



# Building an Enterprise-Ready Chatbot in One Day

Agnieszka Niezgoda  
Maciej Drwięga  
Tomasz Kopacz



# Key steps (easy to replicate for other scenarios)

- (\*)Choose YOUR sample to demonstrate
  - Good: to have CI/CD pipeline (or we can create one using GH Copilot)
- Create an **External** Azure Subscription and tenant
  - Add users
  - Add group "admin" and group "users"
  - Create RG for users
  - Turn off all "Microsoft policies"
  - Add **securityControls=ignore** tag on the resource group per user – this prevents SFI policies.
- Create GitHub Enterprise Trial
  - Assign Azure Subscription as a payment to GHE
  - Create and configure Enterprise (mandatory GHAS, enabled Copilot etc. ) and organization
  - Copy sample code (\*) + your markdown with scenarios
  - Turn on Copilot for all users
- For GH Copilot scenarios to update code – show **example prompts**

The aim of the workshop is to enable you to create applications that are:



modern,



safe



**tailored to the needs of  
large organizations**

as efficiently as possible

RAG is very easy – unless you want to do it right

[Clear chat](#)[Developer settings](#)

# Chat with your data

Ask anything or try an example

What is included in my  
Northwind Health Plus plan  
that is not in standard?

What happens in a  
performance review?

What does a Product  
Manager do?

Type a new question (e.g. does my plan cover annual eye exams?)



In the first hour, we will implement a working chatbot as our starting point.  
We'll also discuss how you can accelerate your organization's application development

# Join us



Agnieszka  
Niezgoda



Maciej Drwięga

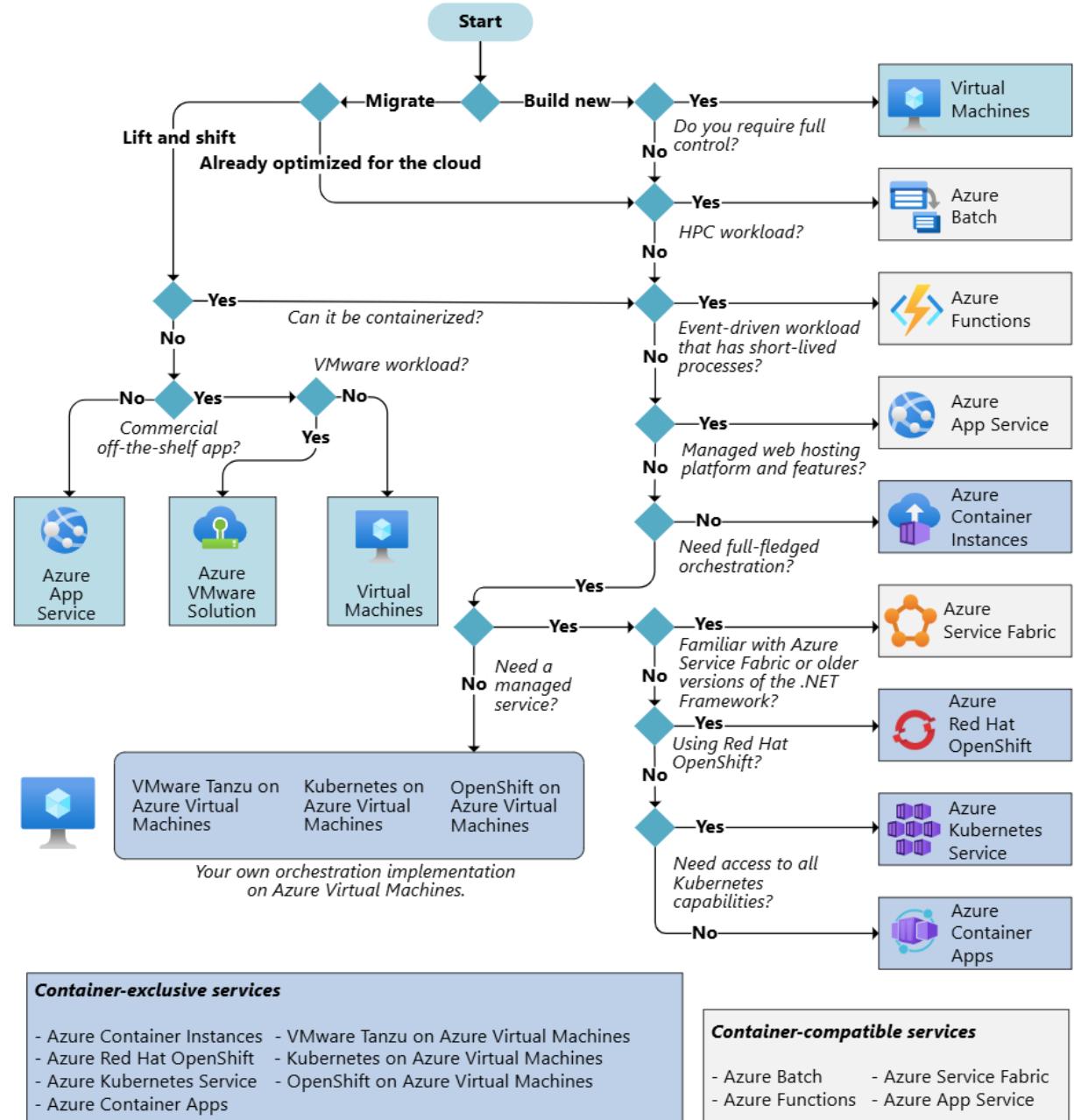


Tomasz Kopacz

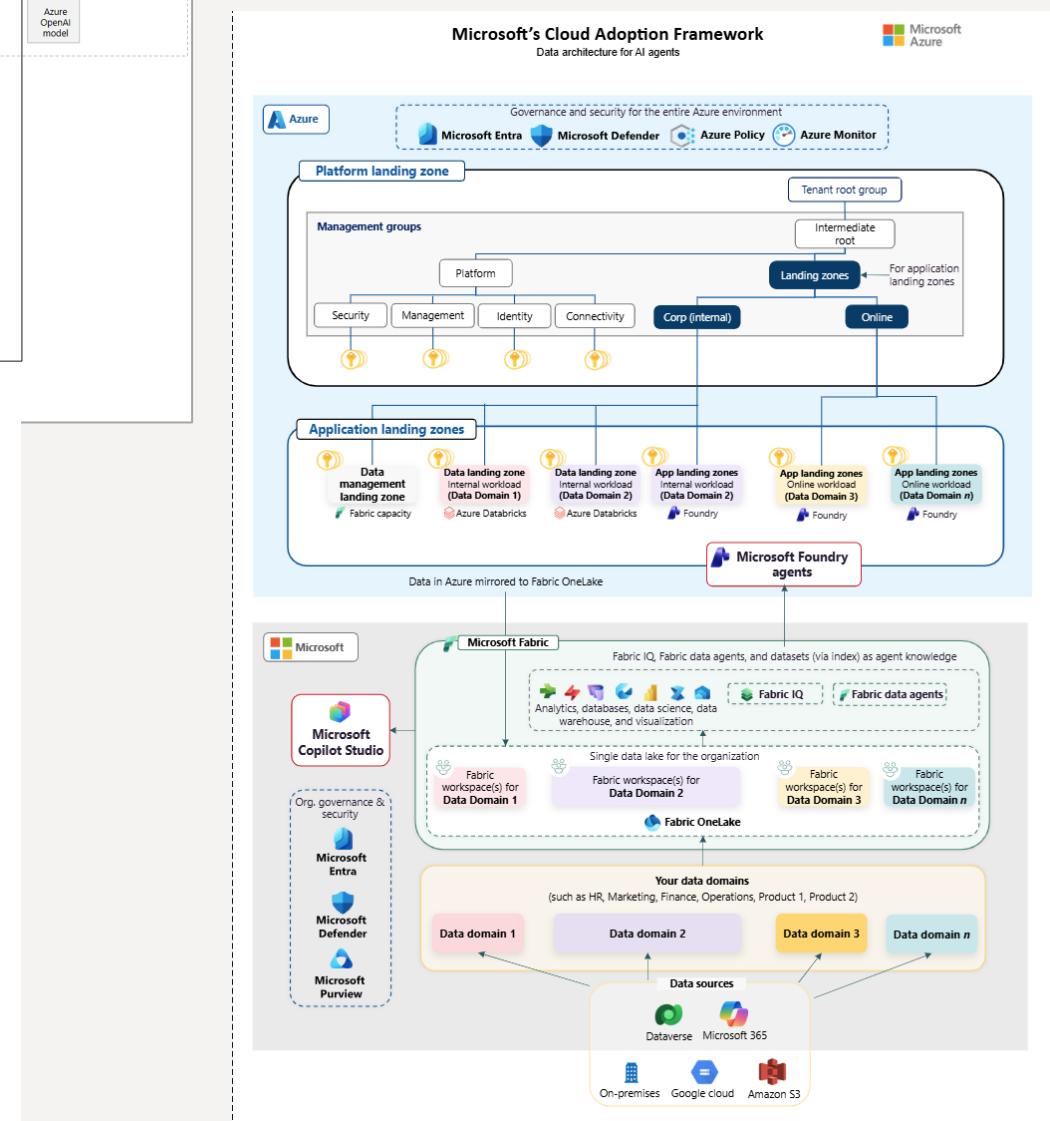
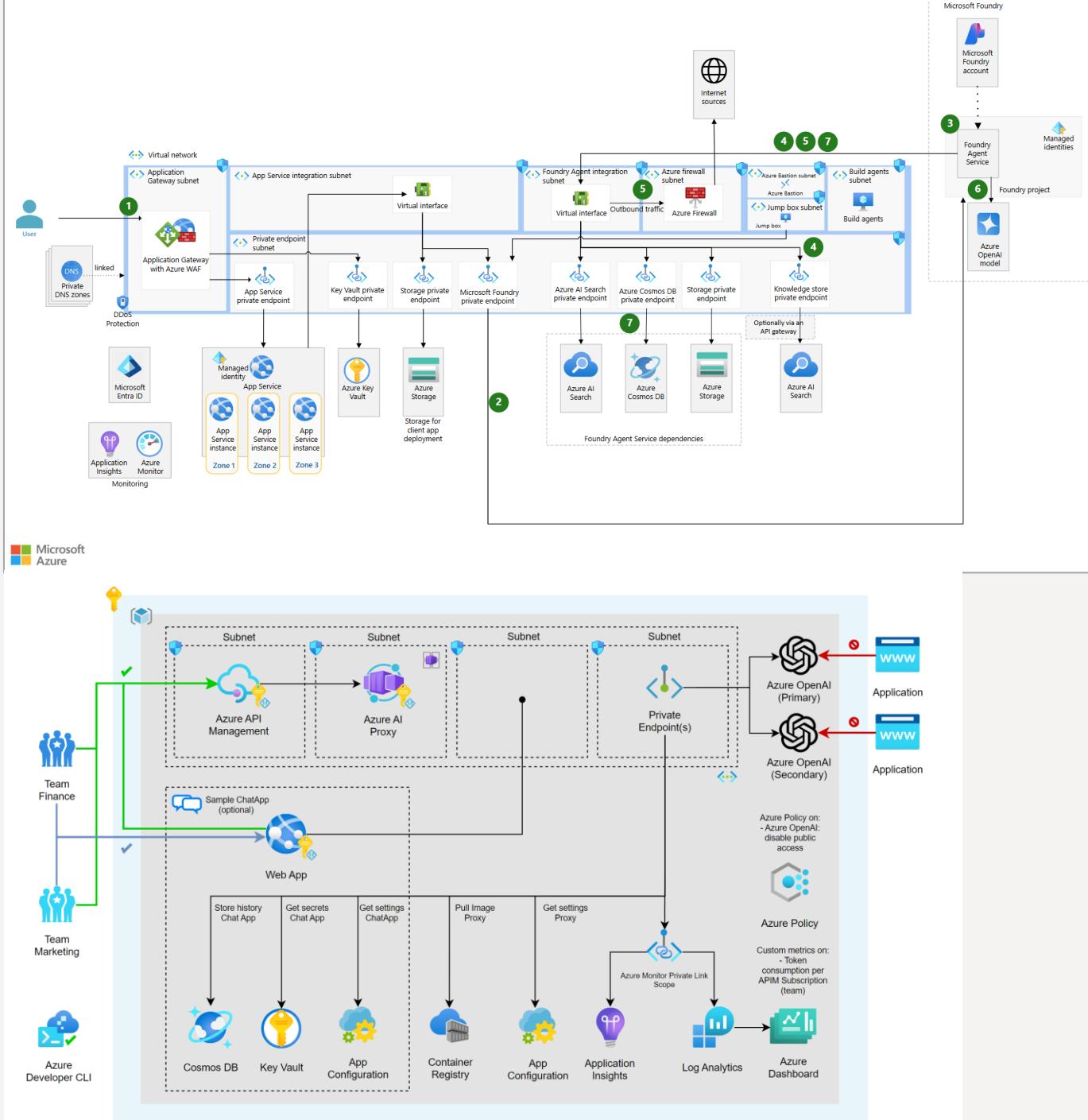
# 1. Introduction. Applications on Azure

30 minutes

**On Azure, you have several different application services to choose from**



# Common application architectures on Azure



# „Czym się aplikacja enterprise różni od hello world”

## Reliability & Fault Tolerance

*Assume everything can fail;  
Fault-tolerant architecture;  
Layered recovery approach;  
Quantitative reliability modelling.*

## Scalability & Performance

*Elastic scaling;  
Use managed services (PaaS);  
Performance tuning & measurement;  
Business metrics and tenant-health signals.*

## Security & Governance

*Enterprise-grade security and governance frameworks;  
RBAC, data governance, auditing, compliance;  
Guardrails for trustworthy operations*

## Observability & Operational Excellence

*Comprehensive logging, tracing, and telemetry;  
Automated operations;  
Telemetry-driven health scoring*

## Data Management & Integrity

*Multi-region data replication strategies;  
Clear data models;  
Secure data access patterns.*

## Maintainability & Modularity

*Well-structured, modular architecture;  
Evolvability;  
Use of standardized services and frameworks.*

# Foundry i API Management w architekturach AI

## API Management:

„AI Gateway” de facto

Kontrola serwerów MCP / A2A

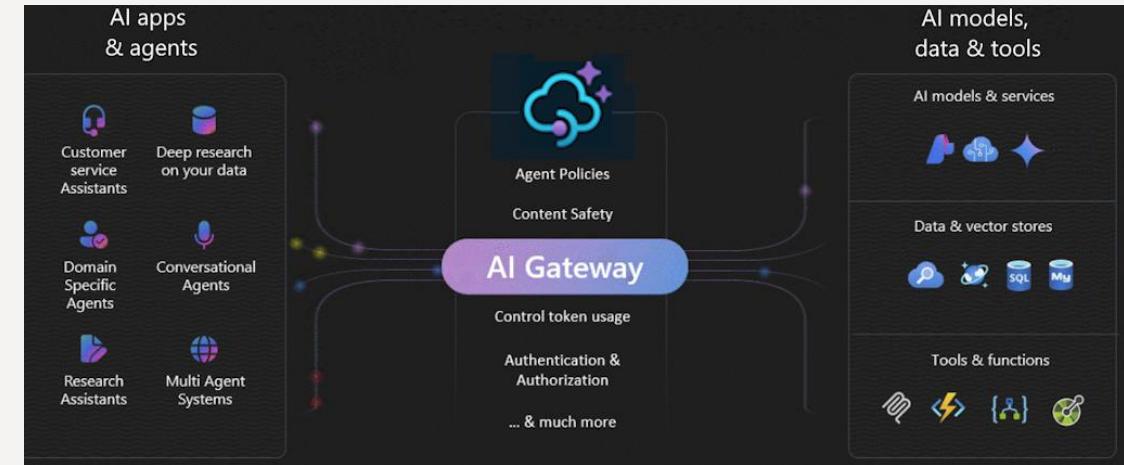
Tłumaczenie na API zgodne z „OpenAI”

Rozkład obciążenia (LB, round-robin)

Centralny widok na używane tokeny

### Semantic Caching

(throttling, autoryzacja, logowanie, ...)



## AI Foundry:

Źródło modeli (GPT-x, LLAMAx, Claude)

Finetuning

Monitoring

Też: platforma do Agentów (technicznie REST API)

Tu (można, ale niewiele doda); też Playground

Historia: AOI, AI Project, AI Foundry

```
azd env set USE_AI_PROJECT=true  
azd provision
```

# Agenda of today's session

1. We start with the implementation of the project with Agnieszka's commentary
2. GitHub: GitHub Actions and Copilot
  1. (Comments on azd and how it works)
  2. Configuration of automatic deployment
  3. Enabling mechanisms to control the quality and security of the code
  4. How Copilot helps you understand code
  5. MCP server configuration for GitHub and ordering Coding Agent to build an alternative UI (e.g. Blazor)
3. Then – new functionalities (modifications to the implemented project)
  1. Changing the chunk and search parameters
  2. Adding Web and Sharepoint source search in addition to AI Search index
  3. Adding a validator
  4. Additional activity logging
  5. Authorization using Entra
4. Finally, remarks on infrastructure and hard security

## 2. Accelerator: functionalities and implementation

1 h

# RAG is used especially in on-your-own-data scenarios

Azure OpenAI + AI Search Chat Ask a question Login

Clear chat + Manage file uploads Developer settings

Does the Northwind Health Plus plan cover eye exams?

Yes, the Northwind Health Plus plan covers vision exams <sup>1</sup>.  
Citations: 1. Benefit\_Options.pdf#page=3

Hearing too?

Yes, the Northwind Health Plus plan covers hearing care services, including hearing tests, evaluations, hearing aids, fittings, adjustments, repairs, and replacements <sup>1</sup>. The plan covers up to \$1,500 every 3 years for hearing aid services, including the cost of the hearing aid itself <sup>1</sup>. However, the plan does not cover the cost of custom ear molds for hearing aids or over-the-counter hearing aids or devices <sup>1</sup>. The coverage is limited to individuals 18 years of age or older <sup>1</sup>.  
Citations: 1. Northwind\_Health\_Plus\_Benefits\_Details.pdf#page=30  
Follow-up questions: What is the process for reimbursement of hearing care expenses?  
Are there any limitations on the number of hearing aid fittings or adjustments covered?  
Are there any restrictions on the providers that can be used for hearing care services?

Type a new question (e.g. does my plan cover annual eye exams?) >

Azure OpenAI + AI Search Login

Clear chat + Manage file uploads Developer settings

Chat with your data

Ask anything or try an example

What is included in my Northwind Health Plus plan that is not in standard?

What happens in a performance review?

Type a new question (e.g. does my plan cover annual eye exams?) >

Azure OpenAI + AI Search Clear chat

Configure answer generation

Override prompt template (1)

Temperature (1)  
0.3

Seed (1)

Minimum search score (1)  
0

Retrieve this many search results: (1)  
3

Exclude category (1)

Suggest follow-up questions (1)

Retrieval mode (1)  
Vectors + Text (Hybrid)

Use oid security filter (1)

Use groups security filter (1)

Stream chat completion responses (1)

ID Token Claims

# Accelerator implementation

1. Repos for the organization:
  - <https://GitHub.com/MH16-App-deployment-on-Azure/base-app>
  - 00TKScenarios\00-GitHub-Start-Codespaces\00-GitHub-Start-CodeSpaces.md
2. Users chat-user-yz | rg-azureclubworkshopchat-yz | chat-user-yz\_mhent
3. Fork from your GH Enterprise
  - Our case: <https://GitHub.com/MH16-App-deployment-on-Azure/base-app>
  - (base) <https://GitHub.com/Azure-Samples/azure-search-openai-demo>
4. Open the forked repo in CodeSpace. We'll be working in CodeSpace – it's a bit more complex to configure the app locally
5. We are using:
  - Domain: [domain name]
  - TenantID: [...]
  - Subscription ID: [...]
6. Environment name and RG: chat-[first letters of your name and surname e.g. AN]
7. Follow the README file to deploy the app – it should take around 20 minutes
  - Helper: 00TKScenarios\00-GitHub-Start-Codespaces\00-GitHub-Start-CodeSpaces.md  
00TKScenarios\01-GitHubActions\01-GitHubActions.md
  - Important: **azd env set AZURE\_RESOURCE\_GROUP pl-tmp1**

# RAG is used especially in on-your-own-data scenarios

Azure OpenAI + AI Search Chat Ask a question Login

Clear chat + Manage file uploads Developer settings

Does the Northwind Health Plus plan cover eye exams?

Yes, the Northwind Health Plus plan covers vision exams <sup>1</sup>.  
Citations: 1. Benefit\_Options.pdf#page=3

Hearing too?

Yes, the Northwind Health Plus plan covers hearing care services, including hearing tests, evaluations, hearing aids, fittings, adjustments, repairs, and replacements <sup>1</sup>. The plan covers up to \$1,500 every 3 years for hearing aid services, including the cost of the hearing aid itself <sup>1</sup>. However, the plan does not cover the cost of custom ear molds for hearing aids or over-the-counter hearing aids or devices <sup>1</sup>. The coverage is limited to individuals 18 years of age or older <sup>1</sup>.  
Citations: 1. Northwind\_Health\_Plus\_Benefits\_Details.pdf#page=30  
Follow-up questions: What is the process for reimbursement of hearing care expenses?  
Are there any limitations on the number of hearing aid fittings or adjustments covered?  
Are there any restrictions on the providers that can be used for hearing care services?

Type a new question (e.g. does my plan cover annual eye exams?) >

Azure OpenAI + AI Search Login

Clear chat + Manage file uploads Developer settings

Chat with your data

Ask anything or try an example

What is included in my Northwind Health Plus plan that is not in standard?

What happens in a performance review?

Type a new question (e.g. does my plan cover annual eye exams?) >

Azure OpenAI + AI Search Clear chat

Configure answer generation

Override prompt template  

Temperature 0.3

Seed  

Minimum search score 0

Retrieve this many search results: 3

Exclude category  

Suggest follow-up questions  

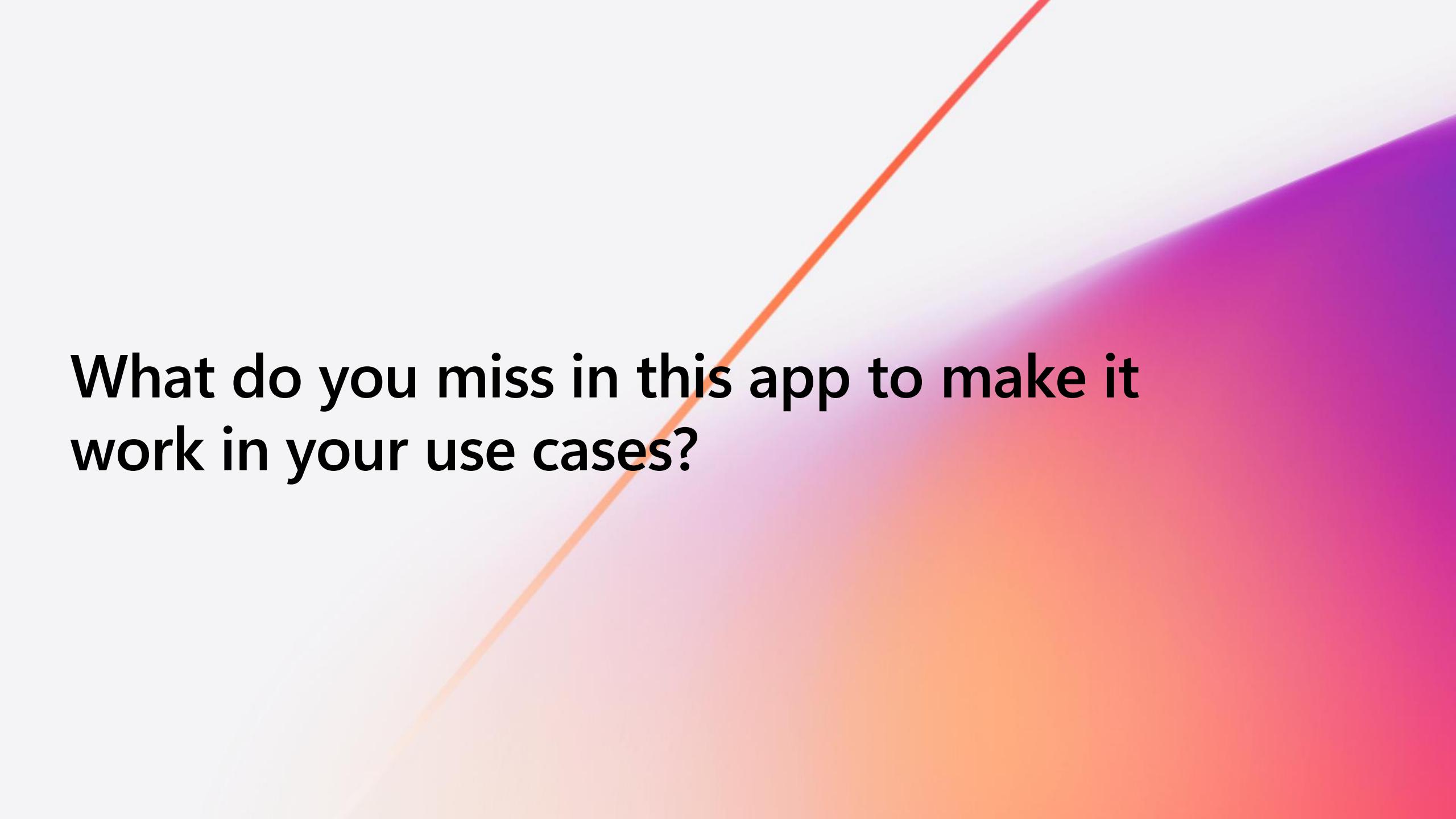
Retrieval mode Vectors + Text (Hybrid)

Use oid security filter  

Use groups security filter  

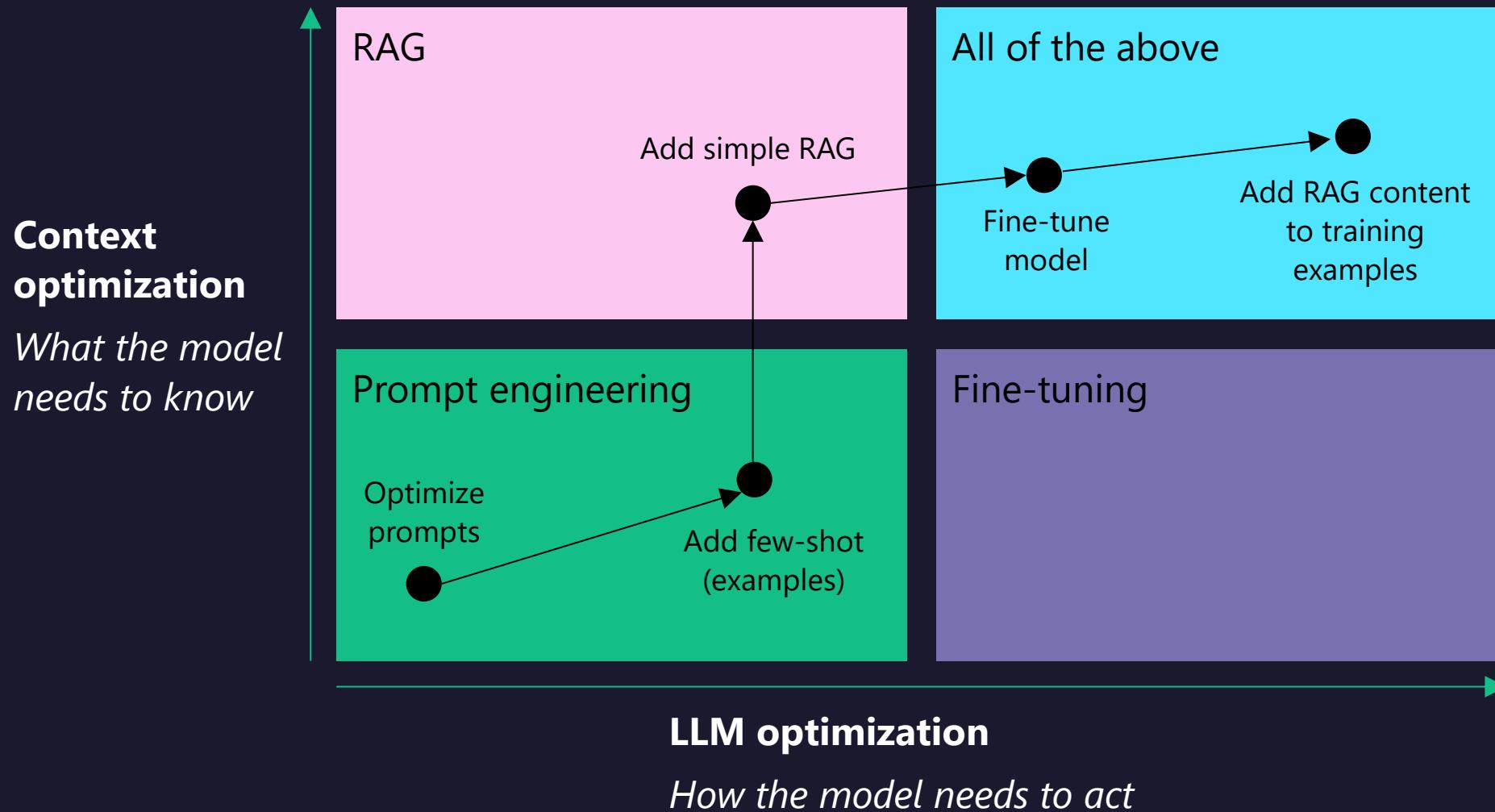
Stream chat completion responses  

ID Token Claims



**What do you miss in this app to make it work in your use cases?**

# The customization approach should depend on your goal



# Problem: generative AI doesn't have context for your data

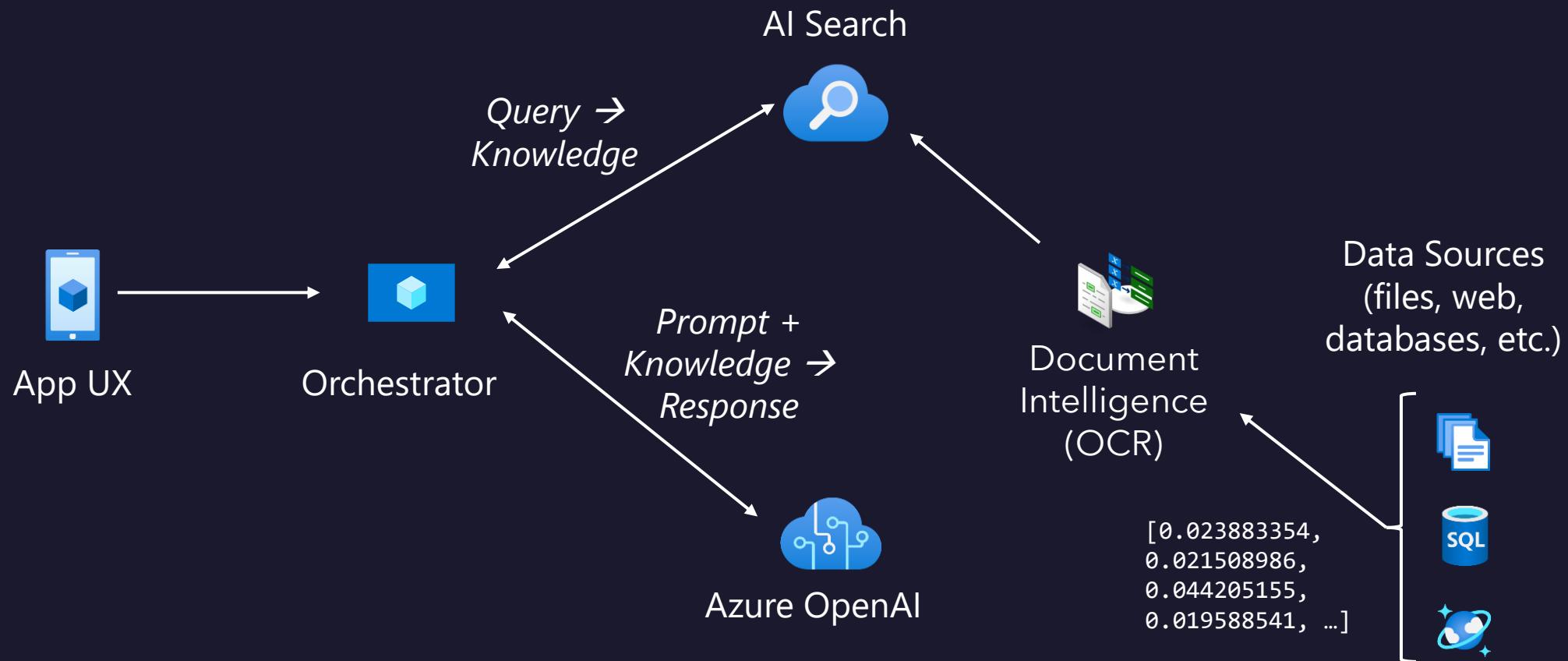
Prompt

Does my health plan  
cover annual eye  
exams?

Response

I'm an AI language model  
and don't have access to  
specific information  
about your health plan

# RAG (retrieval augmented generation) is used especially in on-your-own-data scenarios



# Typical RAG architecture – things to think about

- Frequently asked questions should be cached?

## Logging & monitoring

- User feedback (e.g. thumbs up/down)
- Where to keep the conversation flow
- Conversation statistics

- Chunking method
- Indexing method and model used
- The way of retrieval (agentic?, FoundryIQ? How many chunks?)
- Semantic reranker?



- What interface? Chatbot? Voice bot? Search engine? A bot that understands context?
- Possibility to add new documents by the user?
- Security and authentication

## Orchestrator

- Answers to questions about documentation
- Configuration

- Model Instructions: Style, To-Do's and Don'ts
- How much context should I keep?
- Validating the answer?
- Content filters and other security checks?
- Adding agents and access to MCP servers/functions?

## AI Search



Zapytanie do źródeł  
→ Wiedza

Prompt + wiedza ze źródeł → odpowiedź



## Azure OpenAI

## Application knowledge sources



- Pdfs, Word documents
- Charts, images
- Databases
- Web Resources

- What's the process of supplying sources – especially important when switching to production
- What preprocessing of sources is needed? E.g. creating diagram descriptions after adding them to sources?
- Examples of documents and pages

# Azure AI Search

- Platform-as-a-Service
- Semantic search
- Management free
- Keyword search
- Faceting
- Language analyzers
- Geospatial support
- Suggestions/auto-complete
- Customizable scoring
- Proximity search
- Synonyms
- Cognitive skills
- etc.

The screenshot shows the Microsoft Azure (Preview) interface for a 'Search service'. The service name is 'fsunavala-vector-demo'. The dashboard includes a left sidebar with navigation links like Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems, Search management, Indexes, Indexers, Data sources, Aliases, Skillsets, Debug sessions, Settings, Semantic search (Preview), Knowledge Center, Keys, Scale, Search traffic analytics, Identity, Networking, Properties, Locks, and Monitoring. The main content area displays resource details: Resource group (move) : fsunavala-sandbox, Location (move) : UK South, Subscription (move) : , Subscription ID : , Status : Running, and Tags (edit) : ProjectType : aoai-your-data-service. Below this, there are tabs for Get started, Properties, Usage, and Monitoring. A callout 'Build a full-text search experience with AI and sem...' is visible. The 'Get started' section features 'Connect your data' (with an 'Import' button) and 'Explore your data' (with a 'View' button). The URL for the service is https://fsunavala-vector-demo.azurewebsites.net.

Microsoft Azure (Preview) Report a bug Search resources, services, and docs (G+)

Home > **fsunavala-vector-demo** ★ ...

Search service

Search

+ Add index Import data Search explorer Refresh Delete Move

Would you like to get rewarded for sharing your experience with Azure Cognitive Search? [Write a review & get rewarded](#)

Resource group (move) : [fsunavala-sandbox](#) Url : <https://fsunavala-vector-demo.azurewebsites.net>

Location (move) : UK South Pricing tier : Standard

Subscription (move) : Replicas : 1 (No SLA)

Subscription ID : Partitions : 1

Status : Running Search units : 1

Tags (edit) : ProjectType : aoai-your-data-service

Get started Properties Usage Monitoring

Build a full-text search experience with AI and sem...

Get started building a full-text search experience and learn how to integrate w...

Connect your data Import

Explore your data View

Start here to import your data. Learn how to quickly connect to your data to build your first search index. [Learn more](#)

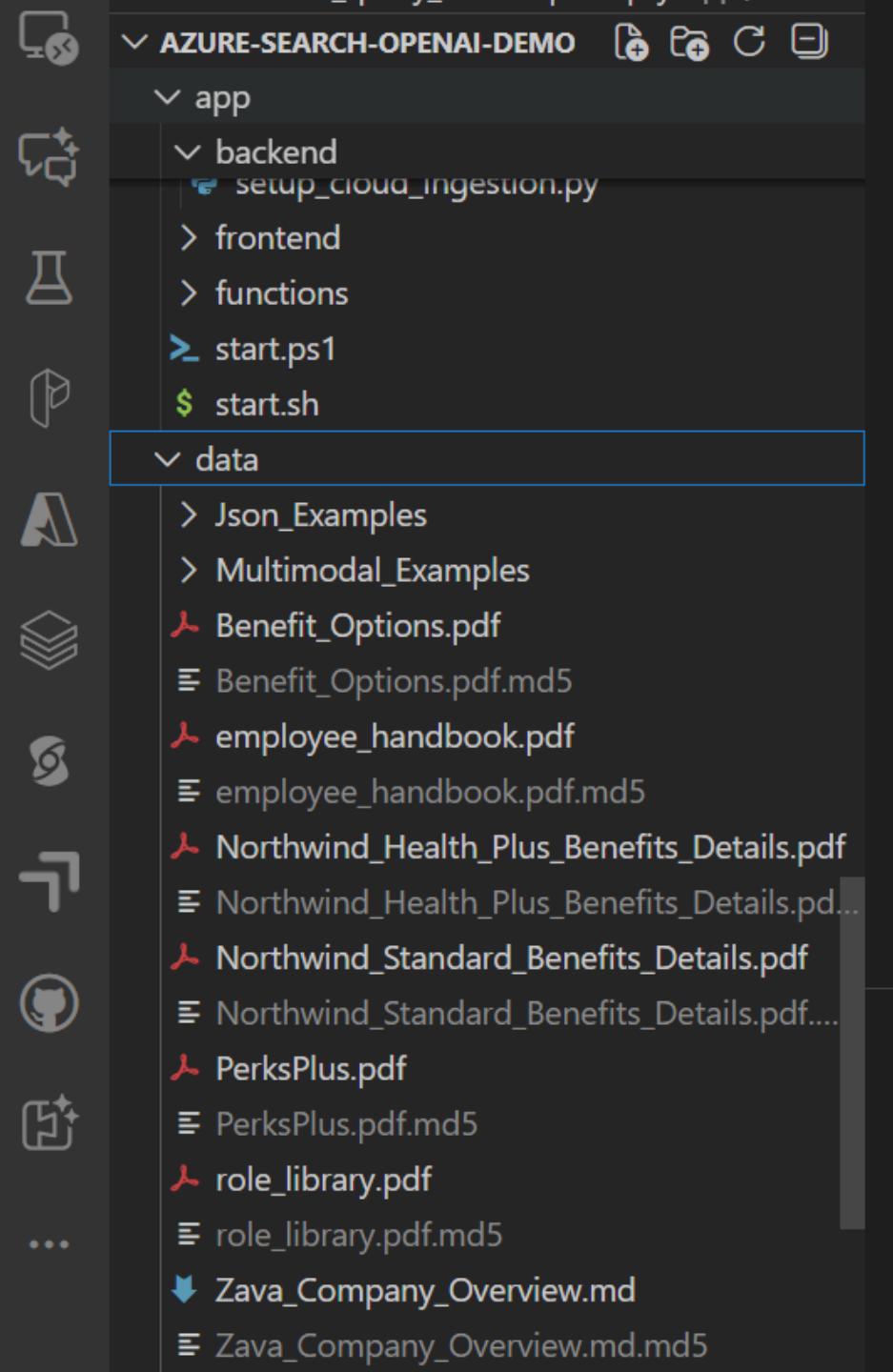
Connect to apps, optimize search results. Leverage features like faceting, filtering, scoring profiles and more. [Learn more](#)

More tools and repl...

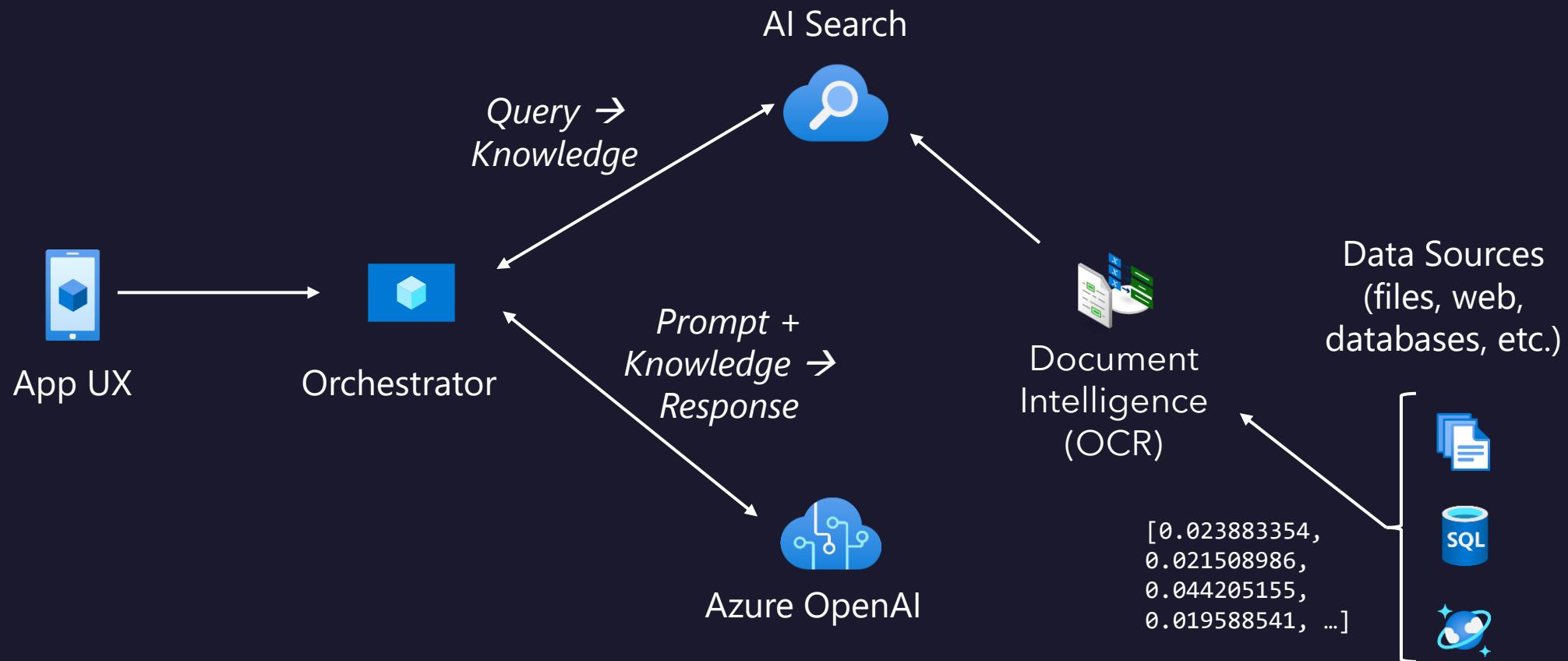
# Default values in the accelerator

1. What infra to create: infra > main.bicep
2. What was created: .azure > [project name] > .env

# The data sources are listed in the data folder



# RAG (retrieval augmented generation) is used especially in on-your-own-data scenarios



# backend>prepdocslib>textsplitter.py defines chunking

The chunking method is defined in [textsplitter.py](#). This file contains:

1. **Base class** [TextSplitter](#) (lines 14-23) - Abstract base class for splitting pages into chunks
2. **Main chunking implementation** [SentenceTextSplitter](#) (starts around line 194) - The primary chunking class that:
  - Splits pages into smaller chunks for embedding models
  - Has configurable [max tokens per section](#) (**hard** token limit for chunks), default 500 tokens
  - Uses [DEFAULT SECTION LENGTH](#) (soft character limit for chunks, can be exceeded by up to 20%; figures are exempt from this limit) of 1000 characters
  - Implements 10% overlap between chunks ([DEFAULT OVERLAP PERCENT](#))
  - Intelligently splits at sentence boundaries or word breaks
3. **Helper classes and constants:**
  - [ChunkBuilder](#) (lines 131-191) - Accumulates text fragments respecting size limits
  - Word breaks and sentence endings for standard and CJK languages

**In the version of the app you have installed, you have 2 approaches**

### **Embedding Generation (lines 877-900)**

compute\_text\_embedding(): Creates text embeddings using Azure OpenAI

compute\_multimodal\_embedding(): Creates image embeddings using Azure AI Vision

# In the version of the app you have installed, you have 2 approaches

**A. Ask Approach:** retrievethenread.py: Simple retrieve-then-read pattern (lines 164-233):

1. Takes user query directly
2. Generates embeddings if using vector search
3. Calls search() method
4. Returns documents and metadata

**B. Chat Approach** *let's focus here*: chatreadretrieveread.py: Multi-step retrieval with query rewriting (lines 339-437):

- 1. Query Rewriting:** Uses GPT to transform chat history + user query into optimized search query
- 2. Embedding Generation:** Creates embeddings from the rewritten query
- 3. Search Execution:** Calls search() method with optimized query
- 4. Returns:** Documents with thought steps showing the process

Microsoft Azure

Search resources, services, and docs (G+)

Copilot

admin@MngEnvMCAP0...  
CONTOSO (MNGENVMCAP0388...)

Home > Resource Manager | Resource groups > rg-azuresearch251208 > gptkb-f4kshm2qzxtu4

### gptkb-f4kshm2qzxtu4 | Indexes

Search service

Search

+ Add index Refresh Delete

Filter by name...

Name	Document count	Vector index quota ...	Total storage size
gptkbindex	758	139.75 KB	10.22 MB

Overview

Activity log

Access control (IAM)

Tags

Diagnose and solve problems

Resource visualizer

Agentic retrieval

Knowledge sources

Knowledge bases

Search management

Indexes

# Embeddings (vectors) are uploaded to the index

You can view the sources in AI Search

Home > Resource Manager | Resource groups > rg-azuresearch251208 > gptkb-f4kshm2qzxtu4 | Indexes > gptkbindex

Save Discard Refresh Create demo app Edit JSON Delete Encryption

Documents	Total storage	Vector index quota usage	Max storage
758	10.22 MB	139.75 KB	15 GB

Search explorer Fields CORS Scoring profiles Semantic configurations Vector profiles

Query options View

dentist

Results

```
1 "@odata.context": "https://gptkb-f4kshm2qzxtu4.search.windows.net/indexes('gptkbindex')/$metadata#docs(*)",
2 "@odata.count": 620,
3 "@search.answers": [
4   {
5     "key": "file-Northwind_Health_Plus_Benefits_Details_pdf-4E6F72746877696E645F4865616C74685F506C75735F42656E656
6     "text": "Northwind Health Plus is a comprehensive health plan that offers coverage for medical, vision, and dental services. It includes benefits like prescription drugs, vision care, and dental coverage. The plan also provides preventive services like annual check-ups and vaccinations at no cost sharing. Northwind Health Plus is designed to help you manage your healthcare needs efficiently and affordably." ,
7     "highlights": "Northwind Health Plus is a comprehensive health plan that offers coverage for medical, vision, and dental services. It includes benefits like prescription drugs, vision care, and dental coverage. The plan also provides preventive services like annual check-ups and vaccinations at no cost sharing. Northwind Health Plus is designed to help you manage your healthcare needs efficiently and affordably." ,
8     "score": 0.9929999709129333
9   },
10  {
11    "key": "file-Northwind_Health_Plus_Benefits_Details_pdf-4E6F72746877696E645F4865616C74685F506C75735F42656E656
12    "text": "...s, such as teeth whitening or veneers . Services that are not medically necessary to relieve pain are not covered under the plan." ,
13    "highlights": "...s, such as teeth whitening or veneers . Services that are not medically necessary to relieve pain are not covered under the plan." ,
14    "score": 0.968999981880188
15  },
16  {
17    "key": "file-Northwind_Standard_Benefits_Details_pdf-4E6F72746877696E645F5374616E646172645F42656E65666974735F
18    "text": "...sia COVERED SERVICES: Dental Injury and Facility Anesthesia The Northwind Standard plan offers coverage for dental services, including emergency care, orthodontics, and oral surgery. It also covers dental injuries and facility anesthesia. The plan includes preventive services like cleanings and exams." ,
19    "highlights": "...sia COVERED SERVICES: Dental Injury and Facility Anesthesia The Northwind Standard plan offers coverage for dental services, including emergency care, orthodontics, and oral surgery. It also covers dental injuries and facility anesthesia. The plan includes preventive services like cleanings and exams." ,
20    "score": 0.9660000205039978
21  },
22 }
```

# What happens when a user asks a question?

User Query



[Optional: Query Rewriting via GPT]



[Generate Embeddings]

- ├→ Text Embedding (Azure OpenAI)
- └→ Image Embedding (Azure AI Vision)



[Azure AI Search Query]

- ├→ Text Search (BM25)
- ├→ Vector Search (HNSW)
- ├→ Hybrid Search (combines both)
- └→ Semantic Ranker (L2 reranking)



[Filter & Score Results]

- ├→ Apply security filters
- ├→ Apply minimum score thresholds
- └→ Extract semantic captions



[Return Documents to Be Used by the Model When Addressing User's Question]

# Retrieval strategies in Azure AI Search



## 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)

# Vector search is good. Hybrid – even better

	Full-text search (BM25)	Pure Vector search (ANN)	Hybrid search (BM25 + ANN)
Exact keyword match	✓	✗	✓
Proximity search	✓	✗	✓
Term weighting	✓	✗	✓
Semantic similarity search	✗	✓	✓
Multi-modal search	✗	✓	✓
Multi-lingual search	✓	✓	✓

# Vector and full-text search have different advantages

Leveraging the Strengths of Lexical and Semantic Approaches in Retrieval

## Discrete (Lexical) Representations

- Advantages:
  - Exact matching
  - Precise control and easy explainability
- Limitations:
  - Struggle to capture nuances in language
  - Limited understanding of conceptual similarity

## Dense (Vector/Semantic) Representations

- Advantages:
  - Capture conceptual similarity
  - Better understanding of language nuances
- Limitations:
  - Not built to match exact terms
  - Reduced explainability compared to discrete

**Bottom Line:** Achieve optimal recall by leveraging the strengths of both discrete and semantic representations for a comprehensive understanding of language

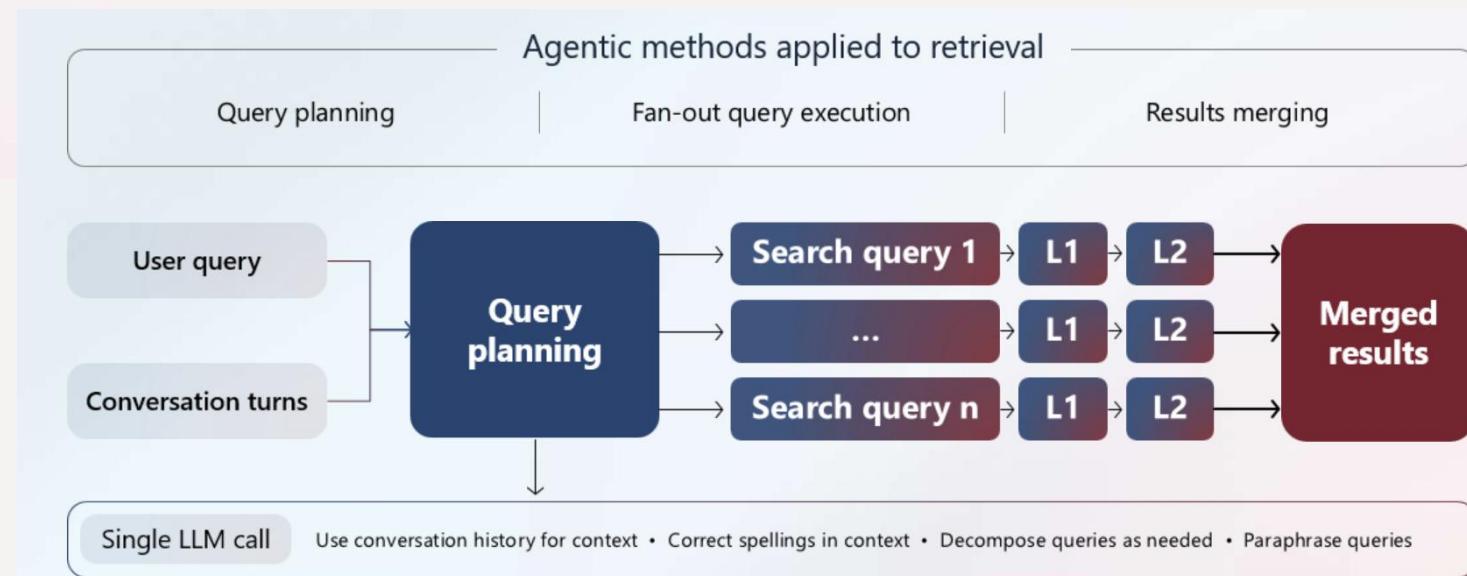
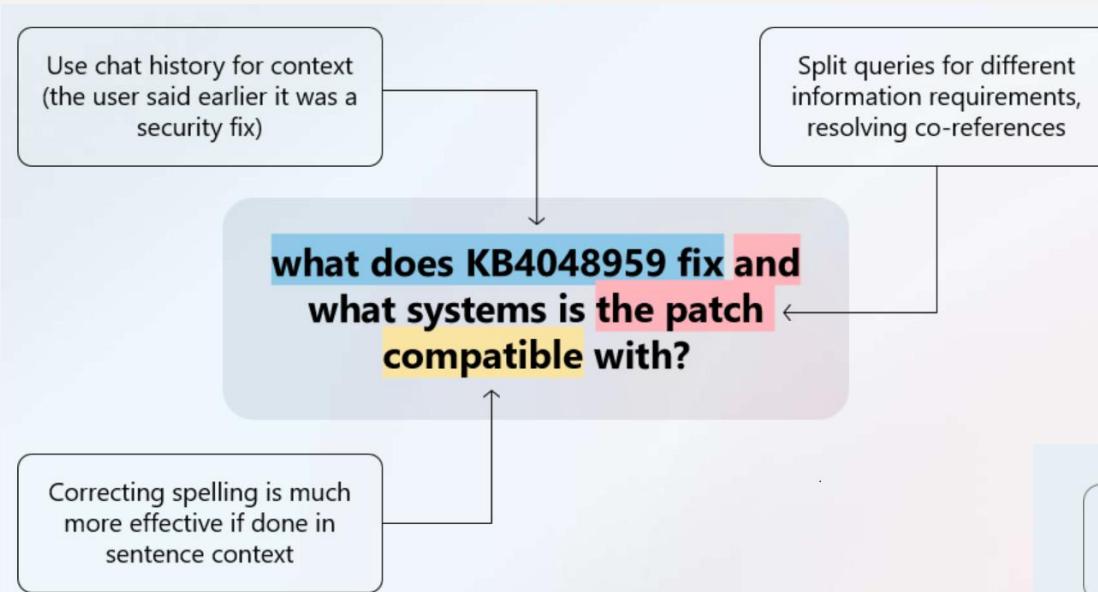
# Semantic ranker additionally organizes the best results, improving the score

Optimize Vector Search with Hybrid and Semantic Ranker Search

- SOTA re-ranking encoder model
- Highest performing retrieval mode
- Free 1000 queries/month
- Multilingual capabilities
- Includes extractive answers, captions, and highlights just like Bing.con

Search Configuration	Customer datasets [NDCG@3]	Beir [NDCG@10]	Multilingual Academic (MIRACL) [NDCG@10]
Keyword	40.6	40.6	49.6
Vector (Ada-002)	43.8	45.0	58.3
Hybrid (Keyword + Vector)	48.4	48.4	58.8
Hybrid + Semantic ranker	<b>60.1</b>	<b>50.0</b>	<b>72.0</b>

# Agentic retrieval in AI Search is a great solution for more complicated questions



# Our app provides 2 alternative search methods: The more advanced one needs to be enabled



**Standard Search Approach (Traditional RAG)**

1. Query Preparation
2. Vector Embedding Generation
3. Azure AI Search Execution
4. Result Processing
5. Ask approach/ chat approach



**Agentic Retrieval Approach (Advanced)**

1. Knowledge Source Setup
2. Reasoning Effort Modes
3. Agentic Retrieval Execution
4. The agent can search across multiple knowledge sources: indexed documents, web, sharepoint
5. Result Aggregation

# There are several important places to set up

## 1. Core Retrieval Logic: backend > approaches > [approach.py](#)

This is the base class that contains the fundamental retrieval methods used by all approaches:

**Main Search Method (lines 295-368):** The [search\(\)](#) method performs the actual Azure AI Search query with:

- **Hybrid search:** Combines text search and vector search
- **Semantic ranker:** Optional L2 reranking for better relevance
- **Semantic captions:** Extracts relevant passages from documents
- **Query rewriting:** Uses Azure AI Search generative query rewriting
- **Filtering:** Applies security filters and custom filters
- **Score thresholding:** Filters results by minimum search/reranker scores (minimum\_search\_score 0-1, minimum\_reranker\_score 0-4; default for both is 0.0); the reranker score is only available when semantic ranker is enabled

## 2. Embedding Generation (lines 877-900)

- [compute\\_text\\_embedding\(\)](#): Creates text embeddings using Azure OpenAI
- [compute\\_multimodal\\_embedding\(\)](#): Creates image embeddings using Azure AI Vision

## 3. Agentic Retrieval (lines 451-650):

The [run\\_agentic\\_retrieval\(\)](#) method implements advanced agentic retrieval that:

- Uses [KnowledgeBaseRetrievalClient](#) from Azure AI Search
- Supports multiple reasoning efforts: minimal, low, medium
- Can query multiple knowledge sources: search index, web, SharePoint
- Performs query rewriting or intent extraction
- Returns structured results with activities and references

# Agentic retrieval offers more than a simple search

When a user asks a question with web search enabled:

## **Step 1: Query Planning** (approach.py line 437-447)

- An AI agent analyzes the conversation
- Decides which sources to search (index, web, or both)
- Generates optimized search queries for each source

## **Step 2: Parallel Searching** (approach.py line 469-477)

```
if use_web_source:  
    knowledge_source_params_list.append(  
        WebKnowledgeSourceParams(  
            knowledge_source_name="web",  
            include_references=True,  
            include_reference_source_data=True,  
            always_query_source=False,  
        )  
    )
```

## **Step 3: Results Processing** (approach.py line 616-622)

- Web results are extracted as KnowledgeBaseWebReference objects
- Each web result contains: id, title, URL, and activity metadata
- Results are serialized with type="web" for frontend display

## **Step 4: Answer Synthesis** (approach.py line 650-663)

- When web source is used, the agent synthesizes an answer
- Reference tokens [ref id:X] are replaced with actual URLs or document names
- The synthesized answer is returned to the user
- Query Plan Visualization: Users can see the "thought process":

# You can find the evaluations in the *evals* folder

## evaluate.py - Main Evaluation Script

This is the primary evaluation tool that measures your RAG app's performance using various metrics.

### Key Features:

- **Custom Metrics:**
  - AnyCitationMetric: Checks if responses contain any citations (e.g., [Document.pdf#page=7])
  - CitationsMatchedMetric: Measures what percentage of ground truth citations appear in the response
- **Built-in Metrics** (via Azure AI Evaluation):
  - gpt\_groundedinness: Whether the answer is grounded in the provided context
  - gpt\_relevance: Whether the answer is relevant to the question
  - answer\_length: Length of the generated answer
  - latency: Response time

### How it works:

1. Loads test questions from ground\_truth.jsonl or ground\_truth\_multimodal.jsonl
2. Sends each question to your running RAG app
3. Evaluates responses using AI-powered metrics
4. Saves results to the `results/` directory with detailed statistics



# The following evaluators are available on Foundry

## Quality

**NEW**

Document Retrieval

Groundedness

Relevance

Coherence

Fluency

Similarity

NLP Metrics (e.g., F1 Score)

**NEW**

AOAI Graders

**NEW****NEW**

## Risk & safety

Indirect Attack Jailbreaks

Direct Attack Jailbreaks

Hate and Unfairness

Sexual

Violence

Self-Harm

Protected Material

Ungrounded Attributes

Code Vulnerability

**NEW**

## Agents

Intent Resolution

Tool Call Accuracy

Task Adherence

Response Completeness

+ Custom evaluators

# If you don't want to create ground\_truth "manually", you can create it in an automated way

## generate\_ground\_truth.py - Test Data Generator

Creates synthetic test datasets automatically from your indexed documents using RAGAS (RAG Assessment).

### **Process:**

- 1. Fetches documents** from your Azure AI Search index
- 2. Builds a knowledge graph** representing relationships between document chunks
- 3. Generates questions** using AI that are answerable from your documents
- 4. Creates ground truth answers** with proper citations
- 5. Saves QA pairs to ground\_truth.jsonl**

### **Key Features:**

- Uses RAGAS library for intelligent test generation
- Creates realistic questions based on your actual data
- Includes proper citations in answers
- Can reuse existing knowledge graphs to save time

# *Evals* also includes security-related evaluators

## safety\_evaluation.py - Safety & Security Testing

Tests your RAG app against adversarial inputs to ensure safety and responsible AI practices.

### What it does:

- 1. Generates adversarial questions** using Azure AI's adversarial simulator
- 2. Tests your app** with potentially harmful queries
- 3. Evaluates responses** for harmful content in 4 categories:
  1. hate\_unfairness: Hateful or unfair content
  2. sexual: Sexual content
  3. violence: Violent content
  4. self\_harm: Self-harm related content
- 4. Generates safety report** with severity scores and statistics
- 5. Saves results to safety\_results.json**

# Evaluators available "out of the box" usually use a scale of 1-5

## Understanding Results

### Evaluation Results (results/ folder)

After running `evaluate.py`, you'll find:

- **Individual question metrics:** Score for each question
- **Aggregate statistics:** Overall averages, totals, rates
- **Citations analysis:** How well your app provides sources
- **Latency metrics:** Response time analysis

Example metrics:

- `gpt_relevance`: 4.5/5 - Answer relevance score
- `citations_matched`: 0.85 - 85% of citations matched
- `any_citation`: true - Response included citations
- `latency`: 1.23s - Response time

### Safety Results (`safety_results.json`)

Contains:

- **Low count:** Number of responses with low/very low severity
- **Low rate:** Percentage of safe responses
- **Mean score:** Average severity score per category

```
evals > results > gpt4omini-ada002 > {} summary.json > ...
1  {
2    "gpt_groundedness": {
3      "pass_count": 44,
4      "pass_rate": 0.88,
5      "mean_rating": 4.62
6    },
7    "gpt_relevance": {
8      "pass_count": 42,
9      "pass_rate": 0.84,
10     "mean_rating": 4.12
11   },
12   "answer_length": {
13     "mean": 922.42,
14     "max": 1616,
15     "min": 193
16   },
17   "latency": {
18     "mean": 3.14,
19     "max": 7.583068,
20     "min": 1.598833
21   },
22   "citations_matched": {
23     "total": 25,
24     "rate": 0.5
25   },
26   "any_citation": {
27     "total": 50,
28     "rate": 1.0
29   }
30 }
```

# Evaluation best practices

1. **Start small:** Test with --numquestions 10 first
2. **Run locally before deployment:** Catch issues early
3. **Regular safety testing:** Run safety eval before major releases
4. **Track metrics over time:** Compare results across versions
5. **Iterate on failures:** Review low-scoring questions to improve prompts
6. **Use both configs:** Test both /ask and /chat endpoints

# Monitoring takes place in app insights and log analytics

The screenshot shows the Microsoft Azure Log Analytics workspace interface. The left sidebar navigation includes Home, Log Analytics workspace, Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems, Logs (selected), Resource visualizer, Settings, Tables, Agents, Usage and estimated costs, Data export, Network isolation, Identity, Linked storage accounts, Properties, Locks, Classic, Legacy activity log connector, and Legacy storage account logs.

The main area displays a search bar at the top with the placeholder "Search resources, services, and docs (G+)" and a Copilot button. Below the search bar is a breadcrumb trail: Home > log-szmvad2z4r7g | Logs. The title "log-szmvad2z4r7g | Logs" is shown with a star icon and three dots.

A "New Query 1+" card is open, showing a table of contents under "Tables". The "LogManagement" section is expanded, listing AppDependencies, AppExceptions, AppMetrics, and AppRequests. The "AppRequests" item is selected, showing its schema:

Type	Name
T	AppRoleInstance (string)
T	AppRoleName (string)
T	AppVersion (string)
T	ClientBrowser (string)
T	ClientCity (string)
T	ClientCountryOrRegion (string)
T	ClientIP (string)
T	ClientModel (string)
T	ClientOS (string)
T	ClientStateOrProvince (string)
T	ClientType (string)

The main content area shows a query results table titled "1 AppRequests". The time range is set to "Last 24 hours" and the results count is "1000 results". The table has columns: TimeGenerated [UTC] (sorted ascending), Id, Name, and Url. The single result row is as follows:

TimeGenerated [UTC]	Id	Name	Url
> 12/15/2025, 4:38:56.368 PM	4cdf190fc672112	POST /ask	http://capps-backend-szmvad2z4r7g.bravecoast-ea9e7930...

At the bottom, there are buttons for "Save", "Share", and "Queries hub", and a "KQL mode" dropdown. The footer includes links for "Query details" and "1 - 13 of 16".

Home > appi-szvmavd2z4r7g

## appi-szvmavd2z4r7g | Performance

Application Insights

Search Refresh Code Optimizations Profiler View in Logs Analyze with Workbooks Copy link Feedback

Diagnose and solve problems Resource visualizer Investigate Application map Smart detection Live metrics Search Availability Failures Performance Agents (preview)

Alerts Metrics Diagnostic settings Logs Workbooks Dashboards with Grafana Usage

admin@MngEnvMCAP0.. CONTOSO (MNGENVMCAP0388..)

Server Browser Local Time: Last 24 hours Roles = All

Operations Dependencies Roles

Operation times: zoom into a range

Avg 50th 95th 99th

1.2 sec  
1000 ms  
800ms  
Request count 20  
0

06:00 PM 09:00 PM Mon 15 03:00 AM 06:00 AM 09:00 AM 12:00 PM 03:00 PM

05:43 PM 05:43 PM 05:43 PM 05:43 PM

Overall Distribution of durations: zoom into a range Scale 99th

Distribution of durations: zoom into a range

Request count Duration

2 50th 1 0.0ms 20ms 140ms 560ms 1.4sec 3.7sec 8.4sec 17.7sec

1.0ms 10ms 100ms 1sec 10sec

Insights (2)

69% COMMON PROPERTIES: resultCode, client\_OS, performanceBucket, cloud\_RoleName, cl...

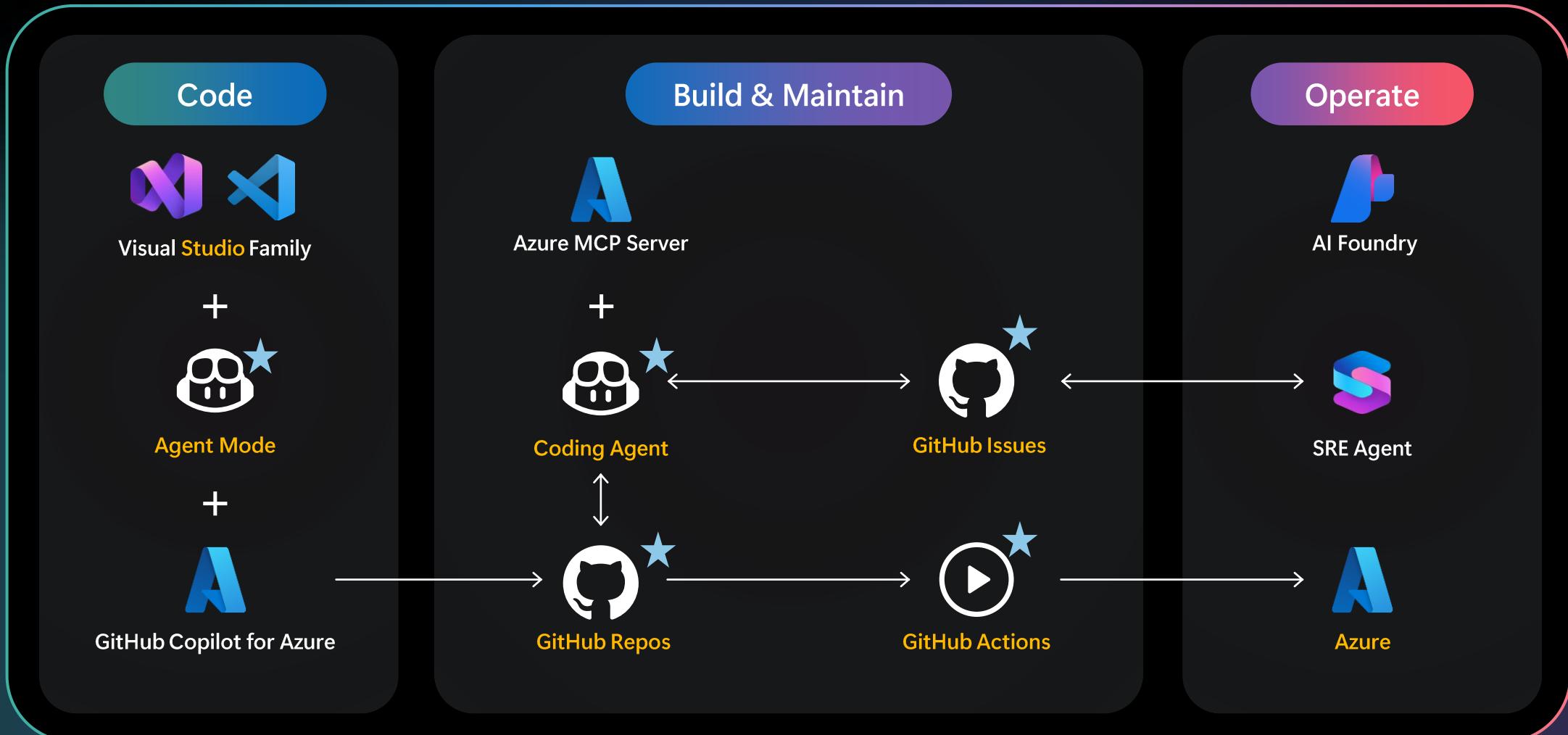
Top 3 Dependencies

Drill into... 16 Samples

### 3. GitHub – Actions and Copilot

45 minut

# Agentic DevOps – key elements



# GitHub Actions

## Goal:

Deployment automation, CI/CD style

Running additional scanners (linter, etc.)

## Scenarios:

00TKScenarios\00-GitHub-Start-Codespaces\00-GitHub-Start-CodeSpaces.md

00TKScenarios\01-GitHubActions\01-GitHubActions.md

# GitHub Advanced Security (i Quality Agent)

## Goal

Analysis of the CODE for security

Analysis of external dependencies (dependency tree)

Source quality analysis – suboptimal structures, inconsistent loops, etc.

## Scenarios

00TKScenarios\02-CodeQuality-AdvSecurity\02-CodeQualit-AdvSecurity.md

# GitHub Copilot

## Goal

Understand the code that was created

Add new functionality faster than usual

## Scenarios:

00TKScenarios\03-Copilot-Questions\03-Copilot-question.md

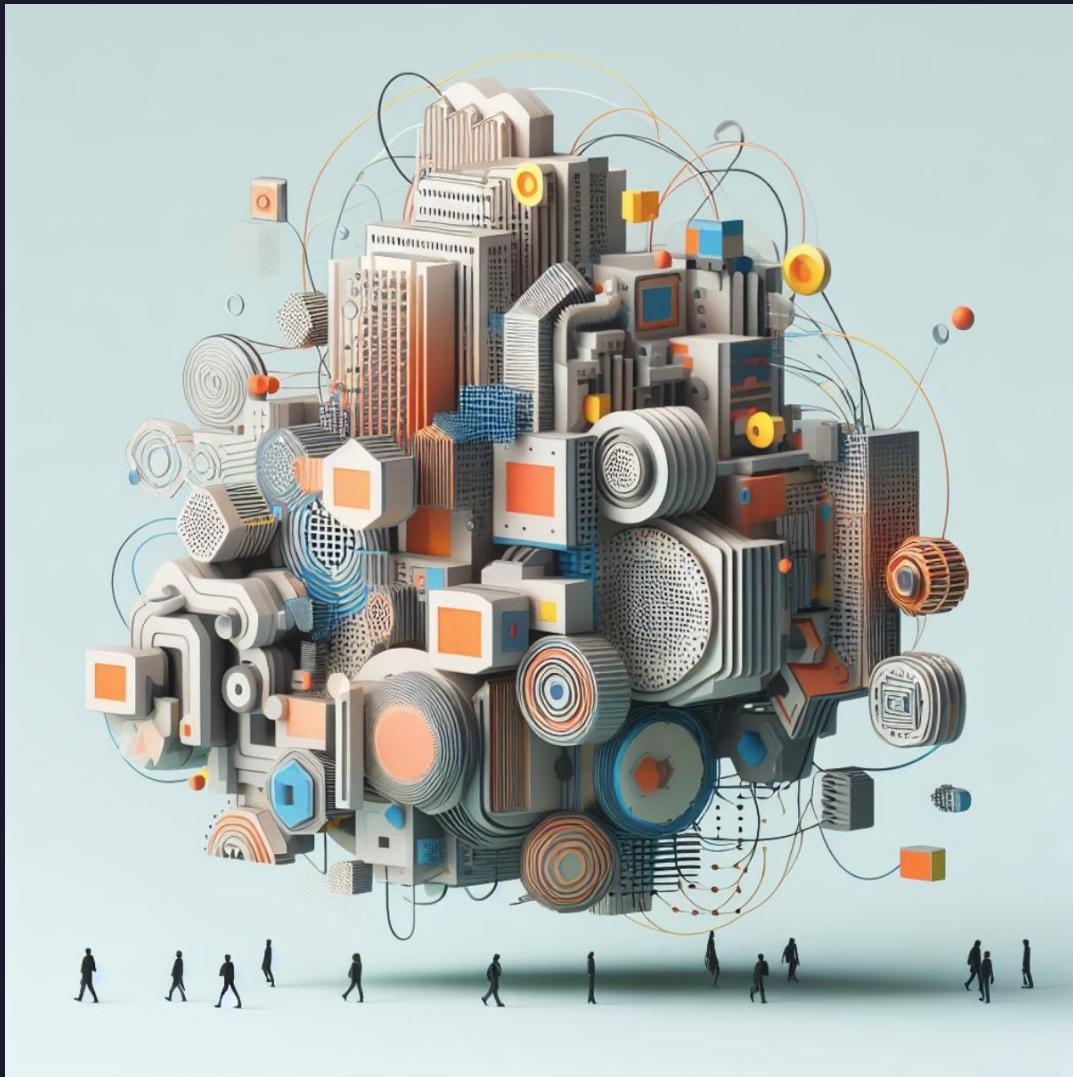
00TKScenarios\04-Santa-Claus\04-Santa-Claus.md

## 4. New functionalities

1h 15 minutes

## 4.1. Changing the chunk and search parameters

# Data preparation is crucial for RAG



1. **Data governance** – How often will new versions of documents appear/ will it be necessary to delete certain documents from the data repository and how to do it
2. How do I ensure that there is no contradiction in our data repository?
3. Method and frequency of indexing/embedding of source documents
4. **Chunk size** and their boundaries
2. **Security and compliance**

# Types of chunking

Key reasons to invest in a strong chunking strategy include:

- Context window constraints
- Improved retrieval efficiency
- Computational optimization

1. **Fixed-Size Chunking**
2. **Semantic Chunking** – This method splits documents at logical boundaries (e.g., sentences, paragraphs, or sections). Consecutive segments that are highly similar may often be merged, resulting in more coherent text blocks.
3. **Recursive Chunking** – This relies on a hierarchy of separators. The algorithm first attempts to split on high-level separators and then moves to increasingly finer separators if the resulting chunks are still too large.
4. **Adaptive Chunking** – This changes the chunk size based on text complexity. Simpler sections become larger chunks, while denser or more intricate sections become smaller chunks.
5. **Context-Enriched Chunking** – This method attaches additional metadata or summaries to each chunk. By doing so, retrieval models have more background context for each chunk, leading to improved understanding during generation.
6. **AI-Driven Dynamic Chunking** – This is AI-based chunking that leverages a Large Language Model (LLM) to detect natural breakpoints in the text. This ensures that each chunk encapsulates complete ideas.

# Remember to add all important information to the index

Example: 2 pages in Excel with identical data for:

- private clients,
- institutional clients

If we put just the title of the file and its content into AI Search, our records will be incomplete!

	A	B	C	D
1	Produkt	Cena		
2	A		32	
3	B		14	
4	C		63	
5	D		35	
6	E		24	
7				
8				
9				
10				
11				
12				
13				

Below the table is a screenshot of an Excel ribbon. The tabs 'klienci\_pryw' and 'klienci\_instytuc' are visible, with 'klienci\_pryw' being the active tab. The status bar at the bottom shows 'Ready' and 'Accessibility: Good to go'.

# When deciding on a chunk strategy, consider

## 1. Document Structure & Type:

- The nature of your source material heavily influences the best approach:
- Structured text (e.g., reports, articles): Semantic or recursive chunking is often recommended. Or AI-driven chunking.
  - Code or highly technical documents: Favor recursive or language-specific chunking.
  - Mixed or unstructured content: AI-driven or context-enriched chunking is typically most effective.

## 2. Query Complexity:

The expected complexity of user queries should guide the chunk size:

- Straightforward, fact-based queries: Require smaller, more direct chunks.
- Multifaceted, analytical queries: Benefit from larger, context-preserving chunks.
- Queries spanning multiple concepts: Use strategies that are designed to keep related data together.

## 3. Model Constraints:

- It's crucial to align the chunking strategy with the limitations of your models:
- Pay attention to the context window sizes of both the LLMs and the embedding models you are using.
  - Keep an eye on token usage to help avoid excessive costs.

## 4. Performance Requirements:

Your desired outcome - speed versus accuracy - determines the necessary sophistication:

- Latency-sensitive use cases: Favor lighter, simpler chunking for faster retrieval times.
- Accuracy-critical domains: Require more advanced or context-enriched chunking.

# The discussed chunking will still not allow the chatbot to answer "summary" questions

## 1. Examples:

1. "List all documents that mention [subject],"
2. "Describe the character of Adam in detail" (based on all excerpts from a longer book)
3. "Describe the course of events in the book step by step" (based on all excerpts from the longer book)

## 2. How to solve it?

1. Will such questions be asked? Is it important for a bot to answer such questions?
2. Preparing additional source materials and uploading them to the index, e.g. summaries of documents
3. Adding GraphRAG

# Current chunking in the app

**Semantic Chunking with sentence-aware boundaries** with elements of recursive chunking and context-enriched chunking

The app uses a custom SentenceTextSplitter class (in textsplitter.py) that implements:

1. **Semantic Chunking** - Splits documents at logical boundaries (sentences, paragraphs) rather than arbitrary character positions
2. **Recursive Chunking** - When text spans exceed the token limit, it recursively splits using a hierarchy of separators:
  - a) First attempts: sentence endings (., !, ? plus CJK equivalents like 。, !, ?)
  - b) Second attempts: word breaks (spaces, punctuation)
  - c) Fallback: midpoint split with 10% overlap
3. **Fixed-Size Constraints** - Enforces hard limits:
  - a) 500 tokens max per chunk
  - b) ~1000 characters soft limit (with 20% overflow tolerance)
4. **Context-Enriched Chunking** - Adds 10% semantic overlap between consecutive chunks to improve retrieval

## Special Features

- **Figure-aware**: Treats <figure> blocks as atomic units that never get split
- **Cross-page repair**: Intelligently merges sentence fragments that were split across PDF page boundaries
- **Adaptive boundaries**: Preserves sentence integrity while respecting token limits

# Let's enter an additional "priority" field into the index

1. Add a "priority" field in the AI Search index
2. Add appropriate processing of source documents: before uploading them to the index, they should be assigned a priority on a scale of 1-3 (for hackathon purposes, this can be a random number)
3. Edit the chatbot prompt so that documents with higher priority are considered more important
4. Delete all .md5 files from the data folder – to re-vectorize the source materials
5. Delete .azure – we create a copy of the application
6. Run azd up
7. Remarks:
  1. Use GitHub Copilot
  2. Note that "in real life" the priority logic should be thought out in the context of the filtering by score used in the code (i.e. how well the chunk fits the user's question)

Individual work. If you're stuck, check out the solution example:  
<https://GitHub.com/aganiezgoda/azure-search-openai-demo-hack>

# If you use semantic ranking, remember about the appropriate semantic configuration

hotels-sample-index ...

Save Discard Refresh Create demo app Edit JSON Delete

Documents Total storage Vector index size Max storage

50 560.69 KB 0 Bytes 15 GB

Search explorer Fields CORS Scoring profiles Semantic configurations

Add semantic configuration Delete

You haven't created any semantic configurations

Create

New semantic configuration

Name \* my-semantic-config

Title field HotelName

Content fields

Insert Delete Move up Move down Move to top

Field name

Description Please select a field

Keyword fields

Insert Delete Move up Move down Move to top

Field name

Tags

Save Cancel

The screenshot shows the Algolia Admin interface for a 'hotels-sample-index'. On the left, there's a summary of the index: 50 documents, 560.69 KB total storage, 0 Bytes vector index size, and 15 GB max storage. Below this are tabs for Search explorer, Fields, CORS, Scoring profiles, and Semantic configurations (which is underlined). A button to 'Add semantic configuration' is visible. The main area says 'You haven't created any semantic configurations' and has a 'Create' button. On the right, a modal window titled 'New semantic configuration' is open. It has a 'Name' field containing 'my-semantic-config' (which is highlighted with a red box). Below it is a 'Title field' dropdown set to 'HotelName'. Under 'Content fields', there are dropdowns for 'Description' (set to 'Please select a field') and 'Tags'. At the bottom of the modal are 'Save' and 'Cancel' buttons, with 'Save' also highlighted with a red box.

# Which search method to choose – the most important

1. Hybrid search with semantic ranker to default
2. Semantic ranker improves results – it also slightly increases the costs of Azure AI Search
3. Semantic configuration is key
4. Documents uploaded to the index can be processed accordingly, e.g. adding fields to them
5. After adding fields, e.g. "valid from", "valid till", you can filter the results accordingly when sending a query

# Choosing a model for chunking

## Model Characteristics

- Task Specificity
- Performance
- Context Awareness
- Model Size and Inference Speed
- Language Support
- Customizability (ability to fine-tune)

## Implementation Considerations

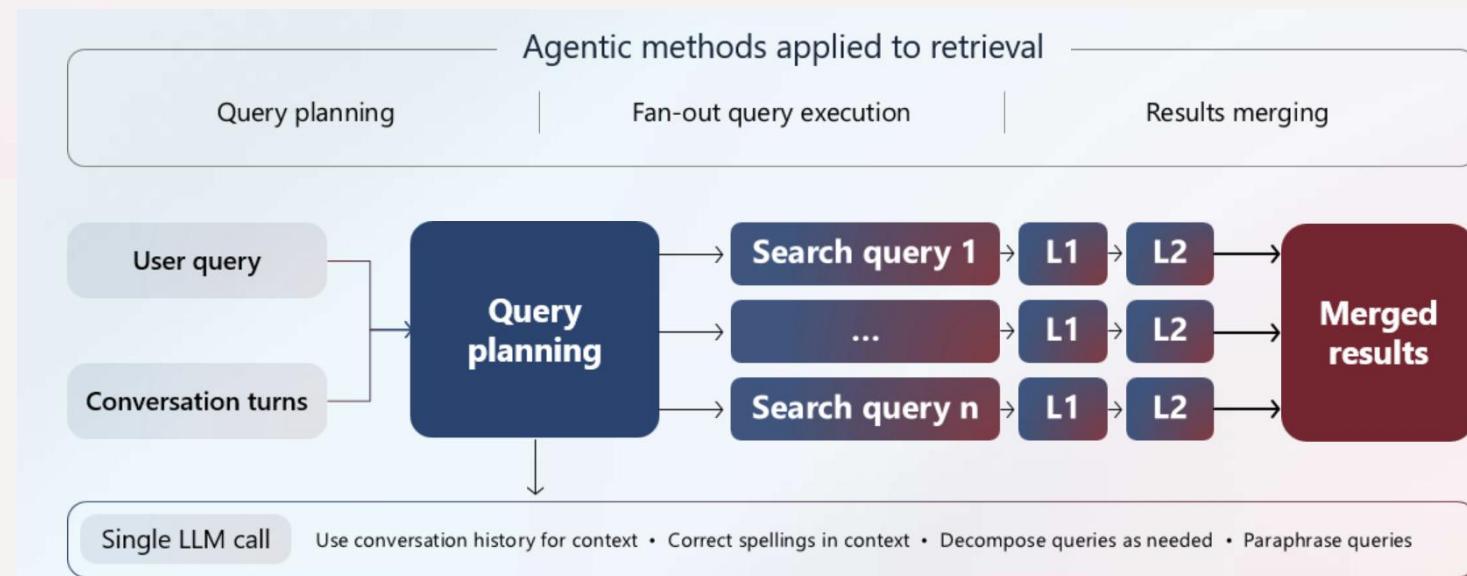
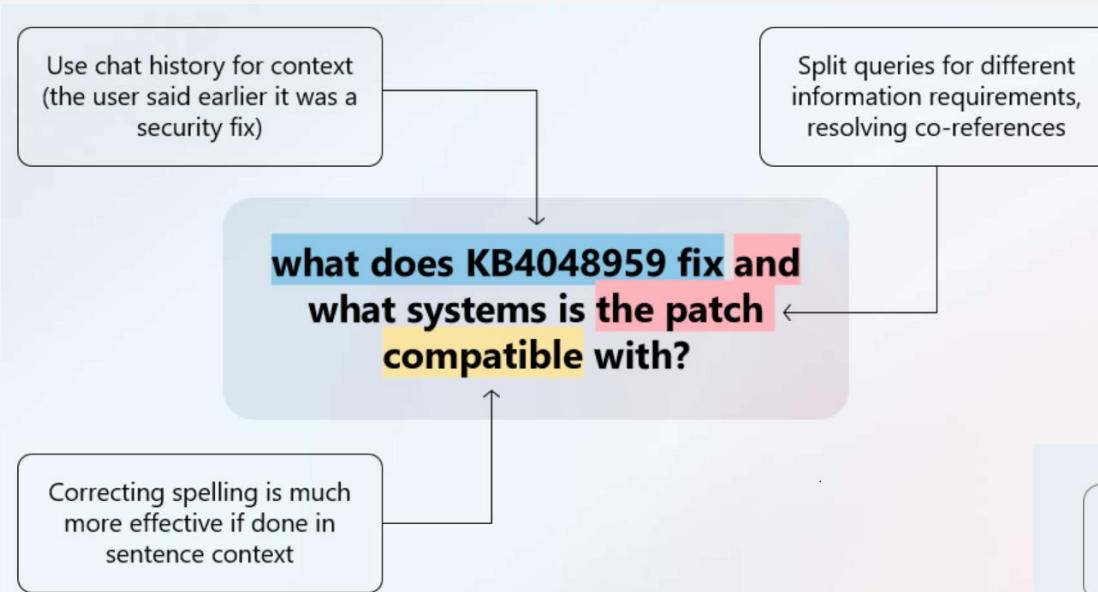
- Training Time and Complexity
- Pre-Trained Models
- Integration
- Community Support and Updates
- Cost

**Ada (text-embedding-ada-002)** produces **1536- dimensional** embeddings. It offers good accuracy at a low cost but has now been surpassed by newer models in both efficiency and quality.

**Text-Embedding-3-Large**, is the **highest-performing OpenAI embedding model** with **3072 dimensions**.

## **4.2. Adding Web and Sharepoint source search in addition to AI Search index**

# Agentic retrieval in AI Search is a great solution for more complicated questions



# Switch search to agent-based by adding web search and SharePoint

```
# Enable web search  
azd env set USE_WEB_SOURCE true
```

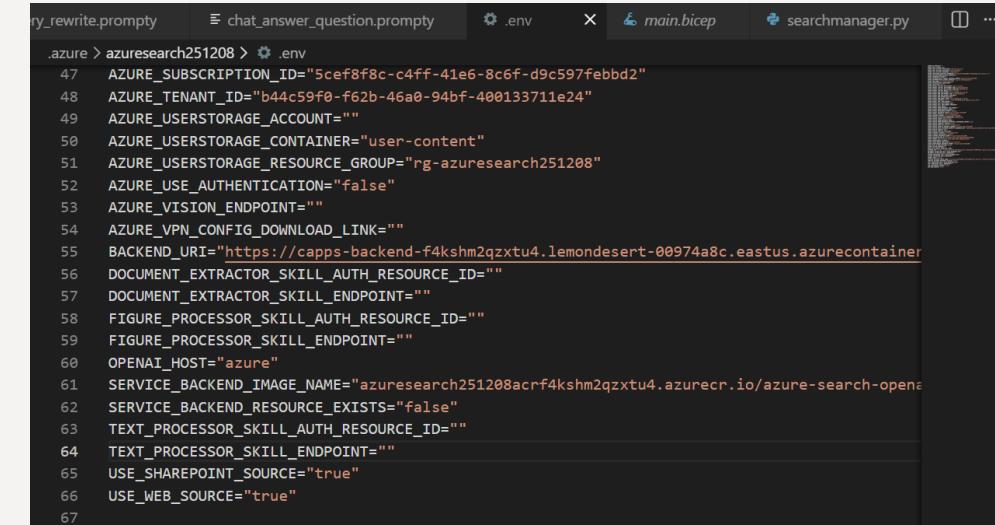
```
# Enable SharePoint search  
azd env set USE_SHAREPOINT_SOURCE true
```

```
# Set the knowledge base model (required for agentic retrieval)  
azd env set AZURE_OPENAI_KNOWLEDGEBASE_DEPLOYMENT knowledgebase  
azd env set AZURE_OPENAI_KNOWLEDGEBASE_MODEL gpt-4.1-mini  
azd env set AZURE_OPENAI_KNOWLEDGEBASE_MODEL_VERSION 2025-04-14  
azd env set AZURE_SEARCH_KNOWLEDGEBASE_NAME "gptkb"
```

```
# Then deploy  
azd up
```

Individual work  
If you're stuck use GH Copilot

Result:



A screenshot of a terminal window titled 'chat\_answer\_question.prompty'. It shows several environment variables being set. The variables include AZURE\_SUBSCRIPTION\_ID, AZURE\_TENANT\_ID, AZURE\_USERSTORAGE\_ACCOUNT, AZURE\_USERSTORAGE\_CONTAINER, AZURE\_USERSTORAGE\_RESOURCE\_GROUP, AZURE\_USE\_AUTHENTICATION, AZURE\_VISION\_ENDPOINT, AZURE\_VPN\_CONFIG\_DOWNLOAD\_LINK, BACKEND\_URI, DOCUMENT\_EXTRACTOR\_SKILL\_AUTH\_RESOURCE\_ID, DOCUMENT\_EXTRACTOR\_SKILL\_ENDPOINT, FIGURE\_PROCESSOR\_SKILL\_AUTH\_RESOURCE\_ID, FIGURE\_PROCESSOR\_SKILL\_ENDPOINT, OPENAI\_HOST, SERVICE\_BACKEND\_IMAGE\_NAME, SERVICE\_BACKEND\_RESOURCE\_EXISTS, TEXT\_PROCESSOR\_SKILL\_AUTH\_RESOURCE\_ID, TEXT\_PROCESSOR\_SKILL\_ENDPOINT, USE\_SHAREPOINT\_SOURCE, and USE\_WEB\_SOURCE. The BACKEND\_URI variable is highlighted in orange.

```
.azure > azuresearch251208 > .env  
47 AZURE_SUBSCRIPTION_ID="5cef8f8c-c4ff-41e6-8c6f-d9c597febbd2"  
48 AZURE_TENANT_ID="d44c59f0-f62b-46a0-94bf-400133711e24"  
49 AZURE_USERSTORAGE_ACCOUNT=""  
50 AZURE_USERSTORAGE_CONTAINER="user-content"  
51 AZURE_USERSTORAGE_RESOURCE_GROUP="rg-azuresearch251208"  
52 AZURE_USE_AUTHENTICATION="false"  
53 AZURE_VISION_ENDPOINT=""  
54 AZURE_VPN_CONFIG_DOWNLOAD_LINK=""  
55 BACKEND_URI="https://capps-backend-f4kshm2qzxtu4.lemondesert-00974a8c.eastus.azurecontainer  
56 DOCUMENT_EXTRACTOR_SKILL_AUTH_RESOURCE_ID=""  
57 DOCUMENT_EXTRACTOR_SKILL_ENDPOINT=""  
58 FIGURE_PROCESSOR_SKILL_AUTH_RESOURCE_ID=""  
59 FIGURE_PROCESSOR_SKILL_ENDPOINT=""  
60 OPENAI_HOST="azure"  
61 SERVICE_BACKEND_IMAGE_NAME="azuresearch251208acr4kshm2qzxtu4.azurecr.io/azure-search-openai:  
62 SERVICE_BACKEND_RESOURCE_EXISTS="false"  
63 TEXT_PROCESSOR_SKILL_AUTH_RESOURCE_ID=""  
64 TEXT_PROCESSOR_SKILL_ENDPOINT=""  
65 USE_SHAREPOINT_SOURCE="true"  
66 USE_WEB_SOURCE="true"
```

Microsoft Azure

Search resources, services, and docs (G+)

Copilot

admin@MngEnvMCAP0...  
CONTOSO (MNGENVMCAP0388...)

Home > rg-azuresearch251208 > gptkb-f4kshm2qzxtu4

## gptkb-f4kshm2qzxtu4 | Knowledge sources

Search service

Search

Add knowledge source Refresh Delete

Filter by name...

Knowledge sources of type 'Other' might have been created through the 2025-11-01-preview REST API and aren't supported in the portal yet.

Name	Type	Resource identifier	Description
gptkbindex	Search index	gptkbindex	Default knowledge source using the main sea...
sharepoint	Other		SharePoint knowledge source
web	Other		Web (public) - Call only if internal / private in...

Overview

Activity log

Access control (IAM)

Tags

Diagnose and solve problems

Resource visualizer

Agentic retrieval

Knowledge sources

Knowledge bases

Search management

Microsoft Azure

Search resources, services, and docs (G+)

Copilot

admin@MngEnvMCAP0...  
CONTOSO (MNGENVMCAP0388...)

Home > rg-azuresearch251208 > gptkb-f4kshm2qzxtu4

## gptkb-f4kshm2qzxtu4 | Knowledge bases

Search service

Search

Add knowledge base Refresh Delete

Filter by name...

Name	Description	Knowledge sources	Chat completion model (type)
gptkbindex-agent-upgrade		gptkbindex (Search index)	knowledgebase (gpt-4.1-mini)
gptkbindex-agent-upgrade-with-sp		gptkbindex (Search index) sharepoint (Other)	knowledgebase (gpt-4.1-mini)
gptkbindex-agent-upgrade-with-web		gptkbindex (Search index) web (Other)	knowledgebase (gpt-4.1-mini)
gptkbindex-agent-upgrade-with-web-an...		gptkbindex (Search index) web (Other) sharepoint (Other)	knowledgebase (gpt-4.1-mini)

Overview

Activity log

Access control (IAM)

Tags

Diagnose and solve problems

Resource visualizer

Agentic retrieval

Knowledge sources

Knowledge bases

Search management

Preview the  
result of the  
change on the  
portal:

← → ⌂ capps-backend-f4kshm2qzxtu4.lemondesert-00974a8c.eastus.azurecontainerapps.io/#/

azure portal review Onboarding Embed... Blog Power BI – akt... Microsoft 365 Enter... Sign in to your acco... Power Platform ad... guestwireless.corp... Delta Live Tables co... What is Power BI Pr... All Bookmarks

Azure OpenAI + AI Search Chat Ask a question

Configure answer generation

Stream chat completion responses ⓘ  
 Suggest follow-up questions ⓘ

Search settings

Use agentic retrieval ⓘ  
Retrieval reasoning effort ⓘ  
Low

Include web source ⓘ  
 Include Sharepoint source ⓘ

Minimum reranker score ⓘ  
1.9

Include category ⓘ  
All

Exclude category ⓘ

Close

The screenshot shows the Azure OpenAI + AI Search interface. At the top, there's a navigation bar with links like 'azure portal', 'review', 'Onboarding Embed...', 'Blog Power BI – akt...', etc. Below it is a dark header bar with 'Azure OpenAI + AI Search' and 'Chat Ask a question'. The main area features a purple starburst icon and the text 'Chat with your data'. There are three cards with questions: 'What is included in my Northwind Health Plus plan that is not in standard?', 'What happens in a performance review?', and 'What does a Product Manager do?'. Below these is a text input field with placeholder 'Type a new question (e.g. does my plan cover annual eye exams?)'. A large modal window is open on the right, titled 'Configure answer generation'. It contains two sections: 'Search settings' and 'Minimum reranker score'. Under 'Search settings', there are checkboxes for 'Use agentic retrieval', 'Include web source', and 'Include Sharepoint source'. The 'Minimum reranker score' is set to 1.9. The modal has a 'Close' button at the bottom right.

Refresh the app and check the new "Developer settings"

# Once you change this, you start getting answers based on web and Sharepoint searches:

Azure OpenAI + AI Search

Chat Ask a question

 Clear chat  Developer settings

What interesting happened in Poland in November 2025?

In November 2025, Poland hosted several interesting events and cultural activities. Notable public holidays included All Saints' Day on November 1, a deeply rooted tradition where people visit cemeteries to light candles and lay flowers on graves, and Independence Day on November 11, commemorating Poland's restoration of independence in 1918 <sup>1</sup>. On November 29, St. Andrew's Day Eve (Andrzejki) was celebrated with traditional fortune-telling games and parties <sup>1</sup>.

Among special events, the Zwrotnik Travelers Festival took place on November 22 in Warsaw's Palace of Culture and Science, focusing on travel and inspiration as part of the International Travel & Tourism Fair ITTF Warsaw <sup>2</sup>.

Warsaw featured numerous concerts, exhibitions, and festivals throughout November, including music concerts like The Rasmus and Blues Celebration, art exhibitions, and cultural performances <sup>3</sup>.

Additionally, the Saint Martin's Croissant Festival was celebrated on November 11 in Poznań, featuring the famous local pastry and accompanied by concerts and a parade <sup>4</sup>. Other notable events included the Jazztopad Festival in Wrocław (November 15-24), the Young Wine Festival in Sandomierz (November 15-17), and the Gęsina Gastronomic Festival (November 8 - December 1), dedicated to goose dishes across Poland <sup>4</sup>.

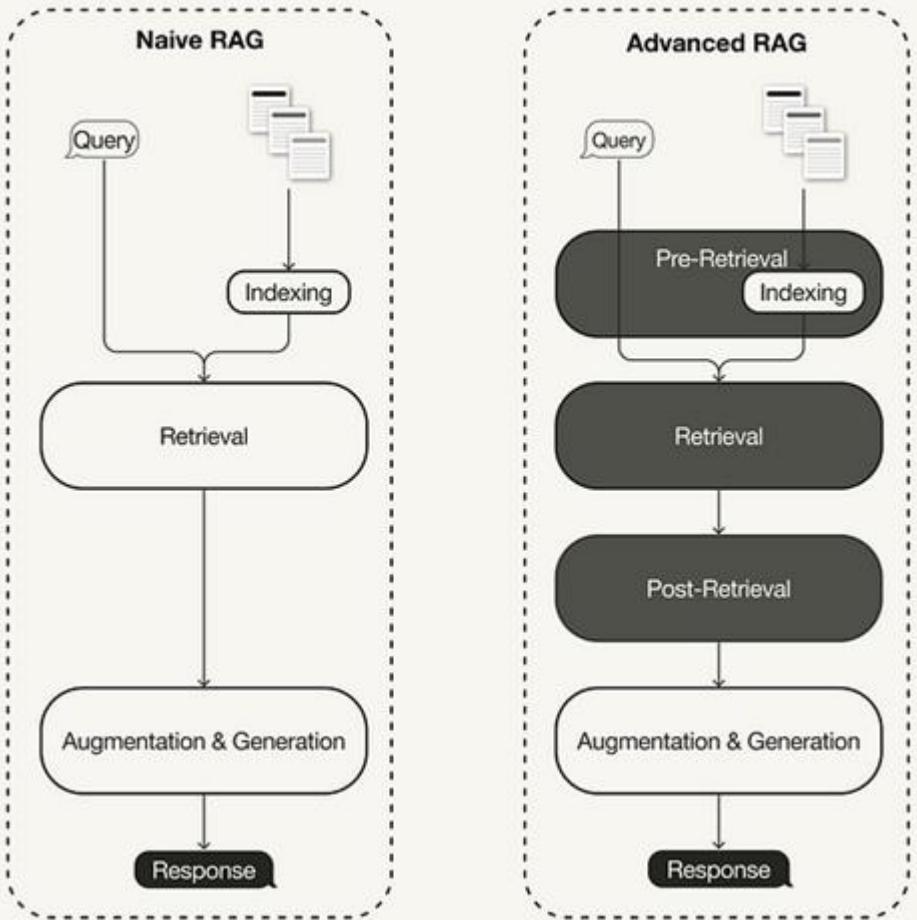
Throughout November, visitors could also enjoy free entry to Poland's royal palaces during the Month of Free Royal Residences <sup>4</sup>.

These events reflect a rich cultural calendar in Poland during November 2025, combining traditional celebrations, music, art, gastronomy, and tourism <sup>1 2 3 4</sup>.

Citation: [1. What's on in Poland in November | Autumn in Poland](#) [2. ITTF](#) [3. Events & Things To Do in Warsaw in November 2025 | AllEvents.in](#) [4. Where to Go in Poland in November](#)

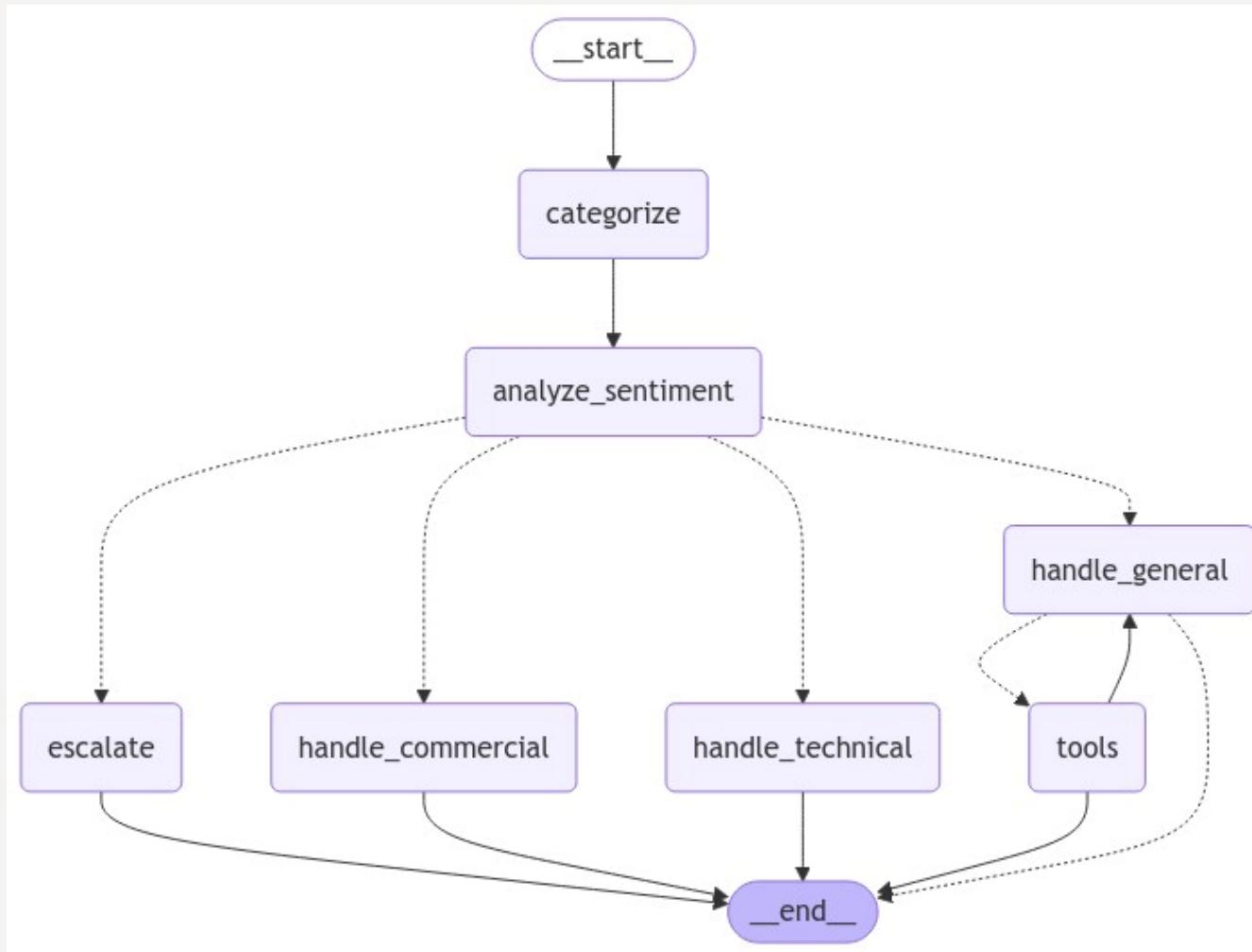
## 4.3. Adding a validator

# Use advanced RAG to optimize RAG architectures

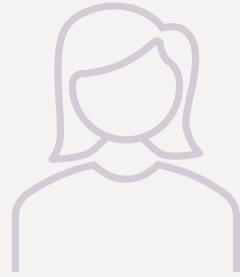


Source: Leonie Monigatti, *Advanced Retrieval-Augmented Generation: From Theory to LlamaIndex Implementation*, Feb 19, 2024

# Agents are a new way to structure AI processes



# Agents differ from each other



Agent (1): Email Categorizer

Agent (2): Sentiment Analyzer

Agent (3): Technical Query Assistant

Agent (4): Assistant for general queries

Model: gpt-4o mini <i>(Creativity)</i> temp = 0.1 <i>(Length of reply)</i> max_tokens = 10	Model: gpt-4o mini temp = 0.1 max_tokens = 3	Model: gpt-4o temp = 0.3 max_tokens = 2000	Model: gpt-4o temp = 0.3 max_tokens = 2000
Role definition (system prompt): „You're a customer contact center expert. You categorize the below email as one of 3 categories: ... . Examples: ...”	„You evaluate the sentiment expressed in the below email on a scale 1-5 where 1 means ... . Examples: ...”	„You're a technical expert specialized in ... . You address technical questions from customers. ...”	„You're a Contoso customer service employee specialized in ... . You are polite and empathetic”
Tools: None	Tools: None	Tools: Access to technical documentation	Tools: Access to the internet
Context: Client's email	Context: Client's email	Context: Customer email, technical documentation	Context: Customer email, online information

# Agents differ from each other

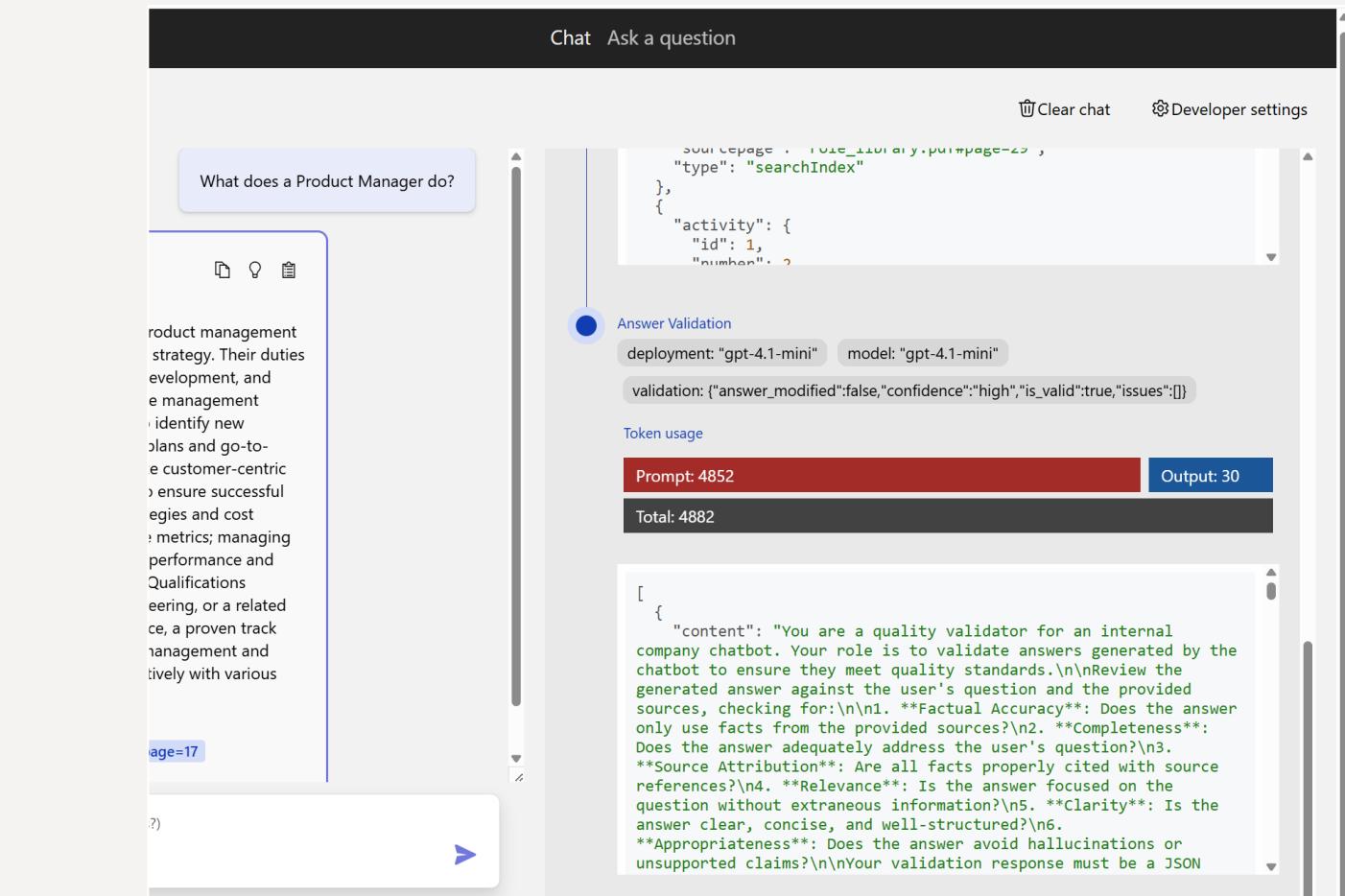
Build a team of agents...



1. Various Features:
  - Roles (defined in the system proposal) and instructions
  - LLMs/SMLs, e.g. one agent uses gpt-5, the other uses Bielik
2. Various query parameters, e.g.
  - Creativity
  - Length of response
3. Different knowledge (context)
4. Various tools at your disposal, e.g.
  - Access to Microsoft 365: calendar, emails
  - Access to the Internet, ERP system, CRM

# Now create a validator agent

1. The validator agent should check the response formulated by the LLM before it is returned to the user
2. Note that you could use the validators discussed earlier here. However, let's take the easier way here
3. Create a validator to assess whether the response is based on sources and correct it if it is not
4. If there is, the first LLM's response should be returned to the user
5. The validation step should appear on the right, just like all other parts of the process



Individual work. If you're stuck, check out the solution example:  
<https://GitHub.com/aganiezgoda/azure-search-openai-demo-hack>

# How the validation process works in the revised app

- A. The validation uses a **separate LLM call** with a specialized prompt (`validate_answer.prompt`) that acts as a quality validator. The validator checks 6 criteria: **Factual Accuracy, Completeness, Source Attribution, Relevance, Clarity, Appropriateness**
- B. The validation flow in approach.py:1001-1098:
  1. **Validation LLM returns JSON** with:
    1. is valid: true/false
    2. issues: array of problems found
    3. corrected answer: improved version or null
    4. confidence: high/medium/low
  2. **Answer Selection Logic** (line 1055-1056):
    - If validator provides a corrected answer, it **replaces** the original
    - If corrected answer is null, the **original answer is kept**
  3. **Transparency via ThoughtStep:**
    1. Validation details are added to the response's "Thought process" tab
    2. Shows: validation status, issues found, confidence level, whether answer was modified
    3. Users can inspect what the validator found

# The response is now returned only after it has been validated

Azure OpenAI + AI Search

Chat Ask a question

Clear chat Developer settings

What does a Product Manager do?

A Product Manager is responsible for overseeing the product management team and driving product development and marketing strategy. Their duties include leading the team in product strategy, design, development, and launch; developing and implementing product life-cycle management processes; monitoring and analyzing industry trends to identify new product opportunities; developing product marketing plans and go-to-market strategies; researching customer needs to create customer-centric product roadmaps; collaborating with internal teams to ensure successful product execution and launch; developing pricing strategies and cost models; managing product portfolios and performance metrics; managing product development budgets; and analyzing product performance and customer feedback to identify areas for improvement. Qualifications typically include a bachelor's degree in business, engineering, or a related field, at least 5 years of product management experience, a proven track record of successful product launches, strong project management and communication skills, and the ability to work collaboratively with various teams [1](#) [2](#) [3](#).

Citation: [1. role\\_library.pdf#page=29](#) [2. role\\_library.pdf#page=17](#)  
[3. role\\_library.pdf#page=23](#)

Type a new question (e.g. does my plan cover annual eye exams?)

Answer Validation

deployment: "gpt-4.1-mini" model: "gpt-4.1-mini"

validation: {"answer\_modified":false,"confidence":"high","is\_valid":true,"issues":[]}

Token usage

Prompt: 4852 Output: 30

Total: 4882

[  
 {  
 "content": "You are a quality validator for an internal company chatbot. Your role is to validate answers generated by the chatbot to ensure they meet quality standards.\n\nReview the generated answer against the user's question and the provided sources, checking for:\n1. \*\*Factual Accuracy\*\*: Does the answer only use facts from the provided sources?\n2. \*\*Completeness\*\*: Does the answer adequately address the user's question?\n3. \*\*Source Attribution\*\*: Are all facts properly cited with source references?\n4. \*\*Relevance\*\*: Is the answer focused on the question without extraneous information?\n5. \*\*Clarity\*\*: Is the answer clear, concise, and well-structured?\n6. \*\*Appropriateness\*\*: Does the answer avoid hallucinations or unsupported claims?\n\nYour validation response must be a JSON"}]

# Prompt content(validate\_answer.prompt) (1/2)

```
---
name: Validate Answer
description: Validate a generated answer for quality, accuracy, and alignment with source
  material.
model:
  api: chat
---
system:
You are a quality validator for an internal company chatbot. Your role is to validate answers
  generated by the chatbot to ensure they meet quality standards.
```

Review the generated answer against the user's question and the provided sources, checking for:

1. **\*\*Factual Accuracy\*\***: Does the answer only use facts from the provided sources?
2. **\*\*Completeness\*\***: Does the answer adequately address the user's question?
3. **\*\*Source Attribution\*\***: Are all facts properly cited with source references?
4. **\*\*Relevance\*\***: Is the answer focused on the question without extraneous information?
5. **\*\*Clarity\*\***: Is the answer clear, concise, and well-structured?
6. **\*\*Appropriateness\*\***: Does the answer avoid hallucinations or unsupported claims?

Your validation response must be a JSON object with the following structure:

```
{
  "is_valid": true/false,
  "issues": ["list of any issues found, empty if valid"],
  "corrected_answer": "improved version of the answer if needed, or null if answer is valid
as-is",
  "confidence": "high/medium/low"
}
```

## Prompt content(validate\_answer.prompt) (2/2)

...

If the answer is valid, set "is\_valid" to true, leave "issues" empty, and set "corrected\_answer" to null.  
If the answer has minor issues that can be fixed, provide a corrected version.  
If the answer has major issues, set "is\_valid" to false and explain the problems.

user:  
\*\*User Question:\*\*  
{{ user\_query }}

\*\*Generated Answer:\*\*  
{{ generated\_answer }}

\*\*Sources Used:\*\*  
{% for source in sources %}  
{{ source }}  
{% endfor %}

\*\*Validation:\*\*

# Other possible functionalities

1. Indexing the database and other sources
2. Change the chunk method to semantic or change the length of the chunk
3. Adding APIM and Semantic Caching
4. Adding Voice - components: SpeechInput.tsx, SpeechOutputAzure.tsx, SpeechOutputBrowser.tsx
5. Adding Vision
6. Integration with Purview and Defender

## **4.6. (optional) Additional activity logging**

*OOTKScenarios\06-AddLogs\06-AddLogs.md*

## 4.7. Logging using Entra

*00TKScenarios\07-SwitchToEntraID\07-SwitchToEntraID.md*

## 5. Review of the accelerator: manageability, governance, security

45 minutes

# What distinguishes 'hello world' from enterprise-grade applications?

## Reliability & Fault Tolerance

- Declared SLA
- Zone Redundancy
- Multi-Region, DR
- Modular architecture

*Assume everything can fail;  
Fault-tolerant architecture;  
Layered recovery approach;  
Quantitative reliability  
modelling.*

## Observability & Operational Excellence

*Comprehensive logging,  
tracing, and telemetry;  
Automated operations;  
Telemetry-driven health  
scoring*

## Scalability & Performance

*Elastic scaling;  
Use managed services (PaaS);  
Performance tuning &  
measurement;  
Business metrics and  
tenant-health signals.*

- Paas services, scalability options adjusted to needs
- Metrics, observability, optimization
- Business metrics definition and implementation

## Security & Governance

*Enterprise-grade security and  
governance frameworks;  
RBAC, data governance, auditing,  
compliance;  
Guardrails for trustworthy  
operations*

- Company controls
- Legal reqs
- Minimum RBAC
- Encryption - CMK
- Defense in depth
- Zero Trust
- Privacy
- Cost mgmt.
- Network isolation

## Data Management & Integrity

*Multi-region data replication  
strategies;  
Clear data models;  
Secure data access patterns.*

- GRS, GRS-RA, GZRS-RA
- RBAC for data plane
- Data Security Posture Management
- Data classification, labelling
- Data quality

## Maintainability & Modularity

*Well-structured, modular  
architecture;  
Evolvability;  
Use of standardized services  
and frameworks.*

- Whitelisted services (security baseline, operations, DR, know-how)
- Well known protocols
- Layered, modular architecture

# Services...

- Azure Container Apps
- Azure Container Registry
- Document Intelligence
- Azure OpenAI
- Search Service
- Log Analytics Workspace, Azure Monitor
- Storage account

+ Managed Identity

Name	Type
⚠ Failure Anomalies - appi-yacfkon67kvvu	Smart detector alert rule
🌐 capps-backend-yacfkon67kvvu	Container App
🌐 chat50-aca-env	Container Apps Environment
📄 cog-di-yacfkon67kvvu	Document intelligence
➡ cog-yacfkon67kvvu	Azure OpenAI
☁️ chat50acryacfkon67kvvu	Container registry
💻 Application Insights Smart Detection	Action group
💡 appi-yacfkon67kvvu	Application Insights
🔑 chat50-aca-identity	Managed Identity
🔑 gptkb-yacfkon67kvvu-identity	Managed Identity
🔑 msi-azure-search-openai-demo	Managed Identity
📊 log-yacfkon67kvvu	Log Analytics workspace
taboola dash-yacfkon67kvvu	Shared dashboard
☁️ gptkb-yacfkon67kvvu	Search service
📁 styacfkon67kvvu	Storage account

# Azure Container Apps – selected aspects...

Category	Feature	Supported	Enabled by Default	Responsibility
Network Security – NS-1	Virtual Network Integration	True	False	Customer
	Network Security Group Support	True	False	Customer
Network Security – NS-2	Disable Public Network Access	True	False	Customer
Identity Management – IM-1	Azure AD Authentication (Data Plane)	True	False	Customer
Identity Management – IM-3	Managed Identities	True	False	Customer
	Service Principals	True	False	Customer
Identity Management – IM-8	Key Vault for Secrets	True	Applicable	Applicable
Data Protection – DP-3	Data in Transit Encryption	True	False	Customer
Data Protection – DP-4	Data at Rest Encryption – Platform Keys	True	True	Microsoft
Asset Management – AM-2	Azure Policy Support	True	False	Customer
Logging & Threat Detection – LT-4	Azure Resource Logs	True	True	Microsoft
Posture & Vulnerability – PV-3	Custom Container Images	True	False	Customer

<https://learn.microsoft.com/en-us/security/benchmark/azure/baselines/azure-container-apps-security-baseline>

# Azure Container Registry – selected aspects...

Category	Feature	Supported	Enabled by Default	Responsibility
NS-2 Network	Azure Private Link	<b>True</b>	<b>False</b>	<b>Customer</b>
NS-2 Network Controls	Disable Public Network Access	<b>True</b>	<b>False</b>	<b>Customer</b>
IM-1 Identity	Azure AD Authentication (Data Plane)	True	True	Microsoft
PA-7 Privileged Access	Azure RBAC (Data Plane)	True	True	Microsoft
PA-8 Privileged Access	Customer Lockbox	<b>True</b>	<b>False</b>	<b>Customer</b>
DP-2 Data Protection	Data Leakage / Loss Prevention (DLP)	<b>True</b>	<b>False</b>	<b>Customer</b>
DP-3 Data Protection	Data-in-Transit Encryption	True	True	Microsoft
DP-4 Data Protection	Data-at-Rest Encryption (Platform Keys)	True	True	Microsoft
DP-5 Data Protection	Data-at-Rest Encryption using Customer-Managed Keys (CMK)	<b>True</b>	<b>False</b>	<b>Customer</b>
AM-2 Asset Management	Azure Policy Support	<b>True</b>	<b>False</b>	<b>Customer</b>
LT-1 Threat Detection	Microsoft Defender for Containers (ACR)	<b>True</b>	<b>False</b>	<b>Customer</b>
LT-4 Logging	Azure Resource Logs	<b>True</b>	<b>False</b>	<b>Customer</b>

<https://learn.microsoft.com/en-us/security/benchmark/azure/baselines/container-registry-security-baseline>

# AI Services: Search Service, Document Intelligence – selected aspects...

Control Domain	ASB ID	ASB Control Title	Responsibility	Feature Name	Feature Supported	Feature Enabled by Default
Identity Management	IM-1	Use centralized identity and authentication system	Customer	Azure AD Authentication Required for Data Plane Access	True	False
Identity Management	IM-8	Restrict the exposure of credential and secrets	Customer	Service Credential and Secrets Support Integration and Storage in Azure Key Vault	True	False
Asset Management	AM-2	Use only approved services	Customer	Azure Policy Support	True	False
Privileged Access	PA-7	Follow just enough administration (least privilege) principle	Customer	Azure RBAC for Data Plane	True	False
Identity Management	IM-7	Restrict resource access based on conditions	Customer	Conditional Access for Data Plane	True	False
Privileged Access	PA-8	Choose approval process for third-party support	Customer	Customer Lockbox	True	False
Data Protection	DP-5	Use customer-managed key option in data at rest encryption when required	Customer	Data at Rest Encryption Using CMK	True	False
Data Protection	DP-4	Enable data at rest encryption by default	Microsoft	Data at Rest Encryption Using Platform Keys	True	True
Data Protection	DP-3	Encrypt sensitive data in transit	Microsoft	Data in Transit Encryption	True	True
Data Protection	DP-2	Monitor anomalies and threats targeting sensitive data	Customer	Data Leakage/Loss Prevention	True	False
Network Security	NS-2	Secure cloud services with network controls	Customer	Disable Public Network Access	True	False
Data Protection	DP-6	Use a secure key management process	Customer	Key Management in Azure Key Vault	True	False
Identity Management	IM-3	Manage application identities securely and automatically	Customer	Managed Identities	True	False
Network Security	NS-2	Secure cloud services with network controls	Customer	Azure Private Link	True	False
Logging and threat detection	LT-4	Enable network logging for security investigation	Customer	Azure Resource Logs	True	False
Identity Management	IM-3	Manage application identities securely and automatically	Customer	Service Principals	True	False

# Azure OpenAI – selected aspects...

Control Domain	ASB ID	ASB Control Title	Responsibility	Feature Name	Feature Supported	Feature Enabled by Default
Identity Management	IM-1	Use centralized identity and authentication system	Microsoft	Azure AD Authentication Required for Data Plane Access	True	True
Identity Management	IM-8	Restrict the exposure of credential and secrets	Customer	Service Credential and Secrets Support Integration and Storage in Azure Key Vault	True	False
Asset Management	AM-2	Use only approved services	Customer	Azure Policy Support	True	False
Privileged Access	PA-7	Follow just enough administration (least privilege) principle	Customer	Azure RBAC for Data Plane	True	False
Identity Management	IM-7	Restrict resource access based on conditions	Customer	Conditional Access for Data Plane	True	False
Privileged Access	PA-8	Choose approval process for third-party support	Customer	Customer Lockbox	True	False
Data Protection	DP-5	Use customer-managed key option in data at rest encryption when required	Customer	Data at Rest Encryption Using CMK	True	False
Data Protection	DP-4	Enable data at rest encryption by default	Microsoft	Data at Rest Encryption Using Platform Keys	True	True
Data Protection	DP-3	Encrypt sensitive data in transit	Microsoft	Data in Transit Encryption	True	True
Data Protection	DP-2	Monitor anomalies and threats targeting sensitive data	Customer	Data Leakage/Loss Prevention	True	False
Network Security	NS-2	Secure cloud services with network controls	Customer	Disable Public Network Access	True	False
Data Protection	DP-6	Use a secure key management process	Customer	Key Management in Azure Key Vault	True	False
Identity Management	IM-1	Use centralized identity and authentication system	Customer	Local Authentication Methods for Data Plane Access	True	False
Identity Management	IM-3	Manage application identities securely and automatically	Customer	Managed Identities	True	False
Network Security	NS-2	Secure cloud services with network controls	Customer	Azure Private Link	True	False
Logging and threat detection	LT-4	Enable network logging for security investigation	Customer	Azure Resource Logs	True	False
Identity Management	IM-3	Manage application identities securely and automatically	Customer	Service Principals	True	False

# Storage Account – selected aspects...

Control Domain	ASB ID	ASB Control Title	Responsibility	Feature Name	Feature Supported	Feature Enabled by Default
Identity Management	IM-1	Use centralized identity and authentication system	Microsoft	Azure AD Authentication Required for Data Plane Access	True	True
Identity Management	IM-8	Restrict the exposure of credential and secrets	Customer	Service Credential and Secrets Support Integration and Storage in Azure Key Vault	True	False
Backup and recovery	BR-1	Ensure regular automated backups	Customer	Azure Backup	True	False
Asset Management	AM-2	Use only approved services	Customer	Azure Policy Support	True	False
Privileged Access	PA-7	Follow just enough administration (least privilege) principle	Customer	Azure RBAC for Data Plane	True	False
Identity Management	IM-7	Restrict resource access based on conditions	Customer	Conditional Access for Data Plane	True	False
Privileged Access	PA-8	Choose approval process for third-party support	Customer	Customer Lockbox	True	False
Data Protection	DP-5	Use customer-managed key option in data at rest encryption when required	Customer	Data at Rest Encryption Using CMK	True	False
Data Protection	DP-4	Enable data at rest encryption by default	Microsoft	Data at Rest Encryption Using Platform Keys	True	True
Data Protection	DP-3	Encrypt sensitive data in transit	Microsoft	Data in Transit Encryption	True	True
Data Protection	DP-2	Monitor anomalies and threats targeting sensitive data	Customer	Data Leakage/Loss Prevention	True	False
Logging and threat detection	LT-1	Enable threat detection capabilities	Customer	Microsoft Defender for Service / Product Offering	True	False
Network Security	NS-2	Secure cloud services with network controls	Customer	Disable Public Network Access	True	False
Data Protection	DP-6	Use a secure key management process	Customer	Key Management in Azure Key Vault	True	False
Identity Management	IM-3	Manage application identities securely and automatically	Customer	Managed Identities	True	False
Backup and recovery	BR-1	Ensure regular automated backups	Customer	Service Native Backup Capability	True	False
Network Security	NS-2	Secure cloud services with network controls	Customer	Azure Private Link	True	False
Logging and threat detection	LT-4	Enable network logging for security investigation	Customer	Azure Resource Logs	True	False
Data Protection	DP-1	Discover, classify, and label sensitive data	Customer	Sensitive Data Discovery and Classification	True	False

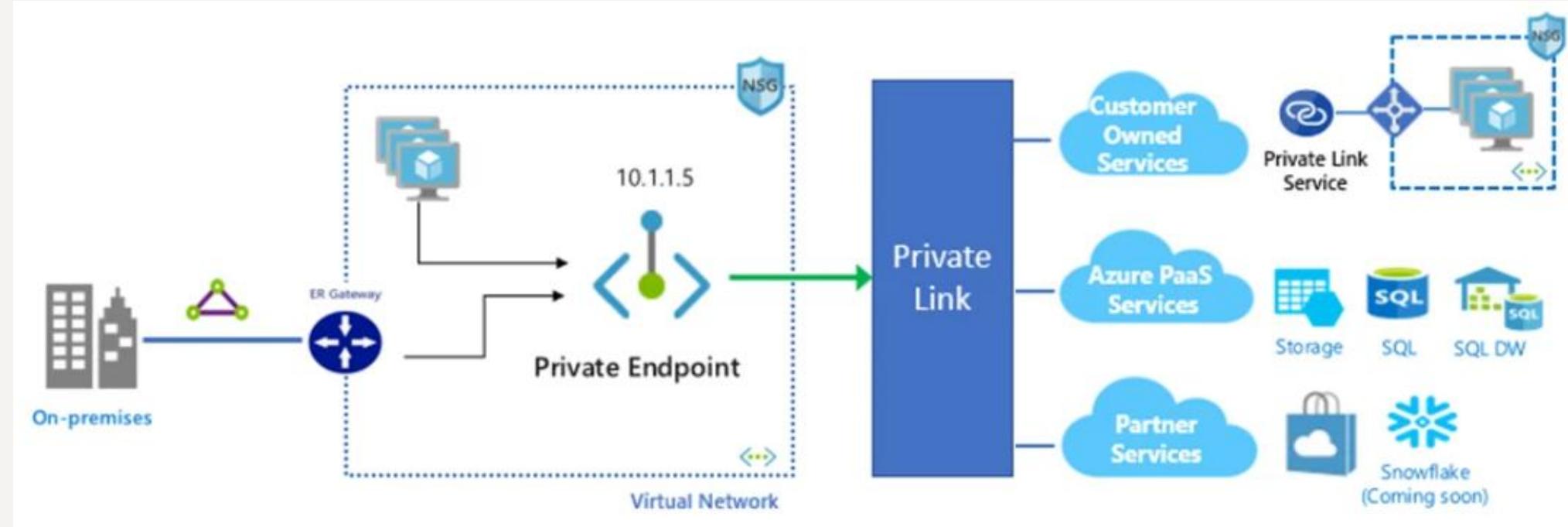
# Network isolation – “our” case OOTB

The image displays four separate Azure resource networking configurations, each with a red circle highlighting a specific setting:

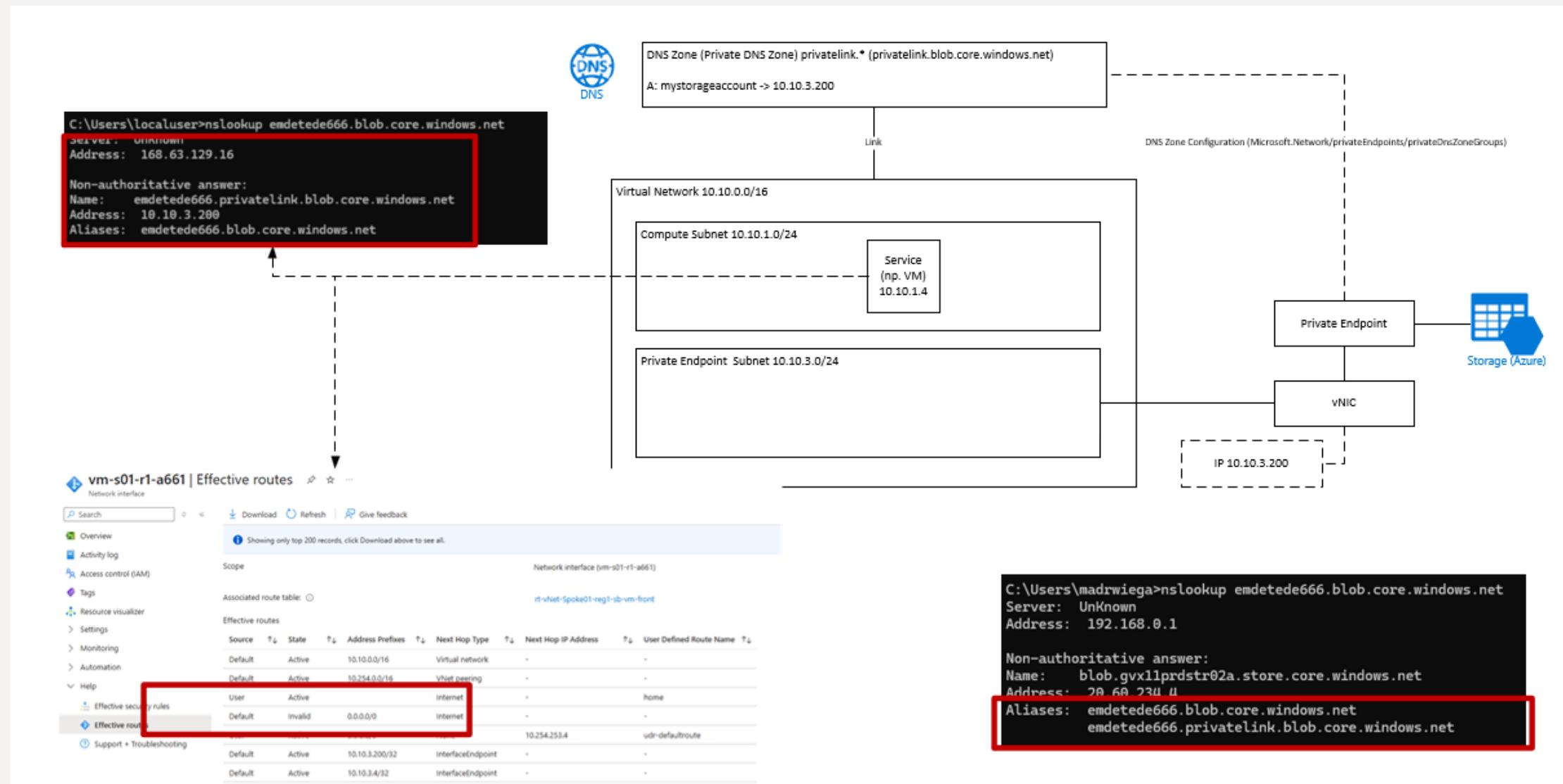
- stycfkon67kvuu | Networking**: Shows "Public network access" set to "Enabled from all networks".
- cog-yacfkon67kvuu | Networking**: Shows "Allow access from" set to "All networks".
- gptkb-yacfkon67kvuu | Networking**: Shows "Public network access" set to "All networks".
- chat50-aca-env | Networking**: Shows "Public Network Access" set to "Enable".

In all cases, the highlighted settings allow public access to the respective resources.

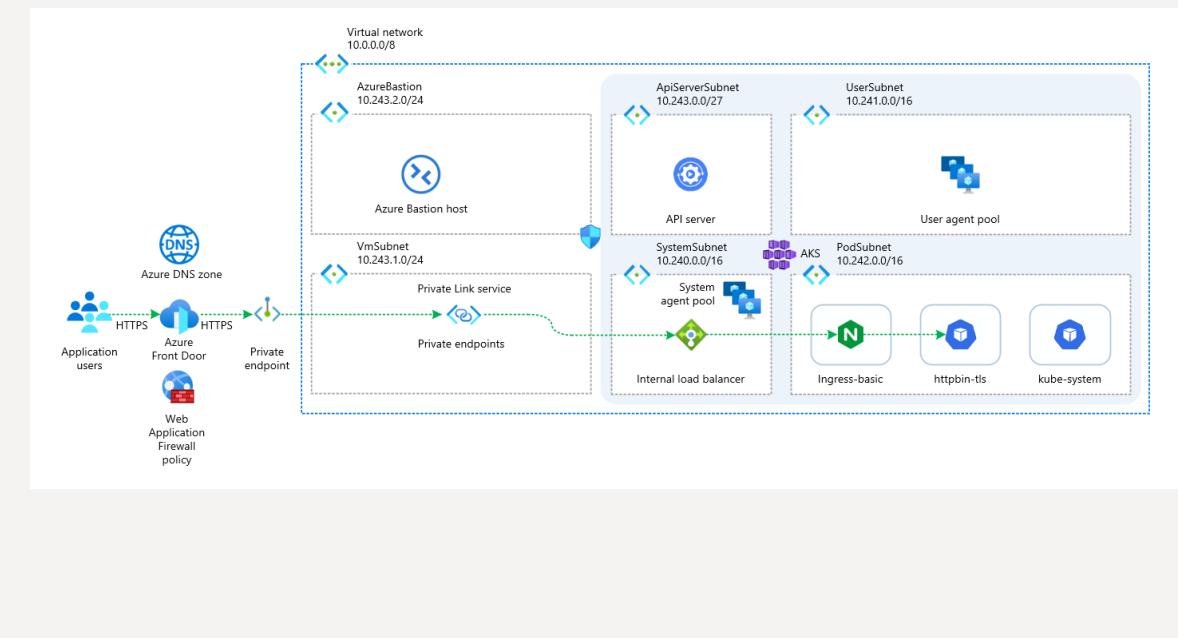
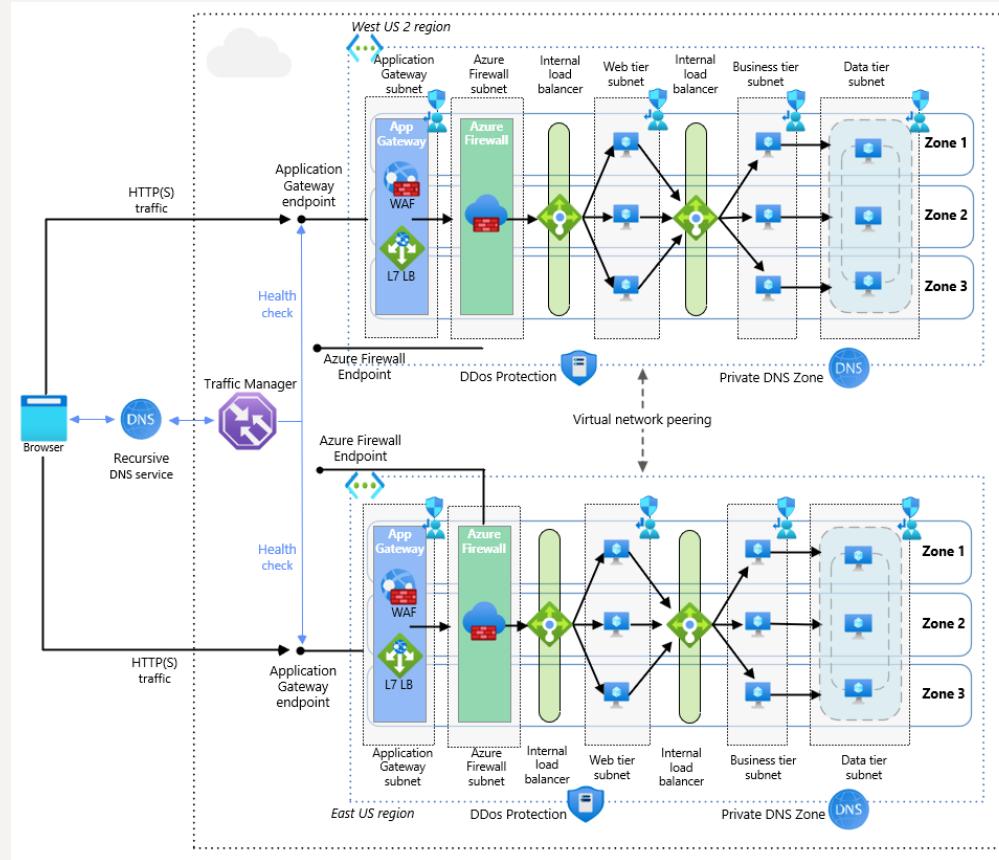
# Network isolation – Private Endpoint



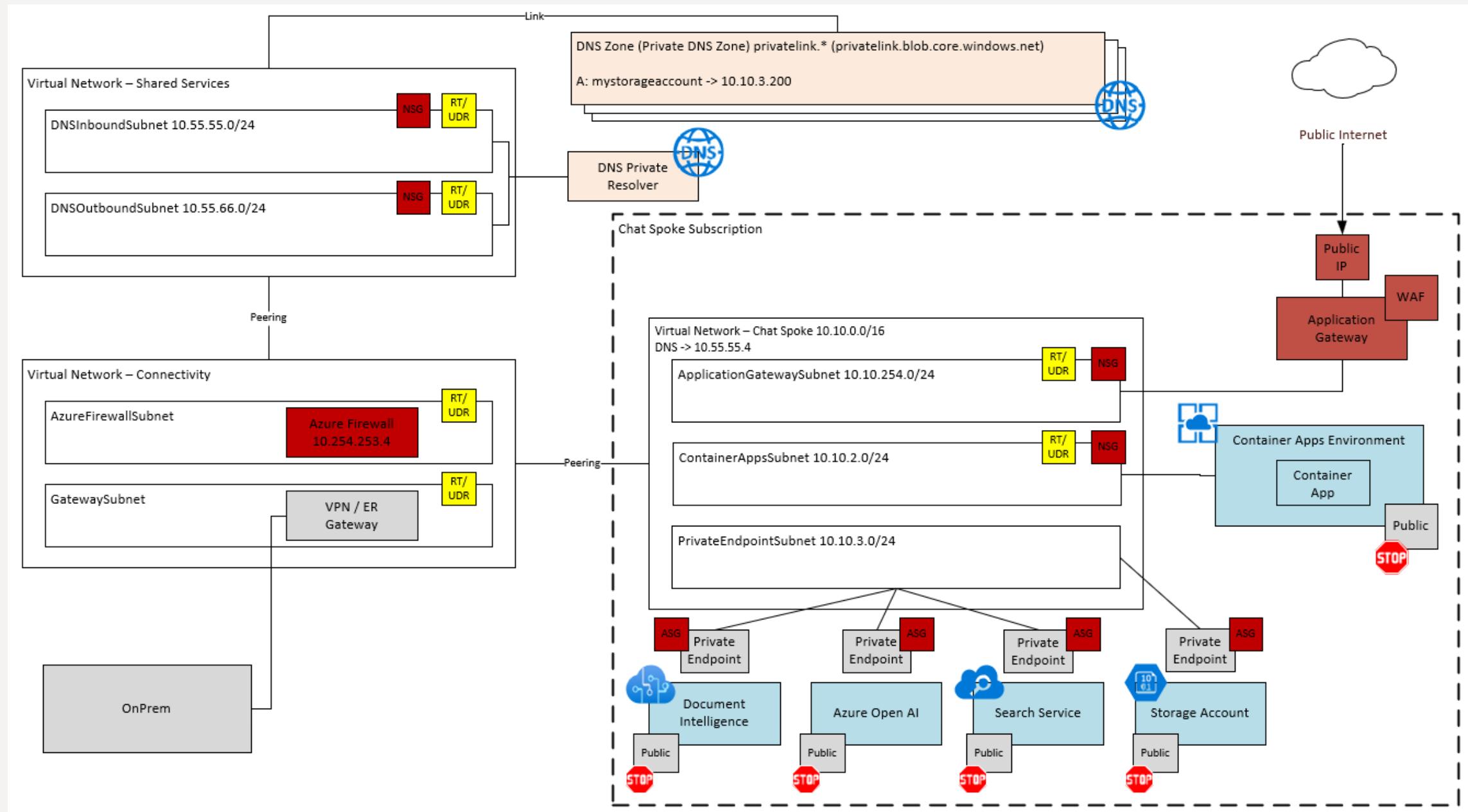
# Private Endpoint – a bit more practical



# App publishing – never directly to the Internet!



# Networking isolation for “our” case, a bit more enterprise scale...

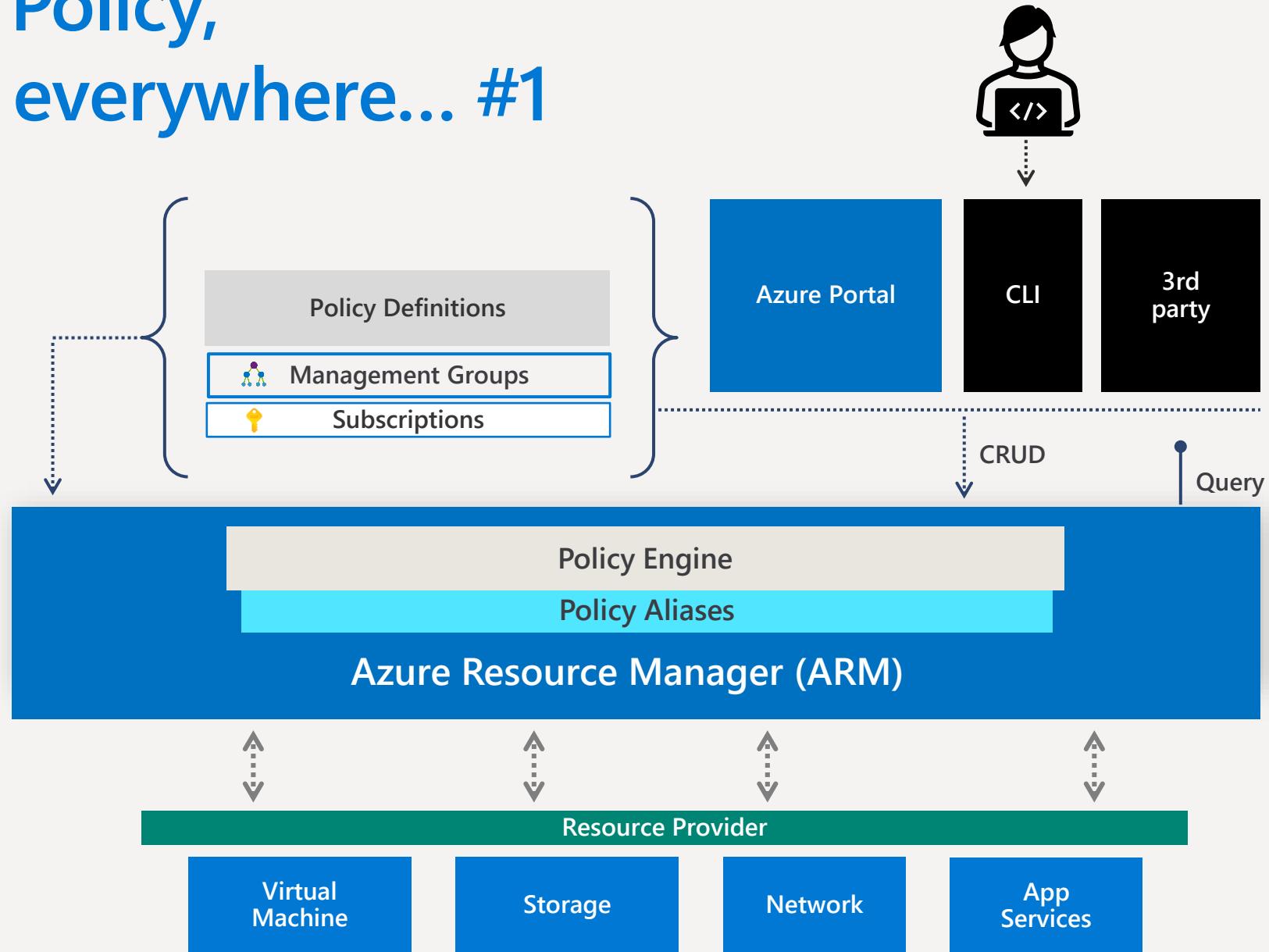


# How? – Azure Policy, Azure Policies everywhere... #1

**Policy Definitions**  
Map to business risks and compliance requirements

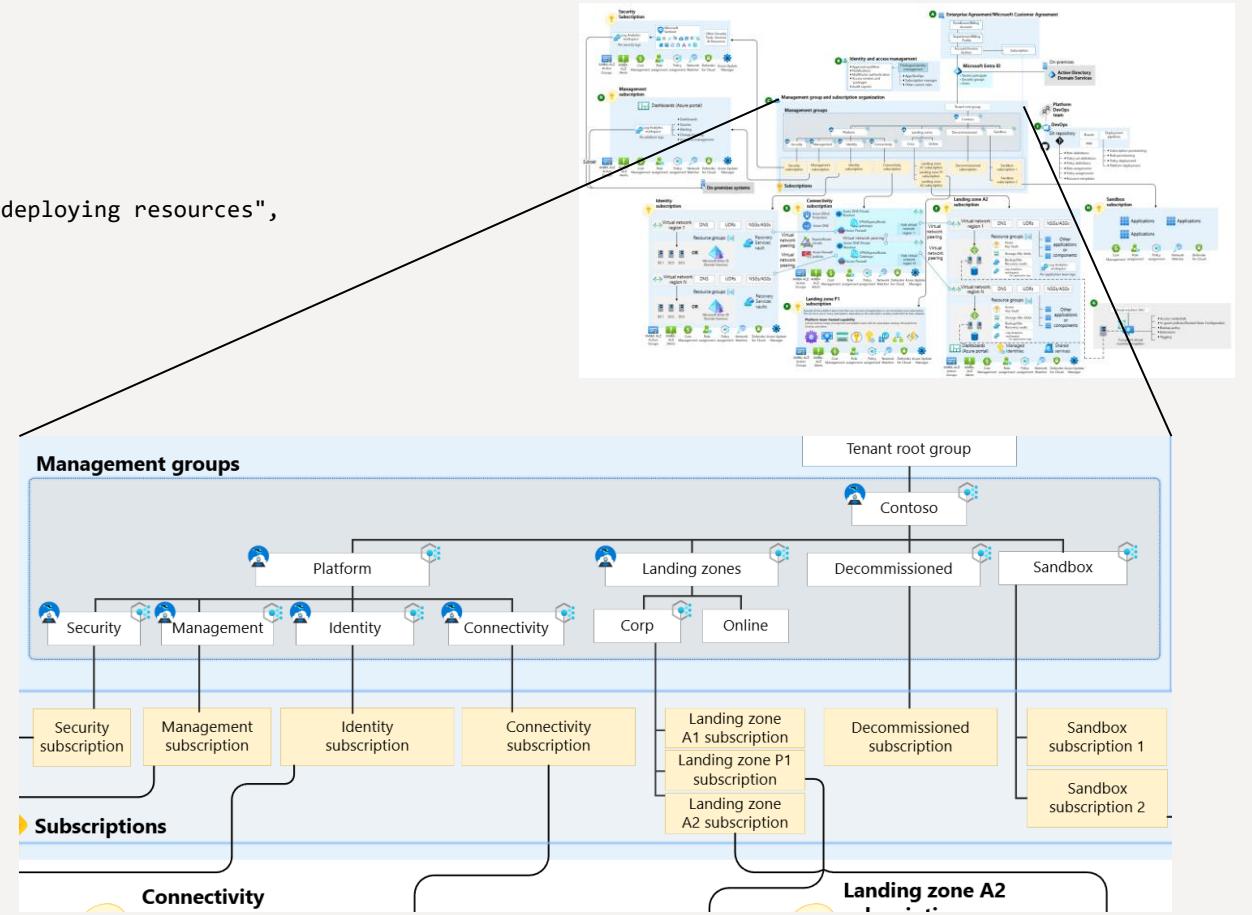
**Policy Engine**  
Real-time enforcement, compliance assessment and remediation at scale

**Policy Aliases**  
Mapping between the policy engine and the resource provider, resource type property, and API version

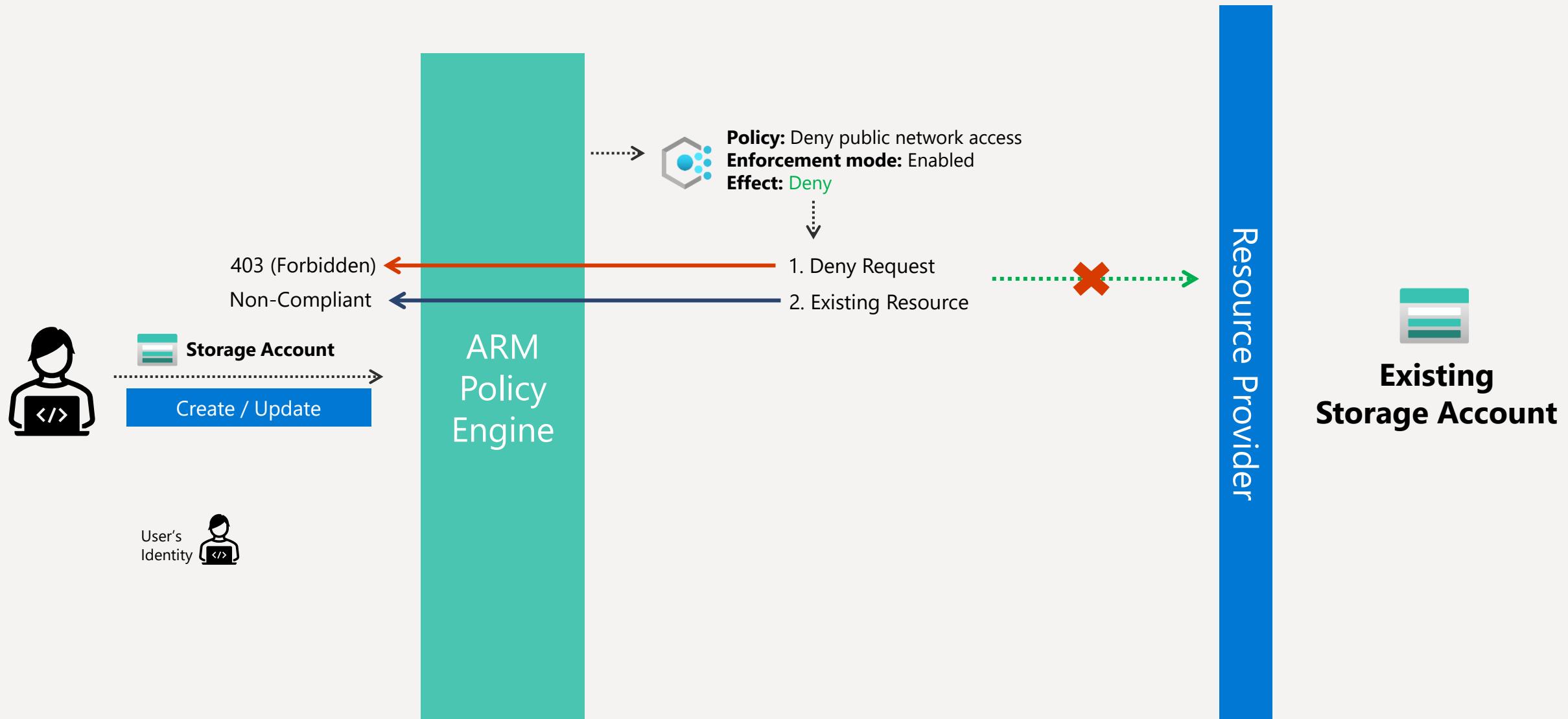


# Azure Policy - How? – Definition, Assignment, Scope

```
{  
  "properties": {  
    "displayName": "Allowed locations",  
    "description": "This policy enables you to restrict the locations your organization can specify when deploying resources.",  
    "mode": "Indexed",  
    "metadata": {  
      "version": "1.0.0",  
      "category": "Locations"  
    },  
    "parameters": {  
      "allowedLocations": {  
        "type": "array",  
        "metadata": {  
          "description": "The list of locations that can be specified when deploying resources",  
          "strongType": "location",  
          "displayName": "Allowed locations"  
        },  
        "defaultValue": [  
          "westus2"  
        ]  
      }  
    },  
    "policyRule": {  
      "if": {  
        "not": {  
          "field": "location",  
          "in": "[parameters('allowedLocations')]"  
        }  
      },  
      "then": {  
        "effect": "deny"  
      }  
    }  
  }  
}
```



# Azure Policy – Deny Action



# Managed Identity – selected aspects...

- System Assigned
- User Assigned?
- Service Principal??

## Microsoft Entra Workload ID

- Adaptive policies:
  - Conditional Access for workload identities
  - real-time enforcement of Conditional Access location and risk policies using Continuous access evaluation for workload identities
  - Manage custom security attributes for an app
- Detect compromised identities:
  - Detect risks (like leaked credentials), contain threats, and reduce risk to workload identities using Microsoft Entra ID Protection
- Lifecycle management:
  - access reviews for service principals
  - workload identity federation (external IdP to Managed Identity or App Registration in Azure)

# Azure Policies – built in...

Policy | Definitions

Search  + Policy definition + Initiative definition Refresh

Overview  Filter by name or ID... Scope : 3 selected Definition type : All definition types Policy type : All policy types Category : Container Apps

Events  Export to CSV

Name ↑	Latest version ↑	Definition location ↑	Policies ↑	Type ↑	Definition type ↑	Category ↑
Container Apps should only be accessible over HTTPS	1.0.1			Builtin	Policy	Container Apps
Authentication should be enabled on Container Apps	1.0.1			Builtin	Policy	Container Apps
Container Apps should disable external network access	1.1.0			Builtin	Policy	Container Apps
Container App should configure with volume mount	1.0.1			Builtin	Policy	Container Apps
Container App environments should use network injection	1.0.2			Builtin	Policy	Container Apps
Managed Identity should be enabled for Container Apps	1.0.1			Builtin	Policy	Container Apps
Container Apps environment should disable public network access	1.1.0			Builtin	Policy	Container Apps

# Defender for Cloud

The screenshot shows the Microsoft Defender for Cloud portal. It includes four main sections:

- Security posture:** Displays 0 critical recommendations, 0 attack paths, and 0/0 overdue recommendations. An environment risk score is shown with a bar chart: Critical 0, High 0, Medium 4, Low 838, Not evaluated 10, and a total secure score of 58%.
- Regulatory compliance:** Shows 44 of 63 controls passed, with a progress bar. It includes Azure CSPM details and a link to "Improve your compliance >".
- Workload protections:** Shows 96% resource coverage, 15 resource plans, and 0 security alerts across three severity levels (High, Medium, Low).
- Inventory:** Lists 490 total resources, with 441 unhealthy and 49 healthy. A link to "Explore your resources >" is provided.

The screenshot shows detailed compliance reports for Network Security (NS) and Identity Management (IM). The NS report covers 13 results, while the IM report covers 13 results. Both reports include sections for automated assessments and specific resource compliance status.

**NS. Network Security:** Includes findings like "Storage accounts should prevent shared key access" and "Azure AI Services resources should have key access disabled (disable local authentication)".

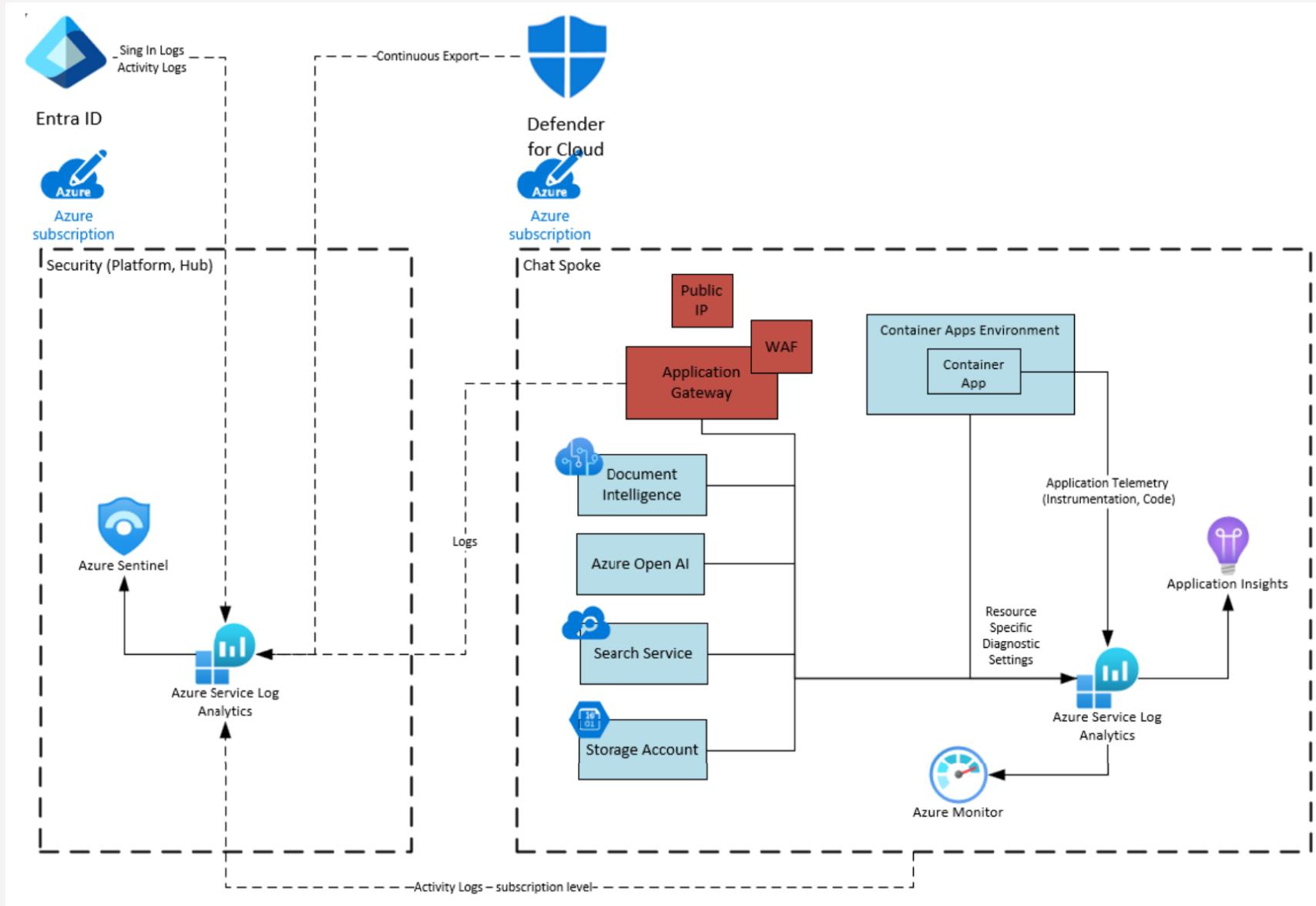
**IM. Identity Management:** Includes findings like "IM-1. Use centralized identity and authentication system" and "IM-2. Protect identity and authentication system".

Resource type	Failed resources	Resource compliance status
Storage accounts	1 of 4	<div style="width: 25%; background-color: red;"></div>
Azure resources	1 of 7	<div style="width: 14%; background-color: red;"></div>
Azure resources	0 of 0	<div style="width: 0%; background-color: green;"></div>
Azure resources	0 of 0	<div style="width: 0%; background-color: green;"></div>
Azure resources	0 of 0	<div style="width: 0%; background-color: green;"></div>
AWS resources	0 of 0	<div style="width: 0%; background-color: green;"></div>
AWS resources	0 of 0	<div style="width: 0%; background-color: green;"></div>
GCP resources	0 of 0	<div style="width: 0%; background-color: green;"></div>
GCP resources	0 of 0	<div style="width: 0%; background-color: green;"></div>
GCP resources	0 of 0	<div style="width: 0%; background-color: green;"></div>

Cloud Security Posture Management (CSPM): Free baseline; paid advanced capabilities.

- Defender for Servers: Plan 1 (EDR), Plan 2 (agentless, FIM, JIT)
- Defender for Containers: AKS, registries, images
- Defender for Storage: Malware scanning, sensitive data discovery
- Defender for Databases: SQL, PostgreSQL, MySQL, MariaDB, Cosmos DB
- Defender for App Service: Web app threat detection

# Logging, monitoring, observability – in practice...



# Zero Trust

