**# GitHub Models Documentation**

**## Table of Contents**

1. [Curl Command](#curl-command)

2. [Available Models](#available-models)

3. [Model Details](#model-details)

**## Curl Command**

The following curl command was used to fetch the list of available models:

```bash

curl -s -X GET "https://models.inference.ai.azure.com/models" -H "Authorization: Bearer ghp\_NTJ3YSkMzFpBThkvVYRr8mBpGJsVJc2SKkcd"

```

**## Available Models**

Here's a list of the available models:

1. AI21-Jamba-Instruct

2. Cohere-command-r

3. Cohere-command-r-plus

4. Cohere-embed-v3-english

5. Cohere-embed-v3-multilingual

6. Meta-Llama-3-70B-Instruct

7. Meta-Llama-3-8B-Instruct

8. Meta-Llama-3.1-405B-Instruct

9. Meta-Llama-3.1-70B-Instruct

10. Meta-Llama-3.1-8B-Instruct

11. Mistral-large

12. Mistral-large-2407

13. Mistral-Nemo

14. Mistral-small

15. gpt-4o

16. gpt-4o-mini

17. text-embedding-3-large

18. text-embedding-3-small

19. Phi-3-medium-128k-instruct

20. Phi-3-medium-4k-instruct

21. Phi-3-mini-128k-instruct

22. Phi-3-mini-4k-instruct

23. Phi-3-small-128k-instruct

24. Phi-3-small-8k-instruct

25. Phi-3.5-mini-instruct

**## Model Details**

Here are the details for each model:

**### AI21-Jamba-Instruct**

- **\*\*ID\*\***: azureml://registries/azureml-ai21/models/AI21-Jamba-Instruct/versions/2

- **\*\*Publisher\*\***: AI21 Labs

- **\*\*Model Family\*\***: AI21 Labs

- **\*\*Model Registry\*\***: azureml-ai21

- **\*\*License\*\***: custom

- **\*\*Task\*\***: chat-completion

- **\*\*Summary\*\***: Jamba-Instruct is the world's first production-grade Mamba-based LLM model and leverages its hybrid Mamba-Transformer architecture to achieve best-in-class performance quality and cost efficiency.

- **\*\*Tags\*\***: chat, rag

**### Cohere-command-r**

- **\*\*ID\*\***: azureml://registries/azureml-cohere/models/Cohere-command-r/versions/3

- **\*\*Publisher\*\***: cohere

- **\*\*Model Family\*\***: cohere

- **\*\*Model Registry\*\***: azureml-cohere

- **\*\*License\*\***: custom

- **\*\*Task\*\***: chat-completion

- **\*\*Summary\*\***: Command R is a scalable generative model targeting RAG and Tool Use to enable production-scale AI for enterprise.

- **\*\*Tags\*\***: rag, multilingual

**### Cohere-command-r-plus**

- **\*\*ID\*\***: azureml://registries/azureml-cohere/models/Cohere-command-r-plus/versions/3

- **\*\*Publisher\*\***: cohere

- **\*\*Model Family\*\***: cohere

- **\*\*Model Registry\*\***: azureml-cohere

- **\*\*License\*\***: custom

- **\*\*Task\*\***: chat-completion

- **\*\*Summary\*\***: Command R+ is a state-of-the-art RAG-optimized model designed to tackle enterprise-grade workloads.

- **\*\*Tags\*\***: rag, multilingual

**### Cohere-embed-v3-english**

- **\*\*ID\*\***: azureml://registries/azureml-cohere/models/Cohere-embed-v3-english/versions/3

- **\*\*Publisher\*\***: cohere

- **\*\*Model Family\*\***: cohere

- **\*\*Model Registry\*\***: azureml-cohere

- **\*\*License\*\***: custom

- **\*\*Task\*\***: embeddings

- **\*\*Summary\*\***: Cohere Embed English is the market's leading text representation model used for semantic search, retrieval-augmented generation (RAG), classification, and clustering.

- **\*\*Tags\*\***: RAG, search

**### Cohere-embed-v3-multilingual**

- **\*\*ID\*\***: azureml://registries/azureml-cohere/models/Cohere-embed-v3-multilingual/versions/3

- **\*\*Publisher\*\***: cohere

- **\*\*Model Family\*\***: cohere

- **\*\*Model Registry\*\***: azureml-cohere

- **\*\*License\*\***: custom

- **\*\*Task\*\***: embeddings

- **\*\*Summary\*\***: Supporting over 100 languages, Cohere Embed Multilingual is the market's leading text representation model used for semantic search, retrieval-augmented generation (RAG), classification, and clustering.

- **\*\*Tags\*\***: RAG, search

**### Meta-Llama-3-70B-Instruct**

- **\*\*ID\*\***: azureml://registries/azureml-meta/models/Meta-Llama-3-70B-Instruct/versions/6

- **\*\*Publisher\*\***: meta

- **\*\*Model Family\*\***: meta

- **\*\*Model Registry\*\***: azureml-meta

- **\*\*License\*\***: custom

- **\*\*Task\*\***: chat-completion

- **\*\*Summary\*\***: A powerful 70-billion parameter model excelling in reasoning, coding, and broad language applications.

- **\*\*Tags\*\***: conversation

**### Meta-Llama-3-8B-Instruct**

- **\*\*ID\*\***: azureml://registries/azureml-meta/models/Meta-Llama-3-8B-Instruct/versions/6

- **\*\*Publisher\*\***: meta

- **\*\*Model Family\*\***: meta

- **\*\*Model Registry\*\***: azureml-meta

- **\*\*License\*\***: custom

- **\*\*Task\*\***: chat-completion

- **\*\*Summary\*\***: A versatile 8-billion parameter model optimized for dialogue and text generation tasks.

- **\*\*Tags\*\***: conversation

**### Meta-Llama-3.1-405B-Instruct**

- **\*\*ID\*\***: azureml://registries/azureml-meta/models/Meta-Llama-3.1-405B-Instruct/versions/1

- **\*\*Publisher\*\***: meta

- **\*\*Model Family\*\***: meta

- **\*\*Model Registry\*\***: azureml-meta

- **\*\*License\*\***: custom

- **\*\*Task\*\***: chat-completion

- **\*\*Summary\*\***: The Llama 3.1 instruction tuned text only models are optimized for multilingual dialogue use cases and outperform many of the available open source and closed chat models on common industry benchmarks.

- **\*\*Tags\*\***: conversation

**### Meta-Llama-3.1-70B-Instruct**

- **\*\*ID\*\***: azureml://registries/azureml-meta/models/Meta-Llama-3.1-70B-Instruct/versions/1

- **\*\*Publisher\*\***: meta

- **\*\*Model Family\*\***: meta

- **\*\*Model Registry\*\***: azureml-meta

- **\*\*License\*\***: custom

- **\*\*Task\*\***: chat-completion

- **\*\*Summary\*\***: The Llama 3.1 instruction tuned text only models are optimized for multilingual dialogue use cases and outperform many of the available open source and closed chat models on common industry benchmarks.

- **\*\*Tags\*\***: conversation

**### Meta-Llama-3.1-8B-Instruct**

- **\*\*ID\*\***: azureml://registries/azureml-meta/models/Meta-Llama-3.1-8B-Instruct/versions/1

- **\*\*Publisher\*\***: meta

- **\*\*Model Family\*\***: meta

- **\*\*Model Registry\*\***: azureml-meta

- **\*\*License\*\***: custom

- **\*\*Task\*\***: chat-completion

- **\*\*Summary\*\***: The Llama 3.1 instruction tuned text only models are optimized for multilingual dialogue use cases and outperform many of the available open source and closed chat models on common industry benchmarks.

- **\*\*Tags\*\***: conversation

**### Mistral-large**

- **\*\*ID\*\***: azureml://registries/azureml-mistral/models/Mistral-large/versions/1

- **\*\*Publisher\*\***: mistralai

- **\*\*Model Family\*\***: mistralai

- **\*\*Model Registry\*\***: azureml-mistral

- **\*\*License\*\***: custom

- **\*\*Task\*\***: chat-completion

- **\*\*Summary\*\***: Mistral's flagship model that's ideal for complex tasks that require large reasoning capabilities or are highly specialized (Synthetic Text Generation, Code Generation, RAG, or Agents).

- **\*\*Tags\*\***: reasoning, rag, agents, multilingual

**### Mistral-large-2407**

- **\*\*ID\*\***: azureml://registries/azureml-mistral/models/Mistral-large-2407/versions/1

- **\*\*Publisher\*\***: mistralai

- **\*\*Model Family\*\***: mistralai

- **\*\*Model Registry\*\***: azureml-mistral

- **\*\*License\*\***: custom

- **\*\*Task\*\***: chat-completion

- **\*\*Summary\*\***: Mistral Large (2407) is an advanced Large Language Model (LLM) with state-of-the-art reasoning, knowledge, and coding capabilities.

- **\*\*Tags\*\***: reasoning, rag, agents

**### Mistral-Nemo**

- **\*\*ID\*\***: azureml://registries/azureml-mistral/models/Mistral-Nemo/versions/1

- **\*\*Publisher\*\***: mistralai

- **\*\*Model Family\*\***: mistralai

- **\*\*Model Registry\*\***: azureml-mistral

- **\*\*License\*\***: custom

- **\*\*Task\*\***: chat-completion

- **\*\*Summary\*\***: Mistral Nemo is a cutting-edge Language Model (LLM) boasting state-of-the-art reasoning, world knowledge, and coding capabilities within its size category.

- **\*\*Tags\*\***: reasoning, rag, agents

**### Mistral-small**

- **\*\*ID\*\***: azureml://registries/azureml-mistral/models/Mistral-small/versions/1

- **\*\*Publisher\*\***: mistralai

- **\*\*Model Family\*\***: mistralai

- **\*\*Model Registry\*\***: azureml-mistral

- **\*\*License\*\***: custom

- **\*\*Task\*\***: chat-completion

- **\*\*Summary\*\***: Mistral Small can be used on any language-based task that requires high efficiency and low latency.

- **\*\*Tags\*\***: low latency, multilingual

**### gpt-4o**

- **\*\*ID\*\***: azureml://registries/azure-openai/models/gpt-4o/versions/2

- **\*\*Publisher\*\***: Azure OpenAI Service

- **\*\*Model Family\*\***: openai

- **\*\*Model Registry\*\***: azure-openai

- **\*\*License\*\***: custom

- **\*\*Task\*\***: chat-completion

- **\*\*Summary\*\***: OpenAI's most advanced multimodal model in the GPT-4 family. Can handle both text and image inputs.

- **\*\*Tags\*\***: multipurpose, multilingual, multimodal

**### gpt-4o-mini**

- **\*\*ID\*\***: azureml://registries/azure-openai/models/gpt-4o-mini/versions/1

- **\*\*Publisher\*\***: Azure OpenAI Service

- **\*\*Model Family\*\***: OpenAI

- **\*\*Model Registry\*\***: azure-openai

- **\*\*License\*\***: custom

- **\*\*Task\*\***: chat-completion

- **\*\*Summary\*\***: An affordable, efficient AI solution for diverse text and image tasks.

- **\*\*Tags\*\***: multipurpose, multilingual, multimodal

**### text-embedding-3-large**

- **\*\*ID\*\***: azureml://registries/azure-openai/models/text-embedding-3-large/versions/1

- **\*\*Publisher\*\***: Azure OpenAI Service

- **\*\*Model Family\*\***: openai

- **\*\*Model Registry\*\***: azure-openai

- **\*\*License\*\***: custom

- **\*\*Task\*\***: embeddings

- **\*\*Summary\*\***: Text-embedding-3 series models are the latest and most capable embedding model from OpenAI.

- **\*\*Tags\*\***: RAG, search

**### text-embedding-3-small**

- **\*\*ID\*\***: azureml://registries/azure-openai/models/text-embedding-3-small/versions/1

- **\*\*Publisher\*\***: Azure OpenAI Service

- **\*\*Model Family\*\***: openai

- **\*\*Model Registry\*\***: azure-openai

- **\*\*License\*\***: custom

- **\*\*Task\*\***: embeddings

- **\*\*Summary\*\***: Text-embedding-3 series models are the latest and most capable embedding model from OpenAI.

- **\*\*Tags\*\***: RAG, search

**### Phi-3-medium-128k-instruct**

- **\*\*ID\*\***: azureml://registries/azureml/models/Phi-3-medium-128k-instruct/versions/3

- **\*\*Publisher\*\***: microsoft

- **\*\*Model Family\*\***: microsoft

- **\*\*Model Registry\*\***: azureml

- **\*\*License\*\***: mit

- **\*\*Task\*\***: chat-completion

- **\*\*Summary\*\***: Same model as Phi-3-medium (4k) but with larger context size. Use this for RAG or few shot prompting.

- **\*\*Tags\*\***: reasoning, understanding, large context

**### Phi-3-medium-4k-instruct**

- **\*\*ID\*\***: azureml://registries/azureml/models/Phi-3-medium-4k-instruct/versions/3

- **\*\*Publisher\*\***: microsoft

- **\*\*Model Family\*\***: microsoft

- **\*\*Model Registry\*\***: azureml

- **\*\*License\*\***: mit

- **\*\*Task\*\***: chat-completion

- **\*\*Summary\*\***: A 14B parameter model. Use this larger model for better quality than Phi-3-mini and with reasoning-dense data.

- **\*\*Tags\*\***: reasoning, understanding

**### Phi-3-mini-128k-instruct**

- **\*\*ID\*\***: azureml://registries/azureml/models/Phi-3-mini-128k-instruct/versions/10

- **\*\*Publisher\*\***: microsoft

- **\*\*Model Family\*\***: microsoft

- **\*\*Model Registry\*\***: azureml

- **\*\*License\*\***: mit

- **\*\*Task\*\***: chat-completion

- **\*\*Summary\*\***: Same model as Phi-3-mini (4k) but with larger context size. Use this for RAG or few shot prompting.

- **\*\*Tags\*\***: reasoning, understanding, low latency

**### Phi-3-mini-4k-instruct**

- **\*\*ID\*\***: azureml://registries/azureml/models/Phi-3-mini-4k-instruct/versions/10

- **\*\*Publisher\*\***: microsoft

- **\*\*Model Family\*\***: microsoft

- **\*\*Model Registry\*\***: azureml

- **\*\*License\*\***: mit

- **\*\*Task\*\***: chat-completion

- **\*\*Summary\*\***: Tiniest Phi-3 model. Optimized for both quality and low latency.

- **\*\*Tags\*\***: reasoning, understanding, low latency

**### Phi-3-small-128k-instruct**

- **\*\*ID\*\***: azureml://registries/azureml/models/Phi-3-small-128k-instruct/versions/3

- **\*\*Publisher\*\***: microsoft

- **\*\*Model Family\*\***: microsoft

- **\*\*Model Registry\*\***: azureml

- **\*\*License\*\***: mit

- **\*\*Task\*\***: chat-completion

- **\*\*Summary\*\***: Same Phi-3-small model but with a larger context size for RAG or few shot prompting.

- **\*\*Tags\*\***: reasoning, understanding, large context

**### Phi-3-small-8k-instruct**

- **\*\*ID\*\***: azureml://registries/azureml/models/Phi-3-small-8k-instruct/versions/3

- **\*\*Publisher\*\***: microsoft

- **\*\*Model Family\*\***: microsoft

- **\*\*Model Registry\*\***: azureml

- **\*\*License\*\***: mit

- **\*\*Task\*\***: chat-completion

- **\*\*Summary\*\***: A 7B parameters model. Use this larger model for better quality than Phi-3-mini and with reasoning-dense data.

- **\*\*Tags\*\***: reasoning, understanding

**### Phi-3.5-mini-instruct**

- **\*\*ID\*\***: azureml://registries/azureml/models/Phi-3.5-mini-instruct/versions/2

- **\*\*Publisher\*\***: microsoft

- **\*\*Model Family\*\***: microsoft

- **\*\*Model Registry\*\***: azureml

- **\*\*License\*\***: mit

- **\*\*Task\*\***: chat-completion

- **\*\*Summary\*\***: Refresh of Phi-3-mini model

- **\*\*Tags\*\***: reasoning, understanding, low latency