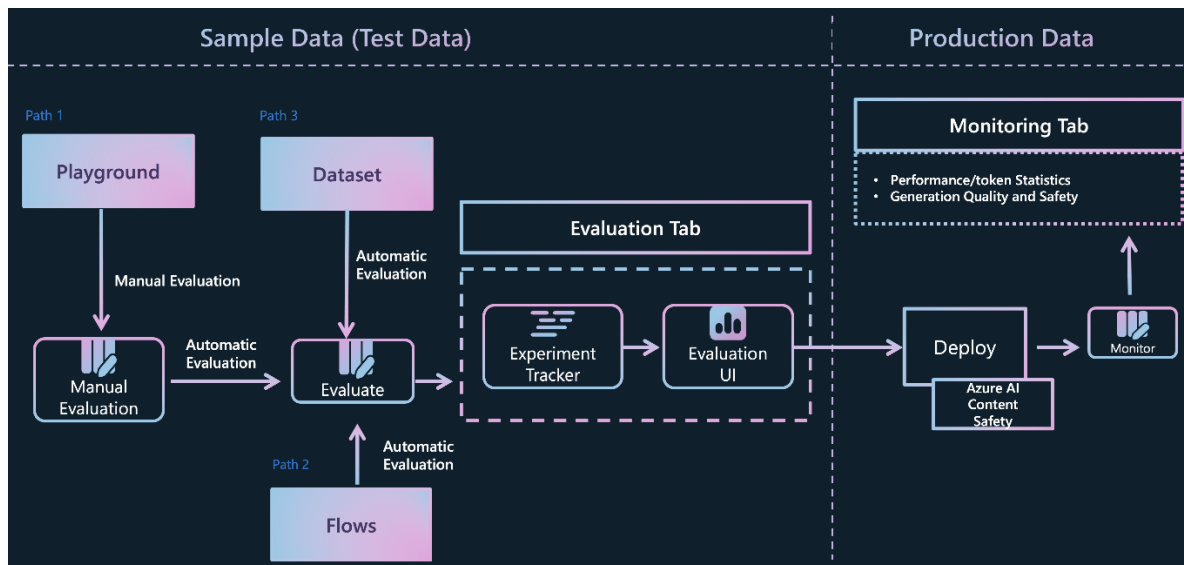


# Evaluating LLM apps with Azure AI Studio

Azure AI Studio supports several distinct paths for generative AI app developers to evaluate their applications:



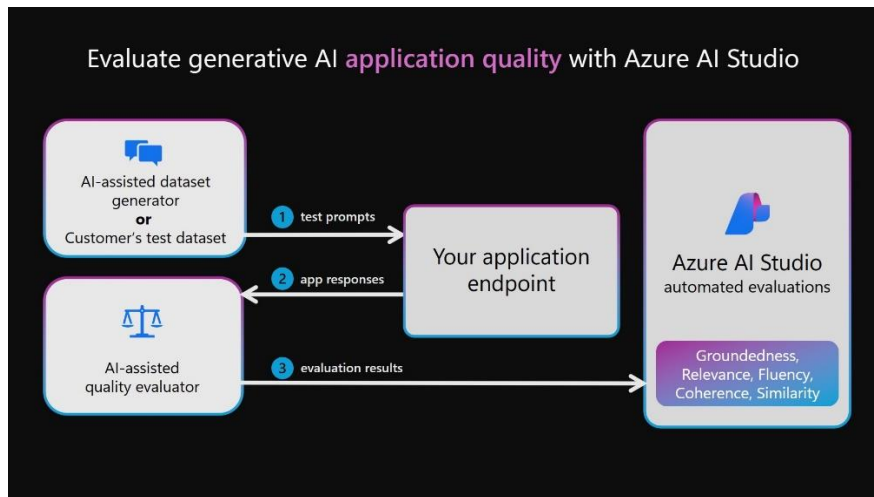
- 1. Playground:** Here, you have the option to select the data you want to use for grounding your model, choose the base model for the application, and provide metaprompt instructions to guide the model's behavior. You can then manually evaluate the application by passing in a dataset and observing the application's responses. Once the manual inspection is complete, you can opt to use the evaluation wizard to conduct more comprehensive assessments.
- 2. Flows:** The Azure AI Studio Prompt flow page offers a dedicated development tool tailored for streamlining the entire lifecycle of AI applications powered by LLMs. With this path, you can create executable flows that link LLMs, prompts, and Python tools through a visualized graph. This feature simplifies debugging, sharing, and collaborative iterations of flows.
- 3. Direct Dataset Evaluation:** If you have collected a dataset containing interactions between your application and end-users, you can submit this data directly to the evaluation wizard within the "Evaluation" tab. This process enables the generation of automatic AI-assisted evaluations, and the results can be visualized in the same tab. This approach centers on a data-centric evaluation method. Alternatively, you have the option to evaluate your conversation dataset using the SDK/CLI experience and generate and visualize evaluations through the Azure AI Studio.

Azure AI Studio provides practitioners with tools for manual and automated evaluation that can help you with the measurement stage. We recommend that you start with manual evaluation then proceed to automated evaluation.

Automated evaluation is useful for measuring quality and safety at scale with increased coverage to provide more comprehensive results. Automated evaluation tools also enable ongoing evaluations that periodically run to monitor for regression as the system, usage, and mitigations evolve. We support two main methods for automated evaluation of generative AI applications: traditional machine learning evaluations and AI-assisted evaluation.

# Introduction to AI-assisted evaluations

Large language models (LLM) such as GPT-4 can be used to evaluate the output of generative AI language systems. This is achieved by instructing an LLM to annotate certain aspects of the AI-generated output. For instance, you can provide GPT-4 with a relevance severity scale (for example, provide criteria for relevance annotation on a 1-5 scale) and then ask GPT-4 to annotate the relevance of an AI system's response to a given question.



To run AI-assisted performance and quality evaluations, an LLM is possibly leveraged for two separate functions. First, a test dataset must be created. This can be created manually by choosing prompts and capturing responses from your AI system, or it can be created synthetically by simulating interactions between your AI system and an LLM (referred to as the AI-assisted dataset generator in the following diagram). Then, an LLM is also used to annotate your AI system's outputs in the test set. Finally, annotations are aggregated into performance and quality metrics and logged to your Azure AI studio project for viewing and analysis.

## Steps to evaluate an LLM application with Azure AI Studio

1. Navigate to the Evaluation tab in Azure AI Studio.
2. Click New evaluation to create new LLM evaluation.

The screenshot shows the "Create a new evaluation" form in Azure AI Studio. The form has a sidebar with four steps: "Basic information" (selected), "Configure test data", "Select metrics", and "Review and finish". The main content area is titled "Add basic information" and includes the following fields and options:

- Evaluation name \***: A text input field containing "evaluation\_hrbenefits\_flow\_qa".
- What kind of scenario are you evaluating? \***: A radio button selection with three options:
  - ☒ **Question and answer with context**: Evaluate single-turn question and answer pairs with context. Below this is a link "Download data template".
  - ☐ **Question and answer without context**: Evaluate single-turn question and answer pairs without context. Below this is a link "Download data template".
  - ☐ **Conversation with context**: Evaluate a single-turn or multi-turn conversation with retrieved documents. Below this is a link "Download data template".
- Select a flow to evaluate (optional)**: A section titled "Which flow do you want to evaluate?" with a text input field containing "Flow-hrbenefits-rag-v1".

At the bottom of the form are "Back" and "Next" buttons.

3. Configure test data. You can add sample test data or use your own test data set. You could use the provided dataset template that aligns with your LLM apps evaluation scenario.

**Configure test data**

Select configuration test data to evaluate \*

Use existing dataset  
Choose from your established dataset collection

Add your dataset  
Upload a file

Choose your existing dataset \*

hrbenefits-qa-with-chathistoryblank (version 1) ▼

Preview of top 3 rows from your dataset

question	truth	chat_history
What is the importance of choosing the right provider in getting ...	Choosing the right provider is an important part of getting the m...	
What should you do when choosing an in-network provider for y...	When choosing an in-network provider for your health care need...	
What range of in-network providers does Northwind Health Plus ...	Northwind Health Plus offers a wide range of in-network provider...	

Dataset mapping for prompt flow \* ⓘ

Name	Type	Value
chat_history	list	<input type="text" value="\${data.chat_history}"/>
query	string	<input type="text" value="\${data.question}"/>

Back

Next

Cancel

4. Configure evaluation metrics. Refer to the documentation for supported evaluation metrics - [Evaluation and monitoring metrics for generative AI - Azure AI Studio | Microsoft Learn](#)

/ Project-amulazuremlws-aoai / Evaluation / Create a new evaluation

Performance and quality metrics curated by Microsoft

☒ **Coherence**  
Measures how well the language model can produce output that flows smoothly, reads naturally, and resembles human-like language.

☐ **Fluency**  
Measure the language proficiency of a generative AI's predicted answer.

☒ **GPT similarity**  
Measures the similarity between a source data (ground truth) sentence and the generated response by a GPT-based AI model.

☒ **F1 score**  
Measures the ratio of the number of shared words between the model prediction and the source data (ground truth).

Connection	Provider	Deployment name/Model
aoaiamulcanadae ▼	AzureOpenAI	gpt-4-turbo-1106preview ▼

Risk and safety metrics curated by Microsoft

How does your dataset map to your evaluation input? \* ⓘ

Name	Description	Type	Data source
answer	The response to question generated by the model as answer	string	<input type="text" value="\${run.outputs.reply}"/>
question	A query seeking specific information	string	<input type="text" value="\${data.question}"/>
ground_truth	The response to question generated by user/human as the true answer	string	<input type="text" value="\${data.truth}"/>

5. Review your evaluation job configuration.

Review your data

Basic information

Evaluation name  
evaluation\_hrbenefits\_flow\_qa

What kind of scenario are you evaluating?  
Question and answer with context

Which flow do you want to evaluate?  
Flow-hrbenefits-tag-v1

Which node do you want to select for variants?  
...

Which variants do you want to evaluate?  
...

Configure test data

Choose your existing dataset  
hrbenefits-qa-with-chathistory/blank (version 1)

Dataset mapping for prompt flow

Name	Type	Value
chat_history	list	\$(data.chat_history)
query	string	\$(data.question)

Select metrics

Select metrics  
Groundedness, Relevance, GPT similarity, Fluency, Coherence

Deployment name/Model  
gpt-35-turbo-16k

How does your dataset map to your evaluation input?

Name	Description	Type	Data source
answer	The response to question generated by the model as answer	string	\$(run.outputs.reply)
context	The source that response is generated with respect to	string	\$(run.outputs.fetched_docs)
question	A query seeking specific information	string	\$(data.question)
ground_truth	The response to question generated by user/human as the true answer	string	\$(data.truth)

6. Once you submit the evaluation run, it will get queued and you will see evaluation job status and detailed results under the Evaluation tab in Azure AI Studio.

Assess and compare AI application performance

Metric evaluations

Manual evaluations

Evaluate your model performance with industry standard metrics to compare and choose the best version based on your need. [Learn more about metrics.](#)

+ New evaluation

Refresh

Cancel

Delete

Compare

Show batch runs

Switch to dashboard view

Evaluations	Status	Created on ↓	Groundedness	Relevance	Retrieval score	Coherence	Similarity	Fluency	F1 score	Violence defe...	Se
evaluation_evaluation_hrbenefits_flow_qa_varia	Completed	Apr 9, 2024 6:40 PM	3.84	4.42	--	4.49	4.3	3.76	--	--	--
evaluation_evaluate_hrbenefits_qa_variant_0	Completed	Apr 9, 2024 5:24 PM	--	--	--	4.78	3.67	--	0.41	--	--

7. Review details of evaluation results – sample below (your results would vary based on your LLM apps, test data etc).

