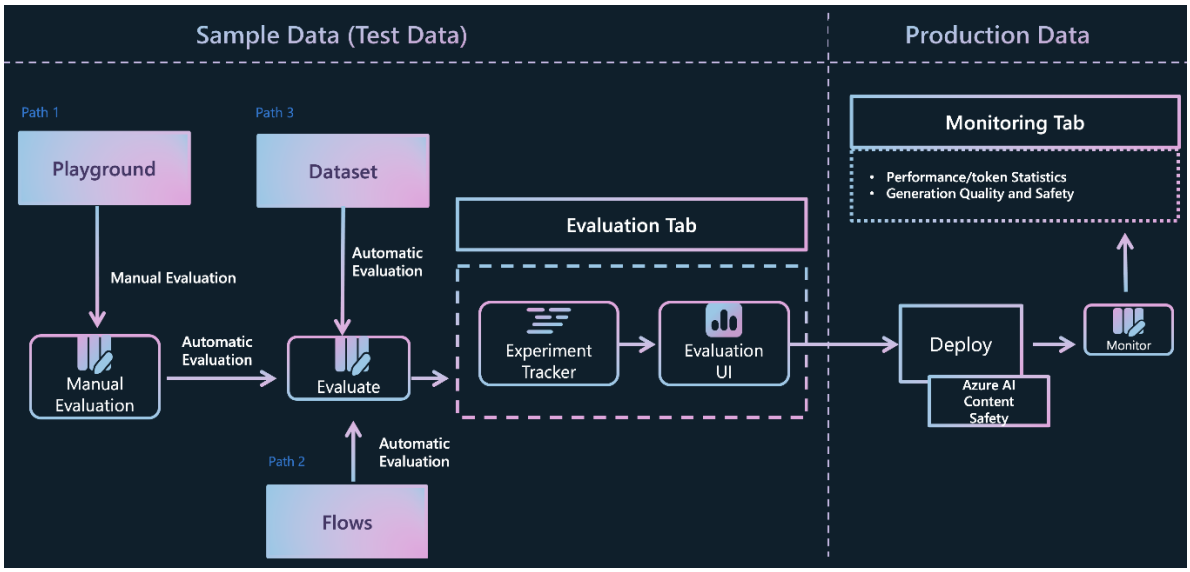
Evaluating LLM apps with Azure AI Studio

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.

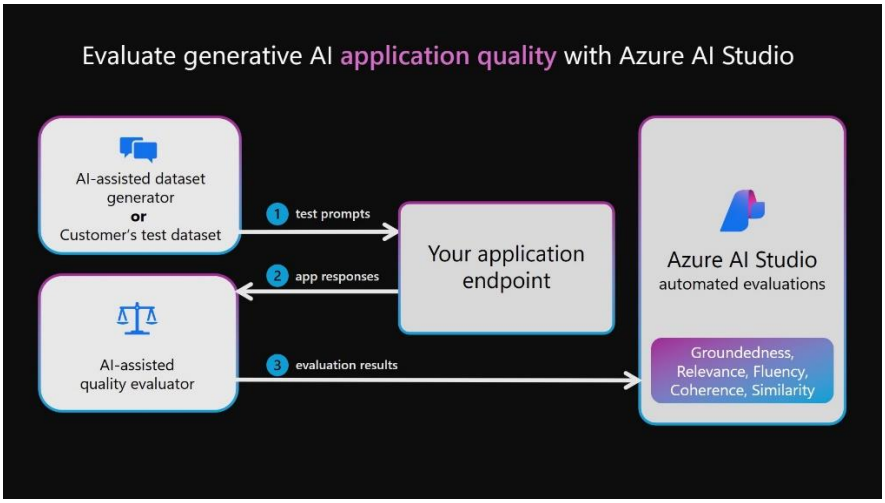
3. **Direct Dataset Evaluation**: If you have collected a dataset containing interactions between your application and end-users, you can submit this data directly to the evaluation wizard within the "Evaluation" tab. This process enables the generation of automatic AI-assisted evaluations, and the results 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

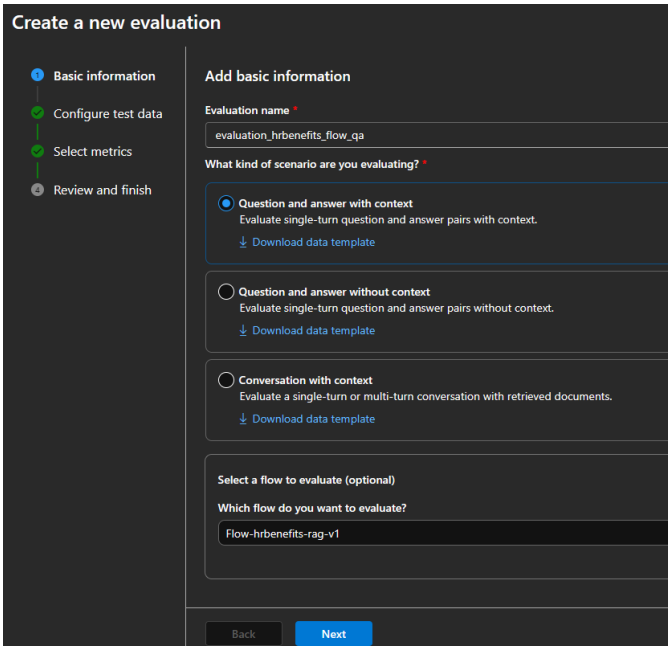
Large language models (LLM) such as GPT-4 can be used to evaluate the output of generative AI language systems. This is achieved by instructing an LLM to annotate certain aspects of the AI-generated output. For instance, you can provide GPT-4 with a relevance severity scale (for example, provide criteria 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.



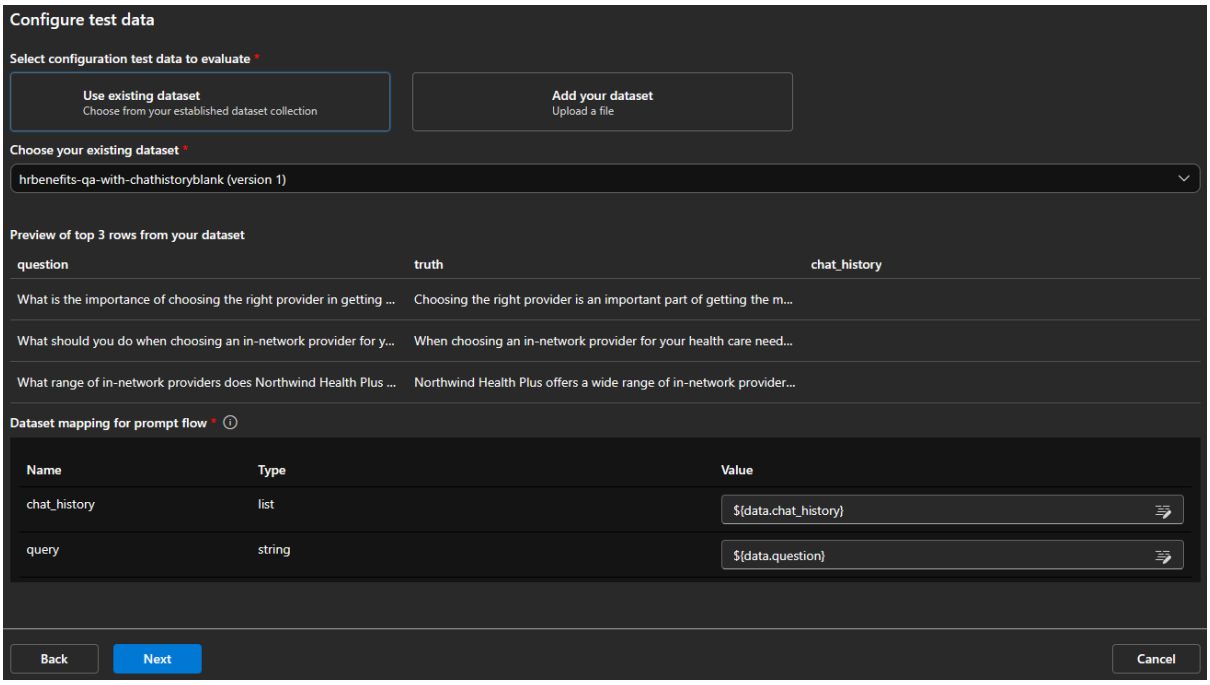
To run AI-assisted performance and quality evaluations, an LLM is possibly leveraged for two separate functions. First, a test dataset must be created. This can be created manually by choosing prompts and capturing responses from your AI system, or it can be created synthetically by simulating interactions 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

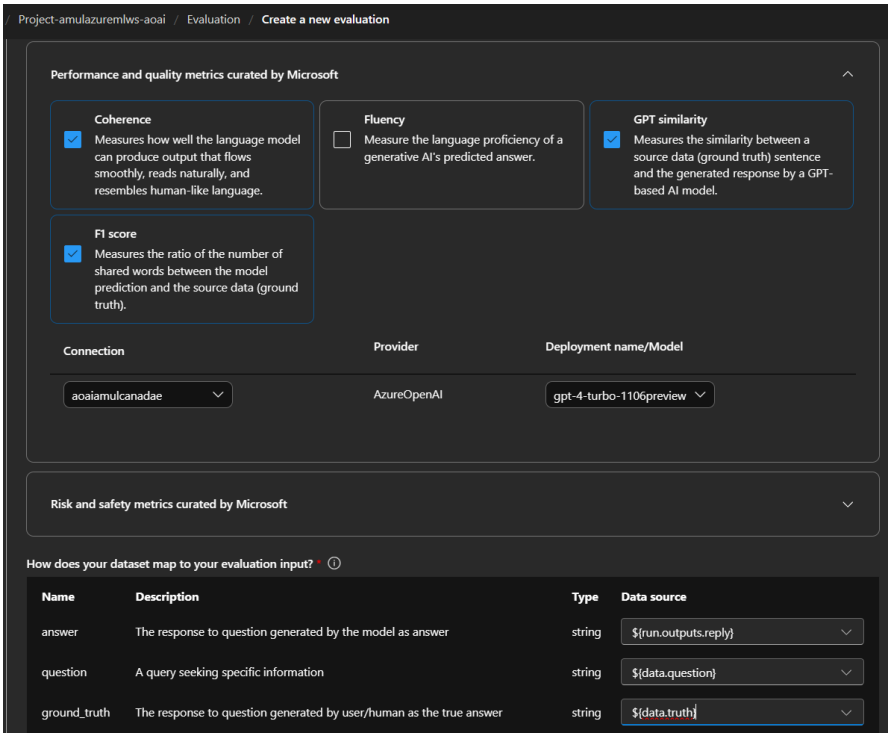
1. Navigate to the Evaluation tab in Azure AI Studio.
2. Click New evaluation to create new LLM evaluation.



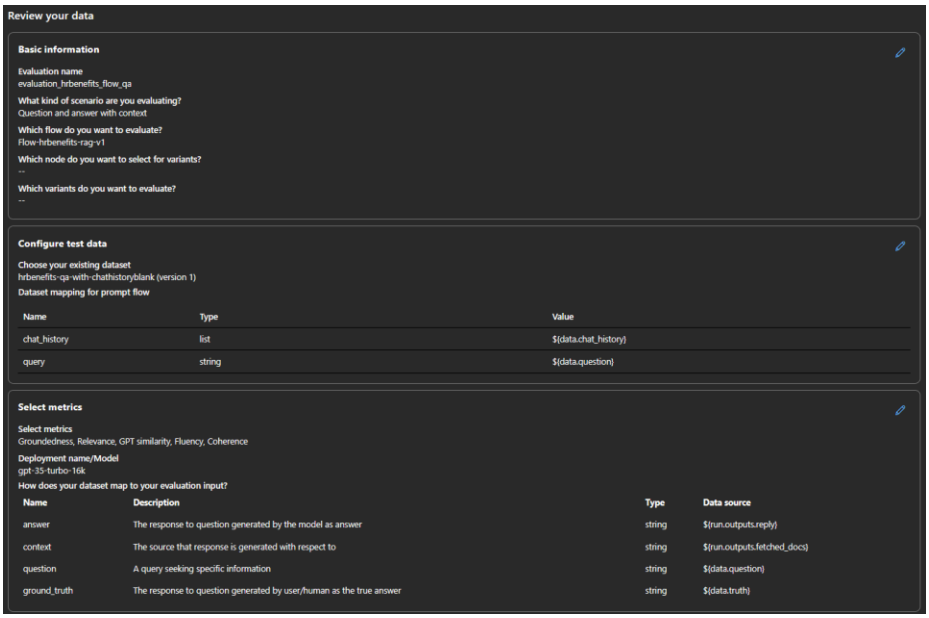
3. Configure test data. You can add sample test data or use your own test data set. You could use the provided dataset template that aligns with your LLM apps evaluation scenario.



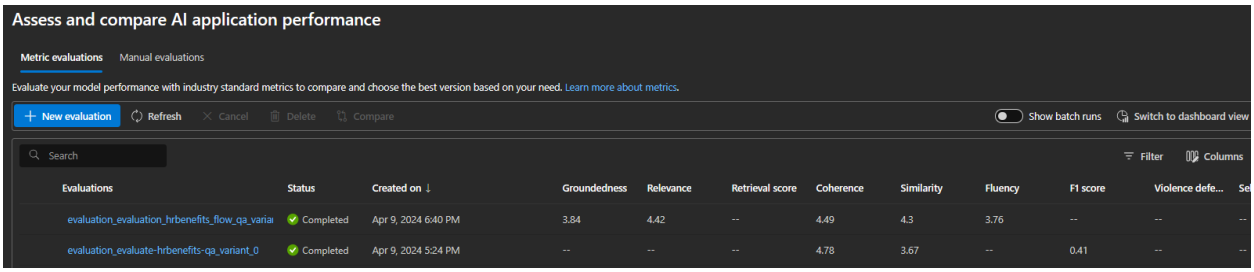
4. Configure evaluation metrics. Refer to the documentation for supported evaluation metrics - [Evaluation and monitoring metrics for generative AI - Azure AI Studio | Microsoft Learn]



5. Review your evaluation job configuration.



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.



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