THE AI PC: WHAT YOU NEED TO KNOW ABOUT THE FUTURE OF COMPUTING AND AI

THE DEMOCRATIZATION OF AI IS ALMOST COMPLETE

EXECUTIVE SUMMARY

A flood of AI PCs will soon be released by the major computer companies. This paper discusses how AI PCs will impact future work environments and how users, developers, and content creators can use AI PCs for improved creativity and productivity. The paper also explores how AI PCs will drive the next stage of AI innovation to bring about the eventual democratization of AI.

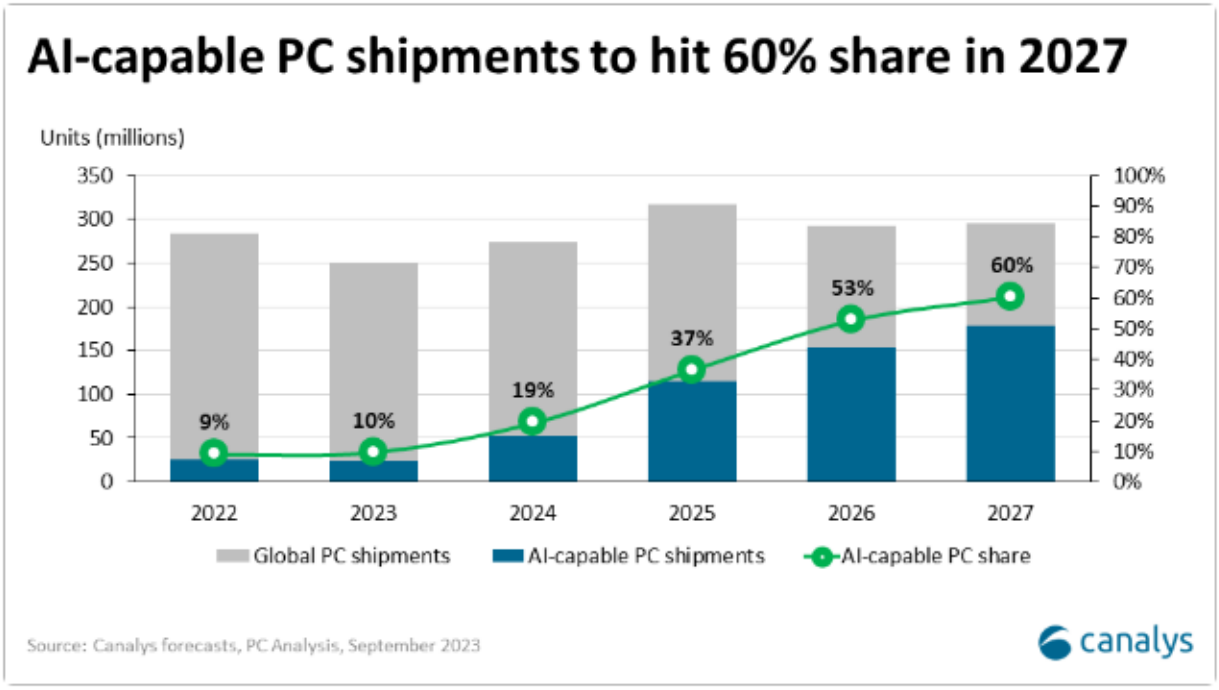
The concept of an AI-enhanced PC came from the need to run applications better on a PC than in the cloud. AI PCs have sophisticated designs and components capable of processing high-performance workloads much better than traditional cloud-dependent PCs. AI PCs use AI algorithms, neural processing units (NPU), and specialized CPUs capable of running large AI loads. Most AI PC models, depending on the application, will also come configured with a GPU. Integrating these components into an AI PC significantly enhances user experiences and provides greater capabilities.

AI PCs are being built using smaller and more efficient AI models. Microsoft's 2.7B-parameter small language model (SLM) Phi-2, discussed in detail later, is a good example of trending small AI models. There are a number of advanced compression[1] techniques being used to shrink model size. A few examples of these techniques include reducing number precision and eliminating redundant parts of neural networks.

Reduced model size is an important factor contributing to AI PCs' ability to train models and run inference accurately. Classic PCs are not equipped to train or run models. They are also hampered by limited bandwidth, the need for cloud connectivity, and security issues. AI PCs[2] offer the freedom to run AI without a cloud connection.

[1] Techopedia
[2] Research Paper: New Era of the PC

FIGURE 1: GROWTH IN AI PC SHIPMENTS 2022 TO 2027



Source: Canalys.com

Growth in AI PC sales over the next five years is expected to be significant. Canalys estimates that 25 million AI PCs were shipped in 2022. That number is expected to exceed 170 million in 2027, representing 60% of all PCs shipped that year.

Those figures indicate a fundamental shift in the market. The sales of AI PCs will almost double from 2023 to 2024 and double again from 2024 to 2025. Instead of buyers settling for CPU- and GPU-equipped PCs, we will see a market dominated by AI PCs equipped with a CPU, GPU, and an NPU. Adding the third component — the NPU — allows AI PCs to participate in every PC market segment, except for PCs costing less than $500.

Once sufficient numbers are deployed in the ecosystem, AI PCs will begin to drive innovations and create a major shift in how consumers and enterprise workers use AI PCs.

A BRIEF SURVEY OF THE AI LANDSCAPE

Although AI has been around for decades, its most significant advancements have been made in the past ten years. These include machine learning, deep learning, large-scale implementations, and, more recently, generative AI.

After OpenAI released ChatGPT in mid-2022, demand for generative AI exploded. Some companies used generative AI to initiate first-mover product enhancements, while others have used it defensively to counter competitive threats.

Intuit[3] is a good example of how AI is being used offensively and expansively across an entire product line. Intuit recently launched a new generative AI-powered assistant called Intuit Assist. It is a generative system that provides personalized assistance for 100 million customers. This major enhancement required years to develop.

Microsoft's major investment in OpenAI allowed it to become one of the first companies to incorporate generative AI into its products. Only a step behind Microsoft, Google countered Microsoft's actions by defensively augmenting generative AI into all its productivity and search offerings.

Both examples require some form of network connectivity for AI to function. Soon, AI PCs will allow similar powerful generative AI applications to be run locally without a network connection — a major step towards AI anywhere and freedom from being tethered to a network connection.

BRINGING AI APPLICATIONS TO PERSONAL COMPUTING

Moor Insights and Strategy expects AI PCs to continue to evolve for many years, largely to accommodate changes in AI.

Changes will be inevitable as next-generation AI models roll out over the next two to three years. New capabilities for both end-users and developers will supplement and enrich existing workflows rather than replace them.

Until then, first-generation AI PCs will be capable of processing and generating multimodal data[4] for text, imaging, speech, and video. Some of its most common applications will include chatbots, video creation and editing, fraud detection, and personal finance. Productivity enhancements will include such things as 3D rendering,

[3] Forbes Oct 2023
[4] Intel

image generation, numerous webcam technologies, and personalized AI — such as those provided by Microsoft Copilot as detailed below.

One of the most promising applications for AI PCs appears to be domain-specific[5] AI assistants. The AI PC's optimized AI software stack, combined with expected increases in hardware performance, will enable accurate, responsive compute experiences comparable to the performance of existing server-based models. Interactive AI agents will be able to function across a wide variety of consumer and business use cases where cloud connectivity is unavailable or is too slow or too expensive.

AI PCs appear to have been tailor-made for content creators and game developers. AI-powered PCs are expected to revolutionize content creation and gaming in many ways, including making these applications more personalized and engaging. According to a recent Forbes article, AI PCs will be able to use AI to follow game developments and analyze the competition's gaming techniques and how they are adapting to the game.

Games use non-player characters (NPCs) that are controlled by the game rather than players and perform routine tasks like giving instructions to players, assigning tasks, and selling game items. With AI, smart NPCs can be more adaptive and more responsive to varying game conditions. Neural networks can be used for personalized adaptation so that each player has an experience unique to them that maximizes engagement, fun, and game challenges.

Adobe, Microsoft, and Google have already incorporated Copilot-like AI into their products. However, most of these applications still require cloud access to function.

DISADVANTAGES OF LARGE-SCALE APPROACHES TO AI

Large-scale, cloud-based approaches to generative AI have significant drawbacks:

- Huge models consume large amounts of compute time, and money which prevents individuals and small organizations from accessing the power of generative AI.
- Large general-purpose models like GPT-4 are one-size-fits-all. It is more efficient and cost-effective to build and run smaller application-specific models, such as those that will be used in AI PCs.

[5] Voicebot.ai

- Security and privacy issues result from unnecessary exposure of proprietary data to the internet caused by data breaches and human errors originating from third-party administrators.

AI PCs will minimize or even eliminate many of these problems through smaller and more efficient AI models.

LIGHTWEIGHT AI MODELS AND APPLICATIONS

Google recently created three new AI models of different sizes optimized to run on everything from data centers to smartphones. The smallest and most efficient model, Gemini Nano6, runs on the Pixel 8 Pro smartphone and provides expanded AI features without a network connection.

Microsoft's Phi-2 small language model

Microsoft recently released a 2.7B-parameter SLM with exceptional reasoning and language understanding capabilities. It is similar to the type of AI models likely to be found in future AI PCs.

The Phi-2 is a transformer-based model that uses 1.4T tokens and took two weeks to train using 96 A100 GPUs. What is astonishing is that the Phi-2 outperforms the Mistral and the 7B and 13B Llama 2 models, as shown in Chart 1. It also outperforms the Google Gemini Nano as shown in Table 1.

TABLE 1: AVERAGE PERFORMANCE ON GROUPED BENCHMARKS COMPARED TO LLAMA-2 AND MISTRAL

Model	Size	BBH	Commonsense Reasoning	Language Understanding	Math	Coding
Llama-2	7B	40.0	62.2	56.7	16.5	21.0
	13B	47.8	65.0	61.9	34.2	25.4
	70B	66.5	69.2	67.6	64.1	38.3
Mistral	7B	57.2	66.4	63.7	46.4	39.4
Phi-2	2.7B	59.2	68.8	62.0	61.1	53.7

[6] Google December 2023

Source: Microsoft

TABLE 2: PHI-2 COMPARED TO GEMINI NANO 2

Model	Size	BBH	BoolQ	MBPP	MMLU
Gemini Nano 2	3.2B	42.4	79.3	27.2	55.8
Phi-2	2.7B	59.3	83.3	59.1	56.7

Source: Microsoft

Benchmark results for small models indicate we are very close to a complete democratization of AI. Democratization[7] will reduce barriers to entry for individuals and organizations and minimize the costs needed to build highly accurate models.

LEVERAGING COPILOT FEATURES FOR AI PCS

A large pool of applications could be created by rewriting existing cloud-dependent PC applications and making them available as AI PC applications. For example, the AI features provided with Microsoft Copilot that require cloud access could be optimized to run on an AI PC platform, resulting in lower latency and fewer concerns about privacy.

A signal of upcoming AI PC applications is the superpower application that provides meeting summaries, transcripts, email automation, contact management, relationships with conversation history summaries, and others. It runs Llama-2 and Whisper[8] entirely on the PC. The smaller Llama 2 7B can run on PCs with at least 8GB of RAM and a good CPU.

Having local AI compute[9] available on the AI PC platform will stimulate the creation of new AI software programs. Increased demand will provide independent developers and startups an opportunity to create and run new advanced algorithms for this market.

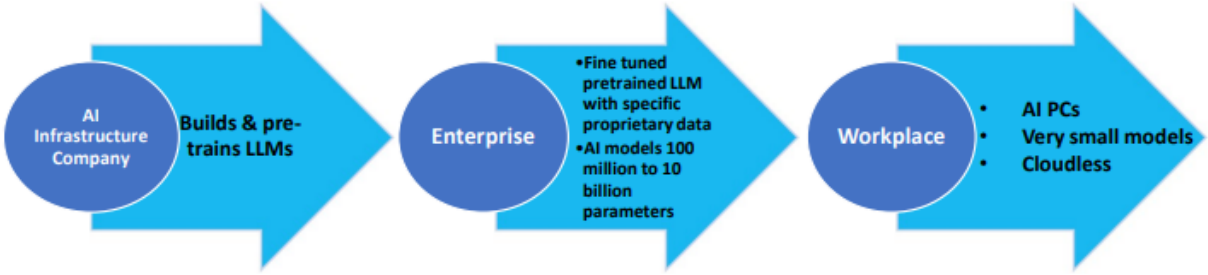
DRIVING THE ADOPTION OF AI PCS

Few companies can afford the resources needed to build large AI models. AI PCs will act as equalizers, allowing companies of all sizes to build and train models for a wide

[7] Turing.com
[8] Whisper and Llama- 7B
[9] Qualcomm June 2023

range of applications and use cases by eliminating infrastructure costs and the restrictions that go with them.

FIGURE 2: THE DEMOCRATIZATION OF AI



Source: Moor Insights & Strategy

AI PCs will allow small-to-medium sized companies to take full advantage of AI and at much lower cost points and with greater efficiency. There is already growing enterprise interest in leveraging AI PC platforms for process optimization, predictive analytics, and other digitization initiatives.

Independent software vendors (ISVs) will have opportunities to increase their revenue through new software licenses and subscriptions written for AI PCs.

Web-based applications using APIs will also need to be rewritten to work with AI-enhanced PCs.

It is important to note that the World Wide Web Consortium (W3C) is developing a low-level API[10] to enable in-browser local inferencing of neural networks similar to WebGL for graphics. The API is a powerful tool that will allow developers to run ML models on the client side without a server. It is being designed to work with CPUs, GPUs, and specialized accelerators such as Google TPUs

A wealth of new AI PC applications should become available from this group. A strong marketing case for AI PCs by Intel, AMD, and other AI PC companies will create a large wave of new business opportunities for ISV partners in the commercial and consumer segments.

Overall, democratization will help AI become more accessible, inclusive, and beneficial for everyone.

[10] W3.org

COMPELLING ECONOMICS OF EMBEDDED AI

Enterprise AI projects typically have high upfront costs, extensive data infrastructure, and specialized modeling expertise. Operational costs can run into the millions of dollars even before applications are deployed. Once trained, delivering AI capabilities to the users of large models still requires datacenter-scale resources.

Companies that choose to build large language models from scratch face other issues:

- Large models require specialized infrastructures equipped with expensive accelerators or dedicated AI chips to obtain an acceptable latency. This means an AI infrastructure must be built from scratch.
- AI toolchains are tied to specific cloud providers. Going all-in means a company will be locked into a vendor without the flexibility to switch in the future.
- Running AI on a vendor's cloud means relinquishing control over data, models, and operations.
- Increased costs — AI models are growing exponentially. The cost to continuously update, expand, and run new models in the cloud can rapidly become exorbitant.

For these reasons, more companies have begun to fine-tune[11] small open-source models for application-specific use because that approach is cheaper and small models requires less time to train and retrain.

This trend and its ease-of-use opens the door for widespread deployment of AI PCs, allowing companies of all sizes to trial AI models and build cost-effective proofs of concept much easier and cheaper than what would be possible with large cloud-based models.

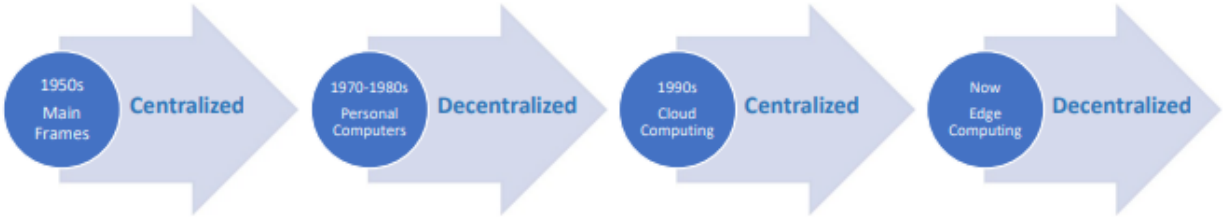
LONG-TERM VIEW OF AI PCS

Large Language Models (LLM) and generative AI models will evolve and grow larger as each generation ingests more data to gain additional capabilities. However, the cloud is the only place where most companies can build very large models, even though it is expensive and open to security issues.

[11] Redhat June 2023

As discussed earlier, the edge is a better place to run models for financial considerations and operational advantages created by decentralization that puts compute close to the data and the problem.

FIGURE 3: THE EVOLUTION OF COMPUTING



Source: Moor Insights & Strategy

Smaller AI models are more accurate, faster, and more economical. AI models with fewer parameters have less redundant knowledge, but that is offset by the depth of knowledge in the specific domain they are fine-tuned for. In addition, smaller models can be fine-tuned faster and retrained more often than huge models.

Although cost is always a consideration, AI cannot be democratized by relying on huge multi-modal general AI models. While there is a place for large models in AI, small AI models like the AI PC are needed to solve business problems requiring access to real-time private data**.**

Democratization of AI will be supported by millions of small generative AI models trained and inferenced on existing and future generations of silicon within enterprise IT infrastructures. These models will be embedded in a wide range of applications and systems designed to run at the edge.

AMD, Intel, and Qualcomm have all designed chips for AI PCs that will push democratization of AI. Intel launched the industry's first[12] AI PC with Core Ultra chips optimized for high-performance computing, data analytics, and other applications.

The first AI PC laptops began shipping at the end of 2023 equipped with a neural processing unit (NPU) and featuring low power consumption. Along with AI PCs, Intel designed a software ecosystem that includes engineering tools and resources for eventual use on more than 100 million PCs by 2025.

[12] PCmag.com October 2023

Intel's plan dovetails with the cyclic nature of how businesses traditionally replace tens of millions of outdated and obsolete laptops, workstations, and other devices every two to six years. As companies roll out new equipment over the coming years, integrating the turnkey acceleration and intelligence of the AI PC into new inventories provides a natural pathway toward operationalizing an AI-enhanced PC environment.

Pilot Programs for AI PCs

There are many AI prototype projects13 that information technology decision-makers (ITDMs) can implement to explore AI capabilities using local resources of an AI PC. However, prior to initiating AI pilots, it is important for managers to emphasize that these projects are focused on testing and scaling AI, not on replacing jobs. It is a good rule to begin with a well-defined project with accurately measured results before deciding to scale up the project.

- **Support chatbots** are among the easiest applications to create, implement, and test. Properly implemented, support chatbots[14] can provide instant responses to questions, send customer inquiries to the right department, and improve overall customer experience.
- **Documentation CoPilots** can greatly increase the accuracy, efficiency, and retrieval of information from massive corporate amounts of documents[15]. Humans cannot match AI's ability to perform contract reviews, invoice processing, or sorting through compliance issues.
- **AI prototypes for specialized data analysis** can be created for a variety of analysis categories such as AutoML, NLP, Data Visualization, recommendation systems, time series, video and image analysis, and many more.
- **SaaS/ERP systems**[16] are good candidates for AI pilot programs. AI has the potential to significantly provide improvements through automated self-healing, identification of performance issues, and providing valuable metric-driven insights. AI analyzes performance metrics, identifies bottlenecks, and predicts potential issues. It can also analyze complex data patterns that are too challenging for humans to understand.

[13] How to launch—and scale—a successful AI pilot project | CIO
[14] How to Test AI Chatbot Automation the Right Way
[15] Chatbot vs Copilot: Do you really need a copilot in your business? 2024
[16] AI in CRM and ERP systems: 2024 trends, innovations, and best practices

CALL TO ACTION

Moor Insights & Strategy believes that users will rapidly come to depend on AI PCs and on innovative AI-based applications capable of running on local resources. Moor Insights also believes AI PCs will come to be regarded as must-have compute devices, dispelling any notion that AI PCs are a passing trend.

Intel, along with its partner OEMs, have introduced AI-enabled PCs powered by the Intel Core Ultra processors. Qualcomm has made significant progress in AI with its Snapdragon X Elite chips optimized for AI workloads. AI PCs are expected to ship later this year.

AI-enhanced PCs' most important value comes from their ability to make AI immediate, responsive, and contextual. That is in direct contrast with PCs dependent on cloud-centric AI and limited by bandwidth, connectivity, and security.

AI PCs will eliminate many technical[17] barriers and allow software to locally interpret the complex and rapidly changing real-world environments.

Considering that AI PCs will have a large catalog of AI applications, increased accessibility for enterprises, and favorable embedded economics, they will have high user adoption and commercial success.

By democratizing and decentralizing AI workloads, companies using AI PCs will be able to create new opportunities across almost every business segment. Both large and small businesses will be able to drive transformative outcomes by taking advantage of opportunities offered by devices like the AI PC.

Every enterprise should now begin planning for the deployment of AI PCs and determining its hardware, software, and data requirements. It should also begin considering which AI pilot tests would be appropriate. AI PCs will create a new era in computing by putting the power of AI on desktops without the restrictions of being connected to the cloud.

[17] Harvard October 2021

IMPORTANT INFORMATION ABOUT THIS PAPER

CONTRIBUTOR

Paul Smith-Goodson VP & Principal Analyst AI & Quantum at Moor Insights & Strategy

PUBLISHER

Patrick Moorhead, CEO, Founder and Chief Analyst at Moor Insights & Strategy

INQUIRIES

Contact us if you would like to discuss this report, and Moor Insights & Strategy will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts but must be cited in-context, displaying author's name, author's title, and "Moor Insights & Strategy". Non-press and non-analysts must receive prior written permission by Moor Insights & Strategy for any citations.

LICENSING

DISCLOSURES

Intel Commissioned this paper. Moor Insights & Strategy provides research, analysis, advising, and consulting to many high-tech companies mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

DISCLAIMER

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Moor Insights & Strategy disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Moor Insights & Strategy and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Moor Insights & Strategy provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.