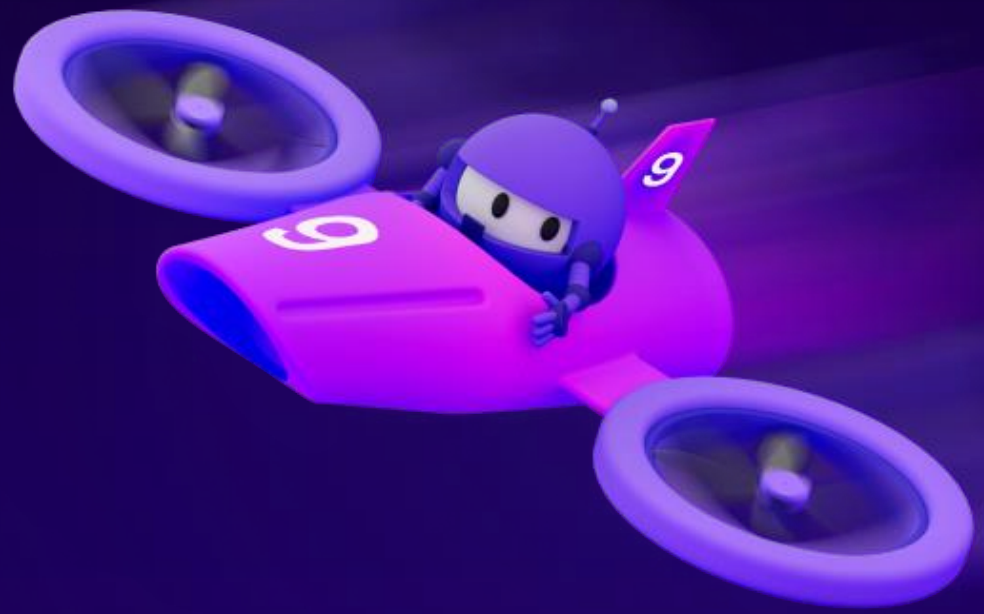
Use .NET Aspire to create GenAI Cloud Native Solutions

Kinfey Lo – Microsoft Senior Cloud Advocate



Agenda

- Introduce AI
- SLM
- Using ONNX Runtime GenAI to inference phi model
- Microsoft.Extension.AI
- Semantic Kerenl
- Sample

01
.NET is cool

Build anything with a unified platform

.NET

Cloud | Web | Desktop | Mobile | Gaming | IoT | AI

Tools


Visual Studio


Visual Studio Code


CLI


GitHub Copilot

\+

Operating system


Windows


Linux

macOS

\+

Ecosystem


NuGet


GitHub


.NET Aspire


Components, tools, library vendors

Announcing .NET 9

aka.ms/get-dotnet-9

Productive

Modern & Secure

Intelligent

Performance

.NET 9 Highlights

Productive

.NET Aspire for improved developer inner loop

Deeper GitHub Copilot integration in Visual Studio

Hot Reload enhancements in VS Code with C# Dev Kit

Refreshed build output highlighting

Modern & Secure

Simplified Blazor authentication

Improved integration with cloud-native services & OpenTelemetry

Enhanced UI controls across Blazor, .NET MAUI, WinUI 3, WPF, and Windows Forms

dotnet restore auditing for known security vulnerabilities

Intelligent

.NET AI Building Blocks with Microsoft.Extensions.AI

AI Tokenizers

Tensor<T>

TensorPrimitives

.NET SDKs for vector databases

Server-Sent Event Parser

Performance

Adaptive Garbage Collection

Up to 25% faster Blazor startup

Optimized static web assets

with ref struct improvements

More Native AOT

Trimming optimizations

New NuGet restore resolver for large projects

02

Let’s talk some AI

.NET Developers questions

- How to use AI model
  in local / cloud
- How to build RAG
- How to build
  AI@Cloud Native application

About Models

Microsoft.Extensions.AI

.NET AI Library

Streamline AI integration with our unified APIs

Common AI Abstractions

Standard Middleware

Interoperability and Extensibility

Available in Preview Today

What is Phi ?

Introducing Phi

Small Language Models

Groundbreaking performance for size, with frictionless availability

Phi-3.5-mini (3.8B)

Phi-3.5-vision (4.2B)

(6. Phi-3.5-MoE 6B active)

Available on

Azure AI Model Catalog

Hugging Face

ONNX Runtime

NVIDIA NIM

Ollama

Phi-3: Groundbreaking SLM performance

Quality

Best quality-cost for size; outperforms models 2x larger

Runs everywhere

CPUs and GPUs; cloud and on-device

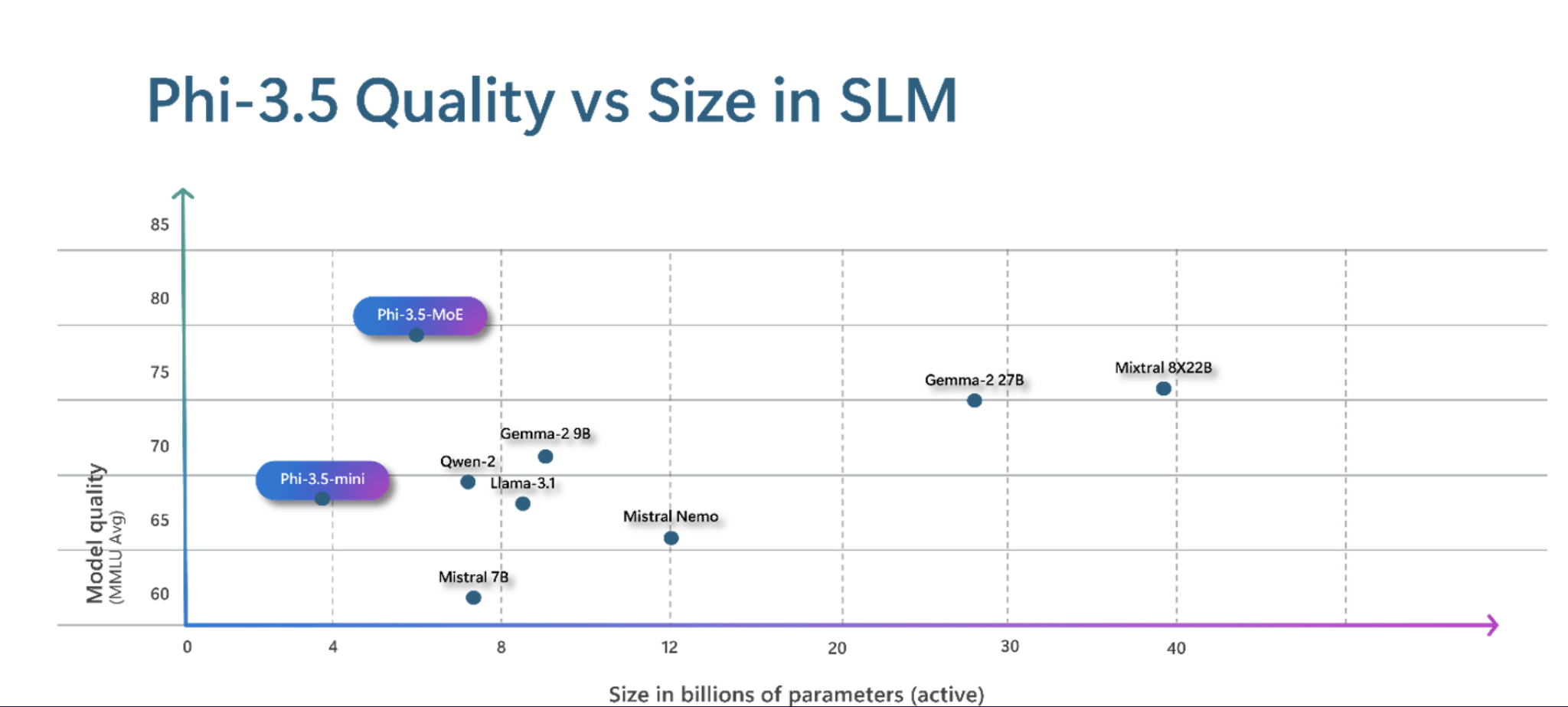
Speed

Blazing fast inference speed

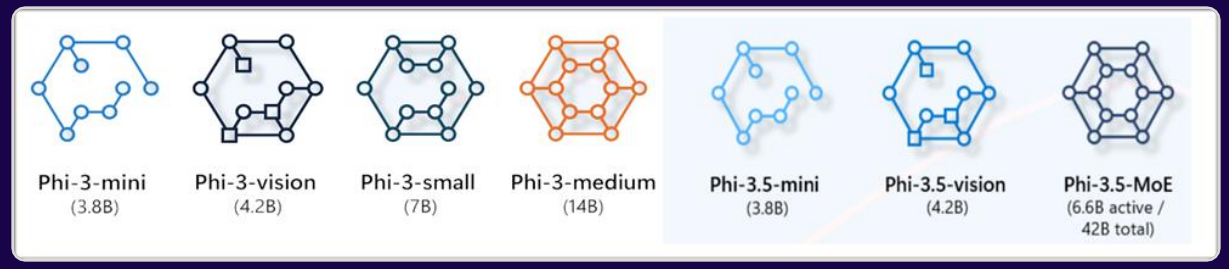
Long-context

Send more context in prompts

Phi offers high quality results at a small size



The opportunity of SLMs apps and devices



Small Language Models

On Cloud

On Edge

Mobile • Edge

(iOS, Android)

NVIDIA

iOS

Windows Copilot Runtime

Applications + Experiences

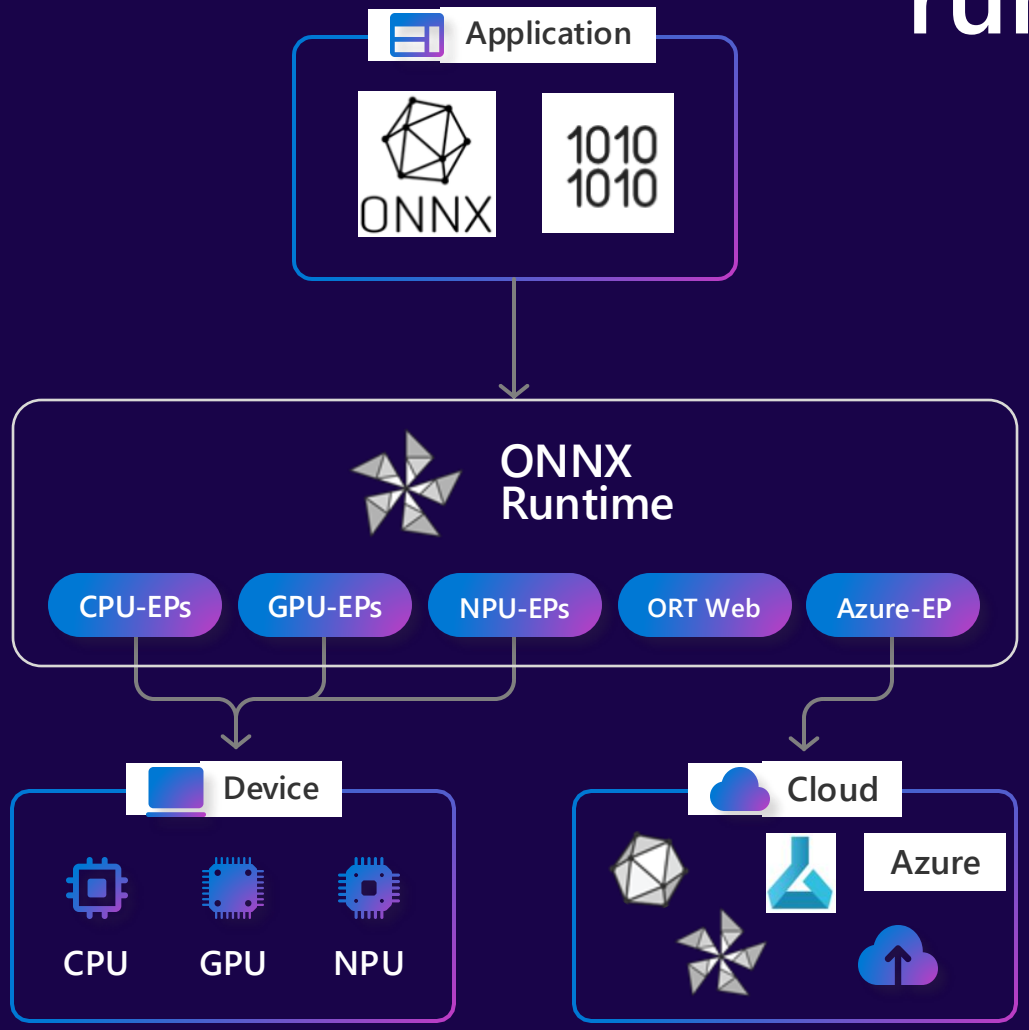
Windows Copilot Library + On-device models

AI frameworks + toolchain

Silicon

aka.ms/wcr

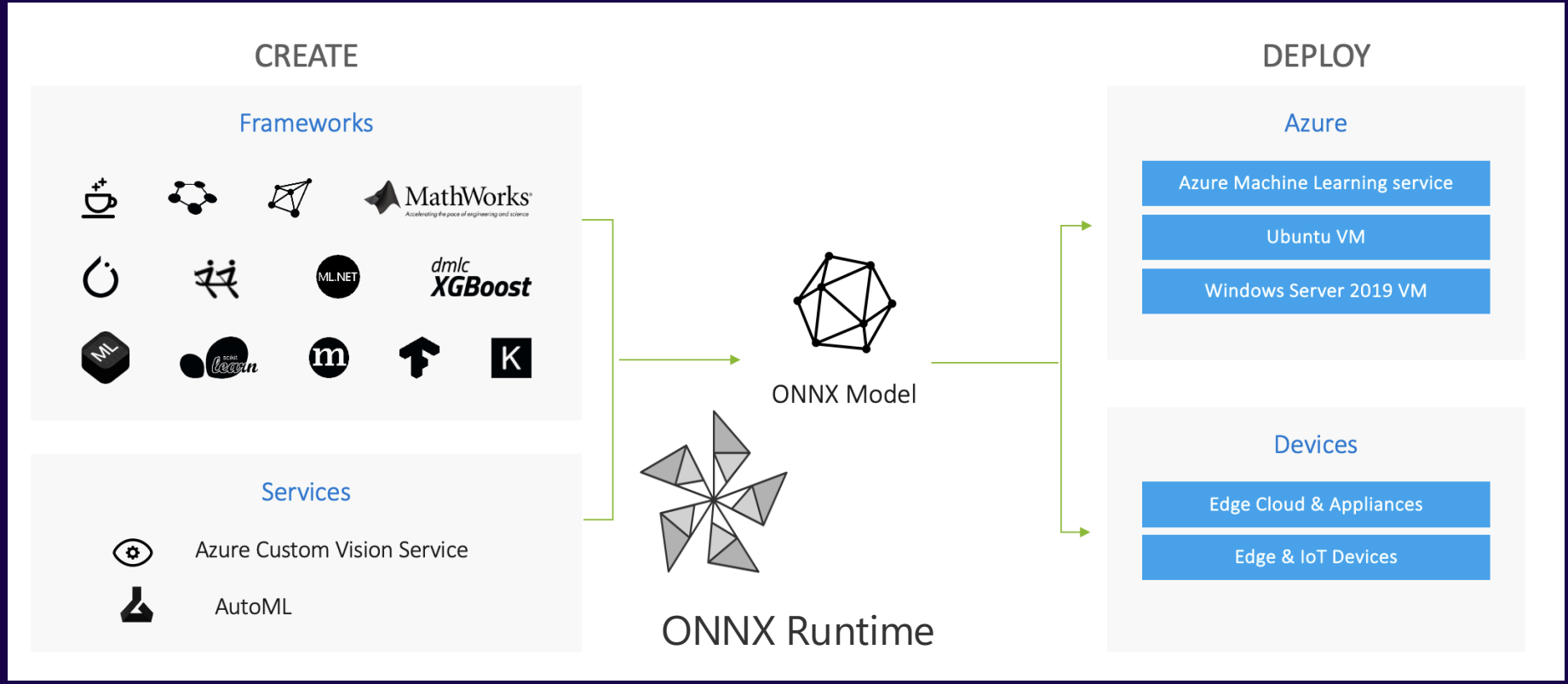
Optimization and performance ONNX runtime



Execution provider	Company
Azure	Microsoft
QNN	Qualcomm
OpenVino	Intel
Vitis AI	AMD
DirectML	Microsoft
WebNN	Microsoft, Intel

Using C# to call local model

.NET
Python ❤️



https://github.com/microsoft/onnxruntime-genai

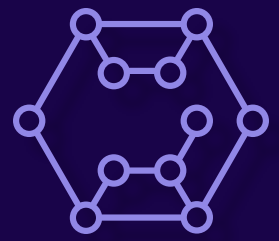
DEMO

03

.NET & AI @Cloud Native

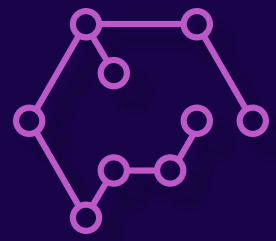
Phi-3-mini-V

(3.8B + 0.3B)



Phi-3-small

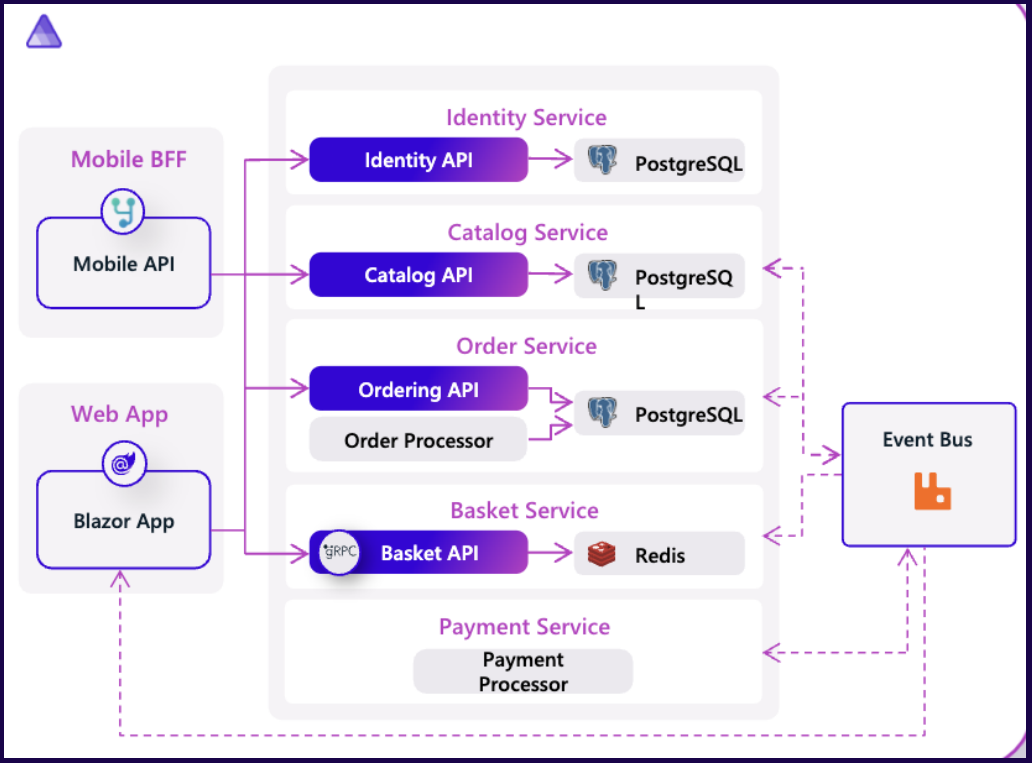
(7B)



Phi-3-mini

(3.8B)

What's .NET Aspire

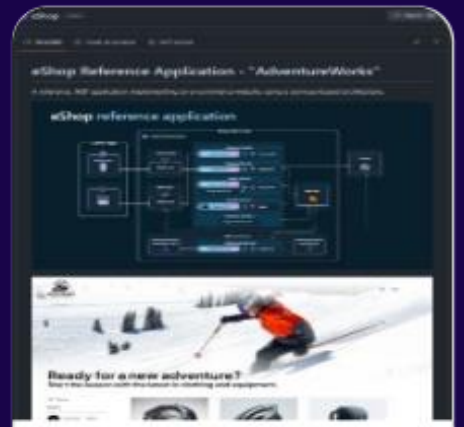


.NET Aspire is designed to improve the experience of building .NET cloud-native apps. It provides a consistent, opinionated set of tools and patterns that help you build and run distributed apps. .NET Aspire is designed to help you with:

- **Orchestration**: .NET Aspire provides features for running and connecting multi-project applications and their dependencies for local development environments.
- **Components:** .NET Aspire components are NuGet packages for commonly used services, such as Redis or Postgres, with standardized interfaces ensuring they connect consistently and seamlessly with your app.
- **Tooling:** .NET Aspire comes with project templates and tooling experiences for Visual Studio, Visual Studio Code, and the dotnet CLI to help you create and interact with .NET Aspire projects.

AI@.NET Aspire

Building Generative AI apps with .NET 8 & 9



Discover end-to-end sample applications and documentation

Learn

Semantic Kernel simplifies consuming AI components

Build

OpenAI

drant

Milvus

Growing AI ecosystem

Ecosystem

.NET Aspire and azd streamline deployment

Deploy

.NET Aspire + Azure + Semantic Kernel Easy monitoring and observability

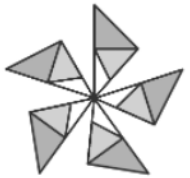
Monitor

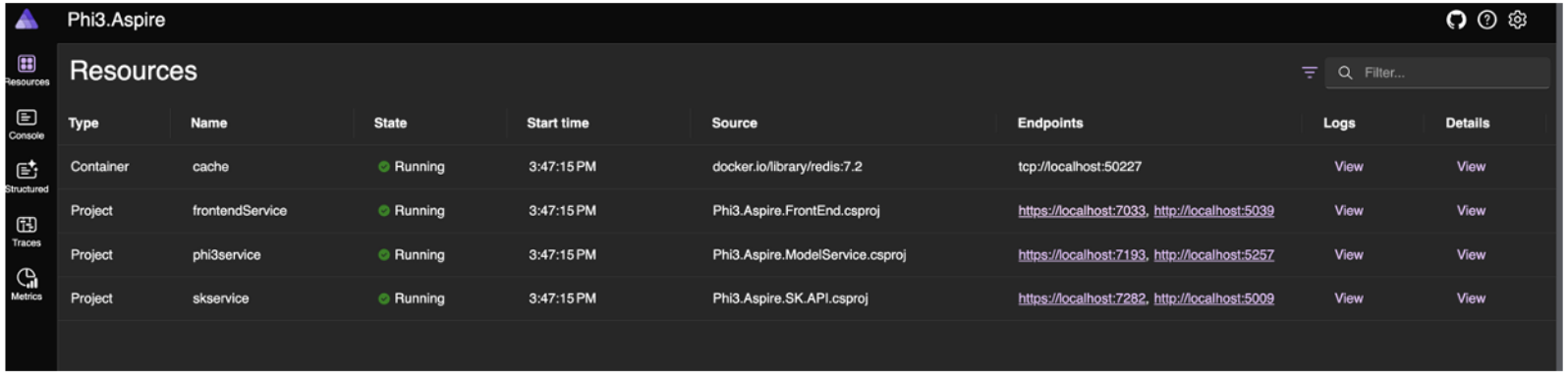
https://devblogs.microsoft.com/dotnet/build-gen-ai-with-dotnet-8/

Ask

question

connect





Demo : .NET Aspire with Phi3-mini

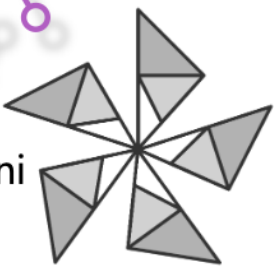
Blazor

ChatUI Service

Semantic Kernel

AI Orchestration-Service

Phi-3-mini
(3.8B)



ONNX Runtime

Phi-3-mini-Service

DEMO

04

Inference SLM with .NET

Microsoft.Extensions.AI

.NET AI Library

Streamline AI integration with our unified APIs

Common AI Abstractions

Standard Middleware

Interoperability and Extensibility

Available in Preview Today

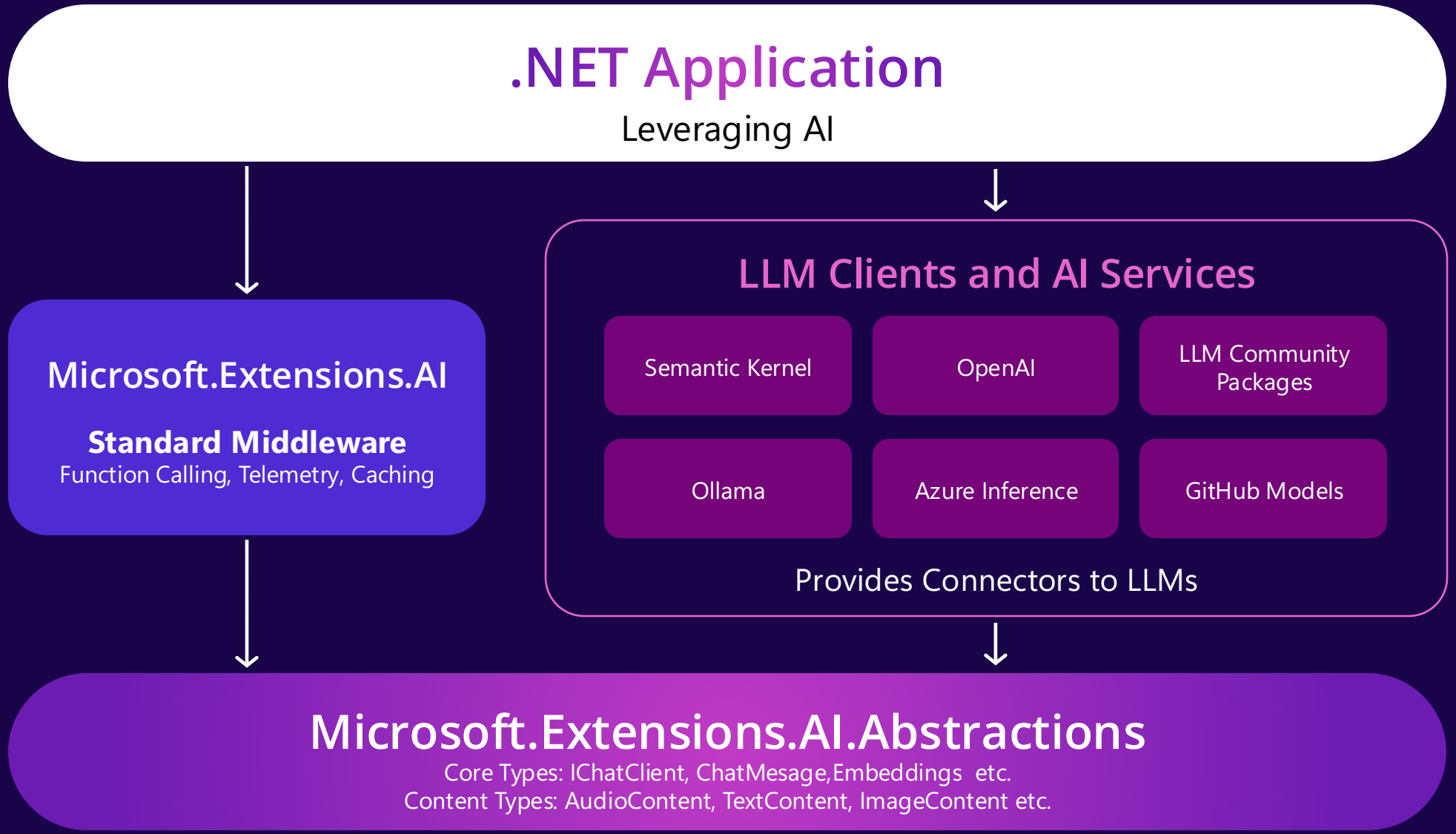
Swap AI Services with Ease

Development

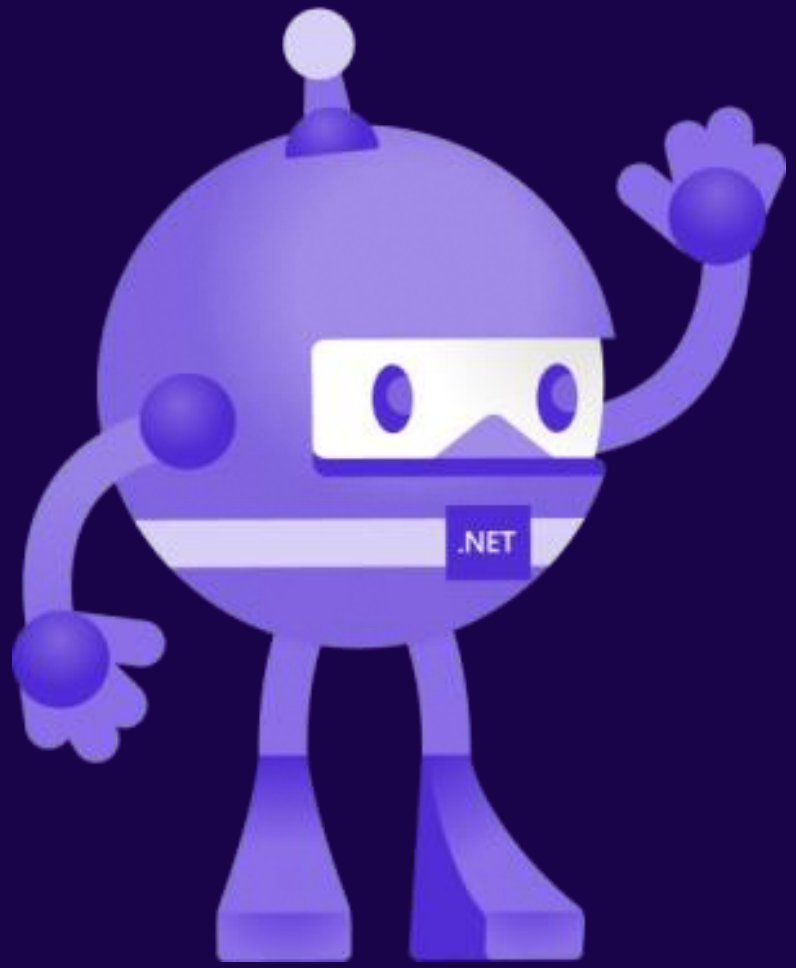
```
using Microsoft.Extensions.AI;

IEmbeddingGenerator<string,Embedding<float>> generator =
    new OllamaEmbeddingGenerator(new Uri("http://localhost:11434/"), "all-minilm");

var embedding = await generator.GenerateAsync("What is AI?");

Console.WriteLine(string.Join(", ", embedding[0].Vector.ToArray()));
```

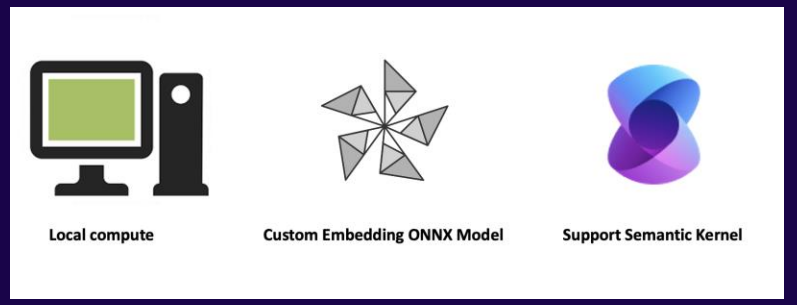
Production

```
using OpenAI;
using Microsoft.Extensions.AI;

IEmbeddingGenerator<string,Embedding<float>> generator =
    new OpenAIClient(Environment.GetEnvironmentVariable("OPENAI_API_KEY"))
        .AsEmbeddingGenerator("text-embedding-3-small");

var embedding = await generator.GenerateAsync("What is AI?");

Console.WriteLine(string.Join(", ", embedding[0].Vector.ToArray()));
```

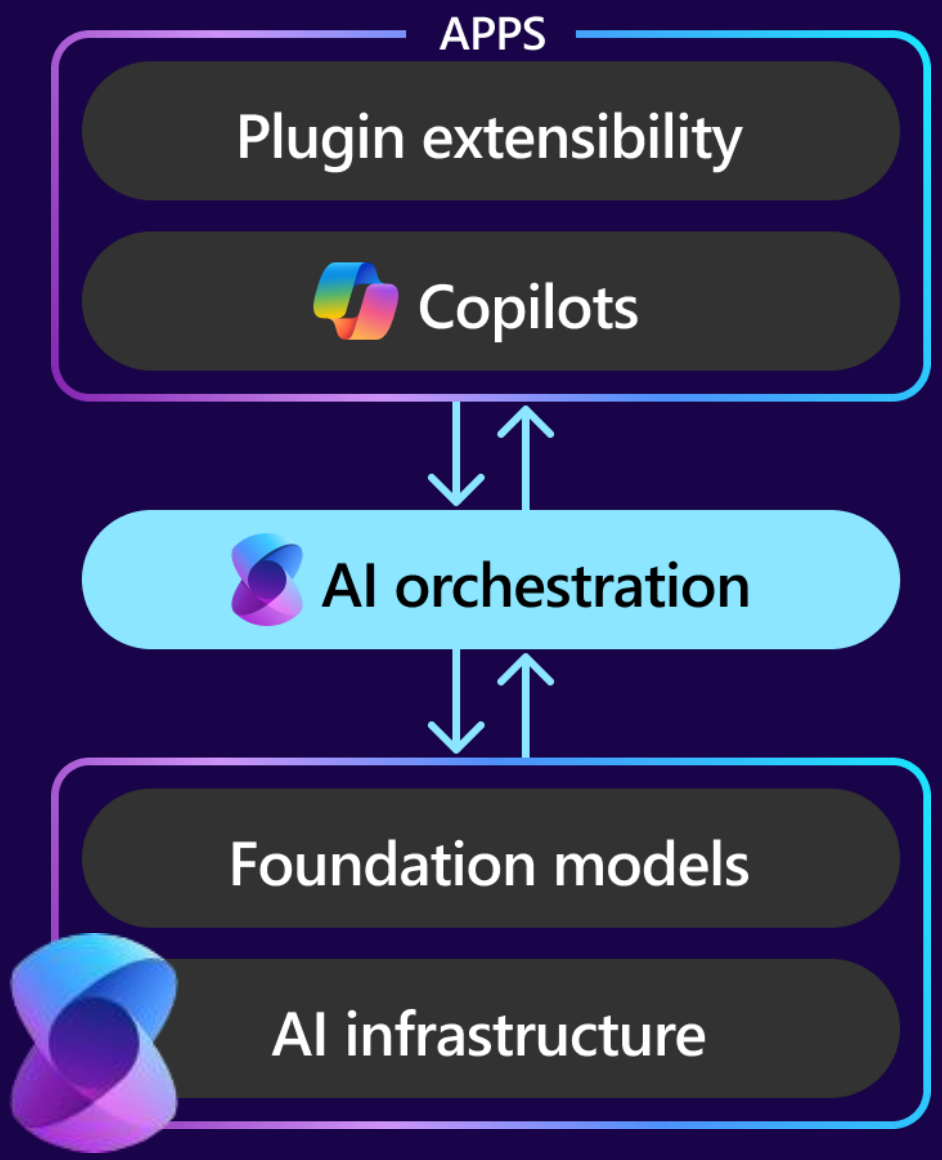

.NET Application
Leveraging AI
Microsoft.Extensions.AI
Standard Middleware
Function Calling, Telemetry, Caching
LLM Clients and AI Services
Semantic Kernel
OpenAI
LLM Community Packages
Ollama
Azure Inference
GitHub Models
Provides Connectors to LLMs
Microsoft.Extensions.AI.Abstractions
Core Types: IChatClient, ChatMesage,Embeddings etc.
Content Types: AudioContent, TextContent, ImageContent etc.

DEMO





Smart Components



Semantic Kernel

This highly extensible, open-source framework empowers you to leverage the latest AI models on top of your existing C#, Python and Java code. You'll be equipped to build custom AI-agents that can automate your business processes.

https://aka.ms/SemanticKernelCookBook

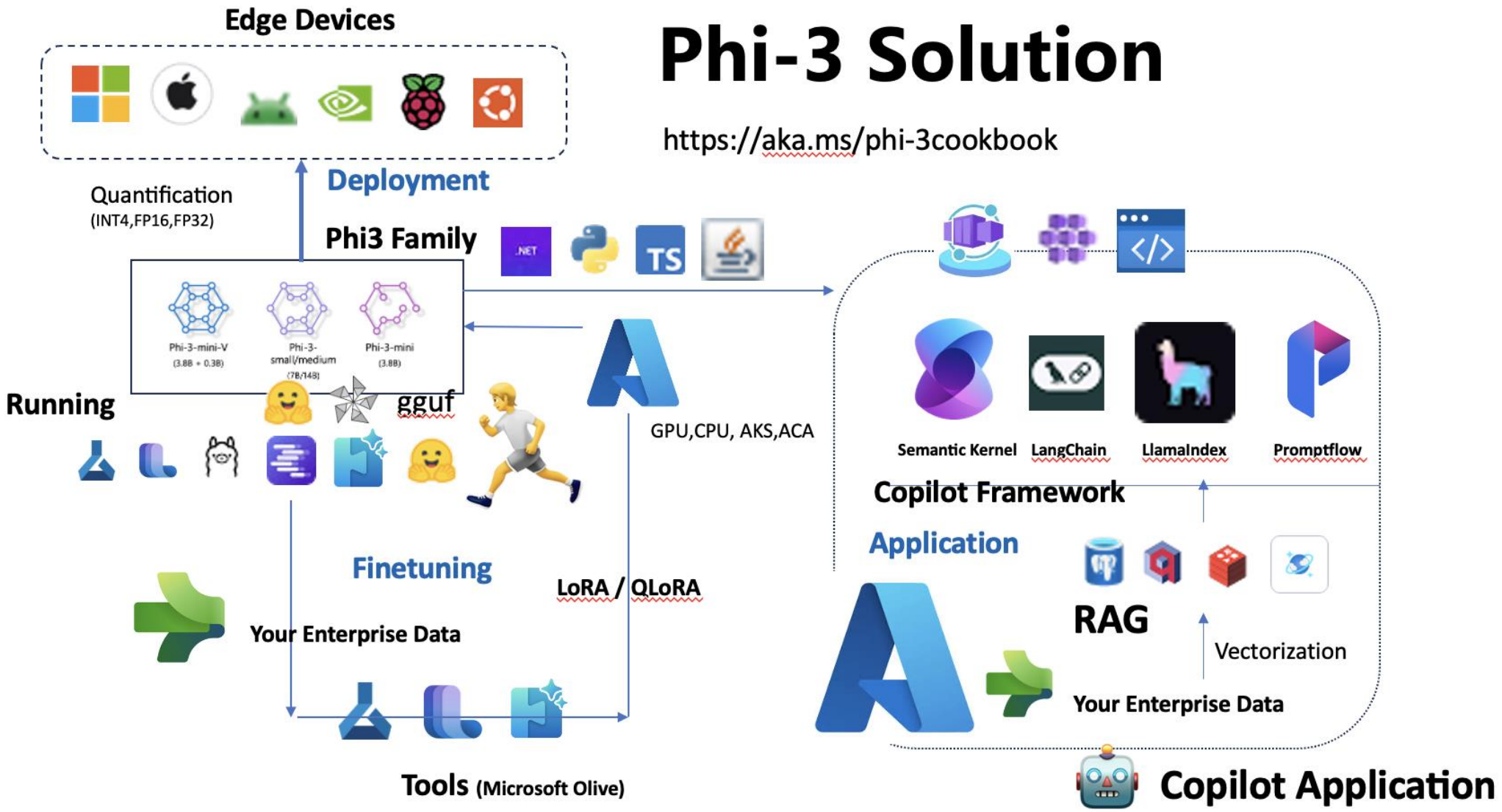
DEMO

05

Create your Copilot Solution

Phi-3 Solution

https://aka.ms/phi-3cookbook



DEMO

Get started with Phi-3

- Learn more: Phi-3 Open Models - Small Language Models | Microsoft Azure
- Phi-3 Cookbook https://aka.ms/phi-3cookbook
- .NET website : https://dotnet.microsoft.com/en-us/

Get .NET 9

Download .NET 9
aka.ms/get-dotnet-9

Thank you