PROWESS

Research Abstract

Optimized Toolkits for LLM Deployment on AI PCs

In research commissioned by Intel, Prowess Consulting evaluated two hardware-optimized toolkits that enable developers to run large language models (LLMs) and AI applications locally.

Developers can now securely deploy models in local desktop environments, eliminating the need for external API and server connections. In addition to reducing costs, local development enables more efficient and flexible deployment of large language models (LLMs), supporting low-latency on-device inference while ensuring data privacy and compliance with security requirements.

Prowess Consulting compared the benefits and challenges of using AI acceleration and hardware-optimized tooling on Dell™ XPS™ 13 devices, powered by Intel® Core™ Ultra processors, and ASUS Zenbook® 14 devices, powered by AMD Ryzen™ AI 7 350 processors. We put the open-source Intel® Distribution of OpenVINO™ toolkit and the AMD Lemonade Server software development kit (SDK) to the test while deploying LLMs locally. This comparison is intended to help developers choose the optimal toolkit for their workflows.

Which Tools Are Best for Working with LLMs?

To find out which development environment came out on top, we created an AI pipeline for each hardware-specific SDK. Our findings indicate that the Intel Distribution of OpenVINO toolkit provides a more efficient and flexible workflow for developers deploying a broad range of LLMs on AI-enabled PCs. For details, read the technical research report.

The Intel Distribution of OpenVINO toolkit outperformed the Lemonade Server SDK in most of the categories we analyzed, making it a good choice for organizations working with LLMs, as shown in Table 1

Table 1. SDK scorecard: the Intel® Distribution of OpenVINO™ toolkit versus the Core ML® framework (using a five-star rating system, from 1 [poor] to 5 [excellent])

Software Development Toolkits	Target Hardware	Platform Compatibility	Model Conversion	Inference	Community Support
Intel® Distribution of OpenVINO™ toolkit	★★★★★	★★★★☆	★★★★★	★★★★☆	★★★★★
Lemonade Server SDK	★★★★☆	★★★☆☆	★★☆☆☆	★★★★☆	★★★☆☆

Primary drivers influencing our SDK ratings in Table 1 include:

- The Intel Distribution of OpenVINO toolkit enabled a straightforward, repeatable pipeline: convert the model, perform quantization (for example, INT8), and then package the optimized build with standard dependencies into a slim container that runs successfully without special workarounds.
- Lemonade Server SDK provided an easy graphical user interface (GUI) on-ramp, but it presented inconsistent model reliability and dev/Python® Package Index (PyPI) roadblocks (including gated neural processing unit [NPU] components, stalled and unsupported quantization, and extra containerization workarounds), which reduce its production readiness.

Lemonade Server NPU and Quantization Roadblocks on AMD Ryzen™ Processors

SOur research focused on optimizing LLMs on Windows® PCs powered by Intel Core Ultra processors and AMD Ryzen AI processors. We compared model conversion, inference, and deployment workflows of the Intel Distribution of OpenVINO toolkit and the Lemonade Server SDK to assess the practical challenges in each.

In our testing, we evaluated both the Lemonade Server SDK's GUI and developer workflows on an ASUS Zenbook 14 (powered by an AMD Ryzen AI 7 350 processor). The GUI offered a quick start—the default CPU chat model was responsive, and an NPU variant of Llama-3.2-3B ran (albeit slower than the CPU model). However, multiple curated or custom models failed to load or crashed.

In the developer path for the Lemonade Server SDK, we couldn't validate NPU use because a required ONNX® Runtime GenAI component wasn't available. Additionally, some NPU components are gated behind the AMD early-access program, requiring an AMD internal technical contact to grant access. CPU quantization stalled, GPU quantization wasn't supported on the tested AMD Radeon™ 860M GPU, and Windows Subsystem for Linux® (WSL) showed stability issues. Containerized deployment ultimately succeeded, but only after extra workarounds.

Key Takeaway

Our findings indicate that the Intel Distribution of OpenVINO toolkit provides a more efficient and flexible workflow for developers deploying a broad range of LLMs on AI-enabled PCs. For more information, read the complete technical research report.

Read the technical research reports:

The Intel® Distribution of OpenVINO™ toolkit vs. the Lemonade Server SDK

The Intel® Distribution of OpenVINO™ toolkit vs. Core ML®

The Intel® Distribution of OpenVINO™ toolkit vs. the Qualcomm® AI Engine Direct SDK

PROWESS

1025/250101