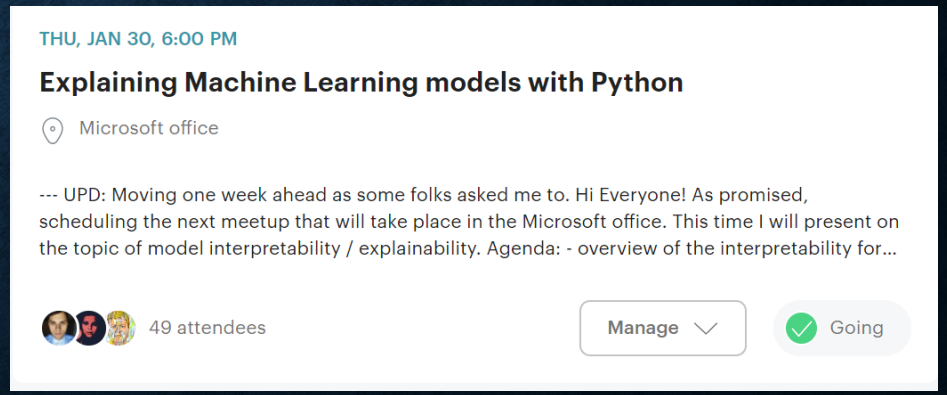
SIMLE AI DEMO TOOLKIT GRADIO + AOAI OYD

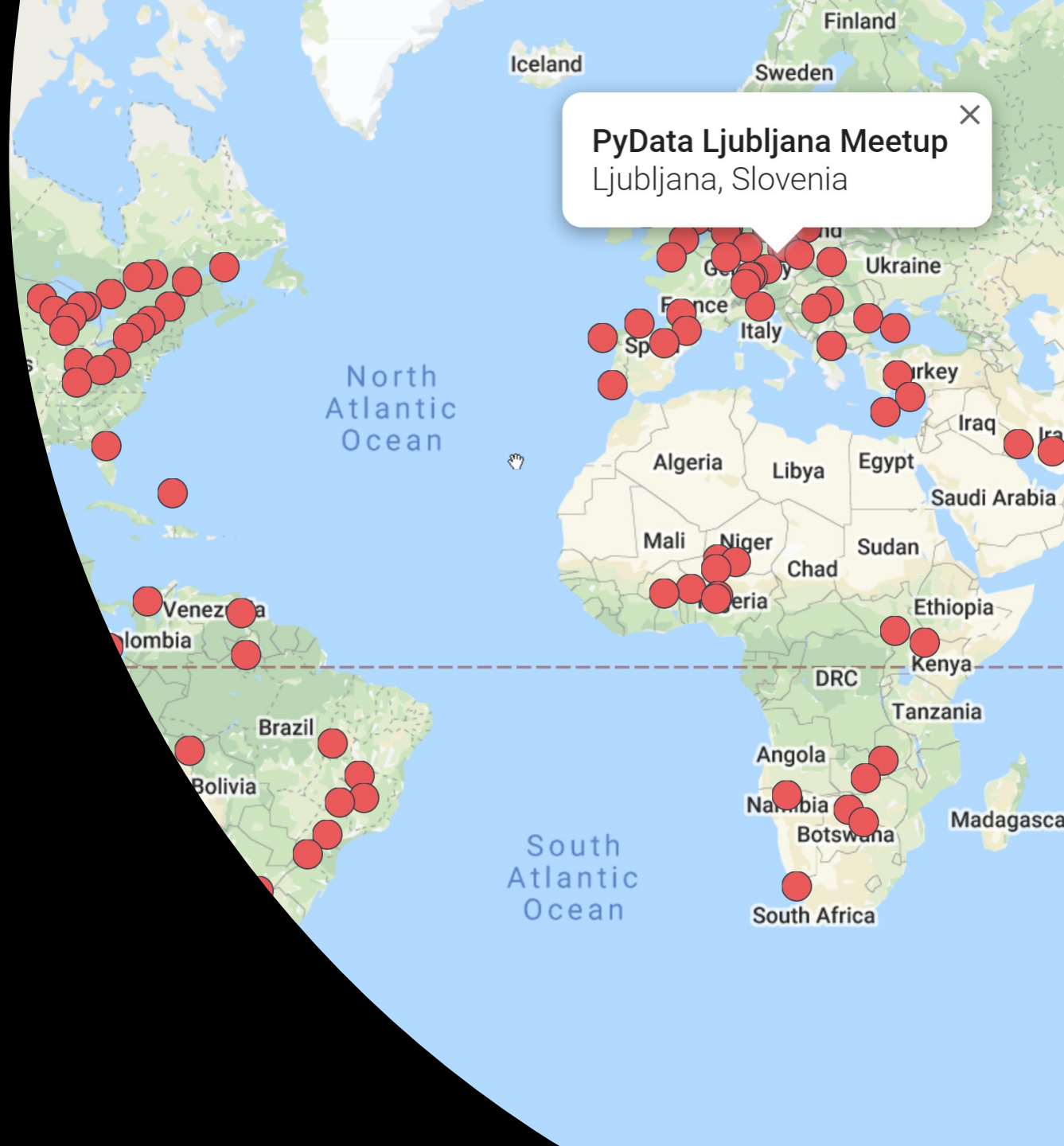
PyData meetup Nov 24

Andrey Vykhodtsev (AI Dev @ Aprova GmbH.)



QUICK INTRO INTO PYDATA

- Numbers
  - 61 countries
  - 164 groups
  - 156065 members
- By NumFocus.org
  https://numfocus.org/programs/pydata
- Since 2017?
- 13th 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

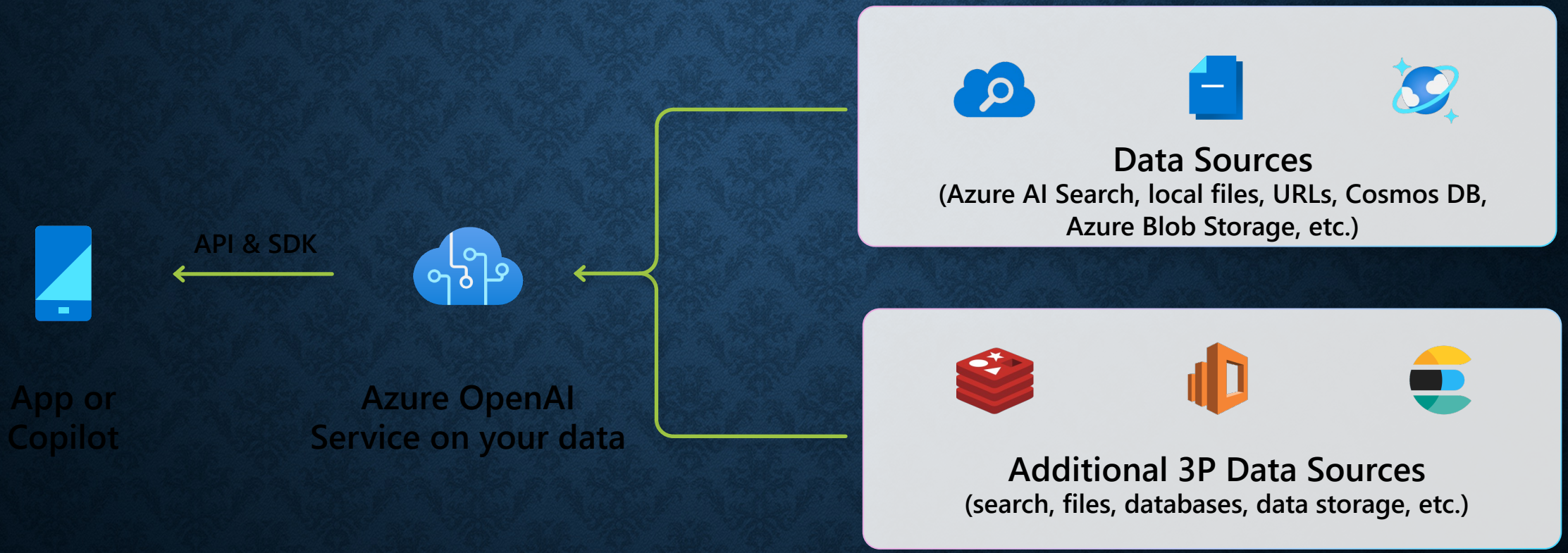
Retrieval augmented generation made easy

Restrict responses to your data

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

Easy deployment as a copilot

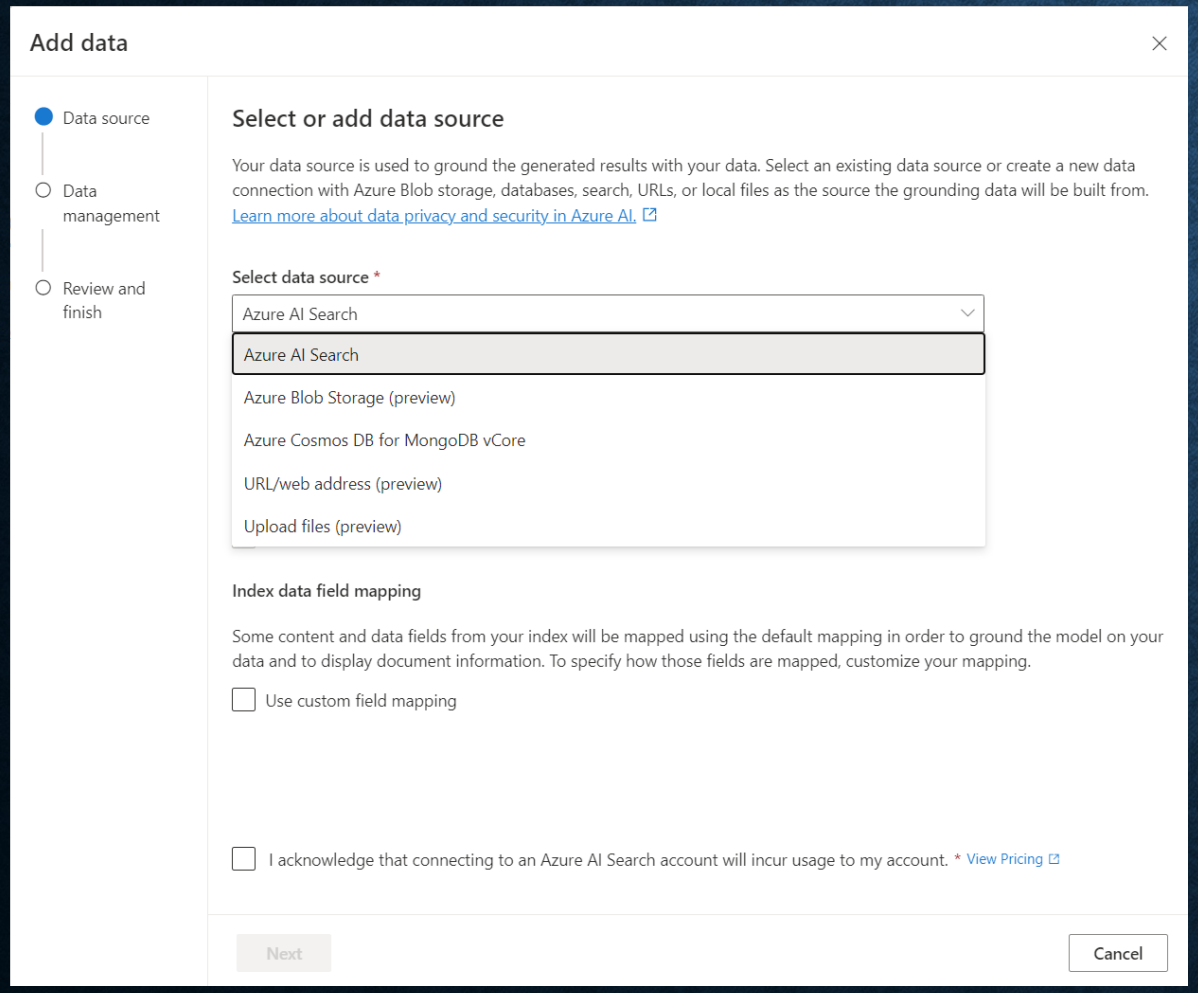
Azure OpenAI on *your* data

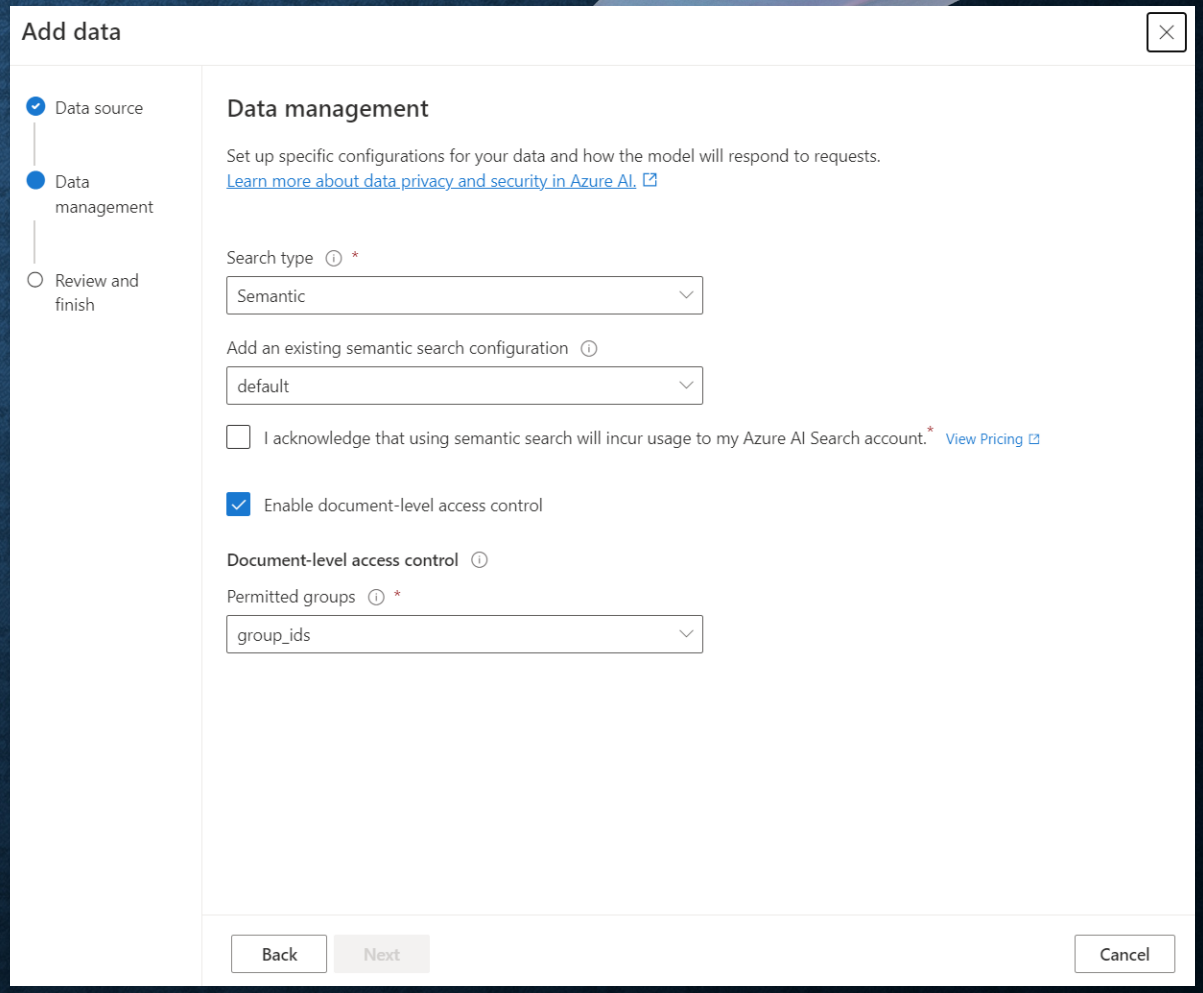


ON YOUR DATA – USE CASES

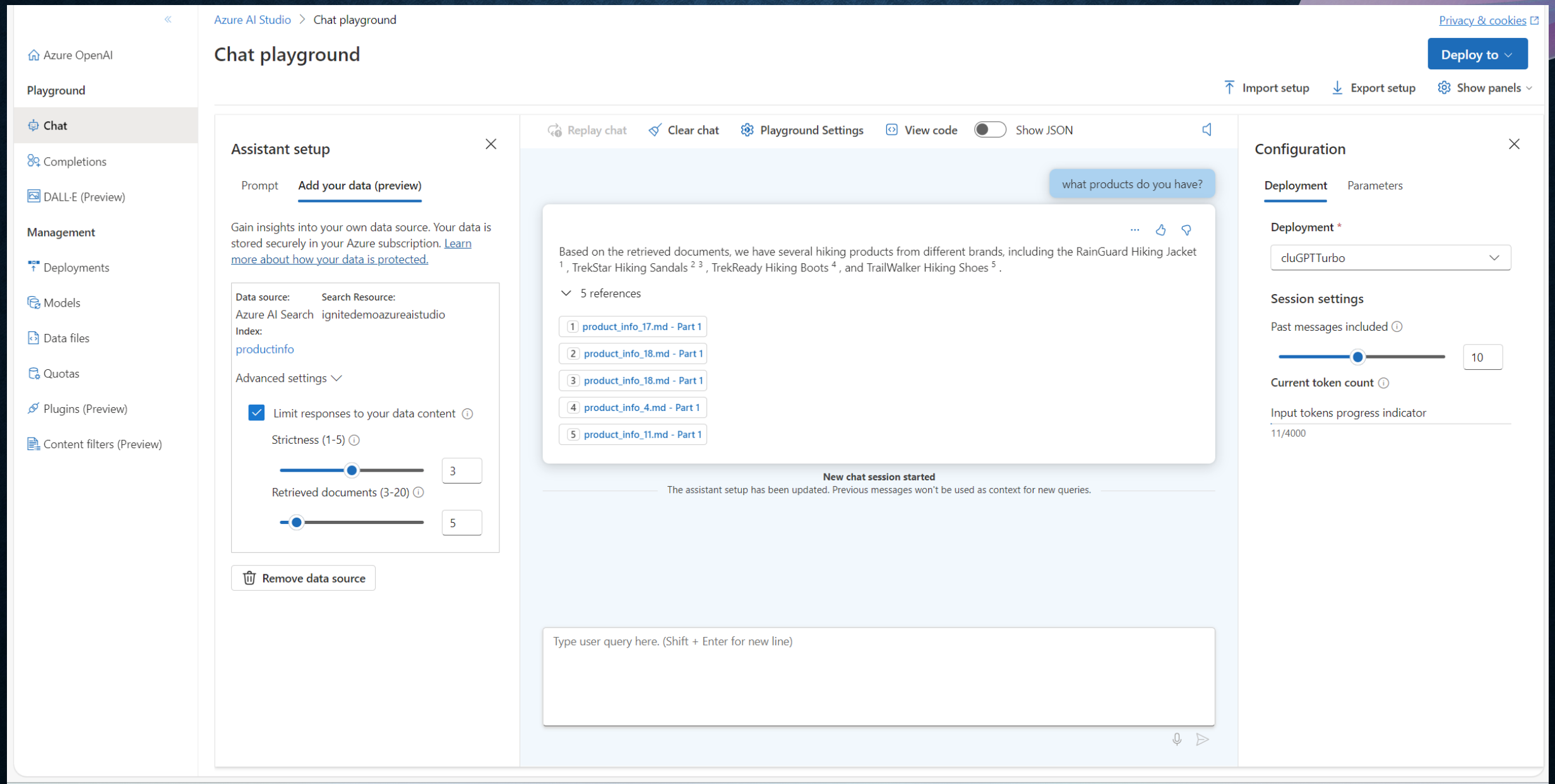
Use Cases	Details
Automated Customer Assistance	Provide quick responses to frequently asked questions and guide users through common troubleshooting steps based on the customers' data.
Real-time Document Search	Provide real-time support to product specifications and guide users through common troubleshooting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

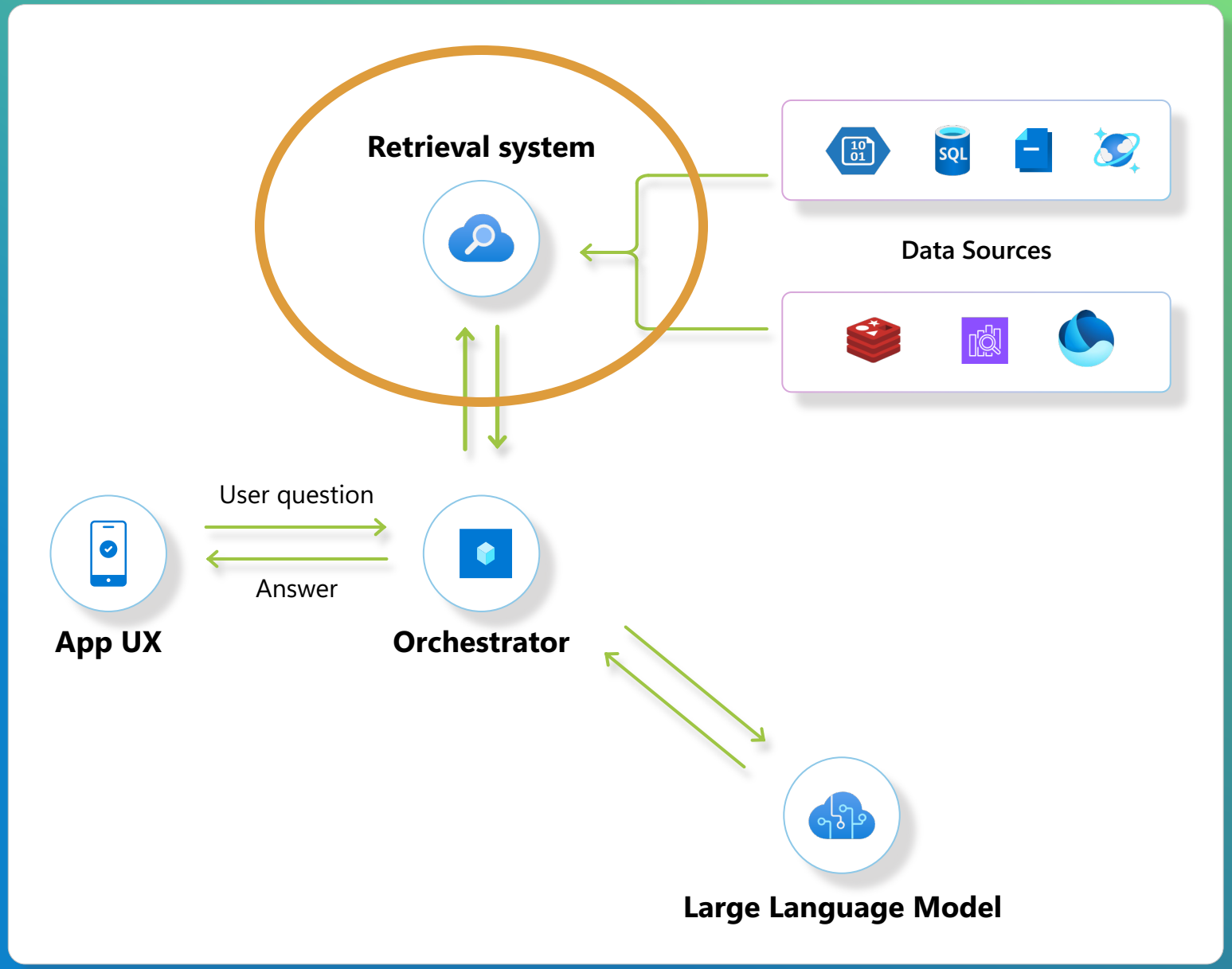




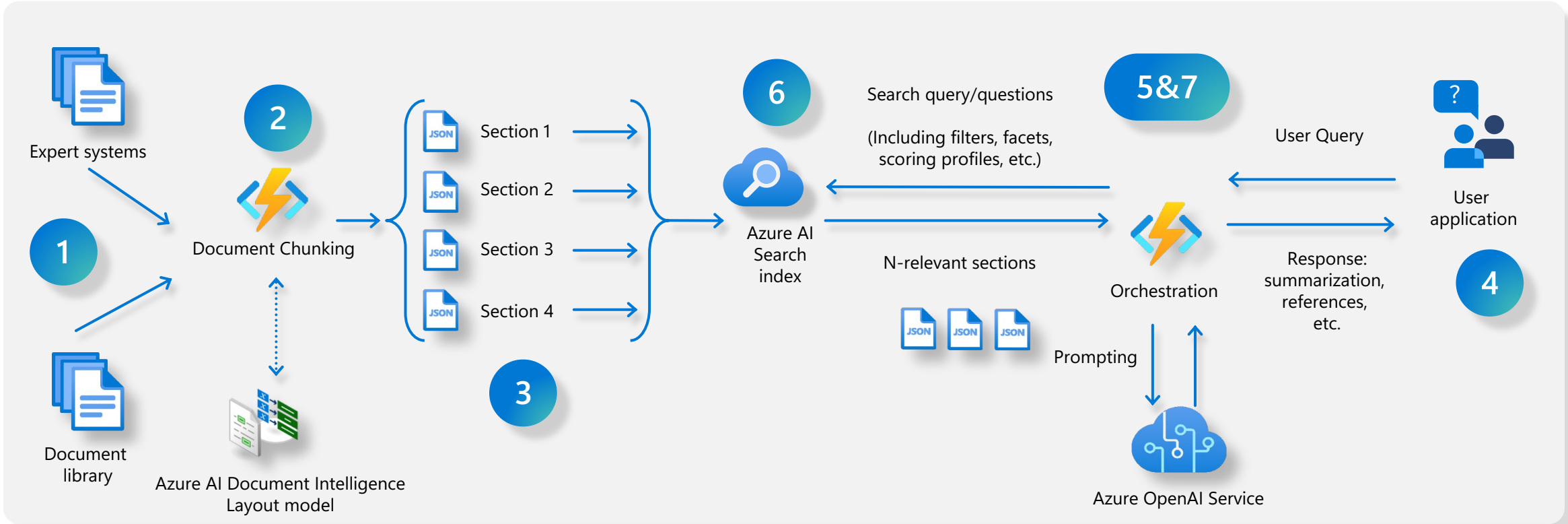
Chat with your data, see citations, tailor the chat experience



RETRIEVAL-AUGMENTED GENERATION

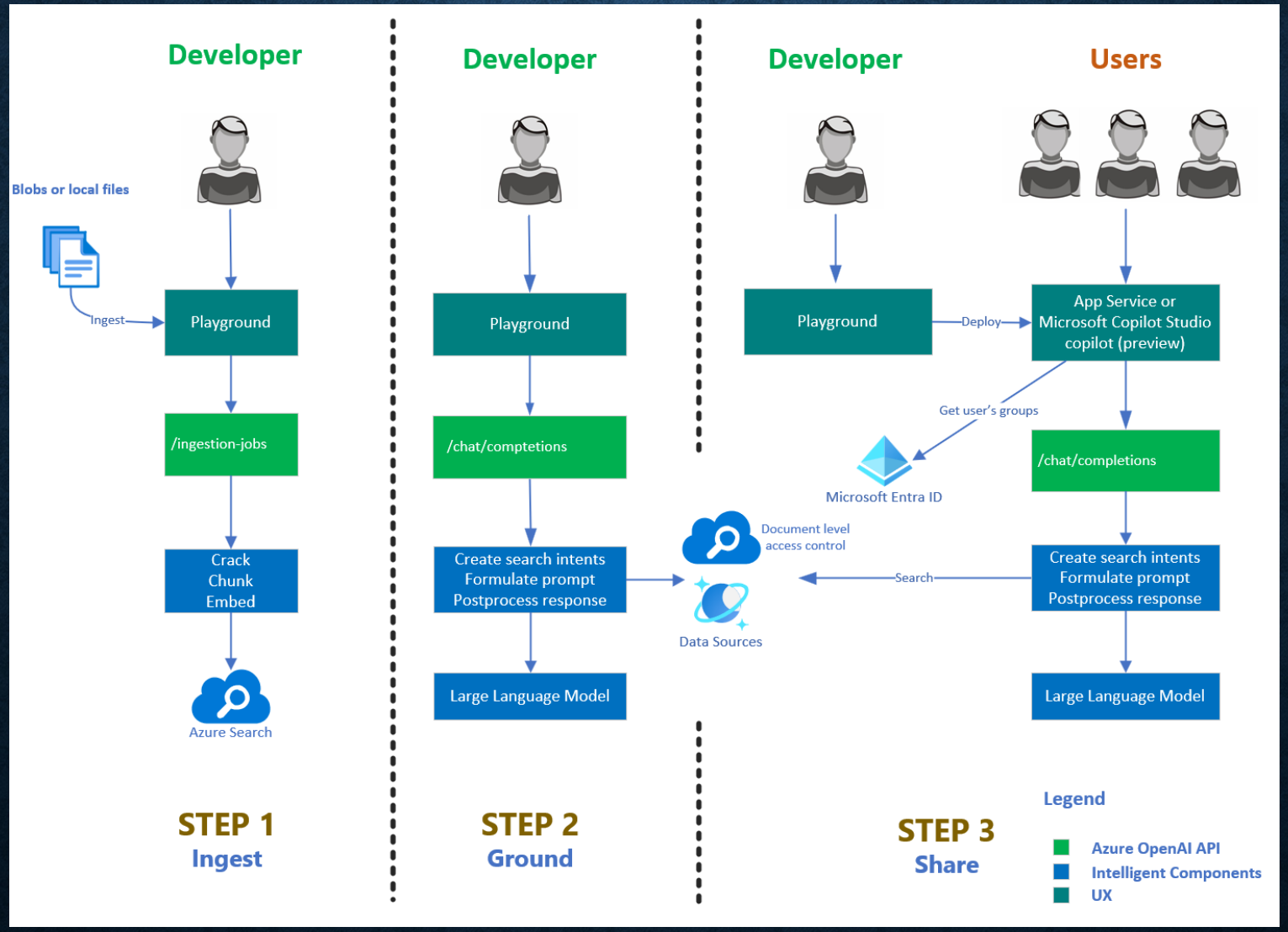


Anatomy of RAG



1. Data ingestion	2. Chunking	3. Indexing	4. User interface	5. Orchestration	6. Data retrieving	7. Orchestration
Different data formats and system of records	What is the best Chunking strategy?	Shall I use vector embeddings data transformation, mappings?	Chatbot for Q&A surfaced to end users	Communication coordination and prompting—Prompt to get retriever query	Shall I use vector, semantic, keyword or hybrid approach?	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

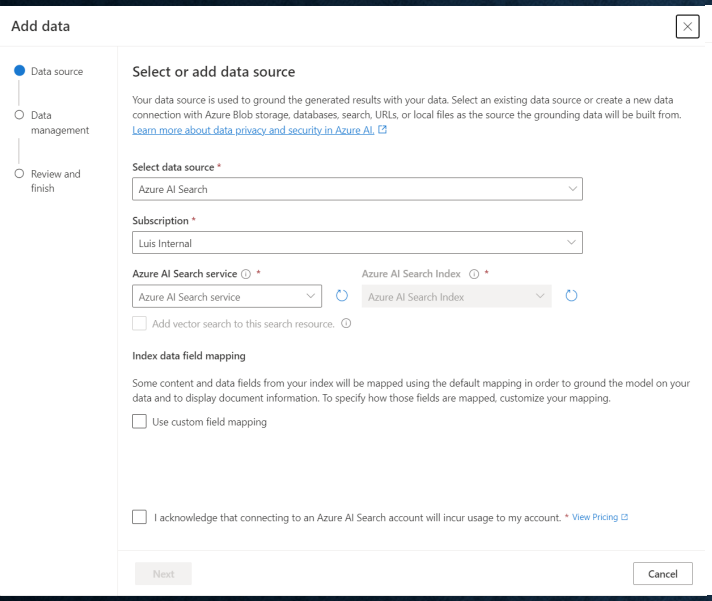
End-to-end data ingestion, chunking, vectorization, and advanced retrieval

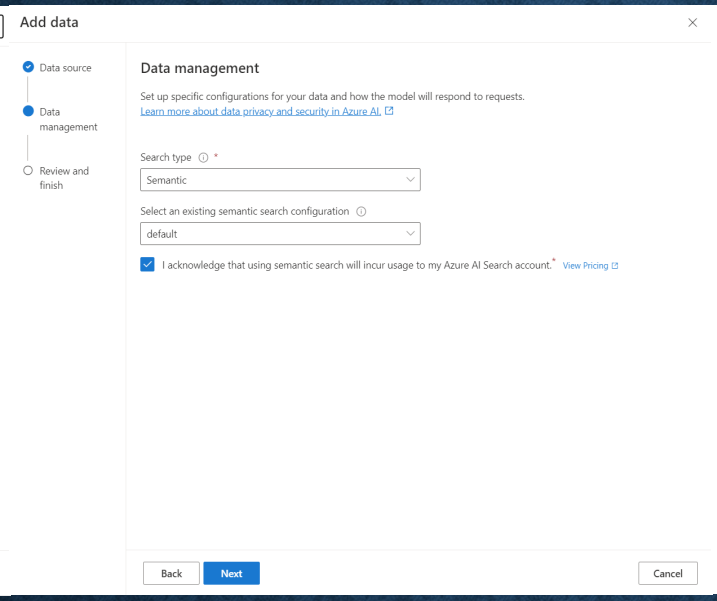
Chunking

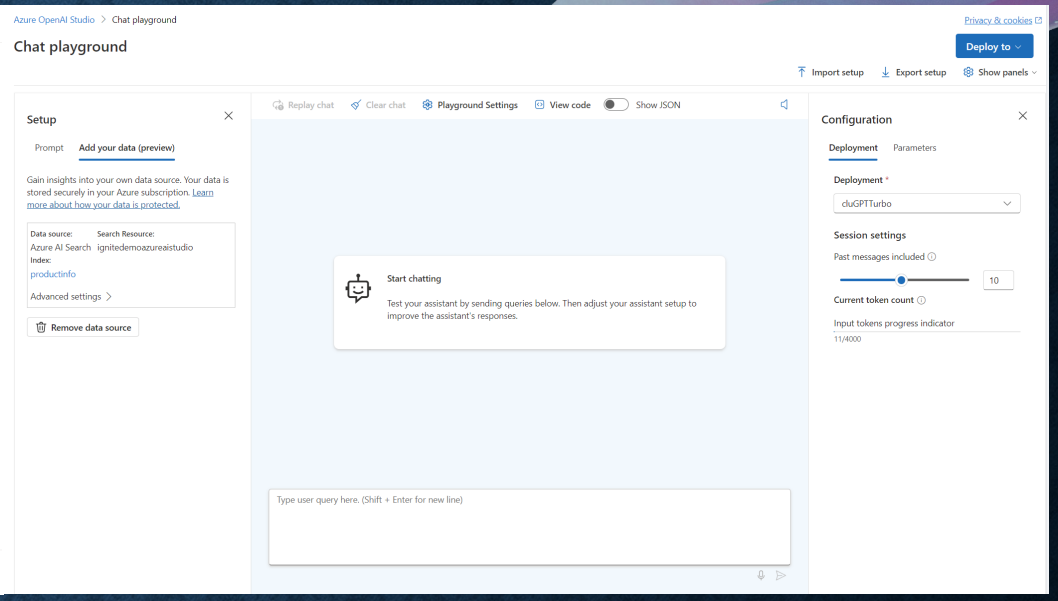
- Built-in Chunking skill (updates to split skill) and updates to index to manage Chunks vs. full documents
- Configure Chunking parameters (e.g., **pages,** overlap window, etc.)
- Automated via Indexer orchestration

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


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
Subscription *
Luis Internal
Azure AI Search service *
Azure AI Search service
Azure AI Search Index *
Azure AI Search Index
Add vector search to this search resource.
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 *
Semantic
Select an existing semantic search configuration
default
I acknowledge that using semantic search will incur usage to my Azure AI Search account.* View Pricing
Back
Next
Cancel


Azure OpenAI Studio
Chat playground
Privacy & cookies
Chat playground
Deploy to
Import setup
Export setup
Show panels
Setup
Prompt
Add your data (preview)
Gain insights into your own data source. Your data is stored securely in your Azure subscription. Learn more about how your data is protected.
Data source: Azure AI Search
Search Resource: ignitedemoazureaistudio
Index: productinfo
Advanced settings
Remove data source
Replay chat
Clear chat
Playground Settings
View code
Show JSON
Start chatting
Test your assistant by sending queries below. Then adjust your assistant setup to improve the assistant's responses.
Type user query here. (Shift + Enter for new line)
Configuration
Deployment
Parameters
Deployment *
cluGPTTurbo
Session settings
Past messages included
10
Current token count
Input tokens progress indicator
11/4000