Towards AI · Following

Model Distillation

Making AI Models Leaner and Meaner: A go-to approach for small and medium businesses | Practical guide to shrinking AI Models without losing their Intelligence

Naveen Krishnan
Published in Towards AI
7 min read · Jan 9, 2025

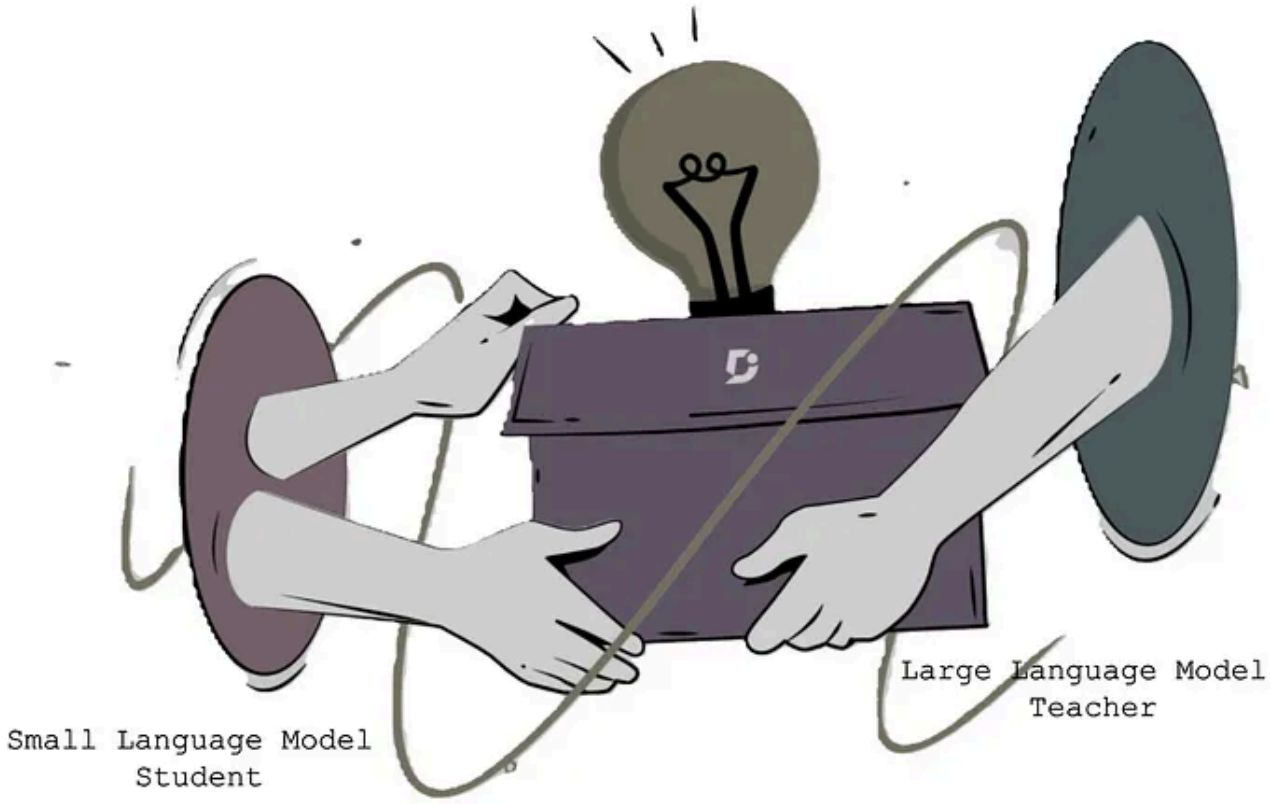

Image Source: Author

1. Introduction

Artificial Intelligence has come this long and especially with the large language models like GPT-4 and others are driving innovation across different industries. These models are powerful, but they are computationally expensive and resource intensive with trillions of parameters along with its multi-modal capabilities. But what if we could get most of their performance in a smaller, faster, and more efficient method 🤔 and that's model distillation.

In this tutorial, we discuss model distillation, its advantages, working and some ways we can implement it. We will also explore several use cases and examples with images and other aids to help you to comprehend the process.

2. What is Model Distillation?

Model distillation is a technique that involves transferring knowledge from a large, pre-trained model (we call it teacher) to a smaller model (we call it Student). The aim is to create a compact model that works almost as well as the large one but uses much less computing power.

Think of it as turning a big encyclopedia into small pocket guides without losing important information.

3. Why Choose Model Distillation?

This technique is going to be very important as more and more industries start adopting AI. Here are the top benefits

- **Cost Efficiency:** Smaller models are always cheap, easy to deploy and maintain.
- **Faster Inference:** The output can be fast which is ideal for real-time applications.
- **Resource Optimization:** Smaller models are the ones which runs on devices with limited computational power, such as smartphones or IoT devices. As we extended the AI capabilities to edge, this plays a vital role.
- **Scalability:** Easier to scale across multiple devices and use cases.

4. How Does Model Distillation Work?

The distillation process involves 2 key steps:

1. The first stage is the synthetic data generation step. In this step, using a training dataset, the teacher model is asked to generate responses for the training data. If there is a validation dataset, the teacher model also generates responses for that dataset as well.
2. The second stage is finetuning. Once the synthetic data is collected, the student model is then finetuned off of the training and validation data created from the teacher model. This transfers the knowledge from the teacher model to the student model.

Here's a visual representation of the distillation workflow:

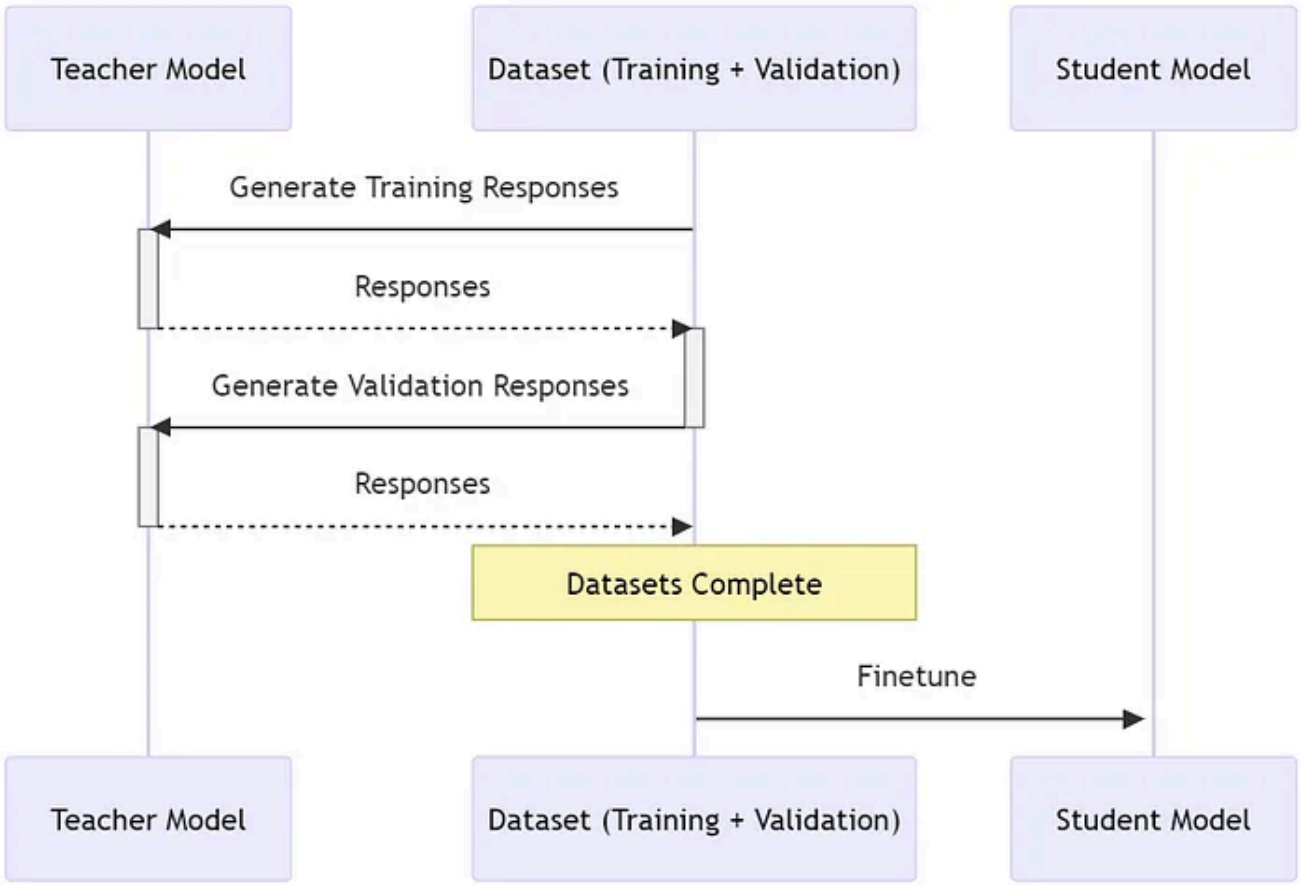

Image Source: github.com/Azure/azureml-examples

5. Distillation Techniques

Standard Knowledge Distillation: This method mainly focuses only on transferring soft predictions from the teacher to student.

Data-Free Knowledge Distillation: This technique is mainly used in cases where the original training data is unavailable and synthetic data is generated using the teacher model.

Feature-Based Distillation: In this technique we will mainly transfer the intermediate features from teacher's layers to the student.

Task-Specific Distillation: This one is primarily optimized for specific tasks like natural language processing, computer vision, or speech recognition etc.

6. Use Cases of Model Distillation

Mobile Applications: As AI matures there will be more need to deploy models on mobile devices for tasks like image recognition, language translation etc. This technique will be help achieve those futuristic use cases on LLM like standards.

Real-Time Systems: This improves the response time in systems like chatbots and recommendation engines.

Edge Computing: Edge devices like AI cameras which can get additional features in future with limited computational power.

Cost Optimization: For Small and Medium business where we need to reducing cloud inference costs for large-scale applications.

Multi-Language Support: Training the SLMs to translate into multiple languages without increasing its size.

7. Implementing Model Distillation

Here's a step-by-step guide, this is more user-friendly quickest approach with Azure AI Foundry.

Azure AI Foundry

Azure AI Foundry

Azure AI Foundryai.azure.com

If you want the notebook approach refer AzureML Model Distillation.

On Azure, these two offerings have got great UI experiences that support an end-to-end distillation flow: **Stored Completions** and **Azure OpenAI Evaluation.**

- **Stored Completions:** This feature helps traffic logging by request and also gives a user-friendly interface for reviewing, filtering and exporting the data that's collected.
- **Azure OpenAI Evaluation:** This gives UI to score data based on predefined criteria.

With these two experiences we can create a comprehensive distillation process:

- Collect live traffic from Azure OpenAI endpoints
- Filter and sublet that traffic in the Stored Completions UI
- Export it to Evaluation UI for quality scoring
- Fine tune from the collected data or a subset based on evaluation scoring.

This simple flow enhances your overall experience by making sure that you can efficiently manage and optimize the data.

7.1 Configure Store Completions

You enable stored completions on Azure OpenAI deployment by adding `store` parameter to `True`. Also, you can use `metadata` parameter to add additonal information to you your stored completion dataset.

```
import os
from openai import AzureOpenAI

client = AzureOpenAI(
    api_key=os.getenv("AZURE_OPENAI_API_KEY"),
    api_version="2024-10-01-preview",
    azure_endpoint = os.getenv("AZURE_OPENAI_ENDPOINT")
```

```
    )

ompletion = client.chat.completions.create(

    model="gpt-4o", # replace with model deployment name
    store= True,
    metadata =  {
    "user": "admin",
    "category": "docs-test",
  },
    messages=[
    {"role": "system", "content": "Provide a clear and concise summary of the t
    {"role": "user", "content": "Ensemble methods combine multiple machine lear
    ]
)

print(completion.choices[0].message)
```

After the stored completions are enabled for an Azure OpenAI deployment, they'll start showing in the Azure AI Foundry portal in the Stored Completions pane like this.

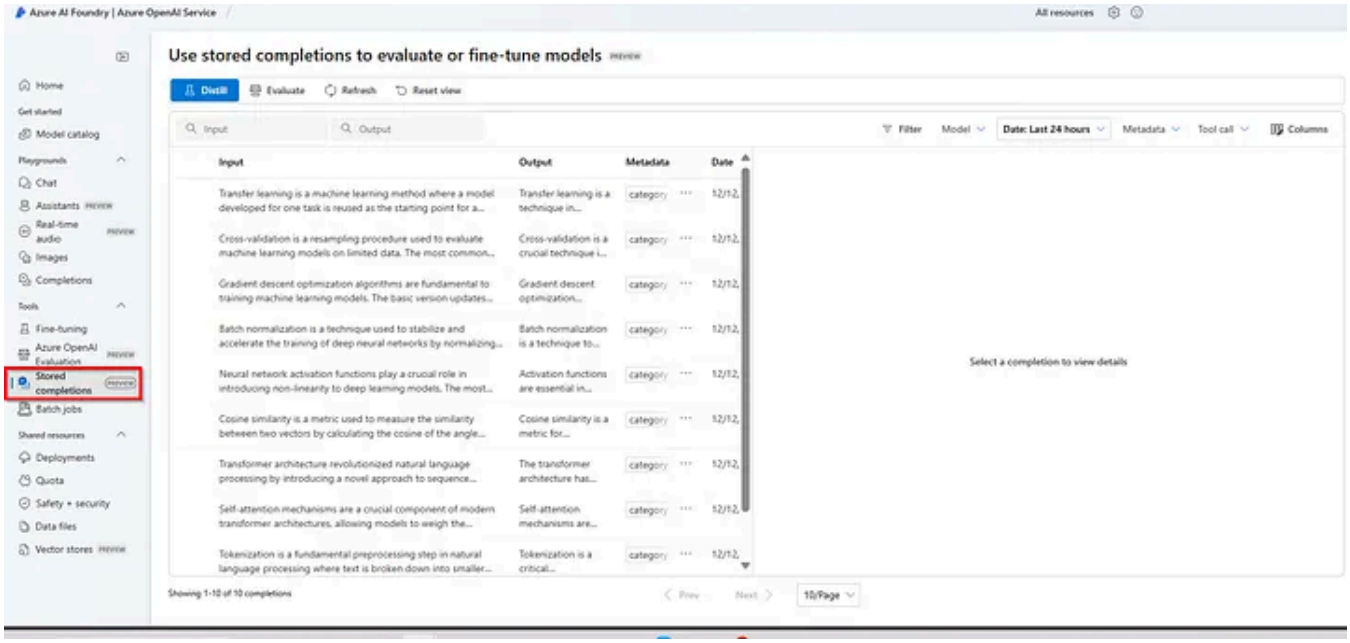

Image source: author

7.2 Distillation

Distillation allows you to turn your stored completions into a fine-tuning dataset. A common use case is to use stored completions with a larger more powerful model for a particular task and then use the stored completions to train a smaller model on high quality examples of model interactions.

Distillation requires a minimum of 10 stored completions, though it's recommended to provide hundreds to thousands of stored completions for the best results.

- From the Stored Completions pane in the Azure AI Foundry portal use the Filter options to select the completions you want to train your model with.
- To begin distillation, select Distill

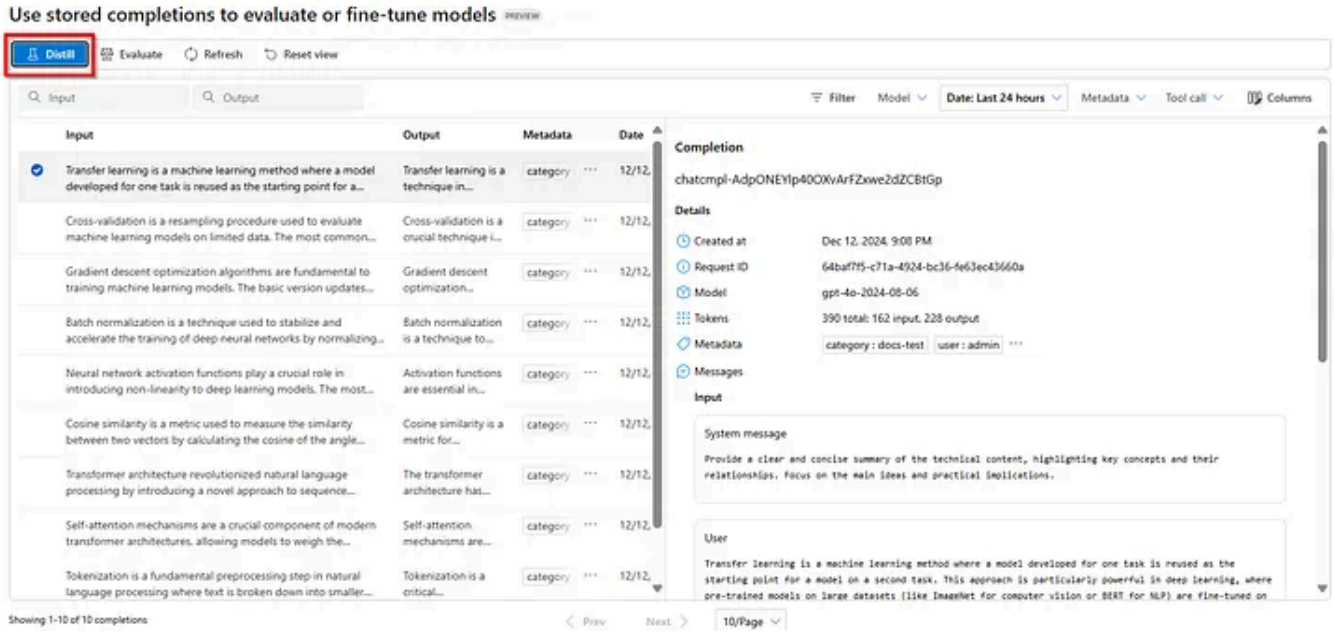

Image source: author

- Pick which model you would like to fine-tune with your stored completion dataset.

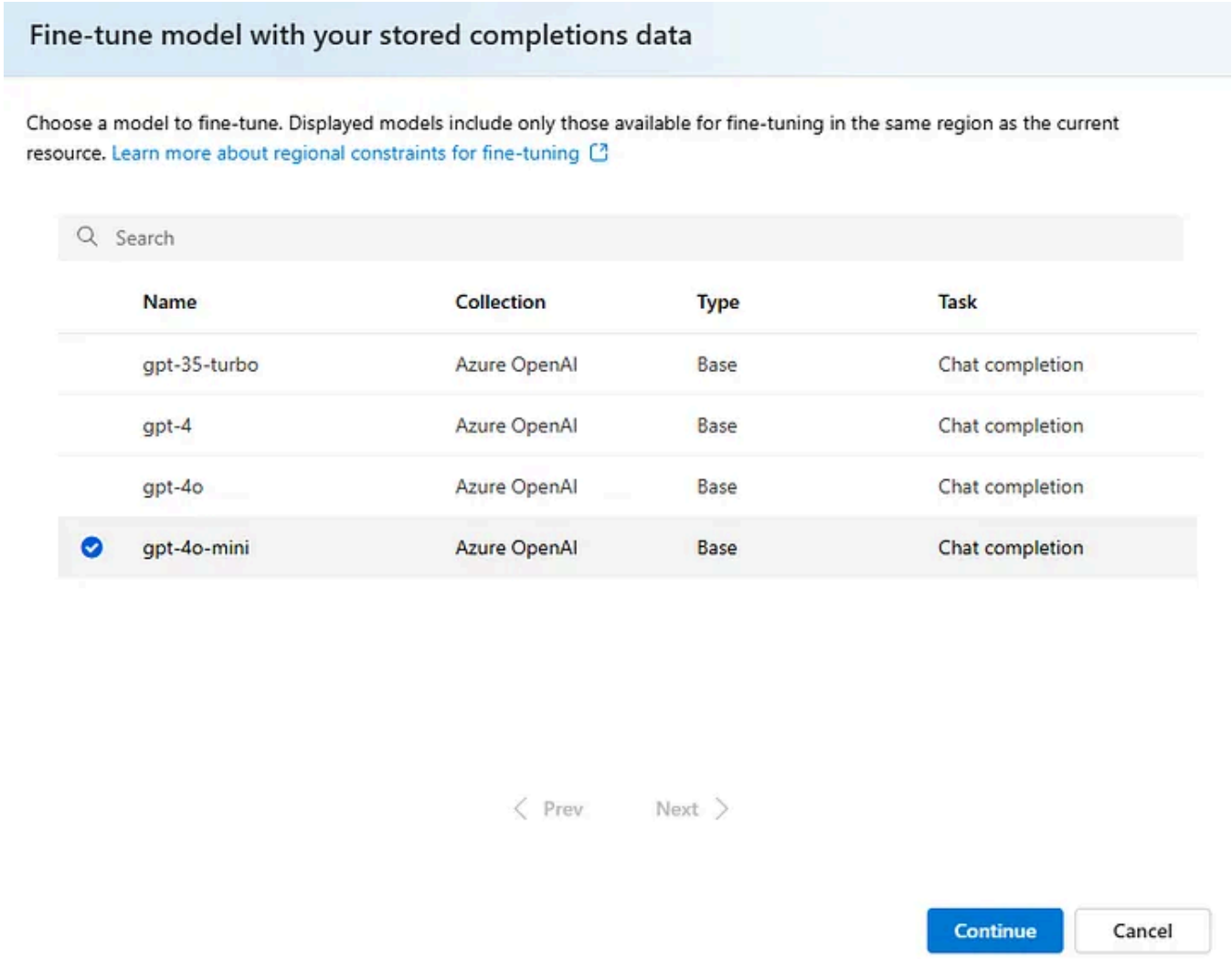

Image source: author

- Confirm which version of the model you want to fine-tune:

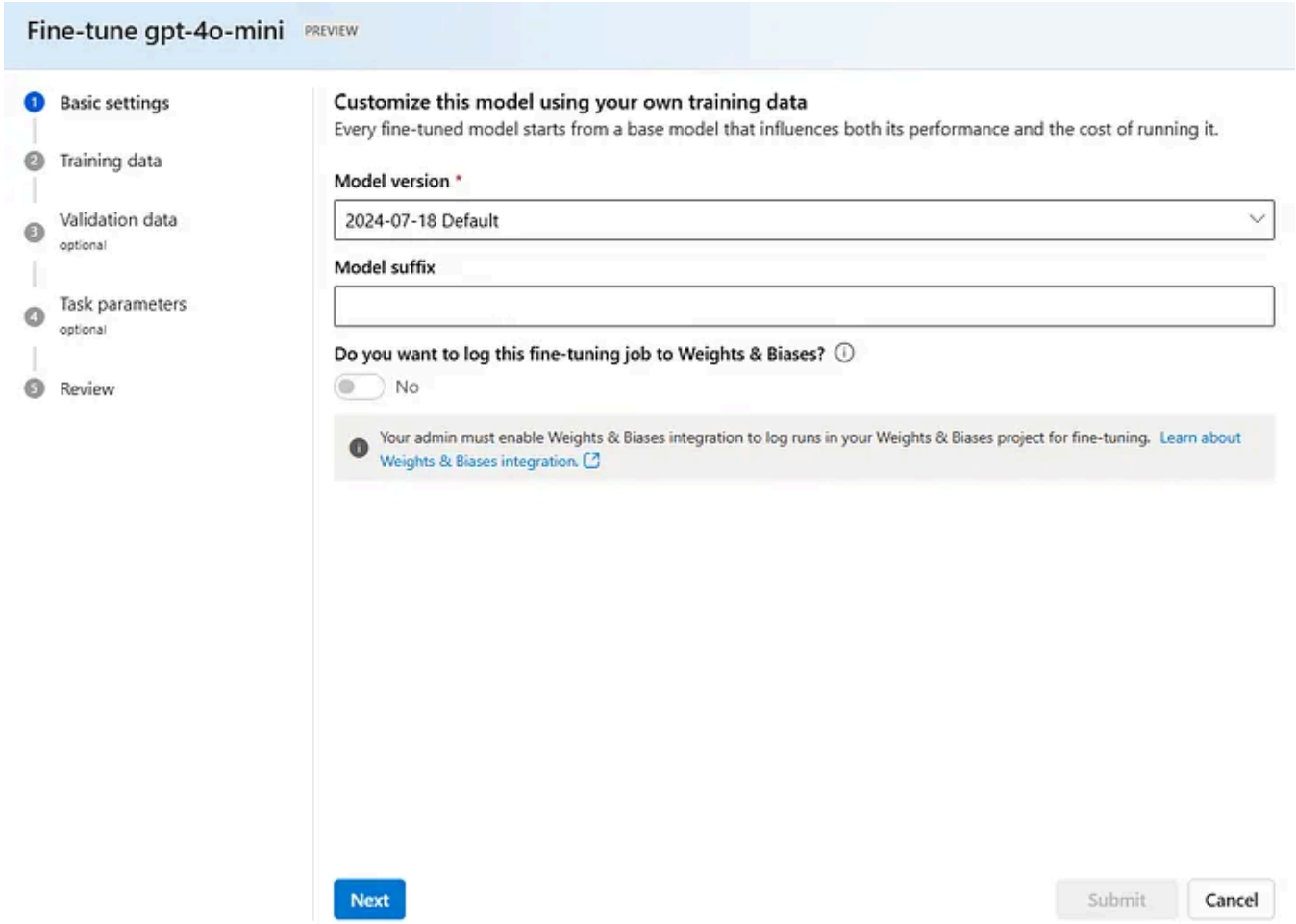

Image source: author

- A `.jsonl` file with a randomly generated name will be created as a training dataset from your stored completions. Select the file > Next.

Note: *Stored completion distillation training files cannot be accessed directly and cannot be exported externally/downloaded.*

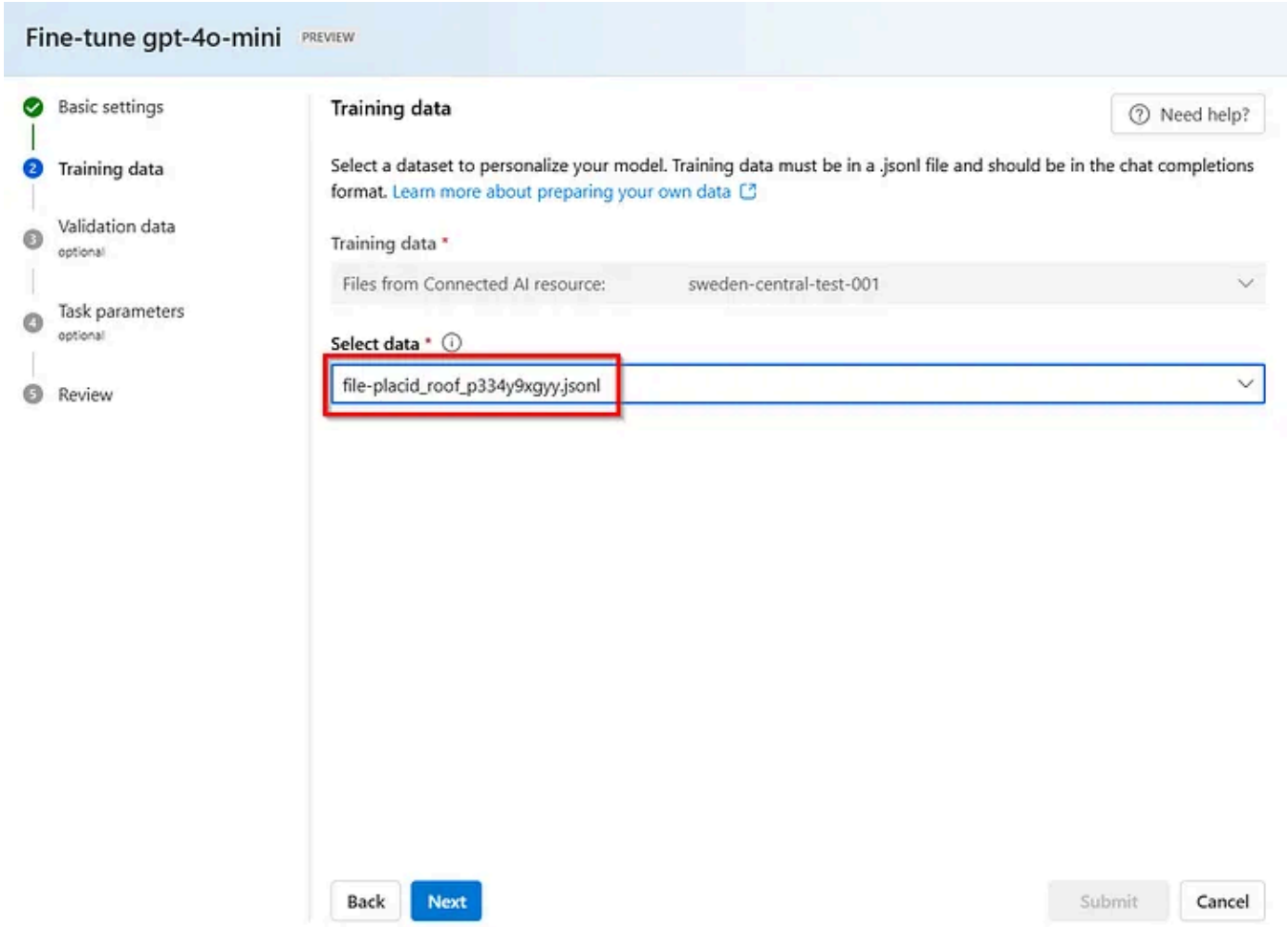

Image source: author

The rest of the steps correspond to the typical Azure OpenAI fine-tuning steps. To learn more, see our fine-tuning getting started guide.

7.3 Evaluation

The evaluation of large language models is a critical step in measuring their performance across various tasks and dimensions. This is especially important for fine-tuned models, where assessing the performance gains (or losses) from training is crucial. Thorough evaluations can help your understanding of how different versions of the model may impact your application or scenario.

Stored completions can be used as a dataset for running evaluations.

- From the Stored Completions pane in the Azure AI Foundry portal use the Filter options to select the completions you want to be part of your evaluation dataset.
- To configure the evaluation, select Evaluate

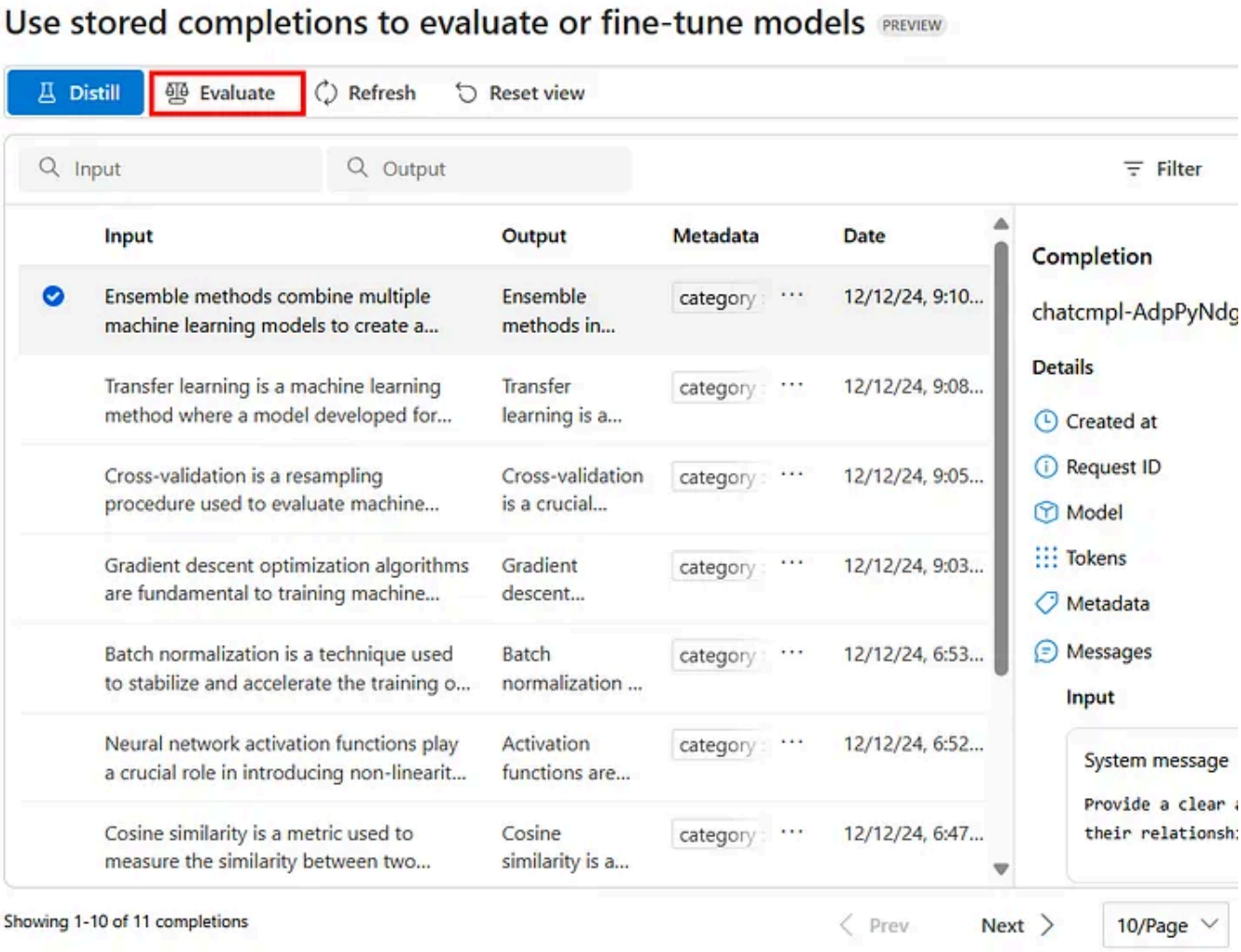

Image source: author

- This launches the Evaluations pane with a prepopulated `.jsonl` file with a randomly generated name that is created as an evaluation dataset from your stored completions.

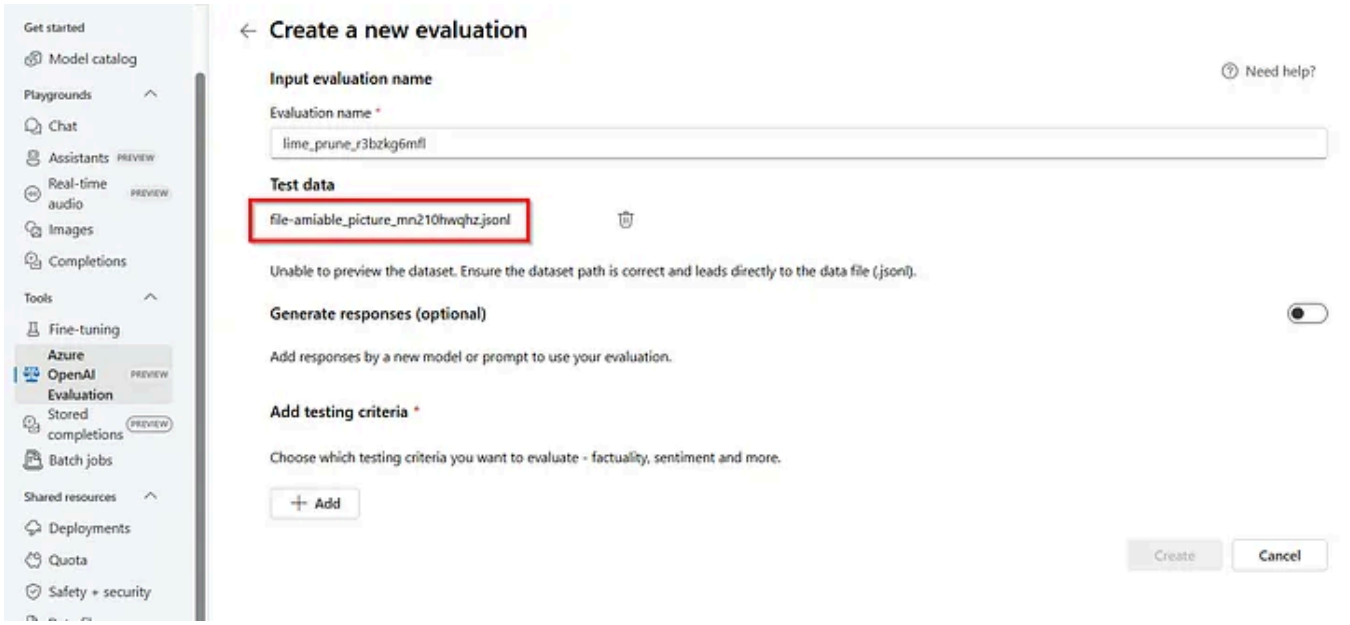

Image source: author

To learn more about evaluation see, getting started with evaluations

8. Challenges in Model Distillation

There are still few challenges:

- **Loss of Performance:** Maintaining balance between size reduction and performance is always not easy.
- **Data Dependency:** We need good quality data as it is very important for right knowledge transfer.
- **Computational Cost:** There is one-time cost involved in training the teacher model at the beginning.

9. Future of Model Distillation

As I already mentioned, AI models can grow up in size to several trillions in upcoming days as we need more multi-modal capabilities and all in one at our hands 😎, and this where distillation will play a crucial role in making them accessible and efficient. We can expect several methods of distillation in near future.

10.Conclusion

Model distillation is a game-changer for AI development and deployment. By creating smaller, efficient models, we can make AI accessible and bring powerful capabilities to a broader audience. 😇 Whether you're working on edge devices, real-time systems or cost-sensitive applications, distillation is a practical solution.

AI | Large Language Models | ChatGPT | Technology | Data Science

Following

Published in Towards AI

77K Followers · Last published 1 day ago

The leading AI community and content platform focused on making AI accessible to all. Check out our new course platform: https://academy.towardsai.net/courses/beginner-to-advanced-llm-dev

Edit profile

Written by Naveen Krishnan

158 Followers · 140 Following

AI Architect @ Microsoft. Passionate about leveraging artificial intelligence to solve real-world problems.

No responses yet

Naveen Krishnan

What are your thoughts?

More from Naveen Krishnan and Towards AI