Evaluate generative AI performance in Microsoft Foundry portal

1. Introduction

https://learn.microsoft.com/en-us/training/modules/evaluate-models-azure-ai-studio/1-introduction

Introduction

Completed

- 2 minutes

Evaluating your generative AI apps is crucial for several reasons. First and foremost, it ensures quality assurance. By assessing your app's performance, you can identify and address any issues, ensuring that it provides accurate and relevant responses. High quality responses lead to improved user satisfaction. When users receive accurate and helpful responses, they're more likely to have a positive experience and continue using your application.

Evaluation is also essential for continuous improvement. By analyzing the results of your evaluations, you can identify areas for enhancement and iteratively improve your app's performance. The ongoing process of evaluation and improvement helps you stay ahead of user needs and expectations, ensuring that your app remains effective and valuable.

In this module, you learn how to use the Microsoft Foundry portal to evaluate your generative AI apps. While you explore some of the features of Microsoft Foundry, the focus is on understanding the importance of evaluation and how it can benefit your app development process.

2. Assess the model performance

https://learn.microsoft.com/en-us/training/modules/evaluate-models-azure-ai-studio/2-assess-models

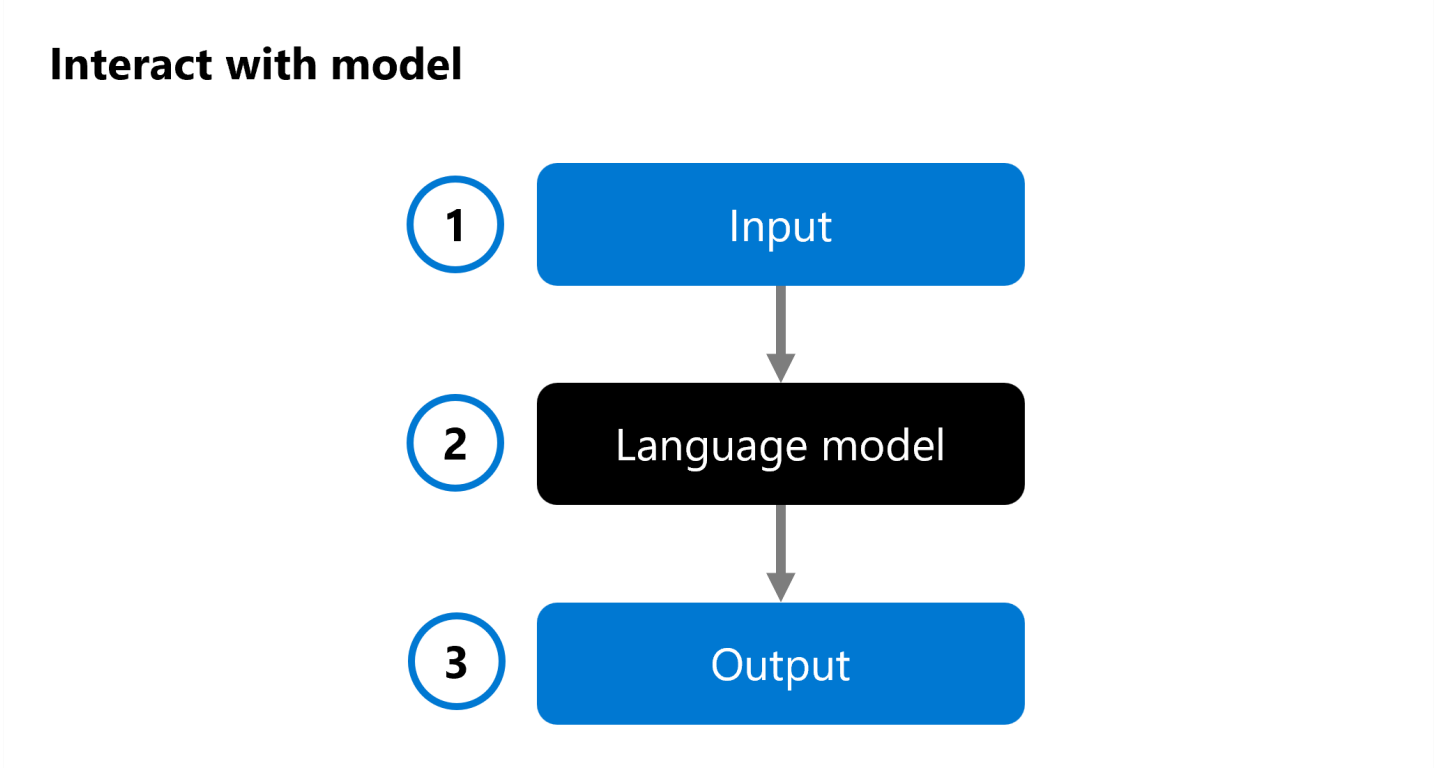
Assess the model performance

Completed

- 6 minutes

Evaluating your model's performance at different phases is crucial to ensure its effectiveness and reliability. Before exploring the various options you have to evaluate your model, let's explore the aspects of your application you can evaluate.

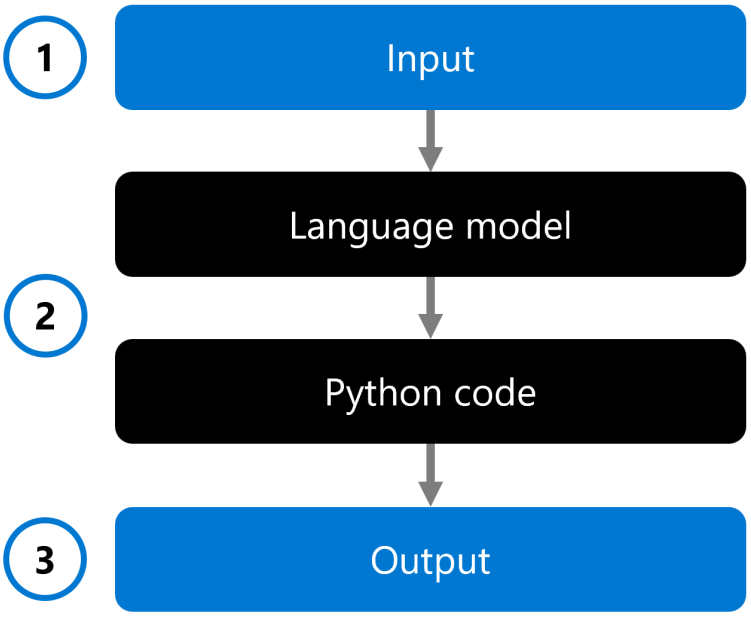
When you develop a generative AI app, you use a language model in your chat application to generate a response. To help you decide which model you want to integrate into your application, you can evaluate the performance of an individual language model:



An input (1) is provided to a language model (2), and a response is generated as output (3). The model is then evaluated by analyzing the input, the output, and optionally comparing it to predefined expected output.

When you develop a generative AI app, you may integrate a language model into a chat flow:

Chat flow



A chat flow allows you to orchestrate executable flows that can combine multiple language models and Python code. The flow expects an input (1), processes it through executing various nodes (2), and generates an output (3). You can evaluate a complete chat flow, and its individual components.

When evaluating your solution, you can start with testing an individual model, and eventually test a complete chat flow to validate whether your generative AI app is working as expected.

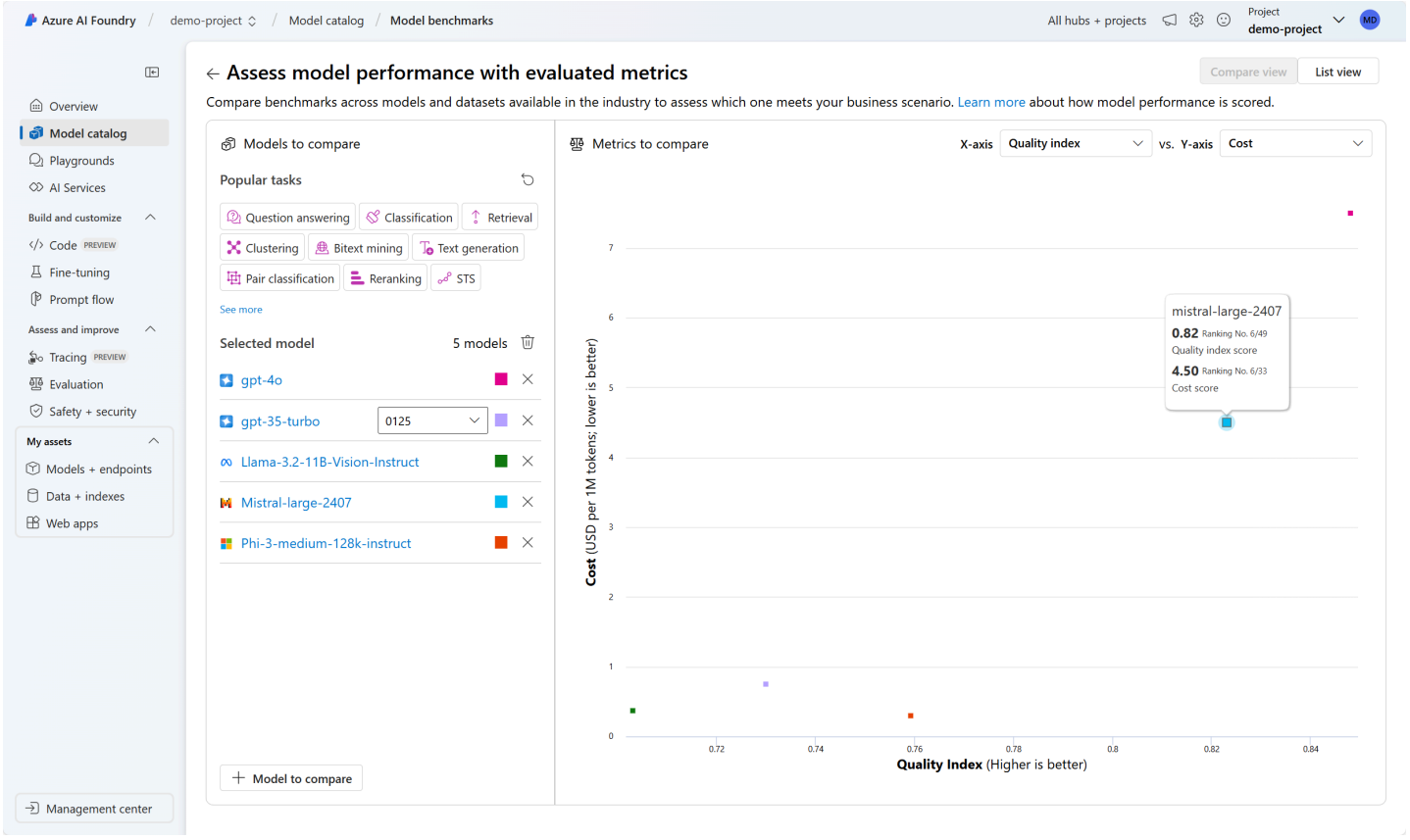
Let's explore several approaches to evaluate your model and chat flow, or generative AI app.

Model benchmarks

Model benchmarks are publicly available metrics across models and datasets. These benchmarks help you understand how your model performs relative to others. Some commonly used benchmarks include:

- **Accuracy**: Compares model generated text with correct answer according to the dataset. Result is one if generated text matches the answer exactly, and zero otherwise.
- **Coherence**: Measures whether the model output flows smoothly, reads naturally, and resembles human-like language
- **Fluency**: Assesses how well the generated text adheres to grammatical rules, syntactic structures, and appropriate usage of vocabulary, resulting in linguistically correct and natural-sounding responses.
- **GPT similarity**: Quantifies the semantic similarity between a ground truth sentence (or document) and the prediction sentence generated by an AI model.

In the Microsoft Foundry portal, you can explore the model benchmarks for all available models, before deploying a model:



Manual evaluations

Manual evaluations involve human raters who assess the quality of the model's responses. This approach provides insights into aspects that automated metrics might miss, such as context relevance and user satisfaction. Human evaluators can rate responses based on criteria like relevance, informativeness, and engagement.

AI-assisted metrics

AI-assisted metrics use advanced techniques to evaluate model performance. These metrics can include:

- **Generation quality metrics**: These metrics evaluate the overall quality of the generated text, considering factors like creativity, coherence, and adherence to the desired style or tone.
- **Risk and safety metrics**: These metrics assess the potential risks and safety concerns associated with the model's outputs. They help ensure that the model doesn't generate harmful or biased content.

Natural language processing metrics

Natural language processing (NLP) metrics are also valuable in evaluating model performance. One such metric is the **F1-score**, which measures the ratio of the number of shared words between the generated and ground truth answers. The F1-score is useful for tasks like text classification and information retrieval, where precision and recall are important. Other common NLP metrics include:

- **BLEU**: Bilingual Evaluation Understudy metric
- **METEOR**: Metric for Evaluation of Translation with Explicit Ordering
- **ROUGE**: Recall-Oriented Understudy for Gisting Evaluation

All of these metrics are used to quantify the level of overlap in the model-generated response and the ground truth (expected response).

3. Manually evaluate the performance of a model

https://learn.microsoft.com/en-us/training/modules/evaluate-models-azure-ai-studio/3-manual-evaluations

Manually evaluate the performance of a model

Completed

- 7 minutes

During the early phases of the development of your generative AI app, you want to experiment and iterate quickly. To easily assess whether your selected language model and app, created with prompt flow, meet your requirements, you can manually evaluate models and flows in the Microsoft Foundry portal.

Even when your model and app are already in production, manual evaluations are a crucial part of assessing performance. As manual evaluations are done by humans, they can provide insights that automated metrics might miss.

Let's explore how you can manually evaluate your selected models and app in the Microsoft Foundry portal.

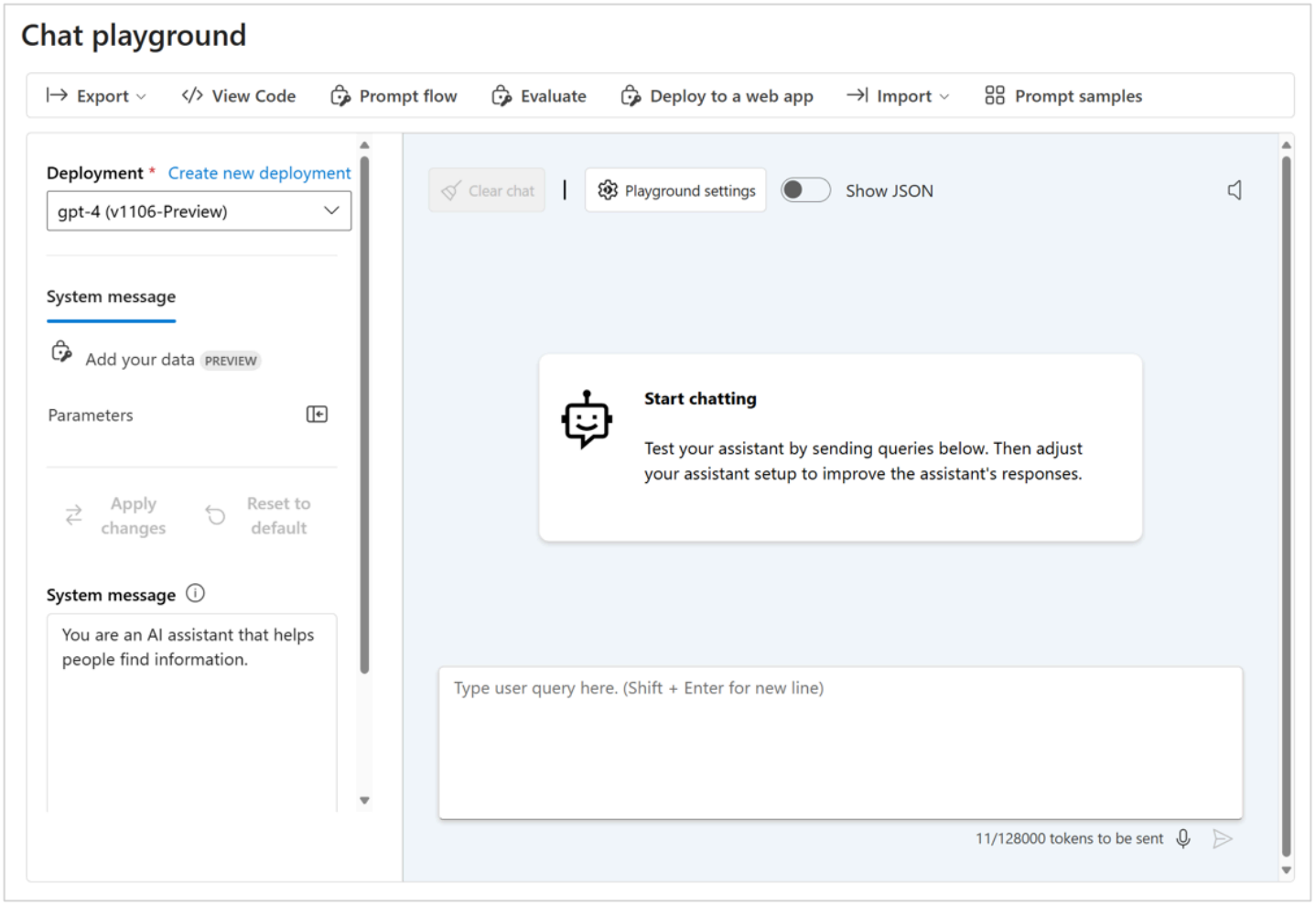
Prepare your test prompts

To begin the manual evaluation process, it's essential to prepare a diverse set of test prompts that reflect the range of queries and tasks your app is expected to handle. These prompts should cover various scenarios, including common user questions, edge cases, and potential failure points. By doing so, you can comprehensively assess the app's performance and identify areas for improvement.

Test the selected model in the chat playground

When you develop a chat application, you use a language model to generate a response. You create a chat application by developing a prompt flow that encapsulates your chat application's logic, which can use multiple language models to ultimately generate a response to a user question.

Before you test your app's response, you can test the selected language model's response to verify the individual model works as expected. You can test a model you deployed in the Microsoft Foundry portal by interacting with it in the **chat playground**.

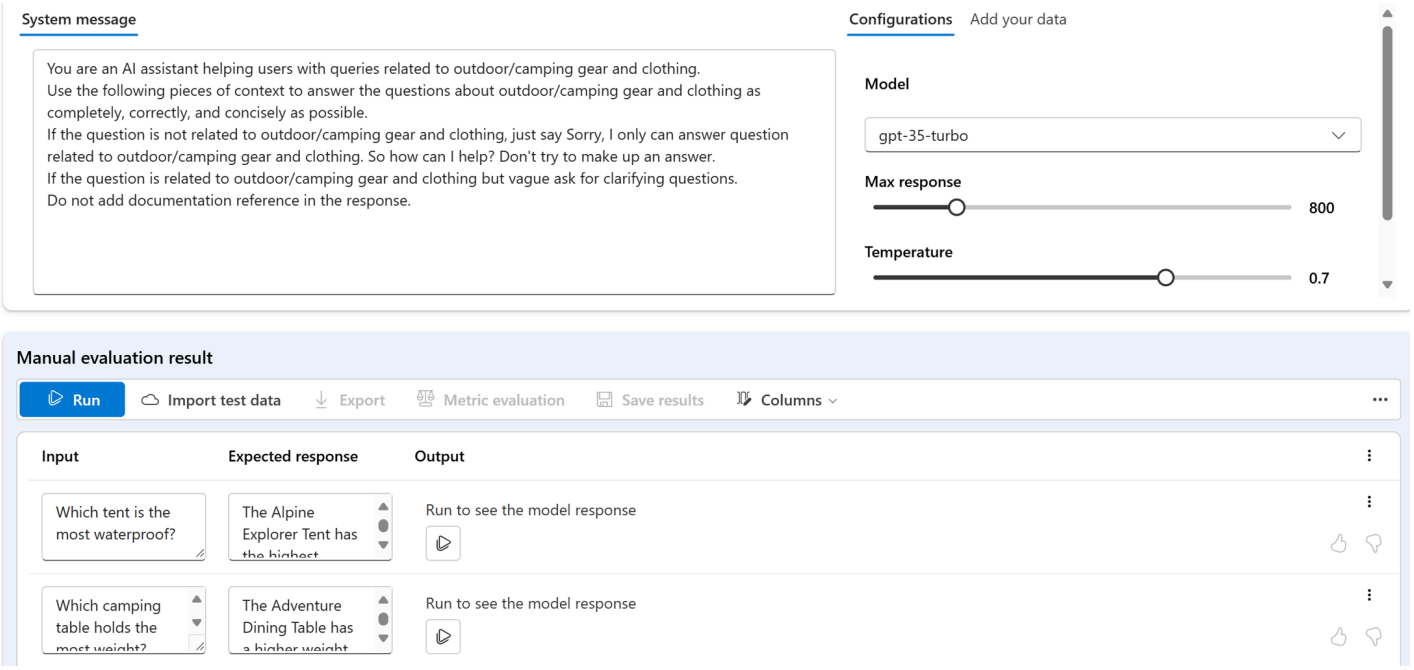


The chat playground is ideal for early development. You can enter a prompt, see how the model responds, and tweak the prompt or system message to make improvements. After applying the

changes, you can test a prompt again to evaluate whether the model's performance indeed improved.

Evaluate multiple prompts with manual evaluations

The chat playground is an easy way to get started. When you want to manually evaluate multiple prompts more quickly, you can use the **manual evaluations** feature. This feature allows you to upload a dataset with multiple questions, and optionally add an expected response, to evaluate the model's performance on a larger test dataset.



You can rate the model's responses with the thumbs up or down feature. Based on the overall rating, you can try to improve your model by changing input prompt, the system message, the model, or the model's parameters.

When you use manual evaluations, you can more quickly evaluate the model's performance based on a diverse test dataset and improve the model based on the test results.

After manually evaluating an individual model, you can integrate the model into a chat application with prompt flow. Any flow you create with prompt flow can also be evaluated manually or automatically. Next, let's explore the evaluation of flows.

4. Automated evaluations

https://learn.microsoft.com/en-us/training/modules/evaluate-models-azure-ai-studio/3b-automated-evaluations

Automated evaluations

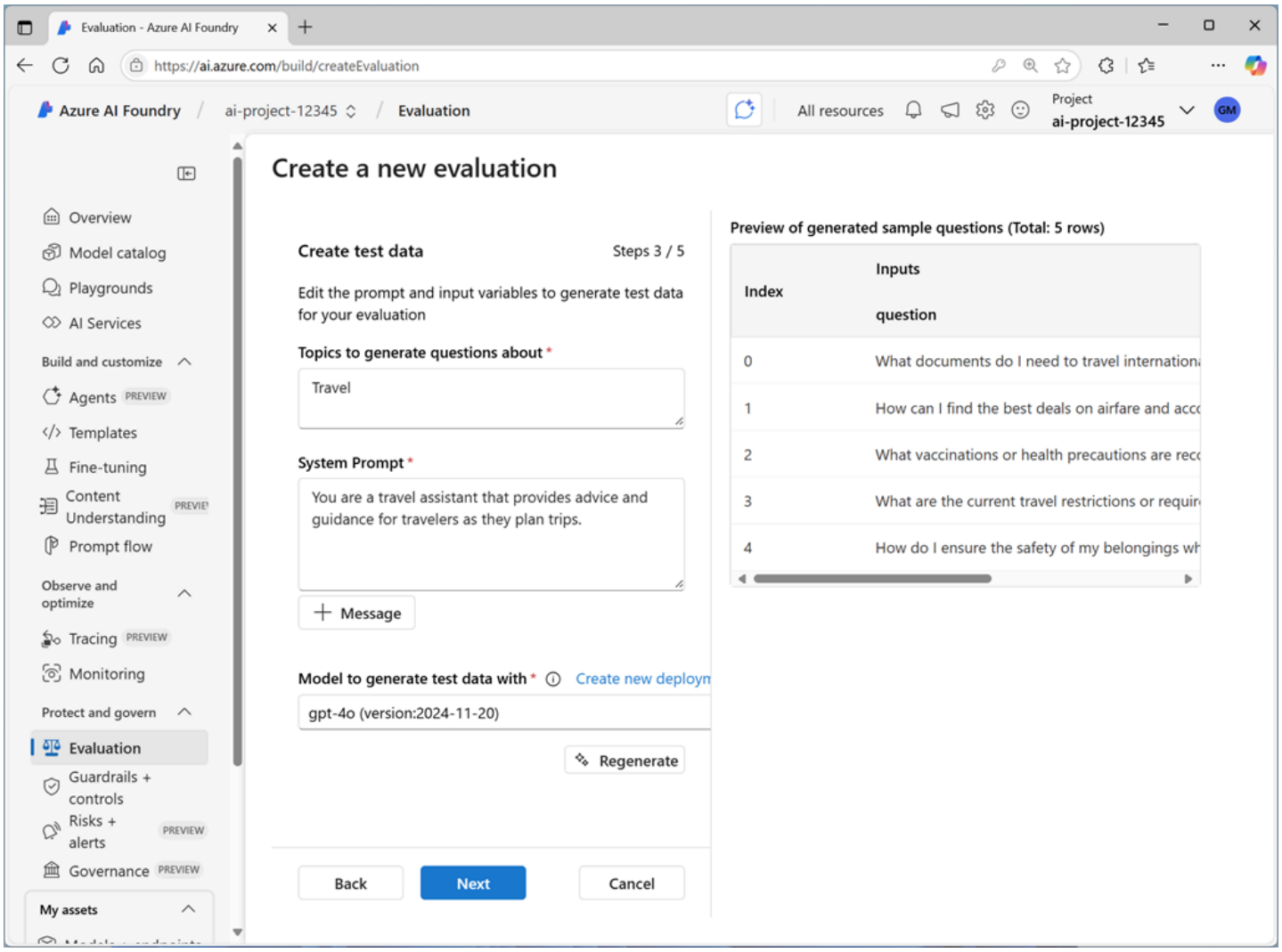
Completed

- 4 minutes

Automated evaluations in Microsoft Foundry portal enable you to assess the quality and content safety performance of models, datasets, or prompt flows.

Evaluation data

To evaluate a model, you need a dataset of prompts and responses (and optionally, expected responses as "ground truth"). You can compile this dataset manually or use the output from an existing application; but a useful way to get started is to use an AI model to generate a set of prompts and responses related to a specific subject. You can then edit the generated prompts and responses to reflect your desired output, and use them as ground truth to evaluate the responses from another model.



Evaluation metrics

Automated evaluation enables you to choose which *evaluators* you want to assess your model's responses, and which metrics those evaluators should calculate. There are evaluators that help you measure:

- **AI Quality**: The quality of your model's responses is measured by using AI models to evaluate them for metrics like *coherence* and *relevance* and by using standard NLP metrics like F1 score, BLEU, METEOR, and ROUGE based on ground truth (in the form of expected response text)
- **Risk and safety**: evaluators that assess the responses for content safety issues, including violence, hate, sexual content, and content related to self-harm.

5. Exercise - Evaluate generative AI model performance

https://learn.microsoft.com/en-us/training/modules/evaluate-models-azure-ai-studio/5-exercise

Exercise - Evaluate generative AI model performance

Completed

- 15 minutes

If you have an Azure subscription, you can use Microsoft Foundry portal to evaluate the performance of a generative AI app.

> **Note**
>
> If you don't have an Azure subscription, and you want to explore Azure AI Studio, you can sign up for an account, which includes credits for the first 30 days.

Launch the exercise and follow the instructions.

Launch Exercise

6. Module assessment

https://learn.microsoft.com/en-us/training/modules/evaluate-models-azure-ai-studio/6-knowledge-check

Module assessment

Completed

- 3 minutes

7. Summary

https://learn.microsoft.com/en-us/training/modules/evaluate-models-azure-ai-studio/7-summary

Summary

Completed

- 1 minute

In this module, you learned to:

- Understand model benchmarks.
- Perform manual evaluations.
- Perform automated evaluations.

Learn more

- Observability in generative AI
- Microsoft Foundry Discord
- Microsoft Foundry Developer Forum