Azure Meetup
TORINO

PA EXPERTISE
RETELIT GROUP

#GlobalAzureTorino

INTR3

TD SYNNEX

BY COMMUNITY • FOR COMMUNITY

#GLOBALAZURE

2024

TORINO

Azure Meetup
TORINO

#GlobalAzureTorino

Utilizzare i propri dati con l'AI grazie a Azure OpenAI, Semantic Kernel e Kernel Memory

Matteo Pagani
Cloud Solution Architect @ Microsoft
Modern Work – AI, Apps & Ecosystem

@qmatteoq @qmatteoq @matteopagani www.developerscantina.com

#GLOBALAZURE 2024

INTR3 TD SYNNEX PA EXPERTISE RETELIT GROUP

What is a copilot?

An experience using generative AI to assist humans with complex cognitive tasks.

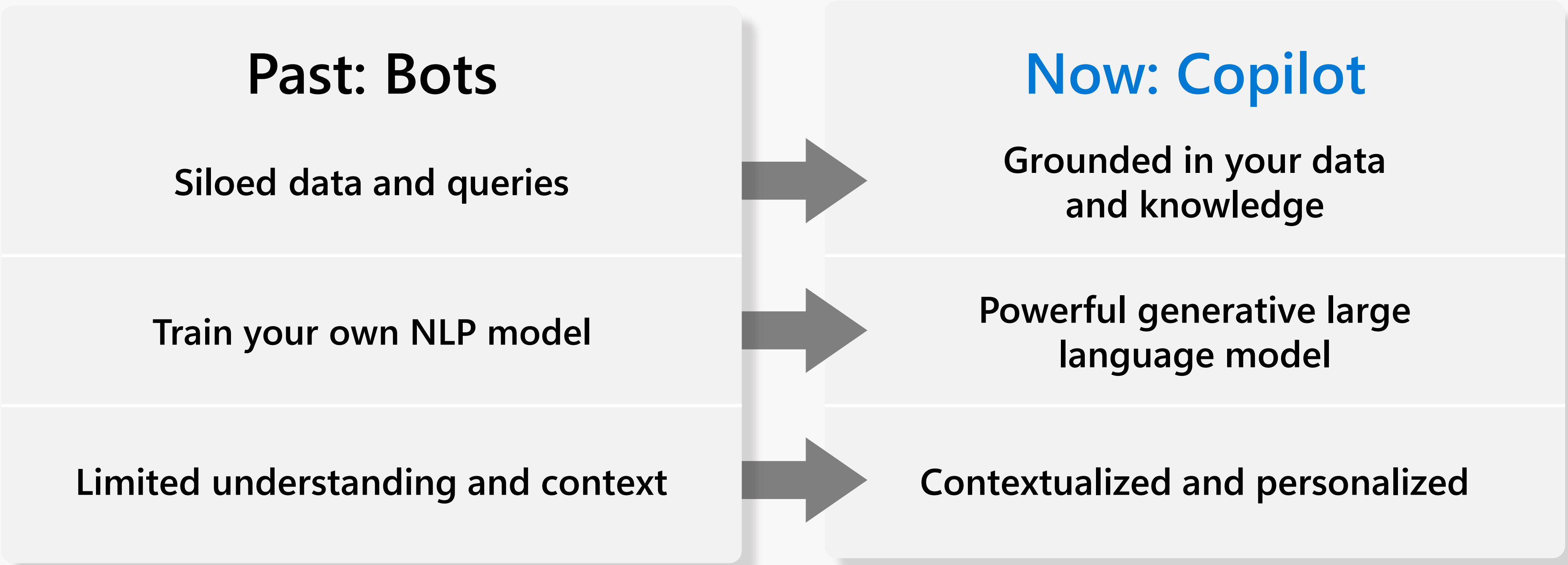
Chat using natural language (and code)

You determine, guide, and approve the output

Its value increases with complexity

Build an AI assistant to help

Copilot: Intelligent AI assistants that use generative AI and large language models to assist humans with complex cognitive tasks.



Past: Bots	Now: Copilot
Siloed data and queries	Grounded in your data and knowledge
Train your own NLP model	Powerful generative large language model
Limited understanding and context	Contextualized and personalized

Transform your business with conversational AI

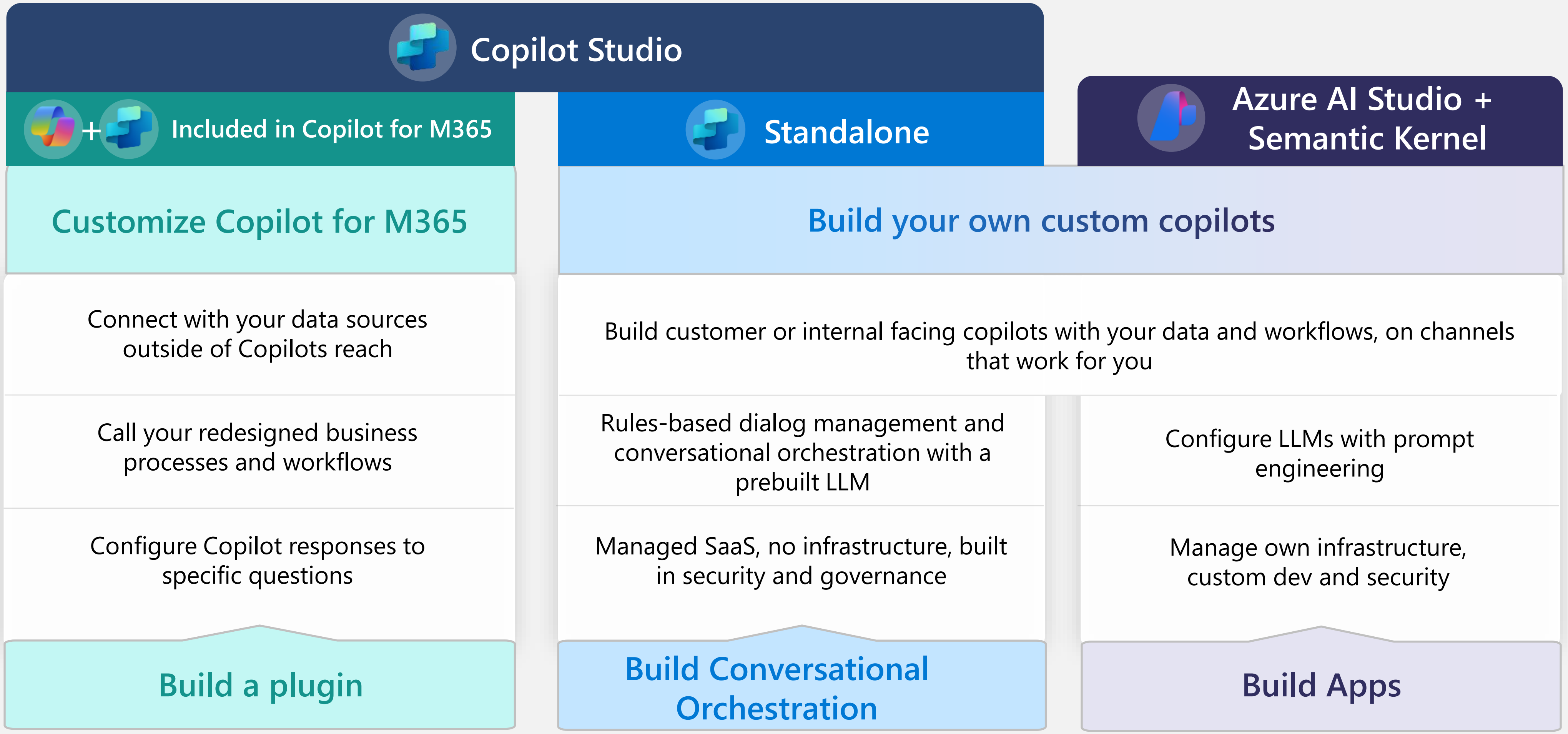
- Enrich employee experiences
- Optimize business processes
- Improve efficiency & time to value
- Bend the curve on innovation
- Reinvent customer engagement
- Reduce development costs & risks

Build your own AI capability

Your apps

Your data

Different building journeys for different needs



Copilot Studio

Included in Copilot for M365

Standalone

Azure AI Studio + Semantic Kernel

Customize Copilot for M365

Build your own custom copilots

Connect with your data sources outside of Copilots reach

Call your redesigned business processes and workflows

Configure Copilot responses to specific questions

Build a plugin

Build customer or internal facing copilots with your data and workflows, on channels that work for you

Rules-based dialog management and conversational orchestration with a prebuilt LLM

Managed SaaS, no infrastructure, built in security and governance

Build Conversational Orchestration

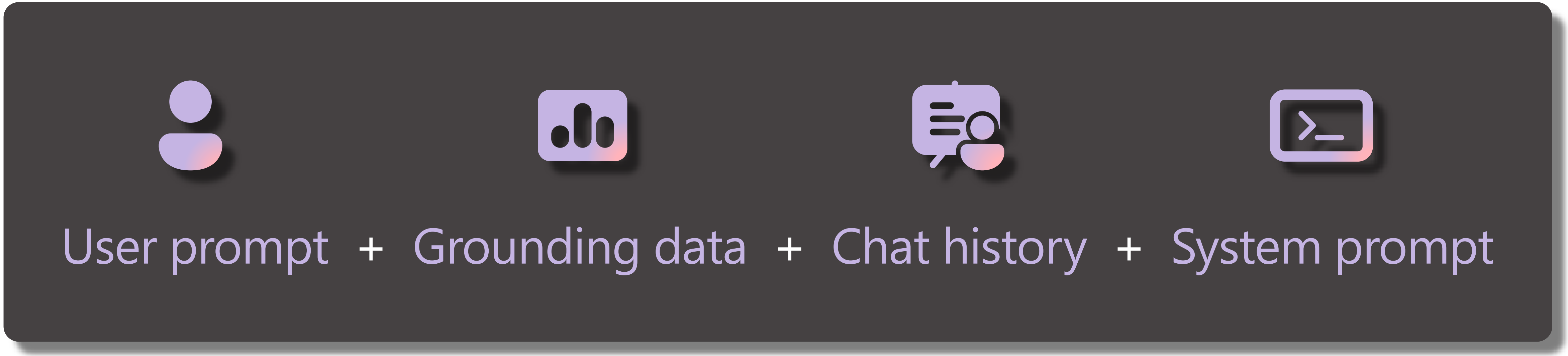
Configure LLMs with prompt engineering

Manage own infrastructure, custom dev and security

Build Apps

Retrieval Augmented Generation (RAG)

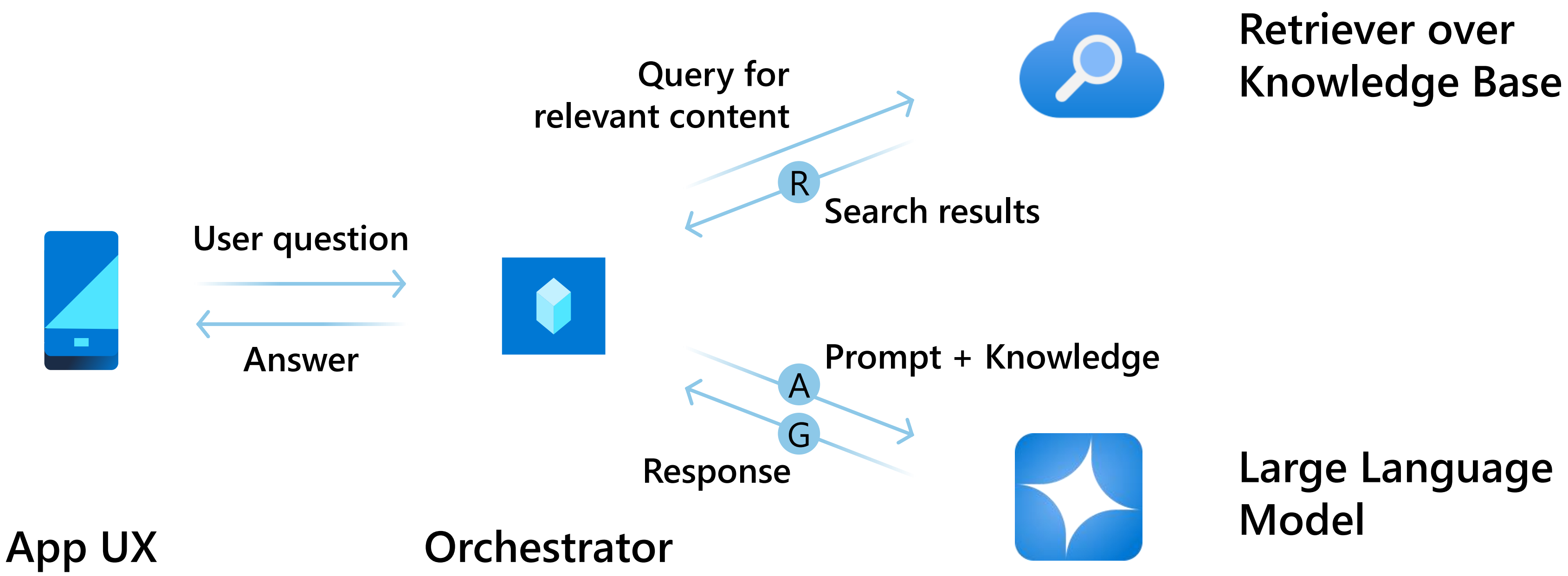
- Enable the LLM to perform tasks on your data
- No need to fine tune the model
- The application identifies the most relevant data based on the user query and sends them together with the user prompt



User prompt + Grounding data + Chat history + System prompt

Retrieval Augmented Generation (RAG)

Anatomy of the workflow



App UX

User question

Answer

Orchestrator

Query for
relevant content

R

Search results

Retriever over
Knowledge Base

Prompt + Knowledge

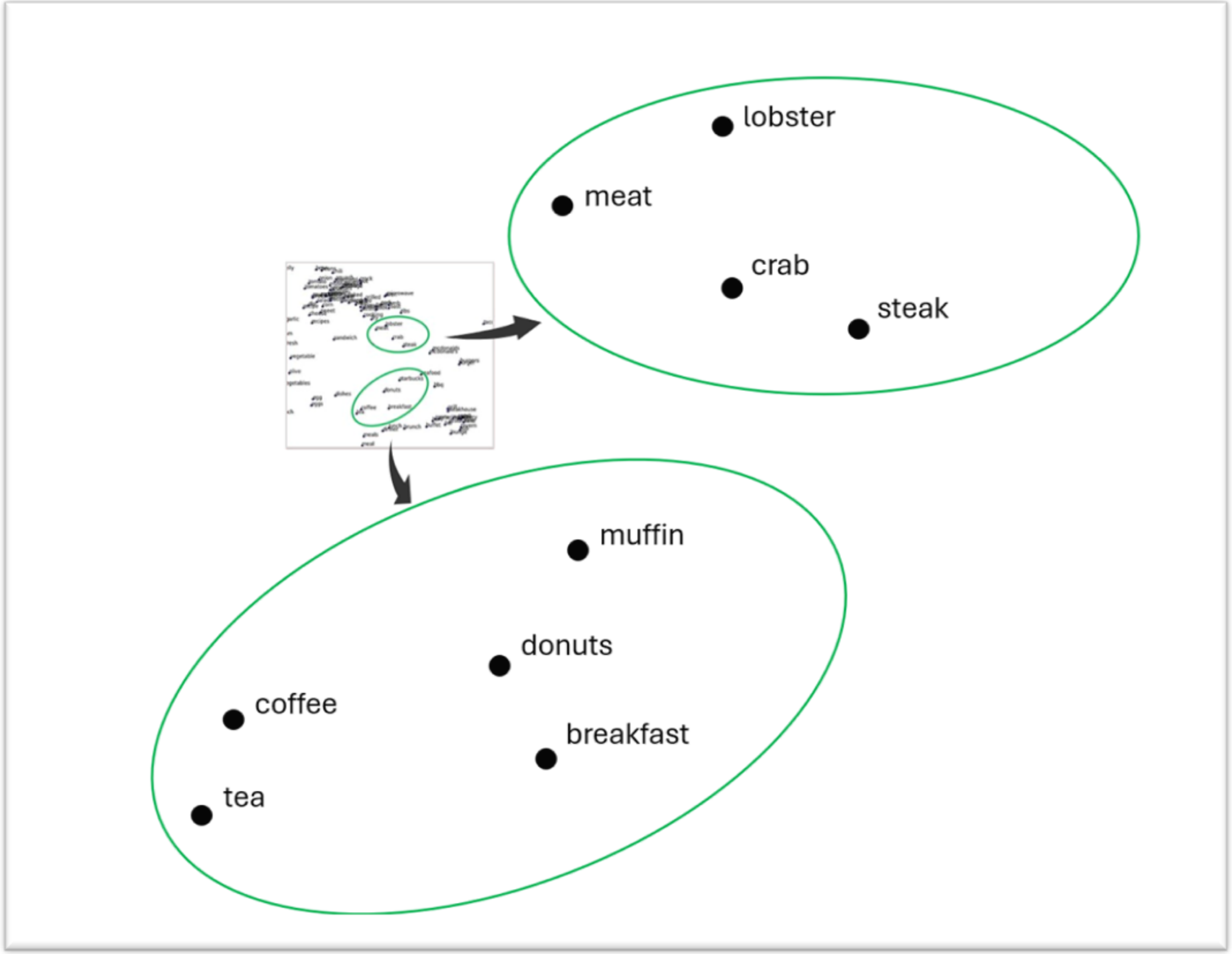
A

G

Response

Large Language
Model

Vector databases



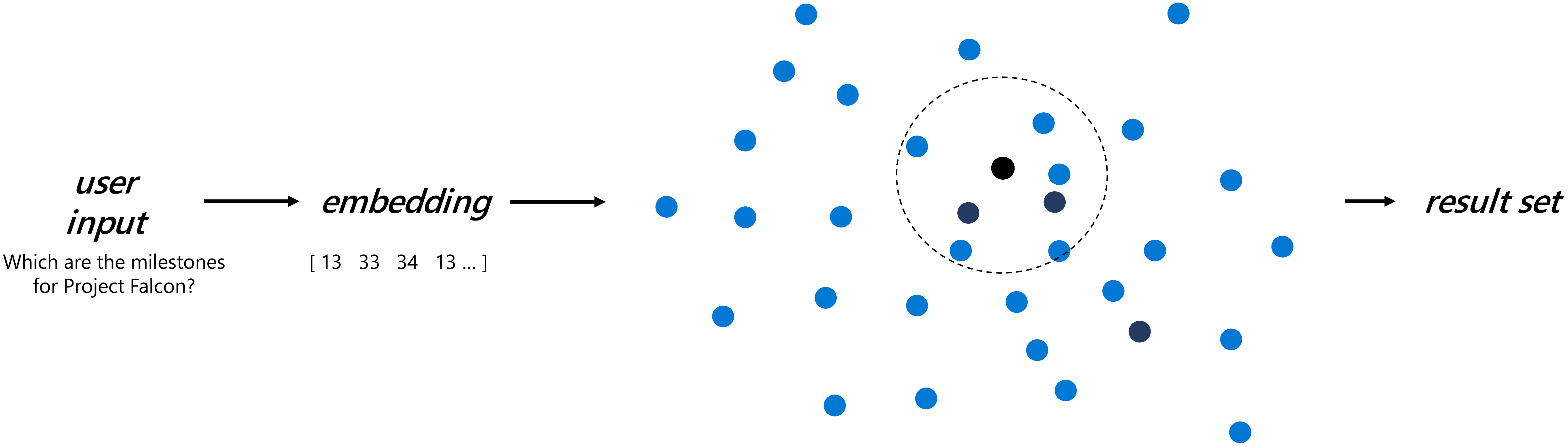
- A vector is a numerical representation of a data point (e.g., word, image or pixel) and it's arranged with close numbers placed in proximity to one another to represent similarity.
- Vectors are stored multi-dimensional spaces where semantically similar data points are clustered together in the vector space.

Benefits

- This approach provides for fast and accurate similarity search and retrieval of data based on their vector distance or similarity.

Similarity Search with embeddings

Once you encode your content as embeddings, you can then get an embedding from the user input and use that to find the most semantically similar content.



user input

Which are the milestones for Project Falcon?

embedding

[13   33   34   13 ...]

result set

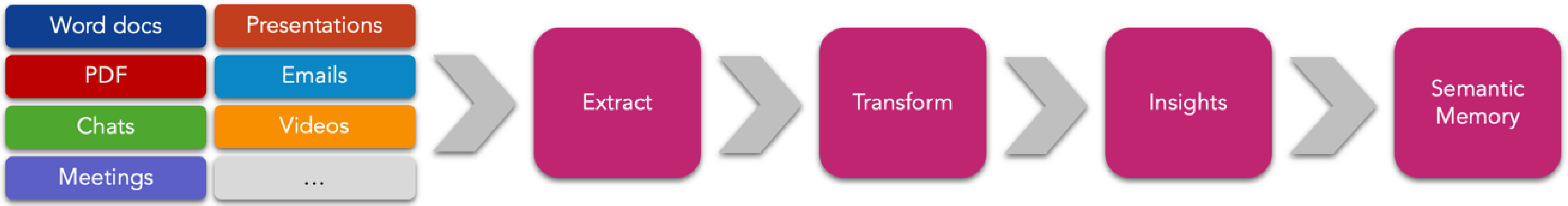
Implementing RAG is expensive

- Build an orchestrator engine:
    - Generating a search intent from a prompt
    - Performing a search on the vector database
    - Including the relevant data into the prompt
    - Send the prompt to the LLM
- Convert your data into vectors
- Implement chunking to process long documents

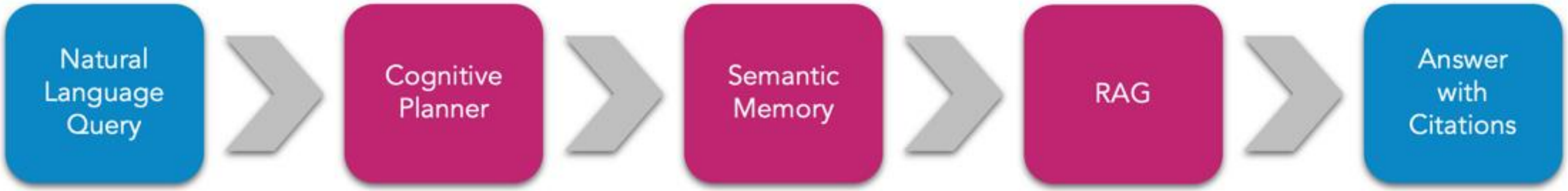
Say hello to Kernel Memory

Simplifies the RAG implementation by automatically taking care of:

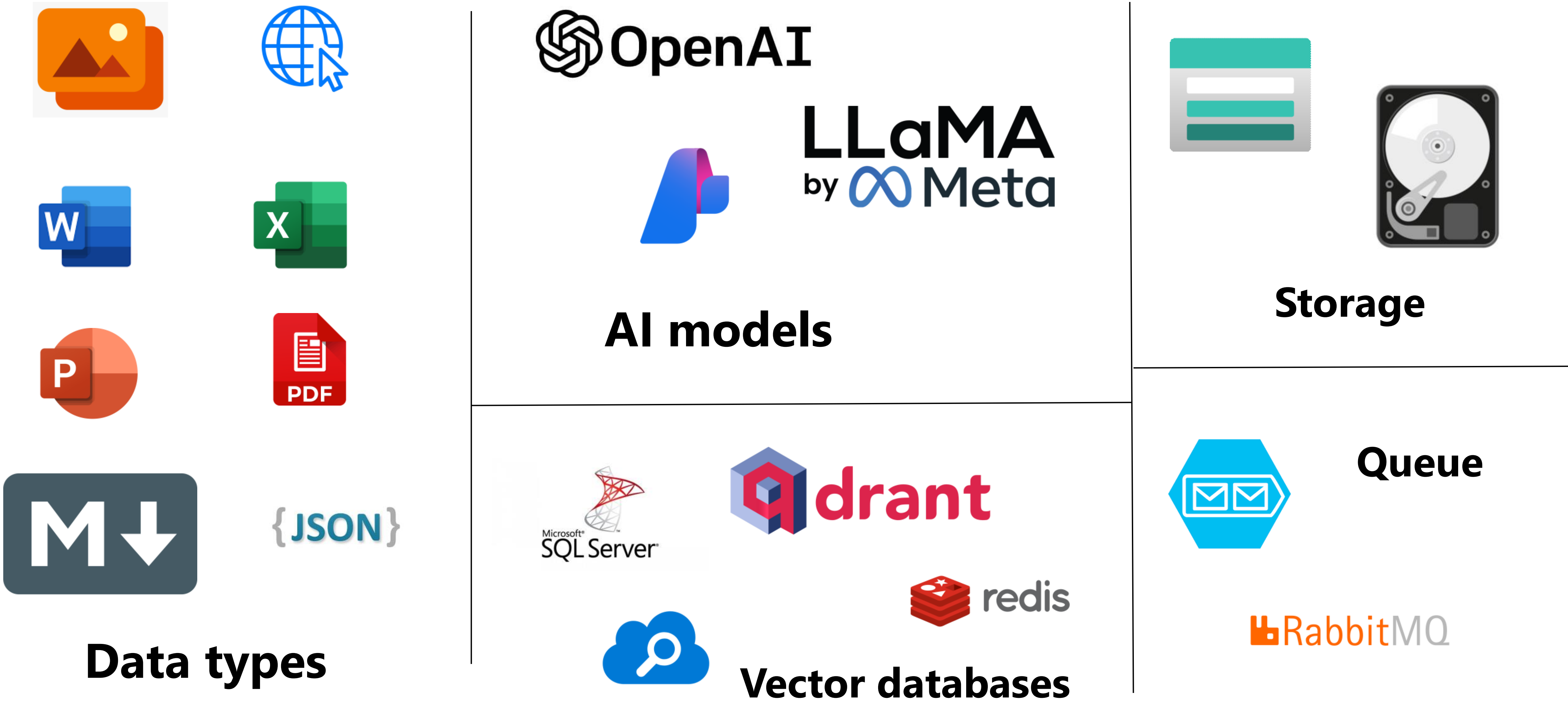
- Converting various data types into embeddings



- Use natural language to ask questions about the ingested content:



Kernel Memory



PDF

{JSON}

Data types

OpenAI

LLaMA by Meta

AI models

Microsoft SQL Server

qdrant

redis

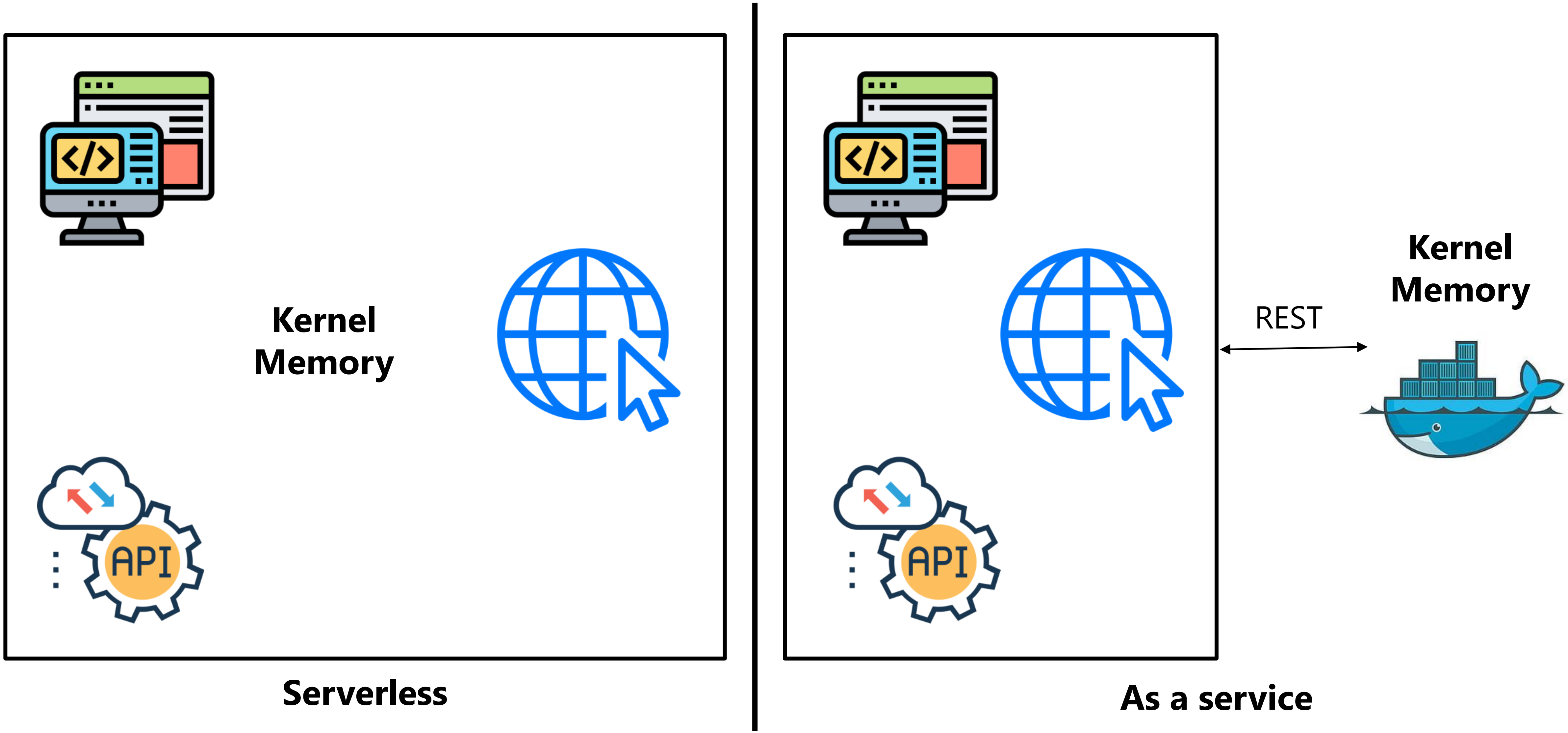
Vector databases

Storage

Queue

RabbitMQ

Two approaches



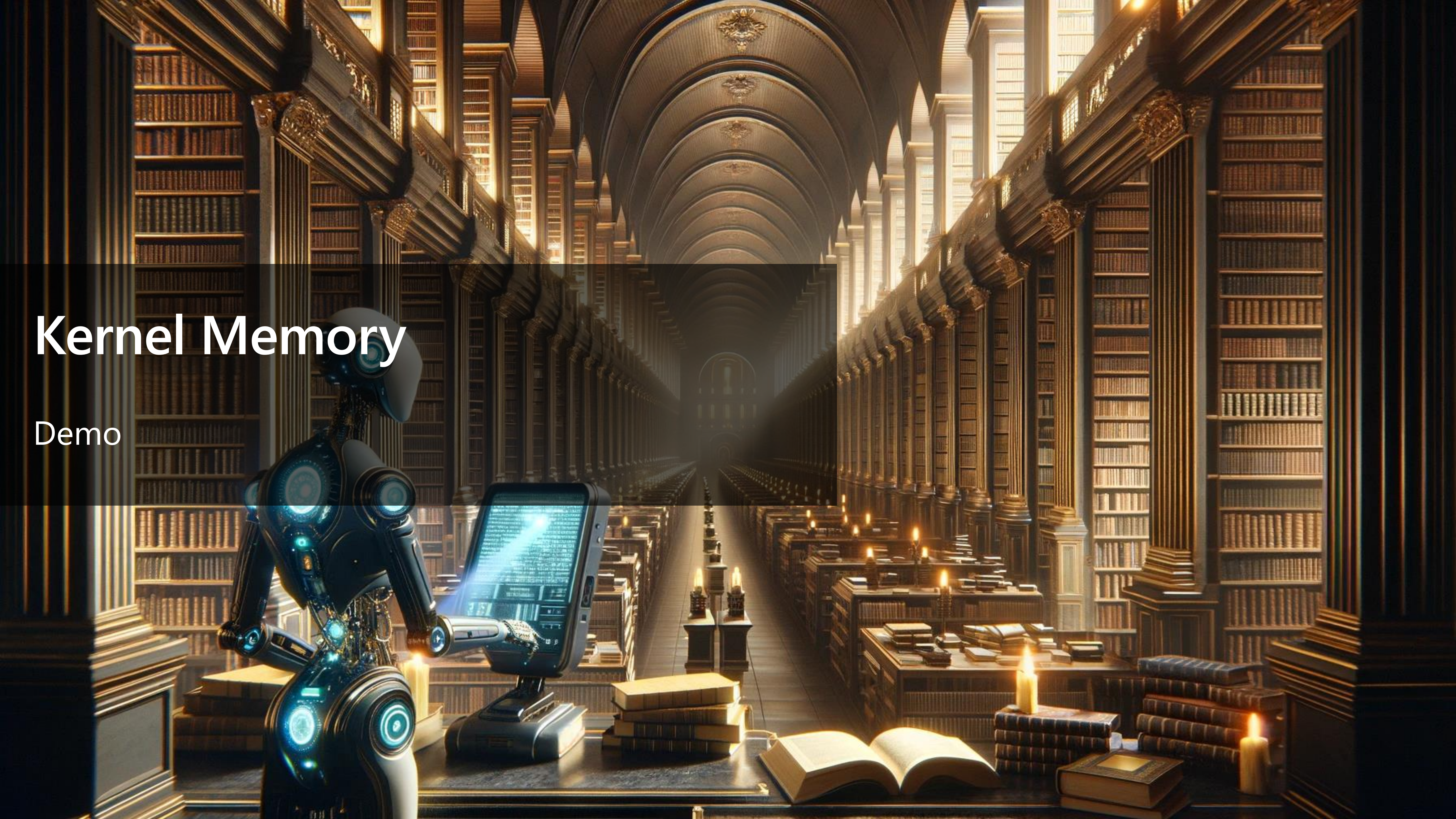
Kernel Memory

Serverless

REST

Kernel Memory

As a service

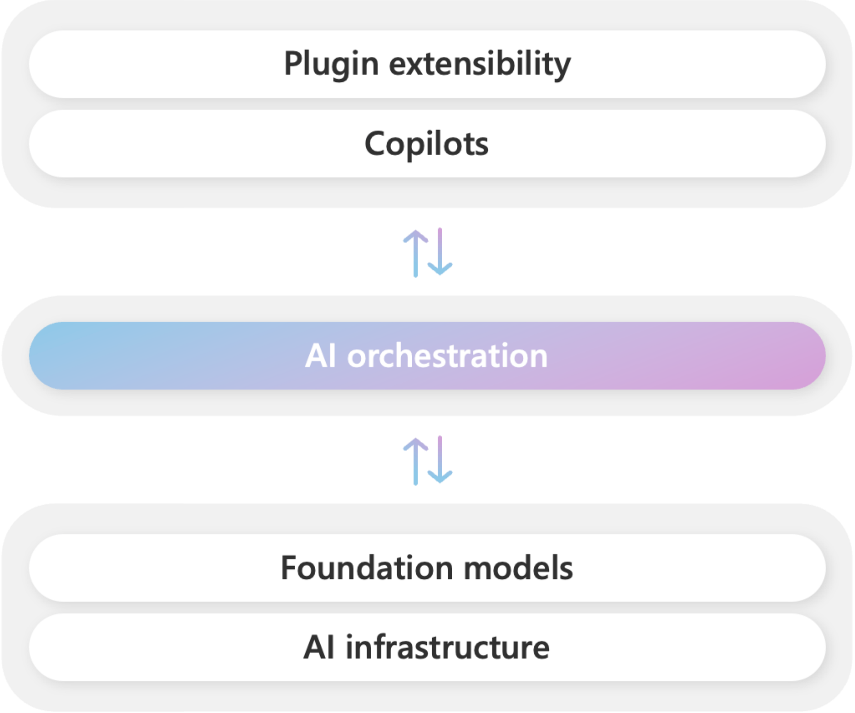


Kernel Memory

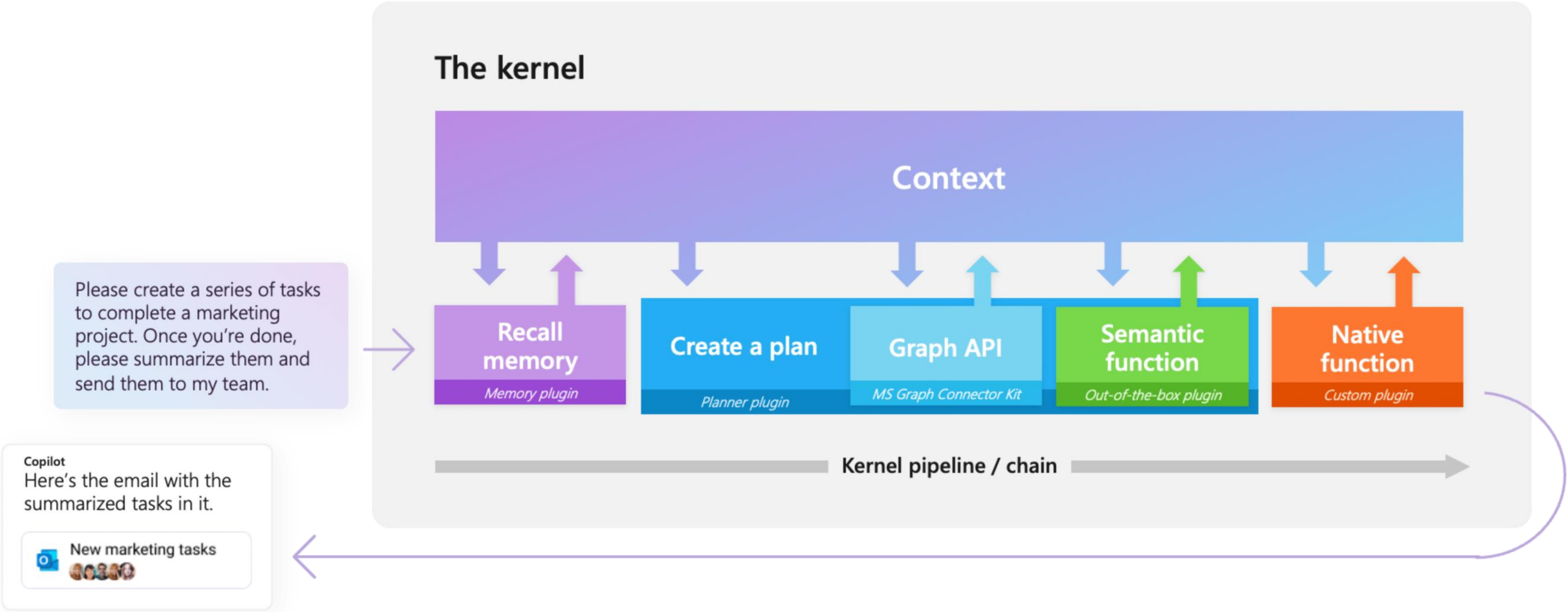
Demo

Introducing Semantic Kernel

- Open-source SDK to build AI agents
- Support for Open AI, Azure Open AI and Hugging Face
- Available in C#, Python and Java
- Support for plugins
- Automatic orchestration with function calling and planners
- Memory and embeddings



Building AI agents with function calling and planners



Please create a series of tasks to complete a marketing project. Once you're done, please summarize them and send them to my team.

The kernel

Context

Recall memory

Memory plugin

Create a plan

Planner plugin

Graph API

MS Graph Connector Kit

Semantic function

Out-of-the-box plugin

Native function

Custom plugin

Kernel pipeline / chain

Copilot

Here's the email with the summarized tasks in it.

New marketing tasks

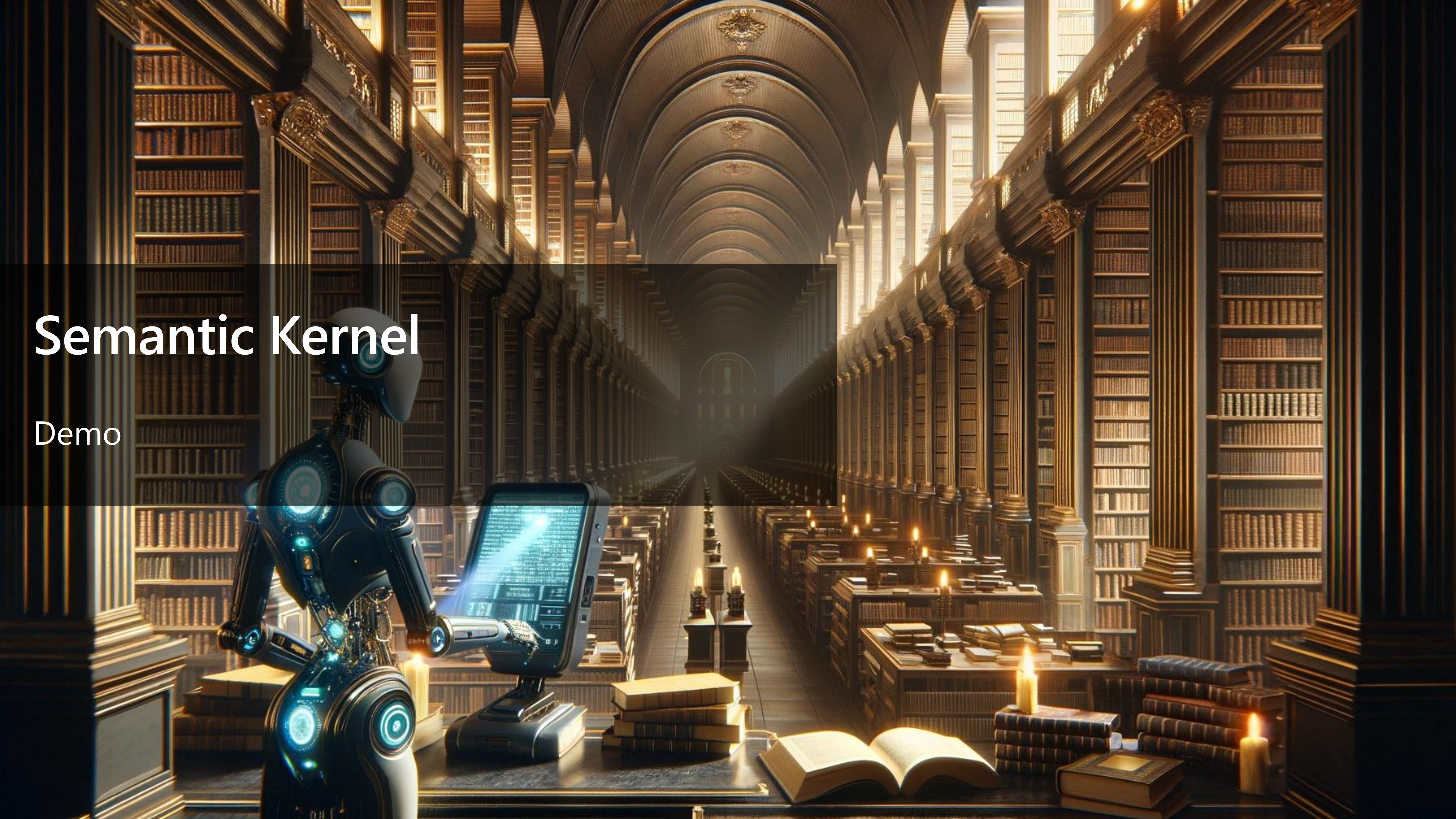
Extend your workflows with plugins

- Integrate external services into your AI workflows
- Reuse prompts and functions across multiple projects

Prompt Functions

Native Functions

OpenAI Plugins

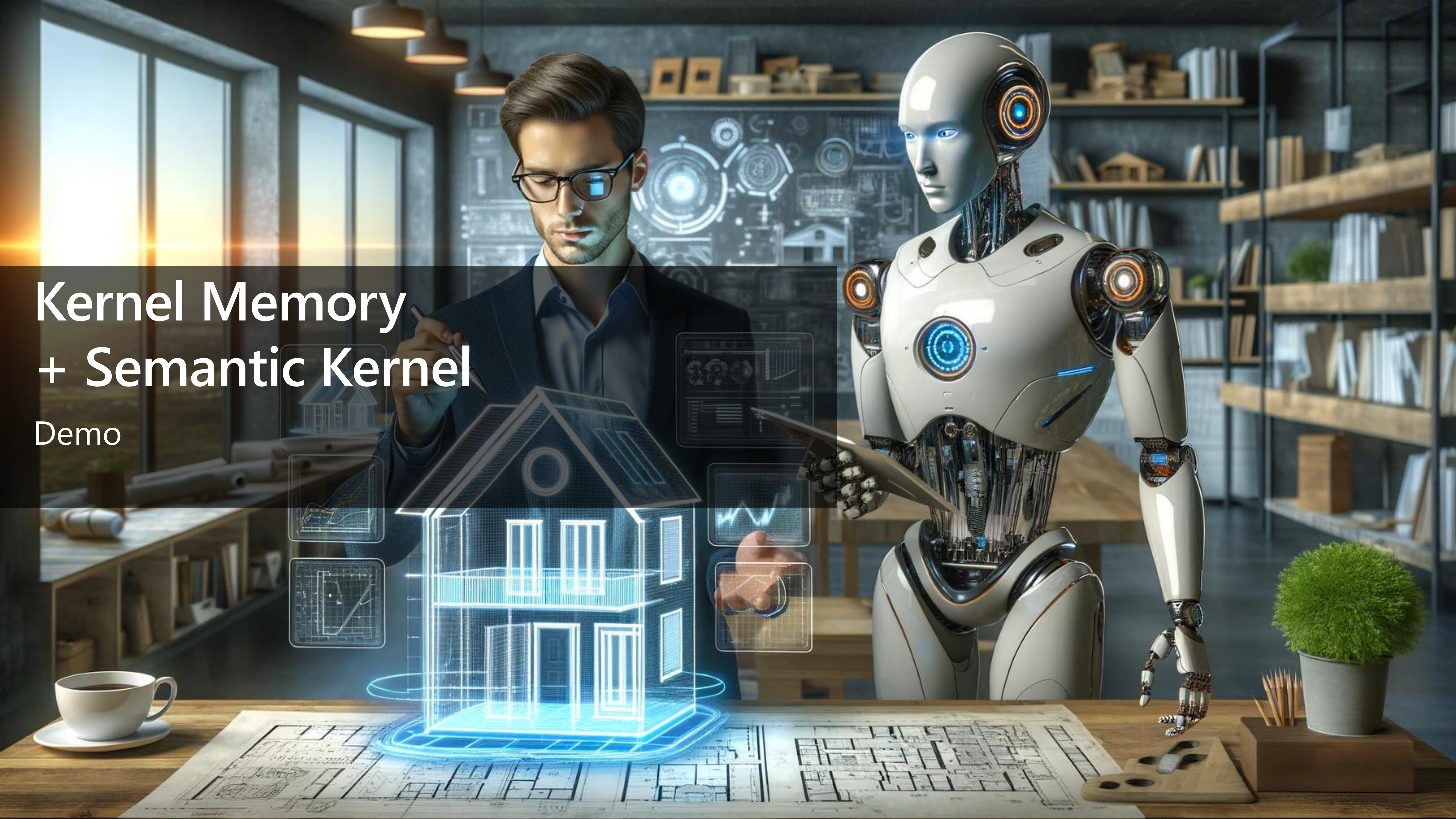


Semantic Kernel

Demo

Kernel Memory + Semantic Kernel

- Kernel Memory is very powerful, but the RAG experience is limited to Q&A
- What if you want to use your data for more complex AI workflows?
  - Combining answers from your data with external services
  - Providing a continuous chat experience with context and history
- The Kernel Memory plugin for Semantic Kernel enables integrating your data into your AI workflows



Kernel Memory
+ Semantic Kernel

Demo

Wrapping up

- Integrating LLM in your apps is simple, but when you to perform more complex workflows many challenges arises
- Semantic Kernel is an open-source library from Microsoft that simplifies the orchestration of AI workflows
- A very common task when you integrate LLMs is performing operations on your data
- Kernel Memory simplifies the RAG implementation, by supporting indexing and querying your documents using an LLM

Resources

- My blog: https://www.developerscantina.com
- Semantic Kernel website: https://learn.microsoft.com/en-us/semantic-kernel/overview/
- Semantic Kernel repository: https://github.com/microsoft/semantic-kernel
- Kernel Memory repository: https://github.com/microsoft/kernel-memory
- Demos: https://github.com/qmatteoq/SemanticKernel-Demos

Azure Meetup
TORINO

#GlobalAzureTorino

BY COMMUNITY • FOR COMMUNITY

#GLOBALAZURE

2024

TORINO

TD SYNNEX

INTR3

PA EXPERTISE
RETELIT GROUP