



Metric Driven AI Development

Robert Hoffmann, PhD

Data Scientist & Cloud Solution Architect





Microsoft Austria

robert.hoffmann@microsoft.com

Goals for this workshop — methodology, tooling, motivation...

- Experience on how to evaluate AI applications
- Theoretical why and how
- Tools and practical approaches using the AI Foundry and Python

Agenda

12:30 PM - 01:30 PM		Check-in & Networking Lunch
01:30 PM - 03:30 PM		Workshop Session
03:30 PM - 03:45 PM		Coffee Break
03:45 PM - 05:00 PM		Workshop Session

Setup

Audience & Requirements

This technical workshop is intended for **developers, data scientists, and program managers**. While evaluation is an advanced and essential part of AI development, only basic experience with LLMs, RAG, and prompt engineering is required to **actively participate**. **Python knowledge is necessary** for the coding portions.

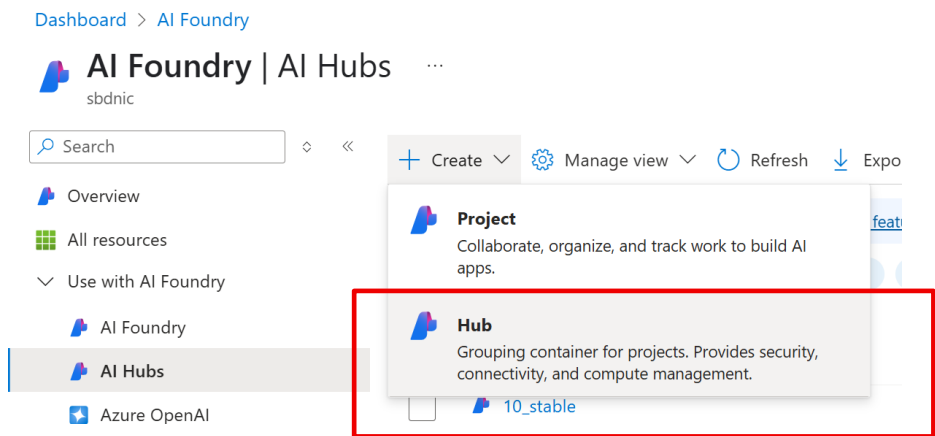
Requirements

Azure: **Azure Subscription**; permissions to create/use an AI Foundry resource; ability to deploy at least OpenAI's GPT-4o model.

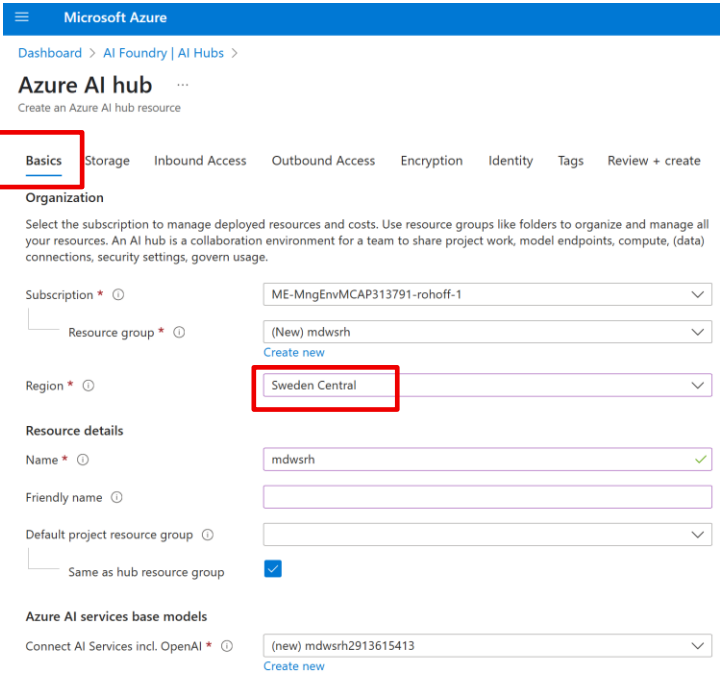
Your Notebook: Azure CLI; Python with virtual environments (e.g., Conda); permissions to install pip packages; Visual Studio Code (or equivalent) for running Jupyter notebooks.

Setup — Azure AI Foundry (Hub*)

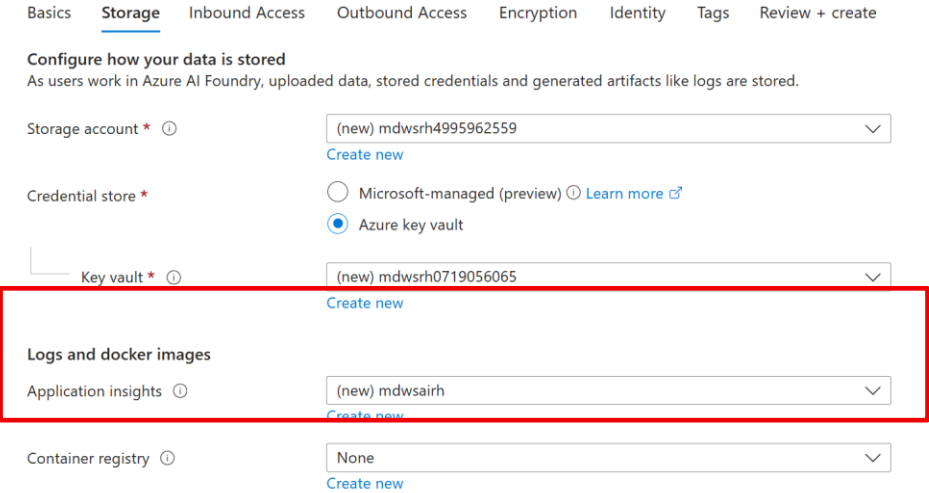
1.



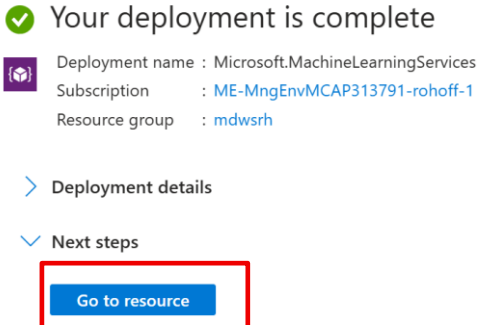
2.



3.




4.



*<https://learn.microsoft.com/en-us/azure/ai-foundry/what-is-azure-ai-foundry#project-types>

Setup — Check AI Foundry (Hub) overview in Azure Portal

 **mdwsrh** ☆ ...
Azure AI hub

Overview

Activity log

Access control (IAM)

Tags

Diagnose and solve problems

Resource visualizer

Events

Settings

Monitoring

Automation

Support + troubleshooting

+ Create project

↓ Download config.json

🗑 Delete

^ Essentials

Resource group

Location

Subscription

Subscription ID

Key Vault

: [mdwsrh](#)

: Sweden Central

: [ME-MngEnvMCAP313791-rohoff-1](#)

: c11caebe-ea81-4036-9e58-ccf406d87ead

: [mdwsrh0719056065](#)

Project resource group (default)

Storage

Container Registry [\(edit\)](#)

Application Insights [\(edit\)](#)

Provisioning State


: [mdwsrh](#)

: [mdwsrh4995962559](#)

: ...

: [mdwsairh](#)

: Succeeded



Govern the environment for your team in AI Foundry

Your Azure AI hub provides enterprise-grade security, and a collaborative environment to build AI solutions. Centrally audit usage and cost, and set up connections to your company resources that all projects can use. [learn more about the Azure AI Foundry](#) ↗

Launch Azure AI Foundry

©Microsoft Corporation
Azure

Setup — Create project in AI Foundry

1.

Govern the environment for your team in AI Foundry



Your Azure AI hub provides enterprise-grade security, and a collaborative environment to build AI solutions. Centrally audit usage and cost, and set up connections to your company resources that all projects can use. [learn more about the Azure AI Foundry](#)

Launch Azure AI Foundry

2.

Azure AI Foundry | Management center / Hub overview

Management center

All resources

Quota

Hub (mdwsrh)

Overview

mdwsrh

New project

Refresh

Delete project

Reset view

3.

Name your project

Your Azure AI Foundry project is where you'll work, collaborate, and connect to data and other services.

Current hub ⓘ

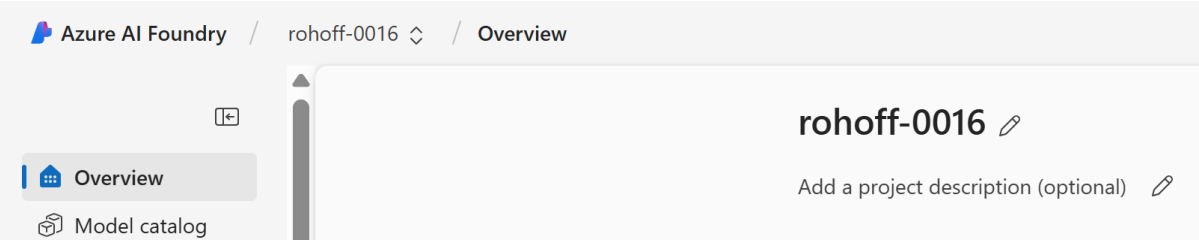
mdwsrh

Project name * ⓘ

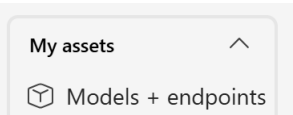
rohoff-0016

Setup — Azure AI Foundry — Model deployment

1.



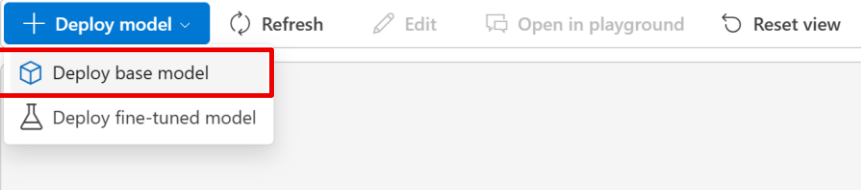
2.



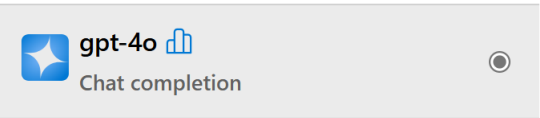
3.

Manage deployments of your models and services

Model deployments Service endpoints



4.



5.

Deploy gpt-4o

Deployment name *

gpt-4o

Deployment type

Global Standard

Global Standard: Pay per API call with the highest rate limits. Learn more about [Global deployment types](#).

Data might be processed globally, outside of the resource's Azure geography, but data storage remains in the AI resource's Azure geography. Learn more about [data residency](#).

Deployment details

Model version upgrade policy

Upgrade once new default version becomes available

Model version

2024-11-20

Resource location

East US

Sweden Central *

450K tokens per minute quota available for your deployment

Tokens per Minute Rate Limit

450K

Corresponding requests per minute (RPM) = 2.7K

A new AI resource will be created for your deployment

A resource location that supports the model has been pre-selected

Create resource and deploy

Cancel

* <https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models?tabs=global-standard%2Cstandard-chat-completions#global-standard-model-availability>

Setup — Clone GitHub repo

1. Clone repo in terminal

git clone <https://github.com/rohoffgit/AI-eval>

2. Enter repo

cd AI-eval

3. Open VSCode

code .

4. Open Terminal in VSCode

5. Optionally create Python environment (Python \geq 3.12), e.g.

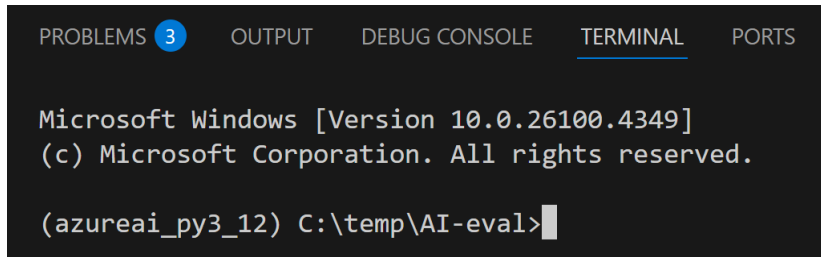
conda create -n azureai_py3_12 python=3.12

Setup — Set up VSCode for environment

1. Set python interpreter / environment in VSCode

Ctrl+Shift+P > 'Python interpreter' > choose correct environment

2. Open new terminal in VSCode to confirm correct environment

A screenshot of a VS Code terminal window. The terminal title bar shows 'PROBLEMS 3', 'OUTPUT', 'DEBUG CONSOLE', 'TERMINAL' (which is selected and underlined), and 'PORTS'. The terminal content displays the Windows command prompt version '10.0.26100.4349' and copyright information '(c) Microsoft Corporation. All rights reserved.'. The current directory is 'C:\temp\AI-eval' and the active Python environment is '(azureai_py3_12)'. A cursor is visible at the end of the command line.

```
PROBLEMS 3 OUTPUT DEBUG CONSOLE TERMINAL PORTS  
  
Microsoft Windows [Version 10.0.26100.4349]  
(c) Microsoft Corporation. All rights reserved.  
  
(azureai_py3_12) C:\temp\AI-eval>
```

3. Install dependencies

python ./00_setup.py

4. Copy .env.template to .env

Setup — Edit .env

Azure AI Foundry

AZURE_SUBSCRIPTION_ID = see <https://portal.azure.com>

AZURE_RESOURCE_GROUP_AI = see <https://portal.azure.com>

AZURE_AI_FOUNDRY_PROJECT_NAME = <https://ai.azure.com/>

AZURE_AI_FOUNDRY_PROJECT_CONNECTION_STRING =

AZURE_AI_INFERENCE_ENDPOINT =

Azure OpenAI

AZURE_OPENAI_ENDPOINT =

AZURE_OPENAI_API_KEY =

AZURE_OPENAI_DEPLOYMENT = gpt-4o

AZURE_OPENAI_API_VERSION = 2025-01-01-preview

Tracing

APP_INSIGHTS_RESOURCE_ID = see <https://portal.azure.com> > App Insights > Overview > JSON view >

The screenshot displays the Azure AI Foundry console interface. At the top, the breadcrumb navigation shows 'Azure AI Foundry / rohoff-0016 / Overview'. The main content area is divided into two sections: 'Project details' and 'Included capabilities'.

Project details

- Project connection string**: swedencentral.api.azureml.ms:c11caebe-ea... (with a copy icon)

Included capabilities

- Azure AI inference** (selected):
 - Use the following endpoint to call all your deployed base models:
 - Azure AI model inference endpoint** (PREVIEW): https://rohof-mc90j3yd-eastus.services.ai.azure.com/models (with a copy icon)
- Azure OpenAI** (selected):
 - Use the following endpoint to call your Azure OpenAI models:
 - Azure OpenAI endpoint**: https://rohof-mc90j3yd-eastus.openai.azure.com/ (with a copy icon)
- Azure AI Services**

API Key

A text input field for the API key, with a copy icon and a refresh icon.

Resource JSON

mdwsairh

Resource ID

/subscriptions/c11caebe-ea81-4036-9e58-ccf406d87ead/resourceGroups/mdwsrh/providers/Microsoft.Insights/...

Let's go!

Metric Driven AI Development

Understanding Metrics

Metrics are essential for evaluating the performance of AI systems, informing development strategies and decisions.

Data-Driven Development

Data-driven approaches utilize metrics to guide the development of AI systems, ensuring they meet user and business needs.

Continuous Improvement

Metric driven development fosters continuous improvement in AI systems, leading to enhanced accuracy and efficiency over time.

Testing in software development

Evaluation — *the process of judging or calculating the quality, importance, amount, or value of something (Cambridge dictionary)*

Tests and metrics are components of evaluation.

e.g. in software development

[unsorted items] -> [sorting functions, show code] -> [sorted items]

[A] -> [A]

[A, B, C] -> [A, B, C]

[C, B, A] -> [A, B, C]

[A, B, C,] -> [A, B, C,]

[1, A, 2, B] -> [A, B, 1, 2]

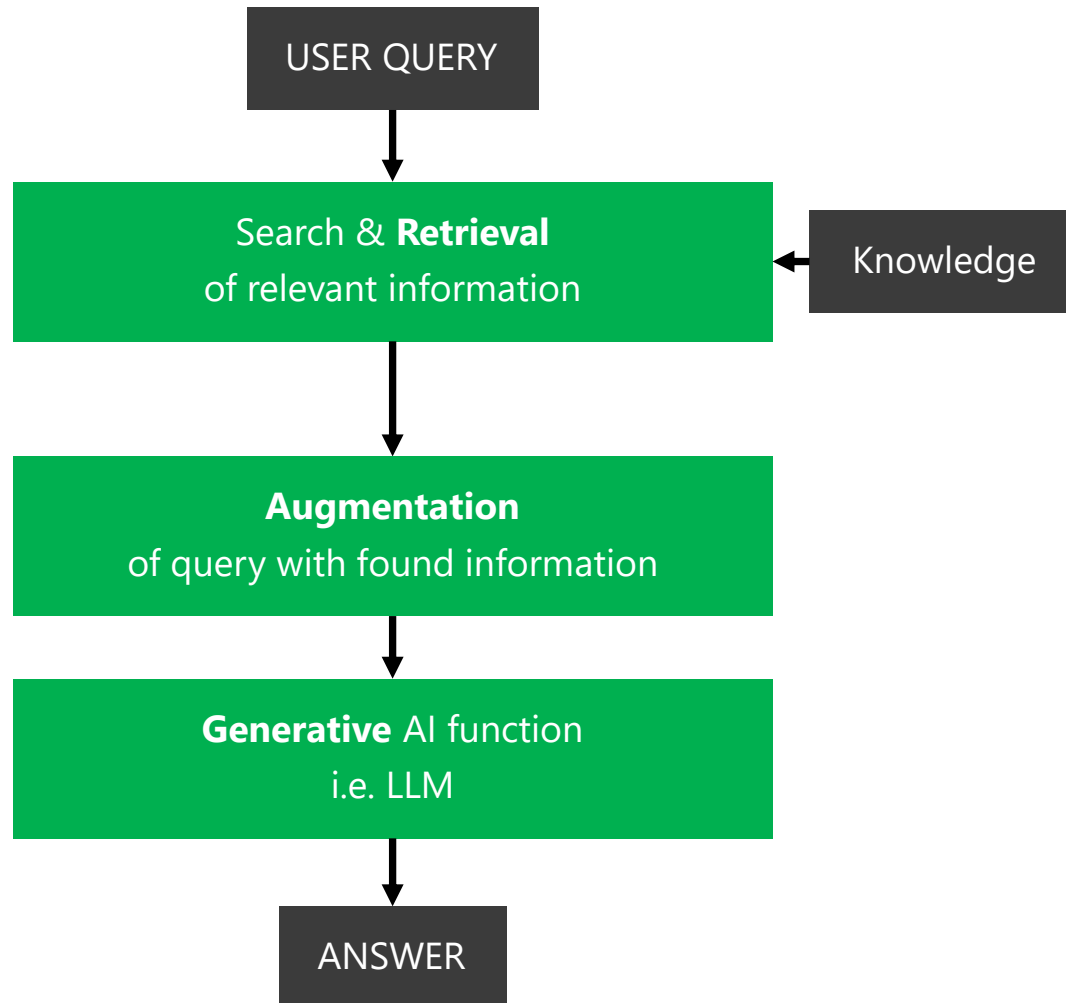
[10, A, 2, B] -> [A, B, 10, 2]

Are AI applications special?

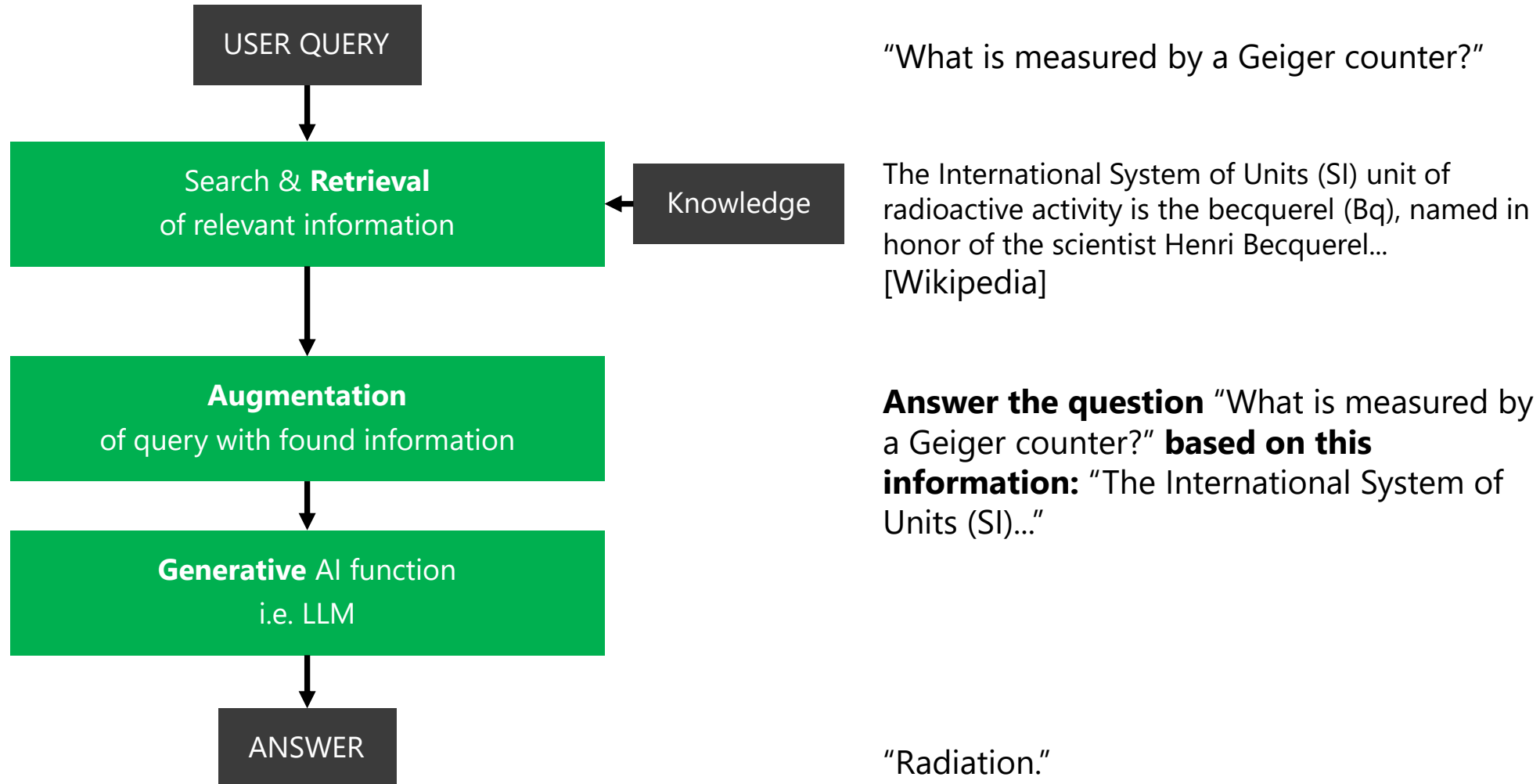
[unsorted items] -> [sorting functions] -> [sorted items]

[input tokens] -> [AI function] -> [output tokens]

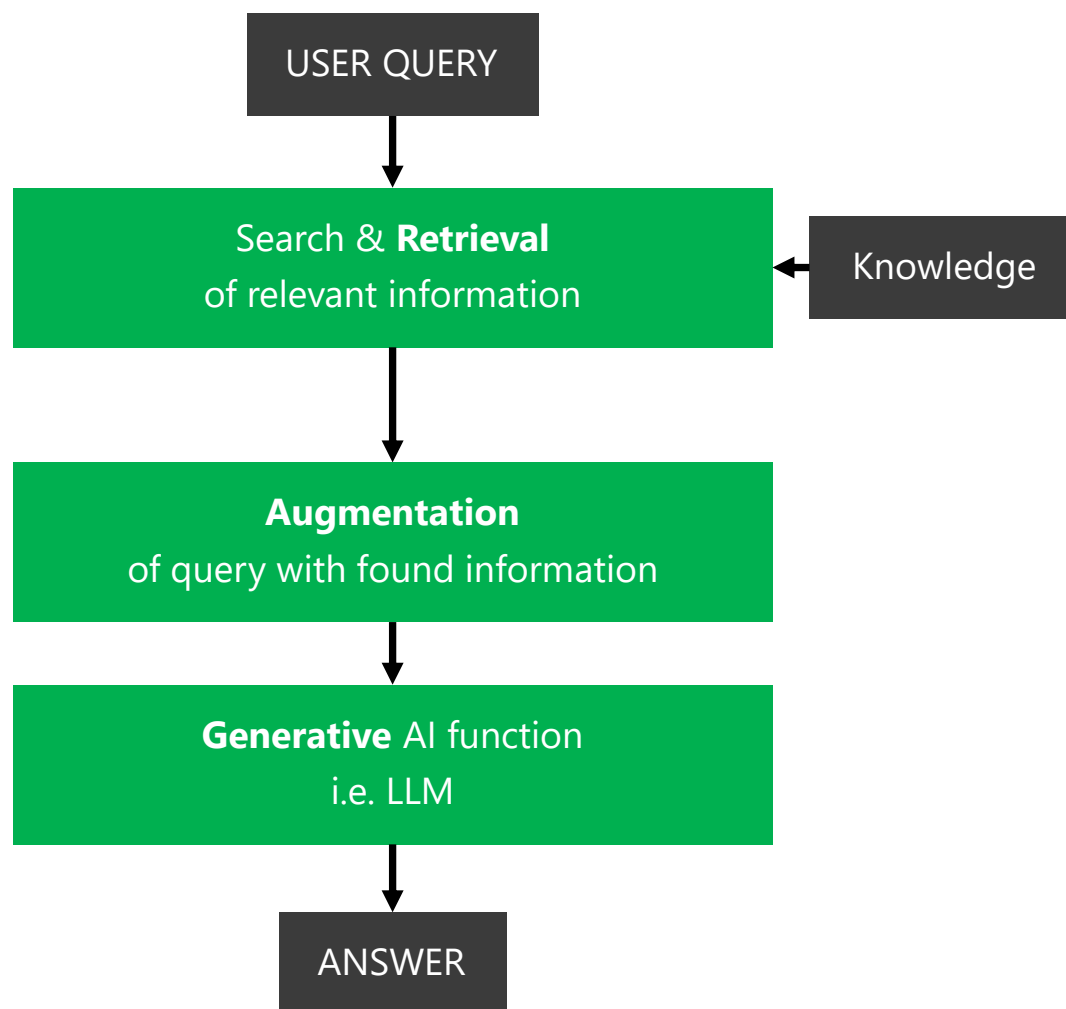
Working example — RAG (Retrieval augmented generation)



Working example — RAG (Retrieval augmented generation)



Working example — RAG (Retrieval augmented generation)



"What is measured by a Geiger counter?"

The International System of Units (SI) unit of radioactive activity is the becquerel (Bq), named in honor of the scientist Henri Becquerel...
[Wikipedia]

Answer the question "What is measured by a Geiger counter?" **based on this information:** "The International System of Units (SI)..."

"Radiation."

- New versions of AI functions and models
- Open-ended set of inputs and thus outputs
- Less deterministic nature of AI
- Subjective or context dependent correct outputs
- Difficult to assess correctness
- ...

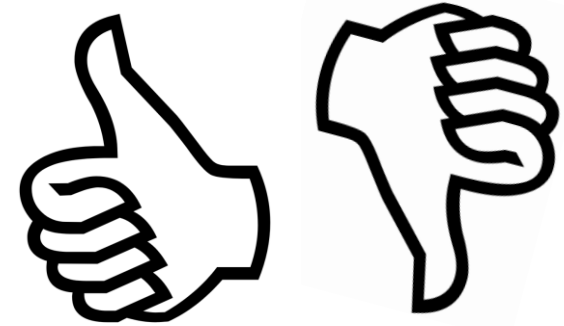
What about other AI applications?

Evaluate Quality of AI Applications and Agents		
Agents	Generation	Custom
<ul style="list-style-type: none">• Intent Resolution<ul style="list-style-type: none">• Correct Intent Identification• Clarification for Ambiguity• Tool Call Accuracy<ul style="list-style-type: none">• Single-Step Call Accuracy• Parameter Extraction Accuracy• Final Response<ul style="list-style-type: none">• Task Adherence• Response Completeness	<ul style="list-style-type: none">• Retrieval Augmented Generation (RAG)<ul style="list-style-type: none">• Retrieval• Groundedness• Relevance• General Evaluators<ul style="list-style-type: none">• Coherence• Fluency <p>Accuracy, precision and recall</p> <ul style="list-style-type: none">• Similarity• F1 score• BLEU/GLEU/ROUGE/METEOR	<ul style="list-style-type: none">• Code-based (rule-based)<ul style="list-style-type: none">• Assertion tests for string match and other criteria• Prompt-based (semantic)<ul style="list-style-type: none">• Off-topic conversations• Friendliness• Competitor mentions

Human feedback — least effort possible

What is the effort that can reasonably be asked from users?

- Drop down
- Free text
- Thumbs-up

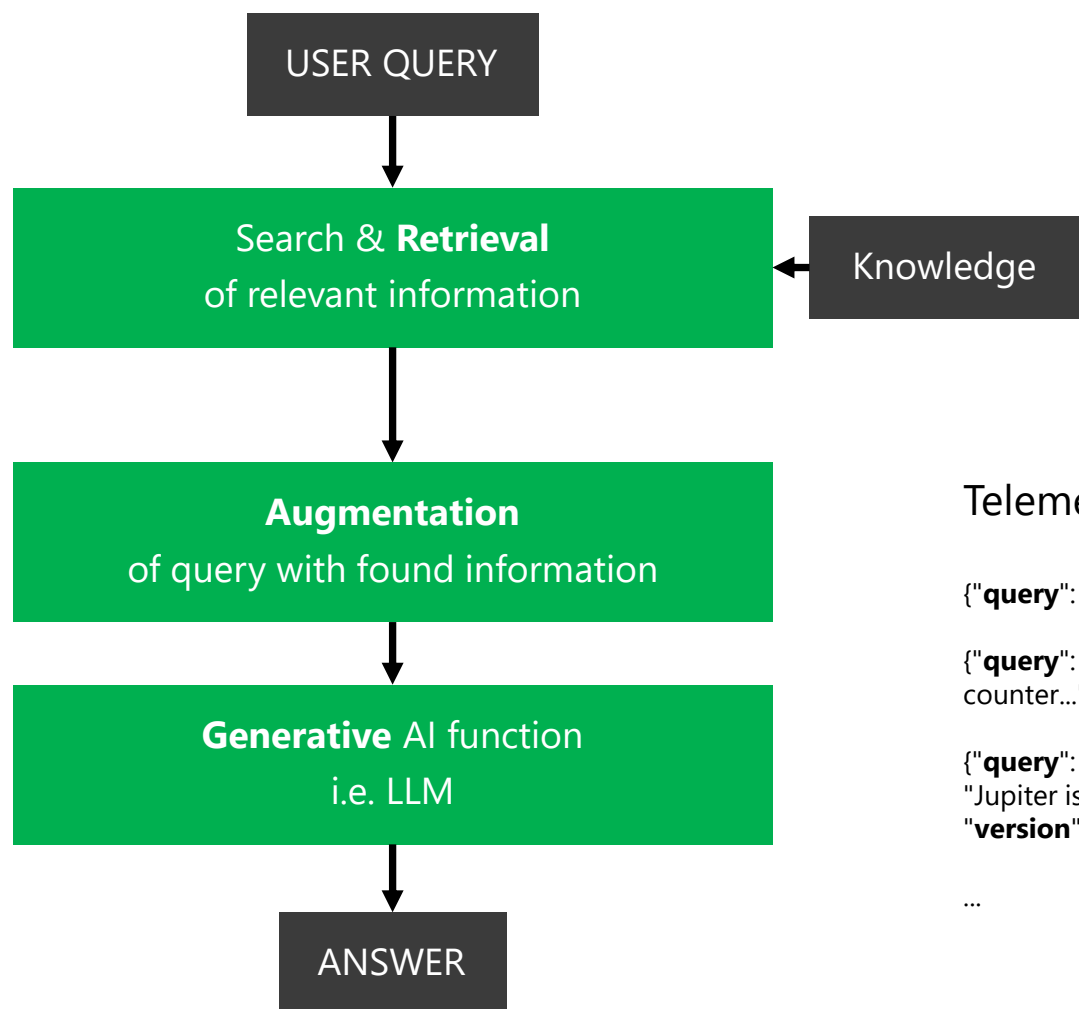


Architecture — Initial telemetry data

Chat application (Dev/Test)

Telemetry data (JSONL)

Working example — RAG (Retrieval augmented generation)



Telemetry / JSONL Format

```
{"query": "...", "ground_truth": "...", "response": "...", "context": ["...", ...], "version": ..., "thumbs_up": ...}
```

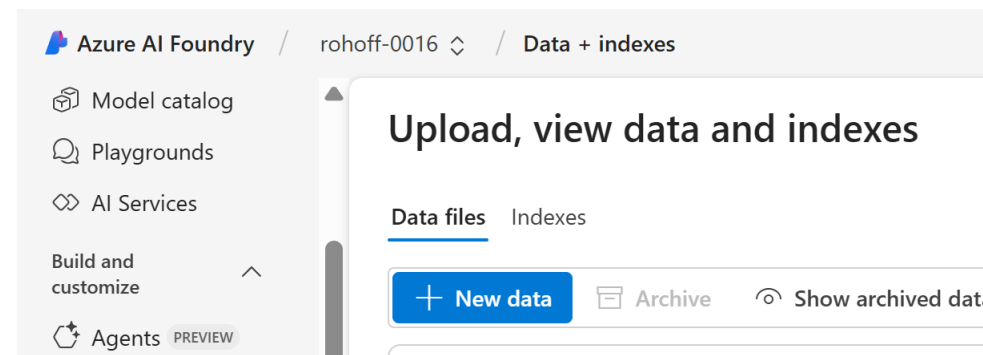
```
{"query": "What is measured by a Geiger counter?", "ground_truth": "Radiation.", "response": "A Geiger counter...", "context": ["The International System of Units (SI)..."], "version": 12, "thumbs_up": true}
```

```
{"query": "Which planet is the biggest in our solar system?", "ground_truth": "Jupiter.", "response": "Jupiter is the biggest planet...", "context": ["Jupiter is the fifth planet from the Sun and the largest..."], "version": 12, "thumbs_up": true}
```


...

Hands-on

- VSCode looking at sample telemetry data, JSONL
- Using Excel to explore JSONL superficially
- Using AI Foundry to upload data



- Evaluation > New Evaluation
- Sideline Discuss Metrics Options

 Evaluate an existing query-response dataset

Qualitative Metrics for Evaluation

Qualitative metrics focus on the human experience, assessing aspects like readability, creativity, and appropriateness.

Metric	Definition	Example	Interpretation
Coherence	The degree to which the output is logically consistent and makes sense.	Assess whether the generated output for an AI creative writing assistant follows a logical sequence and maintains consistent character development throughout the story	Assessed on a scale of 1.0–5.0. High coherence means the content is easy to follow and understand as a whole.
Fluency	The smoothness and readability of the output, with correct grammar and syntax.	Assessing a chatbot's responses to ensure they're grammatically correct and easy to read.	Assessed on a scale of 1.0–5.0. High fluency means the text flows well and sounds natural to a native speaker.
Groundedness	The extent to which the output is based on factual information or given context.	Verifying that a generated news article accurately reflects the facts and sources provided.	Assessed on a scale of 1.0–5.0. High groundedness means the model's output is factually accurate and consistent with the given context or known information.
Groundedness Pro	Detects whether the generated text response is consistent or accurate with respect to the given context in a retrieval-augmented generation question and answering scenario.	When you need to verify that AI-generated responses align with and are validated by the provided context. It's essential for applications where contextual accuracy is key, like information retrieval and question and answering.	False if response is ungrounded and True if it's grounded
Relevance	How well the output aligns with the given context or user query.	Assess how well generated article summaries match users' interests for an AI-powered personalized news aggregator.	Assessed on a scale of 1.0–5.0. High relevance means the content is closely aligned with the user's intent or the subject matter being discussed.
Retrieval	Measures the quality of search without ground truth. It focuses on how relevant the context chunks (encoded as a string) are to address a query and how the most relevant context chunks are surfaced at the top of the list.	Suitable for applications where the model engages in generation using a retrieval-augmented approach to extract information from your provided documents and generate detailed responses, usually multi-turn.	Assessed on a scale of 1.0–5.0 High quality means the output is highly relevant, well ranked, and no bias is introduced.

Quantitative Metrics for Evaluation

Quantitative metrics are often data-driven and based on specific algorithms or statistical analysis.

Metric	Definition	Example	Interpretation
Similarity	The degree of resemblance between the generated output and the reference text.	Assess that the AI-generated content aligns with established legal practices and terminology by comparing AI generated contract clauses with a database of standard legal language.	Assessed on a scale of 1.0–5.0 A high similarity score indicates that the compared texts or concepts have similar meanings or convey essentially the same information, even if using different words.
F1 Score	A measure of a model's accuracy that combines precision (relevance of retrieved items) and recall (completeness of retrieval).	Assess how well a model correctly identifies and classifies various skin conditions for an AI app that assists doctors in diagnosing skin conditions from images.	Assessed on a scale of 0–1. A high F1 score indicates that the model has low false positives and low false negatives.
ROUGE	Recall-Oriented Understudy for Gisting Evaluation, measures the overlap of n-grams (word sequences) between the generated and reference texts. It's useful for assessing if key information is retained in summaries.	Assess how AI-generated summaries of scientific papers capture key findings from the original research papers, comparing them against human-written reviews.	Assessed on a scale of 0–1. Higher ROUGE scores indicate better coverage of the reference content.
BLEU	Bilingual Evaluation Understudy measures how many words overlap between the machine translation and reference translations, considering exact matches and near matches. It focuses on precision and aims to capture translation adequacy and fluency.	Assess the quality of AI-generated posts in various languages for a multilingual content generation tool for a social media platform. Compare the AI generated posts against human-translated versions.	Assessed on a scale of 0–1. Higher BLEU scores suggest better translation quality.
METEOR	Metric for Evaluation of Translation with Explicit ORdering, designed to improve some of the weaknesses of BLEU. METEOR is an automatic metric for machine translation evaluation. It considers synonyms and paraphrases and aligns words between the machine translation and reference.	An e-learning platform developing an AI tutor that explains complex concepts in simpler terms uses METEOR to evaluate the quality of its explanations. This metric helps assess whether the AI-generated explanations effectively convey the same meaning as expert-written materials, even if using different words.	Assessed on a scale of 0–1. A high METEOR score suggests that the generated text closely matches the reference text in terms of content and meaning.
GLEU	A variant of BLEU developed by Google for evaluating machine translation. GLEU is similar to BLEU but calculates the minimum of precision and recall for n-grams, making it more sensitive to changes in translation quality that affect both precision and recall.	A company creating an AI system for generating product descriptions uses GLEU to fine-tune their model. They compare the AI-generated descriptions against professionally written ones, using GLEU's sensitivity to both precision and recall to incrementally improve the system's ability to create compelling and accurate product narratives.	Assessed on a scale of 0–1. A high GLEU score indicates a high degree of overlap in n-grams between the generated text and reference translations. A high GLEU score generally suggests fluency and adequacy in translation, and good precision in word choice and word order.

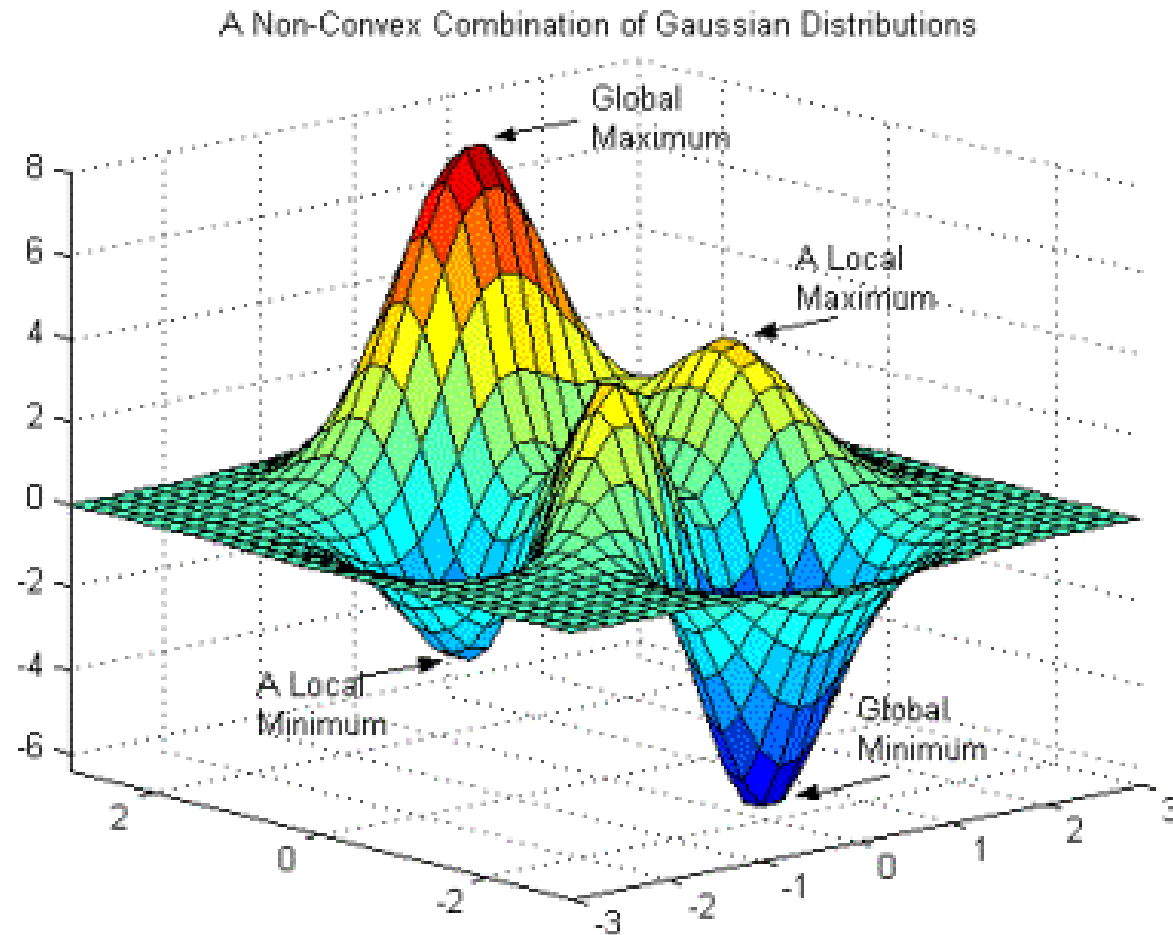
Risk and safety metrics

Metric	What it Measures	Assessment Output
Hateful and Unfair Content Defect Rate	Measures the frequency of AI-generated content that contains hate speech or unfair biases.	Severity level: 0–7, Severity label: Very low, Low, Medium, and High
Sexual Content Defect Rate	Measures the occurrence of AI-generated content containing inappropriate sexual references.	Severity level: 0–7, Severity label: Very low, Low, Medium, and High
Violent Content Defect Rate	Evaluates the frequency of content that depicts or incites violence.	Severity level: 0–7, Severity label: Very low, Low, Medium, and High
Self-Harm Related Content Defect Rate	Measures the generation of content that encourages or glamorizes self-harm.	Severity level: 0–7, Severity label: Very low, Low, Medium, and High
Jailbreak Defect Rate	Measures how often AI systems can be manipulated to bypass safety protocols and restrictions. A jailbreak occurs when a user finds a way to get the AI to produce content it's designed to prevent.	True or False
Indirect Attack Rate	Measures the susceptibility of AI to indirect prompt injections, where seemingly harmless prompts elicit inappropriate responses. An indirect prompt injection occurs when the AI is tricked into generating harmful content through a series of seemingly innocuous prompts.	True or False
Protected Material Defect Rate	Evaluates how often AI systems generate content that infringes upon protected material, such as copyrighted text.	True or False

Hands-on

- Continue Evaluation, select evaluators, i.e. Grounding, Similarity, METEOR
 - Discuss, Report and data
 - Where to go from here?
-
- Show how to download from Logs > instance_results.json
-
- Switch to VSCode > data
 - Explore evaluated file with Excel
 - 10_manual_evaluation_analysis_1_of_2
 - 20_programatic_evaluation_local_dev — why and custom evaluators
 - 30_manual_evaluation_analysis_2_of_2 — check out custom evaluators

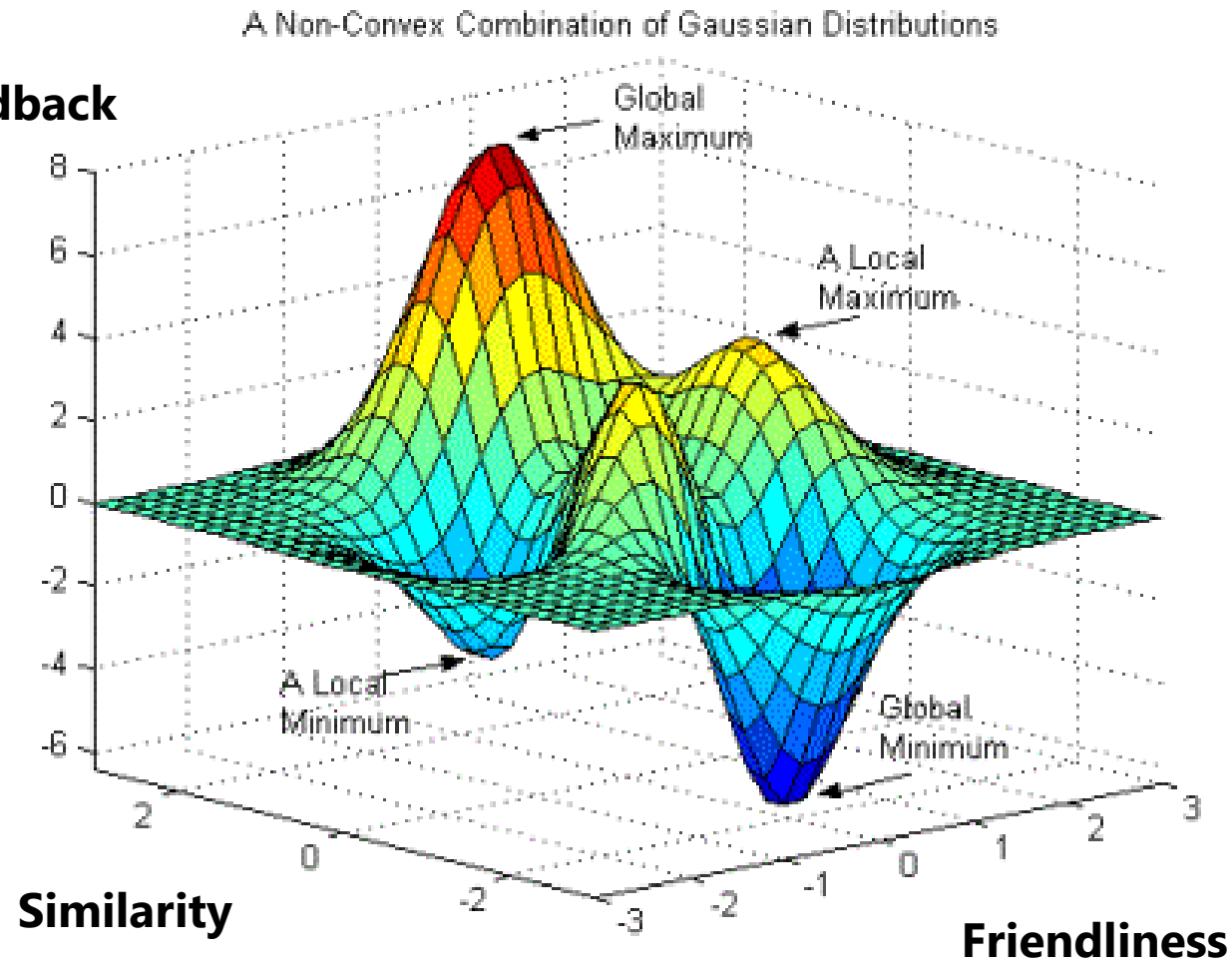
Exploring the optimal configuration space (for an AI application)



https://en.wikipedia.org/wiki/Stochastic_gradient_Langevin_dynamics

Exploring the optimal configuration space (for an AI application)

Neg. human feedback



Architecture — Cloud evaluation

Chat application (Dev/Test)

Telemetry data (JSONL)

Developed custom evaluation (incl. Custom evaluators), Locally

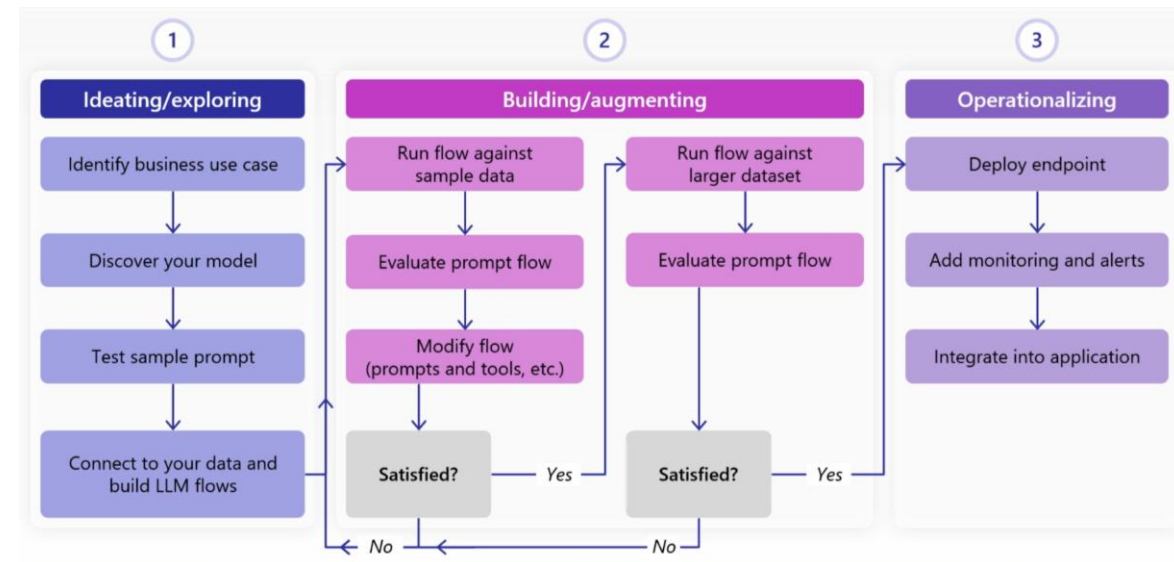
Cloud evaluation incl. custom evaluators

Hands-on

- 40_register_custom_evaluators_for_cloud_evaluation
- 50_cloud_evaluation

Phases of development & evaluation

- Initial ideation
- Productive application development
- Continuous development



Architecture — Tracing and Online evaluation

Chat application (Dev/Test)

Telemetry data (JSONL)

Developed custom evaluation (incl. Custom evaluators), Locally

Cloud evaluation incl. custom evaluators

Tracing

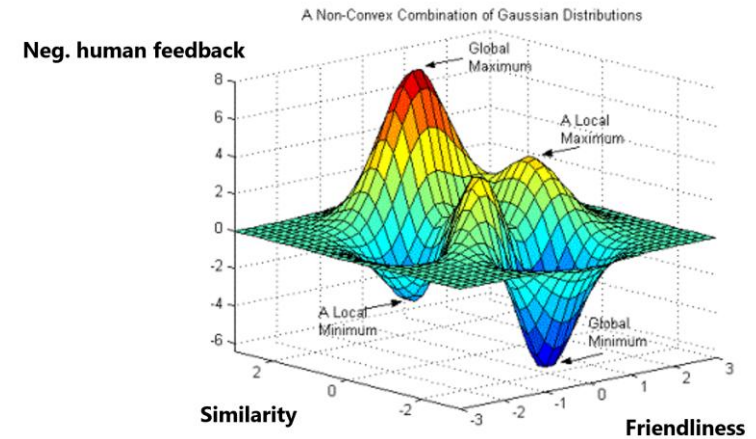
Online cloud evaluation

Hands-on

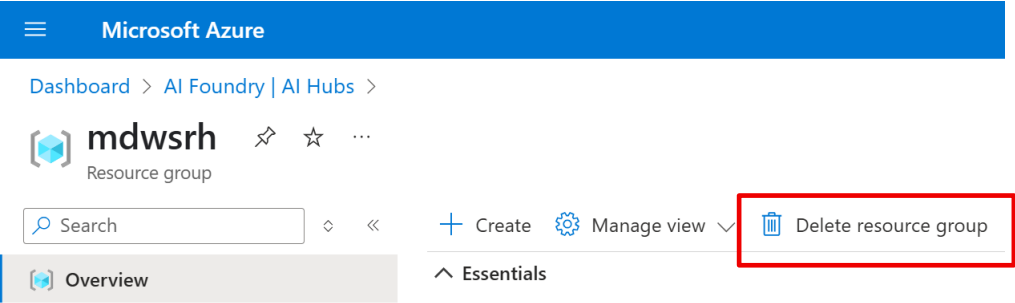
- 60_tracing
- 70_manually_query_traces
- [80_advanced_online_eval]

Evaluation is a continuous and iterative process

- Collect and analyse telemetry data
- Develop & test your hypothesis
- Improve your application
- Keep your progress versioned, documented and organized



Delete resource group — To avoid running costs



Delete a resource group

The following resource group and all its dependent resources will be permanently deleted.

Resource group to be deleted

mdwsrh

Dependent resources to be deleted (10)

All dependent resources, including hidden types, are shown

Name	Resource type
aigsqxiksqmwu	Log Analytics workspace
Application Insights Smart Detection	Action group
mdwsairh	Application Insights
mdwsrh	Azure AI hub
mdwsrh0719056065	Key vault
mdwsrh2913615413	Azure AI Foundry
mdwsrh4995962559	Storage account
mdwsrh4995962559-37cd3fdf-9956-4386-aa70-4;	Event Grid System Topic
rohof-mc90j3yd-eastus	Azure AI Foundry
rohoff-0016	Azure AI project

Enter resource group name to confirm deletion *

Delete

Cancel

Thank you