

Python

JavaScript

C#

.NET

ONNX

DirectML

WebNN

Semantic Kernel

# 當 **Phi-3 SLM** 降落在你的裝置上 你該怎麼玩


Poy Chang





# Poy Chang

目前任職於全美 100 大私人企業，負責企業內部商業應用解決方案的設計與開發，專注於 Azure、.NET 等技術研究

- ✓ STUDY4 社群核心成員
- ✓ Microsoft MVP 微軟最有價值專家
- ✓ .NET Conf Taiwan 的一份子
- ✓ 著作《.NET Conf 總召真心話》
- ✓ 譯著《企業級軟體架構模式》

 [s.poychang.net/blog](https://s.poychang.net/blog)

 [s.poychang.net/github](https://s.poychang.net/github)

 [s.poychang.net/fb](https://s.poychang.net/fb)

首先，你預期的裝置？      目前我**預期**的裝置是...



# LLM vs SLM

---

- 效率

- ✓ 佈署到相對小的設備或是邊緣裝置上
- ✓ 需要較少的計算能力和記憶體配置

- 無障礙性

- ✓ 允許開發者和企業更容易使用
- ✓ 在不需要投資大量基礎架構，讓小型團隊和各別研究人員探索語言模型

- 客製化

- ✓ 為特定領域和任務進行細微調整
- ✓ 專有的處理模型

# 選擇關鍵

---



推理延遲

Inference Latency



隱私保護

Privacy Protection



可控/可靠

Controllable/Reliability



節約成本

Cost Savings

# 選擇關鍵

---



準確度

Accuracy



推理延遲

Inference Latency




吞吐量

Throughput

使用者體驗

成本



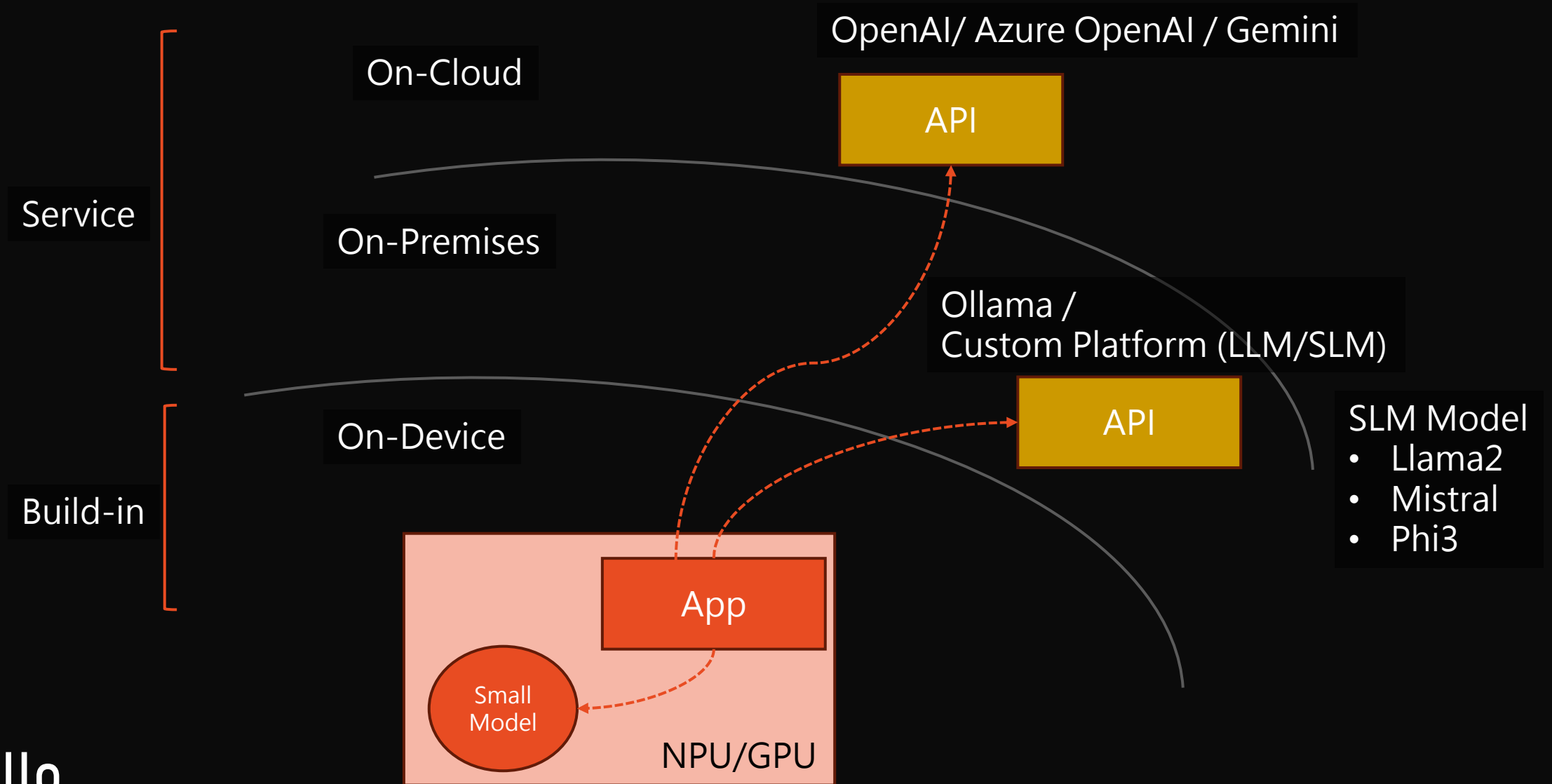
Offline capabilities are a binary choice?  
Where is sweet spot?

0%  
offline capable



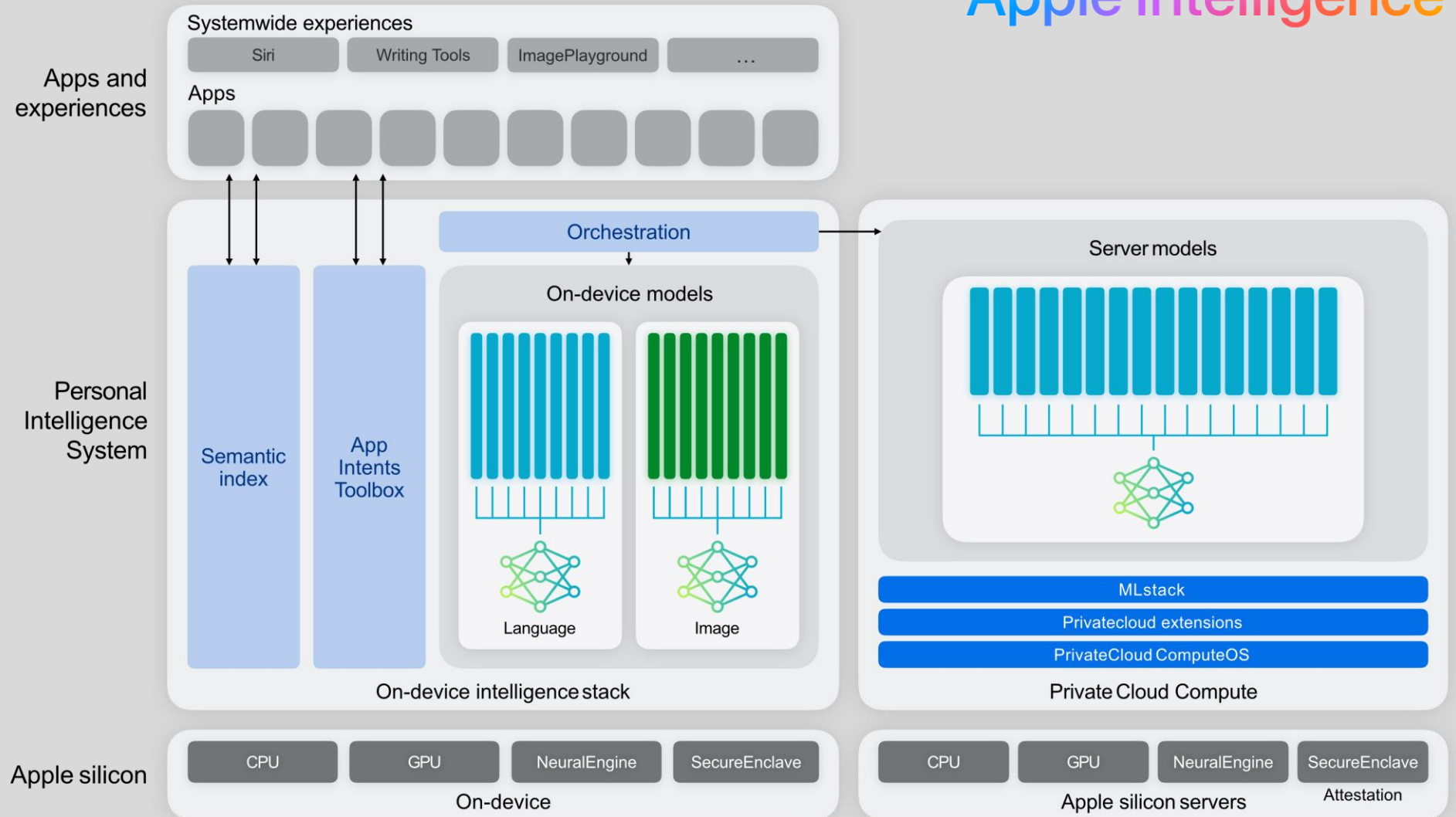
100%  
offline capable



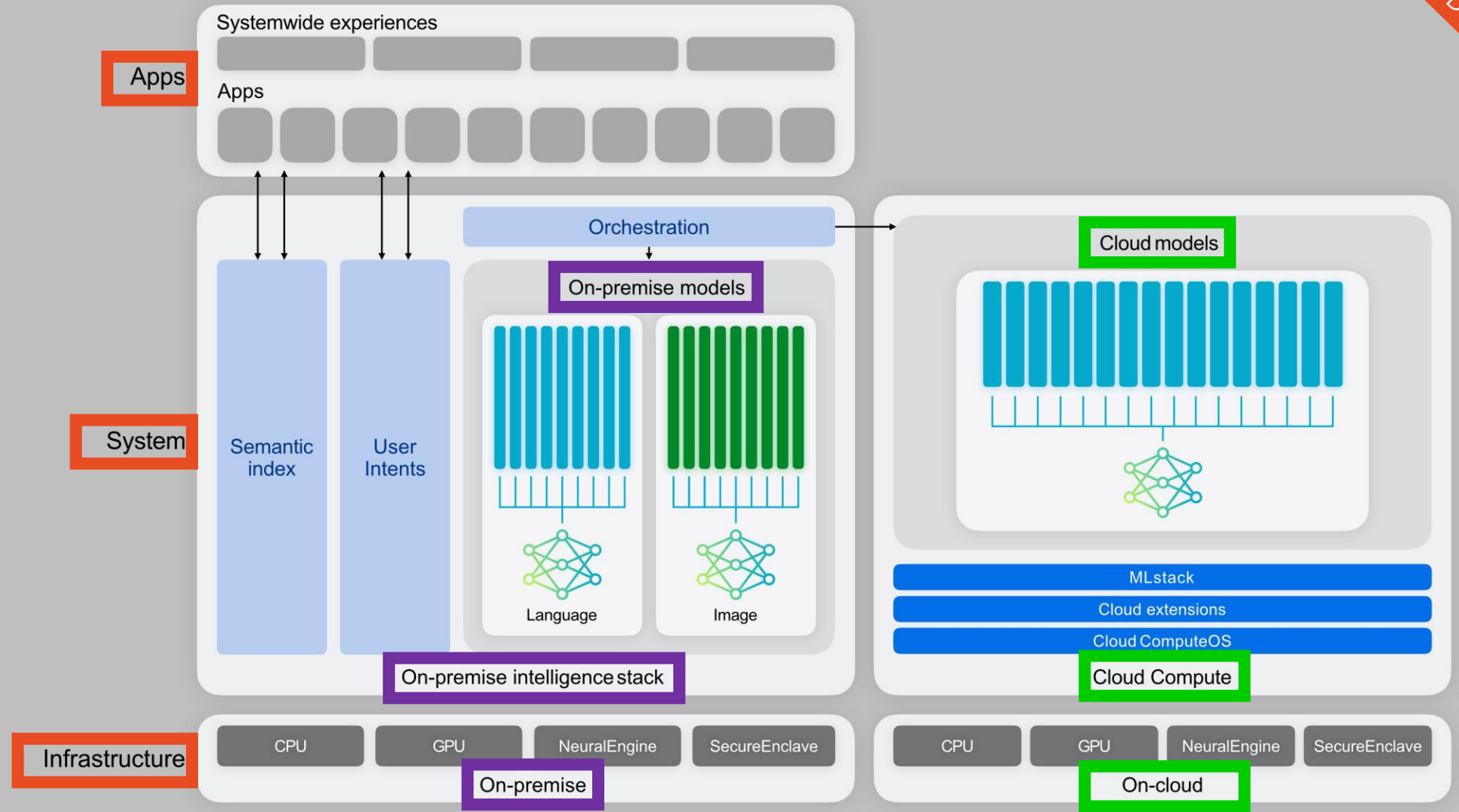




# Apple Intelligence



MODIFY





開發 AI 應用，其實是一場綜合格鬥



# 關於模型這件事...

對於 .NET 生態系的開發者來說，要玩深度學習模型只有 ONNX 一條路

# ONNX ( Open Neural Network Exchange )

---

- ONNX 是一種深度學習模型的格式，用於在不同的框架之間共享模型
  - ✓ 跨平台相容性，在不同的深度學習框架 ( 如 PyTorch、TensorFlow ) 之間轉換
  - ✓ 標準化的格式，可以在不同的硬體加速器和推理引擎上運行
  - ✓ 廣泛支援與開放原始碼
- ONNX Runtime 是一個高性能推理引擎，用於執行 ONNX 格式的機器學習模型
  - ✓ 支援 Python、C++、C#、Java、JavaScript 等程式語言

# ONNX 三部曲

---



Olive

轉換/優化工具

A toolkit for hardware-aware  
AI model optimization.  
Output is an ONNX  
formatted model to run with  
quality and efficiency on the  
ONNX runtime



ONNX

深度學習模型格式

Open and interoperable file  
format for ML and DNN  
models

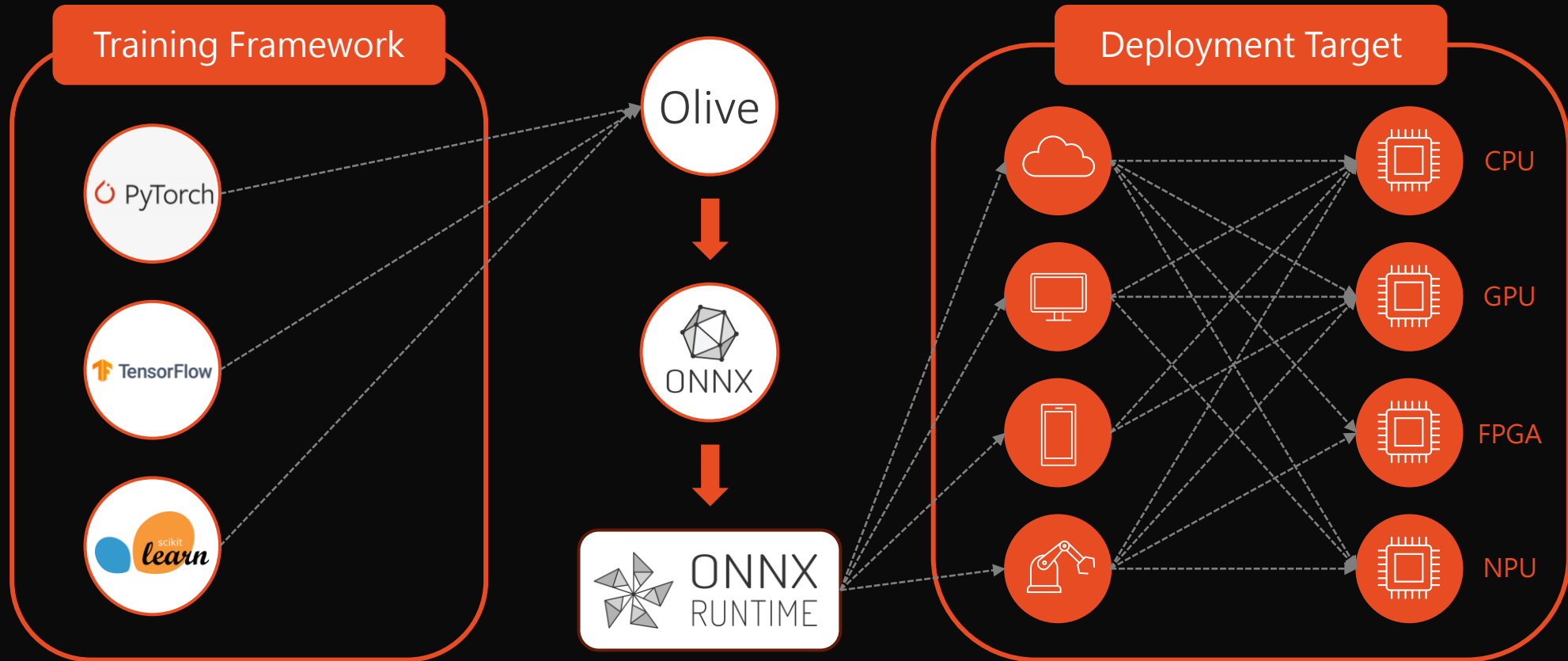


ONNX  
Runtime

推理引擎

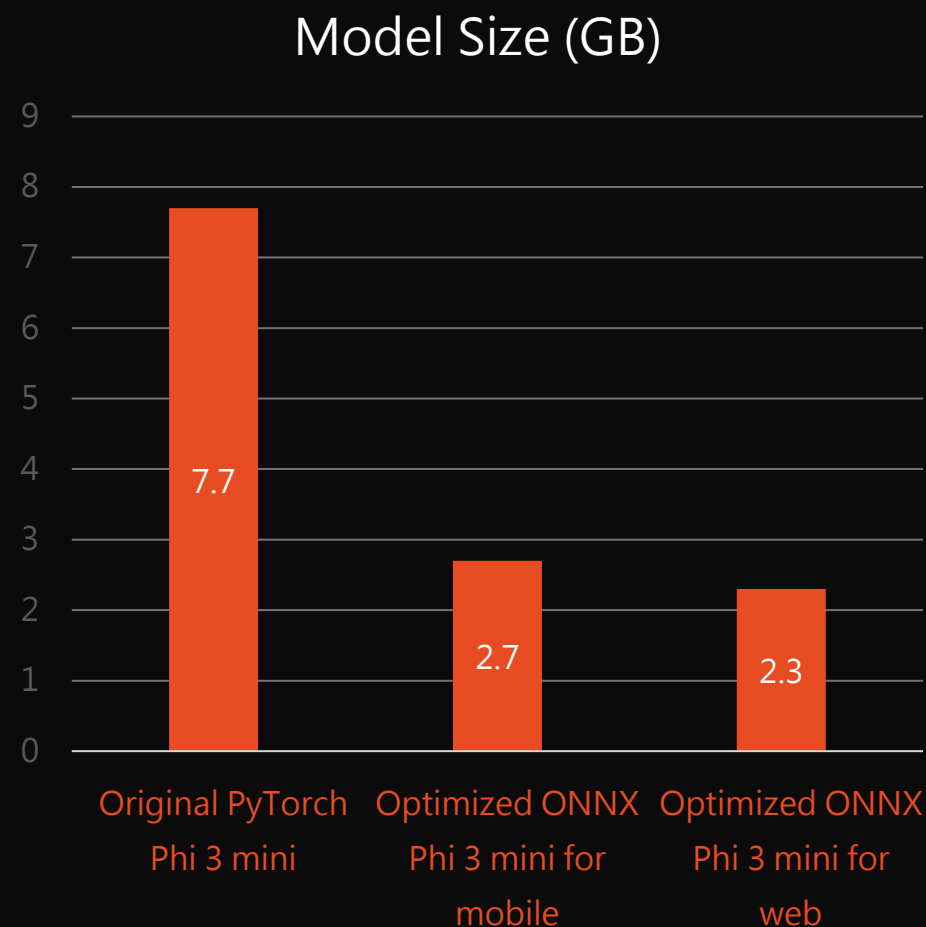
Fast and efficient model  
inference and training engine  
that works across a diverse  
range of hardware  
accelerators

# ONNX 願景：Run Models Everywhere



# Olive ( ONNX Go Live )

- 提供量化模型和優化模型的功能
- 以 Phi 3 mini 為例
  - ✓ 與原始 PyTorch 版本相比，針對行動裝置和 Web 量化的模型尺寸減少了 2.5 倍以上
  - ✓ 從原本的 FP16 量化成 INT4  
<https://github.com/microsoft/Olive/tree/main/examples/phi3>
  - ✓ 採用 QLoRA 算法  
Quantization + Low-Rank Adaptation





# Fine-tuning with AI Toolkit for Visual Studio Code

Preview

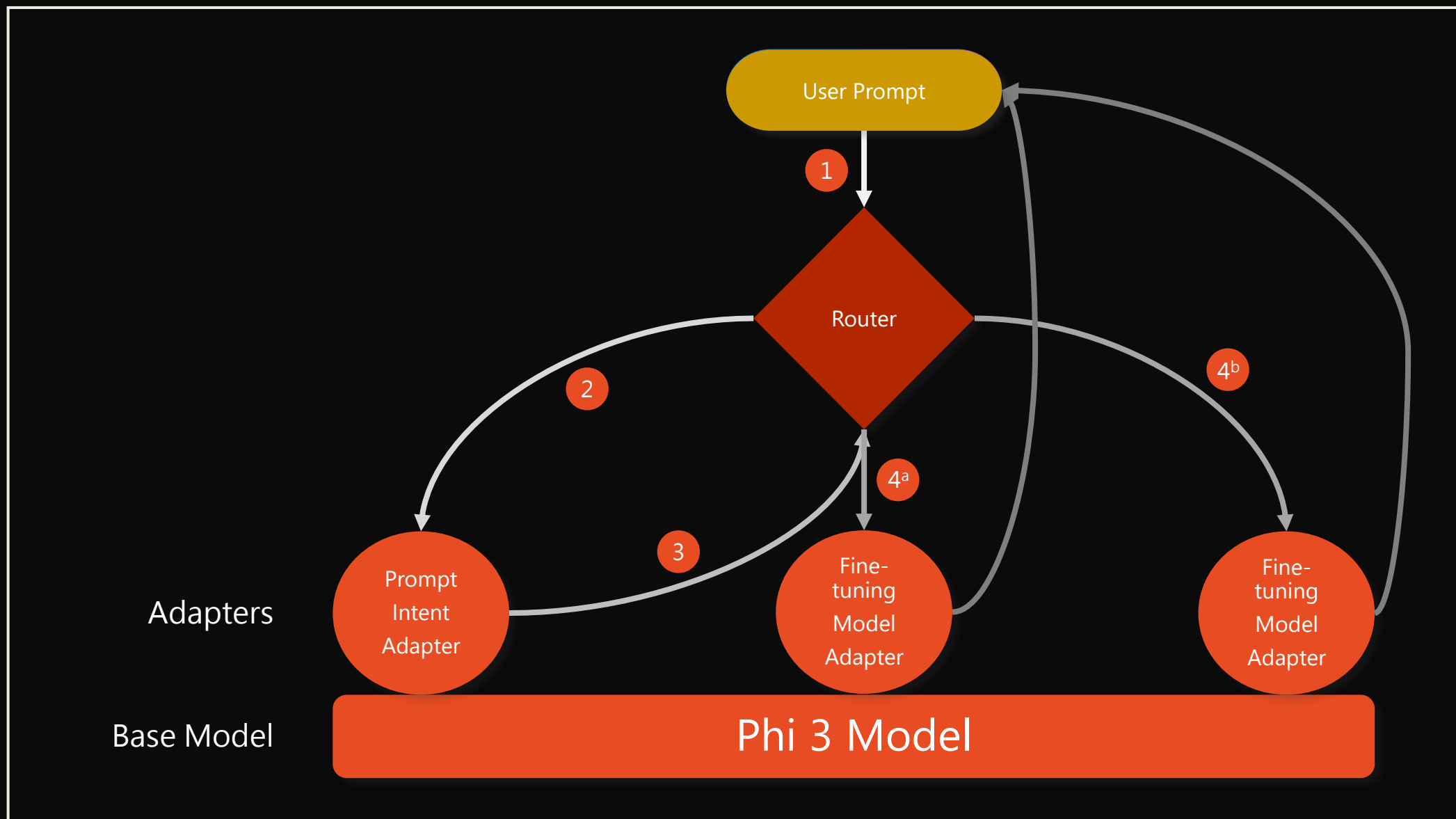
The screenshot shows the AI Toolkit interface within Visual Studio Code. The sidebar on the left contains the 'AI TOOLKIT' section with a 'MODEL CATALOG' sub-section. The main panel displays a grid of AI models for discovery and fine-tuning.

**Discover the best-suited model for your AI project**

**Model catalog**  
Download, explore, and fine-tune models locally and directly in VS Code.

Model Name	Provider	Architecture	File Size	OS Support	Actions
phi-2-int4-cpu	Microsoft Research	3B CPU	2.18 GB	Windows and Linux	View License   Load in Playground
phi-2-int4-cuda	Microsoft Research	3B GPU	1.75 GB	Windows and Linux	View License   Download
mistral-7b-v01-int4-cpu	Mistral AI	3B CPU	4.99 GB	Windows, MacOS and Linux	View License   Download
mistral-7b-v01-int4-cuda	Mistral AI	3B GPU	4.27 GB	Windows, MacOS and Linux	View License   Download
phi-3-mini-4k-cpu-int4-rtn-block-32-acc-level-4-onnx	Microsoft Research	3B CPU	2.72 GB	Windows	View License   Download
phi-3-mini-4k-cpu-int4-rtn-block-32-onnx	Microsoft Research	3B GPU	2.72 GB	Windows	View License   Download

## Concept of Swap Fine-tuning Model



# ONNX 問題

---

- 算子 ( Operator ) 的支援

- ✗ 不同的深度學習框架可能會實作相同算子的不同變體，造成 ONNX 使用的算子與原始框架的略有不同，此差異可能會影響模型推理的結果

- 數值精度

- ✗ 不同框架在處理浮點數計算時，可能會有不同的數值精度（如 FP-32 與 FP-64），這微小的差異會在多層運算後累積，導致推理結果的微小變化

- 量化模型

- ✗ 量化模型的過程中，可能涉及降低神經網路的權重和精度，特別是原始框架和 ONNX Runtime 之間的量化實作不一致時

- 不完全的支援或不相容的特性

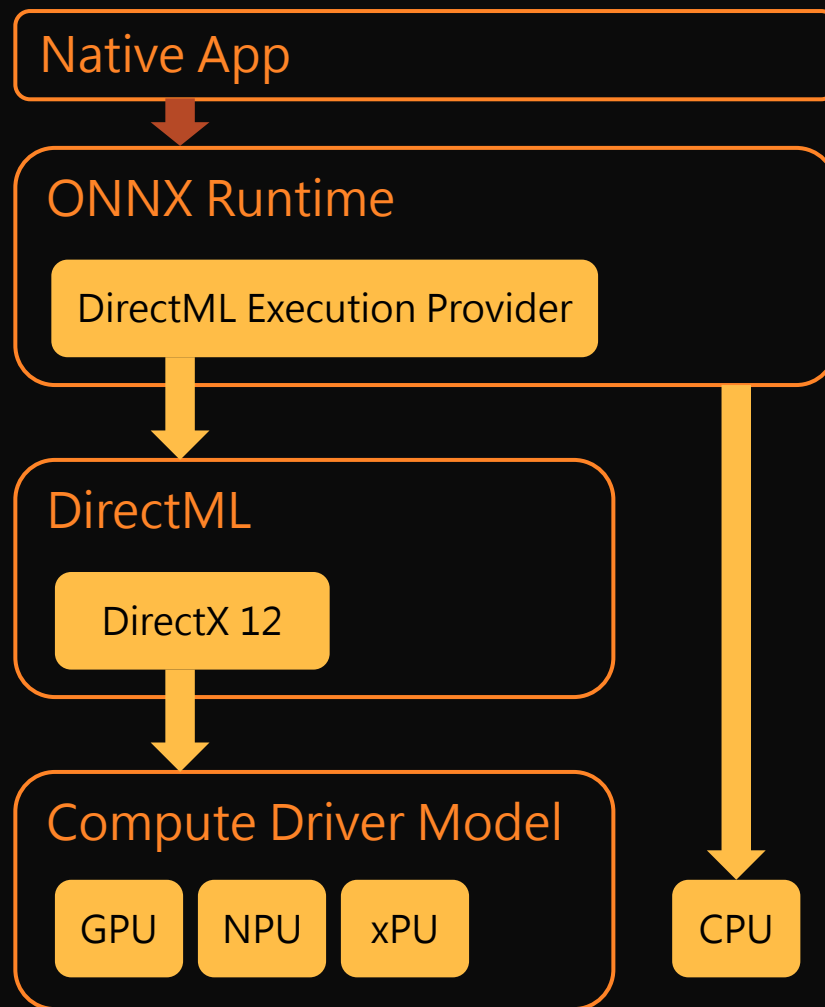
- ✗ 有些框架可能使用 ONNX 尚未完全支援的特性，在轉換過程中，這些部分需要被替換或近似處理，這將會影響模型的精度

# 使用之前，先補足知識點

平台百百款，我們單看 Windows 這一款

# DirectML ( Direct Machine Learning )

- DirectML 提供高性能、跨平台的硬體加速深度學習庫
  - ✓ 設計理念類似 DirectX 圖形處理的作用
  - ✓ 通過統一介面，利用底層硬體的計算能力，提供高效、簡便的開發體驗
  - ✓ 支援多種硬體加速
  - ✓ 基於 DirectX API



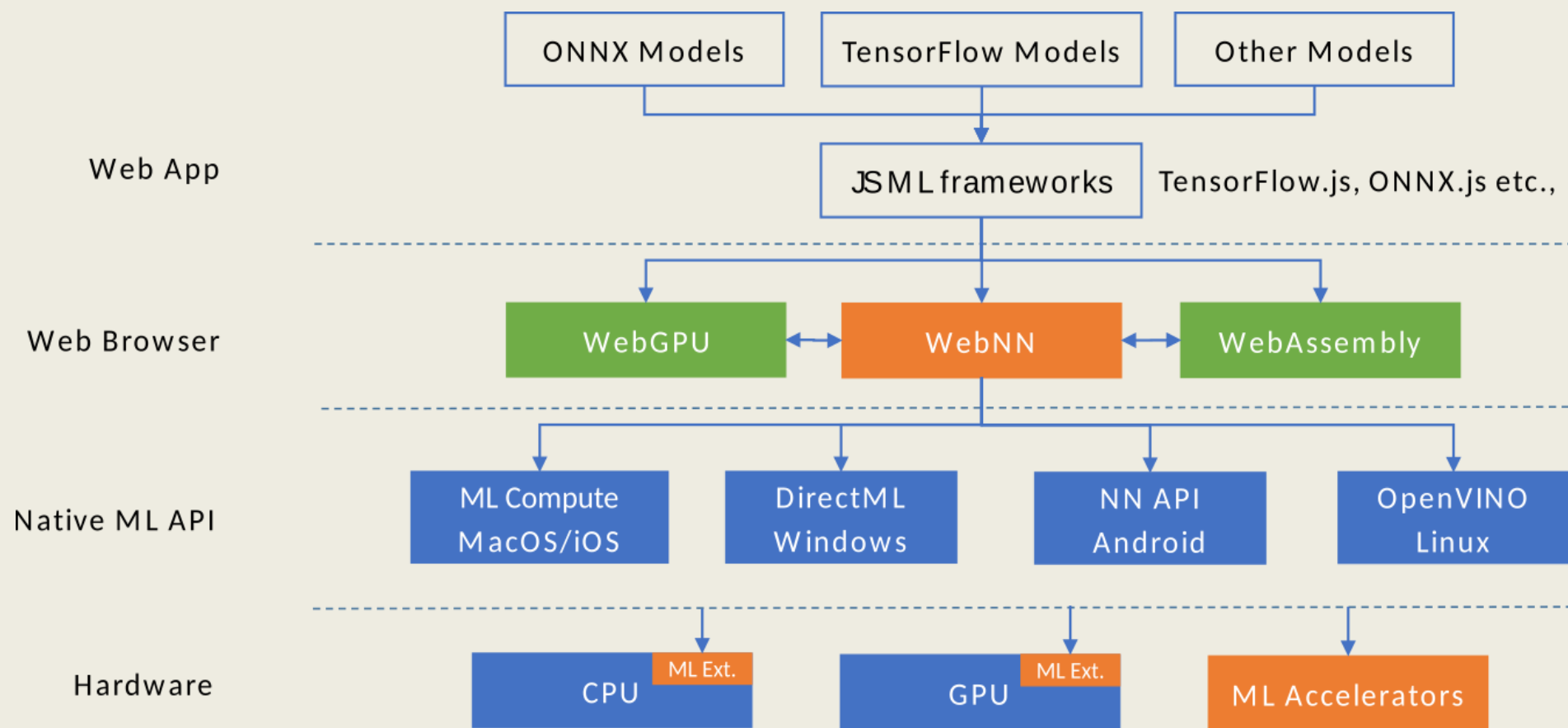
# WebNN ( Web Neural Network )

Developing

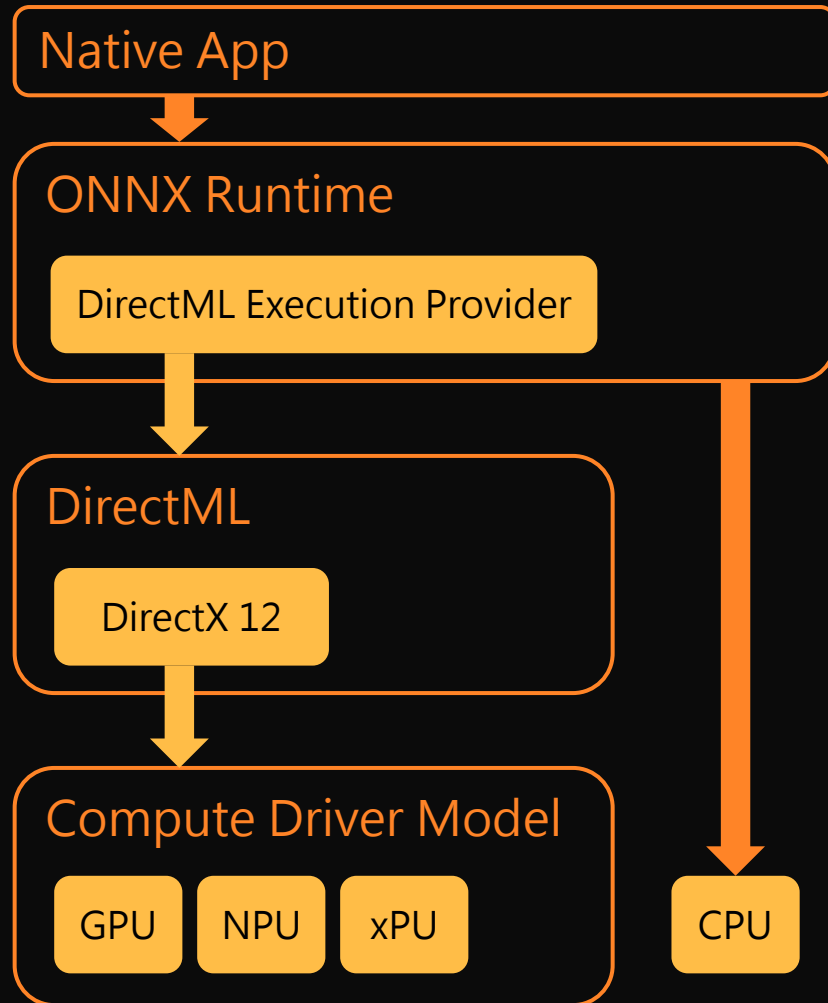
- WebNN 是開發中的 Web 標準，為瀏覽器提供統一的神經網路 API，讓 Web 應用能夠執行機器學習推理
  - ✓ 支援 ONNX 和 TensorFlow.js 模型格式
  - ✓ 不依賴於特定平台
  - ✓ 利用底層的硬體加速



## Take advantage of the native



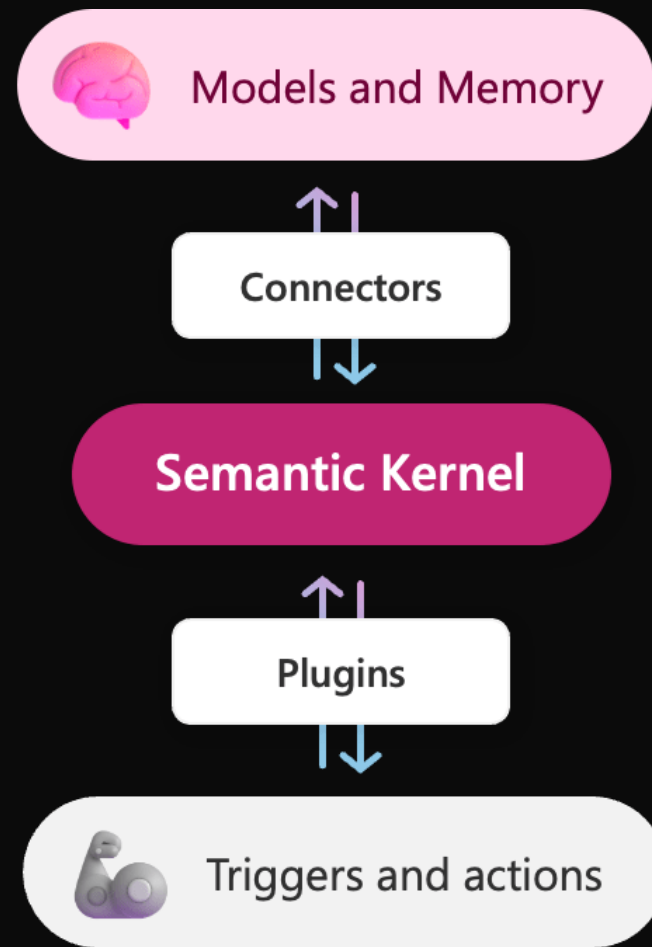
# Native & Web App





# Semantic Kernel

- 由 Microsoft 主導的輕量化且開源的 SDK，通過封裝常見的 AI 操作，並提供靈活且可擴展的框架，讓開發者可以快速建立與語言模型互動的應用程式
  - ✓ 支援開發語意或原生地 Plugins
  - ✓ 藉此開發特定用途的 Agents
  - ✓ 連結向量資料庫或 AI 模型
  - ✓ 可做為開發 AI 協調者 ( Orchestrator ) 的框架



## DEMO

- ✓ 透過 WebNN 運行 Phi 3
- ✓ 使用 OnnxRuntime 運行 Phi 3
- ✓ 借用 Semantic Kernel Connector 運行 Phi 3

Let's run  
Phi 3 locally

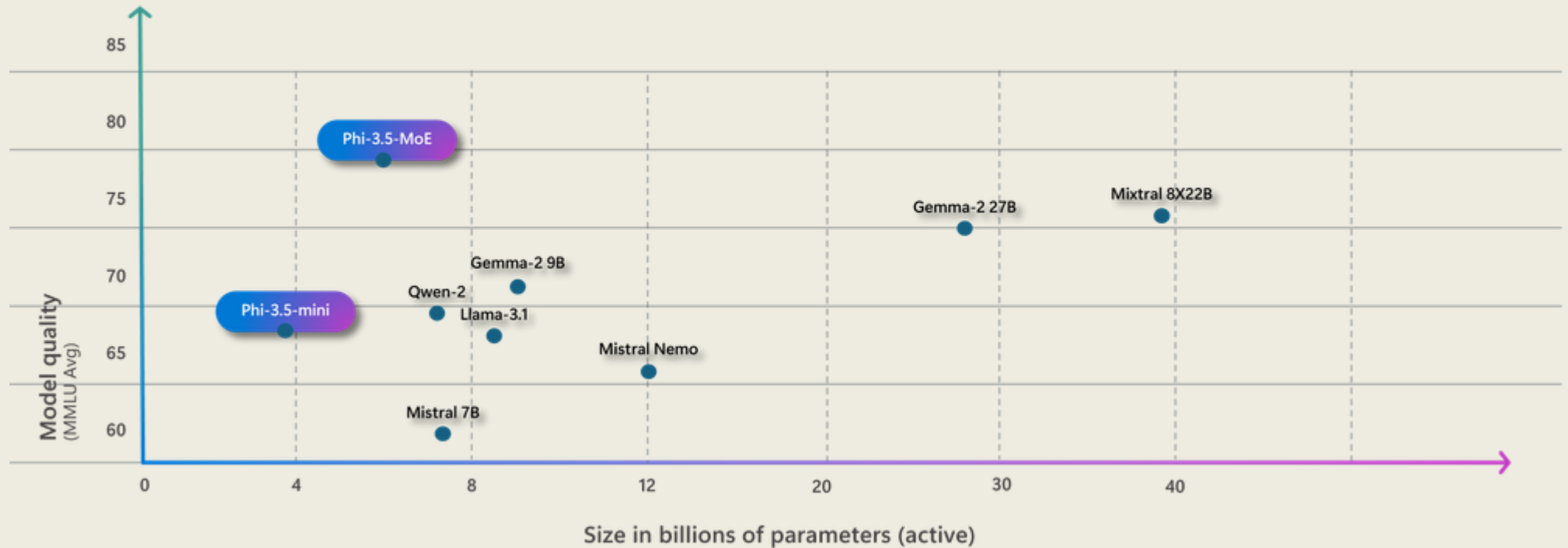
Intel Core i9-13900H CPU

NVIDIA GeForce RTX 4070 Laptop GPU

<https://github.com/poychang/HWDC-PlayWithPhi3>

Phi 3.5 Arrival and ONNX format is coming

## Phi-3.5 Quality vs Size in SLM





thanks for your  
attention!