PROWESS

Research Abstract

Optimized Toolkits for LLM Deployment on AI PCs

In testing commissioned by Intel, Prowess Consulting evaluated the Intel® Distribution of OpenVINO™ toolkit and Core ML® to identify the best LLM deployment pipeline for Intel® and Apple® processors.

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 helping ensure data privacy and compliance with security requirements.

In testing commissioned by Intel, Prowess Consulting compared the benefits and challenges of using hardware-optimized tooling on a Dell™ XPS™ 13 AI PC with an Intel® Core™ Ultra 7 processor and on an Apple® MacBook Pro® with an M4 processor running macOS®. For both systems, we assessed model handling, quantization options, inference paths, and deployment. This comparison is intended to help developers choose the optimal toolkit for their workflows.

Which Tools Are Best for Working with LLMs?

To identify which toolkit delivers the best developer experience, we built an end-to-end pipeline for each. Our research focused on optimizing LLMs on Apple silicon–based Mac® devices using the Core ML® framework and on Windows® devices using the Intel® Distribution of OpenVINO™ toolkit. The Intel Distribution of OpenVINO toolkit outperformed Core ML in all 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	★★★★★	★★★★☆	★★★★★	★★★★☆	★★★★★
Apple® Core ML® framework	★★★★☆	★★★☆☆	★★★★☆	★☆☆☆☆	★★★☆☆

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, INT4 and INT8), and then package the optimized build with standard dependencies into a slim container that runs successfully without special workarounds.
- While Core ML handled model download, conversion, and weights-only INT8 and INT4 quantization (activations remained in floating-point), it required custom Swift®/Xcode® work for inference. There is no documented way to guarantee Apple Neural Engine (ANE)-only execution. As a result, ANE was not engaged in our tests.

LLM Build with the Intel Distribution of OpenVINO Toolkit

In testing the Intel Distribution of OpenVINO toolkit, we packaged an optimized (quantized) Llama®-3.2-3B INT4 and INT8 models together with the application and dependencies, built a container, and ran the solution locally. The build, load, and run steps were completed without requiring any special workarounds beyond the standard Docker® workflow, and deployment was successful.

Roadblocks with Core ML

In testing on an Apple silicon–based Mac, we followed Apple's example workflows and Llama-to–Core ML scripts to convert Llama-3.2-3B. Weights-only INT8 and INT4 quantization completed via coremltools (with activations left in floating-point).

For inference, Core ML requires Swift/Xcode integration. As a result, building a functional chatbot also required custom tokenization and string processing. Developers can request compute-unit preferences, but Core ML ultimately places operations based on compatibility and might fall back to the CPU or GPU. ANE-only targeting cannot be guaranteed.

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

PROWESS

1225/250101