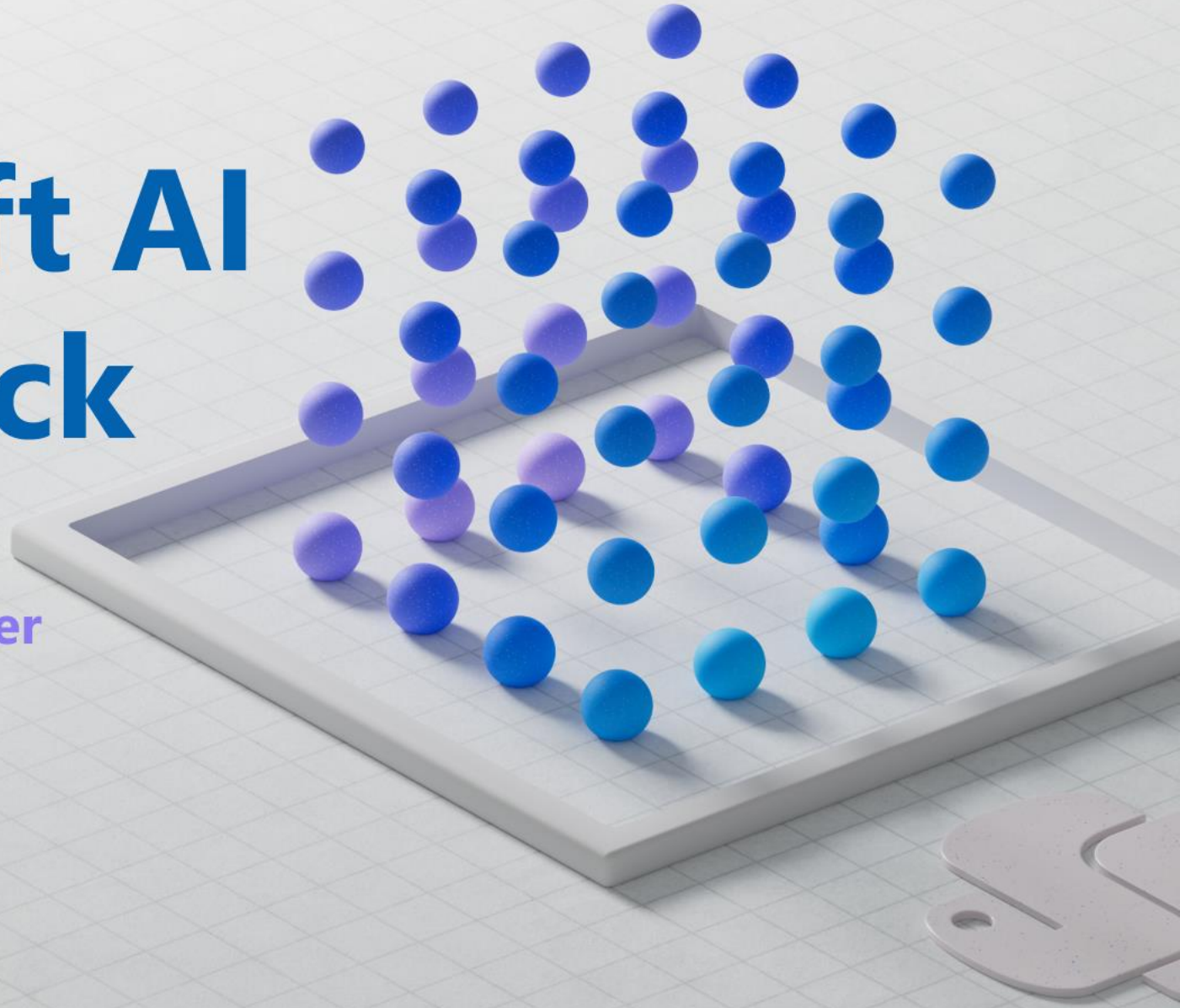
January 29th - February 12th

The Microsoft AI Chat App Hack



Build, innovate, and #HackTogether

aka.ms/hacktogether/chatapp

The AI Chat App Hack

Microsoft

January 29th - February 12th

29th — Building a RAG Chat App in Python

30th — Customizing your RAG Chat App

31st — Azure AI Search Best Practices

1st — GPT-4 with Vision

2nd — HACK HACK HACK

3rd — HACK HACK HACK

4th — HACK HACK HACK

5th — AM: RAG Chat Web Components

PM: Access Control in RAG Chat Apps

6th — Evaluating a RAG Chat App

7th — RAG Chat Special Topic

8th — Continuous Deployment of your Chat App

9th — HACK HACK HACK

10th — HACK

11th — HACK HACK

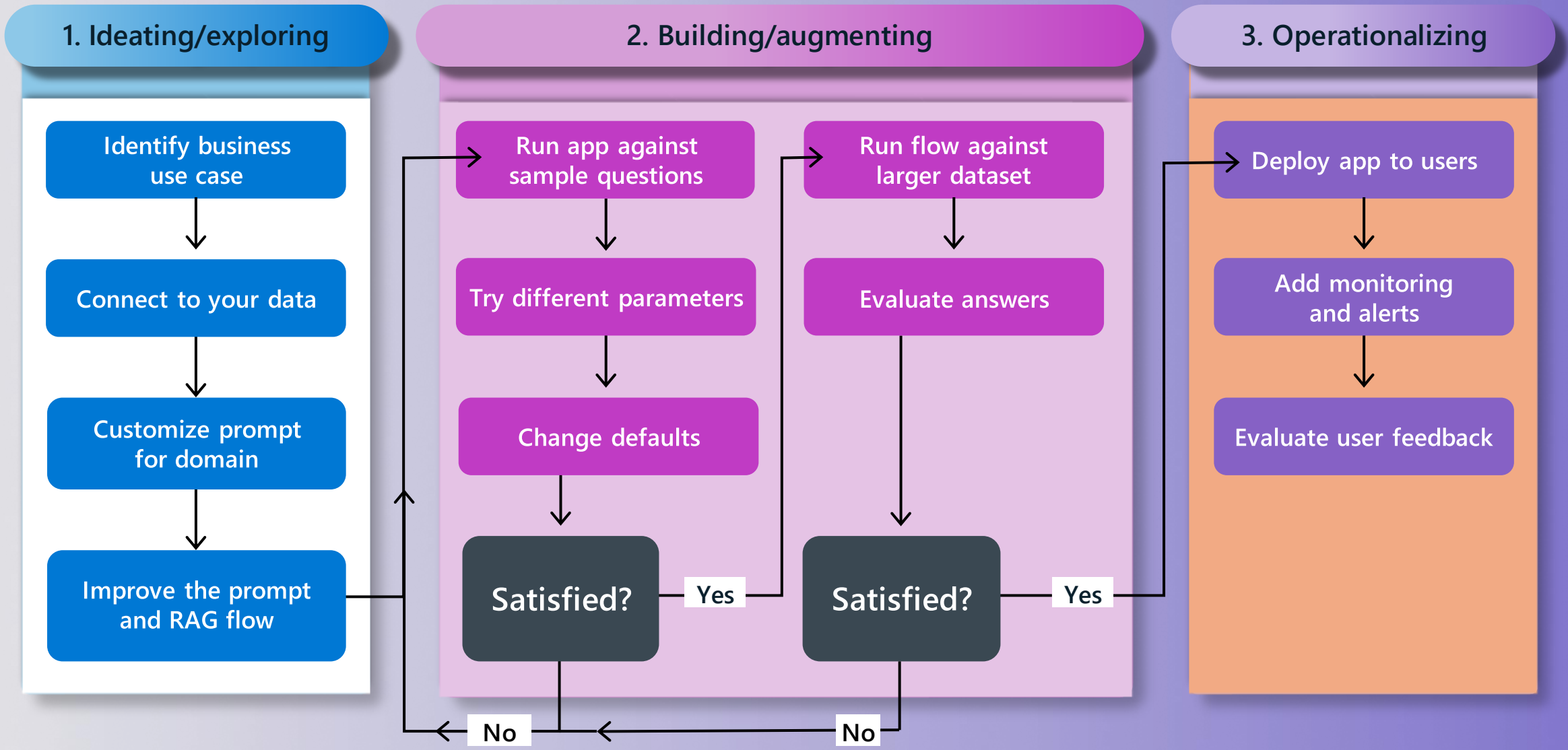
12th — SUBMIT YOUR PROJECT

Build, innovate, and #HackTogether

Microsoft

Evaluating a RAG Chat App

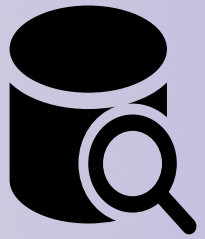
LLM Ops for RAG Chat Apps



RAG: Retrieval Augmented Generation

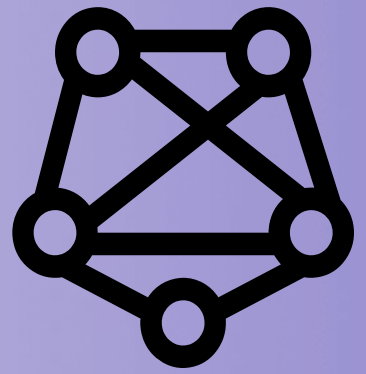
Do my company perks cover underwater activities?

User Question

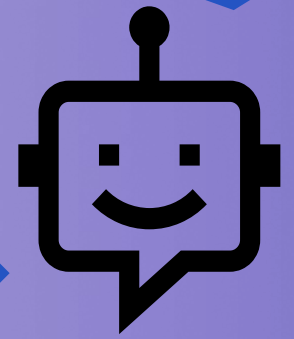


Document Search

PerksPlus.pdf#page=2: Some of the lessons covered under PerksPlus include: · Skiing and snowboarding lessons · Scuba diving lessons · Surfing lessons · Horseback riding lessons These lessons provide employees with the opportunity to try new things, challenge themselves, and improve their physical skills.....



Large Language Model



Yes, your company perks cover underwater activities such as scuba diving lessons [1]

Are the answers high quality?

- Are they correct? (relative to the knowledge base)
- Are they clear and understandable?
- Are they formatted in the desired manner?

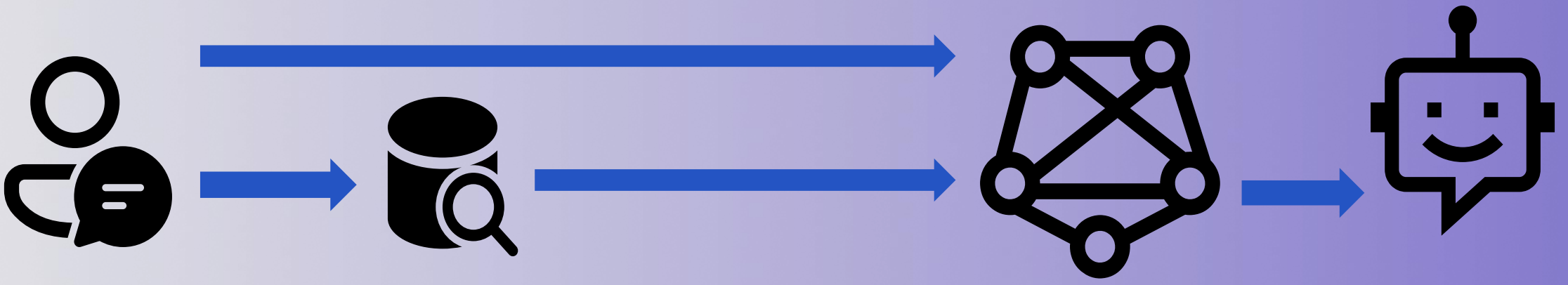
Do the perks cover underwater activities?

Yes, according to the information provided in the PerksPlus.pdf document, underwater activities such as scuba diving are covered under the program.

Yes, the perks provided by the PerksPlus Health and Wellness Reimbursement Program cover a wide range of fitness activities, including underwater activities such as scuba diving. The program aims to support employees' physical health and overall well-being, so it includes various lessons and experiences that promote health and wellness. Scuba diving lessons are specifically mentioned as one of the activities covered under PerksPlus. Therefore, if an employee wishes to pursue scuba diving as a fitness-related activity, they can expense it through the PerksPlus program.

Yes, underwater activities are included as part of the PerksPlus program. Some of the underwater activities covered under PerksPlus include scuba diving lessons [PerksPlus.pdf#page=3].

What affects the quality?



Question

Document Search

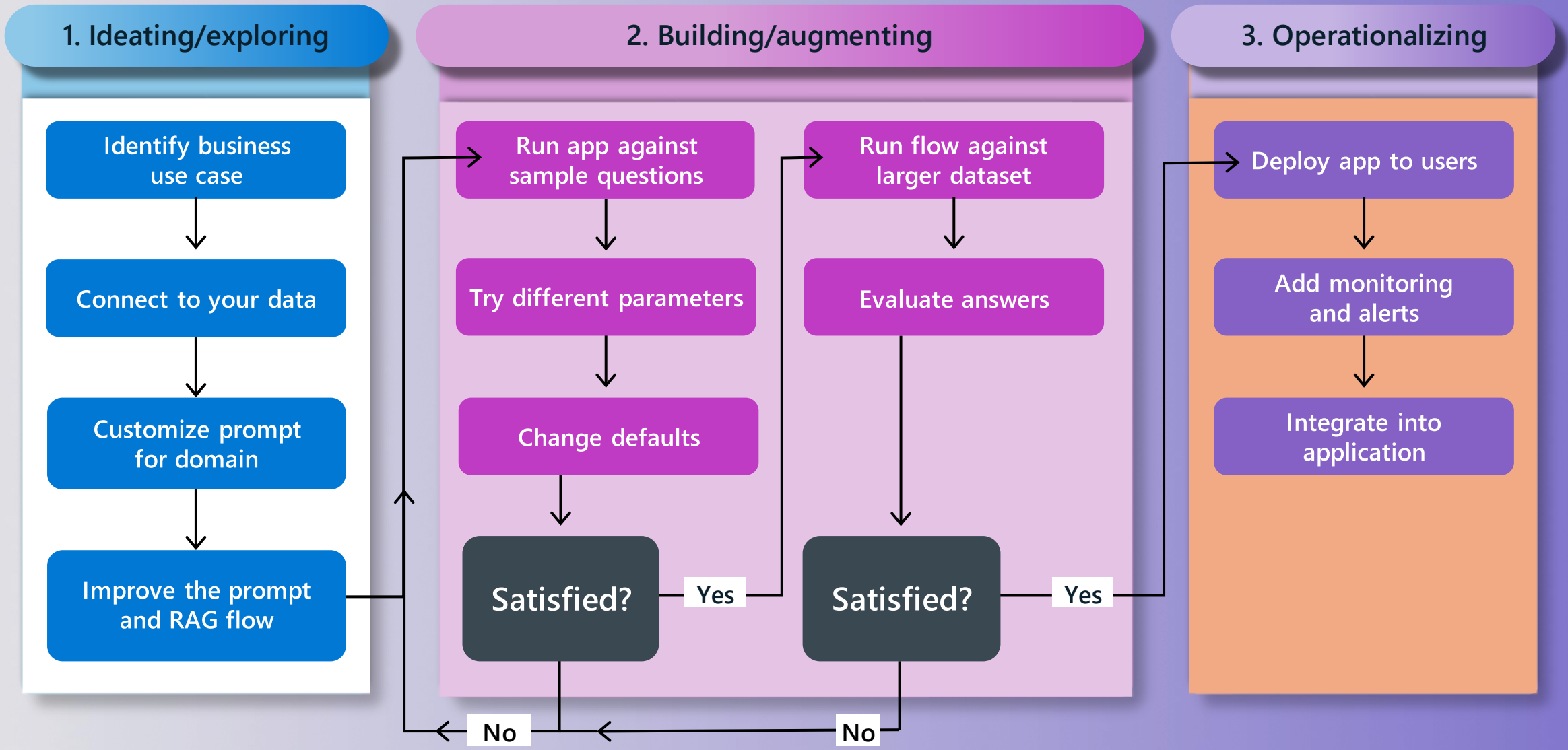
Large Language Model

- Search engine (ie. Azure AI Search)
- Search query cleaning
- Search options (hybrid, vector, reranker)
- Additional search options
- Data chunk size and overlap
- Number of results returned

- System prompt
- Language
- Message history
- Model (ie. GPT 3.5)
- Temperature (0-1)
- Max tokens

Manual experimentation

LLM Ops for RAG Chat Apps

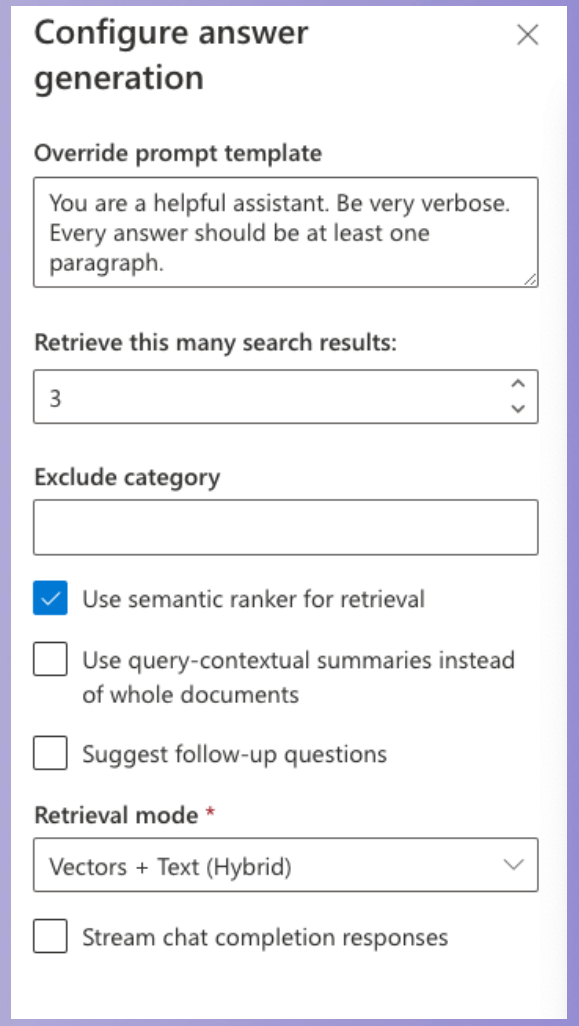


Experimenting with quality factors

Find an easy way to experiment with different settings in your RAG chat app.

For aka.ms/ragchat,

use “Developer Settings”



Prompt Refinement

Types of Prompts

- User Prompt
  - A **user prompt** is the input provided by the person interacting with the language model. It represents the **data,** or the **question** posed to the model.
- System Prompt
  - A **system prompt** serves as a set of **instructions** or **constraints** for the model's response.
  - The system prompt helps guide the AI's behavior. It can define a specific role for the AI (e.g., "be a poet" or "act like Shakespeare") or impose other limitations.

One size fits all? No way!

There are many ways to design a system prompt

- **Simple** – “You are a helpful assistant”

- **Complex** – “You are a talented developer with 8 years expertise in Python and C#. You love to be helpful by answering questions from junior developers.

  Whenever you are asked to provide code examples you pause and ensure you thoroughly understand the question. Then you will construct the code example step-by-step, ensuring you are always following best practices. You will also include comments in the examples that explain it step-by-step.

  In your response you will denote code examples with a header of “Example code:”.”

Prompt Formula

- Inspiration
  - Master the Perfect ChatGPT Prompt Formula (in just 8 minutes)! - YouTube
  - Credit to Jeff Su (LinkedIn / jsu05)

- Jeff's formula is made up of 6 components

Task Context Exemplar Persona Format Tone

Prompt Formula Components

- Task - Articulates the end goal and start with an action verb
  - Answer (a question)
  - Generate (code)
  - Write (a short summary)
  - Etc.
- Context - Use three guiding questions to help structure relevant and sufficient Context.
  - What's the user's background?
  - What does success look like?
  - What environment are they in?
  - **Plus**, content retrieved from search query.

Prompt Formula Components

- Exemplars – examples that can drastically improve the quality of the output by giving specific examples for the AI to follow.
  - Denote citations using [file1.txt][file2.doc]
  - Use the STAR answer framework: Situation, Task, Action, Results
  - Please draft the job description using the format of this existing job description below delimited by triple backticks.
- Persona – Think of who you would ideally want the AI to be in the given task situation.
  - You are an experienced physical therapist with over 20 years of experience.
  - You are a hiring manager looking to fill a [position] on your team.
  - You are a senior product manager responsible for…

Prompt Formula Components

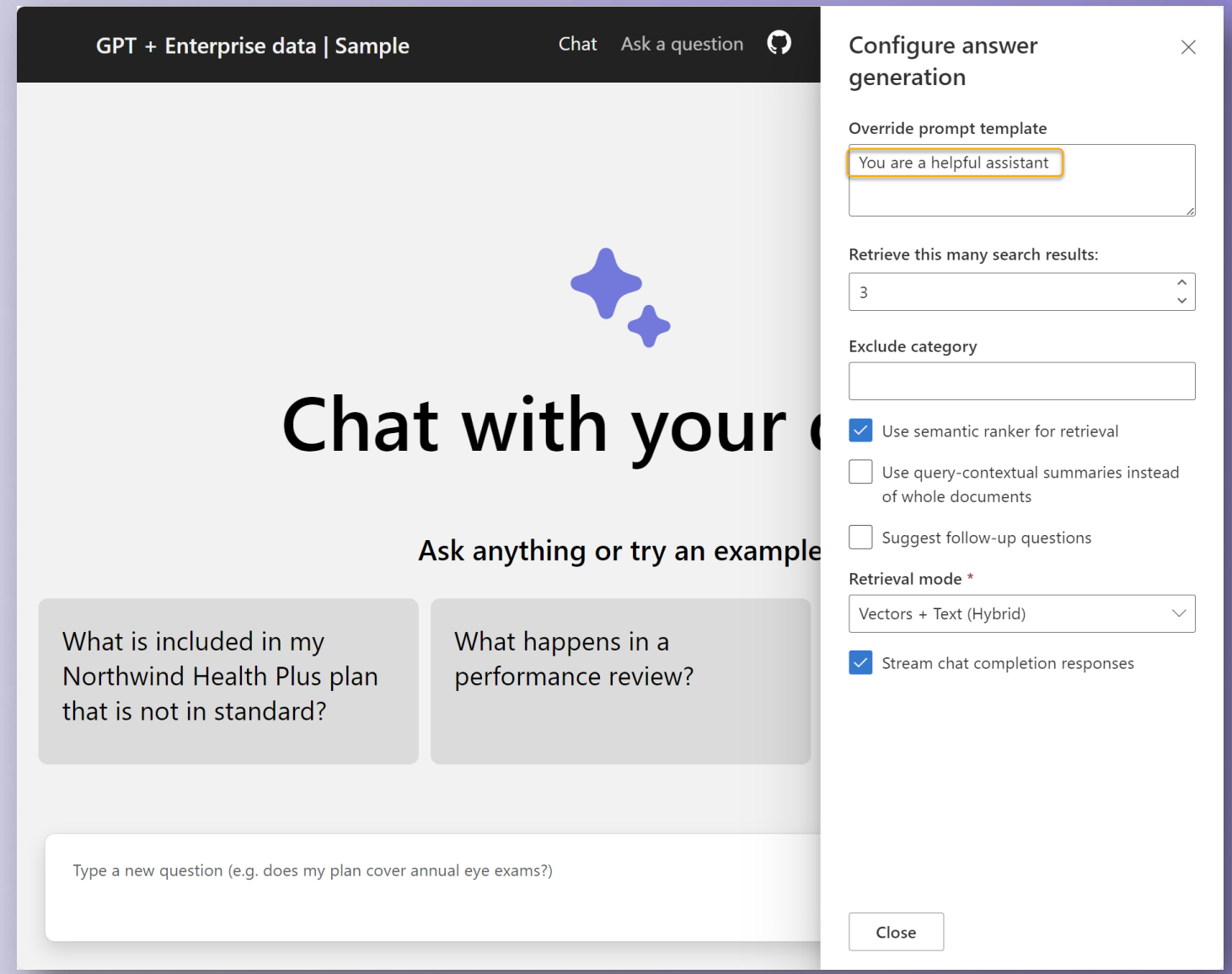
- Format – The layout or organization of the response.
  - Follow your answer to the user's question with citations
  - Don't include markdown
  - Proof-read the document in triple dashes and correct all typos and grammar mistakes and bold all changes you make
- Tone – The AI's attitude or emotional stance towards the subject and the audience.
  - Formal
  - Conversational and intimate
  - Confident and assertive
  - *Tip* - Get tone examples from Copilot or ChatGPT: "Please give me a list of 5 tone keywords to describe serious writing?"

Applying the Prompt Formula to RAG

- "You are a helpful assistant"
  - Surprisingly good, but can miss some important things

Task Context Exemplar Persona Format Tone

Demo 1


GPT + Enterprise data | Sample
Chat
Ask a question
Chat with your
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?)
Configure answer generation
Override prompt template
You are a helpful assistant
Retrieve this many search results:
3
Exclude category
Use semantic ranker for retrieval
Use query-contextual summaries instead of whole documents
Suggest follow-up questions
Retrieval mode *
Vectors + Text (Hybrid)
Stream chat completion responses
Close

Applying the Prompt Formula to RAG

- "You are a helpful assistant"

- What's missing?
  - Information about the person providing the response
  - Context comes from search results, but we want to limit responses to only that context
  - Format the response in the same language as the user prompt
  - We don't want to waste time so let's keep the response brief

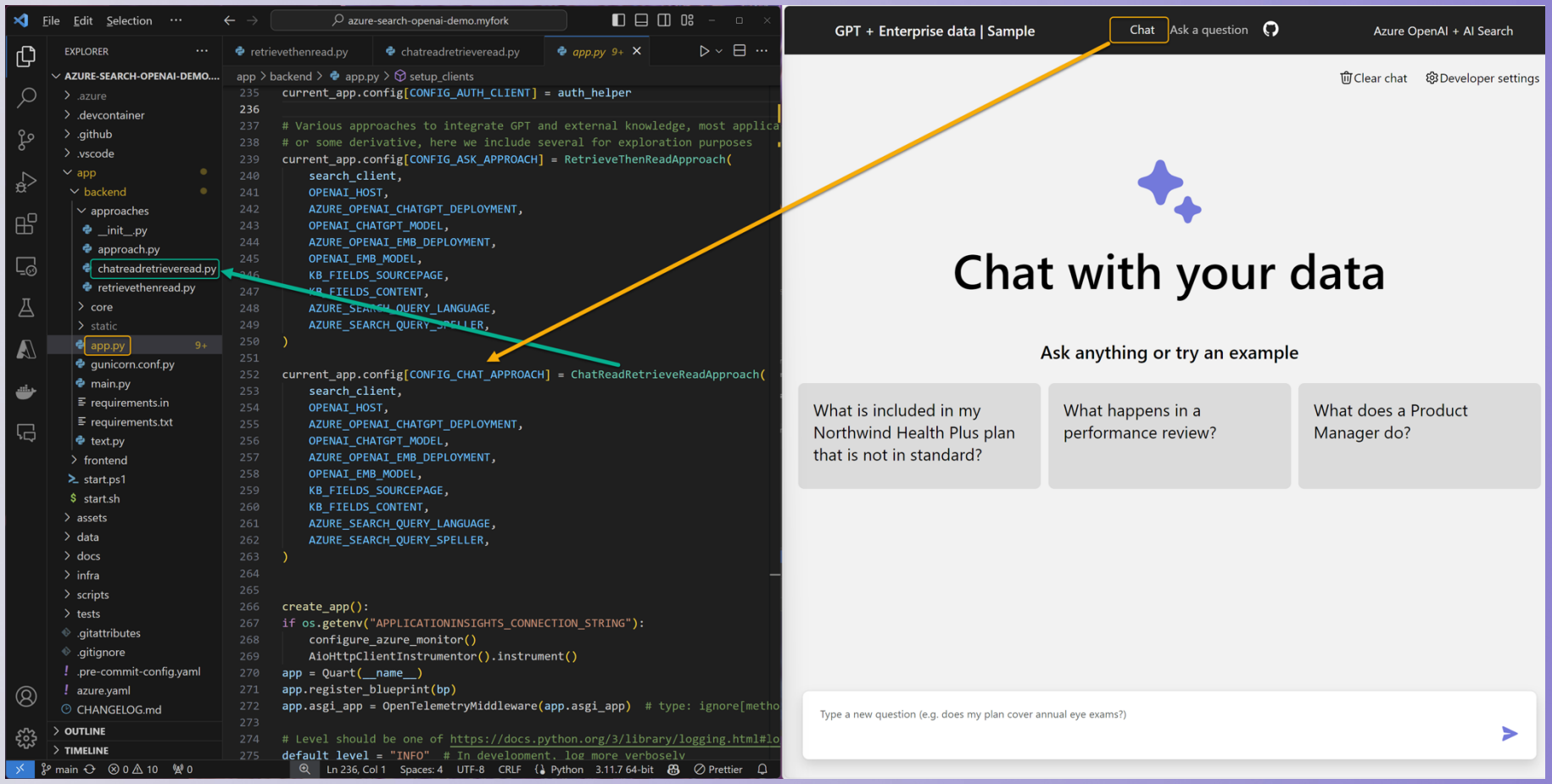
Task Context Exemplar Persona Format Tone

Example System Prompt

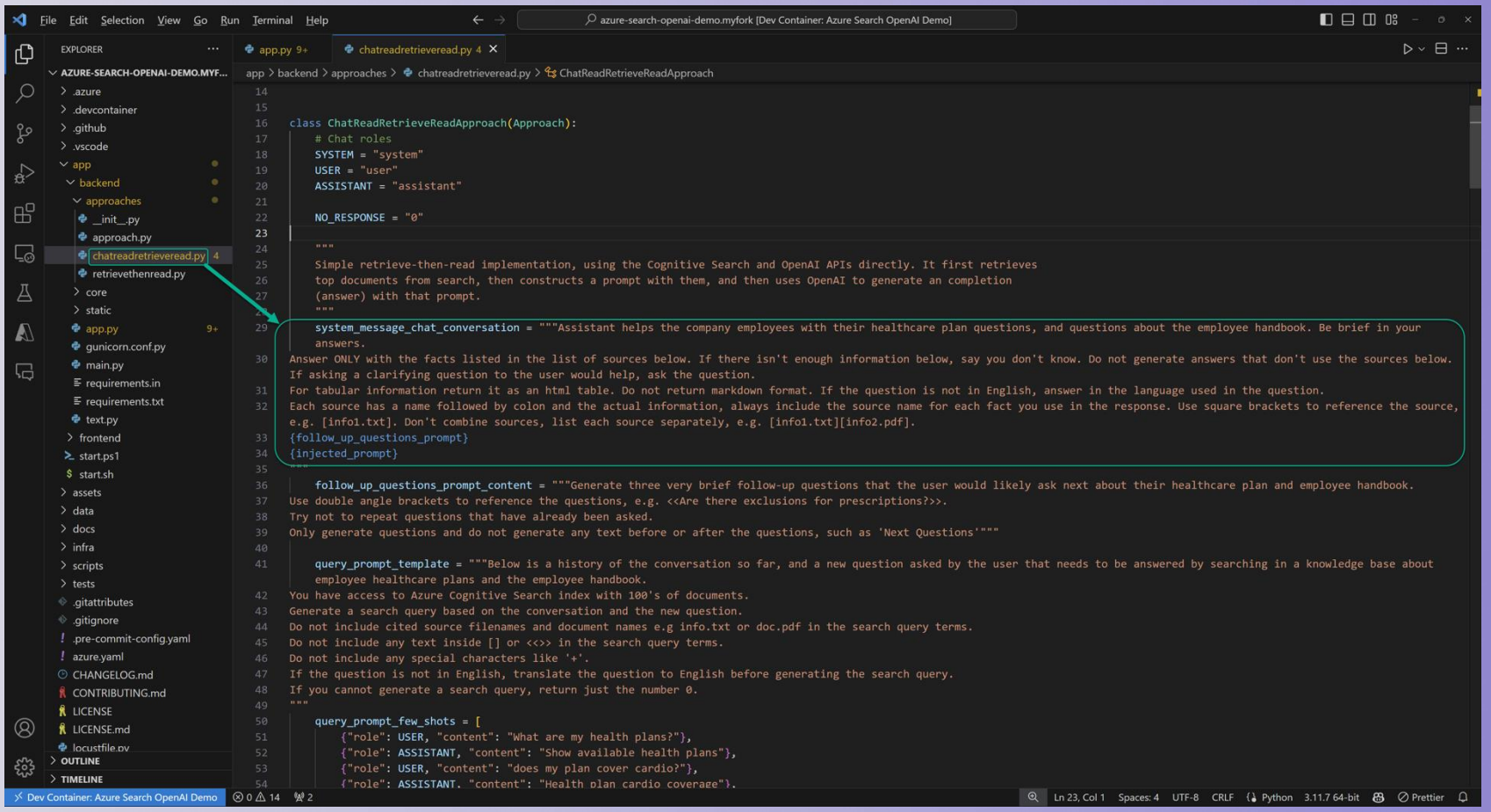
RAG system prompt for aka.ms/ragchat

```
Assistant helps the company employees with their healthcare plan
questions, and questions about the employee handbook. Be brief in your
answers.

Answer ONLY with the facts listed in the list of sources below. If there
isn't enough information below, say you don't know. Do not generate
answers that don't use the sources below. If asking a clarifying question
to the user would help, ask the question.

For tabular information return it as an html table. Do not return markdown
format. If the question is not in English, answer in the language used in
the question.

Each source has a name followed by colon and the actual information,
always include the source name for each fact you use in the response. Use
square brackets to reference the source, e.g. [info1.txt]. Don't combine
sources, list each source separately, e.g. [info1.txt][info2.pdf].
```

How to find the prompt



How to find the prompt



Applying the Prompt Formula

- RAG system prompt for https://aka.ms/azai/py/code
  - Assistant helps the company employees with their healthcare plan questions, and questions about the employee handbook. Be brief in your answers.

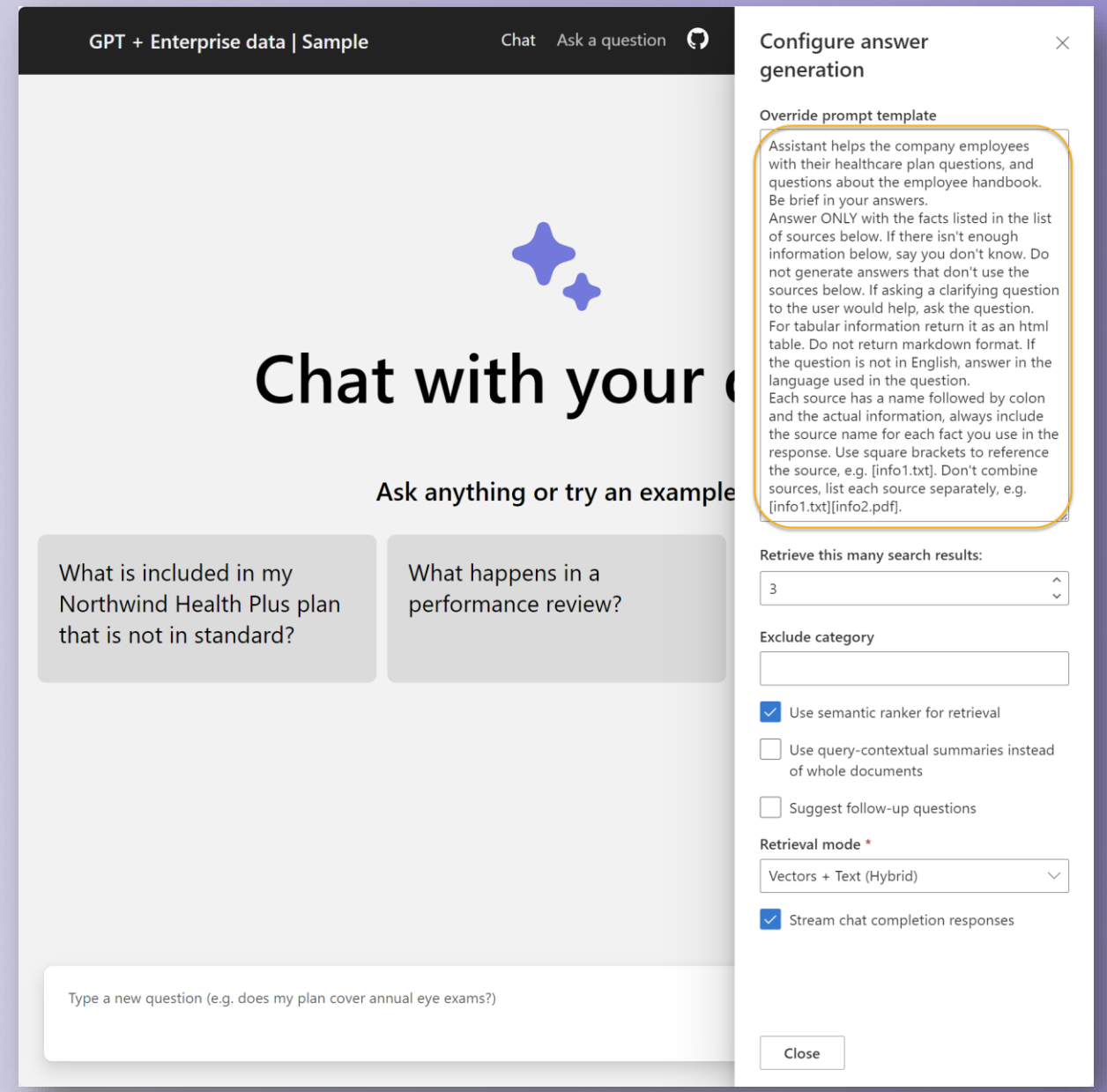
    Answer ONLY with the facts listed in the list of sources below. If there isn't enough information below, say you don't know. Do not generate answers that don't use the sources below. If asking a clarifying question to the user would help, ask the question.

    For tabular information return it as an html table. Do not return markdown format. If the question is not in English, answer in the language used in the question.

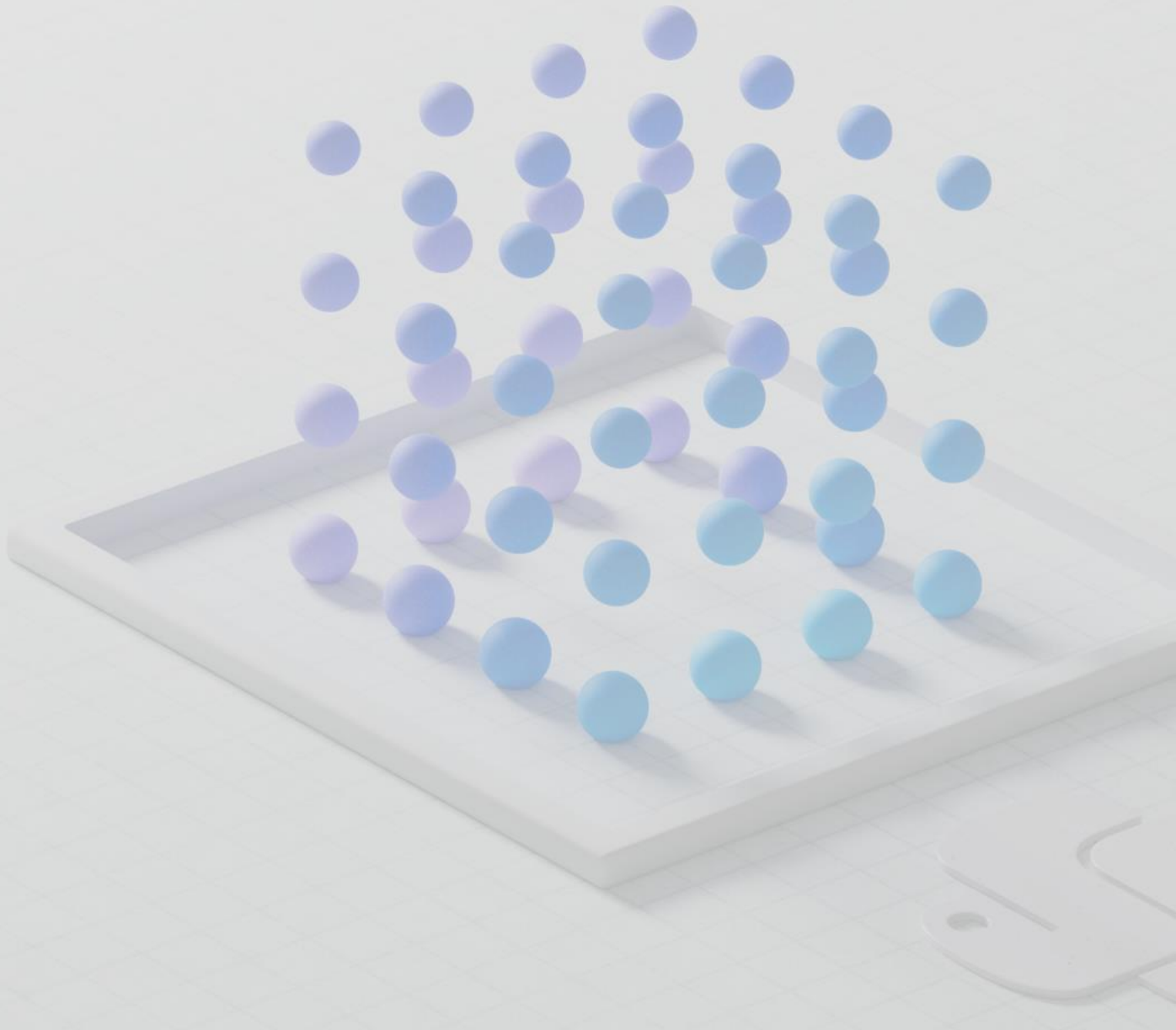
    Each source has a name followed by colon and the actual information, always include the source name for each fact you use in the response. Use square brackets to reference the source, e.g. [info1.txt]. Don't combine sources, list each source separately, e.g. [info1.txt][info2.pdf].

Task Context Exemplar Persona Format Tone

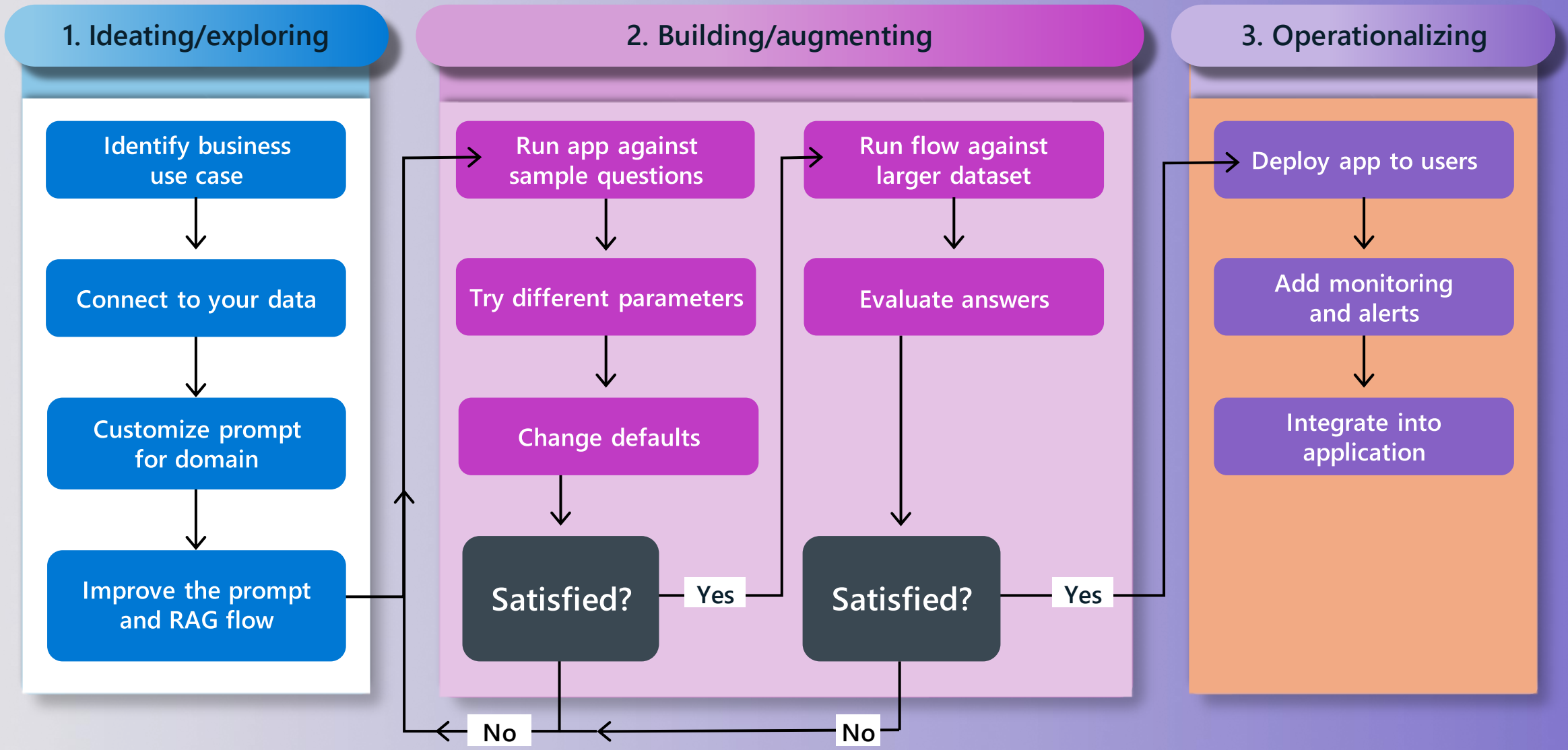
Demo 2


GPT + Enterprise data | Sample
Chat
Ask a question
Chat with your
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?)
Configure answer generation
Override prompt template
Assistant helps the company employees with their healthcare plan questions, and questions about the employee handbook. Be brief in your answers.
Answer ONLY with the facts listed in the list of sources below. If there isn't enough information below, say you don't know. Do not generate answers that don't use the sources below. If asking a clarifying question to the user would help, ask the question.
For tabular information return it as an html table. Do not return markdown format. If the question is not in English, answer in the language used in the question.
Each source has a name followed by colon and the actual information, always include the source name for each fact you use in the response. Use square brackets to reference the source, e.g. [info1.txt]. Don't combine sources, list each source separately, e.g. [info1.txt][info2.pdf].
Retrieve this many search results:
3
Exclude category
Use semantic ranker for retrieval
Use query-contextual summaries instead of whole documents
Suggest follow-up questions
Retrieval mode *
Vectors + Text (Hybrid)
Stream chat completion responses
Close

Automated evaluation



LLM Ops for RAG Chat Apps



AI RAG Chat Evaluator

A set of tools for automating the evaluation of RAG answer quality.

- Generate ground truth data
- Evaluate with different parameters
- Compare the metrics and answers across evaluations

https://github.com/Azure-Samples/ai-rag-chat-evaluator

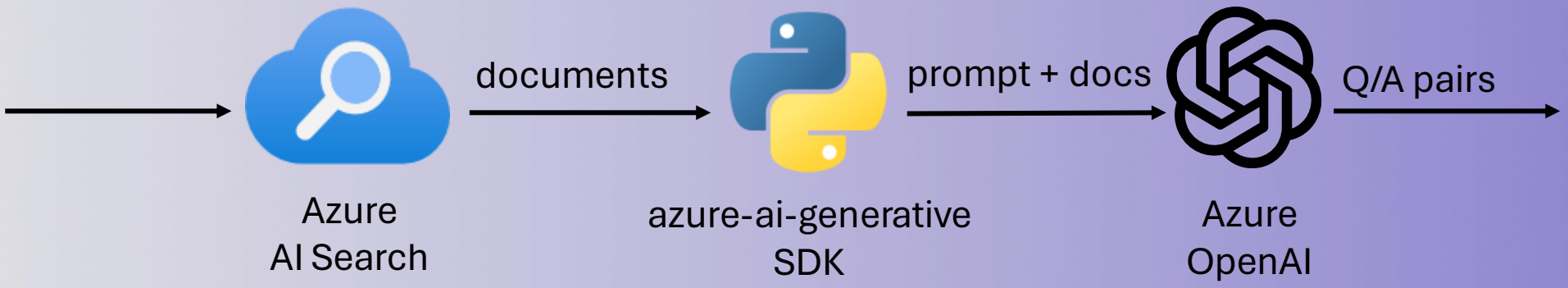
aka.ms/rag/eval

Ground truth data

The ground truth data is the ideal answer for a question.
Manual curation is recommended!

Generate Q/A pairs from a search index:

```
python3 -m scripts generate --output=example_input/qa.jsonl
                            --numquestions=200 --persource=5
```

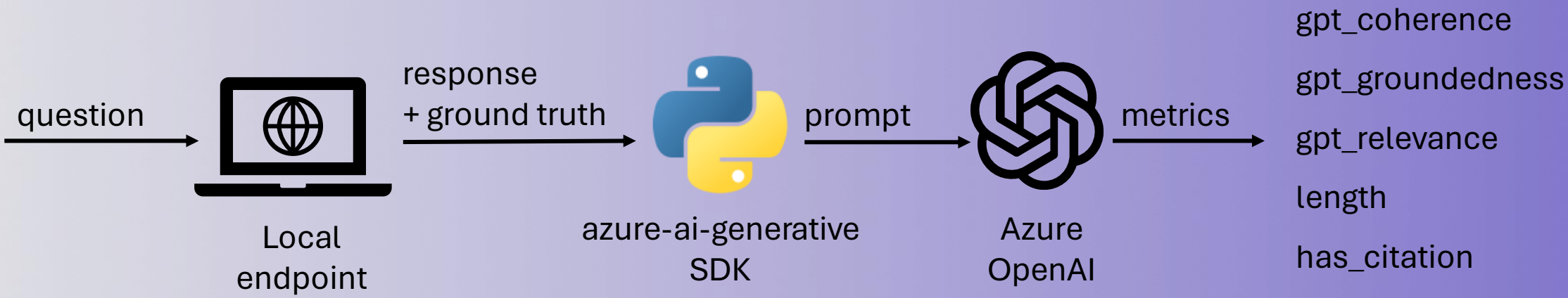


Evaluation

Compute GPT metrics and custom metrics for every question in ground truth.

Evaluate based off the configuration:

```
python3 -m scripts evaluate -–config=example_config.json
```



Review the metrics across runs

After you've run some evaluations, review the results:

```
python3 -m review_tools summary example_results
```

folder	groundedness	%	relevance	%	coherence	%	citation %	length
badprompt_ignoresources_1	4.27	0.81	4.10	0.77	4.60	0.89	0.00	1255.44
baseline_1	4.79	0.93	4.78	0.94	4.84	0.96	0.98	648.97
baseline_2	4.89	0.96	4.76	0.94	4.80	0.96	0.98	659.05
baseline_3	4.83	0.94	4.74	0.93	4.80	0.95	0.98	659.39
generalprompt_1	4.86	0.95	4.84	0.96	4.93	0.98	0.00	1164.14
refinedprompt_1	4.93	0.97	4.88	0.96	4.96	0.99	0.99	1430.57
refinedprompt_2	4.87	0.95	4.86	0.96	4.95	0.99	0.99	1398.83
refinedprompt_trimmed_1	4.90	0.97	4.83	0.95	4.93	0.98	0.99	969.89
refinedprompt_trimmed_2	4.91	0.97	4.89	0.97	4.89	0.98	0.99	975.07
refinedprompt_trimmed_3	4.95	0.98	4.84	0.96	4.92	0.98	0.99	965.01
refinedprompt_trimmed_4	4.93	0.97	4.89	0.97	4.86	0.96	0.98	962.78

Compare answers across runs

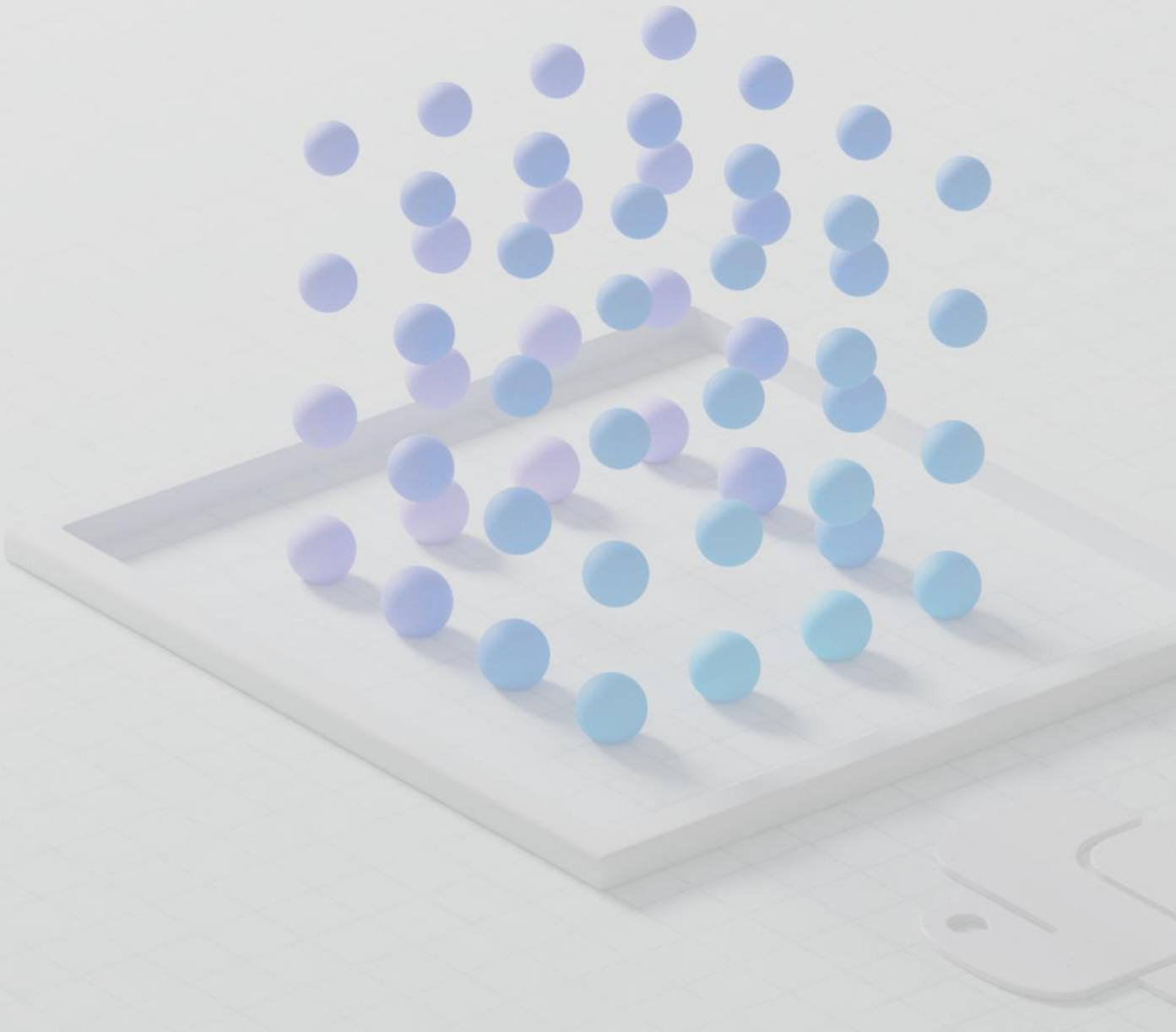
```
python3 -m review_tools diff example_results/baseline_1
                             example_results/baseline_2
```

```
What is the importance of choosing the right provider in getting the most value out of your health
insurance plan?

baseline_1

    Choosing the right provider is important
    for getting the most value out of your
    health insurance plan. In-network
    providers charge lower rates than
    out-of-network providers, so choosing an
    in-network provider can help you save
    money on your healthcare costs
    [Northwind_Standard_Benefits_Details.pdf#
    page=5][Northwind_Health_Plus_Benefits_De
    tails.pdf#page=4]. In addition,
    continuity of care is an important

groundedness  relevance  coherence
5             5          5

baseline_2

    Choosing the right provider is important
    in getting the most value out of your
    health insurance plan for several
    reasons:

     1. Cost savings: In-network providers
        charge lower rates than
        out-of-network providers, so choosing
        an in-network provider can help you
        save money on your healthcare costs.
        [Northwind_Standard_Benefits_Details.

groundedness  relevance  coherence
5             5          5
```

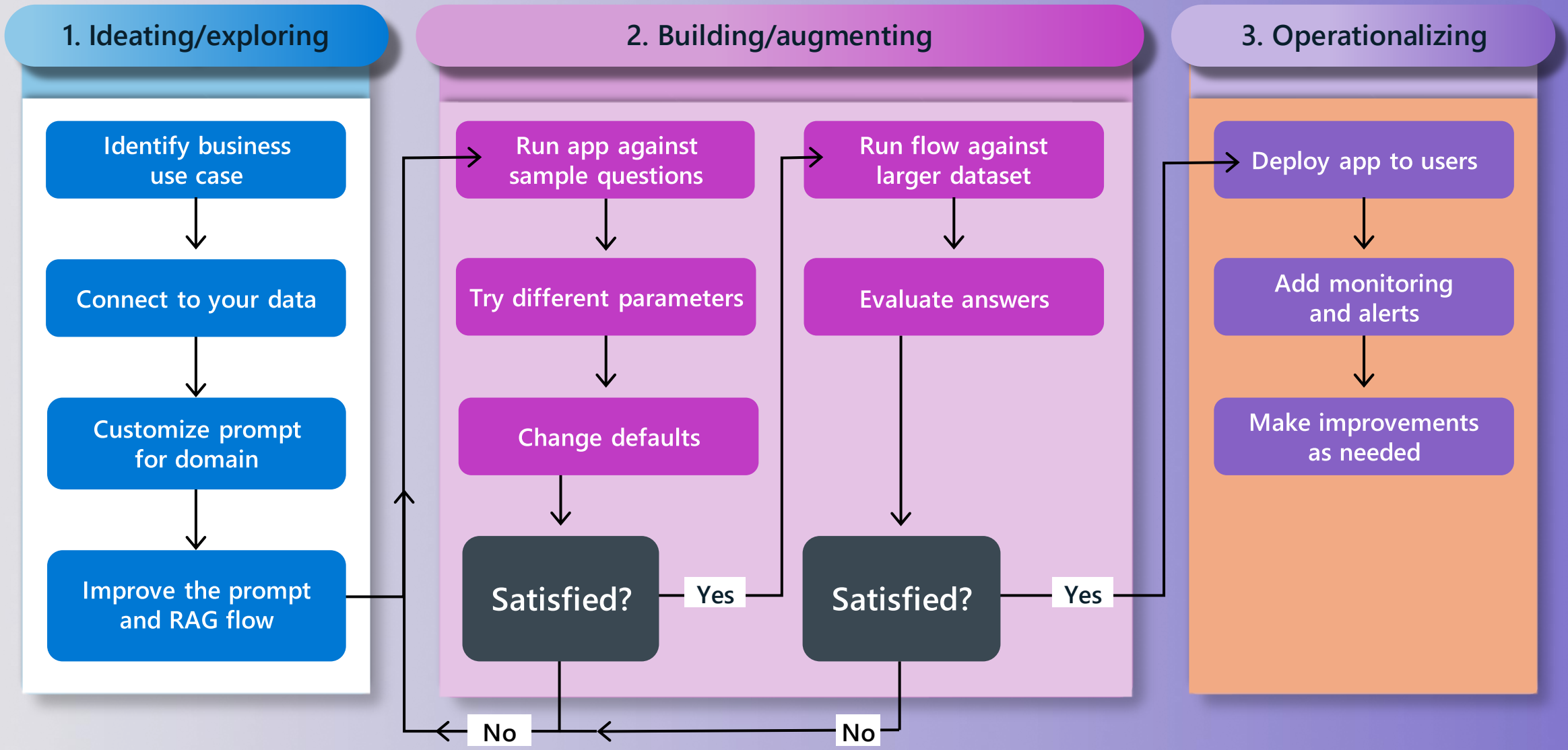
Evaluation approach

- Evaluate at least 200 Q/A pairs.
- Start by evaluating the baseline, the default parameters.
- For each set of parameters, evaluate at least 3x.
  - Consider using seed in the app itself to reduce variation.
- Track evaluation results in a repo, tied to RAG code changes.

Quality monitoring



LLM Ops for RAG Chat Apps



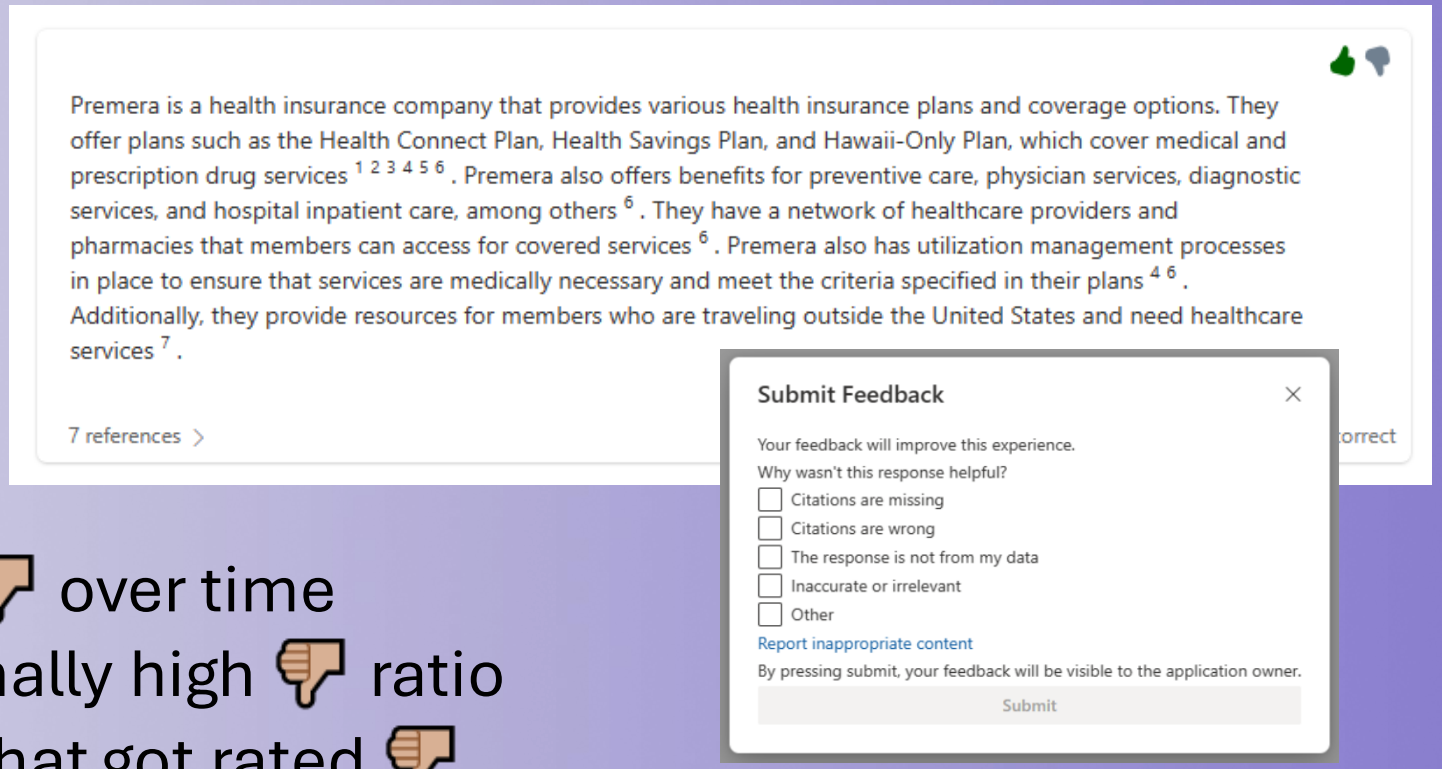
Answer logging

- Log user's questions and answers in a secure database, like CosmosDB.
  - Remove PII first.
  - Un-stream streamed answers.
- Periodically sample questions into evaluation ground truth data set to reflects the topics and question styles used by users. Avoid drift!

https://learn.microsoft.com/azure/ai-services/language-service/personally-identifiable-information
https://learn.microsoft.com/azure/architecture/ai-ml/openai/architecture/log-monitor-azure-openai

Feedback buttons

Add a 👍/👎 button with feedback dialog:



Then you can:

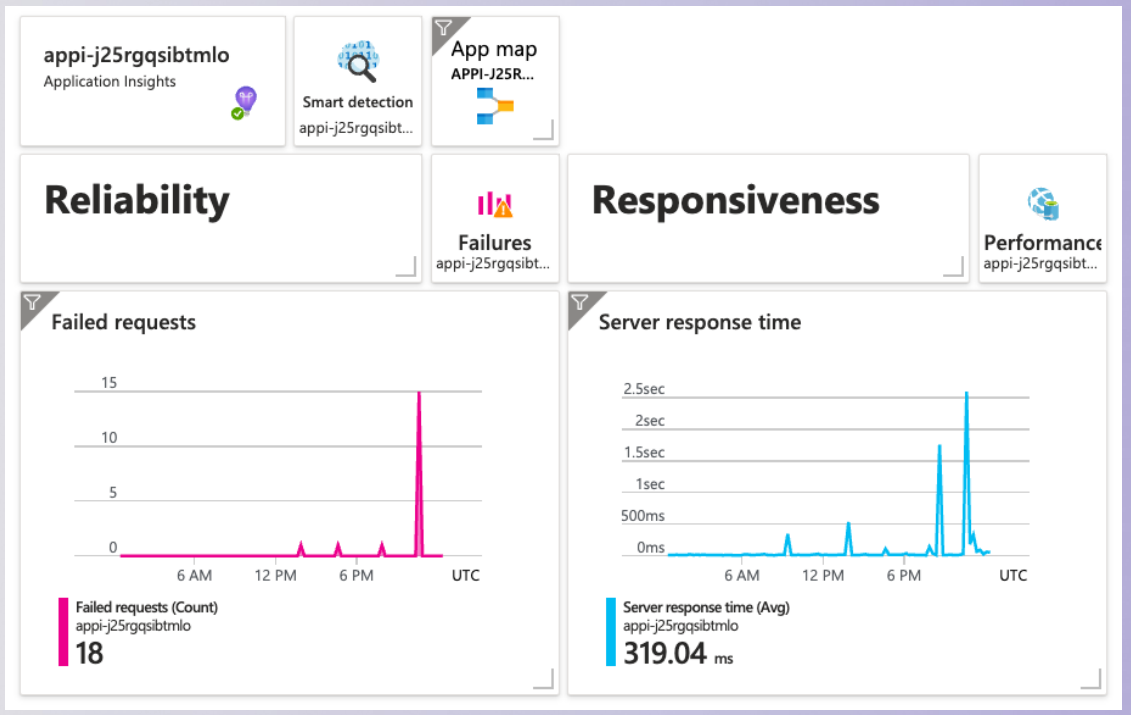
- Monitor ratio of 👍/👎 over time
- Set alerts for abnormally high 👎 ratio
- Debug the answers that got rated 👎
- Use A/B tests on prompt changes with 👍 as goal

https://github.com/microsoft/sample-app-aoai-chatGPT/pull/396

aka.ms/rag/thumbs

Overall application health

Set up dashboards for server latency, errors, OpenAI errors

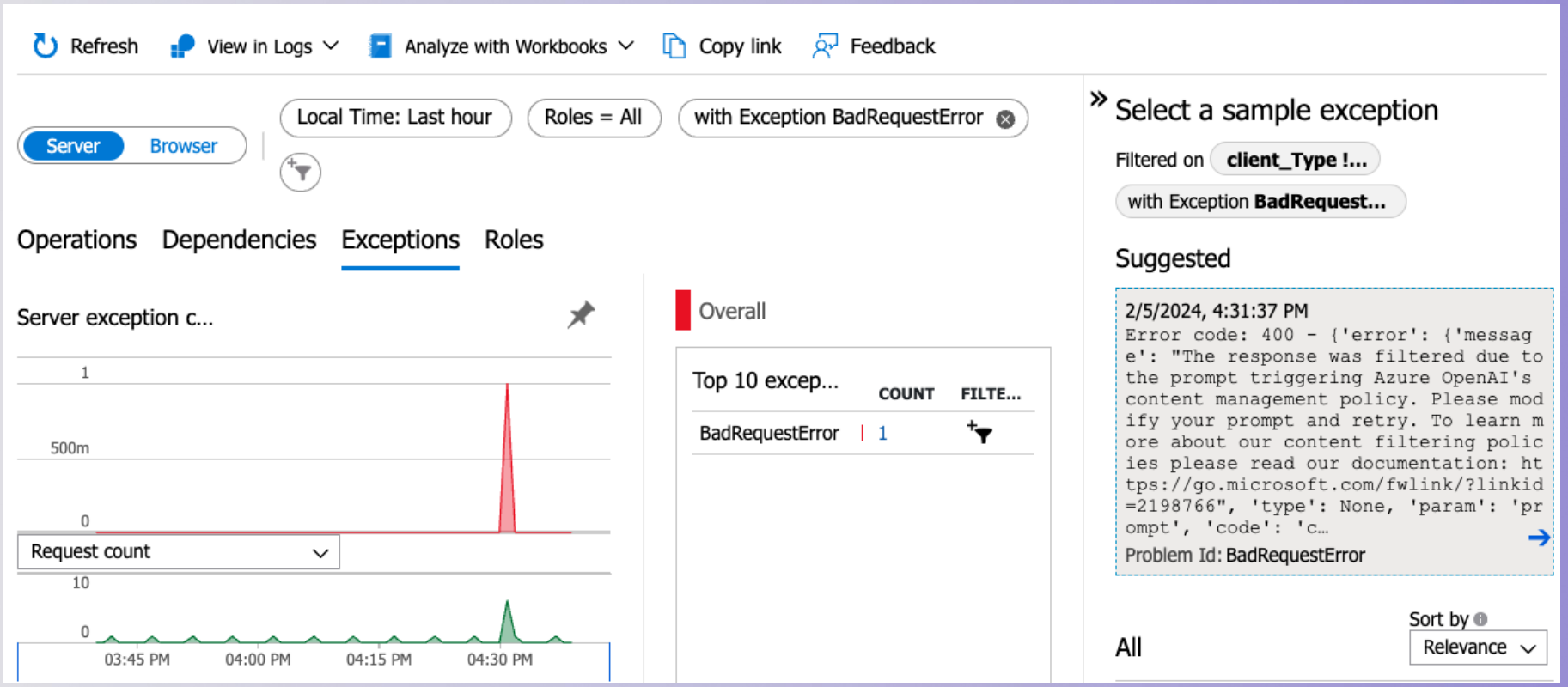


https://learn.microsoft.com/azure/ai-services/openai/how-to/monitoring

aka.ms/openai/monitor

Content safety errors

Azure OpenAI raises a special 400 error for content safety violations.



Next steps

- Register for the hackathon → aka.ms/hacktogether/chatapp
- Introduce yourself in our discussion forum
- Deploy the repo with the sample data
  - See steps on low cost deployment → aka.ms/ragchat/free
- Hack, hack, hack, hack! 👨🏿‍💻 👩🏻‍💻 👱🏼‍💻
- Post in forum if you have any questions. 🙋‍♀️ 🙋🏾‍♀️ 🙋🏻‍♀️
- Submit your project before February 12th to win prizes! 💸 💰 💵
- Join tomorrow's session: "ChatCompletion API Tools & Functions"