Intel AI PC & OpenVINO Introduction

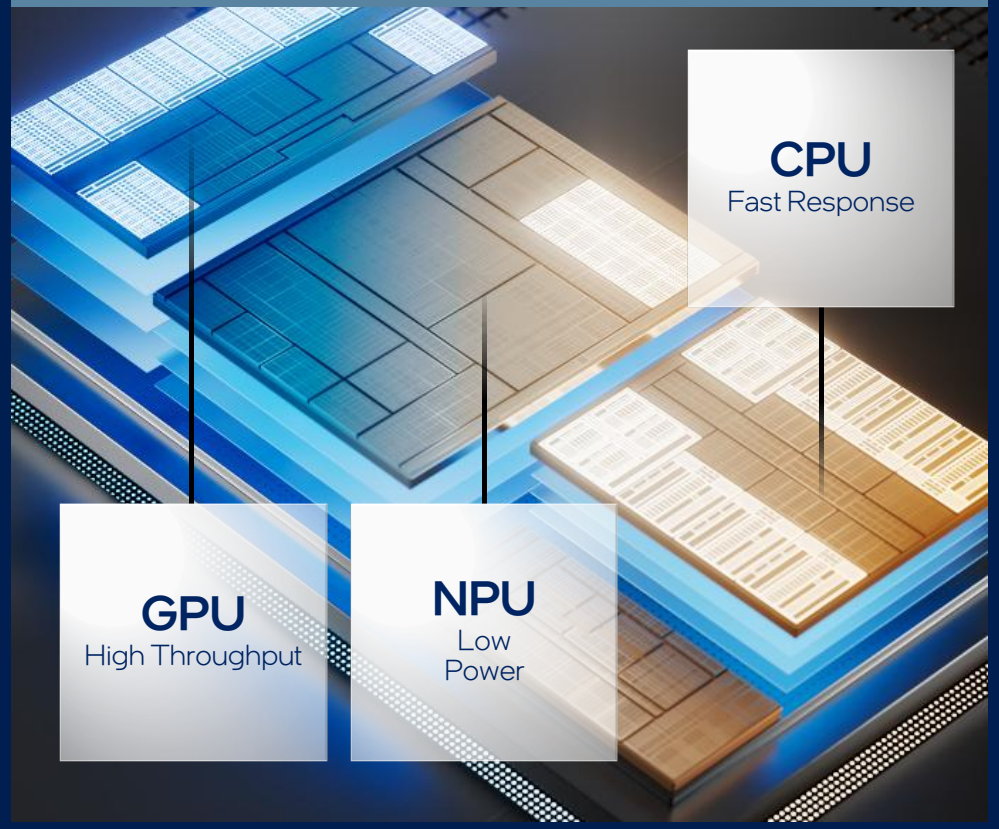
intel ai

Intel AI PC Updated

Intel AI PC

Now AI PC is for everyone

CPU + GPU + NPU



What is an AI PC?

A PC with new NPU silicon that brings new AI experiences in productivity, creativity, and security through a combination of the CPU, GPU, and the NPU.

Protect your data, model and analytics on premise while reduce cloud per token costs

Intel AI PC Roadmap

intel.

Intel® Core™ Ultra
200H series
(Arrow Lake)
for Performance Thin & Light

Intel® Core™ Ultra
200V series
(Lunar Lake)
for Premium Mobile

Intel® Core™ Ultra Series 2

Mobile Segment

intel.

Lunar Lake

Unmatched AI Compute

Up to 120 platform TOPS across CPU, GPU, and NPU

NPU	Up to **48 TOPS**	Dense vector & matrix math	AI assistants & creation
GPU	Up to **67 TOPS**	XMX & DP4a	Gaming & creator AI
CPU	Up to **5 TOPS**	VNNI & AVX	Light AI workloads

Breakthrough x86 Battery Life

	Intel® Core™ Ultra 200V Microsoft Surface Laptop 15″ with Intel Core Ultra 7 268V	Qualcomm Microsoft Surface Laptop 15″ with X Elite-X1E-80-100
Microsoft Teams Microsoft Teams 3x3 with WSE Battery Life	up to **10.4 hours**	up to **8.7 hours**
Microsoft 365 UL Procyon® Battery Life Office Productivity	up to **19.7 hours**	up to **16.3 hours**

Based on measured battery life of comparable PCs with following battery capacities: Microsoft Surface Laptop with Intel® Core ™ Ultra 7 268V– 64Wh; Microsoft Surface Laptop with Snapdragon X EliteX1E-80-100 – 64Wh. System performance will vary significantly with different battery capacity, screen type and other OEM design factors. See intel.com/performanceindex for additional details

intel.

Elite AI Performance

For 447 AI Features and Counting

OpenVINO · ONNX · WindowsML DirectML · W3C WebGPU WebNN

Up to **99** platform TOPS

Intel Core Ultra 200H Series vs.AMD AI 365 (73 platform TOPS)

Up to **77** int8 TOPS	intel ARC	XMX & DP4a Instructions	Sized for gaming and creator AI	**GPU** Powering 40% of AI features in '25
Up to **13** int8 TOPS	**NPU 3**	MAC Arrays	Sized for assistants, webcams, and SLMs	**NPU** Powering 30% of AI features in '25
Up to **9** int8 TOPS	**Skymont Lion Cove**	AVX - VNNI	Always available and low latency	**CPU** Powering 30% of AI features in '25

See intel.com/AIPC for additional details.

Leadership AI Performance

Unlocking the power of AI with Intel® Core™ Ultra

Procyon AI CV Computer Vision, Higher is Better		Intel® Core™ Ultra 7 258V Vendor Preferred Framework OpenVINO	Intel® Core™ Ultra 7 265H Vendor Preferred Framework OpenVINO	AMD AI 7 350 Vendor Preferred Framework ONNX & Ryzen AI	Qualcomm 78-100 Vendor Preferred Framework SNPE
GPU	Int8	1311	1126	56	DNR
	FP16	853	743	292	DNR
	FP32	322	317	199	DNR
NPU	Int8	1830	700	1834	1666
	FP16	1031	380	DNR	DNR
Procyon AI Image Generation Stable Diffusion 1.5, Higher is Better		**Intel Core Ultra 7 258V**	**Intel Core Ultra 7 265H**	**AMD Ryzen AI 7 350**	**Snapdragon X1E-78-100**
GPU	Int8	2315	2020	DNR	DNR
	FP16	379	324	143	DNR
NPU	Int8	3391	842	DNR	DNR
Geekbench-AI-1.2.0 Higher is Better		**Intel Core Ultra 7 258V**	**Intel Core Ultra 7 265H**	**AMD Ryzen AI 7 350**	**Snapdragon X1E-78-100**
GPU	FP16	24568	20205	9906	DNR
	FP32	10370	10029	6569	DNR
NPU	Int8	27886	12879	DNR	26183
	FP16	19719	8713	DNR	15125
Procyon AI Text Higher is Better		**Intel Core Ultra 7 258V**	**Intel Core Ultra 7 265H**	**AMD Ryzen AI 7 350**	**Snapdragon X1E-78-100**
GPU	Microsoft Phi 3.5	1005	727	355	DNR
	Llama 3.1	879	610	208	DNR

DNR = does not run; as of May 2025.

See intel.com/performanceindex for details. Results may vary.

intel.

OpenVINO : Framework To Boost AI Performance on Intel AI PC

Open-Source Software for AI Inference Optimization and Deployment

Google Summer of Code

open source initiative®
Apache 2.0

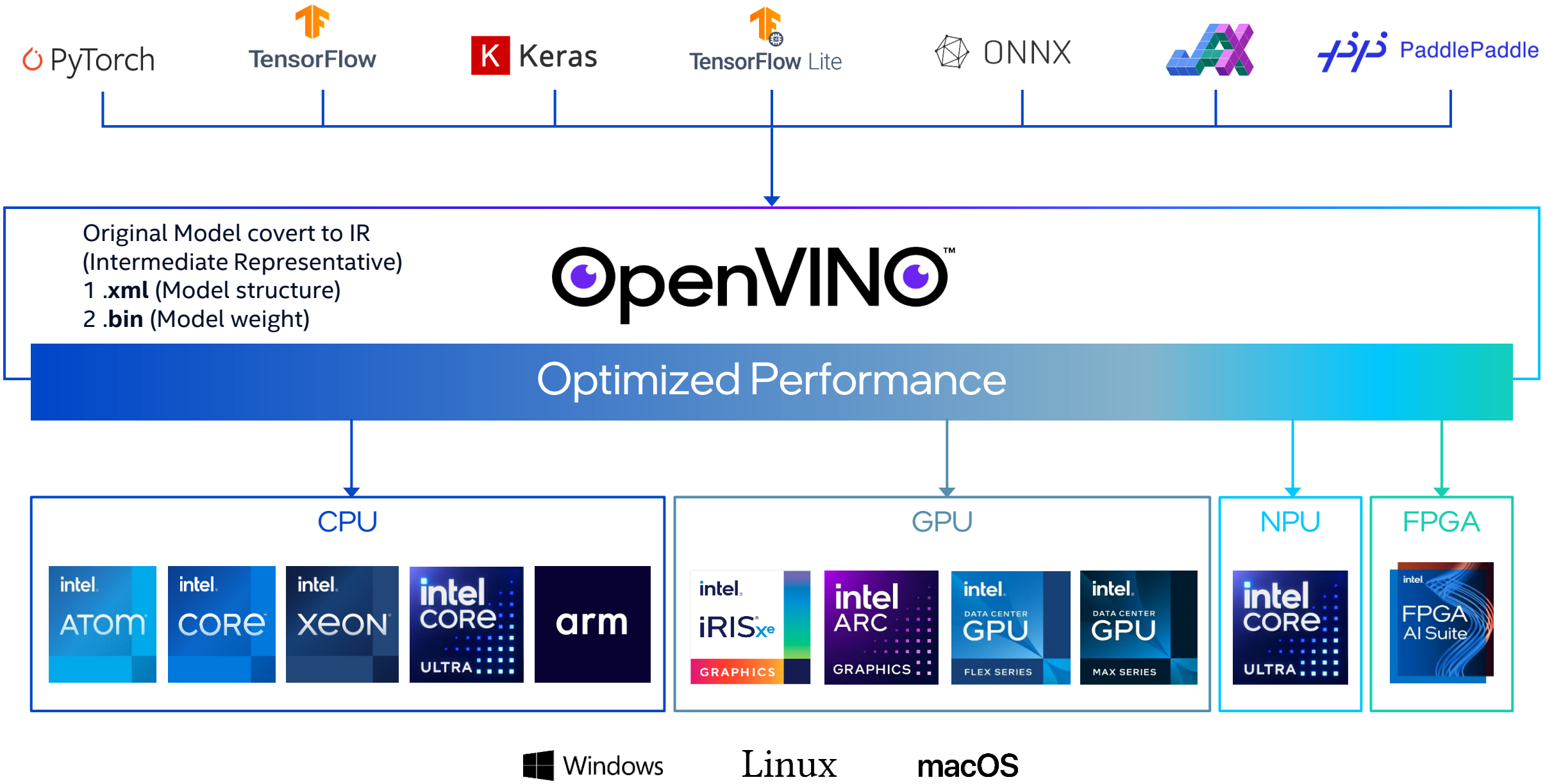
OpenVINO™

- Visual + NLP, Audio, LLM, GenAI, Transformers...
- Inference
- and Neural Network
- Optimization

Powered by oneAPI

OpenVINO™

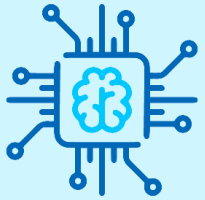
Developer Journey


PyTorch
TensorFlow
Keras
TensorFlow Lite
ONNX
PaddlePaddle
Original Model covert to IR
(Intermediate Representative)
1 .xml (Model structure)
2 .bin (Model weight)
OpenVINO™
Optimized Performance
CPU
GPU
NPU
FPGA
Windows
Linux
macOS

Benefits of Building Applications with OpenVINO™

Build and deploy AI applications in simple steps

Faster inference speed



Maximize AI performance across CPU, GPU, NPU

Smaller model and binary size

Reduce memory footprint

Ability to scale to many nodes with serving

What's New in the 2025.3 Release?

1 Model

- New models supported: Phi-4-mini-reasoning, AFM-4.5B, Gemma-3-1B-it, Gemma-3-4B-it, and Gemma-3-12B.
- NPU support added for: Qwen3-1.7B, Qwen3-4B, and Qwen3-8B.
- LLMs optimized for NPU now available on OpenVINO Hugging Face collection.

2 Optimize

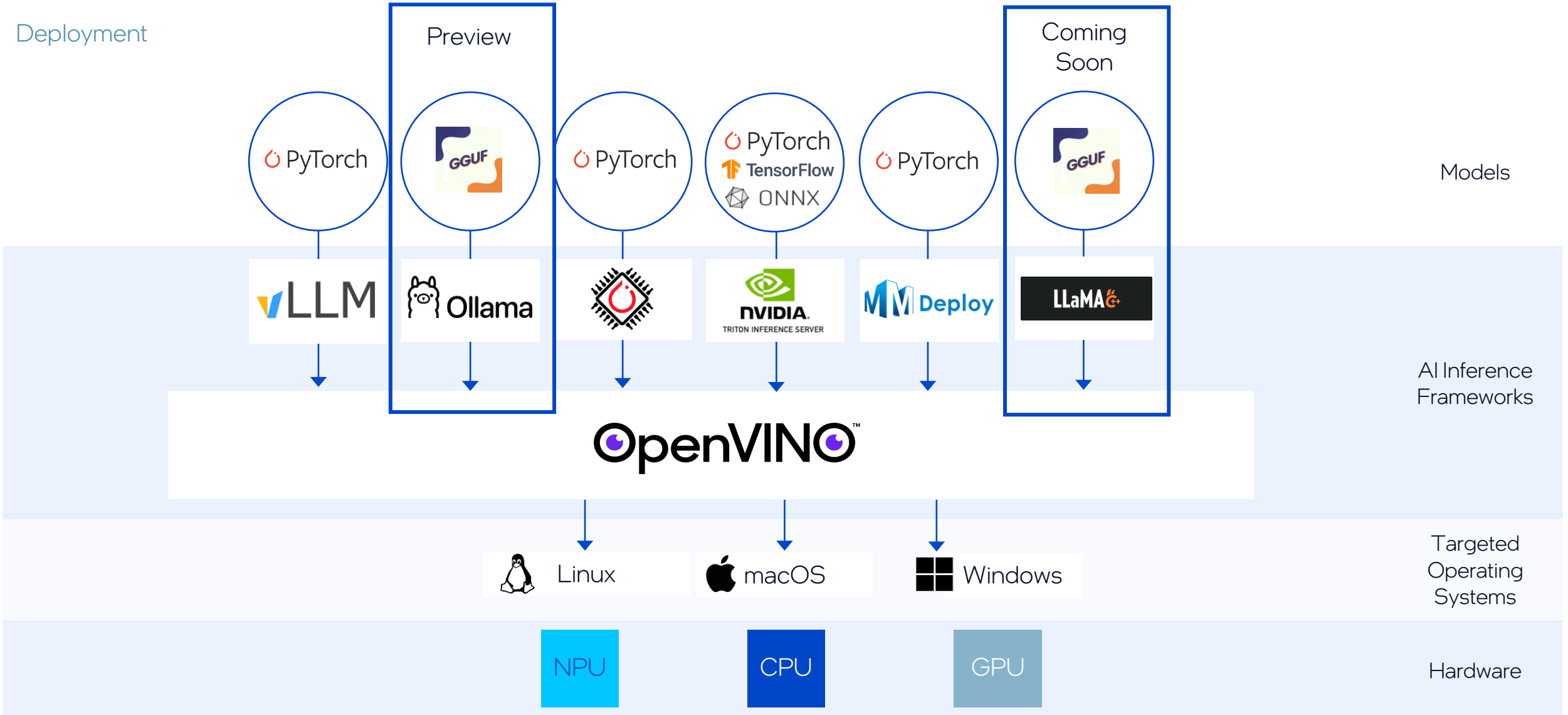
- The NPU support for dynamic batch sizes by using a batch size of 1 and concurrently managing multiple inference requests.
- Key cache compression per channel technique for accuracy improvements for GenAI models on GPUs.
- TextRerankPipeline and Structured Output in OpenVINO™ GenAI.

3 Deploy

- Support for Intel® Arc™ Pro B-Series (B50 and B60).
- The OpenVINO™ Model Server boosts support for agentic AI use cases.
- int4 data-aware weights compression for ONNX models in NNCF.

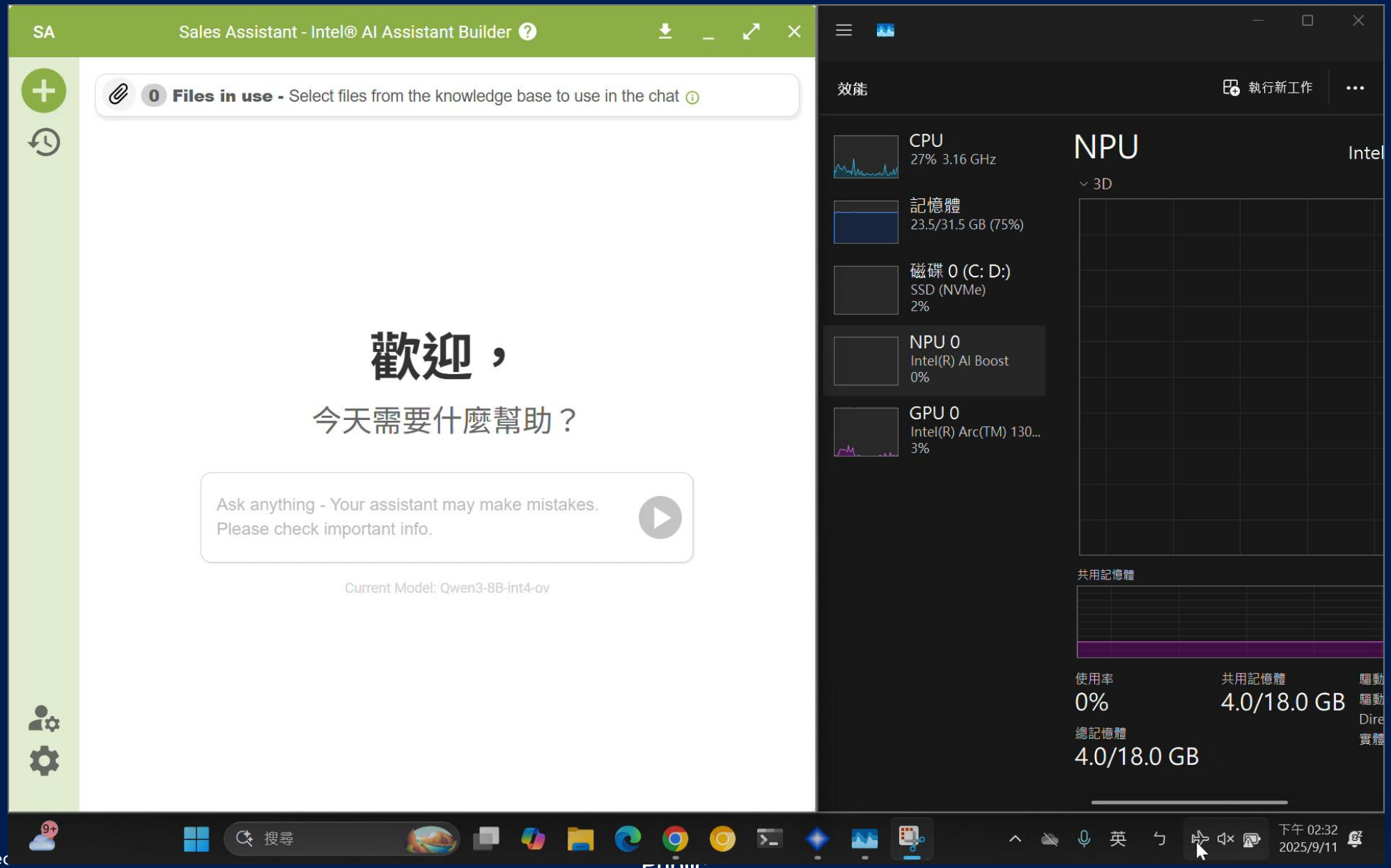
OpenVINO™ as Backend

Deployment

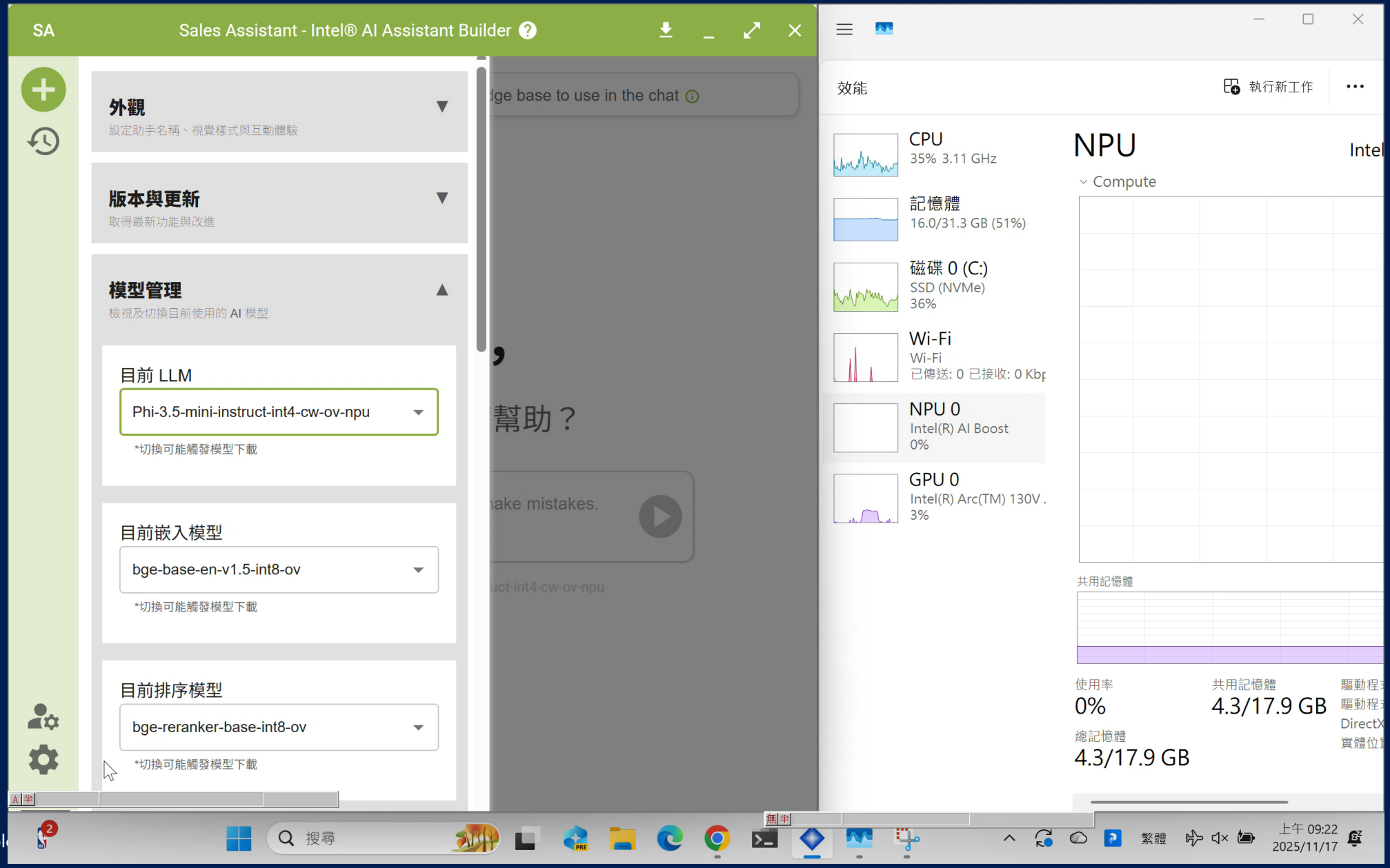


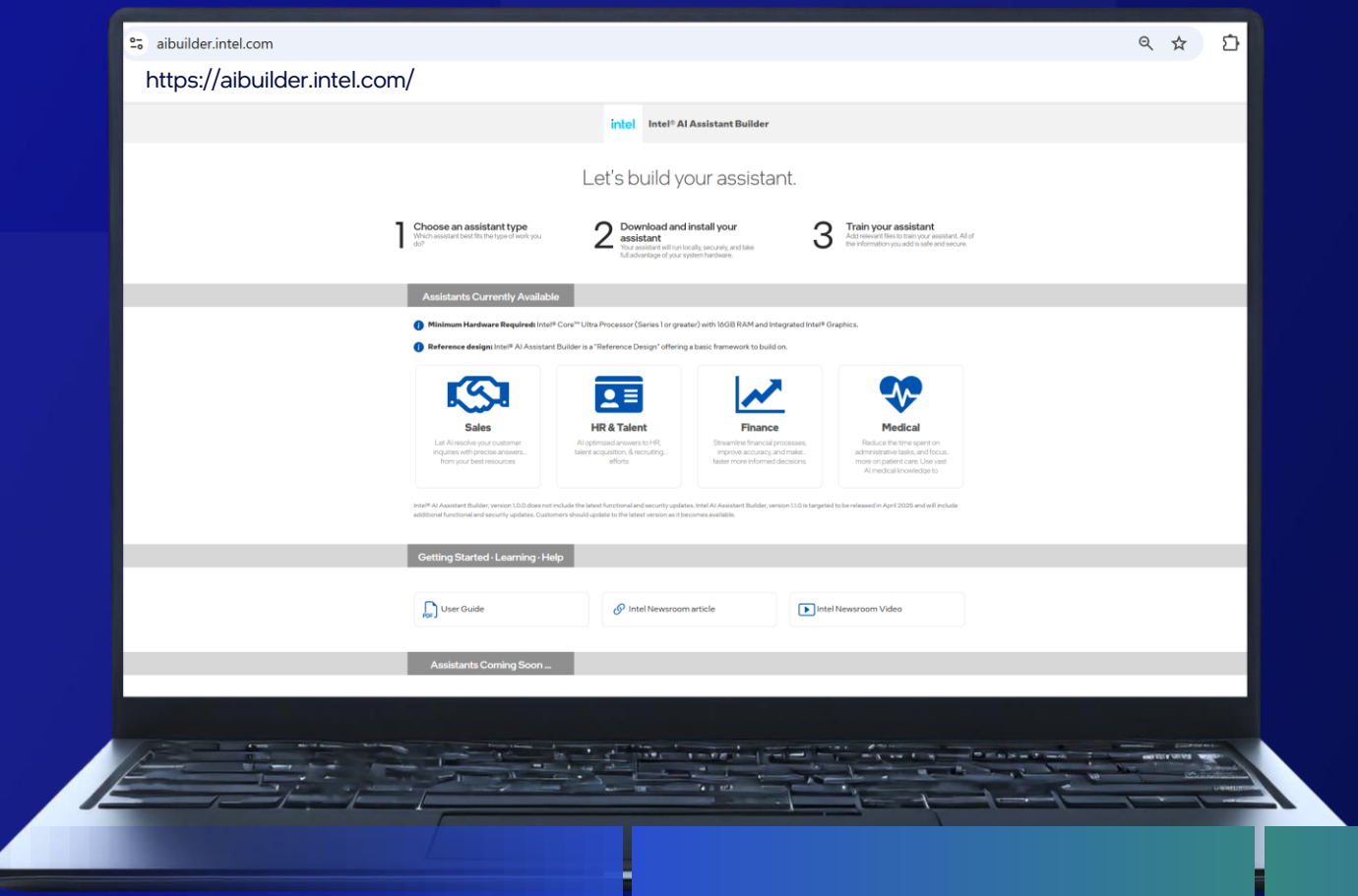

Let’s See How AI Inference
Accelerated on Intel AI PC
with OpenVINO Optimization

Demo: LLM inference on Intel AI PC (Lunar Lake-iGPU)



Demo: LLM inference on Intel AI PC (Lunar Lake-NPU)





Intel® AI Assistant Builder

(SuperBuilder)

Reduce development and deployment **from months to days**

https://github.com/intel/intel-ai-assistant-builder

AI Agents for PC

Multi-Agent Orchestration
Agent with MCP
(Model Context Protocol)

Highly Optimized

LLM Based on System/Work Type
Auto-Model Quantization & Optimization

Custom Workflows

RAG+, Tabular Data, Doc. Scoring, etc.
Model Choice, Hyper Parameter Tuning

Ease-of-Use

Drag & Drop, Simple UI, API
Tune and Customize for YOUR Work

intel

New Line Up of Discrete GPU
Intel® Arc Pro B Series : B50

Introducing intel ARC PRO B50

Bigger Projects

16GB VRAM
128 XMX Engines

Compact Efficient Design

Energy Efficient
Dual Slot Compact Form Factor

Workstation Software

Consumer & Pro Drivers, ISV Certified
Windows & Linux Ready

16
Xᵉ-cores

16GB
Memory

224 GB/s
Memory Bandwidth

170
Peak TOPS[1]

70W
Total Board Power

Gen 5
PCIe (x8)

1: GPU Peak TOPS (trillions of operations per second) represents the peak throughput when running XMX workloads with INT8 datatype and dense models. Performance may vary based on configuration.

Designed for Pro Workloads

Adobe

Ansys
STRATEGIC PARTNER

AUTODESK

SIEMENS

Bentley

SOLIDWORKS*

chaos

ptc

MAXON

D5 RENDER

VECTORWORKS

UNREAL ENGINE

blender

50+ Common ISV
Apps Regularly Certified and Validated

Reliability

Stability

Compatibility

To see the full list of ISV Certified applications visit: intel.com/support/CertifiedGraphics

intel ARC

*A Dassault Systèmes brand

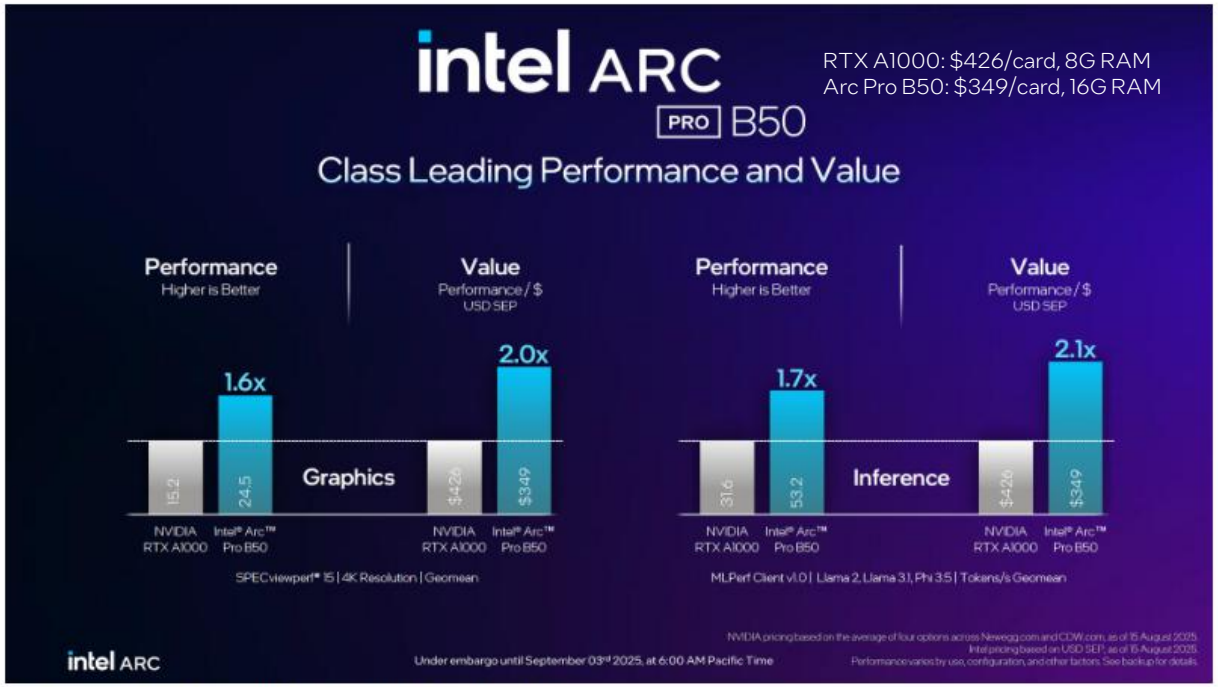
二、總結

Intel Arc Pro B50 單刀直入，剛好插在 SFF 尺寸、70W TBP、16GB VRAM 的專業繪圖卡之中，有著新一代 Xe2 核心架構，支援 AI 推論、光線追蹤、XeSS2、主流影音編解碼與 mini DisplayPort 2.1b UHBR 13.5 影像輸出能力。

對於專業應用的設計、工程、建築、施工與產品設計等軟體，讓小尺寸的工作站可以快速升級的 GPU；更藉由 16GB VRAM 滿足新一代 AI 推論、生成式 AI 模型的最低需求。Intel Arc Pro B50 也是目前專業繪圖卡中有著 16GB 記憶體且最便宜的選擇。

通過上述測試可見 Intel Arc Pro B50 在繪圖或 AI 推論效能上，確實能贏過比較的 RTX A1000，再加上兩者定價的差距，這也讓 Arc Pro B50 無疑是目前專業繪圖卡最佳每元效能的黑馬。

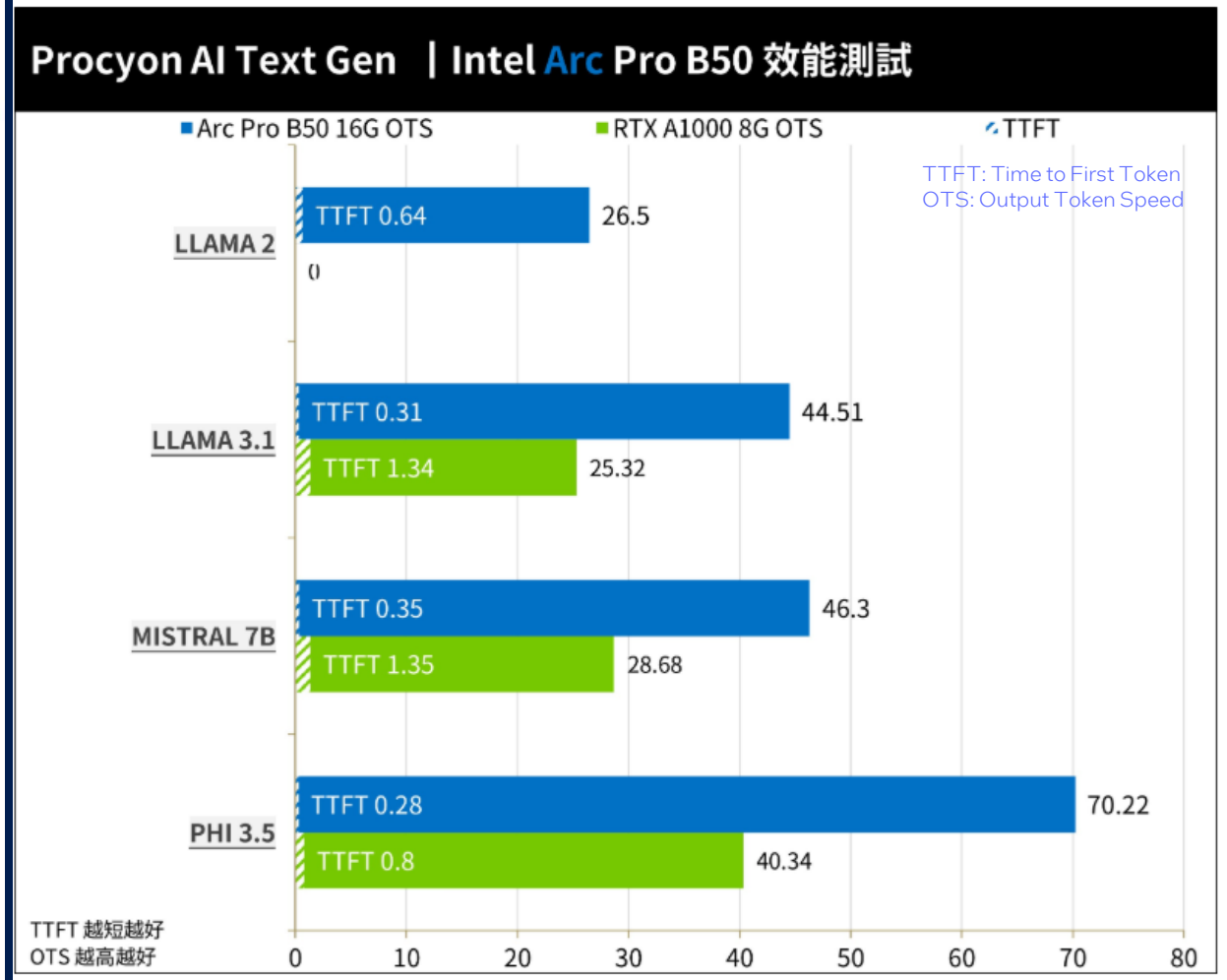
當然要打動專業繪圖市場，還是需要 Intel 穩紮穩打的維護並更新，以及獲得各大 ISV 的認證，慢慢累積 Intel Arc Pro 的硬體實力與軟體開發工具。



↑ 專業繪圖卡最佳每元效能的黑馬 Intel Arc Pro B50。

UL Procyon AI Text Generation Benchmark 提供 ONNX Runtime DirectML 或 OpenVINO 推論引擎，使用 Phi-3.5-mini、Llama-3.1-8B、Mistral-7B 與 Llama-2-13B 等四個模式，每個模型測試 7 個 Prompts 包含 RAG 與非 RAG 的查詢，通過權重後的總分與平均 Time To First Token（TTFT）、平均 Output Token Speed（OTS）提供專業用戶橫量電腦的 AI LLM 推論效能。

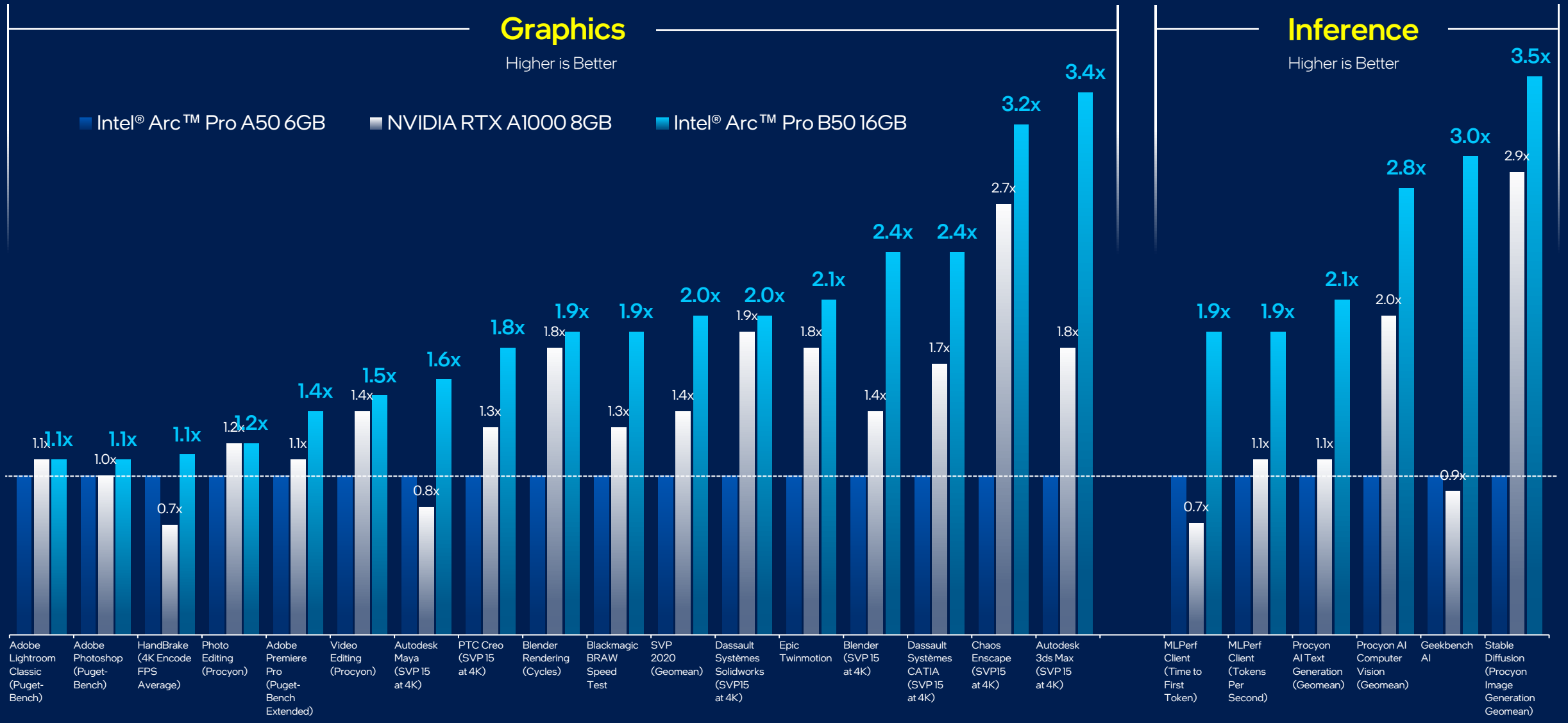
Intel Arc Pro B50 採用 OpenVINO 推論引擎 在四個模型的 TTFT 中獲得最快的輸出時間，以及在 OTS 輸出效能中領先比較的 RTX A1000。



↑ UL Procyon AI Text Generation Benchmark。

Source: https://www.xfastest.com/thread-304612-1-1.html

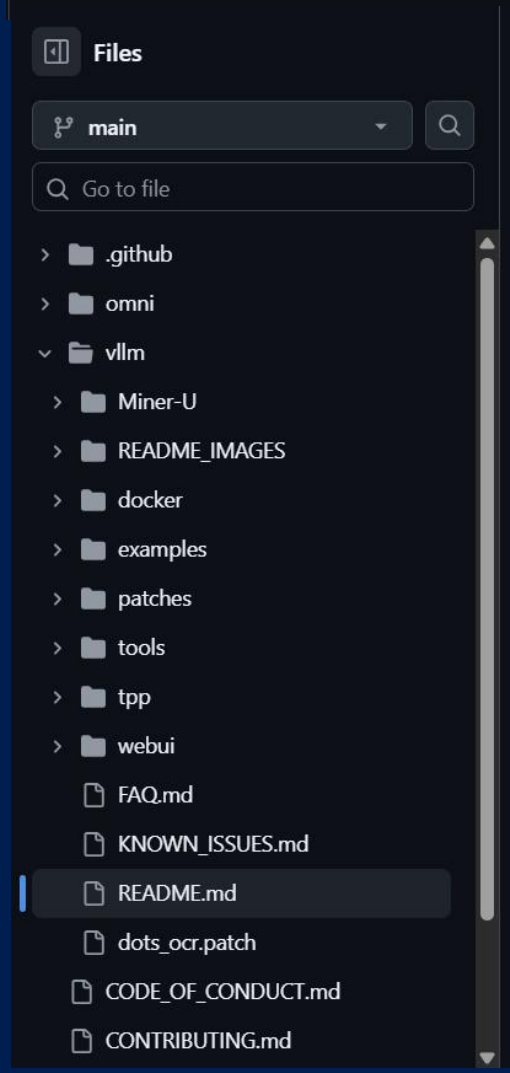
Up to 3x Gen-on-Gen Uplift



LLM Scaler vLLM

llm-scaler-vllm supports running text generation models using vLLM:

- Getting Started
- Features
- Supported Models



llm-scaler / vllm / README.md ↑ Top

Preview | Code | Blame | 2750 lines (2488 loc) · 85.3 KB | Raw

3. Supported Models

Model Name	Category	Notes
DeepSeek-R1-0528-Qwen3-8B	language model	
DeepSeek-R1-Distill-1.5B/7B/8B/14B/32B/70B	language model	
Qwen3-8B/14B/32B	language model	
DeepSeek-V2-Lite	language model	export VLLM_MLA_DISABLE=1
QwQ-32B	language model	
Ministral-8B	language model	
Mixtral-8x7B	language model	
Llama3.1-8B/Llama3.1-70B	language model	
Baichuan2-7B/13B	language model	with chat_template
codegeex4-all-9b	language model	with chat_template
DeepSeek-Coder-33B	language model	
GLM-4-0414-9B/32B	language model	
Seed-OSS-36B-Instruct	language model	
Hunyuan-0.5B/7B-Instruct	language model	follow the guide in here

Source: https://github.com/intel/llm-scaler?tab=readme-ov-file


intel ai

Notices & Disclaimers

Performance varies by use, configuration and other factors. Learn more at www.intel.com/PerformanceIndex.

Results that are based on pre-production systems and components as well as results that have been estimated or simulated using an Intel Reference Platform (an internal example new system), internal Intel analysis or architecture simulation or modeling are provided to you for informational purposes only. Results may vary based on future changes to any systems, components, specifications or configurations.

AI features may require software purchase, subscription or enablement by a software or platform provider, or may have specific configuration or compatibility requirements. Data latency, cost, and privacy advantages refer to non-cloud-based AI apps. Learn more at intel.com/AIPC.

All versions of the Intel vPro® platform require an eligible Intel processor, a supported operating system, Intel LAN and/or WLAN silicon, firmware enhancements, and other hardware and software necessary to deliver the manageability use cases, security features, system performance and stability that define the platform. See www.intel.com/performance-vpro for details.

Codec capabilities may vary by device and configuration. Contact your manufacturer to understand the enabled hardware acceleration and codec capabilities for individual devices.

Overclocking Intel processors with the 200S Boost profile will not void the limited processor warranty provided by Intel. All other warranty terms remain unchanged. This profile does not apply to processors overclocked before the profile launch date. Overclocking results will vary. The 200S Boost profile does not guarantee that any overclocking frequencies will be achievable or stable or that any level of performance will be achievable. Nothing in the Intel® 200S Boost program changes or modifies the performance specifications provided in the product information for any Intel component. Intel's warranty does not cover damage caused to any non-Intel component as a result of overclocking.
The 200S Boost warranty applies to Intel boxed processors. For boxed processors support, contact Intel support or place of purchase. If your PC was purchased through a system builder or OEM, your warranty and support for that warranty are provided exclusively by the system builder or OEM. Please contact your OEM or reseller f

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Built-in Intel® Arc™ GPU only available on select Intel® Core™ Ultra 200V series processor-powered systems; minimum processor power required. OEM enablement required. Check with OEM or retailer for system configuration.

While Wi-Fi 7 is backward compatible with previous generations, new Wi-Fi 7 features require PCs configured with Intel Wi-Fi 7 solutions, PC OEM enabling, operating system support, and use with appropriate Wi-Fi 7 routers/APs/gateways. 6 GHz Wi-Fi 7 may not be available in all regions. Performance varies by use, configuration, and other factors. For details on performance claims, learn more at www.Intel.com/performance-wireless.

No product or component can be absolutely secure. Intel technologies may require enabled hardware, software or service activation.

SPEC®,SPECrate® and SPEC CPU® are registered trademarks of the Standard Performance Evaluation Corporation. See http://www.spec.org/spec/trademarks.html for more information.

intel.