

Deploying Large Language Model (LLM) on Azure

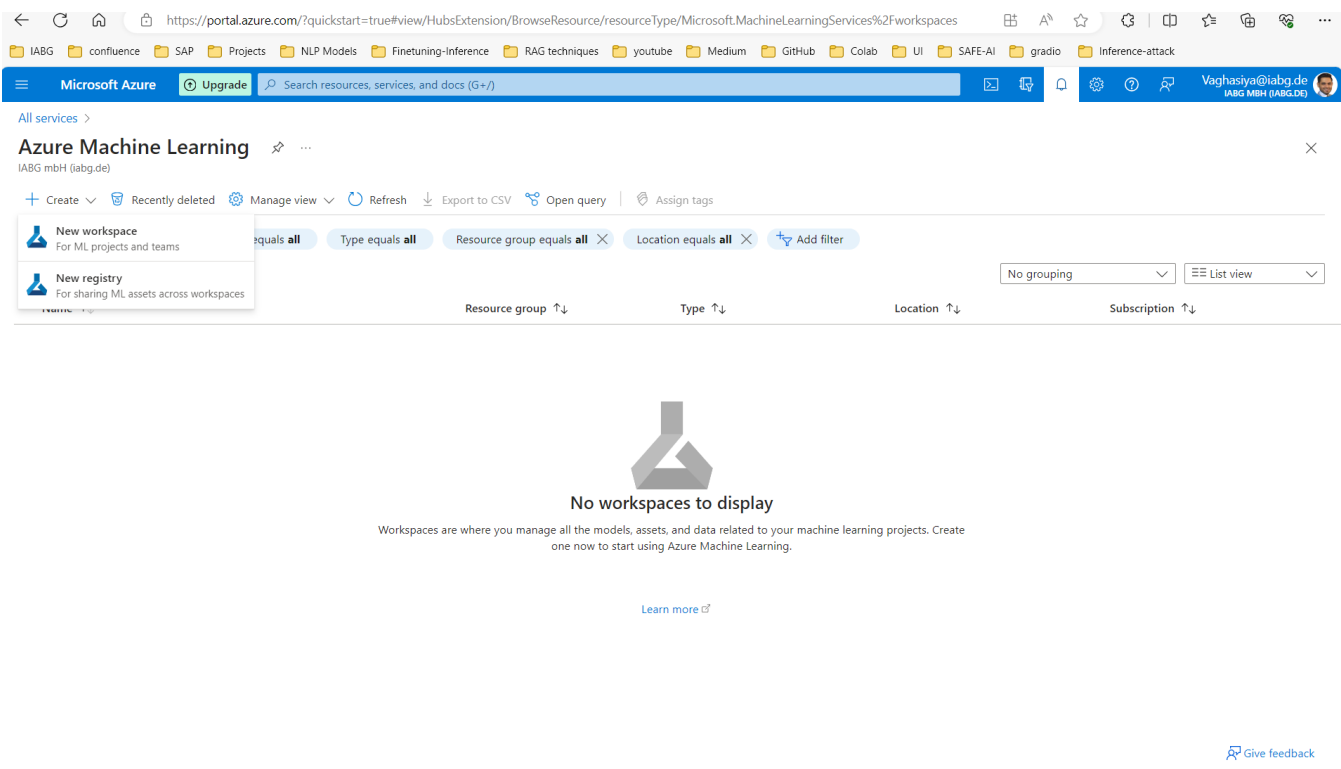
Azure Machine Learning provides a powerful platform for running and deploying large language models (LLMs). This guide will walk you through the steps to deploy your LLM and create an endpoint for your LLM in Azure Machine Learning.

Prerequisites:

- An active Azure account - [Setting Up a free Azure Account - NLP Community - Confluence \(iabg.de\)](#)

Steps to Deploy LLM on Azure:

- **Sign in to Azure Portal:**
 - Open your web browser and navigate to the Azure Portal: <https://portal.azure.com>.
 - Sign in with your Azure account credentials.
- **Create a New Azure Machine Learning Workspace:**
 - In the Azure Portal, search for "**Azure Machine Learning**".
 - Select "**New workspace**".



- Fill out the required information to create a new Azure Machine Learning workspace, such as subscription, resource group, workspace name, and region as mentioned in the following images.
- Keep the default values for **Networking, Encryption, Identity and Tags**.

← ↻ 🔍 https://portal.azure.com/?quickstart=true#view/Microsoft_Azure_MLTeamAccounts/CreateMachineLearningServicesBladeV2/_provisioningContext~/~/%7... IABG confluence SAP Projects NLP Models Finetuning-Inference RAG techniques youtube Medium GitHub Colab UI SAFE-AI gradio Inference-attack

Microsoft Azure Upgrade Search resources, services, and docs (G+)

Home >

Azure Machine Learning

Create a machine learning workspace

Resource details

Every workspace must be assigned to an Azure subscription, which is where billing happens. You use resource groups like folders to organize and manage resources, including the workspace you're about to create.
[Learn more about Azure resource groups](#)

Subscription * ⓘ free-trial

Resource group * ⓘ (New) TVL
[Create new](#)

Workspace details

Configure your basic workspace settings like its storage connection, authentication, container, and more. [Learn more](#)

Name * ⓘ TVL-AI-Experiments ✓

Region * ⓘ East US 2

Storage account * ⓘ (new) tvlaexperimen4964869347
[Create new](#)

Key vault * ⓘ (new) tvlaexperimen4175077230
[Create new](#)

Application insights * ⓘ (new) tvlaexperimen0893206842
[Create new](#)

Container registry ⓘ None
[Create new](#)

[Review + create](#) < Previous Next : Networking

- Click on **"Review + Create"** and then **"Create"** to provision the workspace.

← ↻ 🔍 https://portal.azure.com/?quickstart=true#view/Microsoft_Azure_MLTeamAccounts/CreateMachineLearningServicesBladeV2/_provisioningContext~/~/%7... IABG confluence SAP Projects NLP Models Finetuning-Inference RAG techniques youtube Medium GitHub Colab UI SAFE-AI gradio Inference-attack

Microsoft Azure Upgrade Search resources, services, and docs (G+)

Home >

Azure Machine Learning

Create a machine learning workspace

✓ Validation passed

Basics Networking Encryption Identity Tags Review + create

Basics

Subscription	free-trial
Resource group	(New) TVL
Region	East US 2
Name	TVL-AI-Experiments
Storage account	(new) tvlaexperimen4964869347
Key vault	(new) tvlaexperimen4175077230
Application insights	(new) tvlaexperimen0893206842
Container registry	None

Networking

Connectivity method	Enable public access from all networks
Network isolation	Public

Encryption

Encryption type	Microsoft-managed keys
-----------------	------------------------

Identity

[Create](#) < Previous Next > [Download a template for automation](#)

Home > Microsoft Azure | Overview

Deployment

Search resources, services, and docs (G+)

Overview

Inputs

Outputs

Template

Deployment is in progress

Deployment name : Microsoft.MachineLearningServices
Subscription : free-trial
Resource group : TVL

Start time : 2/27/2024, 5:32:21 PM
Correlation ID : b5351ef7-5b57-4c47-bec1-2eae08916e1c

Deployment details

Resource	Type	Status	Operation details
tvlaexperimen4175077230	Key vault	OK	Operation details
tvlaexperimen4964869347	Storage account	OK	Operation details

Microsoft Defender for Cloud
Secure your apps and infrastructure
[Go to Microsoft Defender for Cloud >](#)

Free Microsoft tutorials
[Start learning today >](#)

Work with an expert
Azure experts are service provider partners who can help manage your assets on Azure and be your first line of support.
[Find an Azure expert >](#)

Home > Microsoft Azure | Overview

Deployment

Search resources, services, and docs (G+)

Overview

Inputs

Outputs

Template

Deployment succeeded

Deployment 'Microsoft.MachineLearningServices' to resource group 'TVL' was successful.

[Go to resource](#) [Go to resource group](#)

Your deployment is complete

Deployment name : Microsoft.MachineLearningServices
Subscription : free-trial
Resource group : TVL

Start time : 2/27/2024, 5:32:21 PM
Correlation ID : b5351ef7-5b57-4c47-bec1-2eae08916e1c

Deployment details

Next steps

[Go to resource](#)

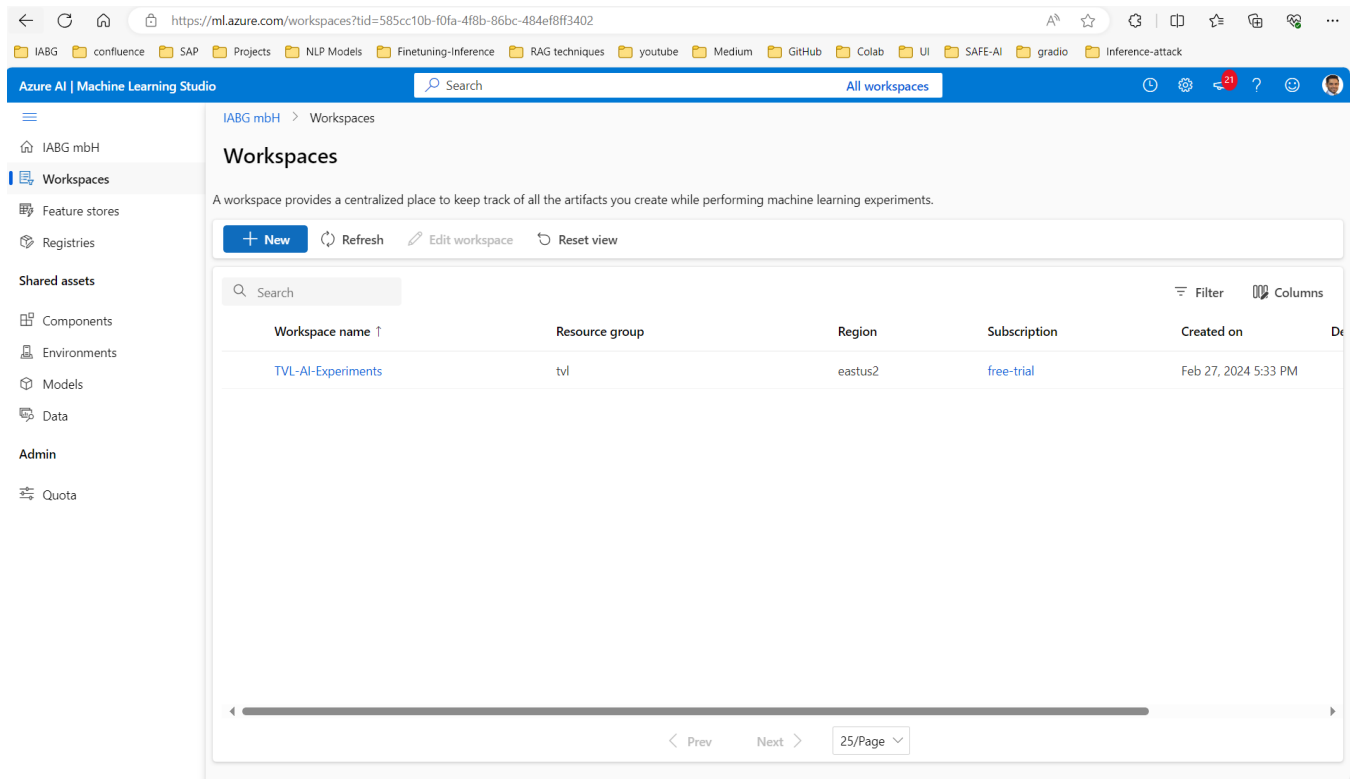
Cost management
Get notified to stay within your budget and prevent unexpected charges on your bill.
[Set up cost alerts >](#)

Microsoft Defender for Cloud
Secure your apps and infrastructure
[Go to Microsoft Defender for Cloud >](#)

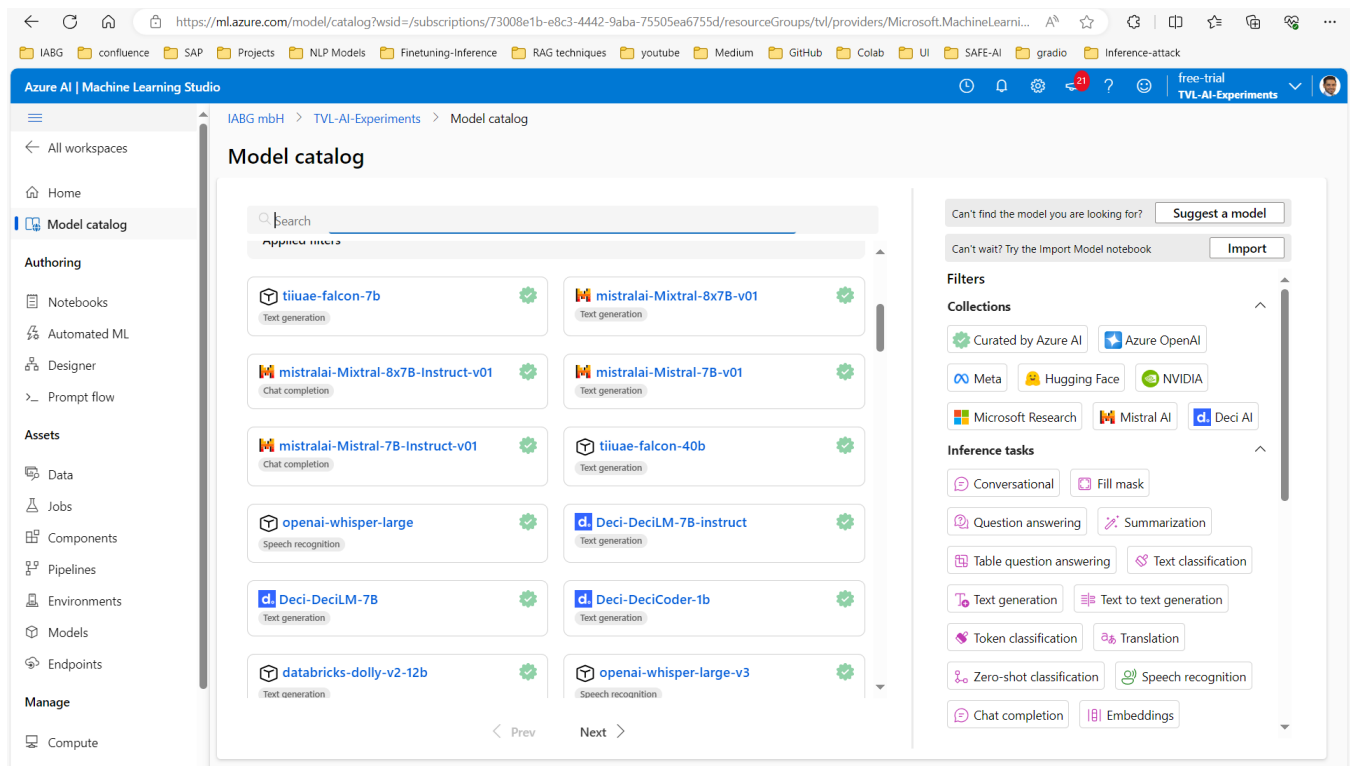
Free Microsoft tutorials
[Start learning today >](#)

Work with an expert
Azure experts are service provider partners who can help manage your assets on Azure and be your first line of support.
[Find an Azure expert >](#)

- **Deploy LLM and create LLM Endpoint:**
 - Navigate to your newly created Azure Machine Learning workspace. <https://ml.azure.com/>.
 - Go to **Workspaces**.
 - Click on the workspace that you recently created.



- Go to **Model Catalogs** and select any Open-Source Model from Huggingface.



- Click on **Deploy Real-Time Endpoint** after selecting the Open-Source LLM(For example: **mistralai-Mistral-7B-Instruct-v01**)

The screenshot shows the Azure AI Machine Learning Studio interface. The left sidebar contains navigation options like 'All workspaces', 'Home', 'Model catalog', 'Authoring', 'Assets', and 'Manage'. The main content area displays the 'mistralai-Mistral-7B-Instruct-v01' model page. The 'Overview' tab is selected, showing the model's task (Chat completion), languages (EN), and license (apache-2.0). The 'Description' section states that the model is a fine-tuned version of Mistral-7B-v0.1. The 'Model Architecture' section lists features like Grouped-Query Attention and Sliding-Window Attention. The 'Limitations' section mentions that the model lacks moderation mechanisms. A 'Real-time endpoint' tooltip is shown over the 'Deploy' button, indicating that the model can be deployed using the real-time endpoint wizard.

- **Virtual Machine and Instance Count:**

- Choose the virtual machine according to your LLM requirements and adjust the instance count as needed.

The screenshot shows the 'Deploy mistralai-Mistral-7B-Instruct-v01:5' dialog box in the Azure AI Machine Learning Studio. The dialog box contains the following fields and options:

- Virtual machine:** Standard_NC24ads_A10... 24 Cores, 220 GB (RAM), 64 GB (Disk), \$3.67/hr
- Instance count:** 1
- Endpoint:** New
- Endpoint name:** tvl-ai-experiments-zgryt
- Deployment name:** mistralai-mistral-7b-instruct-5
- Inferring data collection:** Disabled

 A note at the top of the dialog box states: 'For the selected model, the scoring script and environment are auto generated for you. Learn More'. A warning message indicates: 'You have no dedicated quota. A temporary 168-hour endpoint will be created for you. Alternatively, you can request for quota for persistent endpoints. Learn more about shared quota'. A checkbox is checked: 'I want to use shared quota and I acknowledge that this endpoint will be deleted in 168 hours'.

- **Create LLM Endpoint:**

- Once you click on **Deploy**, you will be redirected to the following page.
- It will take few minutes to deploy the LLM and create endpoint.
- Once it is deployed, **Provisioning State** will show **Succeeded**

The screenshot shows the Azure AI Machine Learning Studio interface. The left sidebar contains navigation options: All workspaces, Home, Model catalog, Authoring (Notebooks, Automated ML, Designer, Prompt flow), Assets (Data, Jobs, Components, Pipelines, Environments, Models), Endpoints (selected), Manage (Compute, Monitoring), and TVL-AI-Experiments. The main content area displays the details for the endpoint 'tv1-ai-experiments-zgryt'. The 'Details' tab is active, showing endpoint attributes and deployment information.

Endpoint attributes:

- Service ID: tv1-ai-experiments-zgryt
- Description: --
- Provisioning state: Succeeded
- Error details: --
- Compute type: Managed
- Created by: Vaghasiya Nehal
- Created on: Feb 27, 2024 5:44 PM
- Last updated on: Feb 27, 2024 5:44 PM
- Authentication type: Key
- Public network access: --

Deployment summary:

- Live traffic allocation: mistralai-mistral-7b-instruct-5 (0%)
- Mirrored traffic allocation: --

Deployment mistralai-mistral-7b-instruct-5:

- Name: mistralai-mistral-7b-instruct-5
- Live traffic: 0%
- Scoring script: Auto-generated
- Provisioning state: Updating
- Error details: --
- SKU: Standard_NC24ads_A100_v4

The screenshot shows the Azure AI Machine Learning Studio interface with the 'Test' tab selected for the endpoint 'tv1-ai-experiments-zgryt'. The left sidebar is the same as the previous screenshot. The main content area displays the 'Test' tab, showing endpoint details and test results.

Endpoint details:

- Created by: Vaghasiya Nehal
- Created on: Feb 27, 2024 5:44 PM
- Last updated on: Feb 27, 2024 5:44 PM
- Authentication type: Key
- Public network access: Enabled
- Swagger URI: https://tv1-ai-experiments-zgryt.eastus2.inference.ml.azure.com/swagger.json
- REST endpoint: https://tv1-ai-experiments-zgryt.eastus2.inference.ml.azure.com/score
- Metrics: View metrics

Tags:

- No tags

Test results:

- Live traffic: 100%
- Scoring script: Auto-generated
- Provisioning state: Succeeded
- Error details: --
- SKU: Standard_NC24ads_A100_v4
- Quota type: Temporary - 6d 23h 38m left
- Egress public network access: Enabled
- Instance count: 1
- Scaling: Configure auto scaling
- Model ID: azureml://registries/azureml/models/mistralai-Mistral-7B-Instruct-v01/versions/5
- Environment: Auto-generated
- Application Insights enabled: --

• Test the Endpoint:

- After deployment, you can test the endpoint by sending requests to it.
- Click on **"Consume"** and you can find **REST endpoint** and **authentication keys**.
- Use the provided endpoint URL and any necessary authentication keys to interact with your LLM programmatically.
- Click on **"Test"** Verify that the endpoint is working as expected by sending sample inputs and analyzing the outputs.

