Get better NPU performance for AI inference workloads

Up to 2.5x higher Procyon AI Computer Vision Benchmark score

Get better CPU performance for on-device AI Chat models

Up to 66% less time for users to wait before seeing an LLM's output

Get better system performance for everyday activities

Up to 11.2% higher PassMark PerformanceTest 11 score

Whisper-quiet under load

30.5 dBA while running a resource-intensive Cinebench 2024 for 30 minutes

This project was commissioned by HP and AMD.

Meet AI challenges head-on with the HP EliteDesk 8 Mini G1a Desktop Next Gen AI PC

In our testing, this AMD Ryzen™ AI 7 PRO 350 processor-powered desktop provided better AI performance than Intel Core Ultra 7 processor-powered Dell and Lenovo desktops

This is a big year. Microsoft discontinued technical support for any computers running Windows 10, and AI adoption is a continuing trend across businesses. By investing in Windows 11 Pro AI PCs with processors that contain integrated neural processing unit (NPU) architecture as well as the usual CPU and GPU architecture, your company opens the door to a world of possibilities. NPUs are designed to accelerate AI inference, computer vision, large language model (LLM), and machine learning (ML) workloads.

To help you decide which AI PC is right for you, we compared productivity and on-device AI performance metrics on these desktops:

- HP EliteDesk 8 Mini G1a Desktop Next Gen AI PC powered by an AMD Ryzen™ AI 7 PRO 350 processor
- Dell Pro Micro Plus Desktop powered by an Intel® Core™ Ultra 7 265 vPro® processor
- Lenovo ThinkCentre M90q Gen 6 powered by an Intel Core Ultra 7 265T vPro processor

As AI use cases continue to unfold, system performance, especially for on-device AI workloads, is becoming more and more important. We found the AMD Ryzen™ AI 7 Pro 350 processor-powered HP EliteDesk 8 Mini G1a Desktop Next Gen AI PC is equipped to help your organization keep up in a rapidly changing business environment.

Our testing

All three Windows 11 Pro AI PCs we evaluated provided built-in AI capabilities, enhanced built-in Windows securities to plug potential pre-Windows 11 vulnerabilities[1], and NPU technology:

HP EliteDesk 8 Mini G1a Desktop Next Gen AI PC

- AMD Ryzen™ AI 7 PRO 350 processor (50 TOPS NPU[2], 8 cores, up to 5.0 GHz)
- AMD Radeon™ 860M graphics
- 64 GB of memory
- 512GB SSD

Dell Pro Micro Plus Desktop

- Intel Core Ultra 7 265 vPro processor (13 TOPS NPU[3], 20 cores, up to 5.3 GHz)
- Intel Graphics
- 64 GB of memory
- 512GB SSD

Lenovo ThinkCentre M90q Gen 6

- Intel Core Ultra 7 265T vPro Processor (13 TOPS NPU[4], 20 cores, up to 5.2 GHz)
- Intel Graphics
- 64 GB of memory
- 1TB SSD

*The results we report reflect the specific configurations we tested. Any difference in the configurations—as well as screen brightness, network traffic, and software additions—can affect these results. For a deeper dive into our testing parameters and procedures, see the science behind the report.

To measure productivity and on-device AI performance, we ran these benchmark tests:

- Geekbench AI
- LM Studio
- MLPerf Client Benchmark
- PassMark PerformanceTest 11
- Procyon® AI Computer Vision Benchmark

We also measured noise output while the desktops ran a sustained Cinebench 2024 workload.

Note: The graphs in this report use different scales to keep a consistent size. Please be mindful of each graph's data range as you compare.

Better cutting-edge and everyday performance

AI insights are transforming industries. Don't be left behind. By prioritizing on-device AI performance suited to your specific needs, you're investing in your company's potential. Getting answers in less time, while keeping control of your sensitive data on-device, is the first step on your road to success. But there's no one test that paints a broad picture of on-device AI and general productivity performance. That's why we examined CPU, GPU, and NPU performance from multiple angles.

Cutting-edge capabilities

To measure GPU performance for on-device ML workloads, we ran the Geekbench AI benchmark. This benchmark used real-world ML apps to provide a multidimensional picture of on-device AI performance.[5] The precision level scores below reflect different AI model requirements: Full Precision (FP32) is the most accurate and the most resource-intensive, Half Precision (FP16) is less accurate but more efficient, and Quantized (INT8) is the most resource-efficient and least accurate of all.[6] For this evaluation, we used the Open Neural Network Exchange (ONNX) open-source AI framework and DirectML AI backend for ML on Windows.

We found the AMD Ryzen™ AI 7 PRO 350 processor-powered HP EliteDesk 8 Mini G1a Desktop Next Gen AI PC outperformed both Intel Core Ultra 7 processor-powered AI PCs.

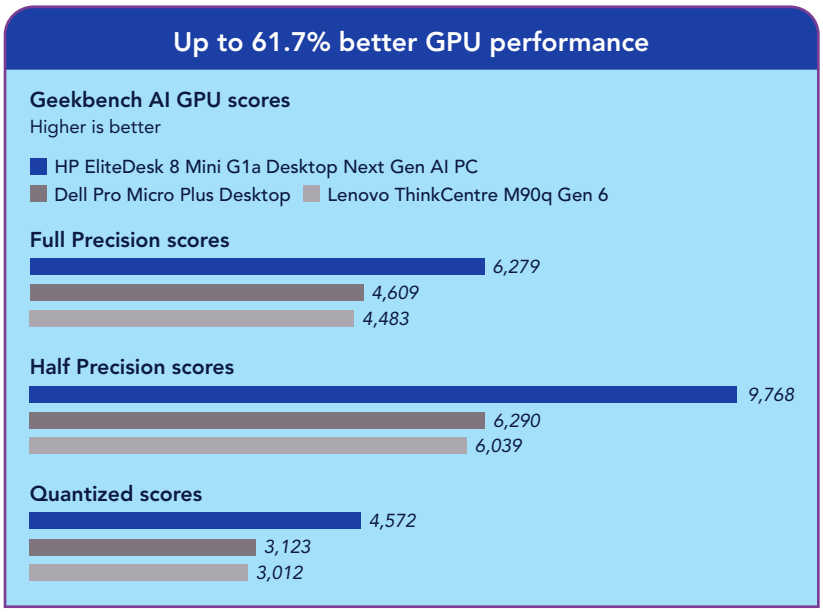

Figure 1: Geekbench AI GPU scores. Source: PT.

Ease your workday with a quiet desktop

The dream for any technology is that it "just works," running so smoothly that it melts away into the background and you can focus on the work it's enabling you to do. That's why it's valuable for your desktops to stay quiet, even when they're under a heavy workload. We measured the noise output of the desktops in this study and found that none of the systems exceeded 31 decibels (dBA) while running a resource-intensive Cinebench 2024 workload for 30 minutes. Given that a whisper-quiet library runs at 30 decibels, a difference of a few decibels is unlikely to cause problems for you.[7] Learn more in the science behind the report.

To measure CPU performance for on-device AI Chat models, we ran LM Studio, which uses local LLMs to capture token metrics.[8] According to Microsoft, LLM tokens are "words, character sets, or combinations of words and punctuation…"[9] In these tests, we ran the Meta-Llama-3.1-8B-Instruct-Q4_K_M model. This LLM predicts what the person typing the query is going to type next based on what they've already input.

When processing the local LLM, **the HP EliteDesk 8 Mini G1a desktop produced the first token and processed the data in significantly less time than the competitors.** The less time it takes an on-device AI chatbot to accurately figure out what you're looking for and produce a result, the less frustrating the user experience.

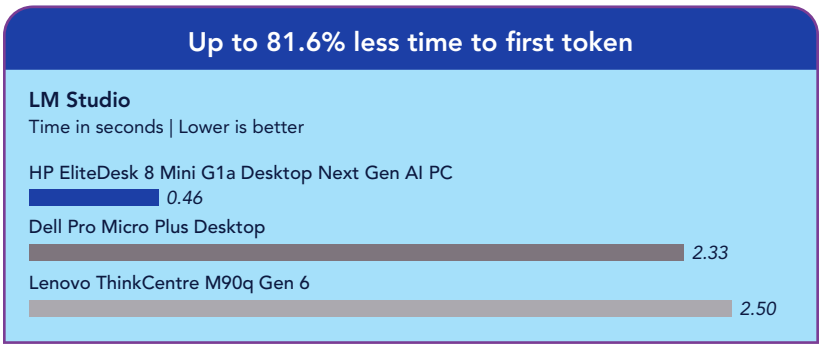

Figure 2: LM Studio results (time to first token). Source: PT.

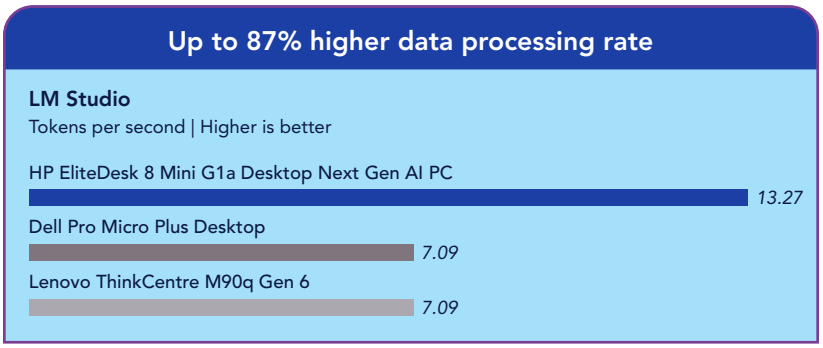

Figure 3: LM Studio results (tokens per second). Source: PT.

To measure NPU performance for AI inference engine models, we used the UL Procyon AI Computer Vision Benchmark.[10] In our integer-optimized (INT8) testing, we used the API optimized for each system's NPU: the AMD Ryzen™ AI API on the AMD processor-based system and the Intel OpenVINO™ inference API on the Intel processor-based systems. The individual AI inference tasks and their use cases were:

- **MobileNetV3, ResNet-50, and Inception-v4**: Convolutional neural networks (CNNs) widely used for image recognition, object detection, and image classification tasks. Essential for research institutions, tech companies, and individuals.[11,12,13]
- **DeepLabv3 and YOLOv3**: Deep neural networks (DNN) that distinguish between different objects and features within images and videos. Used by healthcare providers, manufacturers, and video surveillance companies.[14,15]
- **Real-ESRGAN**: A generator and discriminator network (GAN) that enhances image quality and resolution. Used by digital artists, medical professionals, and real estate firms.[16,17]

These results show the large gap between the overall scores for each AI PC. To see the average wait times for each inference engine on each AI PC, see the science behind the report.

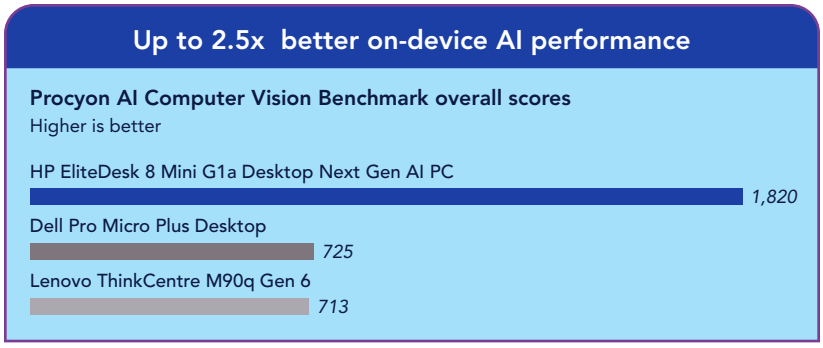

Figure 4: Procyon AI Computer Vision Benchmark overall scores. Source: PT.

About the HP EliteDesk 8 Mini G1a Desktop Next Gen AI PC

The HP EliteDesk 8 Mini G1a is a compact next-gen desktop that delivers flexible deployment options in a tiny footprint. This space-saving choice is also powered by AMD Ryzen™ processors with an up to 50 TOPS NPU.[18] Configurable with up to 64 GB DDR5-5600 MT/s memory and 2 TB PCIe® NVMe® storage, plus fast USB4, Wi-Fi 7, and built-in security via HP Wolf Pro Security, it's a sleek, space-saving choice for tidy workspaces that need next-gen AI power, reliability, and enterprise-grade protection.[19]

To measure NPU performance for on-device GenAI tasks, we used the MLPerf Client Benchmark to perform text and speech generation tasks. MLCommons says, this benchmark "provides clear metrics for understanding how well systems handle generative AI workloads."[20] Because each language-based workload has its own ideal use case, we tested performance with three LLMs:

- **Llama 2 7B Chat**: This chatbot-based model is fine-tuned for conversational dialog and instruction use cases[21] and used in AI-based document summarization and question answering scenarios.
- **Llama 3.1 8B Instruct**: This natural language processing (NLP) model is optimized for multilingual dialog and conversation use cases.[22]
- **Phi 3.5 Mini Instruct**: This NLP model is designed to accelerate long-context tasks, including meeting summarization, long document summarization and QA, and other document-based retrieval tasks.[23]

For these tests, we used the vendor-optimized execution path for each platform: ONNX Runtime GenAI with the AMD Ryzen™ AI SDK in hybrid mode (NPU + iGPU) on the AMD processor-based system and the Intel OpenVINO inference API targeting the Intel NPU on the Intel processor-based systems.

Once again, we found **the AMD Ryzen™ AI 7 PRO 350 processor-powered HP EliteDesk 8 Mini G1a desktop produced the first token and processed the data in significantly less time than the Intel Core Ultra 7 processor-powered AI PCs**. The less time it takes GenAI to summarize a doc or translate dialog, the quicker the user can get up to speed or understand the assignment.

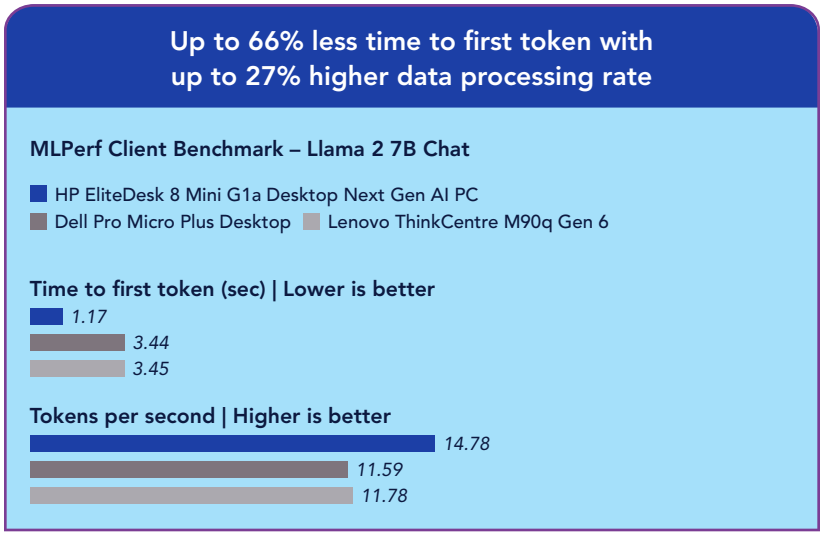

Figure 5: MLPerf Client Benchmark – Llama 2 7B Chat results. Source: PT.

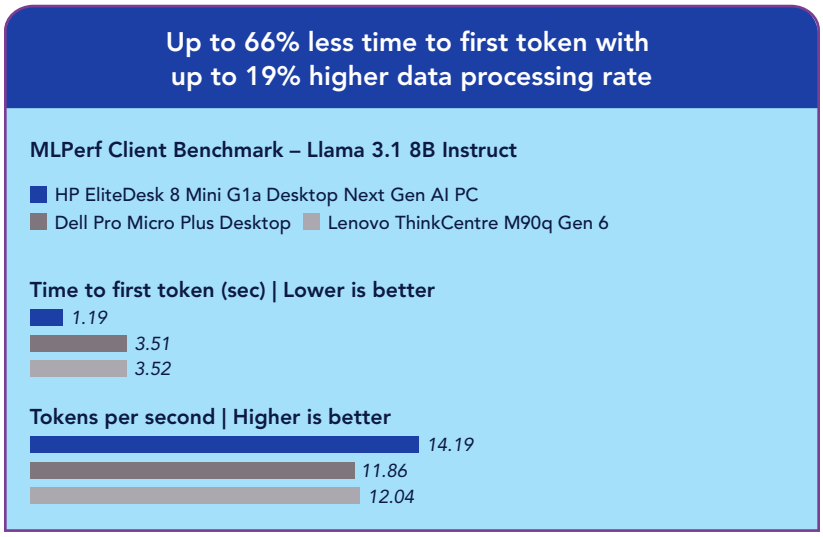

Figure 6: MLPerf Client Benchmark – Llama 3.1 8B Instruct results. Source: PT.

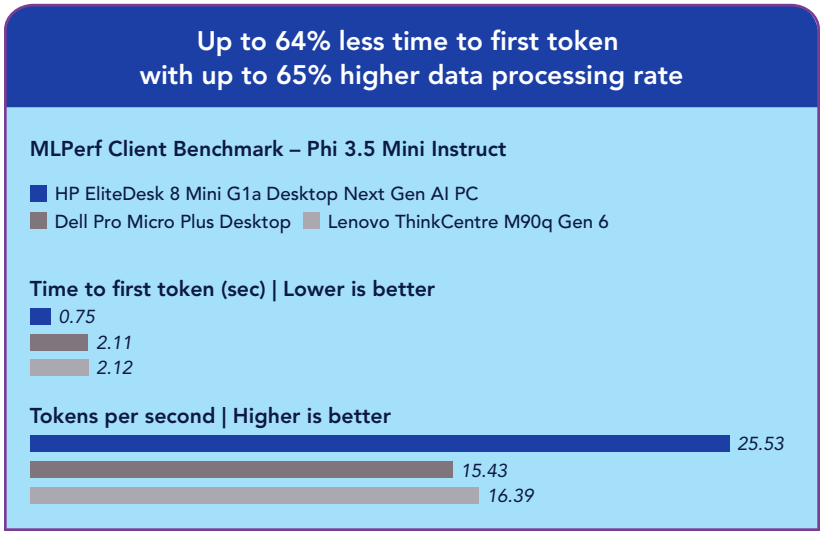

Figure 7: MLPerf Client Benchmark – Phi 3.5 Mini Instruct results. Source: PT.

Day-to-day capabilities

We also wanted to see how the AI PCs performed average everyday tasks. To do that, we ran the PassMark PerformanceTest 11 benchmark. PassMark PerformanceTest 11.0 runs CPU, video card, graphics hardware, and SSD tests and combines the results into an overall rating. This overall rating shows how the PCs handle complex mathematical calculations, 2D and 3D graphics, IOPS, and database operations.[24]

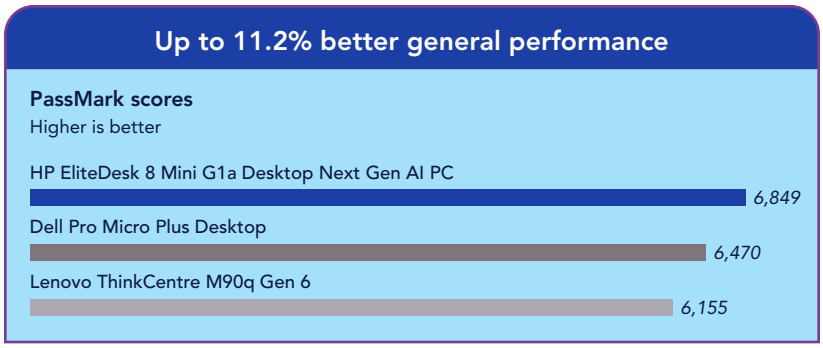

Figure 8: PassMark overall scores. Source: PT.

About the AMD Ryzen™ AI 7 PRO 350 processor

The AMD Ryzen™ AI 7 PRO 350 processor is built on AMD "Zen 5" architecture, delivering enterprise-strength computing plus serious AI acceleration. It has 8 cores and 16 threads, runs at a base clock of 2.0 GHz, and can turbo up to 5.0 GHz while keeping power consumption flexible (15-54 W depending on load). It includes integrated Radeon 860M graphics, supports fast DDR5 or LPDDR5X memory, and is capable of up to 50 TOPS NPU.[25] AMD Ryzen™ AI PRO processors are, according to AMD, purpose-built for "faster, smarter, and more efficient AI-powered workflows."[26]

Conclusion

In our hands-on tests, a HP EliteDesk 8 Mini G1a Desktop Next Gen AI PC with an AMD Ryzen™ AI 7 PRO 350 processor outperformed a Dell Pro Micro Plus Desktop with an Intel Core Ultra 7 265 vPro processor and a Lenovo ThinkCentre M90q Gen 6 with an Intel Core Ultra 7 265T vPro processor. This compact desktop delivered cutting-edge AI capabilities alongside strong general computing power, making it a solid AI PC choice for businesses seeking to harness on-device AI for enhanced productivity and data privacy.

1. Microsoft, "End of support for Windows 10, Windows 8.1, and Windows 7," accessed September 29, 2025, https://www.microsoft.com/en-us/windows/end-of-support.
2. AMD, "AMD Ryzen™ AI 7 PRO 350," accessed October 8, 2025, https://www.amd.com/en/products/processors/laptop/ryzen-pro/ai-300-series/amd-ryzen-ai-7-pro-350.html.
3. Intel, "Intel® Core™ Ultra 7 Processor 265," accessed October 8, 2025, https://www.intel.com/content/www/us/en/products/sku/241068/intel-core-ultra-7-processor-265-30m-cache-up-to-5-30-ghz/specifications.html.
4. Intel, "Intel® Core™ Ultra 7 Processor 265T," accessed October 8, 2025, https://www.intel.com/content/www/us/en/products/sku/241065/intel-core-ultra-7-processor-265t-30m-cache-up-to-5-30-ghz/specifications.html.
5. Geekbench AI, "Introducing Geekbench AI," accessed September 29, 2025, https://www.geekbench.com/ai/.
6. Vishalindev, "Understanding FP32, FP16, and INT8 Precision in Deep Learning Models: Why INT8 Calibration is Essential," accessed September 29, 2025, https://medium.com/@vishalindev/understanding-fp32-fp16-and-int8-precision-in-deep-learning-models-why-int8-calibration-is-5406b1c815a8.
7. "Decibel (Loudness) Comparison Chart," accessed October 8, 2025, https://www.hearingconservation.org/assets/Decibel.pdf.
8. LM Studio, "Model Catalog," accessed September 29, 2025, https://lmstudio.ai/models.
9. Microsoft Ignite, "Understanding tokens," accessed September 29, 2025, https://learn.microsoft.com/en-us/dotnet/ai/conceptual/understanding-tokens.
10. UL Solutions, "Procyon® AI Computer Vision Benchmark," accessed September 29, 2025, https://benchmarks.ul.com/procyon/ai-inference-benchmark-for-windows.
11. Activeloop, "MobileNetV3," accessed September 29, 2025, https://www.activeloop.ai/resources/glossary/mobile-net-v-3/.
12. Petru P., "What is ResNet-50?" accessed September 29, 2025, https://blog.roboflow.com/what-is-resnet-50/.
13. GeeksforGeeks, "Inception-V4 and Inception-ResNets," accessed September 29, 2025, https://www.geeksforgeeks.org/machine-learning/inception-v4-and-inception-resnets/.
14. Isaac Berrios, "DeepLabv3," accessed September 29, 2025, https://medium.com/@itberrios6/deeplabv3-c0c8c93d25a4.

15. Petru P., "What is YOLOv3? An Introductory Guide," accessed September 29, 2025, https://blog.roboflow.com/what-is-yolov3/.
16. Natsnoyuki AI Lab, "Upscaling images with Real-ESRGAN," accessed September 29, 2025, https://medium.com/@natsunoyuki/upscaling-images-with-real-esrgan-db579e9fb68d.
17. Maria Llain, "Restoring Image Quality With AI using Real-ESRGAN and SwinIR," accessed September 29, 2025, https://medium.com/@mariallain/restoring-image-quality-with-ai-using-real-esrgan-and-swinir-20d54c483e39.
18. HP, "Datasheet: HP EliteDesk 8 Mini G1a Desktop Next Gen AI PC," accessed October 8, 2025, https://h20195.www2.hp.com/v2/getpdf.aspx/c09153078.pdf.
19. HP, "Datasheet: HP EliteDesk 8 Mini G1a Desktop Next Gen AI PC.
20. MLCommons, "MLPerf Client Benchmark," accessed September 26, 2025, https://mlcommons.org/benchmarks/client/.
21. Hugging Face, "meta-llama/Llama-2-7b-chat-hf," accessed September 29, 2025, https://huggingface.co/meta-llama/Llama-2-7b-chat-hf.
22. Hugging Face, "meta-llama/Llama-3.1-8B-Instruct," accessed September 29, 2025, https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct.
23. Hugging Face, "Microsoft/Phi-3.5-mini-instruct," accessed September 29, 2025, https://huggingface.co/microsoft/Phi-3.5-mini-instruct.
24. PassMark Software, "PerformanceTest," accessed September 29, 2025, https://www.passmark.com/products/performancetest/?srsltid=AfmBOorBlay-PDlrE7WGgAeFXj_My4yjwSyKEv77SZfnx-CpZwz0bFkP9.
25. AMD, "AMD Ryzen™ AI 7 PRO 350," accessed October 8, 2025, https://www.amd.com/en/products/processors/laptop/ryzen-pro/ai-300-series/amd-ryzen-ai-7-pro-350.html.
26. AMD, "AMD Ryzen™ AI Pro Processors," accessed October 8, 2025, https://www.amd.com/en/products/processors/business-systems/ryzen-ai.html.

Read the science behind this report ▶

Principled Technologies®

Facts matter.®

This project was commissioned by HP and AMD.