

5 USE CASES FOR RUNNING AI *LOCALLY VS. IN THE CLOUD*

AMD Ryzen™ AI-powered PCs can run an impressive range of local models, thanks to a built-in NPU and integrated GPU (iGPU). Local inference can offer major advantages over AI in the cloud. When deciding where your application should live, consider that on-device inference is often best for data privacy, low latency, or offline capabilities.

1

KEEPING SENSITIVE DATA PRIVATE

When you run inference on the PC, data stays on the device. That's especially important if you or your end users are working with sensitive or regulated information, like financial or medical records. Users can help protect their personal data or stay in compliance with their organization's AI policies when working with confidential information and intellectual property.

2

PROVIDING LOW LATENCY FOR FAST RESPONSES

It takes less time to start processing a task on the PC than to send it to a cloud server and wait for results to come back. This can make a difference for latency-sensitive applications like live captioning, real-time translation, and image and video editing.

3

WORKING WITHOUT AN INTERNET CONNECTION

With on-device inference, users can still work even if internet connections are spotty or unavailable. Field service technicians can use local models to diagnose equipment or search technical manuals. Travelers can translate foreign languages while on public transportation. Healthcare providers can transcribe notes or analyze patient information at remote clinics.

4

SELF-HOSTING A MODEL OR APP

For tasks that are specialized and repeatable, fine-tuning a model can yield more consistent results than using a general-purpose model. You can fine-tune and host small or medium models directly on the PC. To get started fast, [AMD AI Playbooks](#) include ready-to-run apps that can also be fine-tuned.

5

USING HYBRID INFERENCE

Hybrid AI can offer the best of both worlds: fast, local inference on the PC and broader capabilities in the cloud. Users might generate drafts or previews on the PC, and then use the cloud for higher-quality refinement and rendering. Digital assistants may rely on a cloud service for basic functionality but use local processing when referencing files stored on the PC.

DEVELOPER TOOLS TO *BRING AI TO THE PC*

AMD Ryzen™ AI Software helps developers build or convert AI applications to run on PCs with select Ryzen processors and the AMD Ryzen™ AI Halo developer platform.

This software leverages the NPU and iGPU for efficient, low-latency inference and supports thousands of pre-trained models, including popular open-source LLMs like gpt-oss and Qwen.

Find resources at the AMD Ryzen AI Software Developer Hub.

[Developer Hub](#)