

# SIMLE AI DEMO TOOLKIT GRADIO + AOAI OYD

PyData meetup Nov 24

Andrey Vykhotsev (AI Dev @ Aprova GmbH.)

THU, JAN 30, 6:00 PM

**Explaining Machine Learning models with Python**

 Microsoft office

--- UPD: Moving one week ahead as some folks asked me to. Hi Everyone! As promised, scheduling the next meetup that will take place in the Microsoft office. This time I will present on the topic of model interpretability / explainability. Agenda: - overview of the interpretability for...

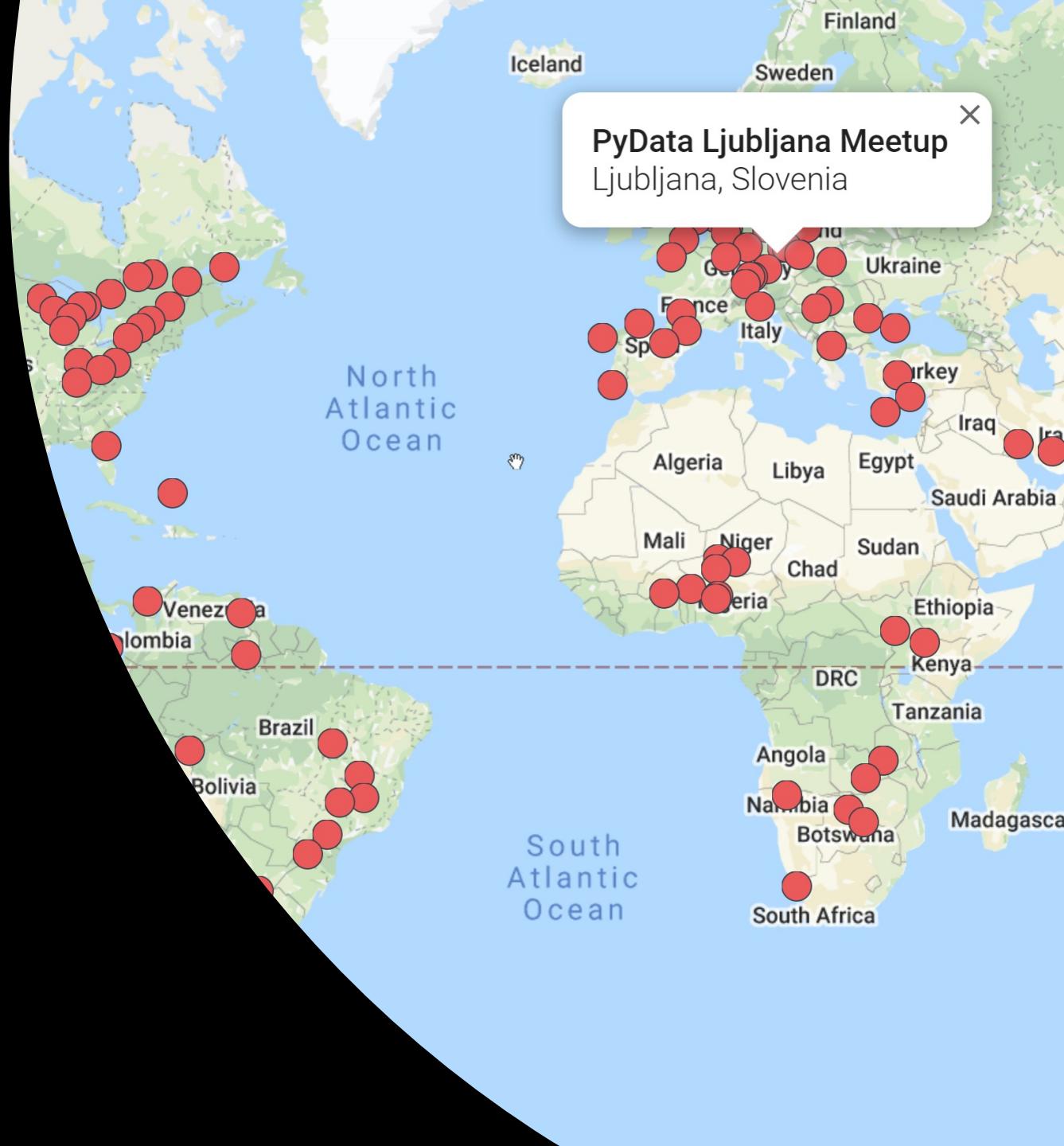
 49 attendees

Manage 

 Going

# QUICK INTRO INTO PYDATA

- Numbers
  - 61 countries
  - 164 groups
  - 156065 members
- By NumFocus.org  
<https://numfocus.org/programs/pydata>
- Since 2017?
- 13<sup>th</sup> in person meetup



# WANT TO BE AN ORGANIZER / PRESENTER?



# **GRADIO FOR BUILDING DEMOS**

# Azure OpenAI on your data



Connect or ingest your data



Ground Azure OpenAI models using your data



Retrieval augmented generation made easy



Restrict responses to your data

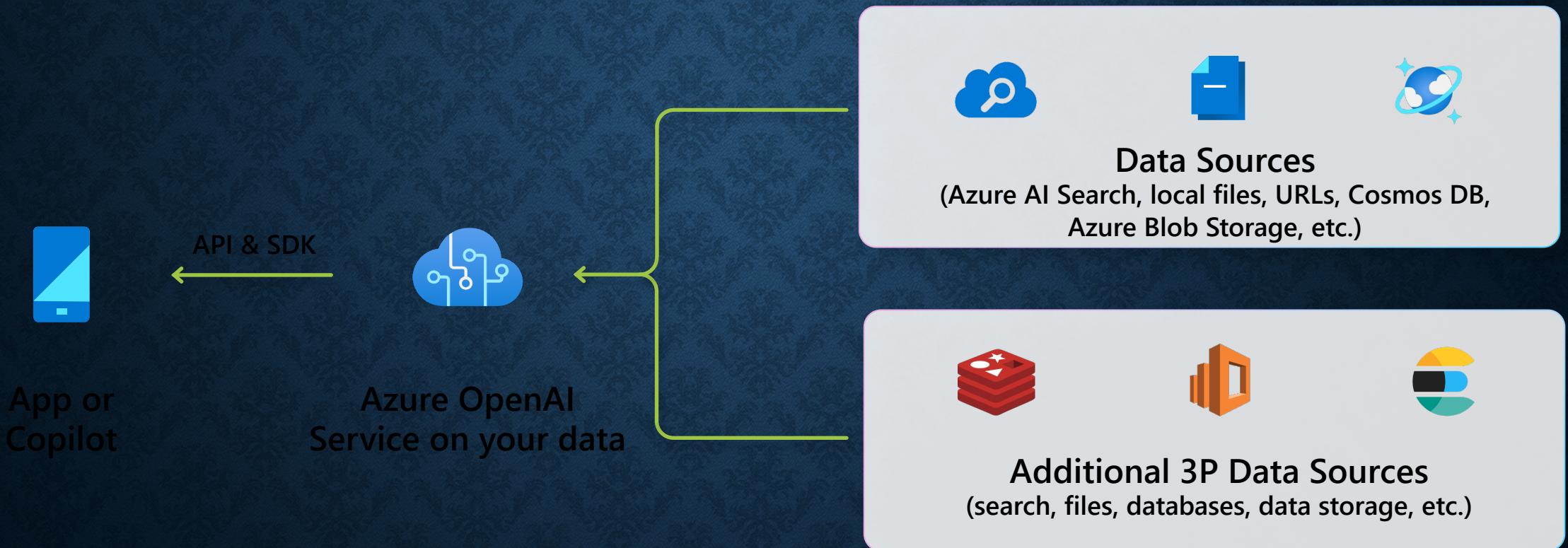


Enterprise grade security including private network, VPN, document level access control, etc.



Easy deployment as a copilot

# Azure OpenAI on your data



# ON YOUR DATA – USE CASES

Use Cases	Details
<b>Automated Customer Assistance</b>	Provide quick responses to frequently asked questions and guide users through common troubleshooting steps based on the customers' data.
<b>Real-time Document Search</b>	Provide real-time support to product specifications and guide users through common troubleshooting steps based on product manuals
<b>Citizen Service</b>	Provide quick responses to frequently asked public service questions and guide users through specific steps based on citizen service support
<b>Learning Assistant</b>	Offer explanations and examples to help users understand academic concepts or learn new skills based on specific curricula
<b>Legal Review</b>	Quick access to legal insights from existing and upcoming legislation to properly advise clients
<b>Marketing Insights</b>	Tap into internal and external resources to respond to internal and external marketing inquiries
<b>Software Development</b>	Generate sample code based on the customer's needs
<b>HR Support</b>	Provide quick responses to frequently asked HR questions based on the customers' HR policy
<b>Industry/Competitive Insights</b>	Tap into publicly available resources to gain insights on the industry and competitors
<b>Health Advice</b>	Provide general information on symptoms, first aid, or healthy living.
<b>Predictive Maintenance</b>	Provide predictive maintenance and customer support based on customer's historic data

# Connect with your own data from various data sources with enterprise-grade security

**Add data**

Data source  
 Data management  
 Review and finish

### Select or add data source

Your data source is used to ground the generated results with your data. Select an existing data source or create a new data connection with Azure Blob storage, databases, search, URLs, or local files as the source the grounding data will be built from. [Learn more about data privacy and security in Azure AI.](#)

**Select data source \***

Azure AI Search  
Azure AI Search (selected)  
Azure Blob Storage (preview)  
Azure Cosmos DB for MongoDB vCore  
URL/web address (preview)  
Upload files (preview)

**Index data field mapping**

Some content and data fields from your index will be mapped using the default mapping in order to ground the model on your data and to display document information. To specify how those fields are mapped, customize your mapping.

Use custom field mapping

I acknowledge that connecting to an Azure AI Search account will incur usage to my account. \* [View Pricing](#)

**Next** **Cancel**

**Add data**

Data source  
 Data management  
 Review and finish

### Data management

Set up specific configurations for your data and how the model will respond to requests. [Learn more about data privacy and security in Azure AI.](#)

**Search type** \*

Add an existing semantic search configuration

I acknowledge that using semantic search will incur usage to my Azure AI Search account.\* [View Pricing](#)

Enable document-level access control

**Document-level access control**

**Permitted groups** \*

**Back** **Next** **Cancel**

# Chat with your data, see citations, tailor the chat experience

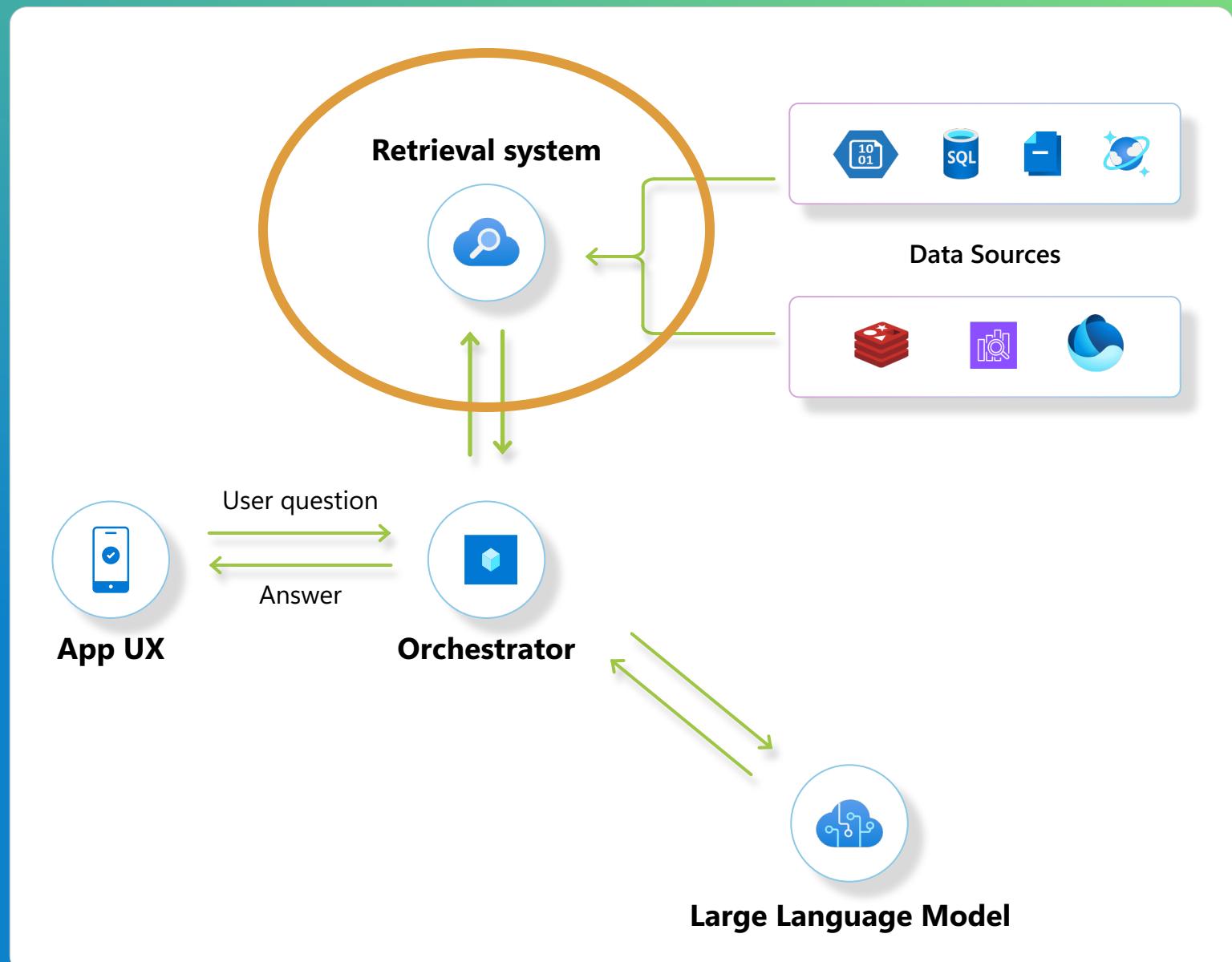
The screenshot shows the Azure AI Studio interface for the Chat playground. On the left, a sidebar lists various features: Azure OpenAI, Playground, Chat (selected), Completions, DALL-E (Preview), Management, Deployments, Models, Data files, Quotas, Plugins (Preview), and Content filters (Preview). The main area has a header with 'Azure AI Studio > Chat playground' and 'Privacy & cookies'. It includes buttons for 'Import setup', 'Export setup', and 'Show panels'. A 'Deploy to' dropdown is also present.

The central part of the screen displays the 'Assistant setup' dialog, which is active. It shows a 'Prompt' section and an 'Add your data (preview)' section. The 'Add your data' section includes a note about secure storage in Azure and links to learn more about data protection. It shows a 'Data source' configuration for 'Search Resource: Azure AI Search' and 'Index: productinfo'. Under 'Advanced settings', there is a checked checkbox for 'Limit responses to your data content' with a 'Strictness (1-5)' slider set to 3, and another slider for 'Retrieved documents (3-20)' set to 5. A 'Remove data source' button is at the bottom of this dialog.

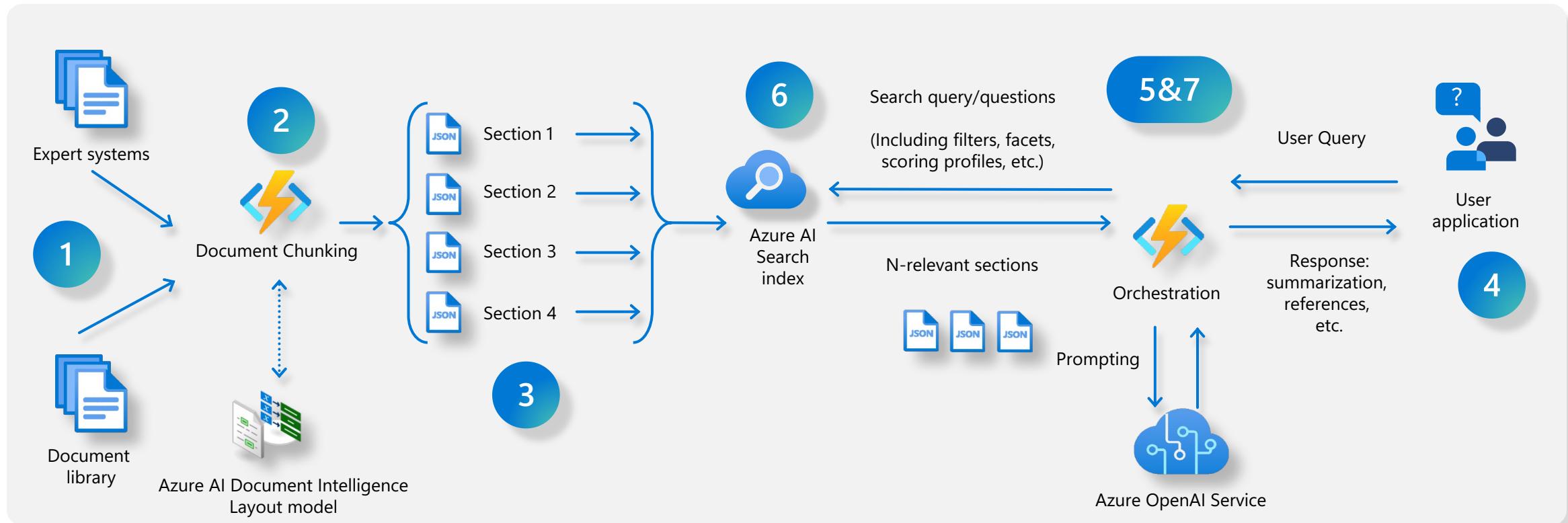
To the right of the setup dialog is a large panel showing a chat session. The user has typed 'what products do you have?'. The AI response is: 'Based on the retrieved documents, we have several hiking products from different brands, including the RainGuard Hiking Jacket<sup>1</sup>, TrekStar Hiking Sandals<sup>2</sup> <sup>3</sup>, TrekReady Hiking Boots<sup>4</sup>, and TrailWalker Hiking Shoes<sup>5</sup>.'. Below this, a list of 5 references is shown, each with a numbered link: 1 product\_info\_17.md - Part 1, 2 product\_info\_18.md - Part 1, 3 product\_info\_18.md - Part 1, 4 product\_info\_4.md - Part 1, 5 product\_info\_11.md - Part 1. A message at the bottom says 'New chat session started' and 'The assistant setup has been updated. Previous messages won't be used as context for new queries.'

On the far right, a 'Configuration' sidebar is open, showing the 'Deployment' tab selected. It includes fields for 'Deployment' (set to 'cluGPTTurbo') and 'Session settings' (with a slider for 'Past messages included' set to 10, 'Current token count' at 10, and an 'Input tokens progress indicator' at 11/4000).

# RETRIEVAL-AUGMENTED GENERATION



# Anatomy of RAG



## 1. Data ingestion

Different data formats and system of records

## 2. Chunking

What is the best Chunking strategy?

## 3. Indexing

Shall I use vector embeddings data transformation, mappings?

## 4. User interface

Chatbot for Q&A surfaced to end users

## 5. Orchestration

Communication coordination and prompting—Prompt to get retriever query

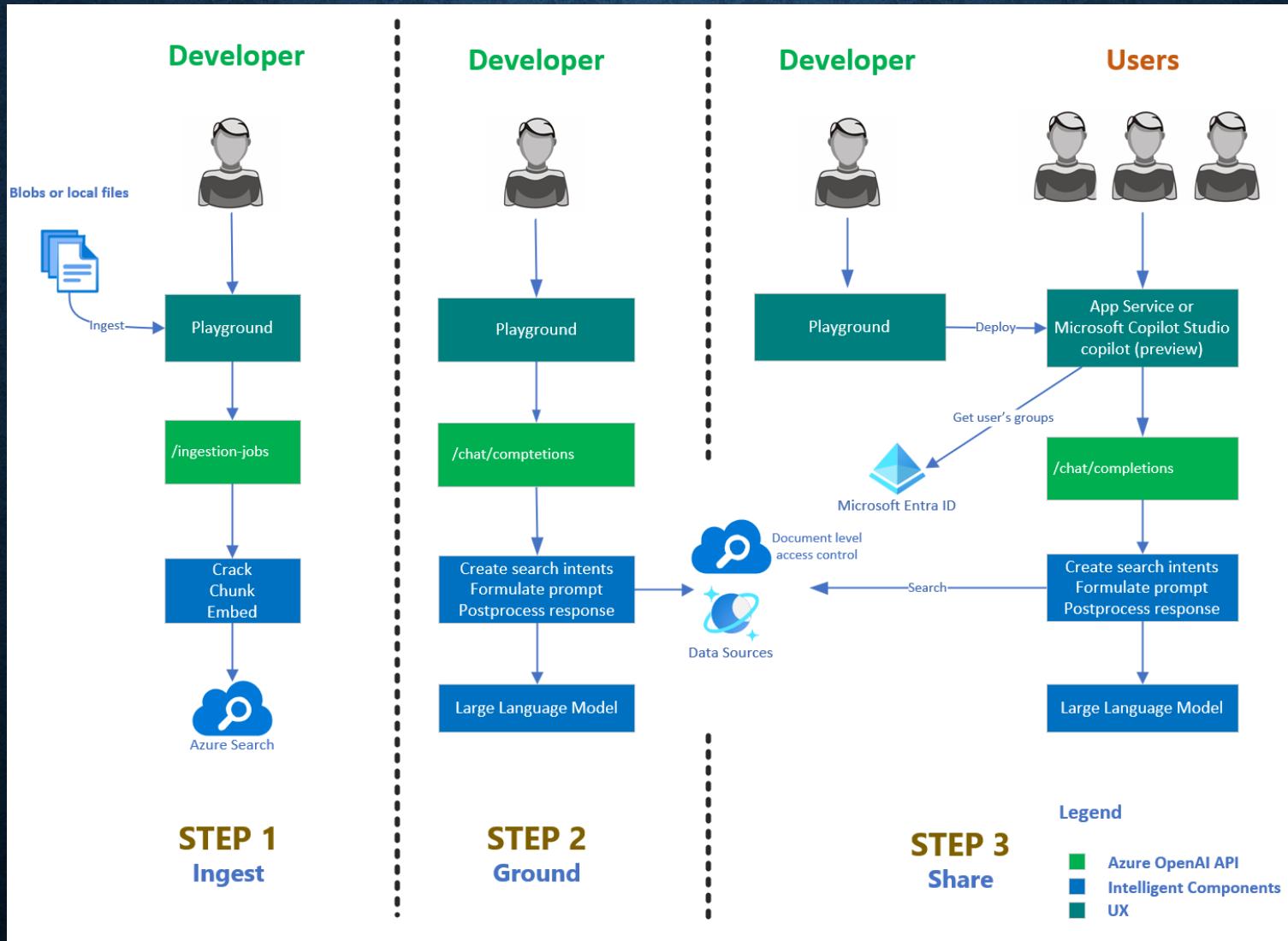
## 6. Data retrieving

Shall I use vector, semantic, keyword or hybrid approach?

## 7. Orchestration

Communication coordination: create user response based on retrieve data and send to User app

# AZURE OPENAI ON YOUR DATA ARCHITECTURE



# Retrieval strategies



## Keyword search

- **For exact, plain text matches**
- “Vocabulary gap” in Q&A systems like Copilot



## Vector search

- **For conceptual similarity, or underlying meaning**
- Weak performance on exact matches (like a product ID or code)



## Hybrid search

- **Best of both vectors and keywords**
- Brings more accurate responses across various scenarios



## Search re-ranking

- **Scores and ranks all retrieved documents by relevance**
- Reranking runs after performing search strategy (can't retrieve information)

# Integrated Vectorization

End-to-end data ingestion, chunking, vectorization, and advanced retrieval



## Chunking

- Built-in Chunking skill (updates to [split skill](#)) and updates to index to manage Chunks vs. full documents
- Configure Chunking parameters (e.g., **pages**, overlap window, etc.)
- Automated via Indexer orchestration



## Vectorization

- Bring-your-own Azure OpenAI endpoint
- Bring-your-own Azure AI Studio model catalog embedding model (Preview)
- Use AI Vision multimodal embeddings model deployed in AI multi-service account (Preview)
- Support for other embedding model REST endpoints
- Query Vectorization capability

