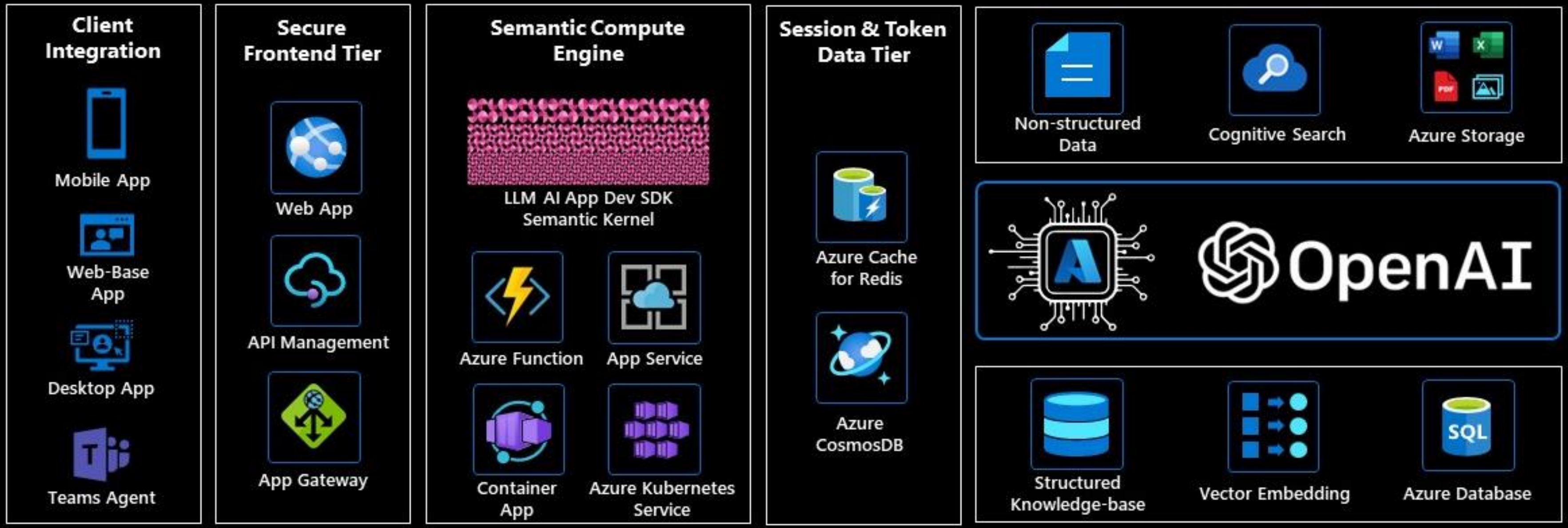
Tech Community Sharing

Azure OpenAI 集成云原生应用

Overview

- Architecture
- Details
- Semantic Kernel
- Demo
- Q & A

Enterprise GPT-ize Intelligent Application Architecture



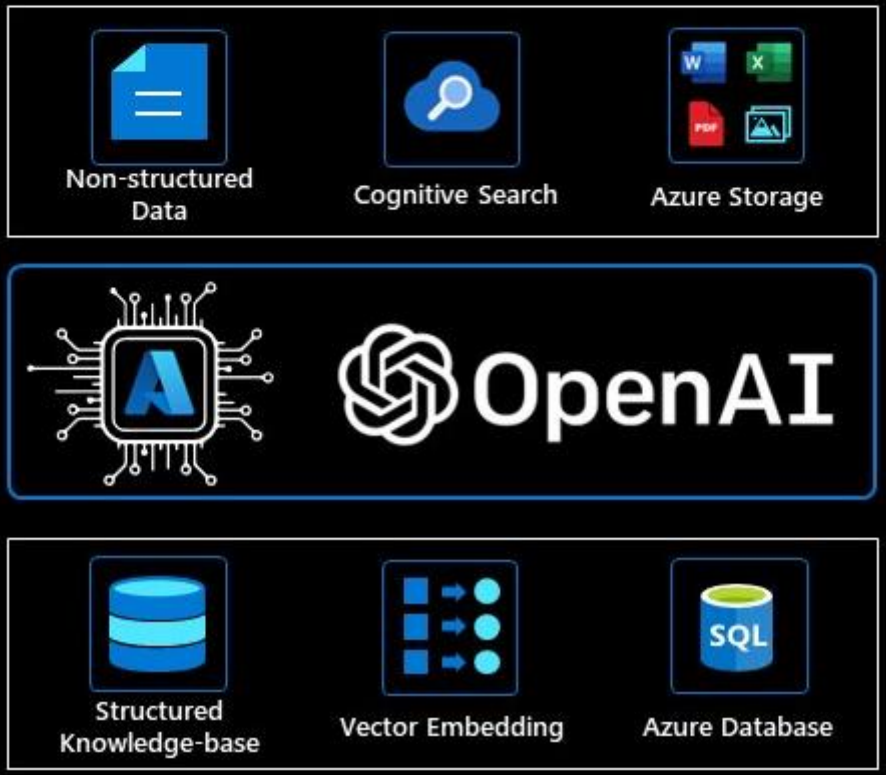
A prompt engineering approach with LLM AI Dev Framework

Azure-OpenAI-App-Innovation-Workshop

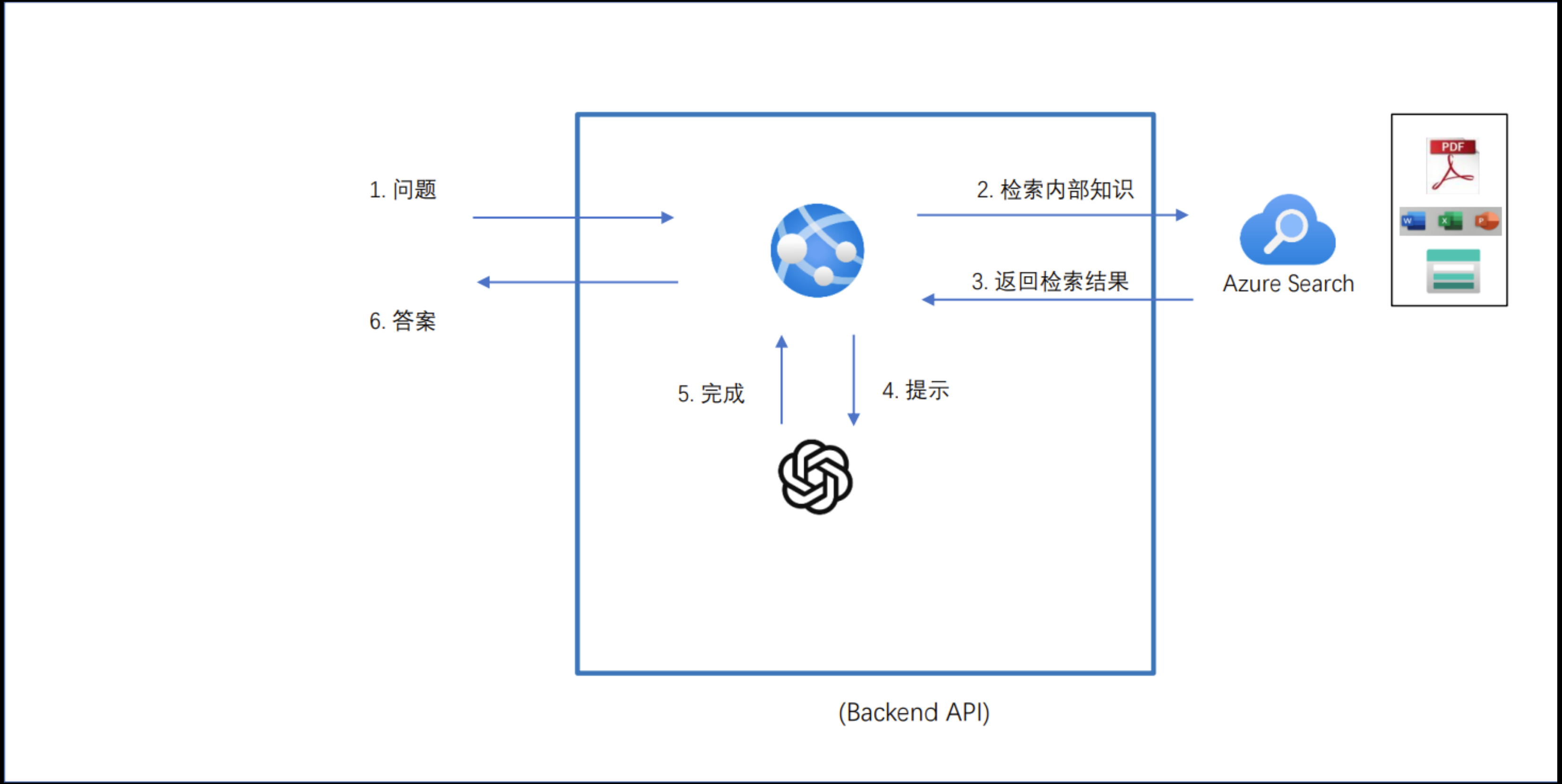
Azure Cognitive Search can quickly index unstructured data such as PDF and WORD files, allowing existing data to be used immediately.

Utilizing Azure Database with vector storage and processing capabilities and combining them with AOAI's embedding vector generation model, the enterprise's existing structured knowledge base can be integrated easily

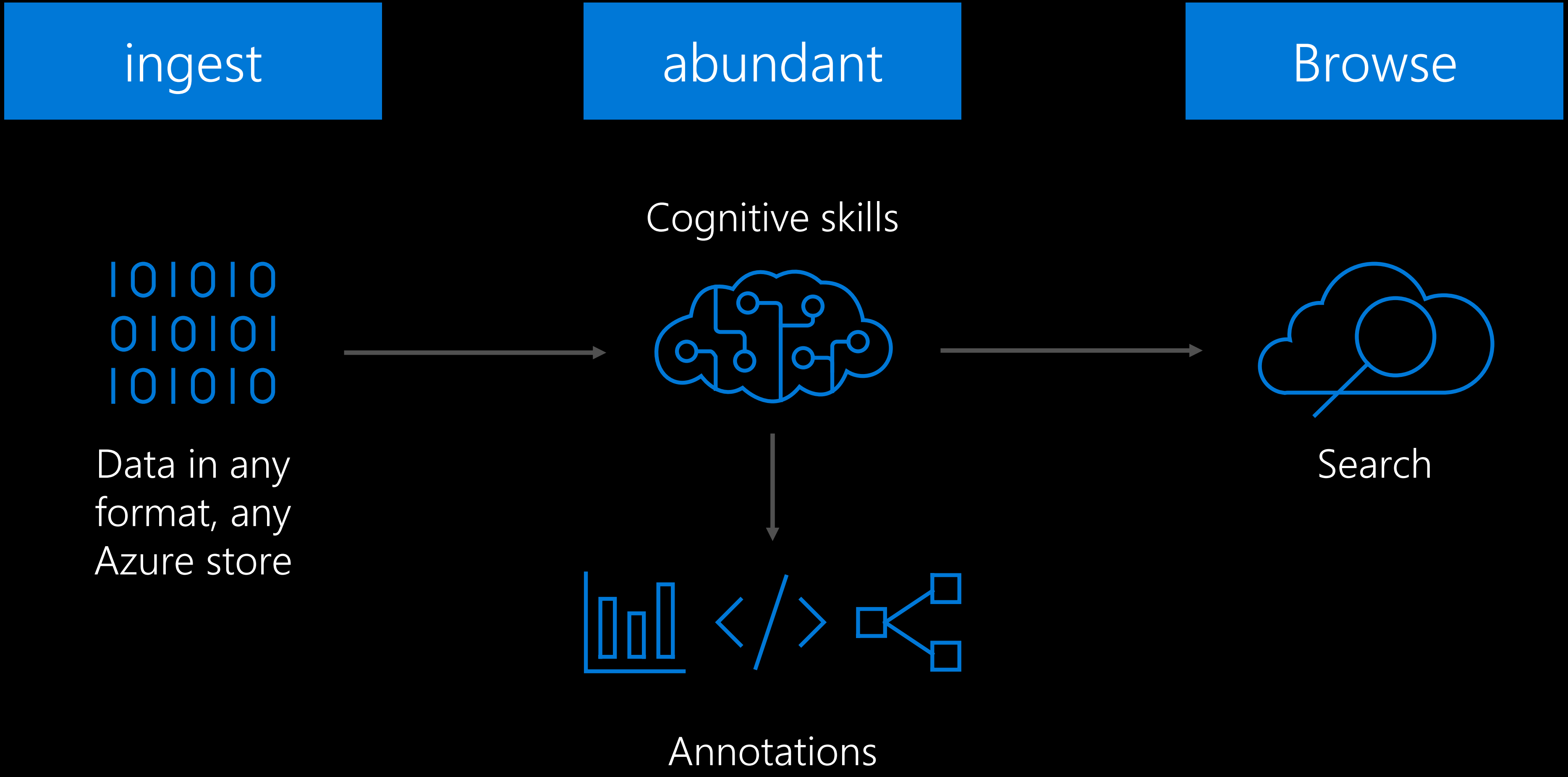
Dual-engine



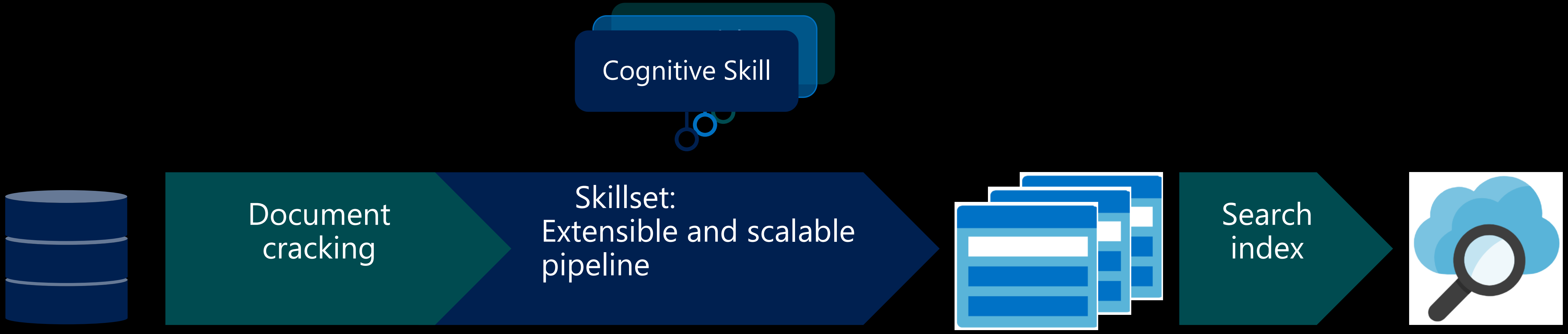
GPT + enterprise knowledge base/data



Azure Cognitive Search



Cognitive Search Architecture



client data

.pdf
.doc
.jpeg
…

Azure has built-in cognitive skills
OCR,
keyword extraction,
Sentiment analyzer, computer vision,
Form Recognizer,
ink recognizer,
entity linking,

Text Analytics
translation
Content Moderator
personalize
QnA manufacturer
speech

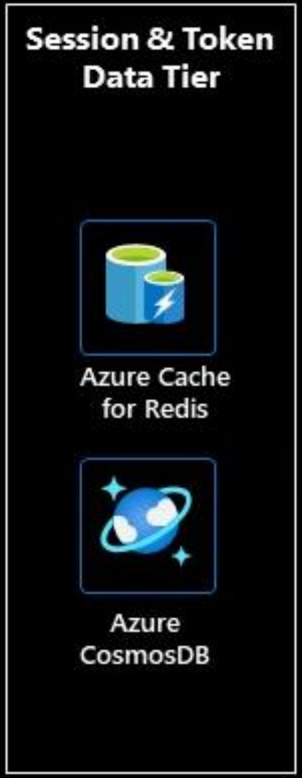
Annotated Documents

Search Index

Web
Mobile
Bots
PowerBI

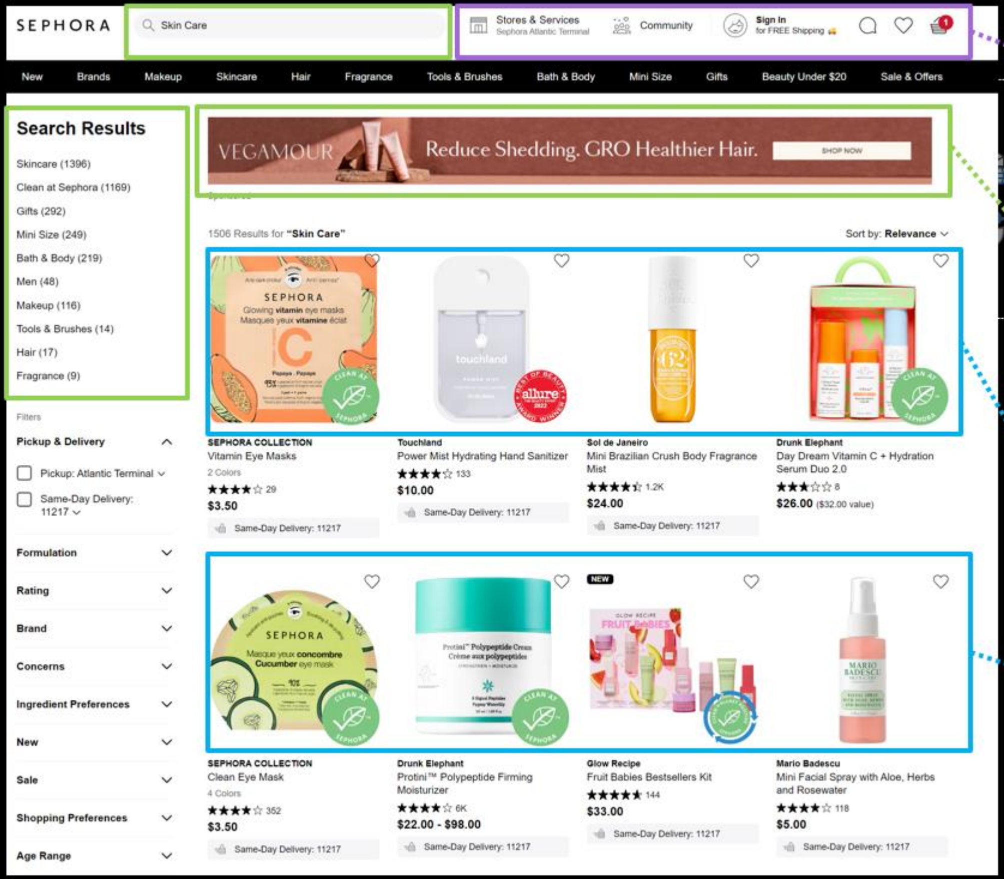
Third Party Enrichers
custom classification model,
custom entity extraction,

Azure machine Learning



- Based on Redis and CosmosDB.
- Add contextual caching, session persistence, prompt persistence, and other capabilities to the application.
- Leave room for future model or engine optimization based on prompts.

How is Azure Cache for Redis Leveraged?



SEPHORA | Skin Care | Stores & Services Sephora Atlantic Terminal | Community | Sign In for FREE Shipping

New | Brands | Makeup | Skincare | Hair | Fragrance | Tools & Brushes | Bath & Body | Mini Size | Gifts | Beauty Under $20 | Sale & Offers

Search Results

- Skincare (1396)
- Clean at Sephora (1169)
- Gifts (292)
- Mini Size (249)
- Bath & Body (219)
- Men (48)
- Makeup (116)
- Tools & Brushes (14)
- Hair (17)
- Fragrance (9)

Filters

Pickup & Delivery

- Pickup: Atlantic Terminal
- Same-Day Delivery: 11217

Formulation | **Rating** | **Brand** | **Concerns** | **Ingredient Preferences** | **New** | **Sale** | **Shopping Preferences** | **Age Range**

VEGAMOUR Reduce Shedding. GRO Healthier Hair. SHOP NOW

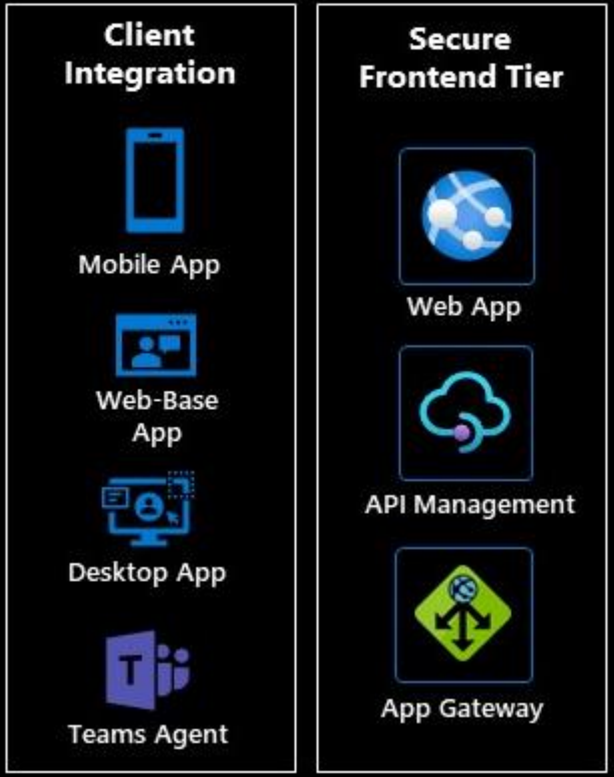
1506 Results for **"Skin Care"** Sort by: **Relevance**

SEPHORA COLLECTION Vitamin Eye Masks, 2 Colors, ★★★★☆ 29, **$3.50**, Same-Day Delivery: 11217	**Touchland** Power Mist Hydrating Hand Sanitizer, ★★★★☆ 133, **$10.00**, Same-Day Delivery: 11217	**Sol de Janeiro** Mini Brazilian Crush Body Fragrance Mist, ★★★★☆ 1.2K, **$24.00**, Same-Day Delivery: 11217	**Drunk Elephant** Day Dream Vitamin C + Hydration Serum Duo 2.0, ★★★☆☆ 8, **$26.00** ($32.00 value)
SEPHORA COLLECTION Clean Eye Mask, 4 Colors, ★★★★☆ 352, **$3.50**, Same-Day Delivery: 11217	**Drunk Elephant** Protini™ Polypeptide Firming Moisturizer, ★★★★☆ 6K, **$22.00 - $98.00**, Same-Day Delivery: 11217	NEW **Glow Recipe** Fruit Babies Bestsellers Kit, ★★★★★ 144, **$33.00**, Same-Day Delivery: 11217	**Mario Badescu** Mini Facial Spray with Aloe, Herbs and Rosewater, ★★★★☆ 118, **$5.00**, Same-Day Delivery: 11217

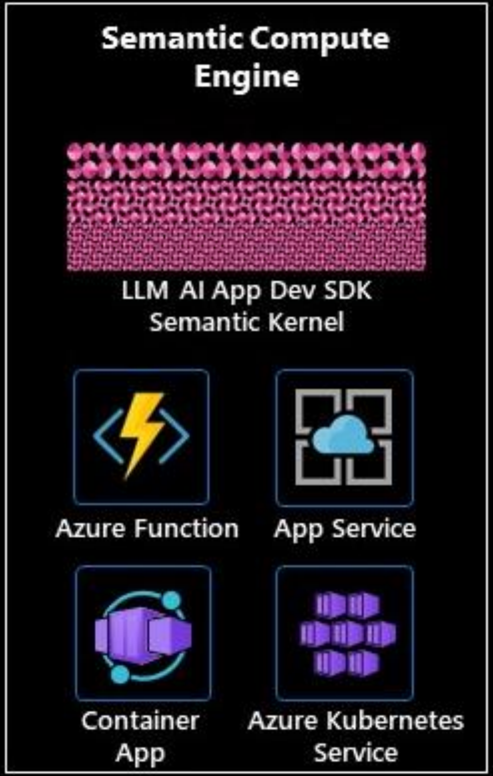
Session Store
Easily resume a user's journey on any mobile or web application by quickly sorting the data of their last session regardless of where they are across the globe

Message Broker Store
Azure Cache for Redis can communicate across all Microservices. In the example, it communicated with its ad server to give a recommended ad for the shopper

Distributed Cache
All images need to be consistently accessed and are static therefore, minimizes load times by keeping all assets into a cache no matter where the user is across the globe



- The API encapsulation, load balancing, and gateway at the front end further enhance application security and reliability.
- Allow the intelligent entity to integrate with various front-end apps in a more secure and stable manner.



- Utilize a flexible semantic compute engine as prompt engine
- Support various deployment forms such as PaaS, Serverless, and containers
- Enable enterprises to optimize the prompt engine with modern LLM AI application development framework
- Elastically scale computing resources according to real-time business needs.

Function as a Service

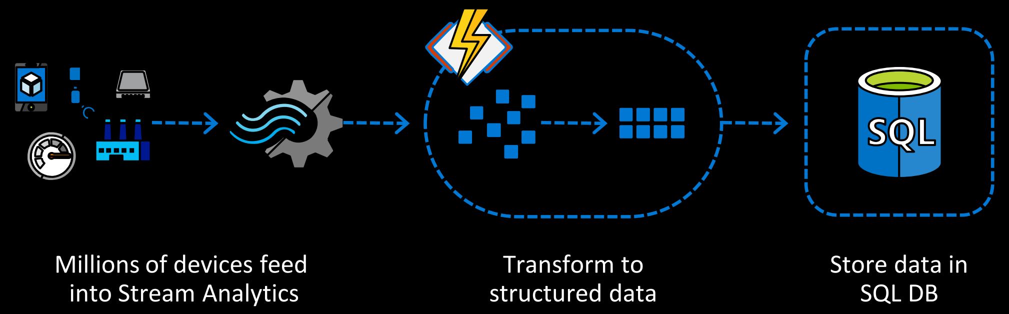
Use serverless code to handle events.

- Writing cloud apps made easy
- Extend functionality based on customer needs
- Develop functions in languages such as C#, Node.js, F#, Python, Java, and more
- Easily schedule event-driven tasks across services
- Exposes the function as an HTTP API endpoint

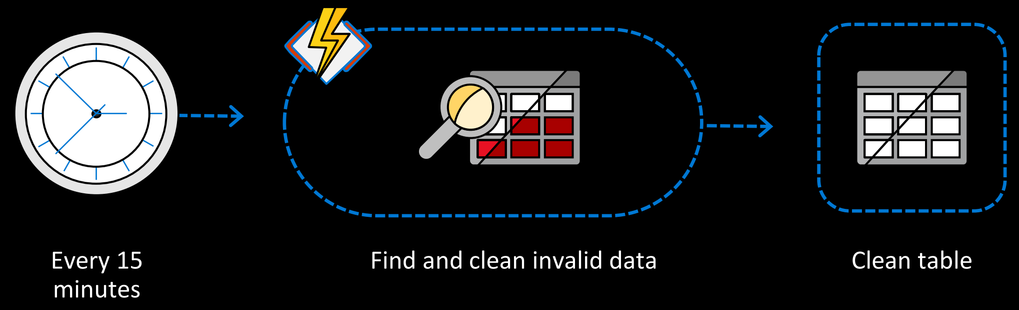
Azure Functions Common application scenarios

Any scenario that requires a response to an event

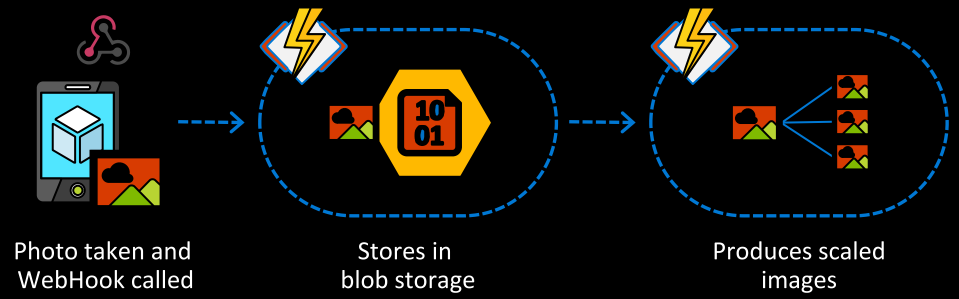
Real-time stream processing


Millions of devices feed into Stream Analytics
Transform to structured data
SQL
Store data in SQL DB

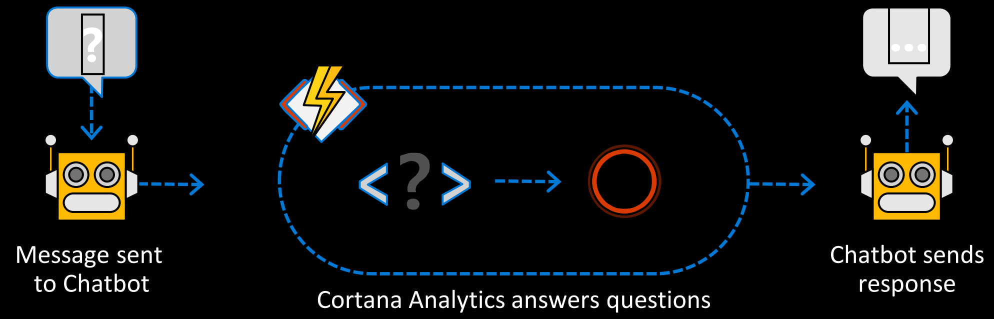
Timer-based processing



Backend (Mobile/IoT/Web)



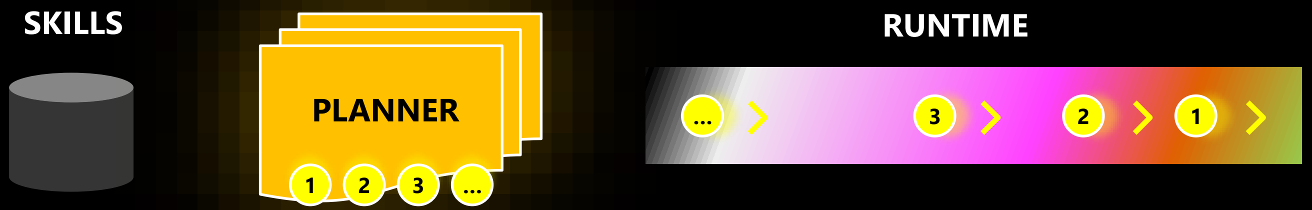
Real-time bot messaging



Semantic Kernel - LLM AI Dev Framework

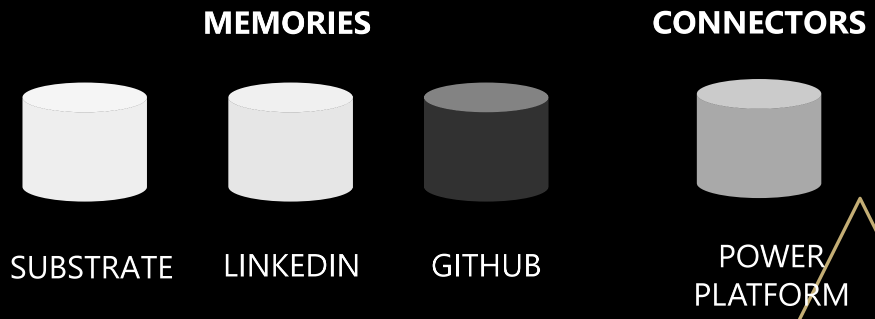
THE CORE

Everything you need to manage complex prompts, chains, long-running tasks, and **planning.**



THE BFD

All the longform **memories** readily available in the MS Graph. Plus, the 900 Power **connectors**.

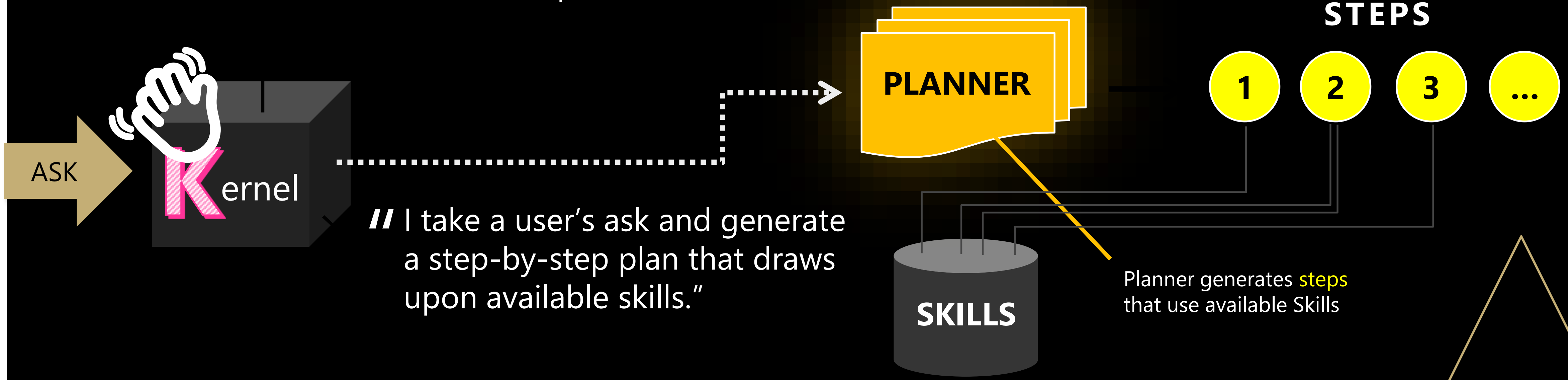


Meet the lightweight Kernel of Semantic Kernel

```
using Microsoft.SemanticKernel;

var myKernel = Kernel.Builder.Build();
```

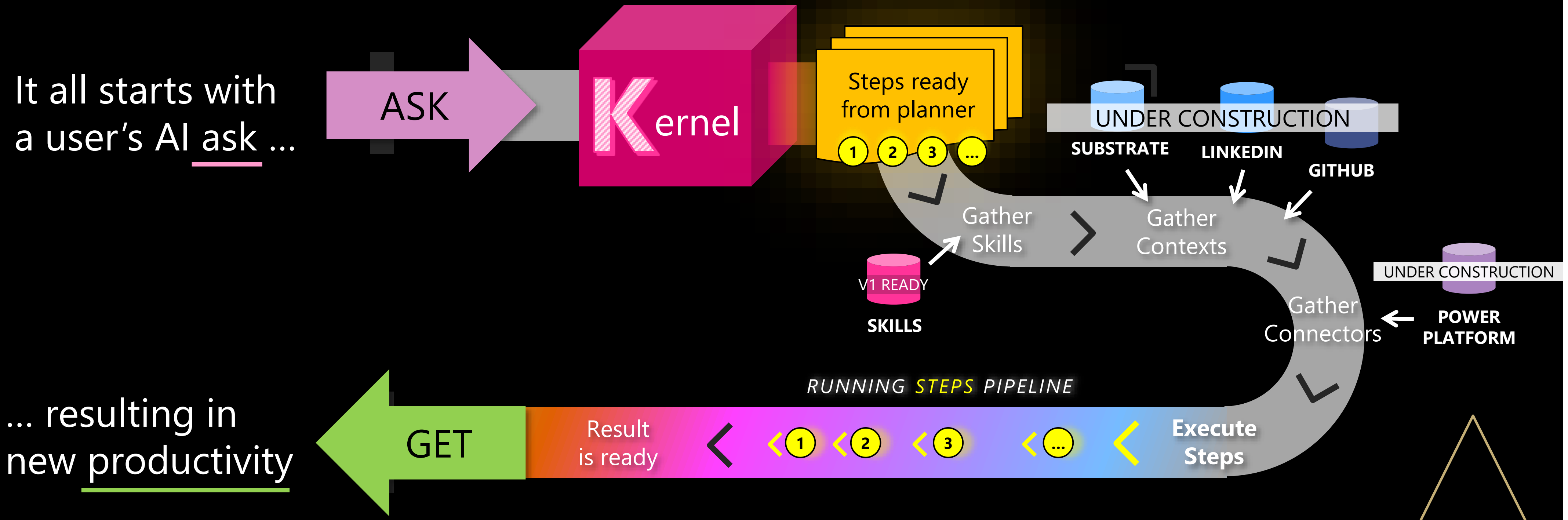
"I've been designed to reduce hallucinations, orchestrate complicated LLM AI prompts combined with native code, use multiple AI models, and … I have a special skill to **PLAN**."

"I take a user's ask and generate a step-by-step plan that draws upon available skills."

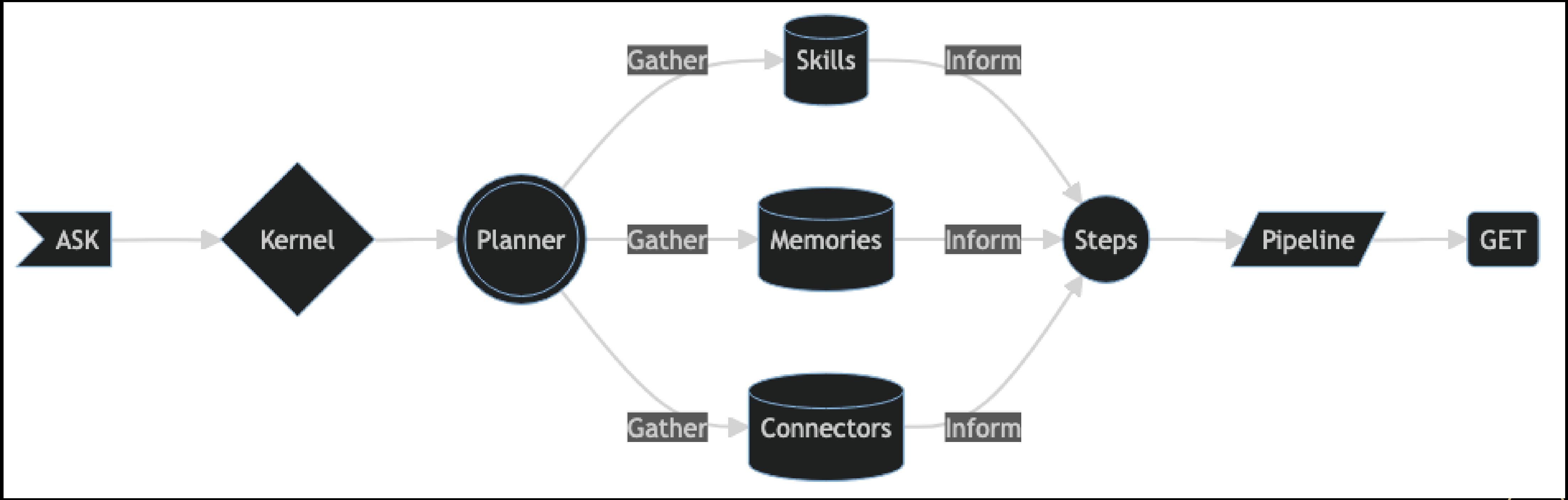


SK R1 Tour

Available NOW

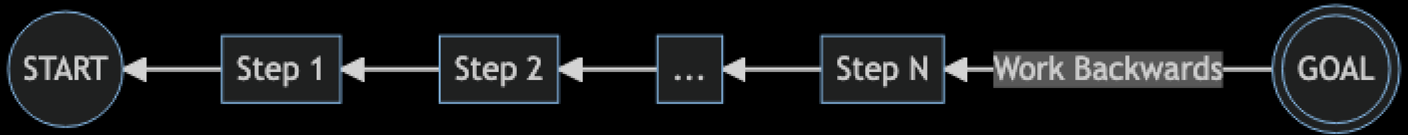


A more "goal-oriented" approach to problem solving



Planner

- The planner will operate within the skills it has available. In the event that a desired skill does not exist, the planner can suggest you create the skill. Or, depending upon the level of complexity the kernel can help you write the missing skill.

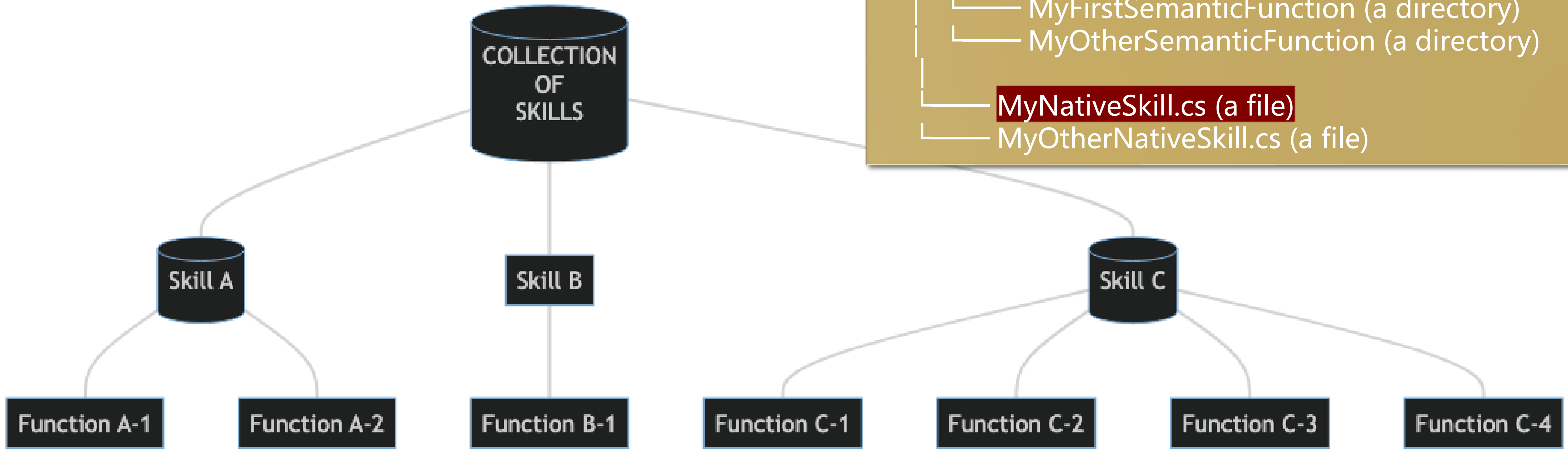


Skill

- an LLM AI prompt — also called a "semantic" function
- native computer code -- also called a "native" function

```
SkillName (directory name)
|
└──── Function1Name (directory name)    skprompt.txt
|                                        config.json
└──── Function2Name (directory name)
```

```
MyAppSource
|
└────MySkillsDirectory
     |
     └──── MySemanticSkill (a directory)
     |   |
     |   └──── MyFirstSemanticFunction (a directory)
     |   └──── MyOtherSemanticFunction (a directory)
     |
     └──── MyNativeSkill.cs (a file)
     └──── MyOtherNativeSkill.cs (a file)
```



Memory

Conventional key-value pairs: Just like you would set an environment variable in your shell, the same can be done when using SK. The lookup is "conventional" because it's a one-to-one match between a key and your query.

Conventional local-storage: When you save information to a file, it can be retrieved with its filename. When you have a lot of information to store in a key-value pair, you're best off keeping it on disk.

Semantic memory search: You can also represent text information as a long vector of numbers, known as "embeddings." This lets you execute a "semantic" search that compares meaning-to-meaning with your query.

Connector

MS Graph Connector Kit

- Add an event to your calendar
- Send an email for you
- Add a file to your OneDrive
- Create a share link to a file in your OneDrive
- Query your organization hierarchy
- Manage your MS To Do list

Out-of-box

- Issue a Bing search query
- Read OpenXML streams (e.g. Word docs)
- Use SQLite as a lightweight database

SK makes app developers' work lives easier

Fast Integration: SK is designed to be embedded in any kind of application, making it easy for developers to add LLM AI functionality to test inside their apps.

Power Prompting: Plain prompts that are fed as API calls can only get you so far. SK provides the abstractions and machinery to unlock your OpenAI or Azure OpenAI API key.

Novel-But-Familiar: For 100% determininism, native code is always available as a first-class partner on your prompt engineering quests. You get the best of both worlds.

How to use Semantic Kernel R1 in just 1 minute

1 MINUTE

Install the nuget package and go

```
#r "nuget: Microsoft.SemanticKernel, *-*"
```

HAVE MORE TIME?

Go deeper with the GitHub repo aka.ms/skrepo

MORE MINUTES

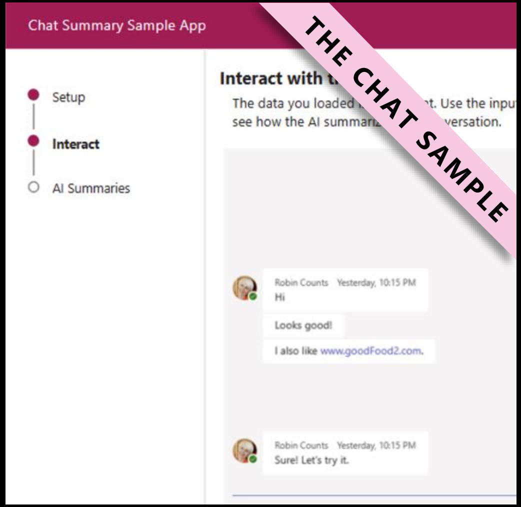
Learn more about its history [aka.ms/sk](aka.ms/sk)

Microsoft **Semantic Kernel**

The Easy Way To Add LLM AI To Your App

The growing set of samples are designed for most use cases.

Easily add chat-themed AI skills into your app.


Chat Summary Sample App

THE CHAT SAMPLE

- Setup
- **Interact**
- AI Summaries

Interact with t
The data you loaded t. Use the inpu
see how the AI summar versation.

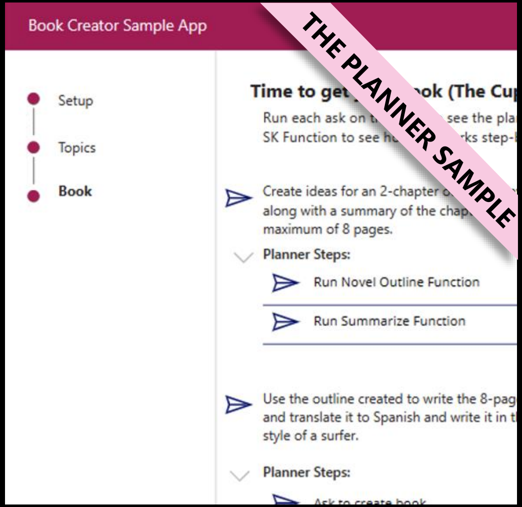
Robin Counts Yesterday, 10:15 PM
Hi

Looks good!

I also like www.goodFood2.com.

Robin Counts Yesterday, 10:15 PM
Sure! Let's try it.

Chain prompts serially or hierarchically via Planner.


Book Creator Sample App

THE PLANNER SAMPLE

- Setup
- Topics
- **Book**

Time to get ok (The Cu
Run each ask on t see the pla
SK Function to see h rks step-

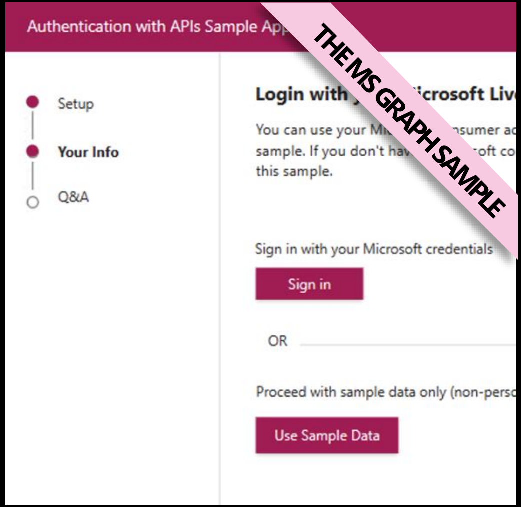
Create ideas for an 2-chapter
along with a summary of the chap
maximum of 8 pages.

Planner Steps:
- Run Novel Outline Function
- Run Summarize Function

Use the outline created to write the 8-pag
and translate it to Spanish and write it in t
style of a surfer.

Planner Steps:

Query MS Graph information with your favorite prompts.


Authentication with APIs Sample Ap

THE MS GRAPH SAMPLE

- Setup
- **Your Info**
- Q&A

Login with icrosoft Liv
You can use your M nsumer ac
sample. If you don't hav oft co
this sample.

Sign in with your Microsoft credentials

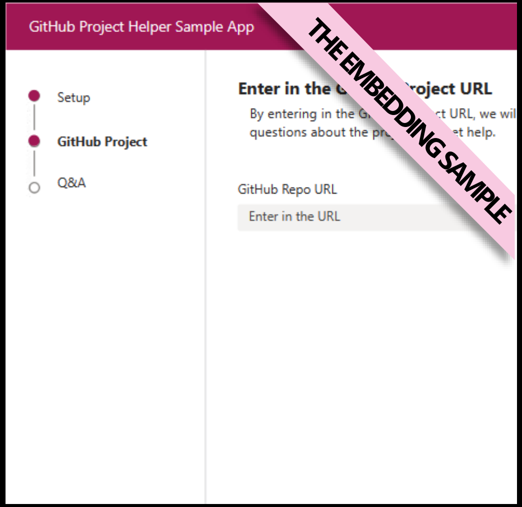
Sign in

OR

Proceed with sample data only (non-perso

Use Sample Data

Chat with your repo of choice using embeddings.


GitHub Project Helper Sample App

THE EMBEDDING SAMPLE

- Setup
- **GitHub Project**
- Q&A

Enter in the roject URL
By entering in the G ct URL, we wil
questions about the pr et help.

GitHub Repo URL

Enter in the URL

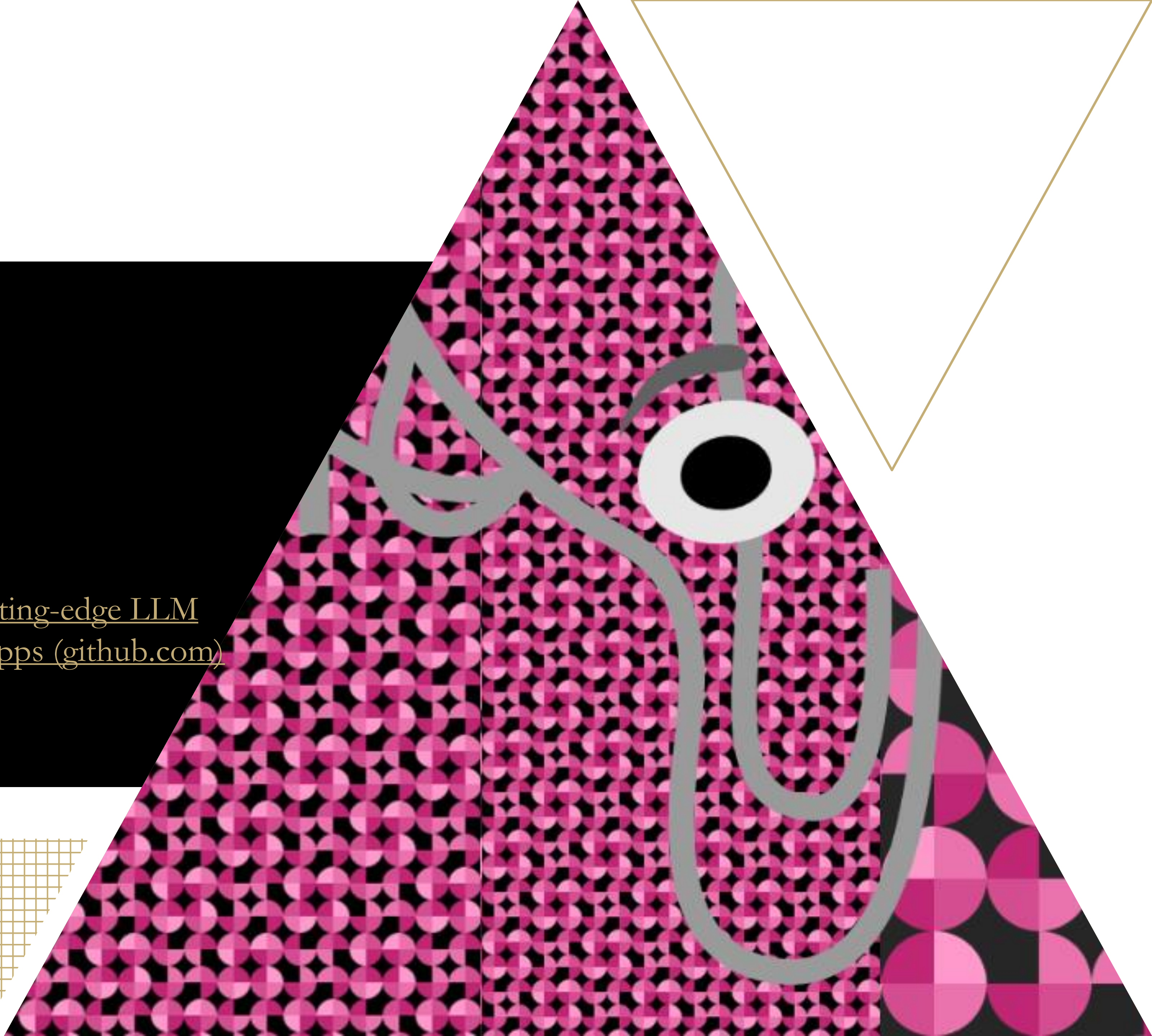
FOR 1P DEV + PRODUCT LEADERS

aka.ms/SK

Demo

SK Demos (sharepoint.com)

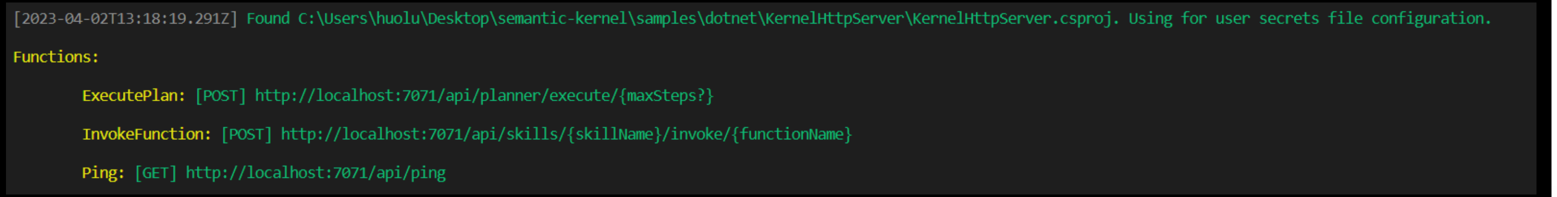
microsoft/semantic-kernel: Integrate cutting-edge LLM technology quickly and easily into your apps (github.com)



Semantic Kernel Service API

semantic-kernel/samples/dotnet/KernelHttpServer

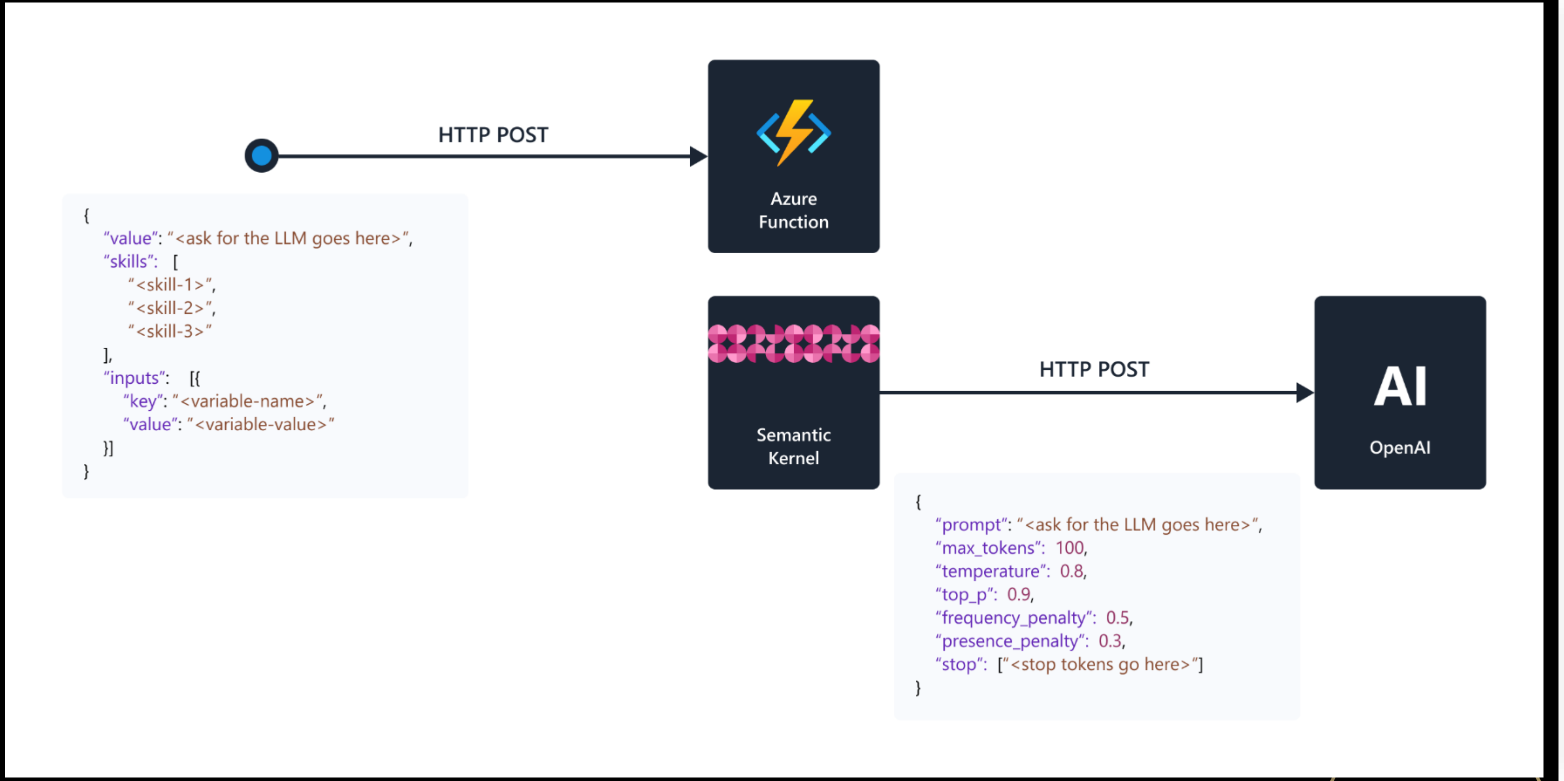
>>func start –csharp



```
[2023-04-02T13:18:19.291Z] Found C:\Users\huolu\Desktop\semantic-kernel\samples\dotnet\KernelHttpServer\KernelHttpServer.csproj. Using for user secrets file configuration.

Functions:

        ExecutePlan: [POST] http://localhost:7071/api/planner/execute/{maxSteps?}

        InvokeFunction: [POST] http://localhost:7071/api/skills/{skillName}/invoke/{functionName}

        Ping: [GET] http://localhost:7071/api/ping
```

- Written in C# against Azure Function Runtime v4
- Expose some Semantic Kernel APIs that you can call via HTTP POST requests

Azure Functions Core Tools

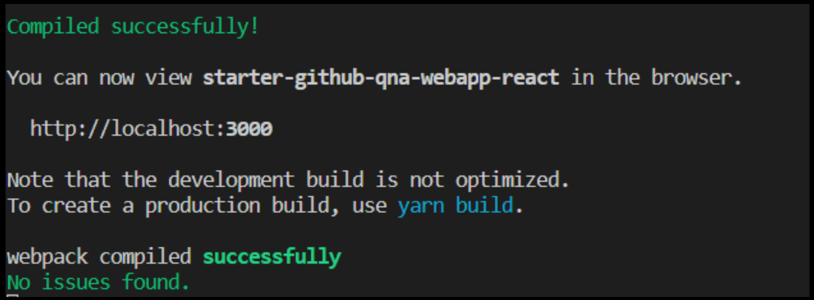


```
{
  "value": "<ask for the LLM goes here>",
  "skills":  [
      "<skill-1>",
      "<skill-2>",
      "<skill-3>"
  ],
  "inputs":   [{
     "key": "<variable-name>",
     "value": "<variable-value>"
  }]
}
```

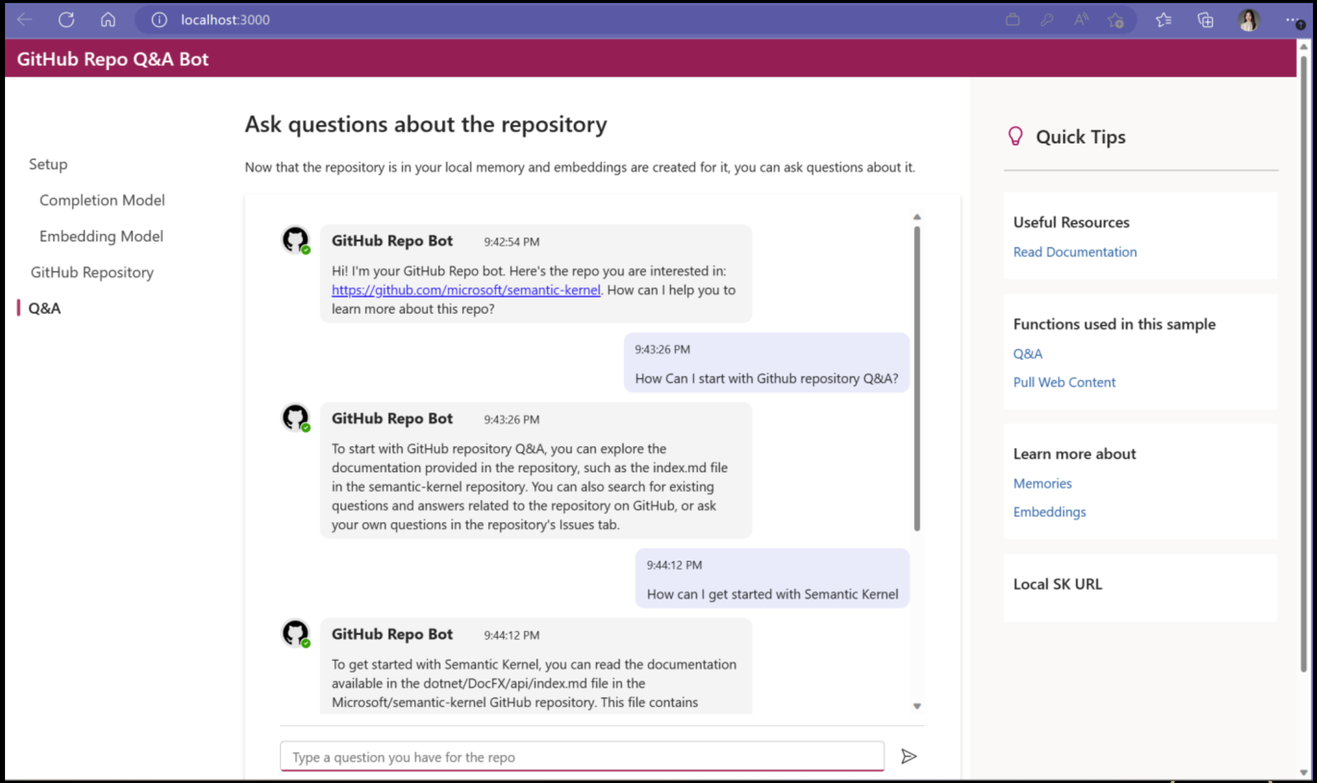
```
{
  "prompt": "<ask for the LLM goes here>",
  "max_tokens":  100,
  "temperature":  0.8,
  "top_p":  0.9,
  "frequency_penalty":  0.5,
  "presence_penalty":  0.3,
  "stop":  ["<stop tokens go here>"]
}
```

GitHub Repo Q&A Bot

```
semantic-kernel/samples/apps/github-qna-webapp-react
>>yarn install
>>yarn start
```



```
Compiled successfully!

You can now view starter-github-qna-webapp-react in the browser.

  http://localhost:3000

Note that the development build is not optimized.
To create a production build, use yarn build.

webpack compiled successfully
No issues found.
```



localhost:3000

GitHub Repo Q&A Bot

Setup
Completion Model
Embedding Model
GitHub Repository
Q&A

Ask questions about the repository

Now that the repository is in your local memory and embeddings are created for it, you can ask questions about it.

GitHub Repo Bot 9:42:54 PM
Hi! I'm your GitHub Repo bot. Here's the repo you are interested in: https://github.com/microsoft/semantic-kernel. How can I help you to learn more about this repo?

9:43:26 PM
How Can I start with Github repository Q&A?

GitHub Repo Bot 9:43:26 PM
To start with GitHub repository Q&A, you can explore the documentation provided in the repository, such as the index.md file in the semantic-kernel repository. You can also search for existing questions and answers related to the repository on GitHub, or ask your own questions in the repository's Issues tab.

9:44:12 PM
How can I get started with Semantic Kernel

GitHub Repo Bot 9:44:12 PM
To get started with Semantic Kernel, you can read the documentation available in the dotnet/DocFX/api/index.md file in the Microsoft/semantic-kernel GitHub repository. This file contains

Type a question you have for the repo

Quick Tips

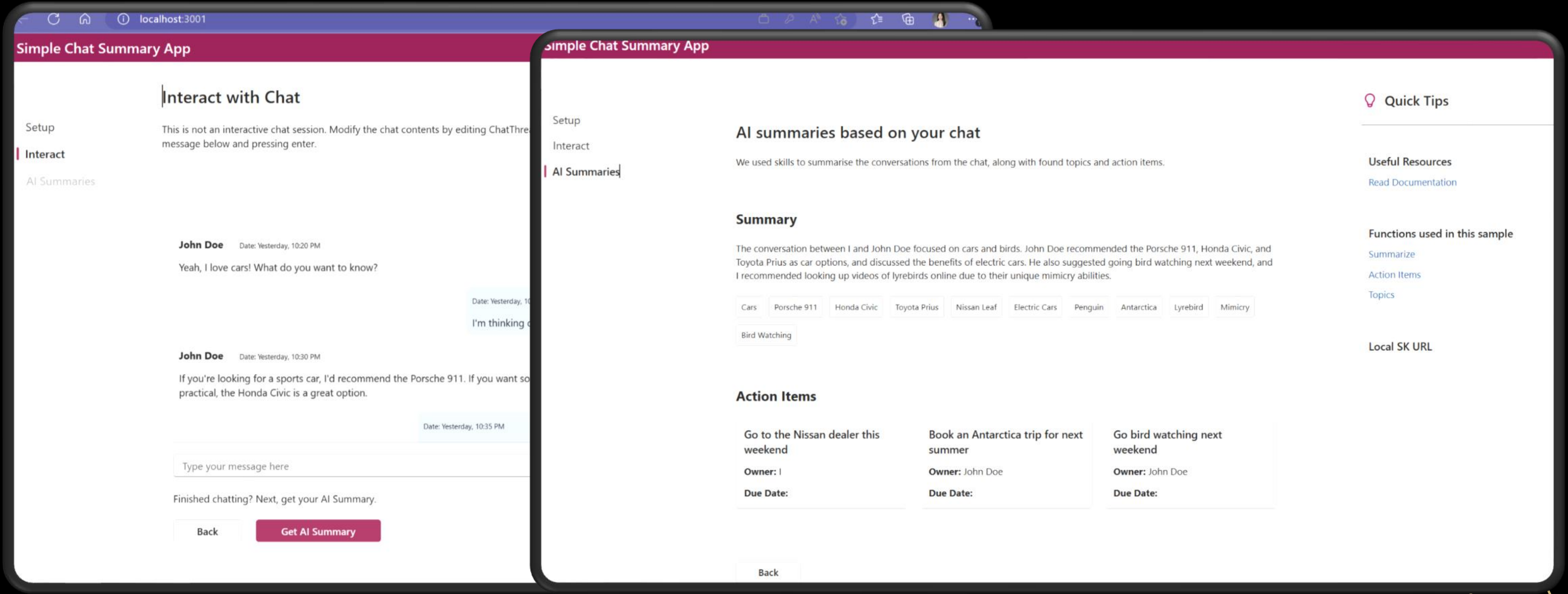
Useful Resources
Read Documentation

Functions used in this sample
Q&A
Pull Web Content

Learn more about
Memories
Embeddings

Local SK URL

Simple Chat Summary App



localhost:3001

Simple Chat Summary App

Setup

Interact

AI Summaries

Interact with Chat

This is not an interactive chat session. Modify the chat contents by editing ChatThre
message below and pressing enter.

John Doe Date: Yesterday, 10:20 PM

Yeah, I love cars! What do you want to know?

Date: Yesterday, 1

I'm thinking

John Doe Date: Yesterday, 10:30 PM

If you're looking for a sports car, I'd recommend the Porsche 911. If you want so
practical, the Honda Civic is a great option.

Date: Yesterday, 10:35 PM

Type your message here

Finished chatting? Next, get your AI Summary.

Back | Get AI Summary

Simple Chat Summary App

Setup

Interact

AI Summaries

AI summaries based on your chat

We used skills to summarise the conversations from the chat, along with found topics and action items.

Summary

The conversation between I and John Doe focused on cars and birds. John Doe recommended the Porsche 911, Honda Civic, and Toyota Prius as car options, and discussed the benefits of electric cars. He also suggested going bird watching next weekend, and I recommended looking up videos of lyrebirds online due to their unique mimicry abilities.

Cars | Porsche 911 | Honda Civic | Toyota Prius | Nissan Leaf | Electric Cars | Penguin | Antarctica | Lyrebird | Mimicry | Bird Watching

Action Items

Go to the Nissan dealer this weekend	Book an Antarctica trip for next summer	Go bird watching next weekend
Owner: I	**Owner:** John Doe	**Owner:** John Doe
Due Date:	**Due Date:**	**Due Date:**

Back

Quick Tips

Useful Resources

Read Documentation

Functions used in this sample

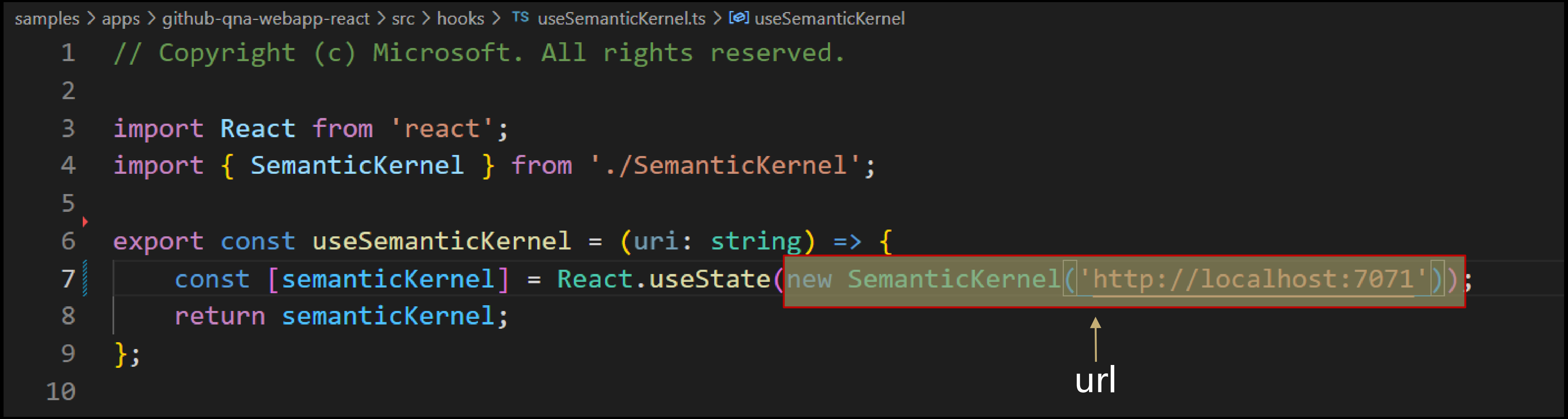
Summarize

Action Items

Topics

Local SK URL

Workaround(Optional)



samples > apps > github-qna-webapp-react > src > hooks > TS useSemanticKernel.ts > useSemanticKernel

```
// Copyright (c) Microsoft. All rights reserved.

import React from 'react';
import { SemanticKernel } from './SemanticKernel';

export const useSemanticKernel = (uri: string) => {
    const [semanticKernel] = React.useState(new SemanticKernel('http://localhost:7071'));
    return semanticKernel;
};

```

Q & A

- Semantic Kernel - Home (sharepoint.com)
- Azure-OpenAI-App-Innovation-Workshop/Workshop Content EN at main · xuhaoruins/Azure-OpenAI-App-Innovation-Workshop (github.com)