Running AI Model Locally & Microsoft Sematic Kernel for C#

Carlos Bolivar
Principal Software Engineer

WHO AM I?

I work as a Principal Software Engineer at DASI (Aircraft Company) and have more than 18 years of experience in the software industry, primarily focused on Microsoft dotnet products.

AGENDA

01 BASIC CONCEPTS

02 RUN AI MODEL LOCALLY

03 SEMATIC KERNEL

04 DEMO

05 Q&A

BASIC CONCEPTS

What is a LLM?

A large language model (LLM) is a type of machine learning model desgined for natural language processing task such as language generation.

Key Features

- Natural Language Understanding (NLU)
- Text Generation
- Code Writing
- Summarization
- Conversational Abilities
- Question Answering
- Data extractions
- Text Search (RAG)

Major LLM

- ChatGPT (OpenAI)
- DeepSeek R1
- Gemini (Google)
- Llama (Meta AI)
- Mistral (Mistral AI SAS)
- Claude (Antropic)
- Phi-3 (Microsoft)
- Qwen (Alibaba)

CLOUD-BASED PLATFORMS

Multiples Models Deploy

Azure AI Foundry	https://ai.azure.com
Github Marketplace	https://github.com/marketplace
Together AI	https://www.together.ai
Hugging Face	https://huggingface.co
Amazon Bedrock	https://aws.amazon.com/bedrock

Single Model Deploy

OpenAI	https://platform.openai.com/docs/overview
Google AI Studio	https://aistudio.google.com
Deepseek	https://platform.deepseek.com
Llama	https://www.llama.com
Mistral	https://mistral.ai

ON-PERMISES & SELF-HOSTED PLATFORMS

Ollama

Ollama is an open-source command-line tool that lets developers run large language models (LLMs) locally.

It provides a lightweight framework for downloading, running, and managing LLMs directly on your machine, with minimal setup.

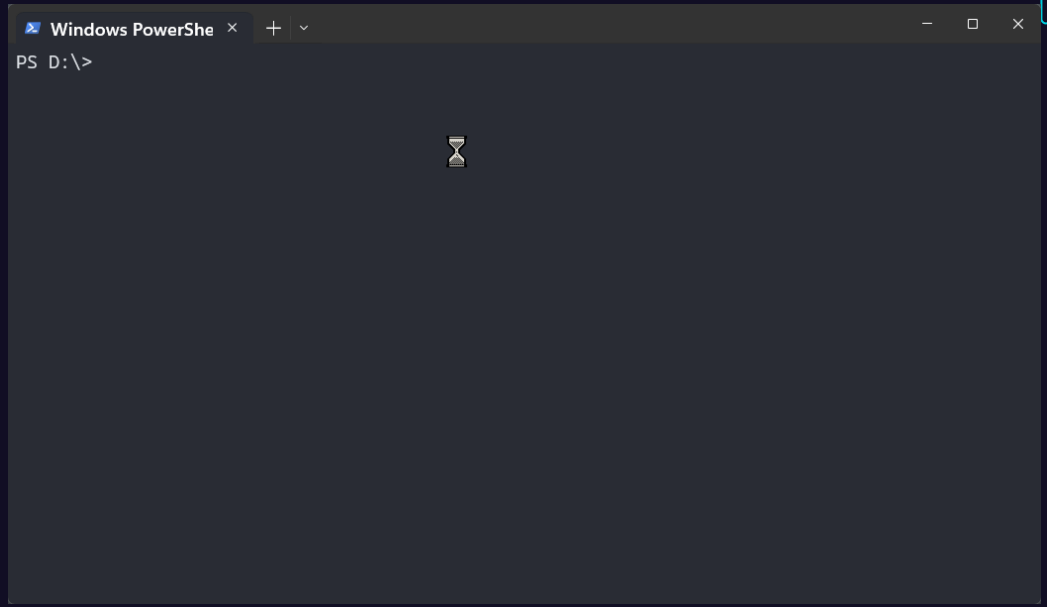
https://ollama.com/

RUN MODEL LOCALLY WITH OLLAMA

MODEL LIBRARY

Model
Gemma 3
Claude
DeepSeek-R1
Llama 3
Phi 4
Mistral
.. And more

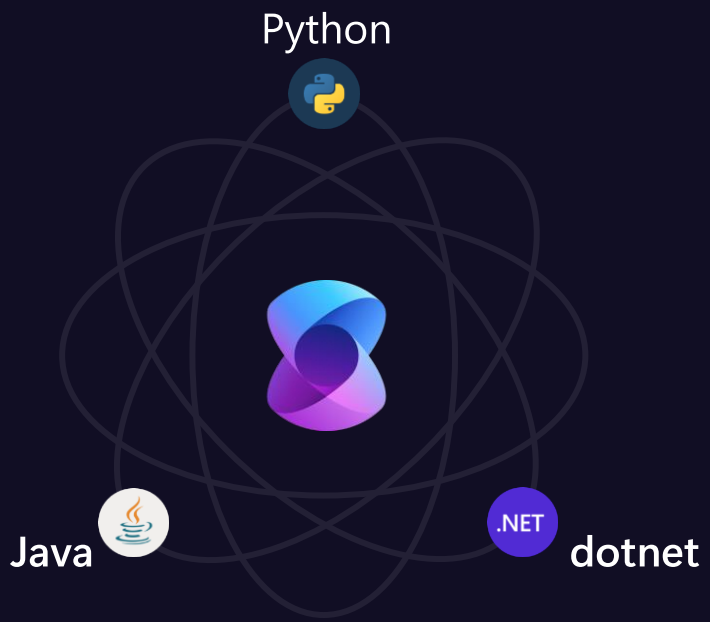
RUNNING A MODEL LOCALLY



MICROSOFT SEMATIC KERNEL

WHAT IS IT?

An open-source SDK that lets you easily build agents that call your existing code. As a highly extensible framework, it can be used with models from OpenAI, Azure OpenAI, Hugging Face, and more!



https://learn.microsoft.com/en-us/semantic-kernel/overview/

AI SERVICES OF SEMATIC KERNEL

Services	C#	Python	Java
Chat Completion	✅	✅	✅
Text Generation	✅	✅	✅
Embedding Generation (experimental)	✅	✅	✅
Text-to-image (experimental)	✅	✅	❌
Image-to-text (experimental)	✅	❌	❌
Text-to-audio	✅	✅	❌
Audio-to-text	✅	✅	❌

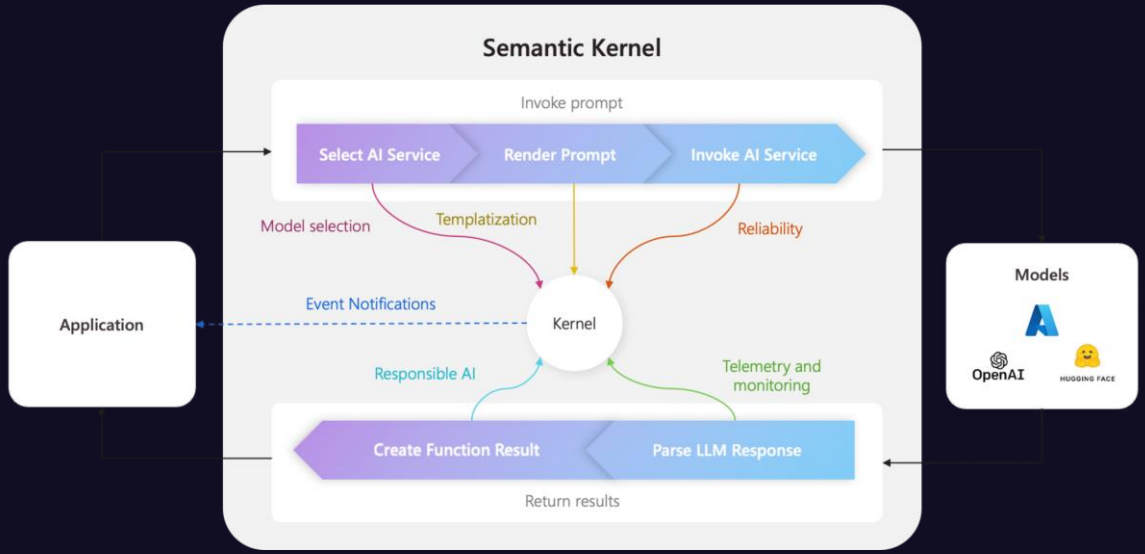
https://learn.microsoft.com/en-us/semantic-kernel/overview/

SEMATIC KERNEL – THE KERNEL

WHAT IS THE KERNEL?

The kernel is the central component of Semantic Kernel. It has all of the services and plugins necessary to run both native code and AI services, it is used by nearly every component within the Semantic Kernel SDK to power your agent.

"All from a single place…"



https://learn.microsoft.com/en-us/semantic-kernel/overview/

SEMATIC KERNEL - COMPONENTS

SERVICES

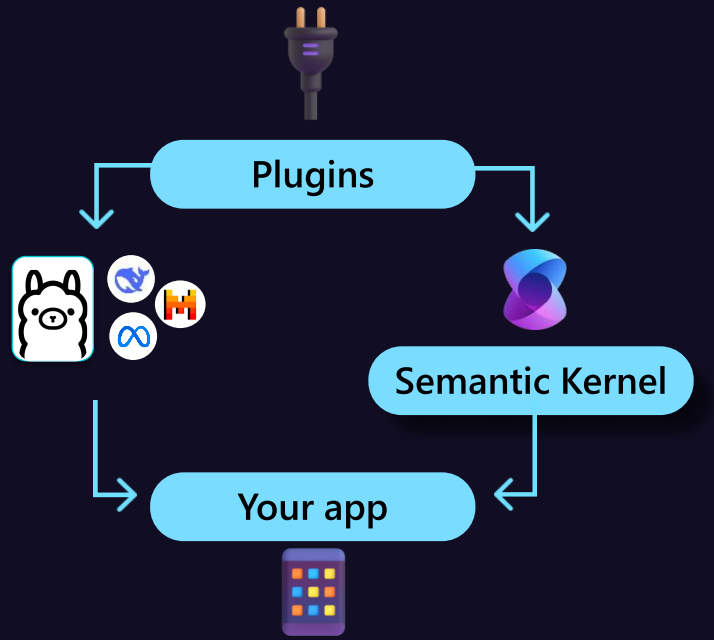
These include both AI services and additional services essential for operating your application.

This was designed based on the Service Provider pattern used in .NET to facilitate dependency injection across all languages.

PLUGINS

These are the components that are used by your AI services and prompt templates to perform work.

For example, can use plugins to retrieve data from a database or call an external API to perform actions.



https://learn.microsoft.com/en-us/semantic-kernel/overview/

DEMO

THANKS!

Do you have any questions?

https://www.linkedin.com/in/cbolivar82/en

https://github.com/cbolivar82/ollama-sematic-kernel

SCAN ME!