AI-Infused RAG applications with Advanced Artificial Intelligence

April 16/17th, 2025

Aleksandra Butneva
Elisa Piccin
Niccolò Benetti

Cloud Solution Architects
Data & AI Microsoft

Agenda (1)

9.30-17.30 - 16 Aprile 2025

- Introduzione dei partecipanti e controllo degli ambienti
- AI Foundry features & capabilities per lo sviluppo delle applicazioni GenAI-infused con AI Foundry SDK
  - **Inizia a sviluppare la tua app con AI Foundry Project / Hub / AI Services**
    - Gestione di Progetti AI con AI Foundry Management Center (Utenti, Connessioni, Infra) e la sua integrazione con Azure
    - Azure Governance
      - Azure Landing Zone intro/recap - Design principles for Azure applications - Azure Architecture Center | Microsoft LearnAzure Well-Architected Framework - Microsoft Azure Well-Architected Framework | Microsoft Learn
      - Azure Developer CLI intro/recap - Azure Developer CLI commands overview | Microsoft Learn
      - Framework di orchestrazione LLM intro/recap (LangChain, Semantic Kernel, …)
    - Esercizio UI-first/Demo: creazione di risorse AI Foundry
  - **Scopri il giusto modello LLM/SLM: Model catalog di AI Foundry**
    - Model comparison & benchmarking
    - Esercizio UI-first: creazione di deployment dei modelli LLM/SLM
  - **Pausa**
  - **Predisposizione di conda environment su Azure Machine Learning** per proseguire con gli esercizi code-first
  - **Prova il modello della tua scelta senza grounding: Chat Playground di AI Foundry**
    - Esercizio code-first: AI Foundry SDK Basic Operations
  - **Estendi il modello della tua scelta con grounding: Agent Playground di AI Foundry**
    - Esercizio code-first: Agent Service Basics: LLM grounding con Bing Search
    - Esercizio code-first: RAG OOB con Agent Service & grounding con File Search

Agenda (2)

- **Pranzo**
  - **Prepara la KB del RAG chatbot con smart document processing: Content Understanding Playground / Document Intelligence Studio**
    - Introduzione al servizio: Slide/ Esercizi UI-first
    - Esercizi code-first:
      - Modalità di utilizzo dei modelli LLM multimodali
      - Modalità di utilizzo di DocIntelligence Client
      - Modalità di utilizzo di Content Understanding Client
- **Pausa**
  - **Valuta i modelli con LLM Evaluation SDK di AI Foundry**
    - Introduzione su RAI (Responsible AI approach di Microsoft)
    - Esercizi code-first: Evaluator library/evaluation SDK di AI Foundry

9.00(15)-13.00 - 17 Aprile 2025

- Focus: Scenario RAG
  - UniBologna - presentazione dei dati previsti per il PoC (data perimeter)
  - RAG basics: MustKnow del pattern
  - Esercizio UI-first: creazione dell'indice AI Search sul portale Azure
  - Esercizio code-first: RAG con Agent Service & grounding con l'AI Search predisposto dal portale ("Chat with your data")
- **Pausa**
- Demo RAG end-to-end code-first con AI Foundry SDK con focus sulle pratiche di logging, groundedness delle risposte e osservabilità

Modern applications aim to maximize the value of enterprise data with AI

A unified platform to explore, build, test, deploy, and monitor generative AI applications

A single platform for your data gravity across your organization to ground your AI on your data

AI Powered

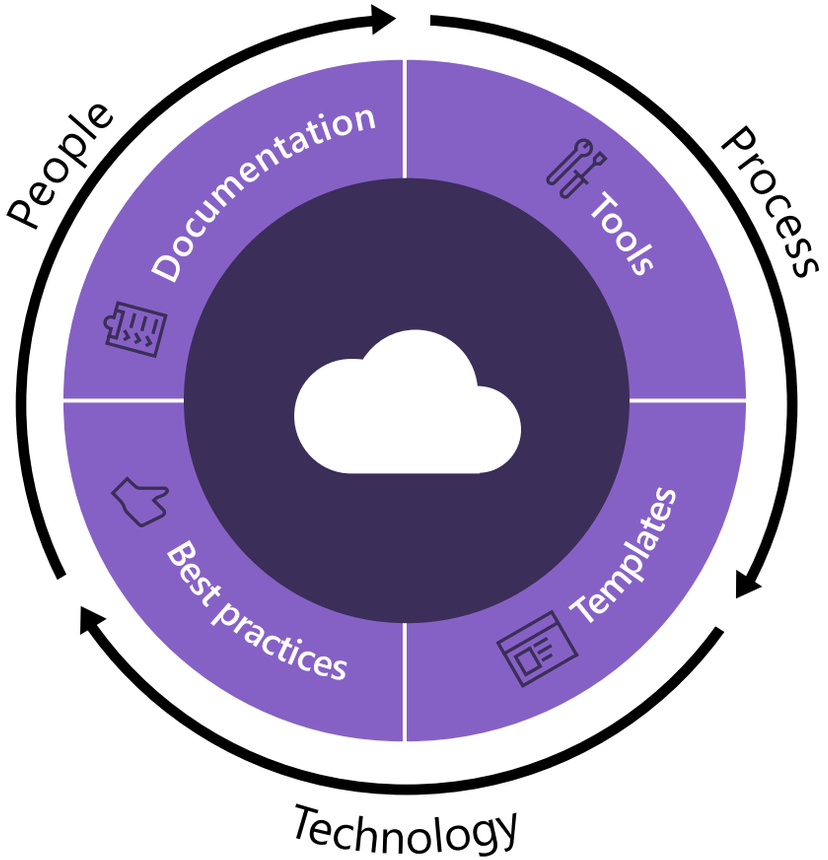
(Gen) AI accelerates your data journey in Fabric

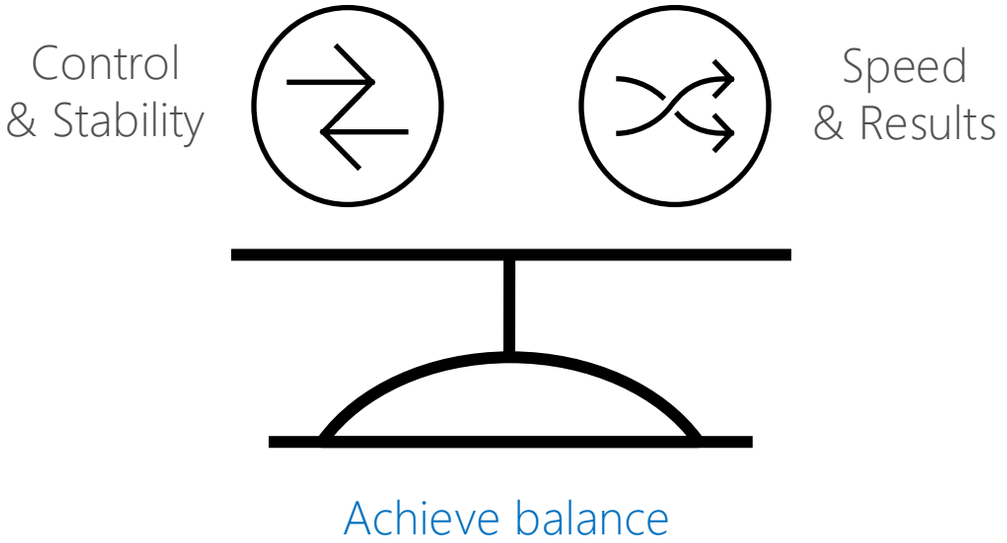
Copilot accelerated experiences

AI-driven insights

Custom generative AI for your data

Microsoft Cloud Adoption Framework for Azure





Align business, people and technology strategy.
Achieve business goals with actionable, efficient, and comprehensive guidance.
Deliver fast results with control and stability.

What is greenfield development?

Developing a new environment without dependencies

- Beginning of your cloud adoption journey
- Your new cloud environment
- No production workloads are deployed
- Environment can be overwritten, and workloads moved without business impact
- No business processes are impacted by changes to operations, controls, or access

> **We have not deployed anything in Azure and want a start-to-finish plan based on best practices.**

What is “brownfield development?”

Developing and deploying code in existing environments

- Workloads are already in use in the cloud
- Migration events have already executed
- Users are accessing services within the environment
- Production workloads are running

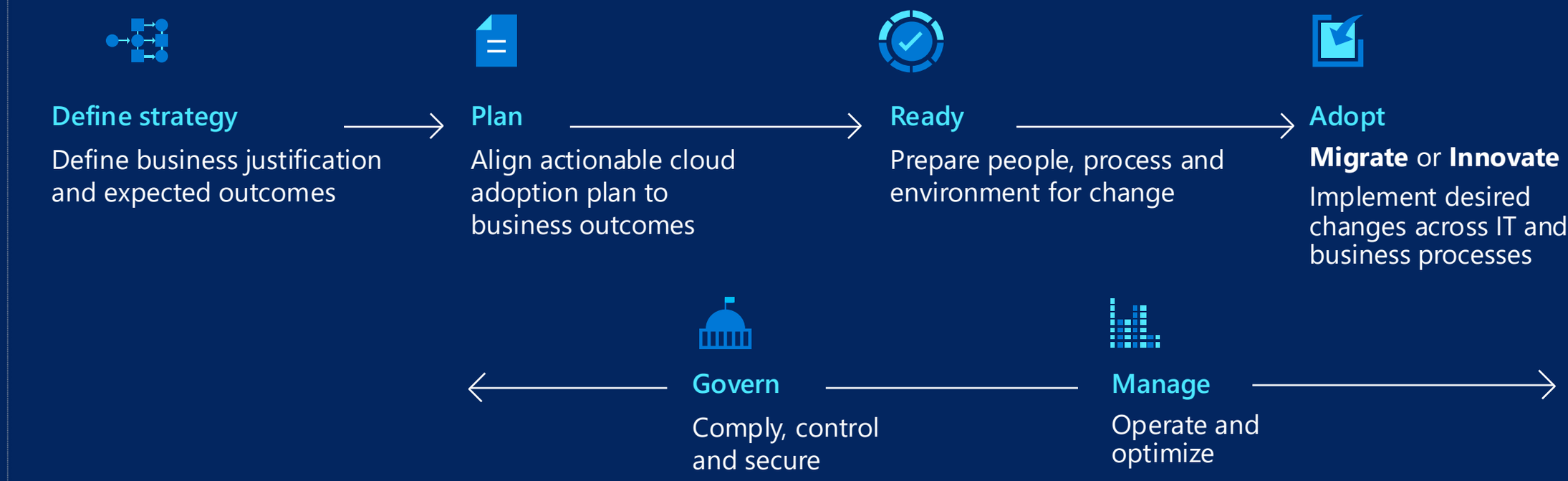
> “We have environments and landing zones already, but our requirements have changed, and we need to move towards a different type of architecture.”

Azure basics

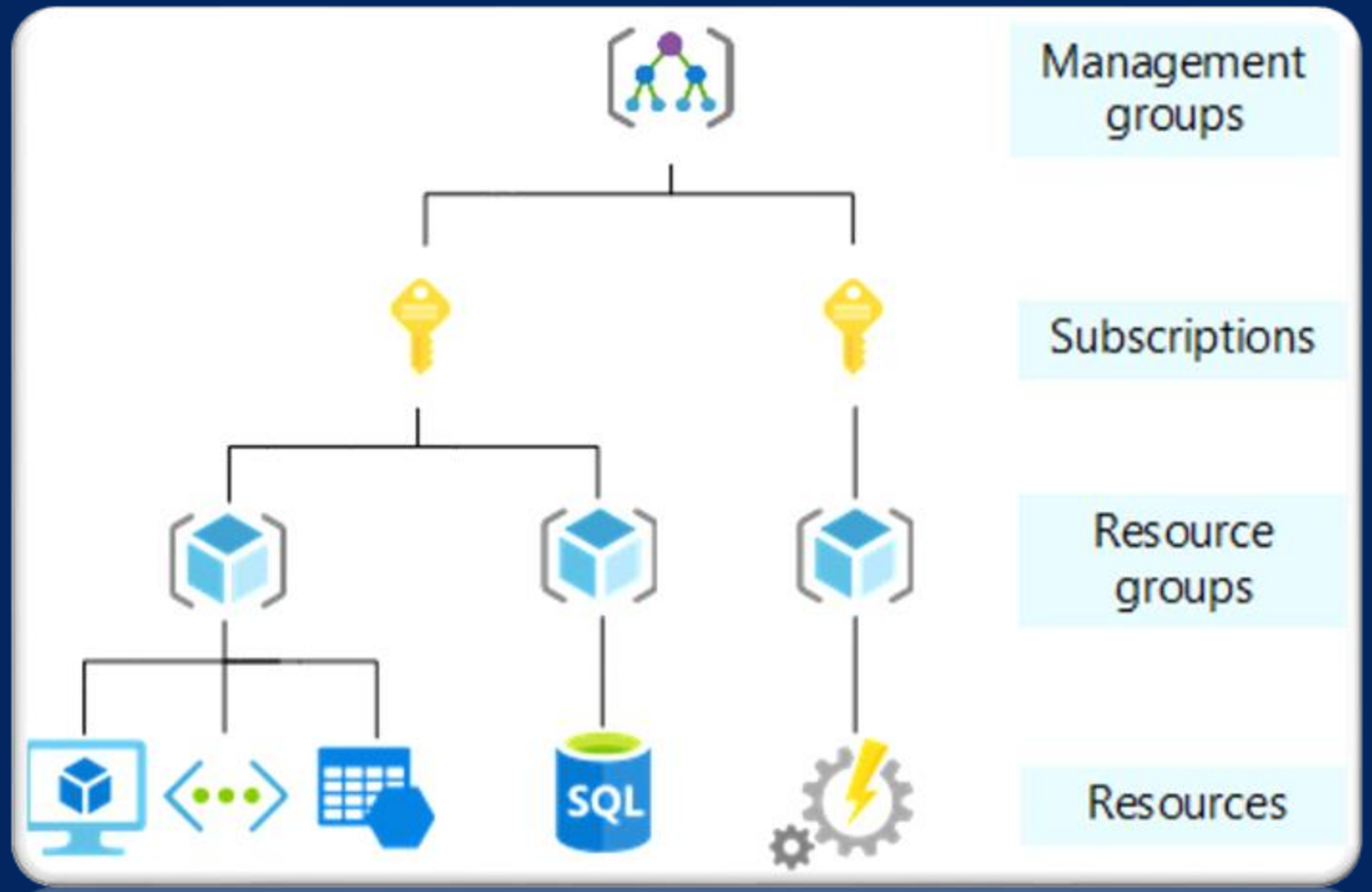
How are Azure AI Services organized

Implement business, people & technology strategies to adopt the cloud with confidence

Microsoft Cloud Adoption Framework for Azure



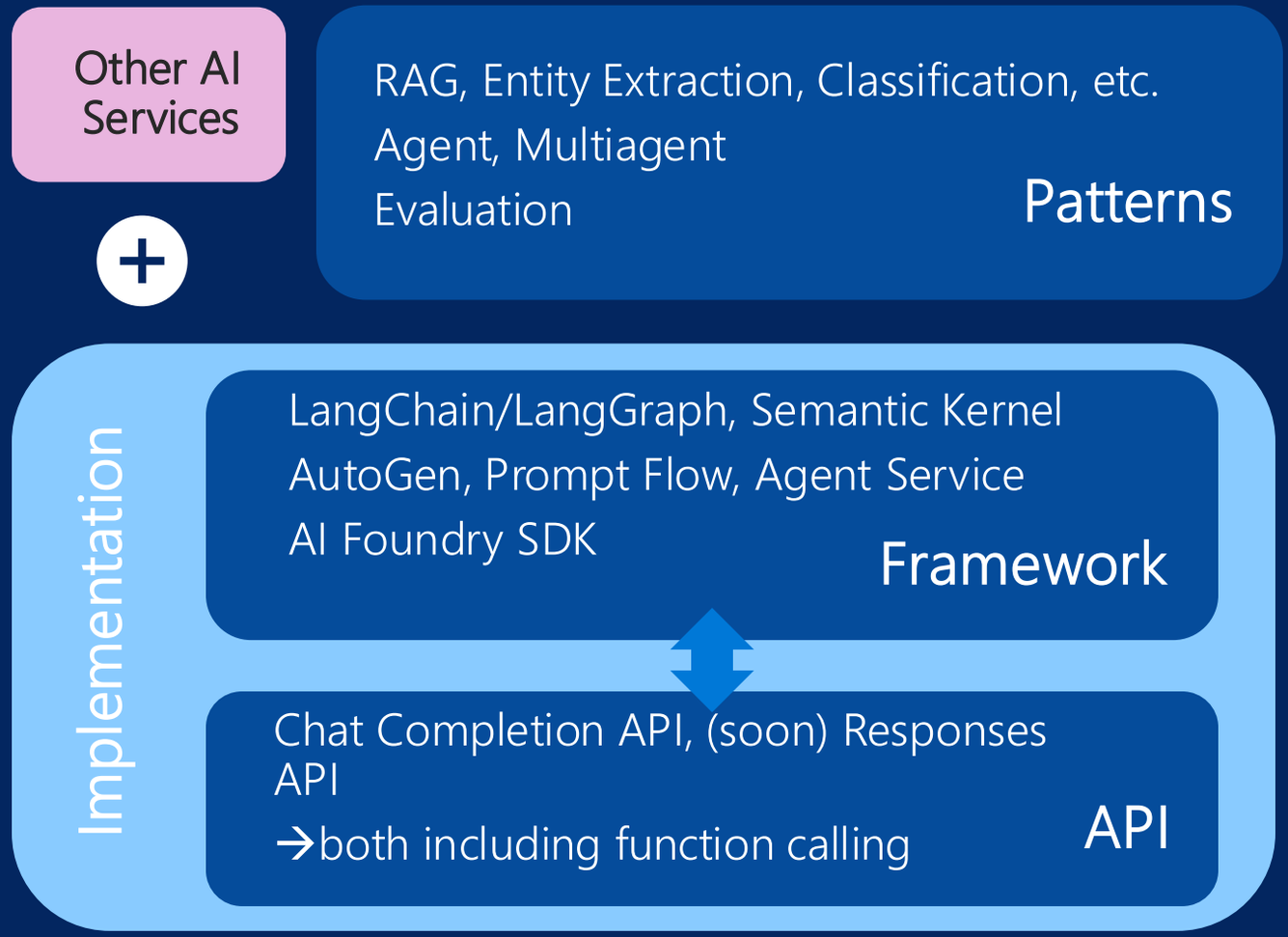
Scope levels for organizing Azure resources



Organization and governance recommendations

- Treat subscriptions as a unit of management aligned with your business needs and priorities.
- Make subscription owners aware of their roles and responsibilities.
- Do a quarterly or yearly access review for Azure AD Privileged Identity Management to ensure that privileges don't proliferate as users move within your organization.
- Take full ownership of budget spending and resources.
- Ensure policy compliance and remediate when necessary.
- Reference the following principles as you identify requirements for new subscriptions:
- Scale limits: Subscriptions serve as a scale unit for component workloads to scale within platform subscription limits. Large specialized workloads like high-performance computing, IoT, and SAP should use separate subscriptions to avoid running up against these limits.
- Management boundary: Subscriptions provide a management boundary for governance and isolation, allowing a clear separation of concerns. Different environments, such as development, test, and production, are often removed from a management perspective.
- Policy boundary: Subscriptions serve as a boundary for the Azure Policy assignments. For example, secure workloads like PCI typically require other policies in order to achieve compliance. The other overhead doesn't get considered if you use a separate subscription. Development environments have more relaxed policy requirements than production environments.
- Target network topology: You can't share virtual networks across subscriptions, but you can connect them with different technologies like virtual network peering or Azure ExpressRoute. When deciding if you need a new subscription, consider which workloads need to communicate with each other.

Do we just need a model?



Some useful considerations

- Project requirements (time and effort)
- Programming language preference (Python, C# etc.)
- Open-source vs proprietary framework
- Application scalability
- Target user experience
- Continuous Development and Integration logics

Azure AI

Best-in-class AI foundation models

Azure Traditional and GenAI first AI Services

Pre-trained, turnkey solutions for intelligent applications

Azure Machine Learning

Full-lifecycle tools for designing and managing AI models

Responsible AI Tooling

Build and manage apps that are trustworthy by design

Azure AI Foundry

A comprehensive platform to develop and deploy custom copilots

Azure OpenAI Service

Large, pretrained AI models from OpenAI to unlock new scenarios

Custom AI models fine-tuned with your data and hyperparameters

Built-in responsible AI to detect and mitigate harmful use

Enterprise-grade security with role-based access control (RBAC) and private networks

Azure AI Search

Secure, scalable solution that revolutionizes information retrieval of user-owned content.

Comprehensive search engine that supports vector, full-text, and hybrid searches across a richly indexed database.

Works closely with other Azure services, allowing for automated data ingestion from Azure data sources and incorporation of consumable AI for advanced processing tasks

Advanced features like semantic ranking, relevance tuning, and comprehensive query syntax support.

Azure AI Content Safety

Azure AI Content Safety uses AI to help you create safer online spaces.

- With cutting edge AI models, it can detect hateful, violent, sexual, and self-harm content and assign it a **severity score,** allowing businesses to prioritize what content moderators review.
- **Azure AI Content Safety can handle nuance and context**, which can assist human content moderator teams.
- Azure AI Content Safety isn't one-size-fits-all—**it can be customized to help businesses implement their policies.** Plus, its multi-lingual models enable it to **understand many languages simultaneously.**

1

Azure AI Content Safety classifies harmful content into four categories:

Hate Sexual Self-harm Violence

2

Next, it returns a four or eight severity level for each category:

Hate: 0 – 2 – 4 – 6 or 0-1-2-3-4-5-6-7
Sexual: 0 – 2 – 4 – 6 or 0-1-2-3-4-5-6-7
Self-harm: 0 – 2 – 4 – 6 or 0-1-2-3-4-5-6-7
Violence: 0 – 2 – 4 – 6 or 0-1-2-3-4-5-6-7

3

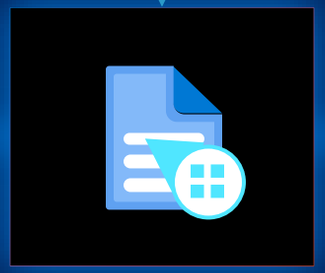
Then, users take actions based on the severity levels:

Auto allowed
Auto rejected
Send to human moderator

Traditional Azure AI Services for Business Process Automation

Computer / AI Vision

- Read API
- Optical character recognition
- Content tagging
- Image detection



AI Document Intelligence

- Layout / generic document understanding
- Generative Entity Extraction
- Key value pairing
- Custom Model training

AI Speech

- Speech to Text
- Speech translation & diagnostics
- Text to Speech

Azure ML Studio

An immersive experience for managing the end-to-end machine learning (and LLM Ops with AML Prompt Flow) lifecycle

Azure Machine Learning Workspace

Authoring

- Notebooks
- Automated ML
- Designer
- Prompt flow
- Tracing PREVIEW

Assets

- Data
- Jobs
- Components
- Pipelines
- Environments
- Models
- Endpoints

Manage

- Compute
- Monitoring
- Data Labeling
- Linked Services PREVIEW
- Connections PREVIEW

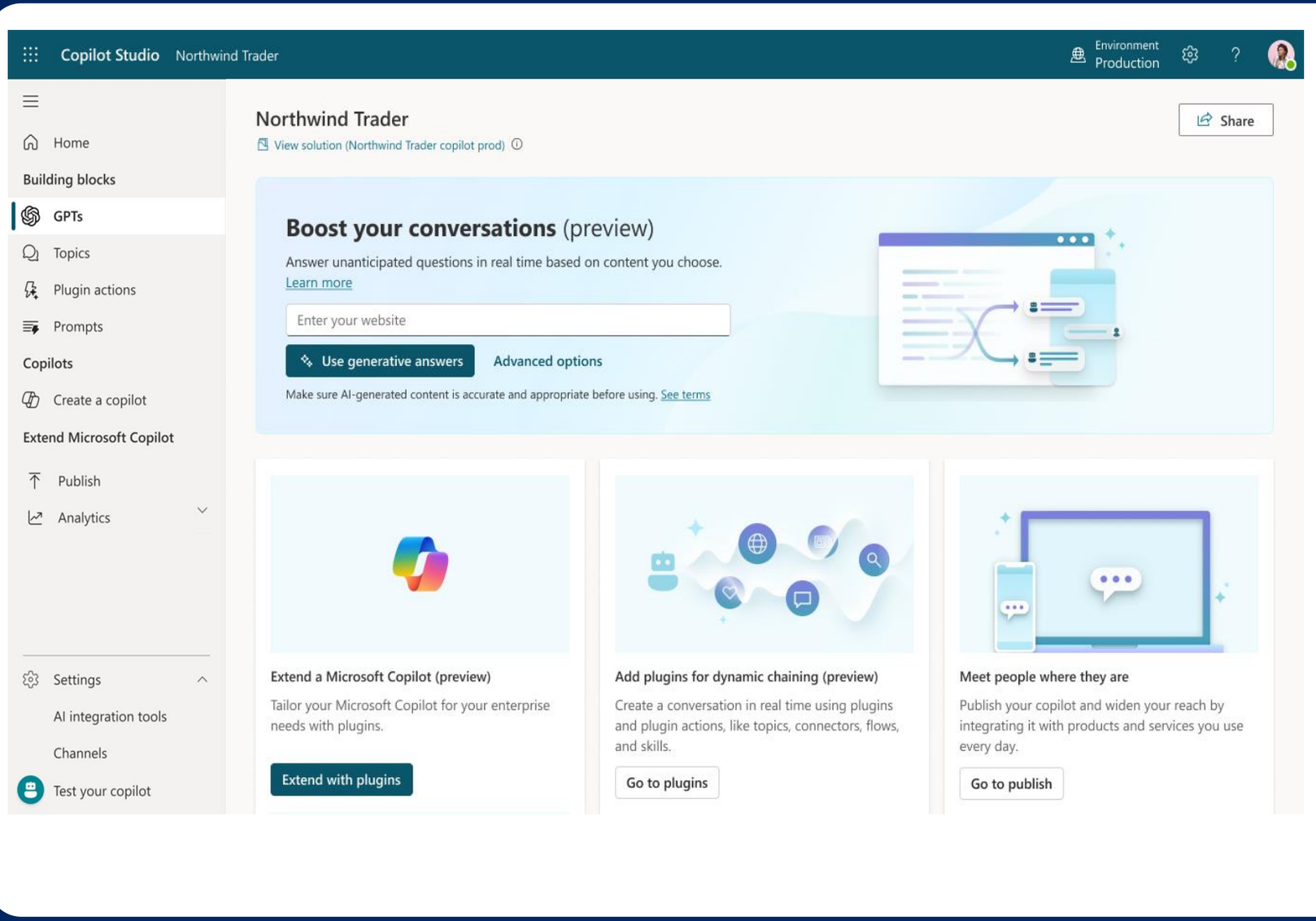
Microsoft Copilot Studio

Build your own copilot
Create and publish a custom copilot for your organization using the intuitive building experience enhanced with large language models and generative AI

Customize Microsoft Copilot
Extend and customize 1st party Microsoft Copilots with your own enterprise scenarios. Copilot Studio will be included with the Microsoft 365 Copilot SKU.

Connected platform
Integrates and exposes various Microsoft's conversational AI technology stacks - integrated with Azure AI Studio, Azure Cognitive Services, Azure Bot Framework, Power Platforms AI models and more

Manage copilot experiences
Governance and control features to monitor usage with full visibility of customizations, standalone copilots as well as who is building and customizing them.

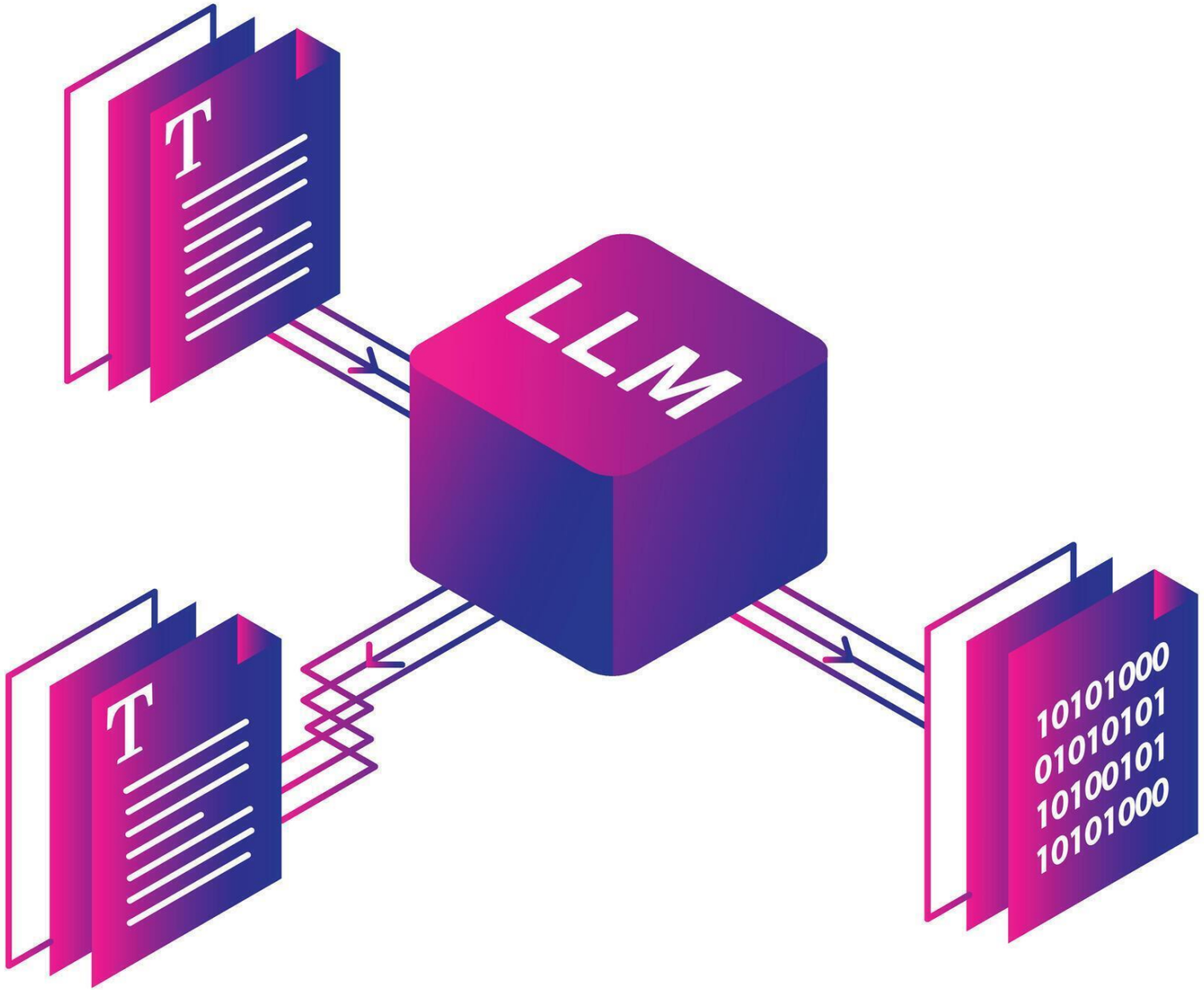


Summing up your Copilot options

I want a generative AI enabled chat application

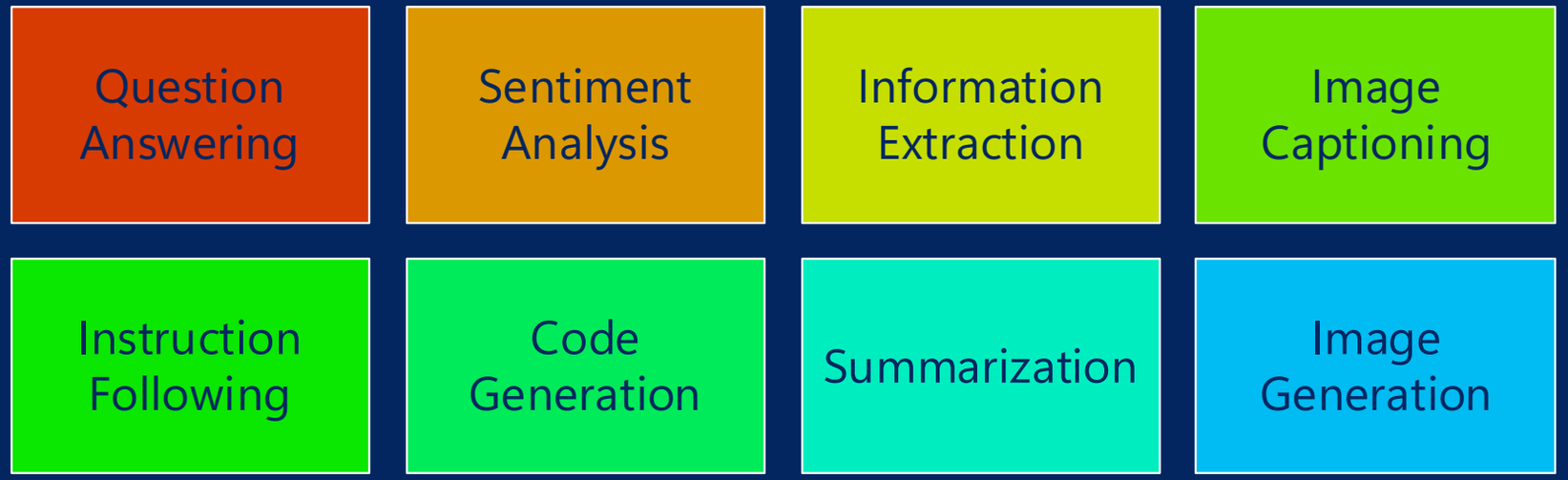
Generative AI

Making LLMs and SLMs work

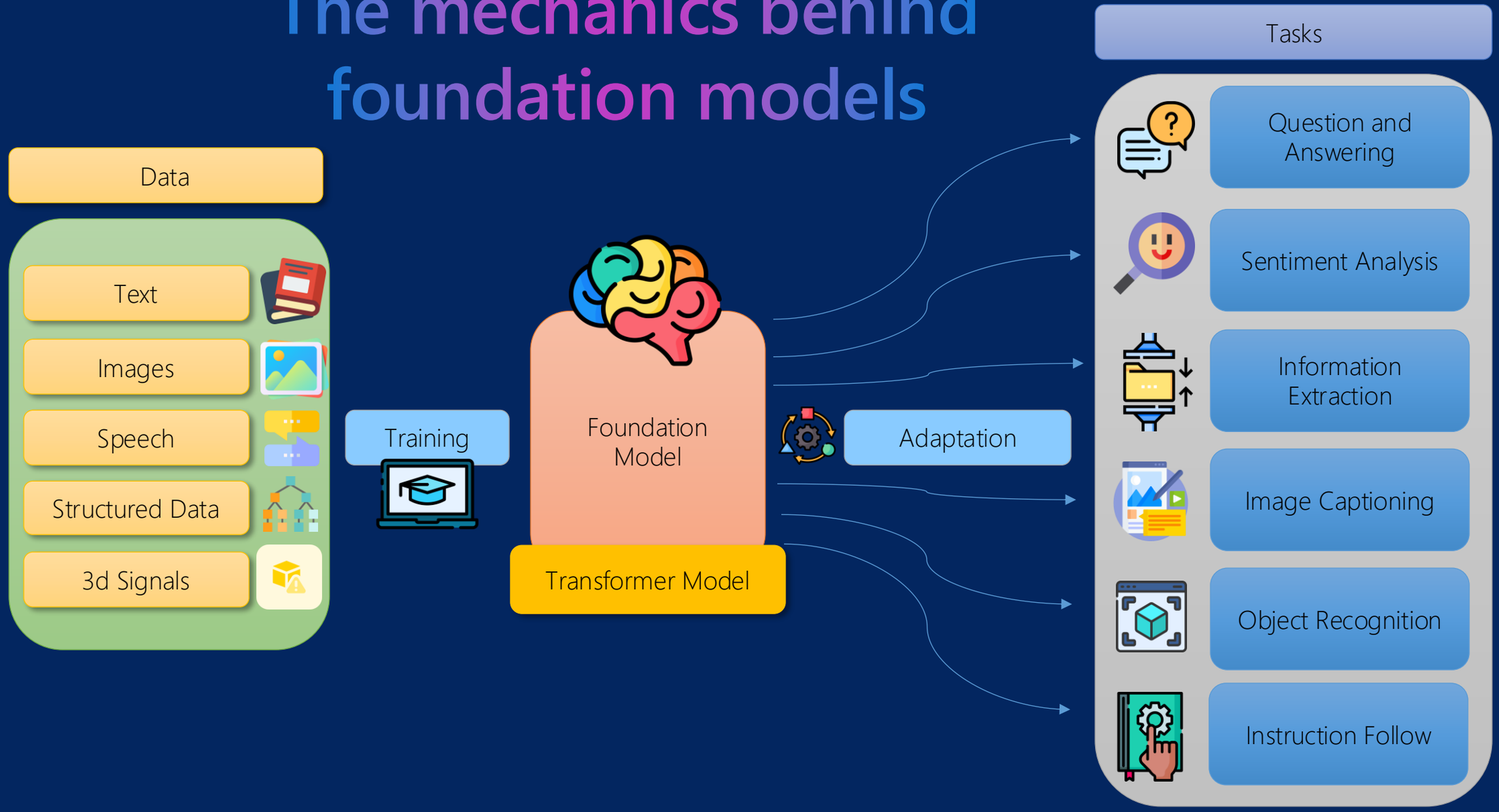


In the world of Generative AI, once there was … a foundation model

Train one model on a huge amount of data and adapt it to many applications

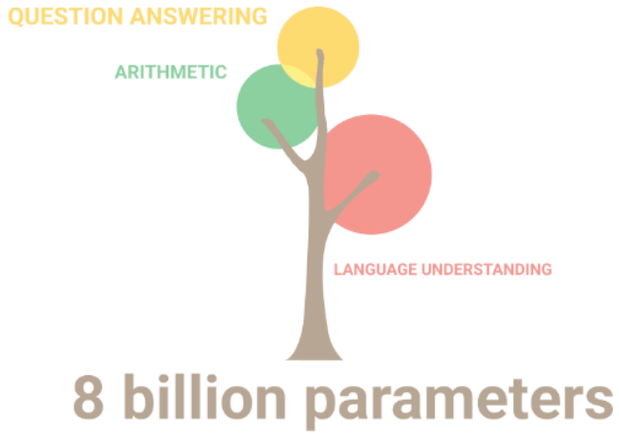


The mechanics behind foundation models



Foundation models exhibit emergent behavior

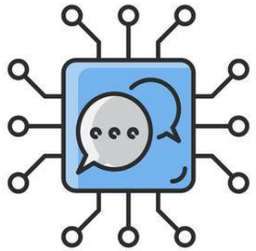
As the scale of the model increases, the performance improves across tasks while also **unlocking new capabilities that were not anticipated**. For example, a model trained on a large language dataset might learn to generate stories on its own, or to do arithmetic, without being explicitly programmed to do so.



Are foundational models always that big?

A small language model (SLM) is a lightweight generative AI model.

The label "small" in this context refers to the size of the model's neural network, the number of parameters the model uses to decide, and the volume of data the model is trained on.



SLMs require less computational power and memory than large language models (LLMs). This makes them suitable for on-premises and on-device deployments, one card on an A100, deploying to V100 and T4's, and capable of running on CPU and GPU scenarios on client machines with sufficient memory.

For more complex scenarios, we orchestrate models with frameworks

LangChain / LangGraph / LangSmith

What

An agentic, chain-based LLM framework with support for function calling

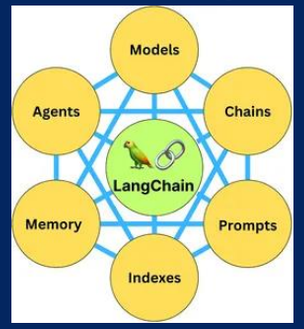
How

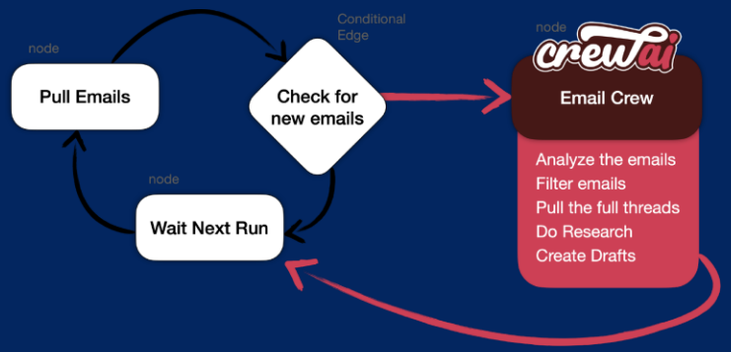
By defining a "state graph" for an agent implementation = LangGraph
By offering a UI (LangSmith) for monitoring the status of chain executions

Main concepts

Graph consists of 3 objects: Agent State, Edges, and Graph

Graph = *pipeline, workflow*
Agent State = *conversation history*
Edges = *connections between agents and nodes*





Semantic Kernel

What

An agentic, kernel-based LLM framework with support for function calling

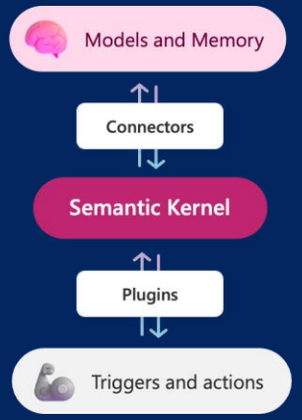
How

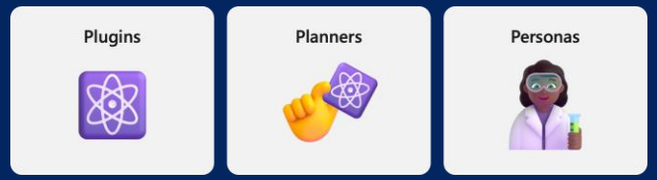
By populating a "kernel" for an agent implementation

Main concepts

The SK stack consists of 4 objects: Kernel, Plugins, Planners and Connectors

Kernel = *like an OS, responsible for managing resources that are necessary to run "code" in an AI application*
Plugins = *conversation history*
Planners = *a function that takes a user's ask and returns a plan on how to accomplish the request.*
Connectors = *connections to LLM models and memory*





Microsoft

Azure AI Foundry

Unified AI Development Platform

Powered by Azure AI Stack

Get to know Azure AI

Azure AI Foundry

One place for building and deploying AI solutions

- Model Catalog
- Complete AI Toolchain
- Responsible AI Practices
- Enterprise-Grade Production at Scale

Cutting-Edge Models

Access to the latest foundation and open-source models

- Model Catalog
- Models As a Service
- GPT Model Family
- Open-Source Models
- Small Language Models

Azure AI Services

Pre-trained, turnkey solutions for intelligent applications

- Azure OpenAI Service
- Azure AI Search
- Azure AI Speech
- Azure AI Vision
- Azure AI Content Safety
- Azure AI Document Intelligence
- Azure AI Language
- Azure AI Translator

Azure Machine Learning

Full-lifecycle tools for designing and managing responsible AI models

- Prompt Flow Orchestration
- Responsible Model Design
- Model Fine-Tuning
- Model Training

Azure AI Infrastructure

State-of-the-art silicon and systems for AI workloads

- High-Bandwidth Networking
- Microfluidic Cooling
- Azure Maia Silicon

Azure AI Foundry

A unified platform for developing generative AI apps and custom copilot experiences

Unified Platform

Multimodal AI Tools
Code-centric
Developer
Experiences

Data Integration

Ground models
using your
own data
Microsoft Fabric

Hybrid & Semantic Search

Retrieval Augmented
Generation (RAG)
Vector support

Full Development Lifecycle

Model Catalog
Prompt flow
LLMOps

Safe & Responsible AI

Content
Classification
Model Monitoring
Jailbreak Risk
Detection

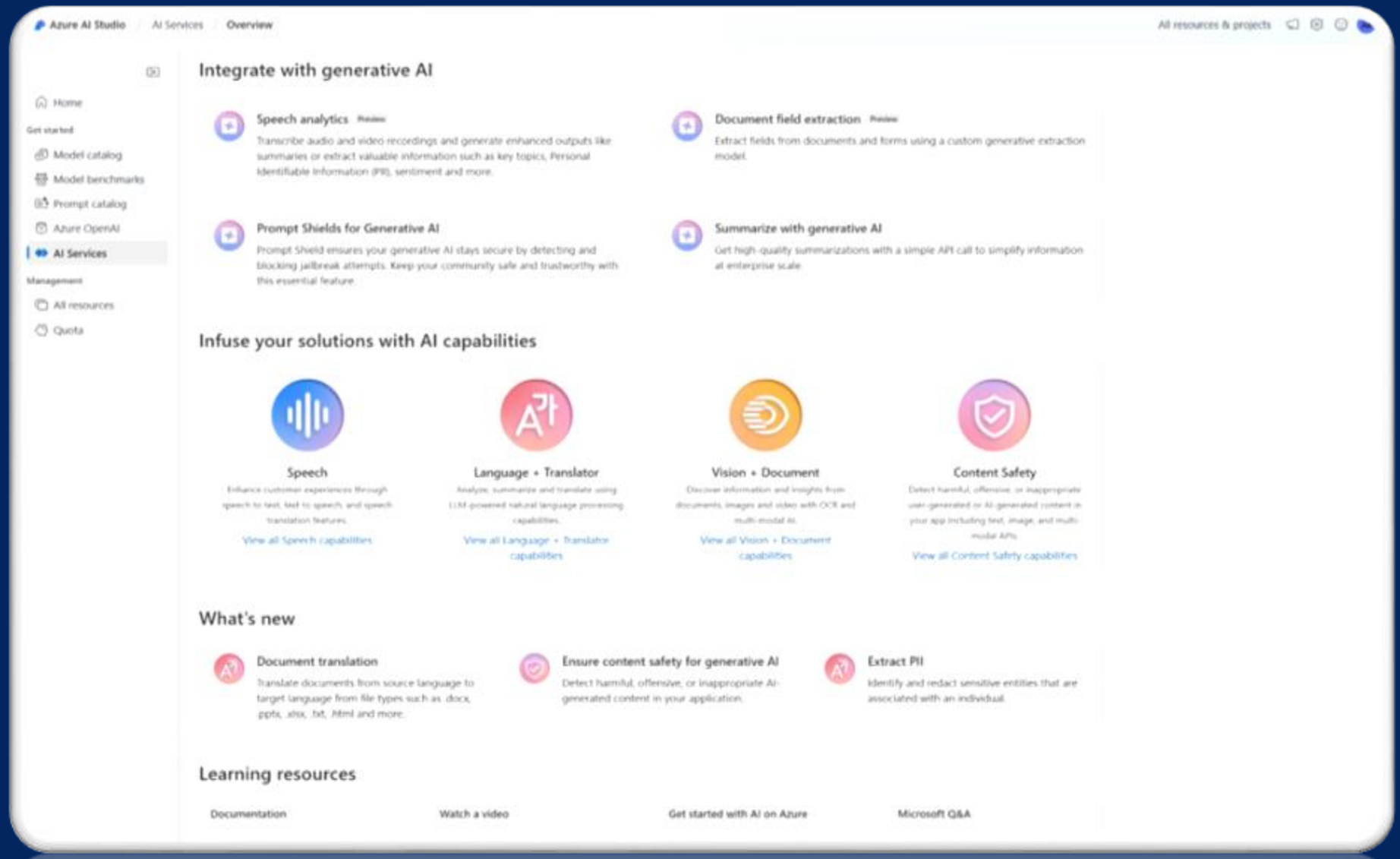
Explore, build, evaluate, and deploy AI responsibly

Complete AI Toolchain

Access collaborative, comprehensive tooling to support the development lifecycle and differentiate your apps

Orchestrate and debug AI workflows	Streamline model and app evaluations	Monitoring & observability
Streamline app development with easy-to-use prompt orchestration, tracing, and debugging via interactive visual and code-first workflows.	Quickly and continuously iterate on your application and track the impact of ongoing changes to improve application quality and safety.	Makes it easier for developers to proactively monitor the quality, safety, and operational metrics for application while in production.

Integrated cutting-edge AI Services



Demo

Azure AI Foundry

https://ai.azure.com

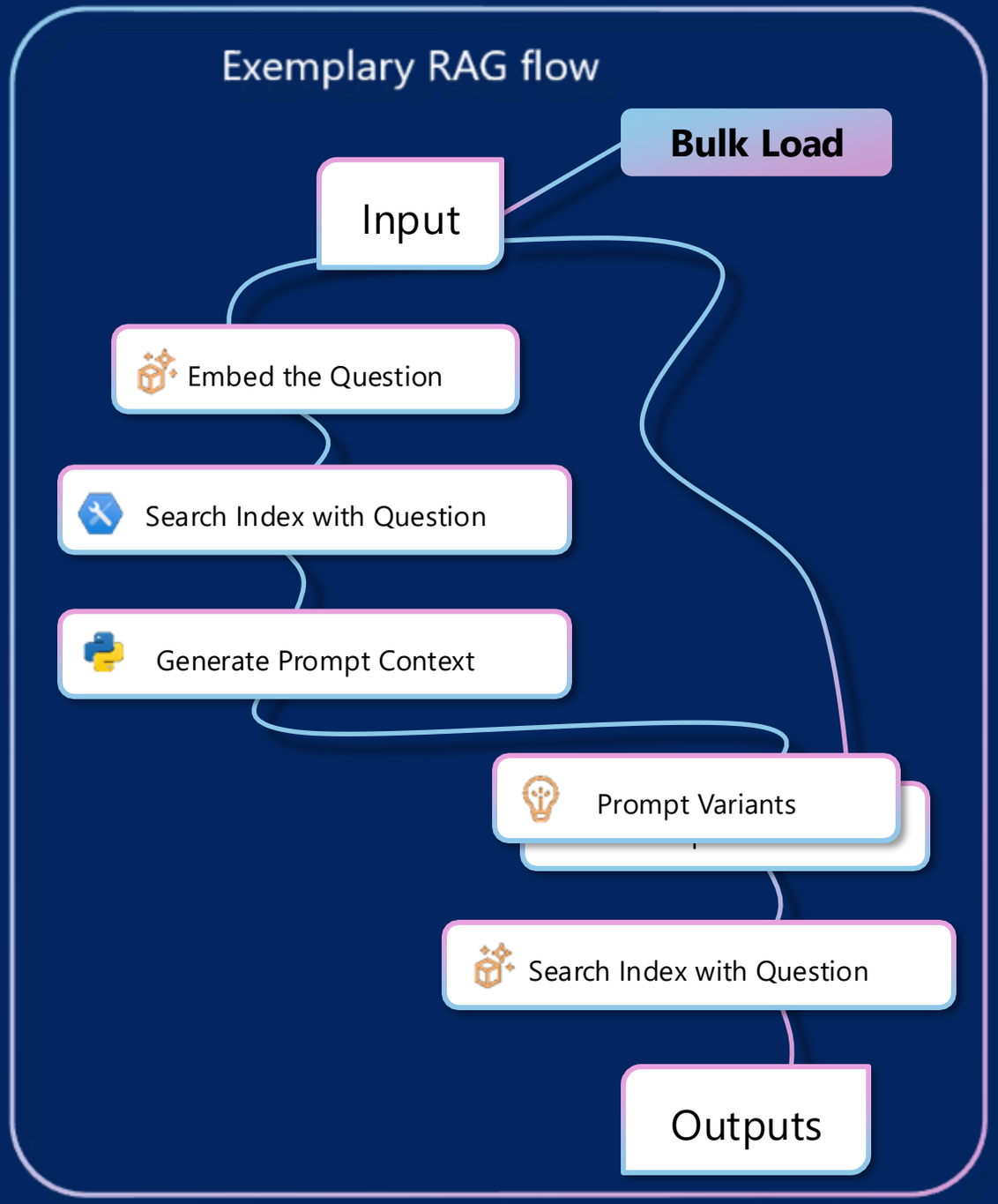
Microsoft

Azure Machine Learning / AI Foundry Prompt Flow

AI Solution Prototyping & Acceleration

Prompt Flow Overview

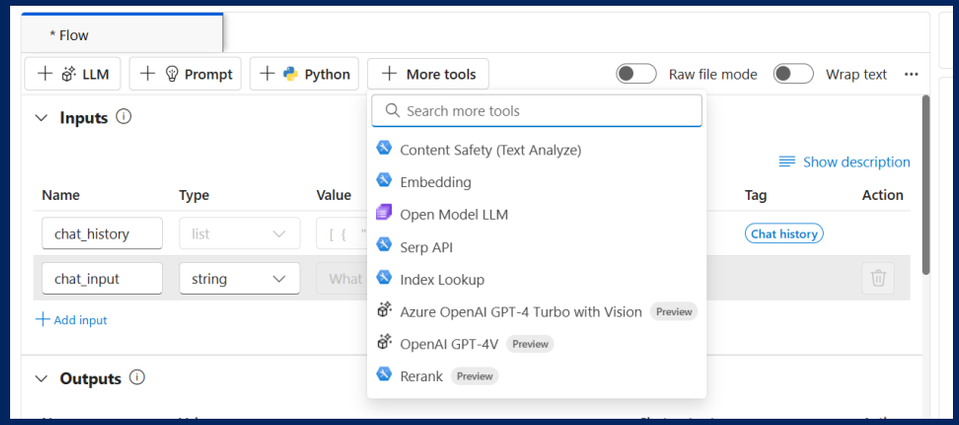
- Orchestrates AI models, prompts, and APIs
- Support for prompt tuning with variations and versions
- Evaluate the AI quality of the workflows with metrics like performance, groundedness, and accuracy
- Test and compare with blue/green deployments and testing
- Manage APIs and external connections, with support to Semantic Kernel, LangChain and Plugins



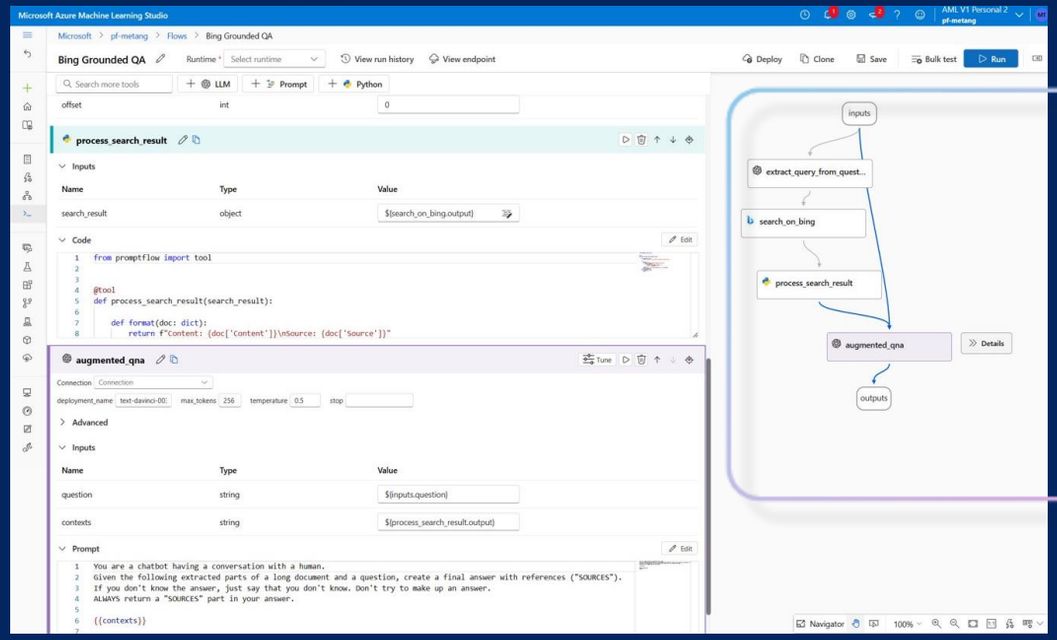
Azure Machine Learning prompt flow

Capabilities

- Develop workflows
  - Accelerate development by introducing tools – logical containers that standardize connections to various language models, external data sources, and custom code
  - Standardize development with a DAG [Directed Acyclic Graph] format for application verticals
- Test and evaluate
  - Test flows with large datasets in parallel with batch runs
  - Evaluate the AI quality of the workflows with metrics like GPT similarity, groundedness, and accuracy
- Prompt tuning
  - Easily tune prompts with variants and versions
- Compare and deploy
  - Visually compare across experiments
  - One-click deploy to a managed endpoint for rapid integration



Tools



DAG graph

Azure Machine Learning Prompt Flow

Prompt flow authoring

Develop your LLM flow from scratch

- Construct a flow using pre-built tools
- Support custom code
- Clone flows from samples
- Track run history

LLM apps can be defined as Directed Acyclic Graphs (DAGs) of function calls. These DAGs are flows in prompt flow UI.

A DAG flow is a DAG of functions (we call them *tools*). These functions/tools connected via input/output dependencies are executed based on the topology by prompt flow executor. Also conditional function logic is supported. The graph structure is defined in flow.dag.yaml file.

Alternatively, you can define a FLEX flow (defined in the flow.flex.yaml file) that points to flow entry (entry:function_name or entry:ClassName) and is less rigid than DAG, not requiring the use of pre-defined tools. This enables testing, running, or viewing traces via the Prompt Flow VS Code Extension or any IDE of your choice.

The term Flex is a shorthand for flexible, indicating its adaptability to most scenarios with minimal adjustments. A FLEX flow is a type of flow that uses a Python function or class as the entry point, which encapsulates the LLM app logic. These entries can be manipulated with pure code experience.

LLM Ops in Practice

From an idea to a full-scale solution

LLM Ops Maturity Model

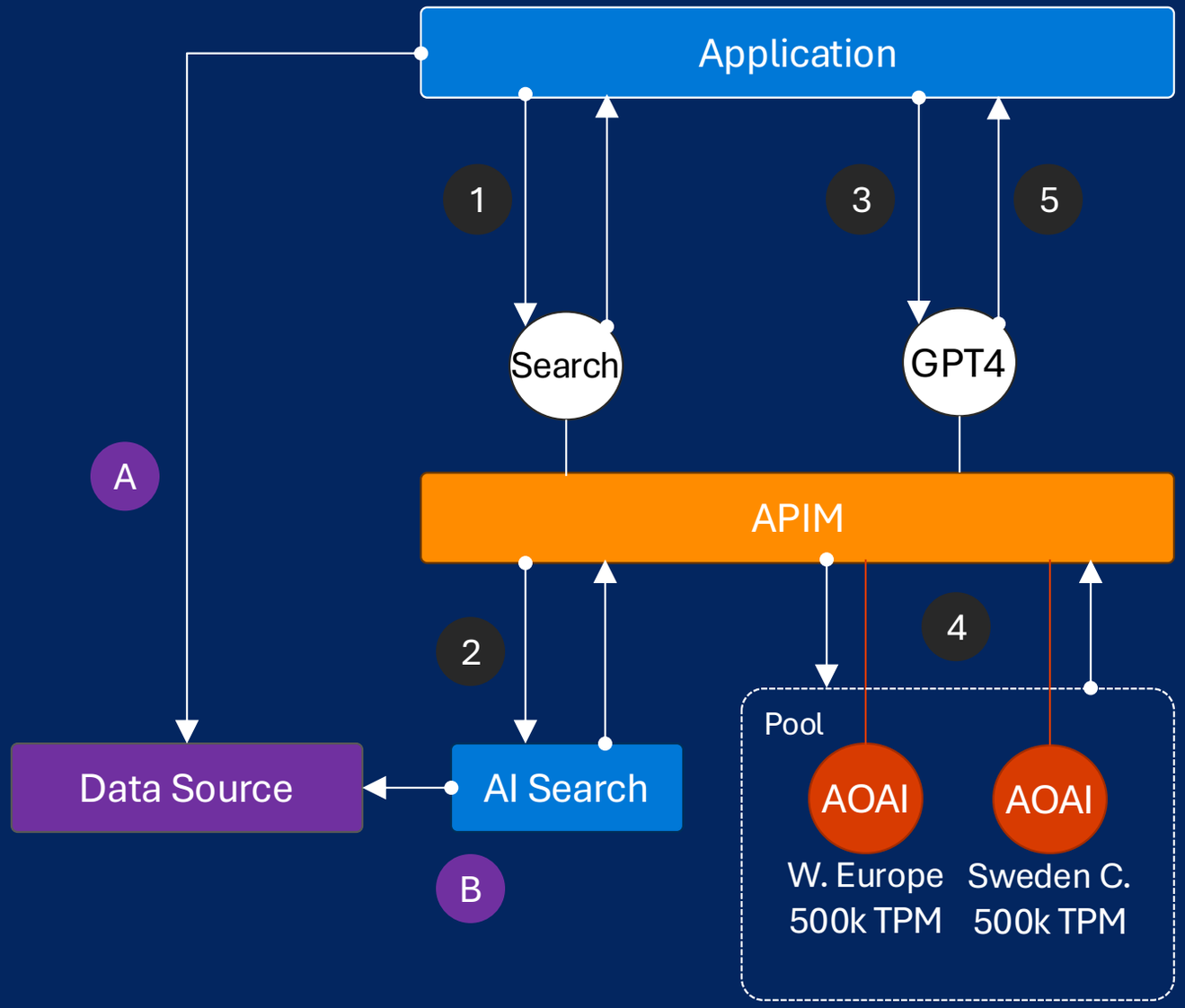
Achieve generative AI operational excellence with the LLMOps maturity model | Microsoft Azure Blog

Scalable RAG Architecture

A RAG workflow using Application components, Azure AI Search, and Azure OpenAI to deliver Generative AI

- APIM (API Manager) orchestrates all API activity to secure, monitor, and scale organizations of any size
- Load balancer pool with Azure Open AI with TPM cost model with guaranteed throughput to provide a resilient and cost-efficient solution
- Upgrade paths for AI service and Application are simplified when orchestrating with APIM along with settings

Example of an LLM RAG App



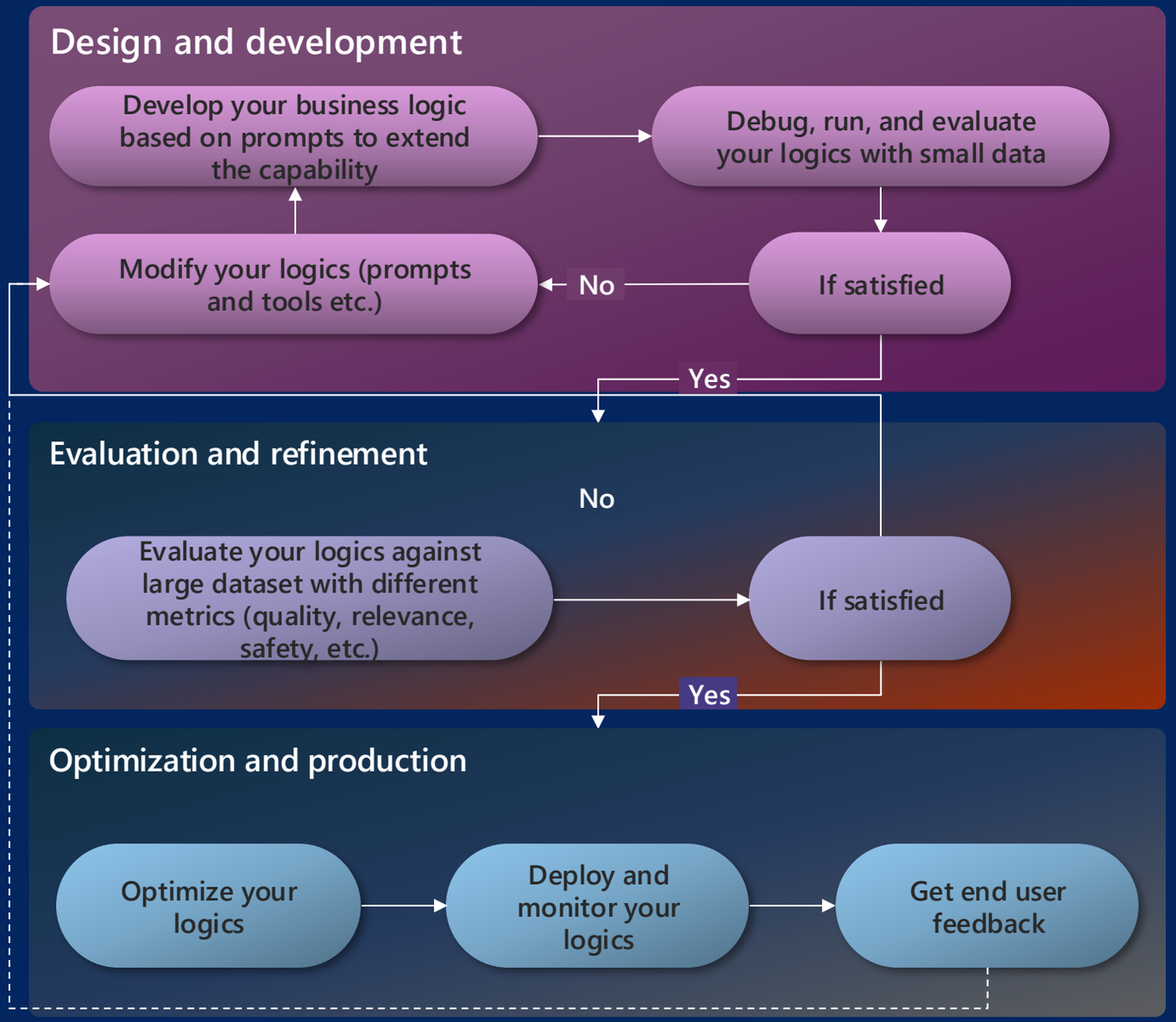
- **A** Admin uploads data.
- **B** Search indexer + skills chunks new files, vectorizes the chunks (aka embedding), and tags chunks and embeddings.
- **1** User submits prompt (aka question).
- **2** Message send to Search to find the most relevant data from the indexed copies of the data in the source.
- **3** Retrieved + prompt + chat history is sent to Azure OpenAI model .
- **4** Model uses all the information sent to it to process the user prompt and generate a response.
- **5** Response is sent to the user.

Operationalize LLM app development

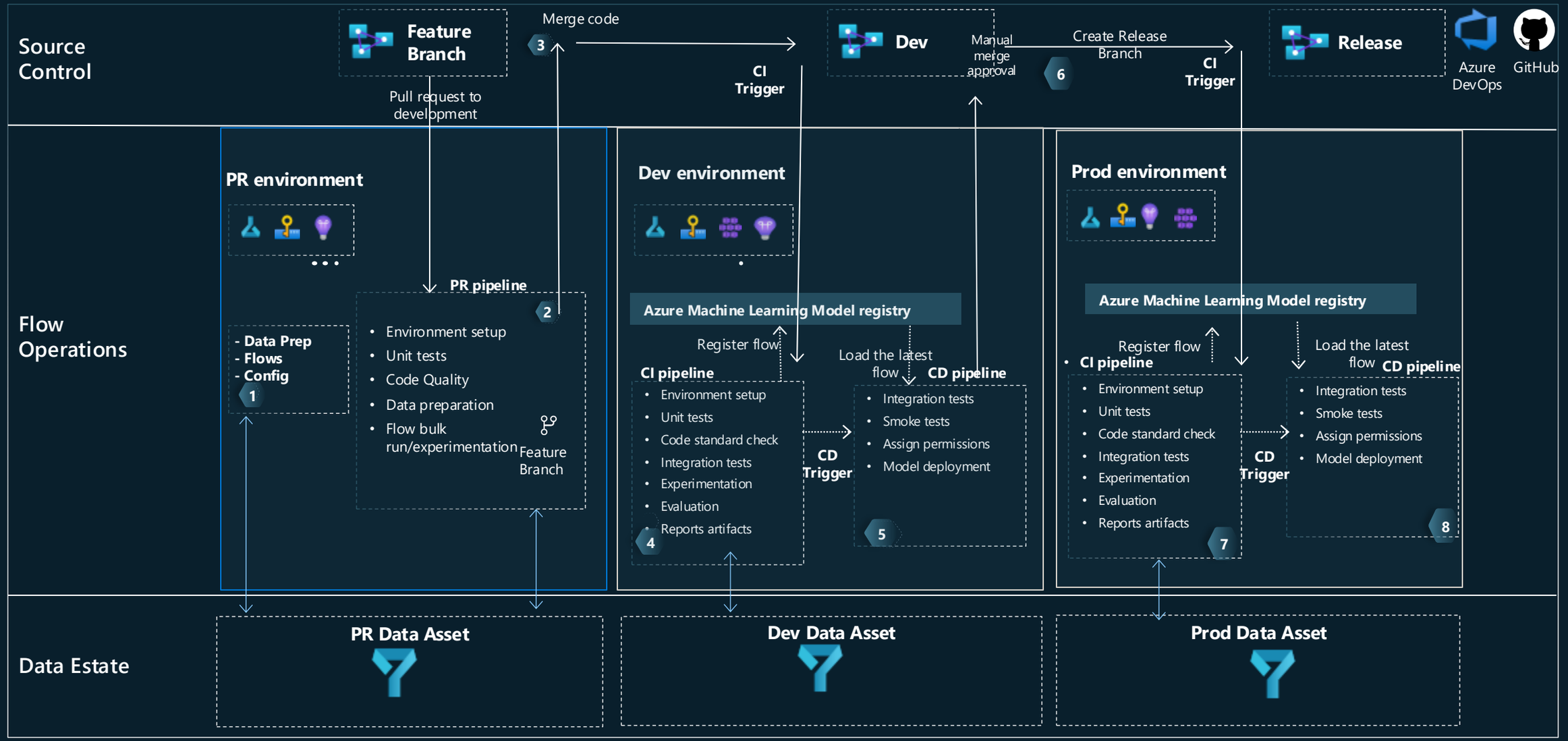
LLMOps is a complex process.

Customers want:

- Private data access and controls
- Prompt engineering
- CI/CD
- Iterative experimentation
- Versioning and reproducibility
- Deployment and optimization
- Safe and Responsible AI



CI/CD Process Flow: An Example



Source: LLMOps process Flow

Microsoft

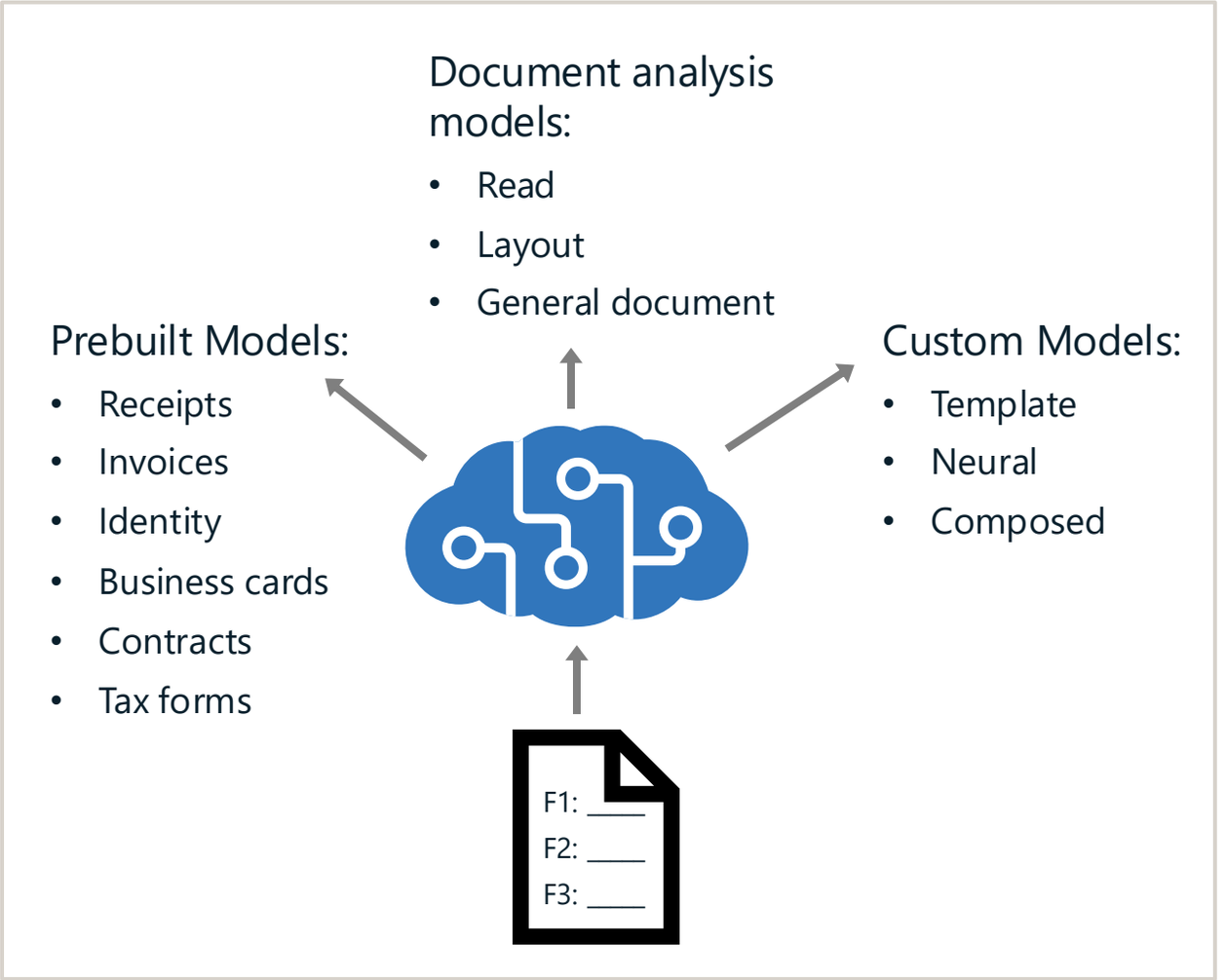
Azure AI Document Intelligence

The Document Intelligence Service

Data extraction from forms and documents:

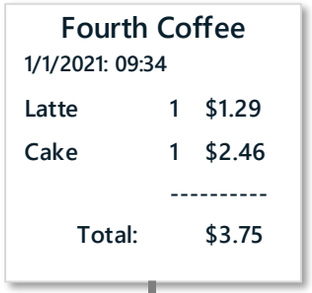
- Document analysis from general documents
  - Read: OCR for printed and written text
  - Layout: Extract text and structure
  - General document: Extract text, structure, and key-value pairs
- Prebuilt models for common form types
- Train custom models for your own forms
  - Custom template: Extract data from static layouts
  - Custom neural: Extract data from mixed-type documents
  - Custom composed: Collection of multiple models assigned to a single model

Provision as single-service **Document Intelligence** resource or multi-service **Azure AI Services** resource



Prebuilt models

Receipt

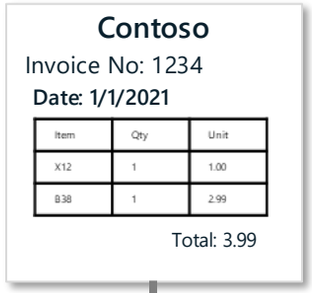


Analyze Receipt

Get Analyze Receipt Result

```
{
  "MerchantName": "Fourth Coffee",
  "TransactionDate": 2021-01-01,
  "TransactionTime": 09:34,
  "Items" [
    ...],
  "Total": 3.75
}
```

Invoice

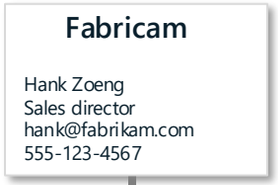


Analyze Invoice

Get Analyze Invoice Result

```
{
  "VendorName": "Contoso",
  "InvoiceNumber": 1234,
  "InvoiceDate": 2021-01-01
  "Tables" [
    ...],
  "TotalInvoiceAmount": 3.99
}
```

Business Card



Analyze Business Card

Get Analyze Business Card Result

```
{
  "ContactNames": [
    {
      "FirstName": "Hank",
      "LastName": "Zoeng"
    }],
    ...
}
```

Calling the API

- Each request is configured with your resource endpoint and needs your resource key
- Send the request, which when successful returns a poller to get the results
  - REST returns it in `Operation-Location` header
  - SDKs return an object from the request
- Query the poller received for the extracted data

REST

```
Request POST:
{endpoint}/documentintelligence/documentModels/prebuilt-layout:analyze?api-version={version}

Operation-Location:
{endpoint}/documentintelligence/documentModels/prebuilt-layout/analyzeResults/ab12345c-12ab-23cd-b19c-2322a7f11034?api-version={version}
```

C#

```
AnalyzeDocumentOperation operation = await client.AnalyzeDocumentFromUriAsync(WaitUntil.Completed, "prebuilt-layout", fileUri);

AnalyzeResult result = operation.Value;
```

Python

```
poller=document_analysis_client.begin_analyze_document_from_url("prebuilt-document", docUrl)

result = poller.result()
```

API response

- Response it broken down by page, lines, and words
- Subset of REST response included here
- SDK response objects have similar structure, broken down similarly
- Additional data about detected text or selection marks, such as bounding box and handwritten style


```
{
    "analyzeResult": {
        "apiVersion": “{version}",
        "modelId": "prebuilt-invoice",
        ...
        "pages": [{
            "pageNumber": 1,
            "angle": 0,
            "width": 8.5,
            "height": 11,
            "unit": "inch",
            "words": [{
                "content": “Margie’s",
                "boundingBox": [
                    0.5911,
                    0.6857,
                    1.7451,
                    0.6857,
                    1.7451,
                    ...
                ],
                "confidence": 1,
                "span": {...}
            }],
        }]
    }
}
```

Types of custom models

Custom classification

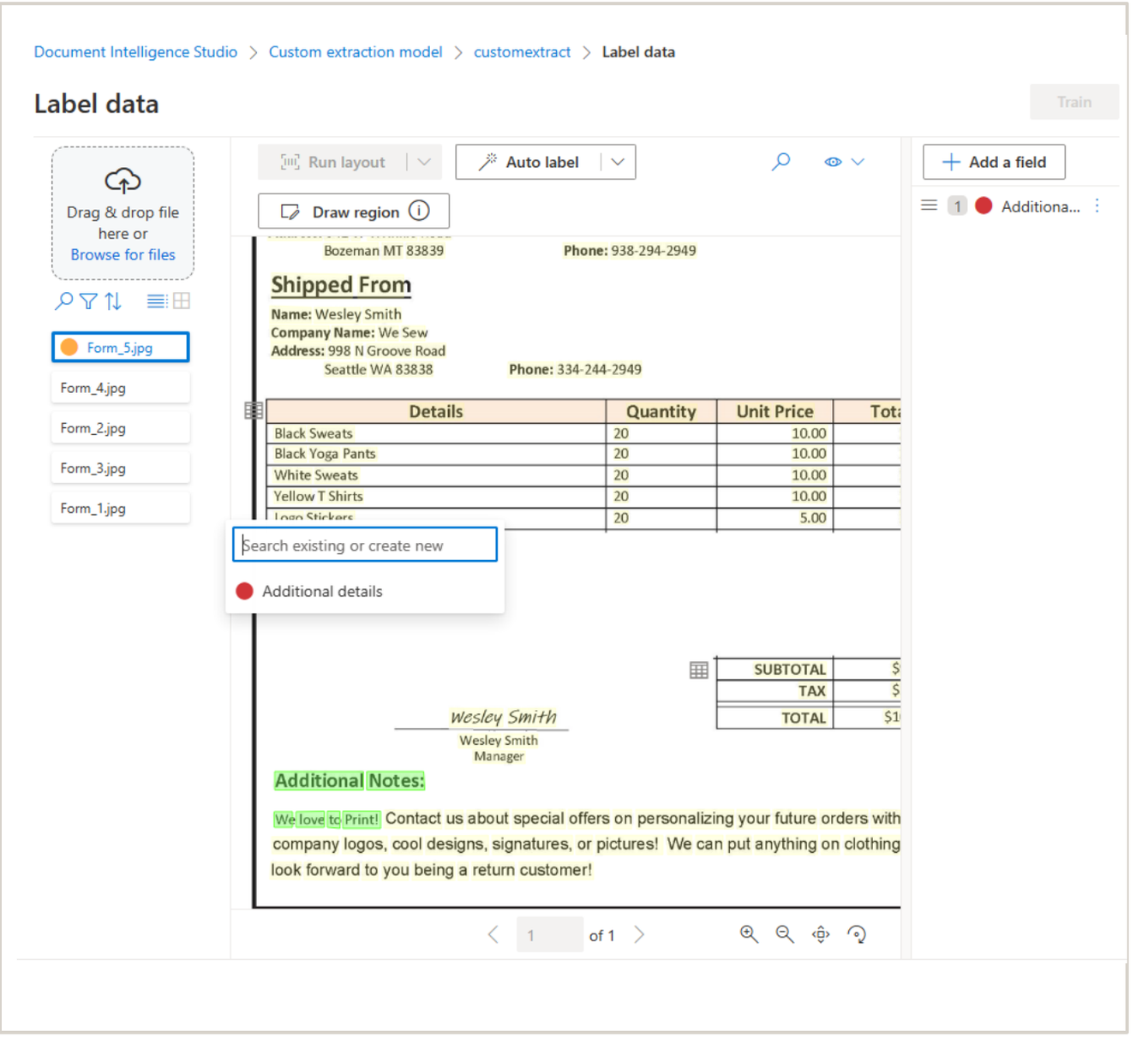
- Apply a label to the entire document
- Ideal for sorting large numbers of incoming documents into types
- Requires two different classes, and a minimum of five labeled documents per class
- One type of training model

Custom extraction

- Apply label to specific text
- Ideal for extracting custom labels from documents
- Requires five examples of the same document type
- Two training methods:
  - **Custom template (custom form)**
    - Training time: 1-5 minutes
    - Document structure: forms, templates, other structured documents
  - **Custom neural (custom document)**
    - Training time: 20-60 minutes
    - Document structure: structured and unstructured documents

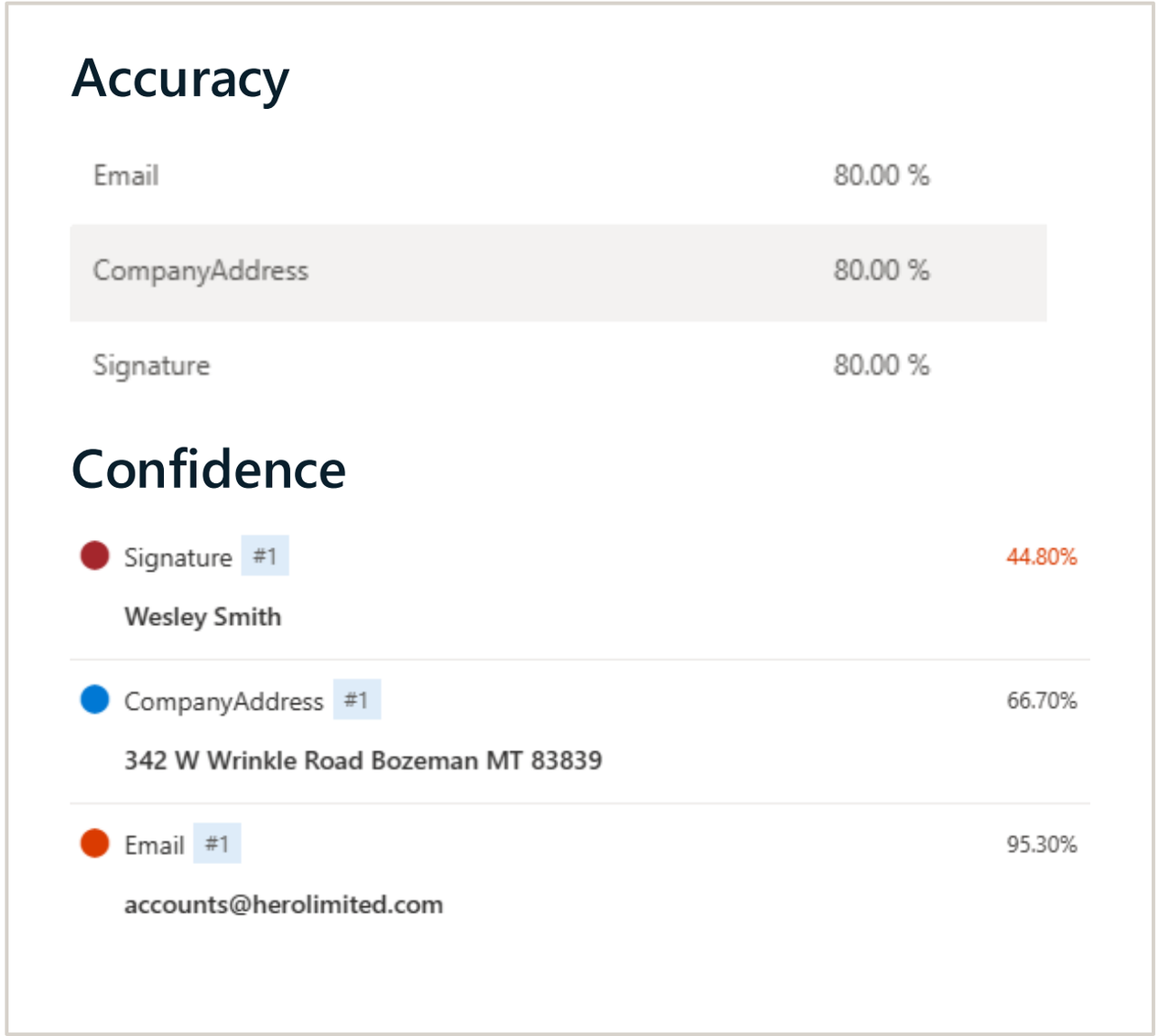
Training Custom Models

1. Create project and upload training files to your project, or connect to blob storage containing files
2. Add data type (such as field or signature) to start labeling your dataset
3. Select a word in the document, and assign one of the fields to label it
4. Repeat for all fields and files in your dataset
5. Layout and auto label (using a prebuilt model) can assist in this process
6. Train the model, providing a Model ID used in API requests



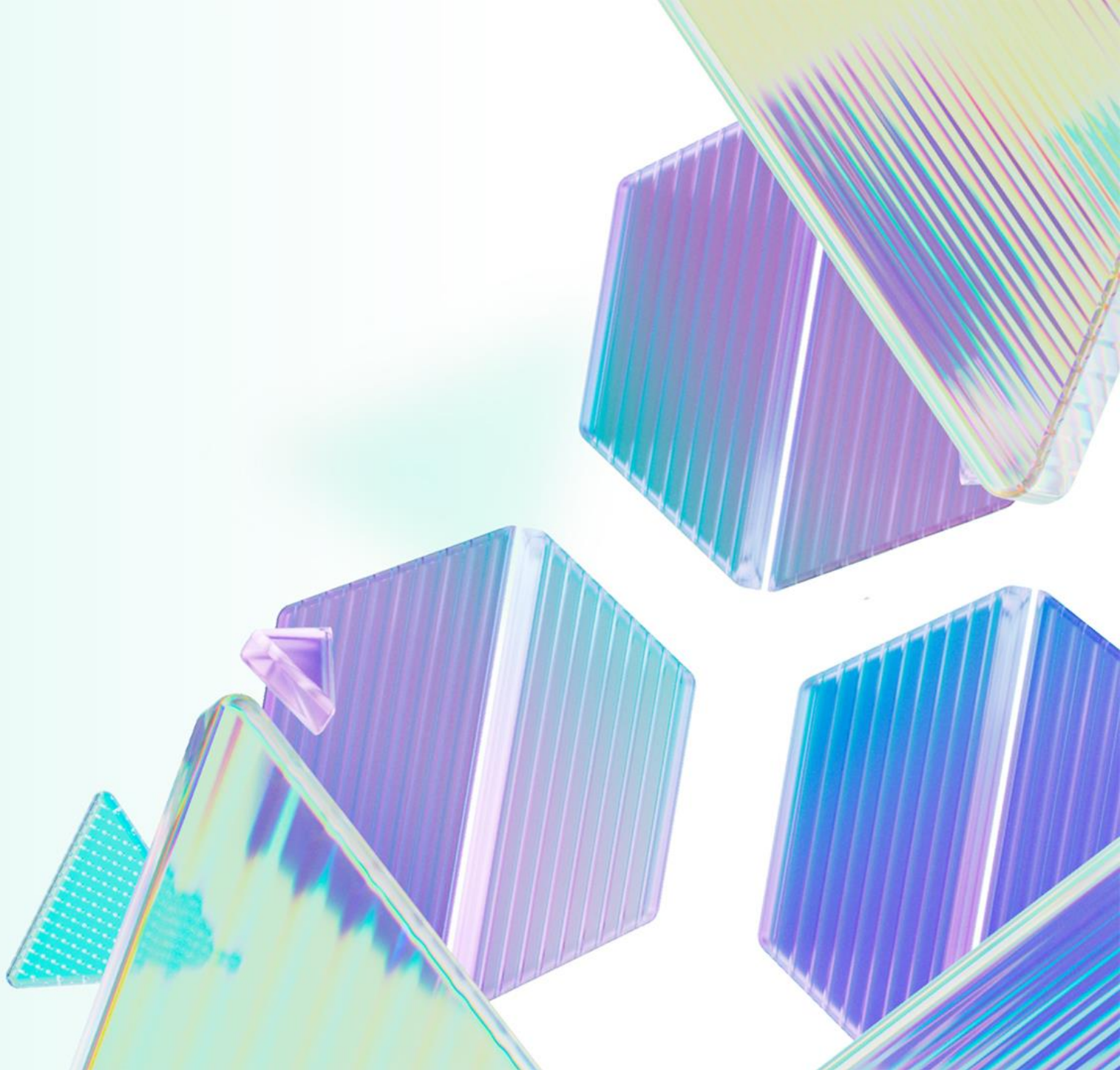
Accuracy and confidence scores

- After training, a custom model has an estimated accuracy score
- Score is calculated by running combinations of training data predictions against the labeled values
- Confidence score is the same as using prebuilt models, indicated how accurate the model thinks that specific prediction is
- Confidence scores are provided in the response from the model for each predicted label



Microsoft

Azure AI Content Understanding



Content Understanding

1

No complex prompt engineering, define task-specific schema

2

Results grounded to input content when extracted

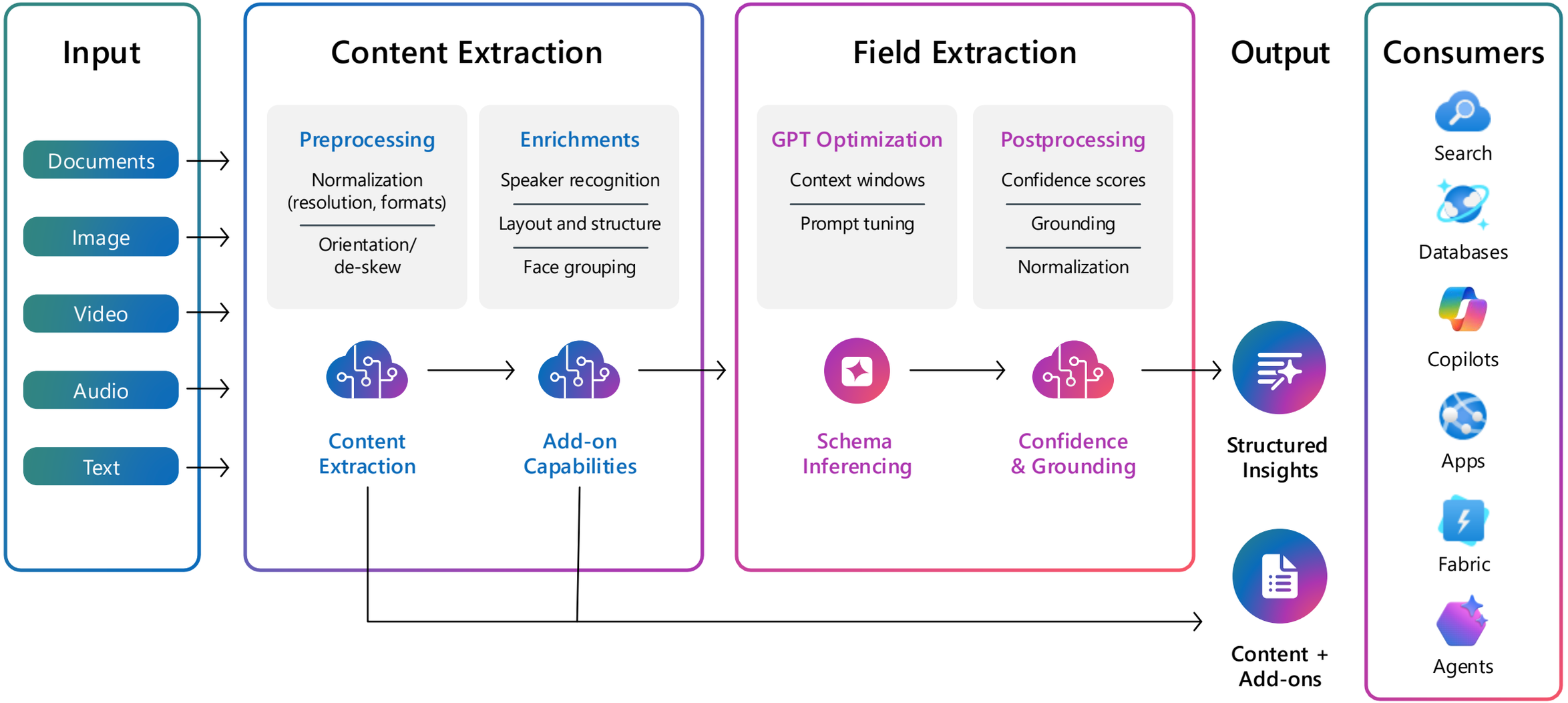
3

Confidence scores for automation and validation

4

Enrich the content when needed

Components of Content Understanding



Best practices for accelerating time to value

Improve Accuracy

Guide Output with Field Descriptions/Fix Mistakes by Editing

Use Classification Fields for Specific Outputs

Implement the "Deliberation Before Decision" pattern for complex Tas

Select Generative Vs Extractive methods based on Use Case

Specify the Desired Output Language for Generative Fields

Edit and improve field details to correct mistakes

Set language for audio/video when known, esp. for single-language content

Reduce Costs

Grounding and use of confidence scores reduces manual intervention

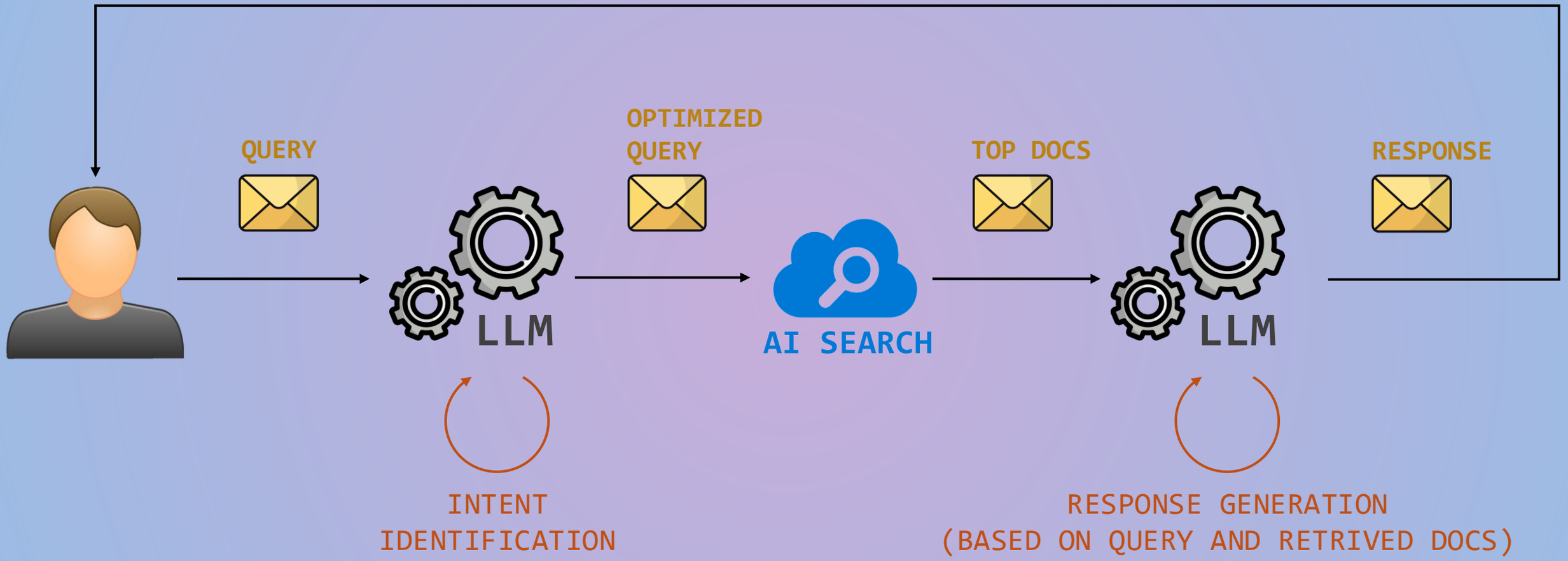
Resources

Announcement Blog	aka.ms/content-understanding-launch-blog
Breakout session @ Microsoft Ignite 2024	youtu.be/vYjAg27aHkA?si=k73mzRiA7tgEOhoz
Product Page	aka.ms/content-understanding
Pricing Page	azure.microsoft.com/pricing/details/content-understanding
Docs	aka.ms/content-understanding-mslearn
AI Studio	aka.ms/content-understanding-aistudio
Samples Repositories	aka.ms/cu-samples github.com/Azure-Samples/azure-ai-content-understanding-python github.com/Azure-Samples/azure-ai-search-with-content-understanding-python github.com/Azure-Samples/azure-ai-content-understanding-with-azure-openai-python

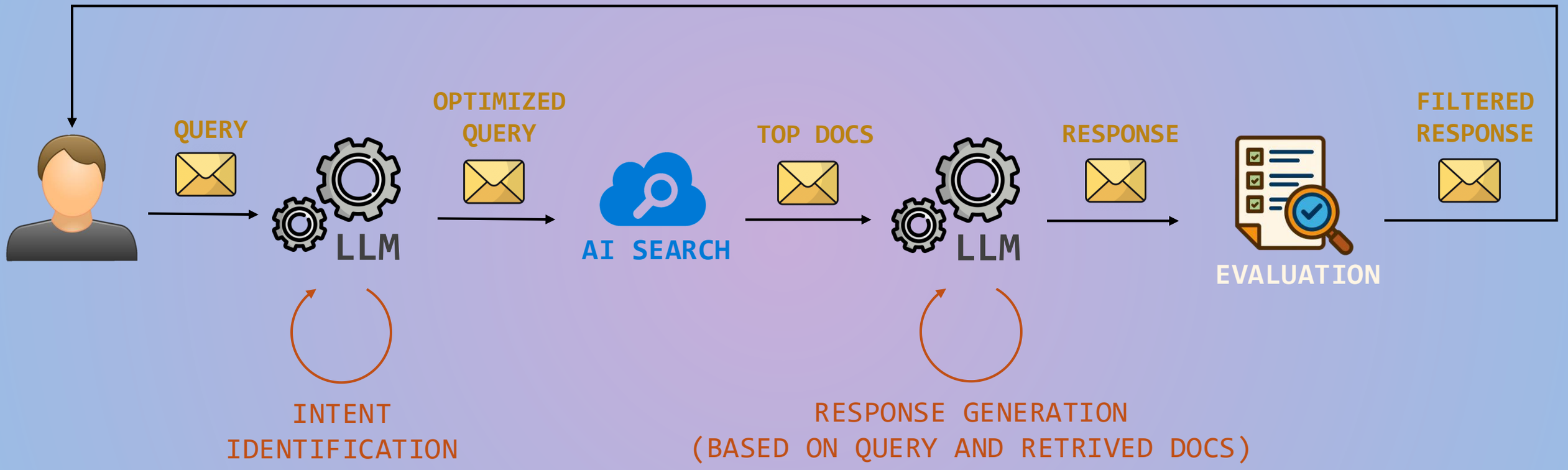
Q&A

ADVANCED RAG WITH AI FOUNDRY SDK

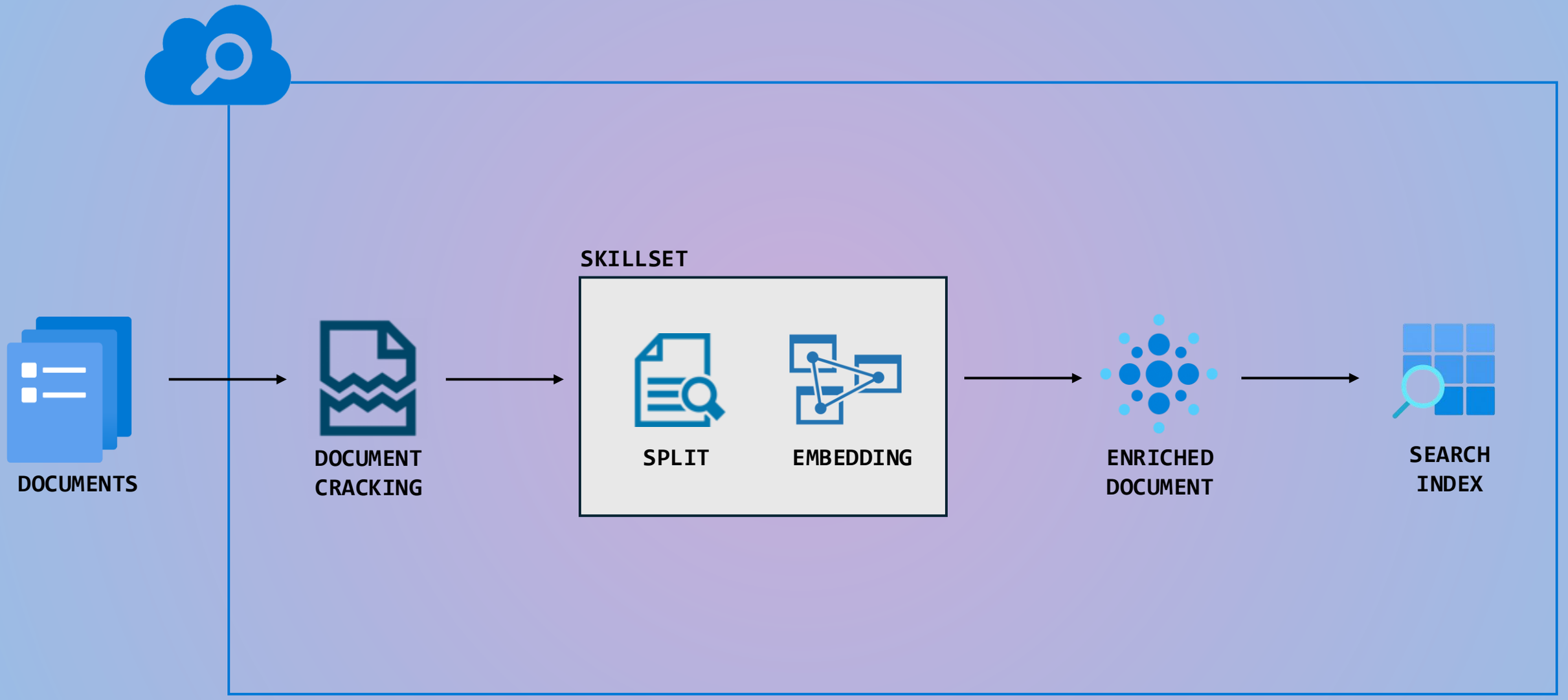
CHATBOT APP MESSAGE FLOW



CHATBOT APP MESSAGE FLOW (WITH EVALUATION STEP)



DOCUMENT ENRICHMENT PIPELINE IN AZURE AI SEARCH

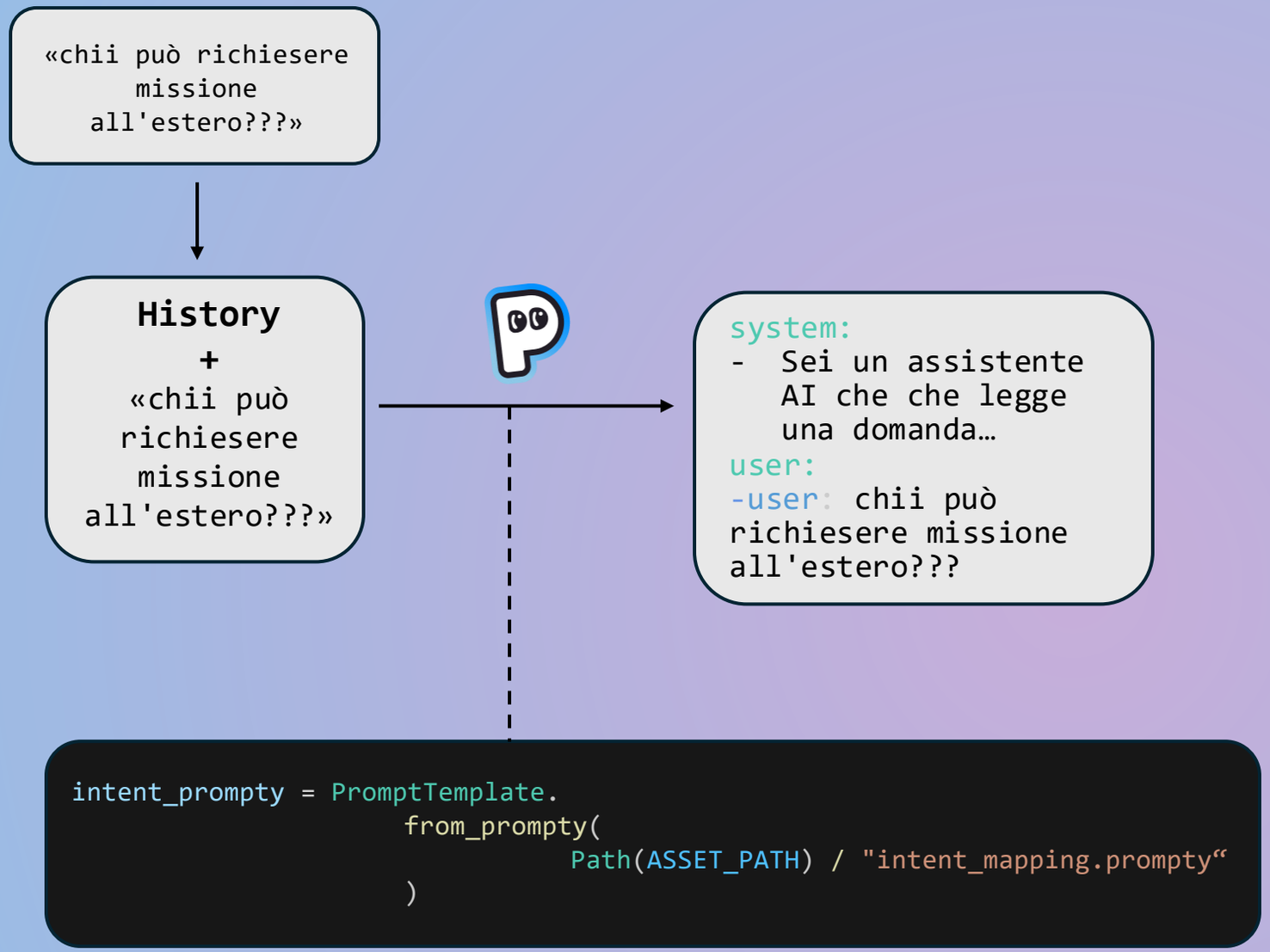


«chii può richiesere
missione
all'estero???»

«chii può richiesere missione all'estero???»

```
history.append({"role": "user", "content": user_input})
```

History
+
«chii può richiesere missione all'estero???»


«chii può richiesere missione all'estero???»
History + «chii può richiesere missione all'estero???»
system:
- Sei un assistente AI che che legge una domanda…
user:
-user: chii può richiesere missione all'estero???
intent_prompty = PromptTemplate.
from_prompty(
Path(ASSET_PATH) / "intent_mapping.prompty“
)

«chii può richiesere missione all'estero???»

History
+
«chii può richiesere missione all'estero???»

```
intent_mapping_response = chat.complete(
        model=os.environ["INTENT_MAPPING_MODEL"],
        messages=intent_prompty.create_messages(conversation=messages),
        **intent_prompty.parameters,
    )
```

system:
- Sei un assistente AI che che legge una domanda…

user:
-user: chii può richiesere missione all'estero???

intent: «l'utente vuole sapere chi ha diritto alla richiesta di missioni all'estero»

search_query: «chi può richiedere missioni all'estero»

«chii può richiesere missione all'estero???»

History
+
«chii può richiesere missione all'estero???»

system:
- Sei un assistente AI che che legge una domanda…

user:
-user: chii può richiesere missione all'estero???

intent: «l'utente vuole sapere chi ha diritto alla richiesta di missioni all'estero»

search_query: «chi può richiedere missioni all'estero»

```
search_results = search_client.search(
        search_text=search_query,
        vector_queries=[vector_query],
        select=["chunk_id", "title", "chunk"]
    )
```

3 documents retrieved:

Document 1 : Reg_Mis_2025.pdf
Articolo 3. Soggetti aventi titolo al conferimento dell'incarico di…

Document 2 : Reg_Mis_2025.pdf
La missione è strettamente personale, con la conseguenza che…

Document 3 : REG_mis-web.pdf
L'incarico e l'autorizzazione a compiere la missione devono…

«chii può richiesere missione all'estero???»

History
+
«chii può richiesere missione all'estero???»

system:
- Sei un assistente AI che che legge una domanda…

user:
-user: chii può richiesere missione all'estero???

intent: «l'utente vuole sapere chi ha diritto alla richiesta di missioni all'estero»

search_query: «chi può richiedere missioni all'estero»

3 documents retrieved:

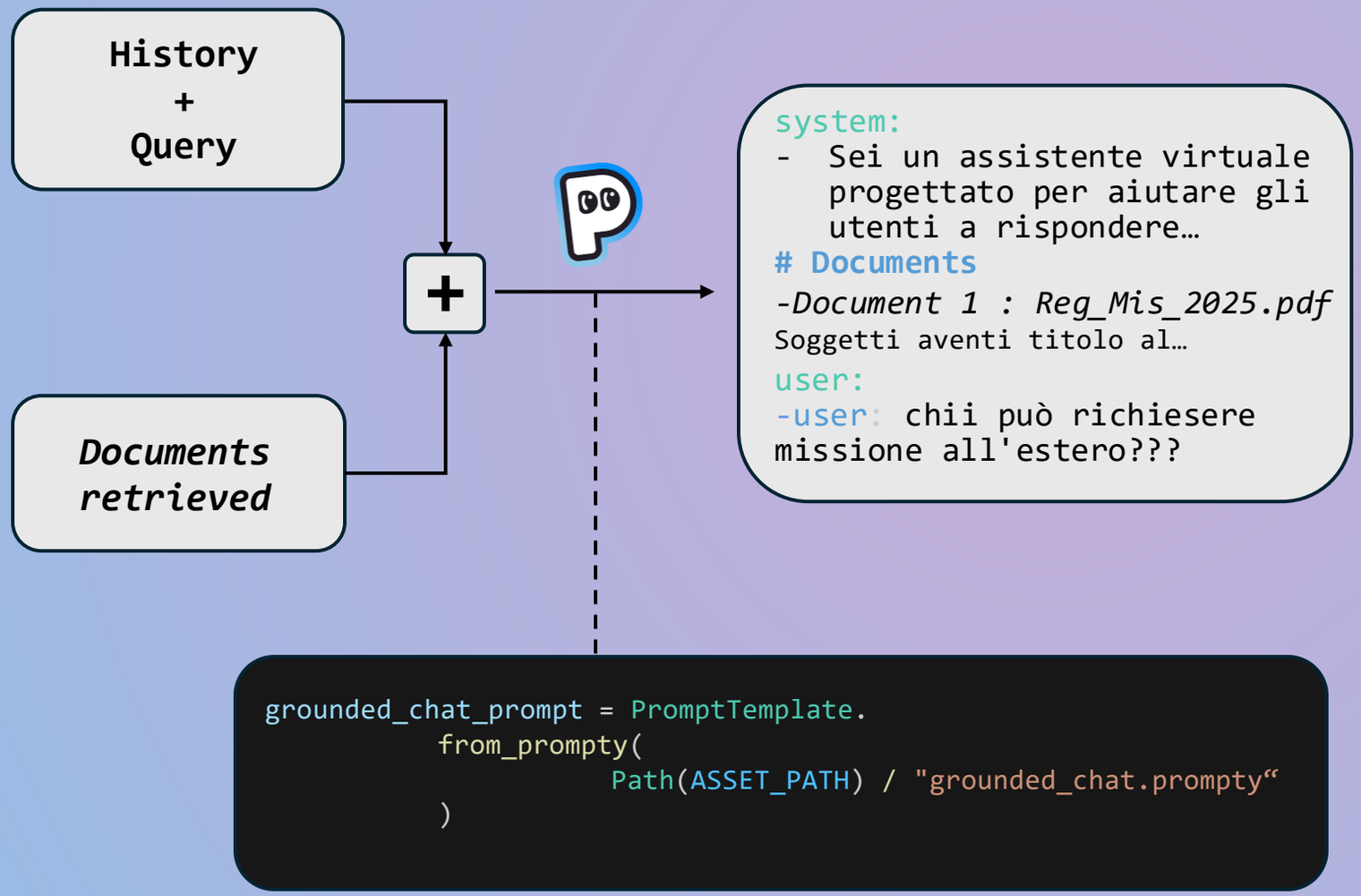
Document 1 : Reg_Mis_2025.pdf
Articolo 3. Soggetti aventi titolo al conferimento dell'incarico di…

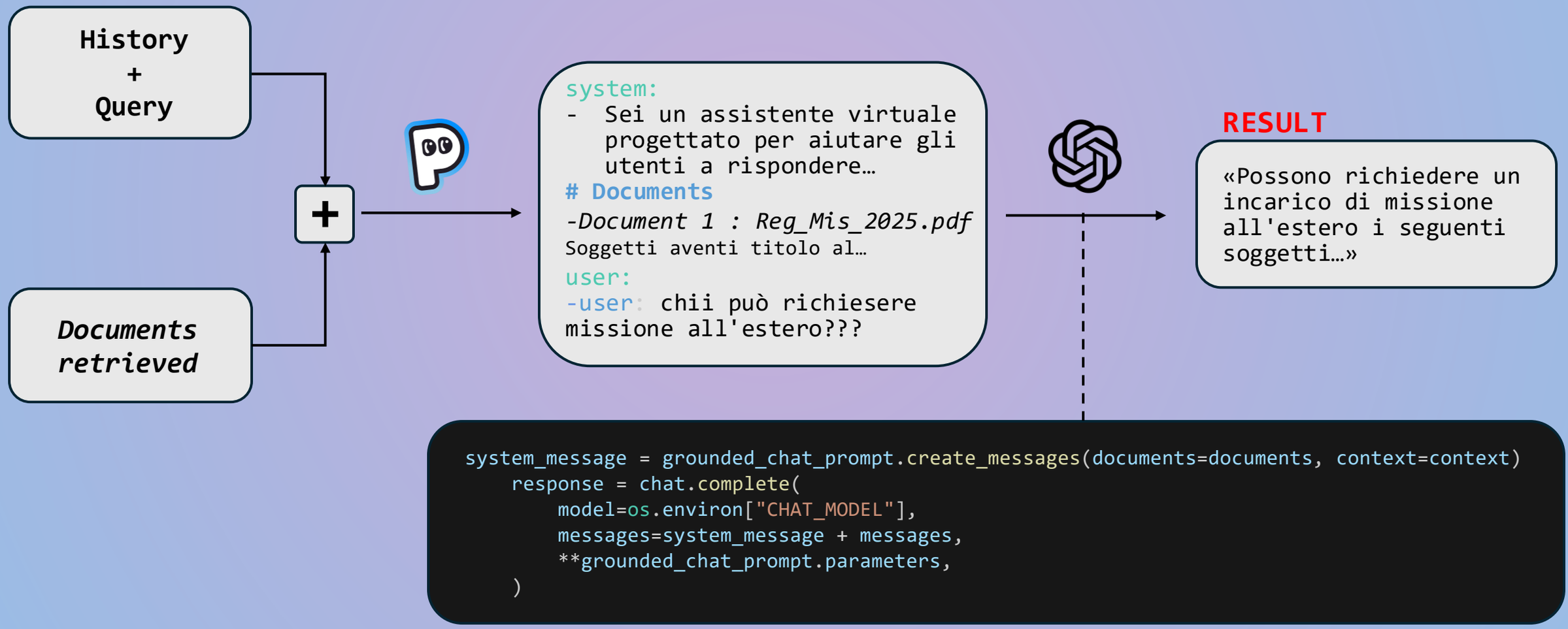
Document 2 : Reg_Mis_2025.pdf
La missione è strettamente personale, con la conseguenza che…

Document 3 : REG_mis-web.pdf
L'incarico e l'autorizzazione a compiere la missione devono…

History
+
Query

Documents retrieved


History
+
Query
Documents retrieved
+
system:
- Sei un assistente virtuale progettato per aiutare gli utenti a rispondere…
Documents
-Document 1 : Reg_Mis_2025.pdf
Soggetti aventi titolo al…
user:
-user: chii può richiesere missione all'estero???
grounded_chat_prompt = PromptTemplate.
from_prompty(
Path(ASSET_PATH) / "grounded_chat.prompty“
)


History
+
Query
Documents retrieved
system:
- Sei un assistente virtuale progettato per aiutare gli utenti a rispondere…
Documents
-Document 1 : Reg_Mis_2025.pdf
Soggetti aventi titolo al…
user:
-user: chii può richiesere missione all'estero???
RESULT
«Possono richiedere un incarico di missione all'estero i seguenti soggetti…»
system_message = grounded_chat_prompt.create_messages(documents=documents, context=context)
response = chat.complete(
model=os.environ["CHAT_MODEL"],
messages=system_message + messages,
**grounded_chat_prompt.parameters,
)

Microsoft

Wrap-up