Deploying Large Language Model (LLM) on Azure

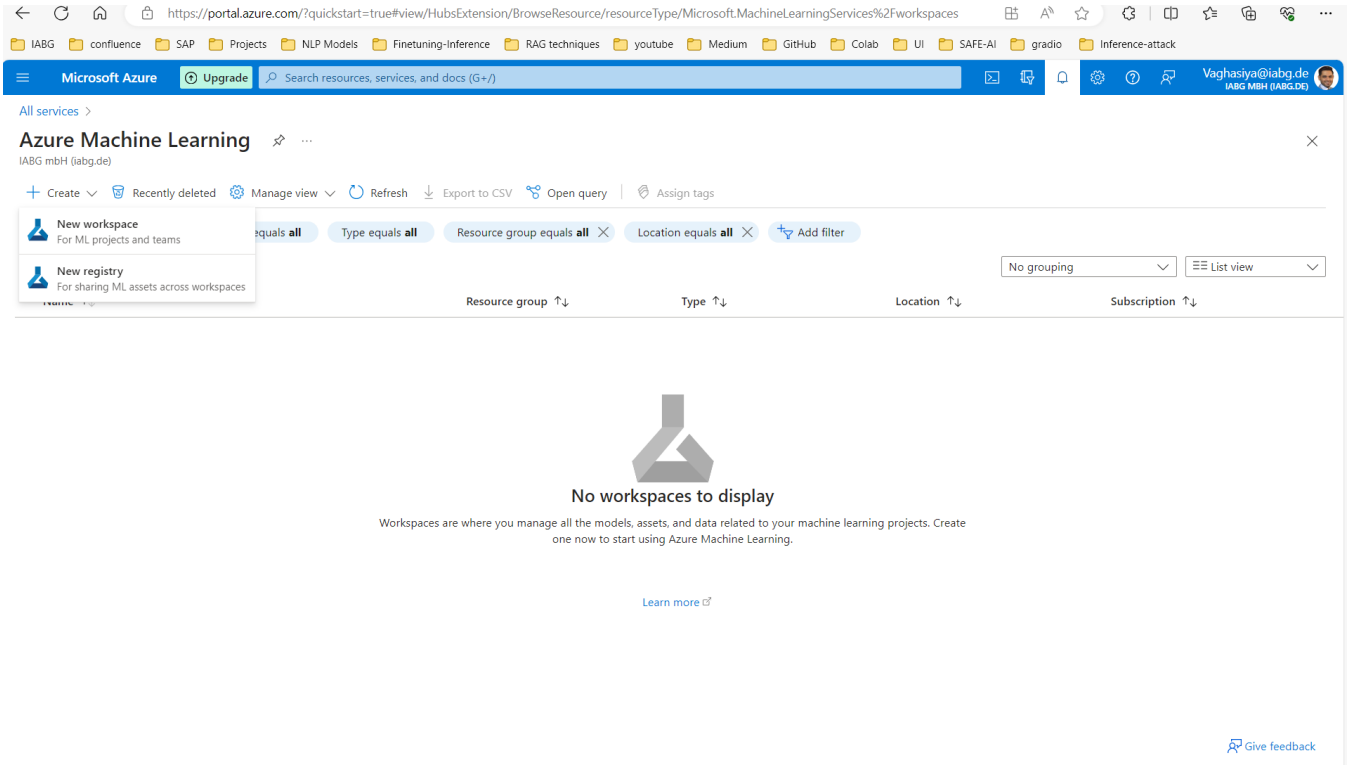
Azure Machine Learning provides a powerful platform for running and deploying large language models (LLMs). This guide will walk you through the steps to deploy your LLM and create an endpoint for your LLM in Azure Machine Learning.

Prerequisites:

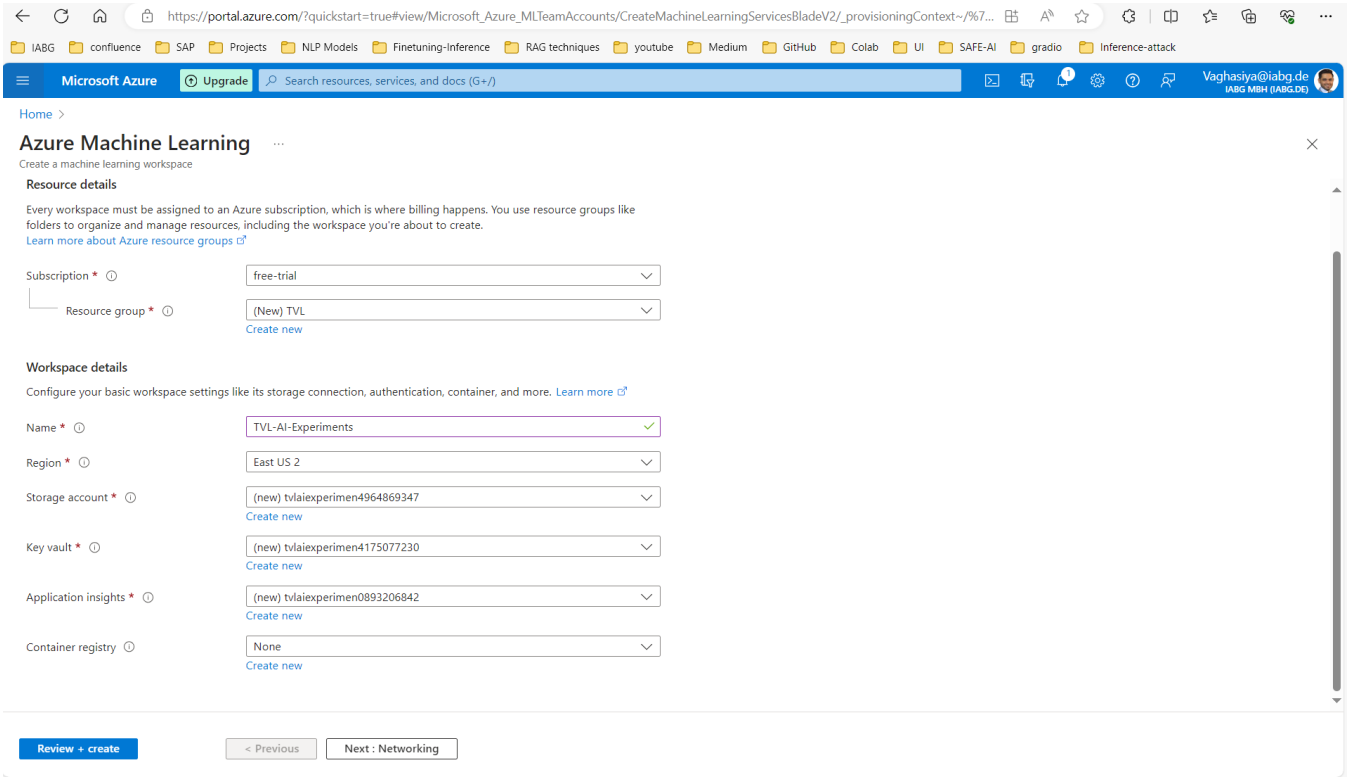
- An active Azure account - Setting Up a free Azure Account - NLP Community - Confluence (iabg.de)

Steps to Deploy LLM on Azure:

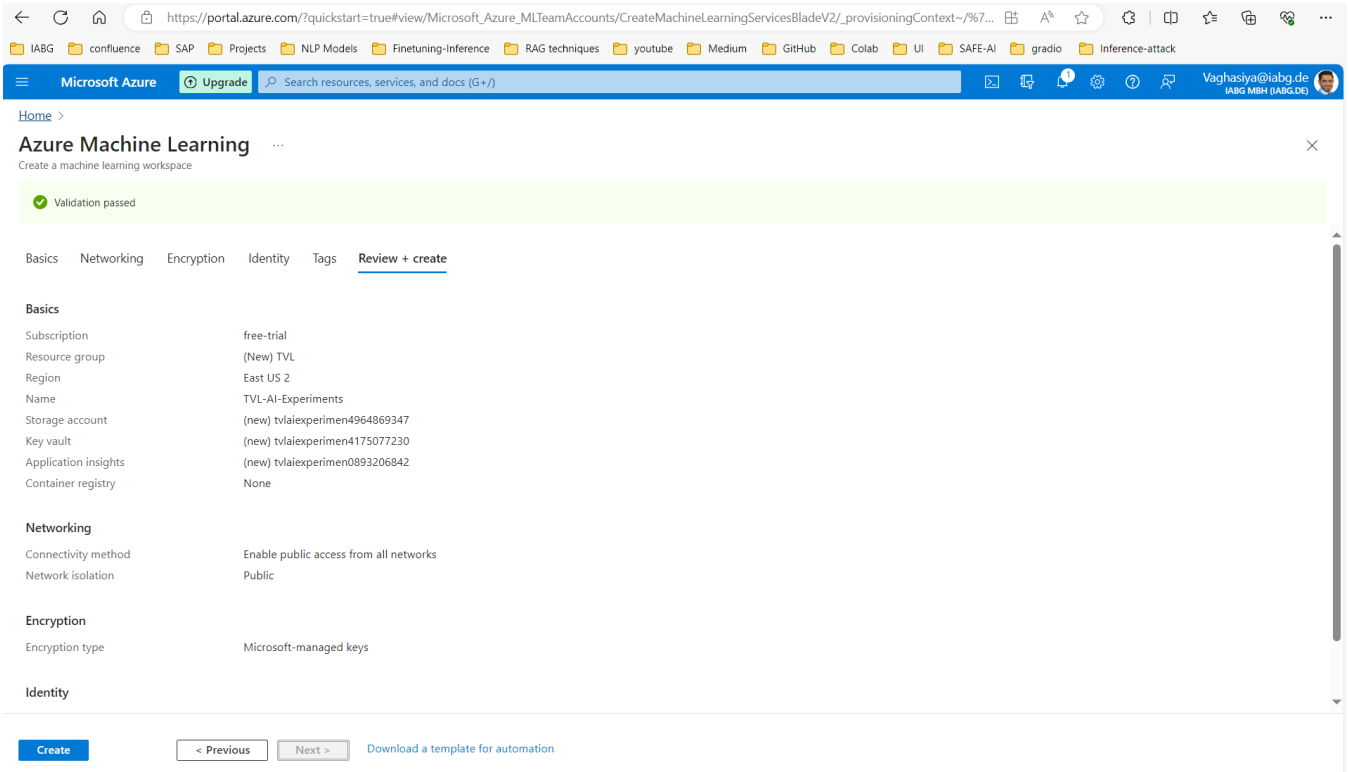
- **Sign in to Azure Portal:**
    - Open your web browser and navigate to the Azure Portal: https://portal.azure.com.
    - Sign in with your Azure account credentials.
- **Create a New Azure Machine Learning Workspace:**
    - In the Azure Portal, search for "**Azure Machine Learning**".
    - Select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

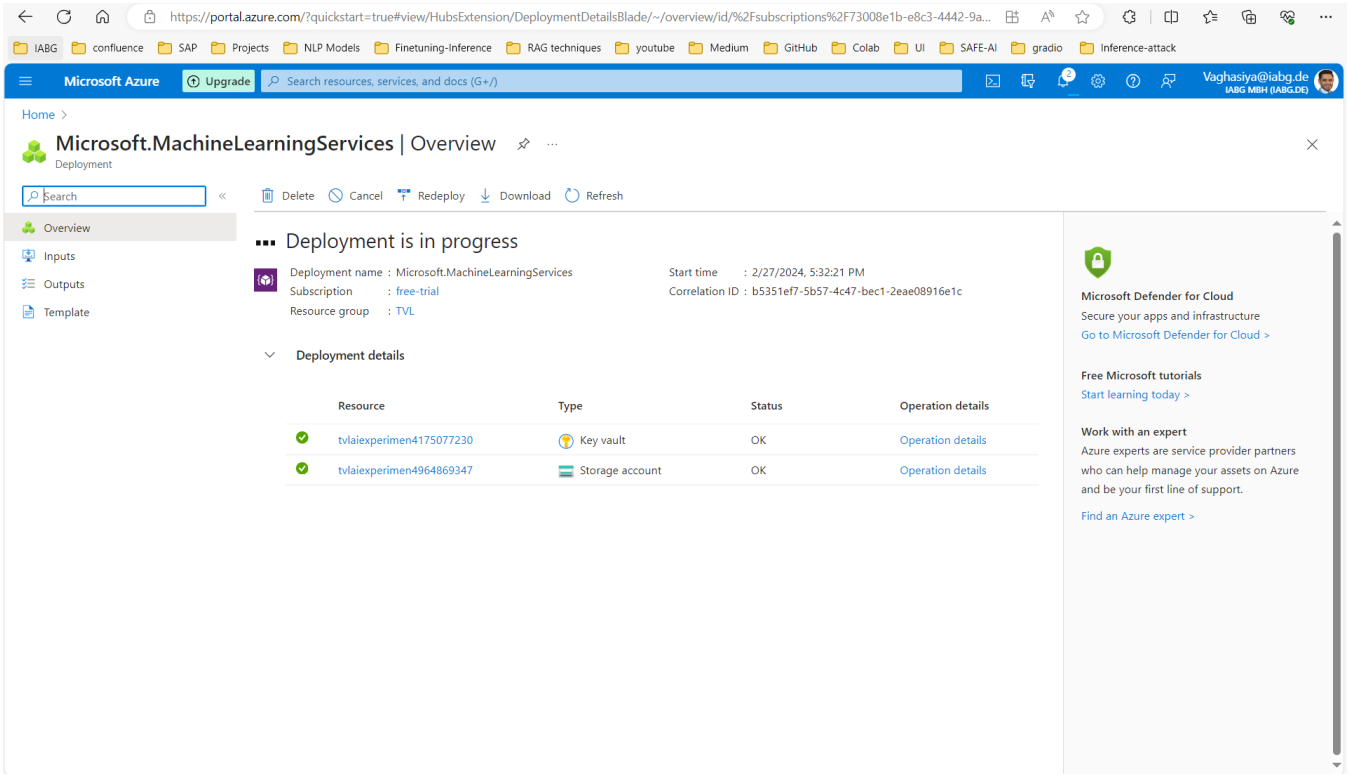


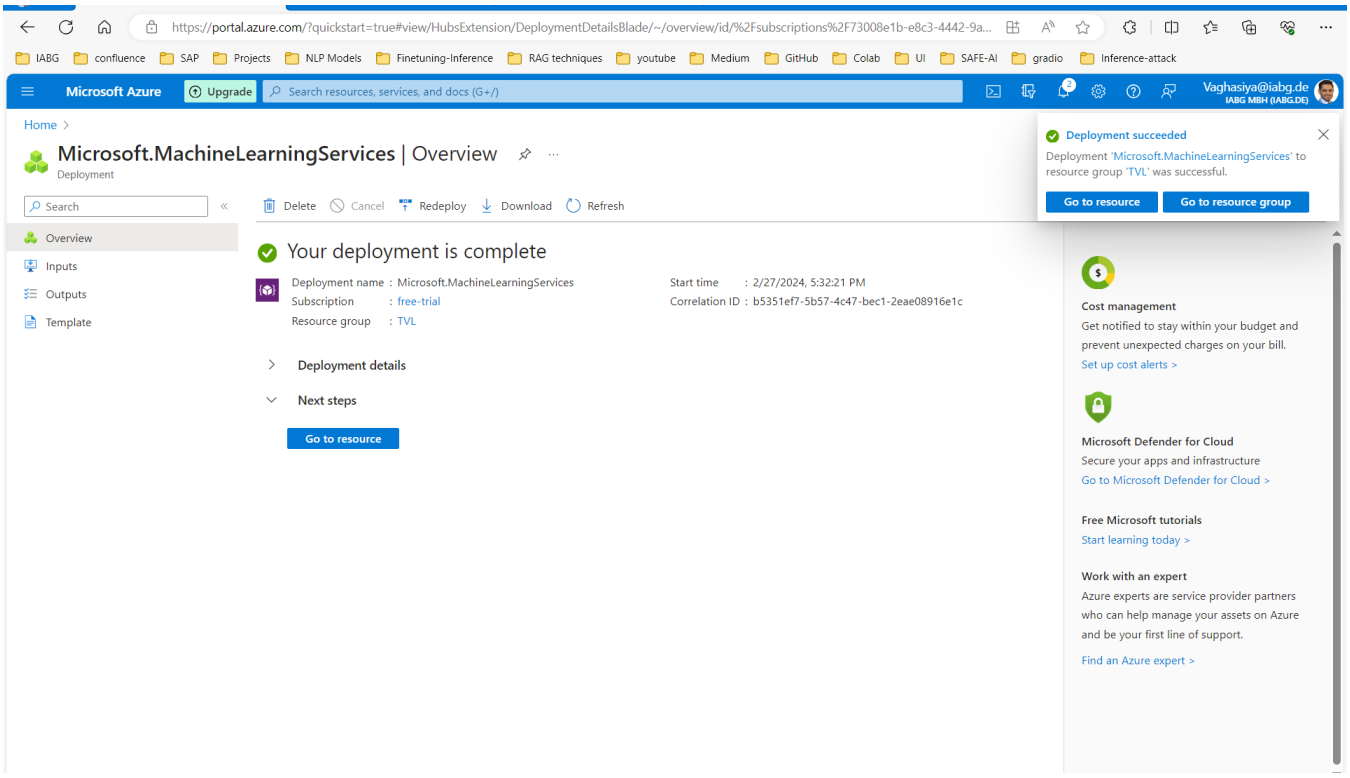
    - Fill out the required information to create a new Azure Machine Learning workspace, such as subscription, resource group, workspace name, and region as mentioned in the following images.
    - Keep the default values for **Networking**, **Encryption**, **Identity** and **Tags**.



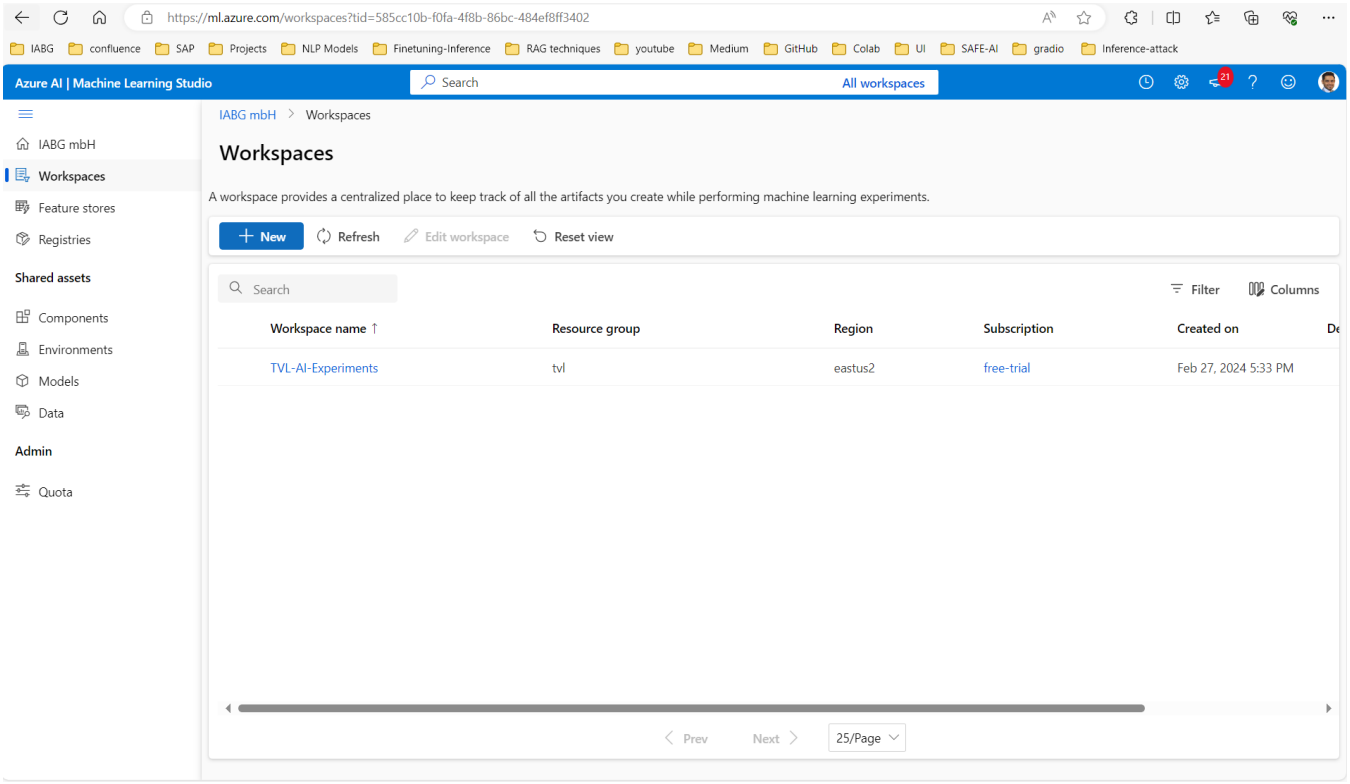
- Click on "**Review + Create**" and then "**Create**" to provision the workspace.



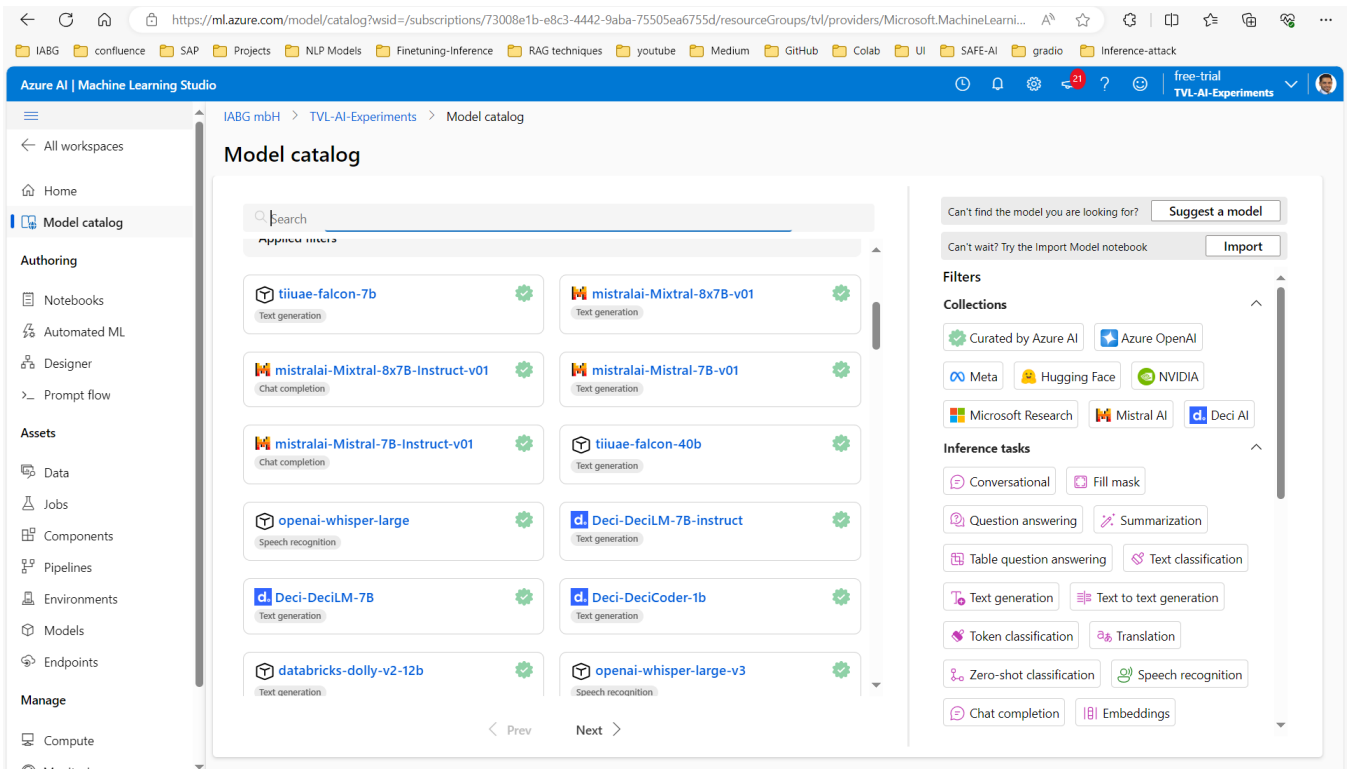




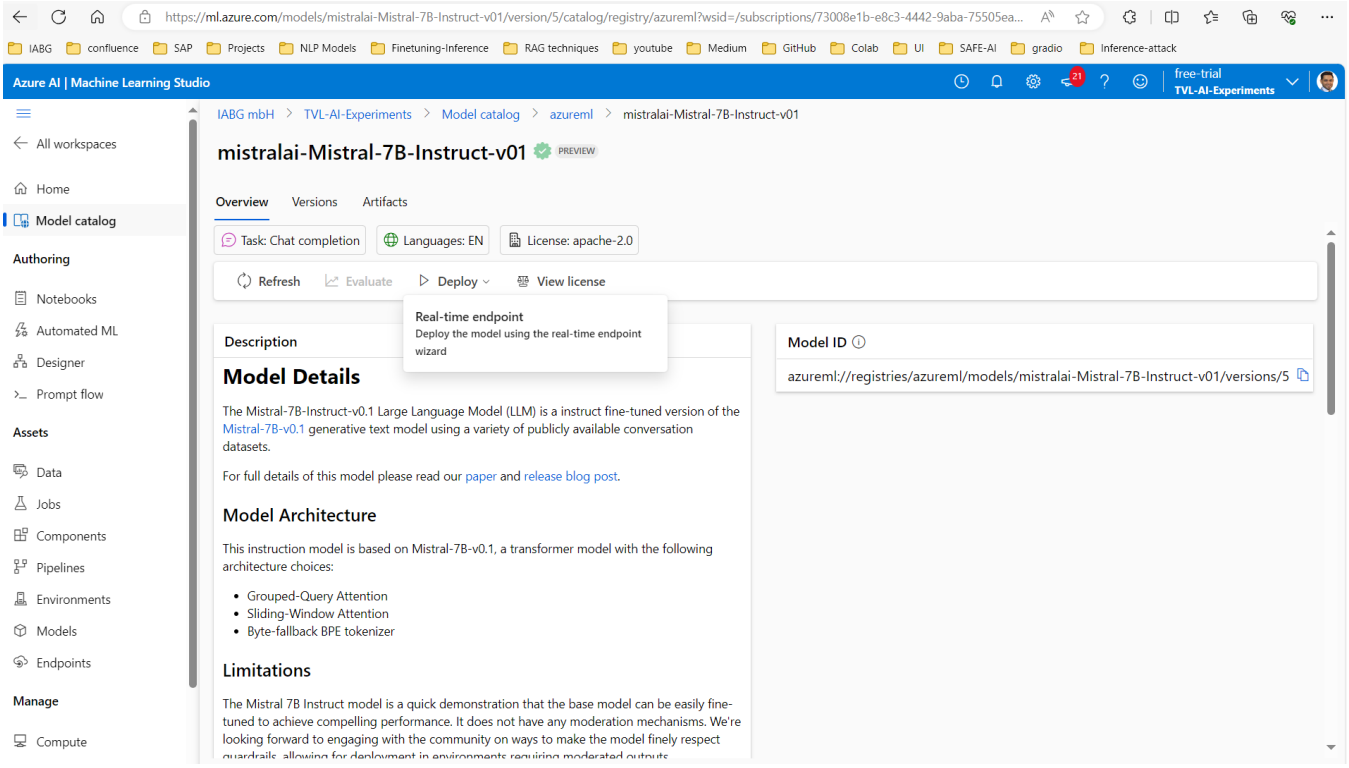
- **Deploy LLM and create LLM Endpoint:**
  - Navigate to your newly created Azure Machine Learning workspace. https://ml.azure.com/.
  - Go to **Workspaces.**
  - Click on the workspace that you recently created.



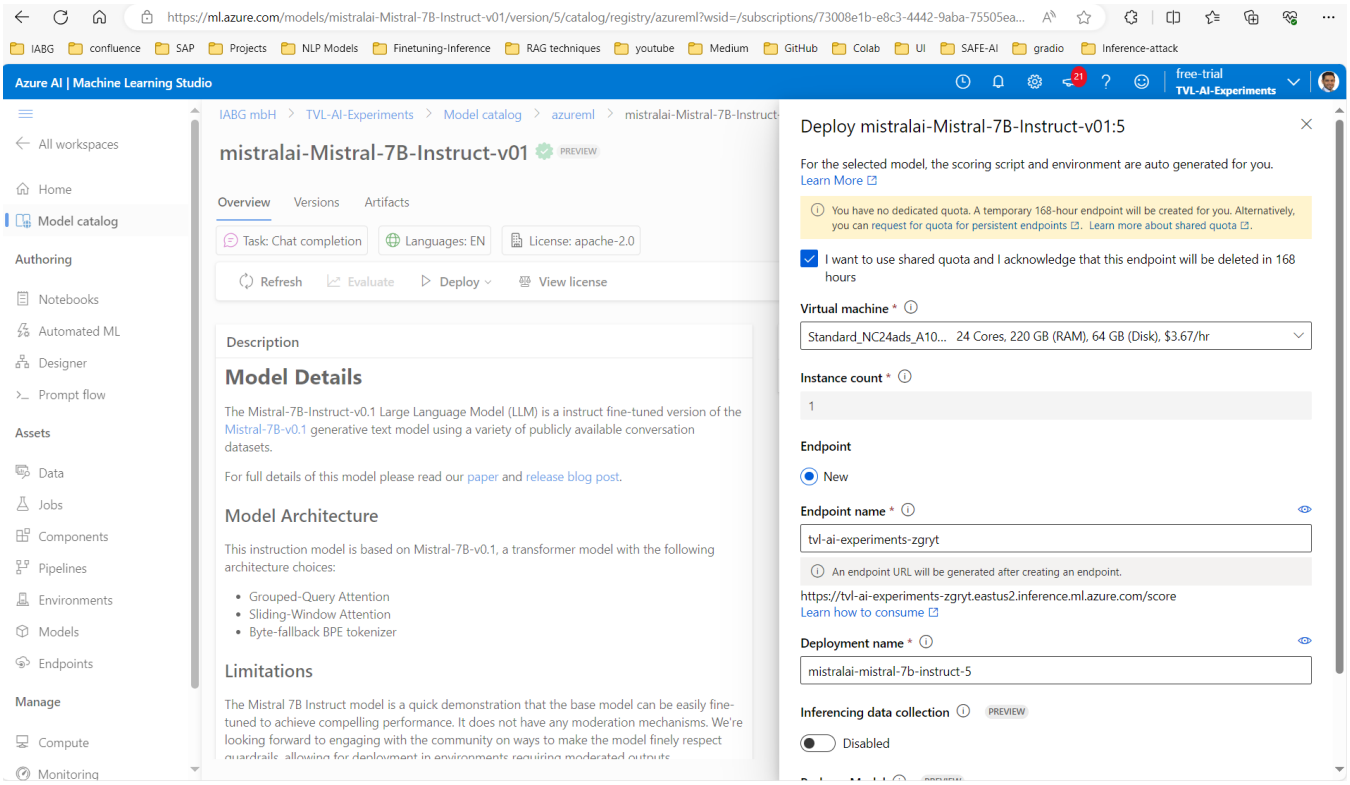
- Go to **Model Catalogs** and select any Open-Source Model from Huggingface.



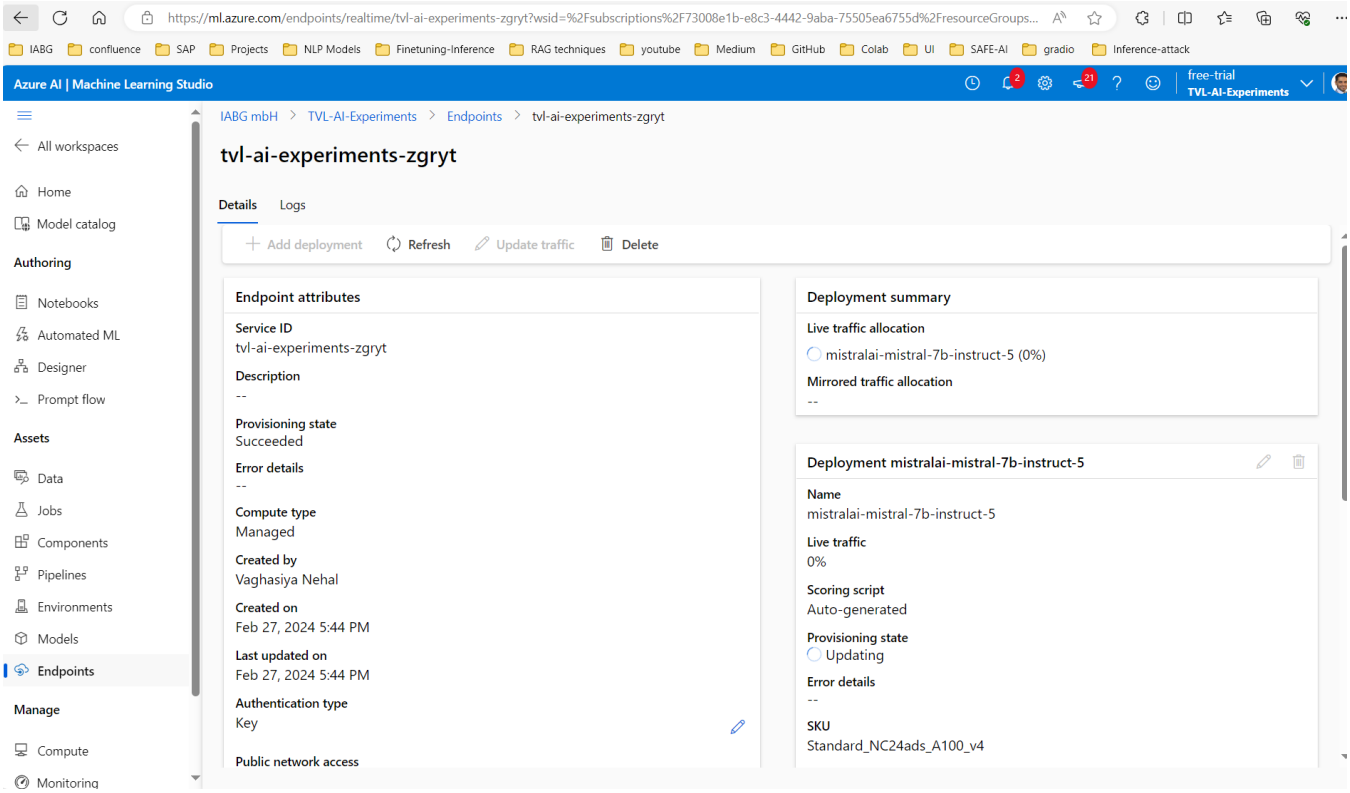
- Click on **Deploy Real-Time Endpoint** after selecting the Open-Source LLM(For example: **mistralai-Mistral-7B-Instruct-v01**)

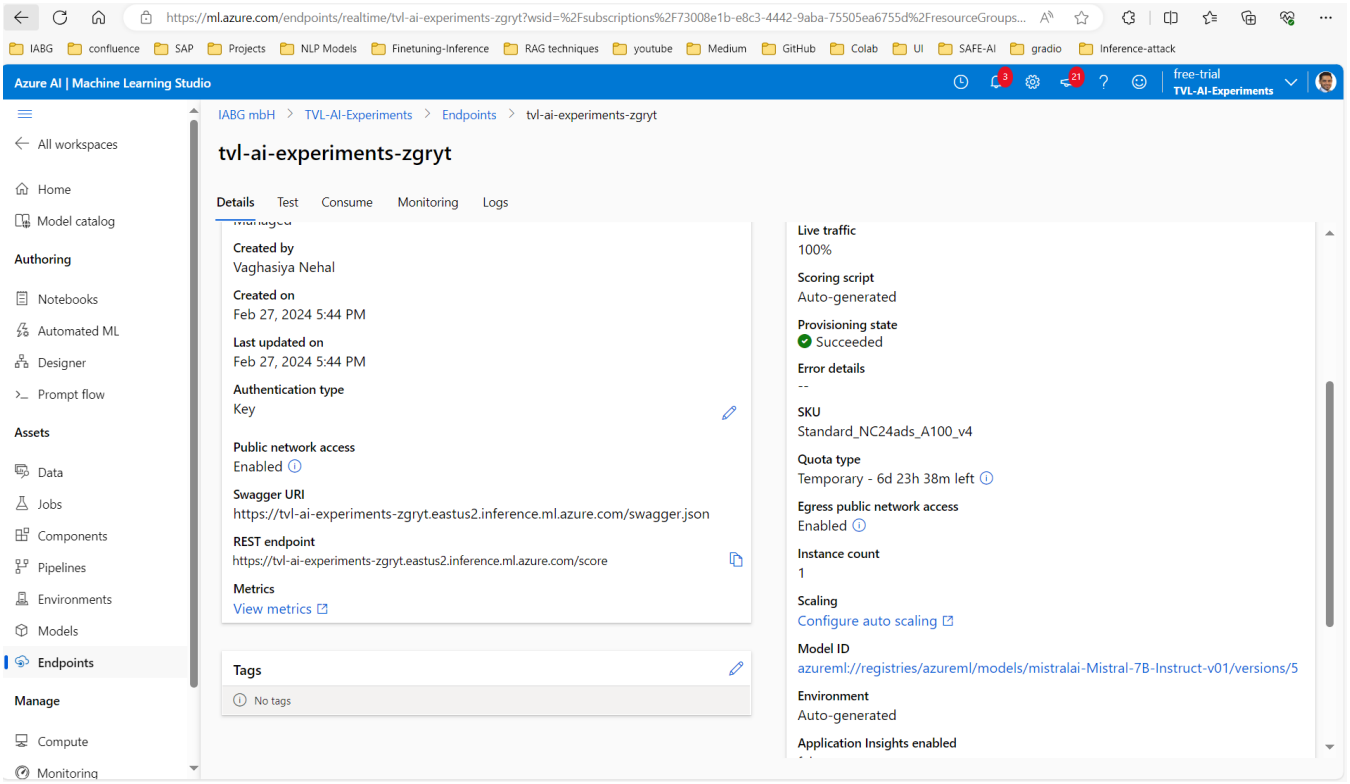


- **Virtual Machine and Instance Count:**
  - Choose the virtual machine according to your LLM requirements and adjust the instance count as needed.

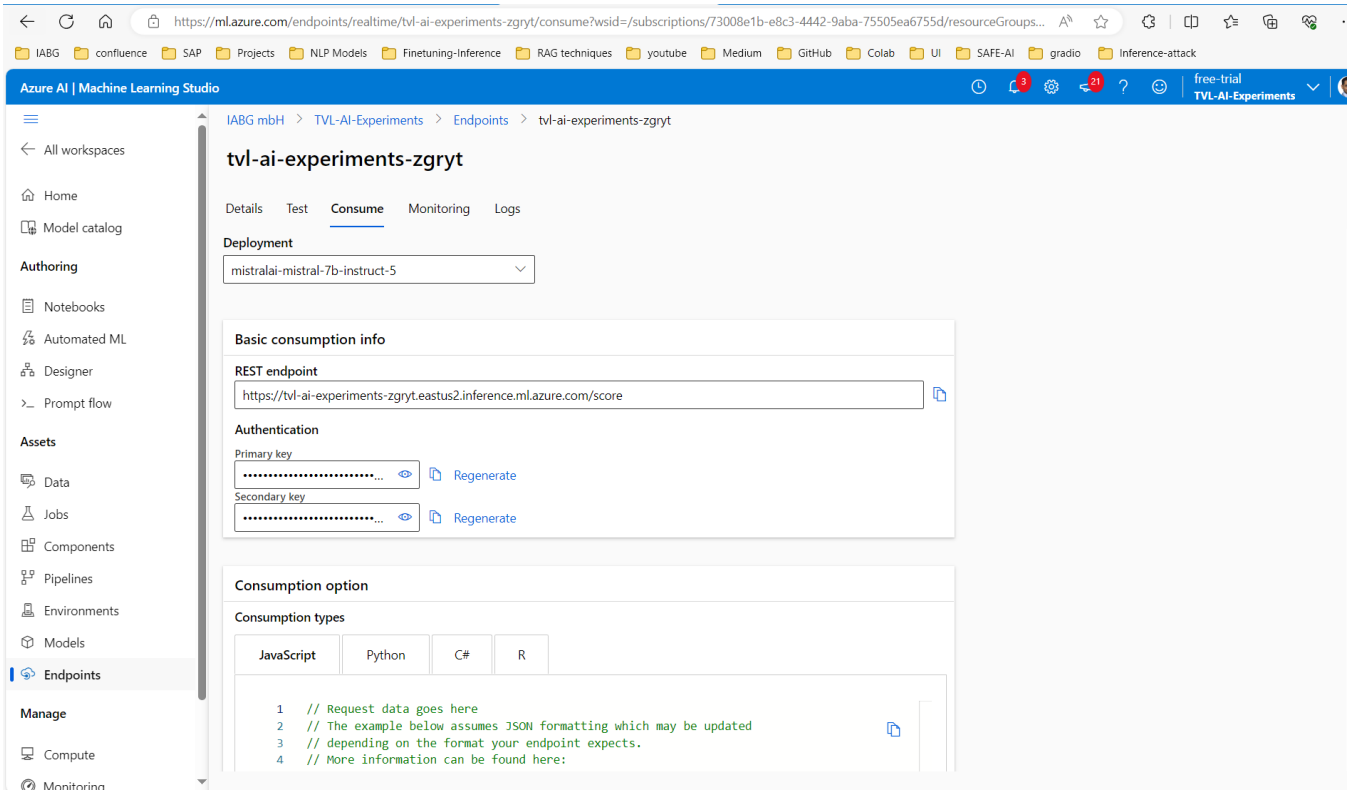


- **Create LLM Endpoint:**
  - Once you click on **Deploy**, you will be redirected to the following page.
  - It will take few minutes to deploy the LLM and create endpoint.
  - Once it is deployed, **Provisioning State** will show **Succeeded**





- **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.


https://ml.azure.com/endpoints/realtime/tvl-ai-experiments-zgryt/consume?wsid=/subscriptions/73008e1b-e8c3-4442-9aba-75505ea6755d/resourceGroups...
IABG
confluence
SAP
Projects
NLP Models
Finetuning-Inference
RAG techniques
youtube
Medium
GitHub
Colab
UI
SAFE-AI
gradio
Inference-attack
Azure AI | Machine Learning Studio
free-trial
TVL-AI-Experiments
All workspaces
Home
Model catalog
Authoring
Notebooks
Automated ML
Designer
Prompt flow
Assets
Data
Jobs
Components
Pipelines
Environments
Models
Endpoints
Manage
Compute
Monitoring
IABG mbH > TVL-AI-Experiments > Endpoints > tvl-ai-experiments-zgryt
tvl-ai-experiments-zgryt
Details
Test
Consume
Monitoring
Logs
Deployment
mistralai-mistral-7b-instruct-5
Basic consumption info
REST endpoint
https://tvl-ai-experiments-zgryt.eastus2.inference.ml.azure.com/score
Authentication
Primary key
Regenerate
Secondary key
Regenerate
Consumption option
Consumption types
JavaScript
Python
C#
R
1 // Request data goes here
2 // The example below assumes JSON formatting which may be updated
3 // depending on the format your endpoint expects.
4 // More information can be found here: