient chatbots, surgeon training with AI-enabled Extended Reality (XR)	Intelligent route mapping, supply chain tracking, XR Heads-Up Display (HUD)	Dynamic instructions for field technicians; digital assistant and chatbot for technicians, XR HUD	Intelligent productivity tools powered by generative AI; enhanced computing experiences; local virtual assistant	Running personal copilot during off-site projects; digital assistants; productivity applications on the move
What devices will AI be deployed on?	IoT, wearables, XR	IoT, wearables, smartphones, XR, PCs	IoT, wearables, smartphones, XR, automotive	IoT, XR, smartphones, laptops	PCs, smartphones	Smartphones, PCs

KEY CHALLENGES

A clear value proposition exists for on-device generative AI across consumers and enterprises, but moving inferencing to the device level brings commercial and technical challenges that need to be addressed. These challenges are addressed in Figure 6.

Figure 6: Perceived Challenges of Enterprise On-Device AI

(Source: ABI Research)



Most commercially impactful among these challenges is the application ecosystem. Currently, AI applications development is focusing on experience with tools enabling automatic photo editing, for example. However, these will not build a strong enough case for device upgrades in a stagnant or new device market. Instead, applications should target "productivity" and ease of use with quantifiable time or money savings.

ON-DEVICE GENERATIVE AI BY DEVICE TYPE

The following section provides an assessment of the applications, opportunities, and challenges for on-device generative AI across different form factors.

SMARTPHONES AND TABLETS

Smartphones and tablets are no strangers to AI with voice assistants like Siri or Google Assistant leveraging cloud-based NLP models. However, these applications, often relying on the cloud, have created little commercial value—offering poor user experience with significant latency and network connectivity requirements. To mitigate unoptimized experiences provided by cloud-based AI, modern high-end smartphones now incorporate on-device AI accelerators to handle inference for voice control, AI-enhanced imaging, and other basic AI applications locally. However, the value added by these traditional AI applications alone may not be sufficient to substantially impact consumer decisions to upgrade smartphones.

AI's commercial success in smartphones will hinge on the ability of these devices to support generative and general models that can power a wide variety of generative AI applications. These devices need to be equipped with adequate on-device processing engines to accomplish the following:

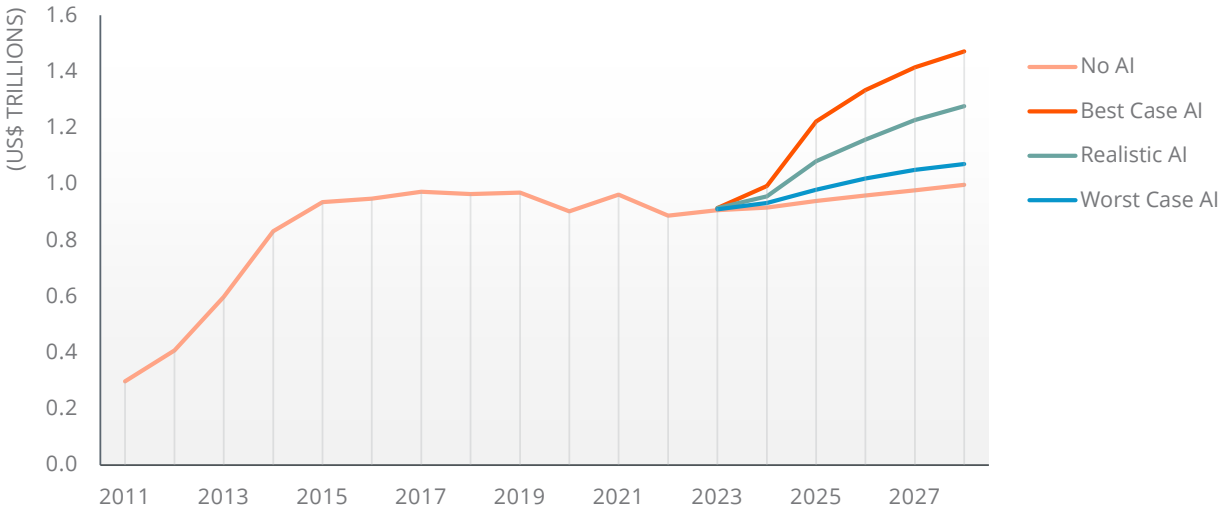
- **Develop Intuitive, Natural Interactions with AI:** Reducing latency for ASR, NLP, and generative AI models to enable real-time responses that enhance/personify digital assistants.
- **Support "Personal AI" Models:** Local models that are fine-tuned, for now in the cloud, based on data inflows from sensors and end users to provide personalized responses/insights without compromising data privacy.
- **Enable Richer Application Development Environment:** Developers will be able to create a plethora of innovative applications beyond voice control and enhanced imaging. Smartphone users will be able to enjoy these applications that will improve the overall user experience, while helping them be more creative and more productive.
- **Establish Data Sovereignty and Data Privacy:** Ensure that data regulation and governance are not compromised, which is especially applicable for enterprises with employees distributed across multiple regions.
- **Enable Always-on AI Regardless of Whether the Smartphone Is Connected or Not:** Model continues to run all AI workloads from basic and use case-specific to larger generative AI models and provide unprompted recommendations in real time. Local processing limits upstream data flows and ensures that applications work when disconnected from networks.

Growth in the smartphone market has remained limited over the last 8 years, which resulted from a lack of compelling innovation and led to growing replacement cycles, as well as consumer decisions to upgrade to higher-end devices and hold onto them for a longer period of time. As the consumer relationship with their smartphone is shifting from one based on entertainment to one that could help them enhance their creativity and productivity in everyday life, unlocking growth in this market will not be achieved by only enabling "experience" applications like generated, personalized wallpaper. Instead, development should target productive AI; for example, consumers will appreciate using generative AI applications that help them manage smart homes, optimize energy usage, complete administrative tasks, and see savings from their utility bills. These applications will help consumers achieve significant savings in terms of time and cost. Some vendors are already recognizing the importance of productive AI for on-device generative AI's value proposition; one example is Qualcomm and Samsung's partnership focused on bringing generative AI capabilities to the Galaxy S24 series.

Chart 1 explores how on-device generative AI could impact the smartphone market size in four different scenarios (these scenarios will also apply to the laptop/PC market in Section 4.4.2).

Chart 1: Smartphone Market Size with and without On-Device Generative AI World Markets: 2011 to 2027

(Source: ABI Research)



ABI Research expects that maximizing market growth requires a strongly aligned hardware and software proposition that includes productivity AI applications targeting consumer and enterprise pain points. Hardware-optimized productive AI targeted at specific pain points will provide time and/or money savings sufficient to encourage end users to buy new devices. This will decrease the refresh rate, while new hardware will increase Average Selling Prices (ASPs), creating new growth in both the smartphone and PC markets. Below, ABI Research provides the qualitative methodology that informed each situational forecast:

- **Best Case:** End-to-End (E2E) optimized on-device generative AI proposition that combines hardware, software, and verticalized, productivity-focused applications. Builds a strong productivity AI proposition that incentivizes the purchase of new devices at a higher ASP, significantly increasing shipments and leading to long-term market growth.
- **Realistic AI:** Application ecosystem mainly focused on experience (not productivity) creates some, but limited device market growth.
- **Worst Case AI:** Market remains relatively flat as on-device AI hardware is not brought to market with consumer or enterprise applications.
- **No AI:** AI is not deployed on-device and the market remains stagnant over the forecast period.

Achieving the "best case" scenario requires stakeholders to promote the development of productivity AI applications for smartphones. Section 5 highlights the key software innovation and partnership alignment that will help maximize market growth.

PCS AND LAPTOPS

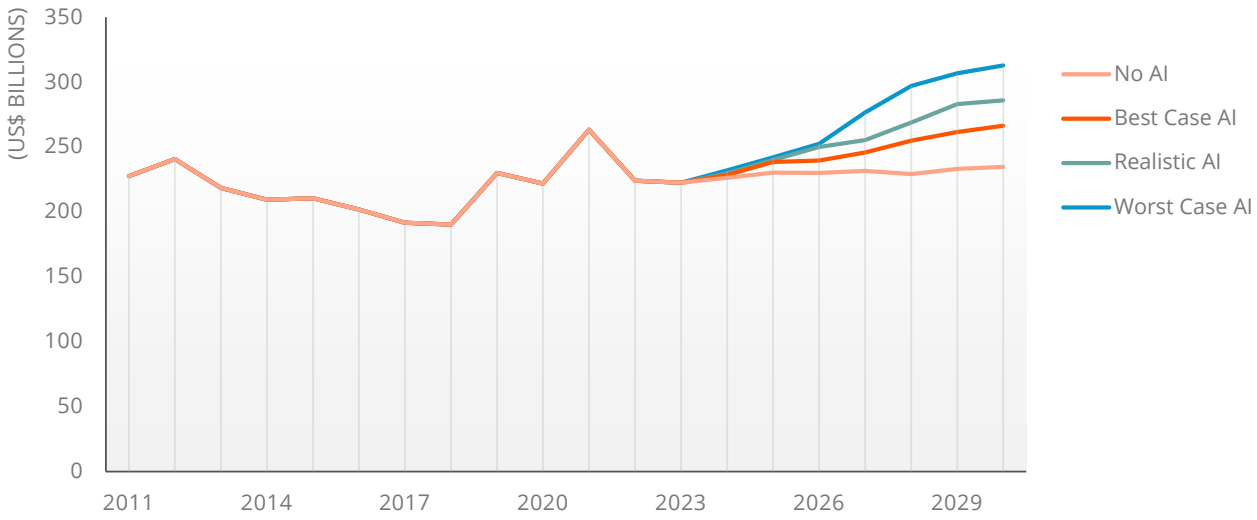
Deploying AI natively on PCs or laptops will create enormous value around offline productivity, data privacy, enhanced user-device communication, and model personalization. It will have applicability for consumers and enterprises as it supports productivity anywhere. For example, professional service providers will be able to run Microsoft Copilot or other creativity tools powered by generative AI on the move to augment their day-to-day workflows and reduce costs, while traveling to client sites.

Stakeholders have been bullish around this opportunity with hardware innovation through the integration of NPUs into PC processors, but also recognize the importance of building a full-stack value proposition. Qualcomm has continued to expand its Snapdragon compute ecosystem with ISV partners, including Microsoft. Meanwhile, competitors like Intel are developing PC AI through the AI accelerator program, providing funding and expertise for ISVs. These strategies encapsulate the understanding that "productivity-focused software" will be the key differentiator for on-device generative AI.

Chart 2 shows the market size for PCs and provides a forecast in four different scenarios, all related to on-device generative AI deployment and productivity AI (see Section 4.3.3 for scenario breakdown).

Chart 2: PC Market Size with and without On-Device Generative AI World Markets: 2011 to 2029

(Source: ABI Research)



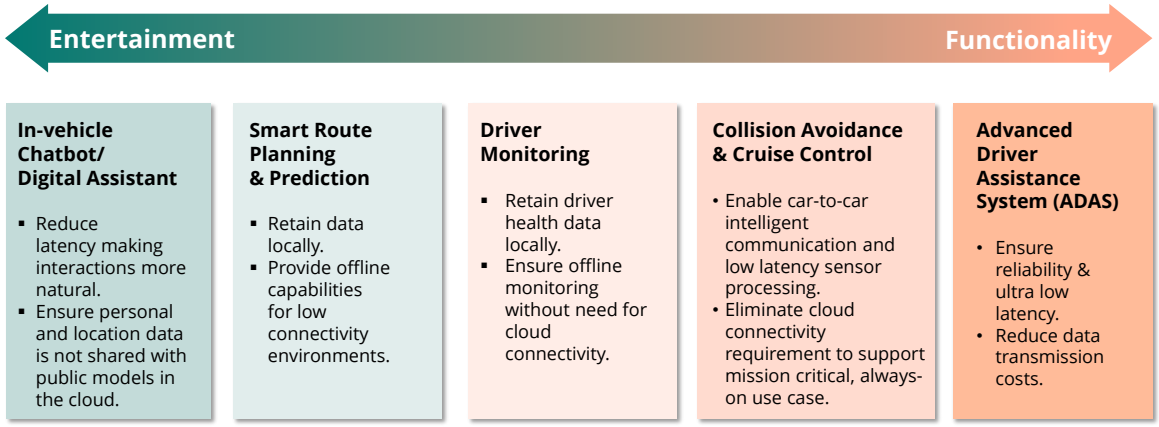
Like the smartphone market, unlocking growth in the stagnated PC market requires productive AI applications closely aligned with consumers and enterprise pain points with a clear ROI case based on time or money savings. For example, productive AI applications that target employee use cases like scheduling, note taking, copilot, and automated contract creation could save time every day. Assuming these applications save 30 minutes, and the employee costs US$20/hour, this productive AI application will translate into daily savings of US$10, and yearly savings of US$2,210 per year (based on a 231-day working year). Assuming that enterprises reinvest these savings into devices capable of running on-device generative AI, it will reduce refresh rates from, on average, 4 to 5 years to less than 2 years. This will stimulate massive growth in the PC market. For vendors to unlock this, they must target a productive AI strategy. In Section 5, ABI Research explores how to develop this through software innovation and stakeholder engagement.

AUTOMOTIVE

Unlike other devices, AI is already natively deployed in vehicles supporting a variety of use cases, including self-driving functionality, which runs MV models locally to understand, interpret, and respond to incoming sensor data. Subsequently, innovation is focused on upgrading native AI capabilities to support larger, more complex, and performant models (e.g., LLMs) to enhance applications and use cases across entertainment and functionality like Telsa's Grok AI that will provide an onboard intelligent assistant for drivers. These use cases are highlighted in Figure 7.

Figure 7: Automotive On-Device AI Use Cases

(Source: ABI Research)



Given the device constraints, native processing will be necessary to protect user data, ensure quality of experience, enable mission-critical use cases, and alleviate associated connectivity costs. OEMs are already targeting entertainment-focused use cases. Opel, in partnership with Qualcomm, is integrating LLMs into the infotainment system of its new Corsa range to provide a contextually-aware and adaptive cockpit system that adjusts to driver preferences; and BMW and Mercedes are deploying chips to support generative applications in their infotainment systems.

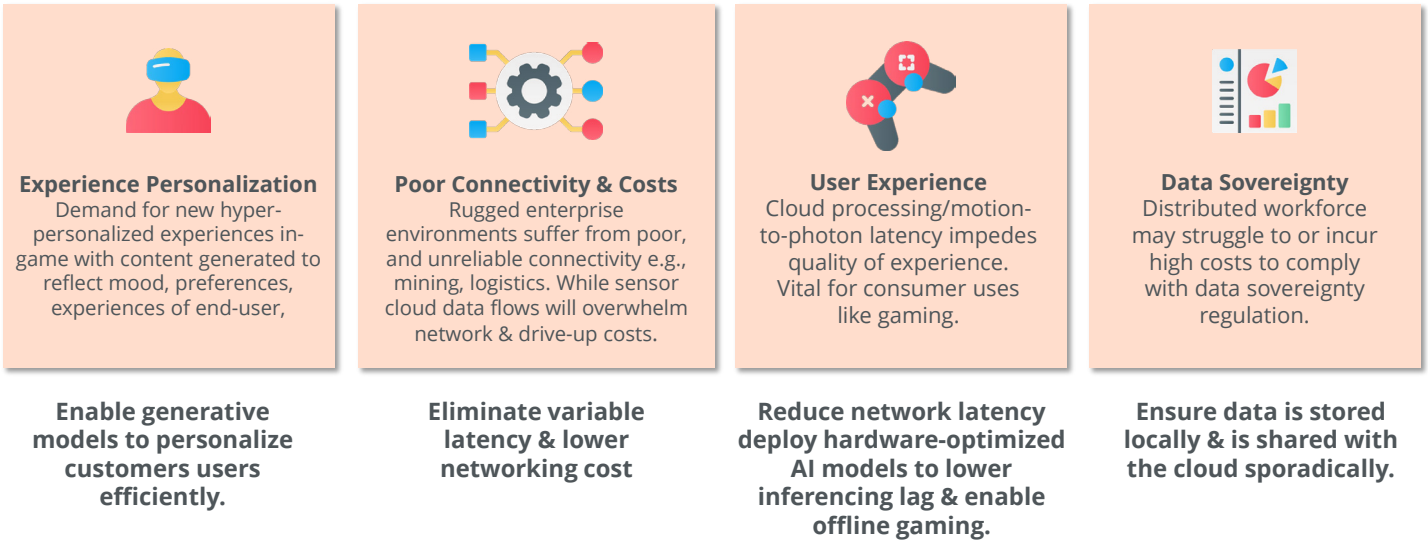
EXTENDED REALITY (XR) (INCLUDING AUGMENTED REALITY (AR), MIXED REALITY (MR), AND VIRTUAL REALITY (VR))

AI deployment will help transform XR in consumer (e.g., gaming/entertainment) and enterprise (e.g., manufacturing and telecommunications) use cases. Generative, NLP, and ASR models can offer a more immersive, productive, and intuitive experience for users enabling a range of new cases—metaverse image generation, 3D content generation, always-on digital assistant for factory workers, or automated form filling based on user prompts and sensor data. So far, most XR devices still rely on the cloud for running generative AI models and applications. But cloud processing will create challenges, and meeting user expectations will require on-device AI processing, as highlighted by Figure 8.

Meta has been a first mover by leveraging Qualcomm's Snapdragon processor to enable native processing for generative models in Quest 3.

Figure 8: Factors Driving On-Device Generative AI in XR

(Source: ABI Research)



In conjunction with local, on-device AI processing, XR will use a distributed workload processing framework that runs different workloads across connected devices, e.g., connecting smartphones. This will optimize processing speeds and power consumption.

Enterprise XR deployment has been relatively slow, but "killer" productivity AI applications supported by on-device generative AI processing could finally build a strong proposition for implementation in healthcare, manufacturing, logistics, and many more verticals.

INTERNET OF THINGS AND WEARABLES

The IoT is a network of devices that can sense and use trigger actions in response to certain conditions. These devices use simple rule-based AI, but use cases are increasingly emerging for powerful predictive or generative models deployed in the cloud. This brings challenges for certain deployment environments, including the following:

- **Billions of data points ingested that will congest networks and drive up associated costs.**
- **Deployments often in low/no-connectivity locations meaning unreliable cloud connectivity.**
- **Latency-sensitive applications that cannot rely on variable network connections.**

Deploying generative AI natively can mitigate these challenges and enable valuable enterprise use cases: predictive maintenance, problem management, lights-out manufacturing, traffic monitoring, inventory management/scheduling, and processing planning/optimization. However, given IoT compute and power constraints, this remains challenging.

Wearables can use generative AI models to transform incoming sensor data into insights and recommendations. Currently, these use simple graph-based or tabular predictive AI models to understand sensor data trends, so inference for some of these models could easily be executed on-device using Tiny Machine Learning (TinyML) frameworks. However, generative models that offer a significant opportunity for insight and recommendation personalization in the consumer sector can only be used by these devices in connection with cloud services. An example is WHOOP Coach, which uses the GPT-4 application to provide conversational and personalized responses to user inquiries. This still relies on smartphone connectivity to the cloud.

Running generative AI natively on wearables will be challenging, given the power and resource constraints. In most cases, wearables will distribute AI workloads across connected devices. This distributed AI compute framework leverages a multi-device approach to ensure that AI workloads are processed effectively. This will enable consumer and enterprise use cases ranging from patient care to spatial computing, and wearables could leverage active monitoring and real-time insights, while mitigating networking, latency, and privacy risks. On-device wearable generative AI remains in its very early stages with Meta and Qualcomm supporting AI applications on Ray-Ban glasses; Humane is working with OpenAI and Qualcomm to bring Ai Pin to market; and Zebra is demonstrating on-device generative AI powered by Qualcomm.

HOW WILL SOFTWARE INNOVATION SUPPORT ON-DEVICE GENERATIVE AI?

On-device generative AI hardware must be complemented with the development of productive AI applications. These applications need to offer cost or time savings sufficient to incentivize purchasing new devices; for example, AI applications that can save consumers time with their personal finances through smart bill management. These applications must be multimodal to enhance human accessibility and offer more intuitive interfaces for user engagement with AI. They should also be capable of running multiple models concurrently; for example, using voice control engines for translating voice to text for feeding generative AI requests and prompts.

HOW DOES SOFTWARE NEED TO EVOLVE?

Encouraging the development of productive AI applications will require greater developer accessibility. This is exemplified by historic market growth for web-based and smartphone applications. These were unlocked through the transition to visual coding and the introduction of SDKs/native programming languages, respectively. With this in mind, ABI Research has identified four areas of AI software that need attention from hardware vendors to promote the development of on-device "productive AI" applications:

- **AI Model Optimization Techniques:** The market must invest in Research and Development (R&D) targeting new and improved techniques like compression, quantization, model imitation, pruning, or distillation. Increasingly, existing techniques will struggle with LLMs given their intricacy and complexity. The goal should be to further reduce model size, improve power efficiency, and lower memory burden, while maintaining consistent accuracy; enabling more complex, performant models to be deployed on-device. Qualcomm's AI Hub offering a library of optimized models and its open-sourcing of the AI Model Efficiency Toolkit (AIMET) is a step in the right direction and will increase support for application development on device. Beyond optimization, one tool that will support enterprise applications with "small," targeted models is Retrieval Augmented Generation (RAG). RAG enables models to retrieve information from specific datasets without model retraining and with lower memory demands, which reduces model hallucination, improving reliability and accuracy.

- **Open Source:** Reduce barriers and enable innovation across software layers from models, tools, and applications. Generative AI models are already benefiting from open source with rapid improvements in the performance and deployment of "small" models like Meta's Llama (13B, 7B), Microsoft's Orca (13B) offering similar performance to GPT-4, and the Mistral-7B. By open-sourcing Machine Learning (ML) tools and datasets, start-ups/ISVs will have greater access to market-leading technologies supporting the development of productive AI applications optimized for on-device generative AI.

- **Unified Software Stack:** A comprehensive integrated set of software tools, frameworks, libraries, and technologies that can streamline application development and improve interoperability within an ecosystem. The Qualcomm AI Stack covers its mobile, automotive, XR, PC, IoT, and cloud platforms. It offers a full stack from integration with AI frameworks like PyTorch and TensorFlow to OSs. It enables developers to build applications and run them anywhere across different AI hardware (CPU, GPU, NPU).

- **SDKs:** Chip vendors must release SDKs to help accelerate the ecosystem of applications for on-device processing. SDKs lower developer barrier by reducing the time and effort required to optimize applications to new chip architectures. The Qualcomm AI Stack, which includes the Neural Processing SDK and AI Engine Direct SDK, is a good example.

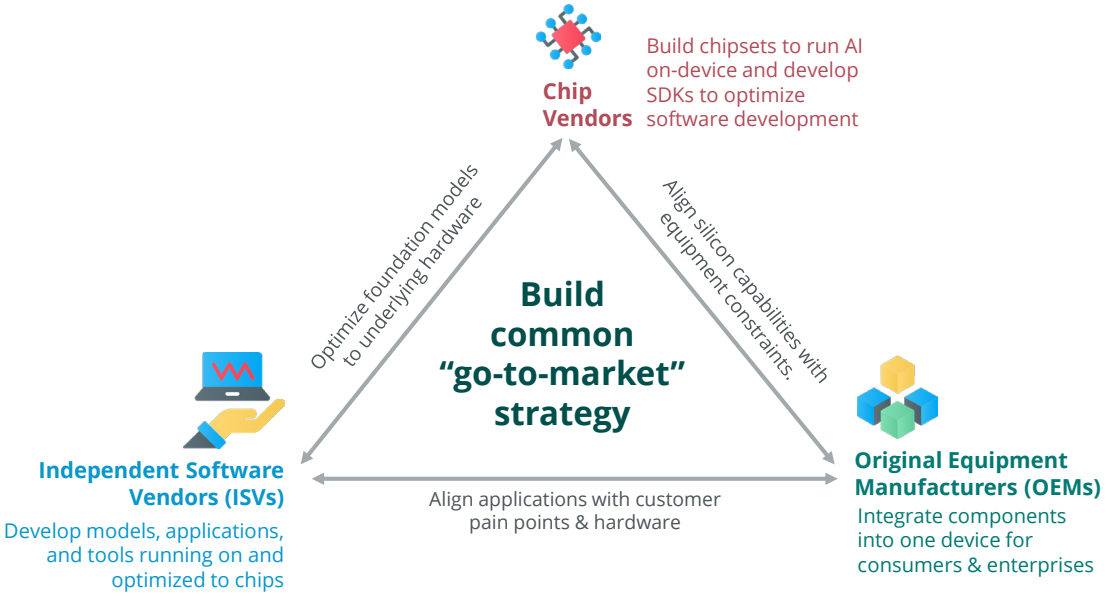
WHAT ROLE WILL STAKEHOLDERS PLAY?

A strong on-device generative AI go-to-market strategy should include energy-efficient, device-optimized chips, and small models supporting productive AI applications. Achieving this requires three stakeholders that need to

align closely on R&D and go-to-market strategy. Figure 9 provides an overview of this combined value proposition.

Figure 9: On-Device Generative AI Stakeholder Partnership

(Source: ABI Research)



Successful vendors will be vertically integrated (across devices, chipsets, and software) like Google or partner effectively to combine these components into a full-stack go-to-market strategy. Some notable partnerships have already emerged:

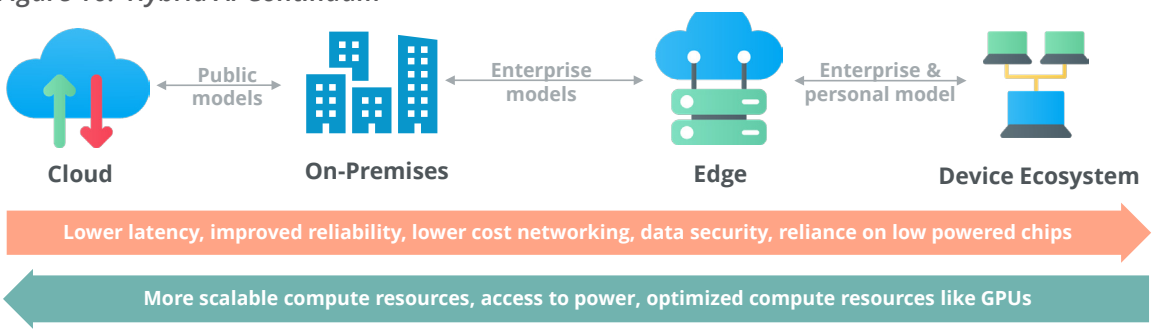
- Humane has partnered with Qualcomm's Snapdragon and OpenAI to develop Ai Pin.
- Meta has partnered with Qualcomm to build Meta Ray-Bans and Quest 3 (Snapdragon XR 2 Gen 2).
- Microsoft has partnered with Qualcomm's Snapdragon X Elite on Windows AI PCs.

WHAT IS HYBRID AI AND HOW DO WE GET THERE?

Consumers and enterprises will see significant value from on-device generative AI, but that is not the end of the story. The market will move toward integrating workloads from device to cloud within a hybrid AI framework.

A hybrid AI architecture distributes and coordinates AI workloads from device to edge to cloud depending on the use case's commercial (cost, governance) and technical (security, power usage requirements, latency sensitivity, model type/size) requirements. Figure 10 provides a blueprint for ABI Research's expectation of the hybrid AI framework.

Figure 10: Hybrid AI Continuum



Training will mostly reside in the cloud to maximize price to performance, given hardware and power constraints, but inferencing and fine-tuning workloads will be processed in the location that aligns with the constraints and requirements of each use case. This hybrid framework will support joint processing through which workloads are split up and run in parallel across a combination of devices or locations. For example, a digital assistant may run ASR locally to minimize latency, while generative models will be run at the edge or locally to lower device-level power consumption. This system design will spread power consumption, lessen memory-related workload bottlenecks, and optimize performance for each application.

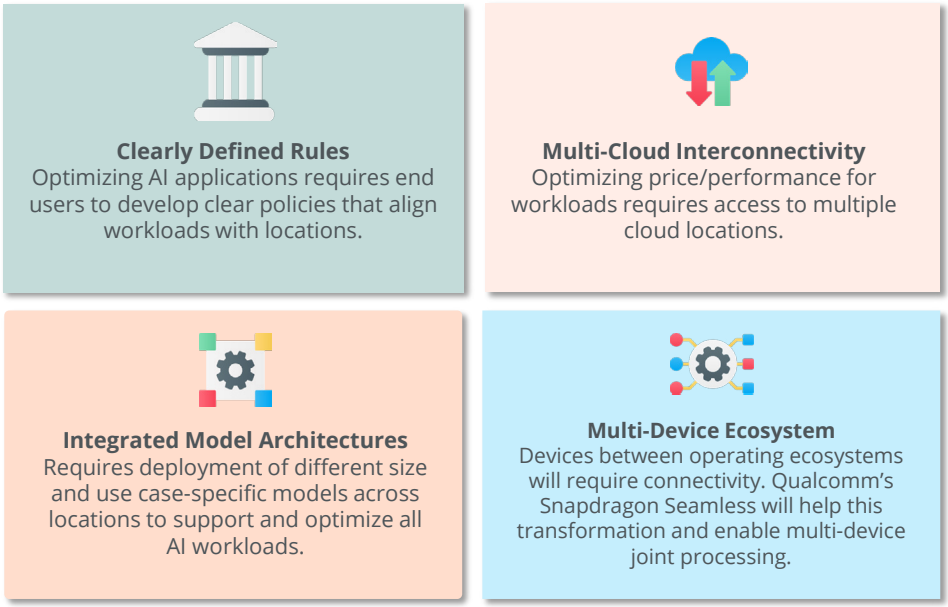
Initially, hybrid frameworks will use "rule-based" methods to align workload deployment with predetermined criteria. But moving forward, recommender models will be deployed to shift workloads automatically to understand where and when to place workloads depending on behavioral predictions, changing cost/performance expectations and application requirements. This framework will be able to understand when and where workloads should be processed, enabling improved "cloud economics" through batch processing.

As part of this hybrid AI framework, different models will be deployed that offer diverse features to end users. Very large public models will reside in the cloud and be used for generalized search functions; enterprise or private models will be deployed from edge to cloud and offer specialized functions tailored to internal datasets; and personal models will reside on the device and be tuned to users through ingested unstructured sensor data and prompts.

To move from on-device to hybrid AI, ABI Research has identified four fundamental building blocks, shown in Figure 11, aside from efficient on-device AI hardware, that need attention from the industry.

Figure 11: Building Blocks for Hybrid AI

(Source: ABI Research)

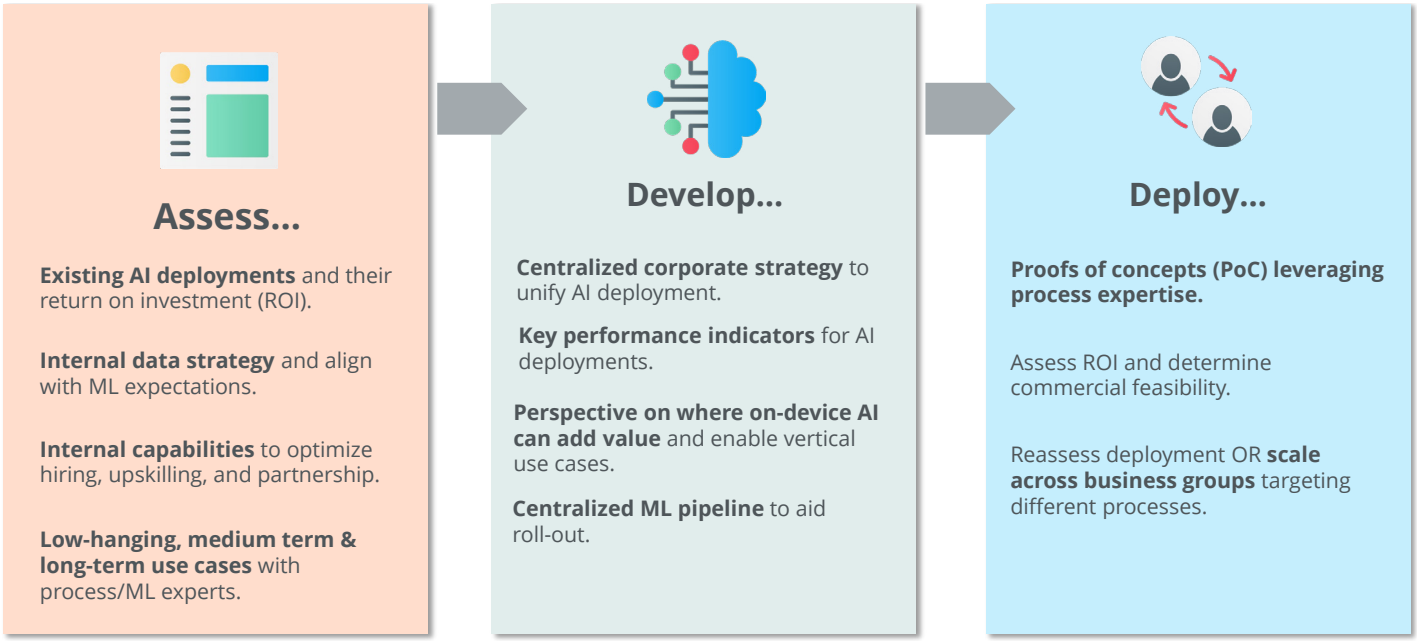


HOW SHOULD ENTERPRISES FUTURE-PROOF THEIR AI STRATEGY?

Enterprises already have fragmented AI deployments across business units and use cases with most lacking a future-proof, centralized strategy that guides effective implementation and value creation. In Figure 12, ABI Research provides a foundational framework to help enterprises start building a future-proof AI strategy.

Figure 12: Enterprise AI Strategy Breakdown

(Source: ABI Research)



Published May 2024

157 Columbus Avenue

New York, NY 10023

+1.516.624.2500

About ABI Research

ABI Research is a global technology intelligence firm uniquely positioned at the intersection of technology solution providers and end-market companies. We serve as the bridge that seamlessly connects these two segments by providing exclusive research and expert guidance to drive successful technology implementations and deliver strategies proven to attract and retain customers.