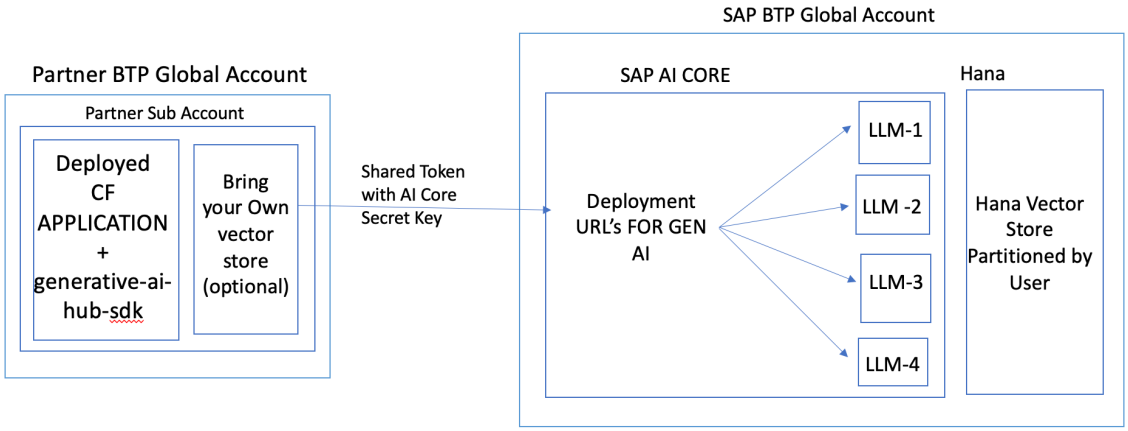
Starter Guide

Reference overview architecture is as the following.



We can deploy the following models on the service Here is the list of available deployment ID's:

- **gpt-35-turbo:** chat model Azure Open AI GPT 3.5
- **gpt-4:** Azure Open AI GPT 4
- **text-embedding-ada-002-v2:** Azure Open AI ADA embedding used to generate text embeddings.
- **falcon-40B:** Opensource Flacon model deployed on AI core.

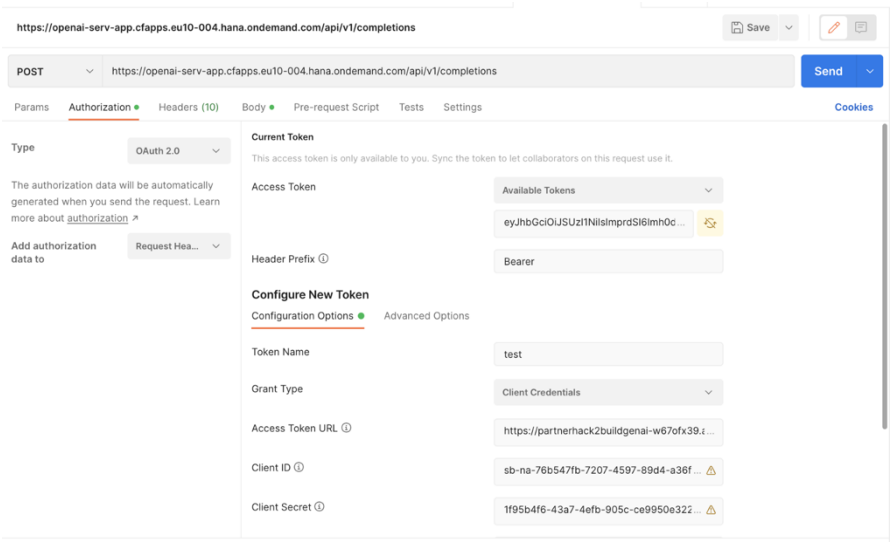
Please note to access the LLM's a deployment needs to be created in given instance of AI core. Please refer to tutorial.

1.Configuring Postman to Generative AI Hub: Please use client_id and client_secret of AI core instance provided to you.

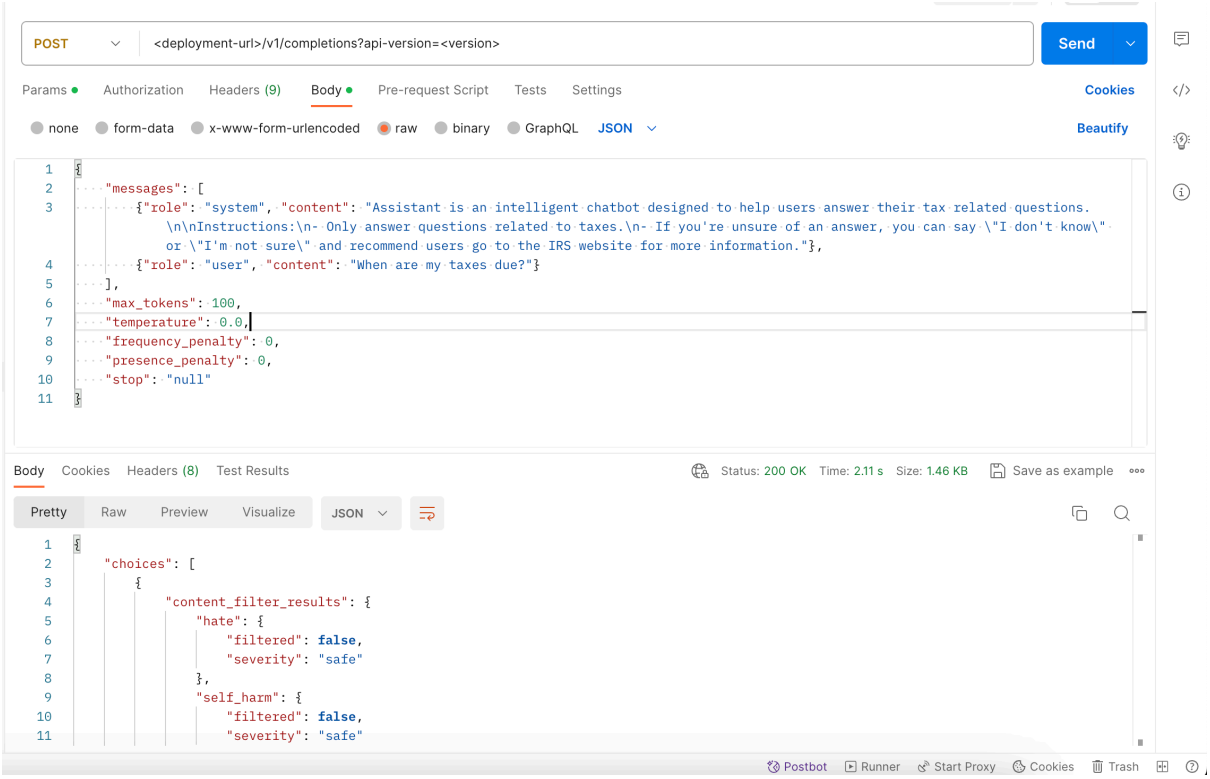
You need to make a post request to API_url. for example, here we used third party tool postman to validate if connection is happening successfully. (You can use any other tools to validate. It is mandatory to generate JWT token before validation)

Setup Authentication using oAuth 2.0.
Here the access token URL should be in the following format:

- https://<auth_url>/oauth/token?grant_type=client_credentials
- Add your client ID and client secret and click on Get access token



Below screenshot shows the sample usage of GPT-4 model



2.[Demo Video](#) of connecting to LLM through AI API using Postman.
3. [model JSON body for using different LLM's:](#) having payloads for each available LLM in AI Core.
4. [Generative hub Download and documentation](#) **:** hosted on PYPI.
5. [How to use and connect to LLM's using Gen AI HUB SDK](#): containing tutorial on how to install SAP Gen AI Hub SDK and code on how to use various LLM's.
6. [How to use Hana Vector store :](#) Via Gen AI Hub SDK.