

# SAP BTP - CAP RAG - AI Workshop

## Hands-on Guide

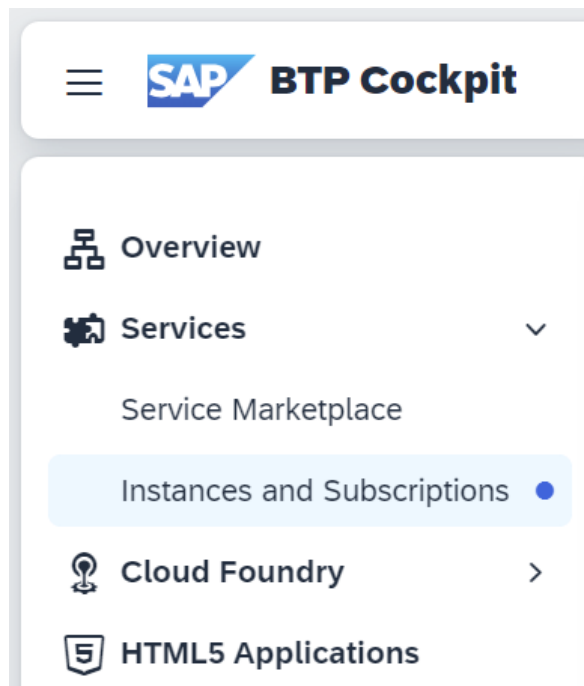
### Contents

Configure LLM for Usage in AI Launchpad .....	2
Clone Template Project from GitHub .....	8
Update AI Core Access Details in the Codebase.....	13
Create the Database Layer .....	14
Create the Embeddings OData Service .....	15
Create the Chat OData Service .....	16
Build and Deploy Application to SAP BTP Cloud Foundry Runtime .....	18
Build .....	18
Log in to Cloud Foundry .....	18
Deploy .....	21
Use the Embeddings App .....	22
Use the Chat App .....	23

# Configure LLM for Usage in AI Launchpad

We are going to create two configurations, one for a chat model and another for embedding model.

In the SAP BTP Cockpit, navigate to **Services > Instances and Subscriptions**.



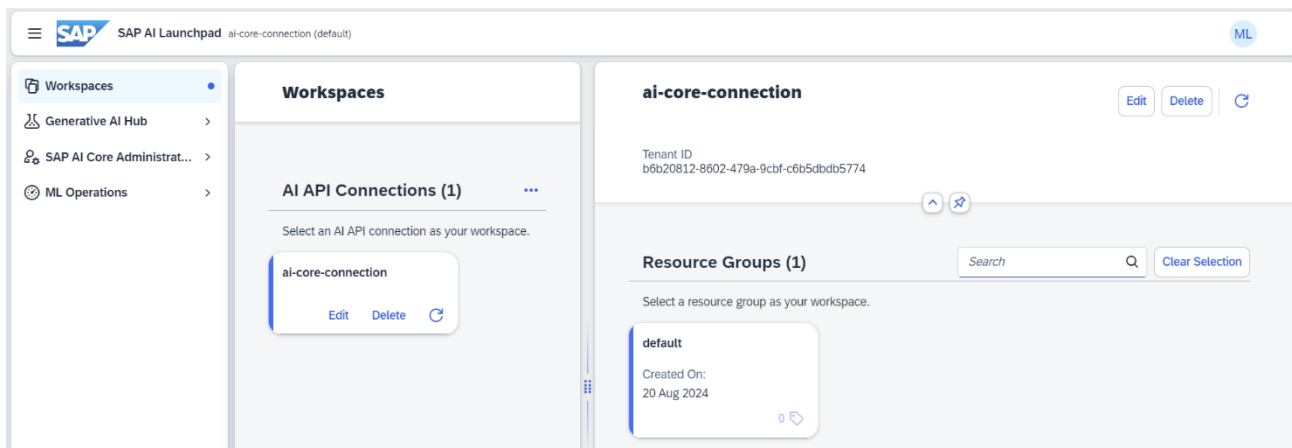
In the **Subscriptions tab**, click on **SAP AI Launchpad**. **SAP AI Launchpad** will open in a new tab, navigate to it.

Subscriptions (4)   Instances (4)   Environments (1)

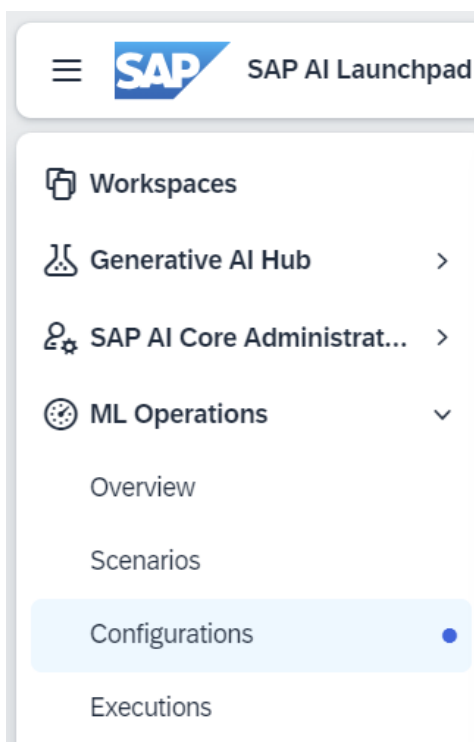
Applications to which your subaccount is currently subscribed

Application	Plan	Created ...	Changed...	Status	
<a href="#">SAP AI Launchpad</a>	standard	20 Aug 2...	20 Aug 2...	Subscribed	>
<a href="#">Continuous Integration &amp; Delivery</a>	default	20 Aug 2...	20 Aug 2...	Subscribed	>
<a href="#">SAP HANA Cloud</a>	tools	20 Aug 2...	20 Aug 2...	Subscribed	>
<a href="#">SAP Business Application Studio</a>	standard-edition	20 Aug 2...	20 Aug 2...	Subscribed	>

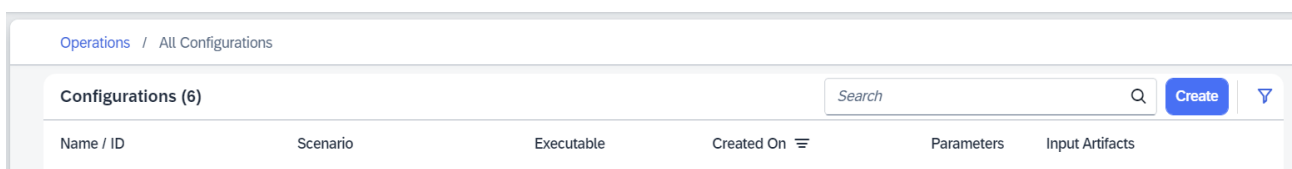
Select the **Resource Group** (default) created before.



Expand the **ML Operations** menu and click on **Configurations**.



Click on the **Create** button.



Enter a **Configuration Name** (gpt-4o), select the **Scenario** (foundation-models), select the **Version** (0.0.1), select the **Executable** (azure-openai). Click on the **Next** button.

Operations / All Configurations / Create Configuration

1 Enter Name and Executable 2 Input Parameters 3 Input Artifacts 4 Review

### 1. Enter Name and Executable

Configuration Name:\*

Scenario:\*

Version:\*

Executable:\*

Enter **modelName** (gpt-4o) and **modelVersion** (latest). Click on the **Next button**.

Operations / All Configurations / Create Configuration

1 Enter Name and Executable 2 Input Parameters 3 Input Artifacts 4 Review

### 2. Input Parameters

Configuration Parameters (2) [Reset](#) [Enable Description](#)

modelName:  Default: gpt-35-turbo

modelVersion:  Default: latest

Click on the **Review button**.

Operations / All Configurations / Create Configuration

1 Enter Name and Executable 2 Input Parameters 3 Input Artifacts 4 Review

### 3. Input Artifacts

No input artifacts are defined for this executable. Choose "Review" to proceed.

Click on the **Create button**.

Operations / All Configurations / Create Configuration

1 Enter Name and Executable

2 Input Parameters

3 Input Artifacts

4 Review

#### 4. Review

##### 1.Name and Executable

Configuration Name: gpt-4o  
Scenario Name: foundation-models  
Scenario Version: 0.0.1  
Executable Name: azure-openai

Edit

##### 2.Input Parameters

modelName: gpt-4o  
modelVersion: latest

Edit

##### 3.Input Artifacts

No input artifacts are defined for this executable.

Previous

Create

Cancel

From the newly created configuration, click on the **Create Deployment** button.

Operations / All Configurations / Configuration Details

## gpt-4o

Create Deployment

Configuration ID:  
4480898c-a889-4368-bd6c-594f975f2f6e

Scenario:  
foundation-models  
Version 0.0.1

Created On:  
today 5:57:48 pm

Executable:  
azure-openai

Parameters (2)

Search

Name	Description	Value
Optional Parameters		
modelVersion		latest Default: latest
modelName	supportedModels: gpt-35-turbo, gpt-4o, gpt-4, gpt-35-turbo-16... <a>Show More</a>	<b>gpt-4o</b> Default: gpt-35-turbo

Select the **Standard** option for **Duration**. Click on the **Review** button.

Operations / All Deployments / Create Deployment

1 Select Scenario

2 Select Executable

3 Select a configuration

4 Duration (Optional)

5 Review

4. Duration

Choose a timeframe for the deployment to be active.

☒ Standard

Use the standard duration which defaults from your environment.

☐ Custom

Click on the **Create button**.

Operations / All Deployments / Create Deployment

1 Select Scenario

2 Select Executable

3 Select a configuration

4 Duration (Optional)

5 Review

5. Review

1.Scenario

Scenario Name: foundation-models

Edit

2.Executable

Executable Name: azure-openai

Scenario Version: 0.0.1

Edit

3.Configuration

Configuration Name: gpt-4o

Edit

4.Duration

Duration: Standard

Edit

Previous

Create

Cancel

The **Deployment** will be created, and it will be available for usage when its status is **RUNNING**. The **Deployment ID** (dc9a9804353026a2) is the key used later by the applications you build.

Operations / All Deployments / Deployment Details

dc9a9804353026a2

Update

Stop

Delete

Current Status:

Created On:

Changed On:

Submitted On:

Started On:

Duration:

RUNNING

today 7:35:23 pm

today 7:37:38 pm

today 7:36:08 pm

today 7:37:13 pm

0 minutes 29 seconds

Target Status:

Running Until:

URL:

RUNNING

Unlimited

<https://api.ai.prod.ap-southeast-2.aws.ml.hana.ondemand.com/v2/inference/deployments/dc9a9804353026a2>

Now, repeat the same **Configuration** and **Deployment** steps to create the **embedding model**.

Operations / All Configurations / Configuration Details

text-embedding-ada-002

Create Deployment

Configuration ID:  
cc43bf55-6309-454b-9b3d-fa32344fc146

Scenario:  
foundation-models  
Version 0.0.1

Created On:  
26 Aug 2024, 10:34:57 pm

Executable:  
azure-openai

Parameters (2)

Search

Name	Description	Value
Optional Parameters		
modelVersion		latest Default: latest
modelName	supportedModels: gpt-35-turbo, gpt-4o, gpt-4, gpt-35-turbo-16... <a>Show More</a>	text-embedding-ada-002 Default: gpt-35-turbo

Operations / All Deployments / Deployment Details

dc8d095f102905b1

UpdateStopDeleteRefresh

Current Status:  
RUNNING

Created On:  
today 7:40:44 pm

Changed On:  
today 7:48:53 pm

Submitted On:  
today 7:42:10 pm

Started On:  
today 7:43:20 pm

Duration:  
5 minutes 51 seconds

Target Status:  
RUNNING

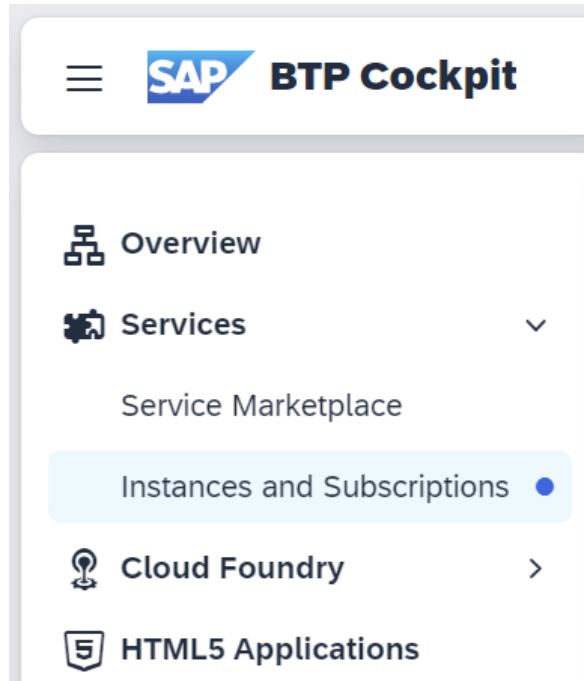
Running Until:  
Unlimited

URL:  
https://api.ai.prod.ap-southeast-2.aws.ml.hana.ondemand.com/v2/inference/deployments/dc8d095f102905b1

# Clone Template Project from GitHub

**SAP BAS Dev Space:** <https://help.sap.com/docs/bas/sap-business-application-studio/dev-spaces-in-sap-business-application-studio>

In the SAP BTP Cockpit, navigate to **Services > Instances and Subscriptions**.



On the **Subscriptions tab**, click on SAP Business Application Studio (BAS) to launch it in a new tab.

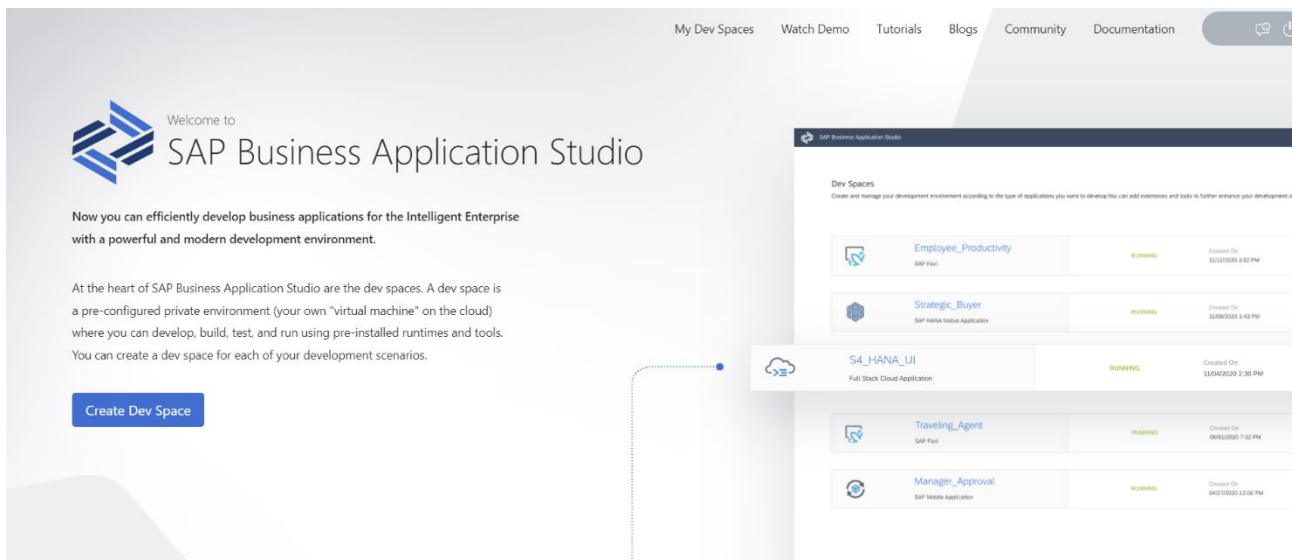
Subscriptions (5)   Instances (9)   Environments (1)

Applications to which your subaccount is currently subscribed

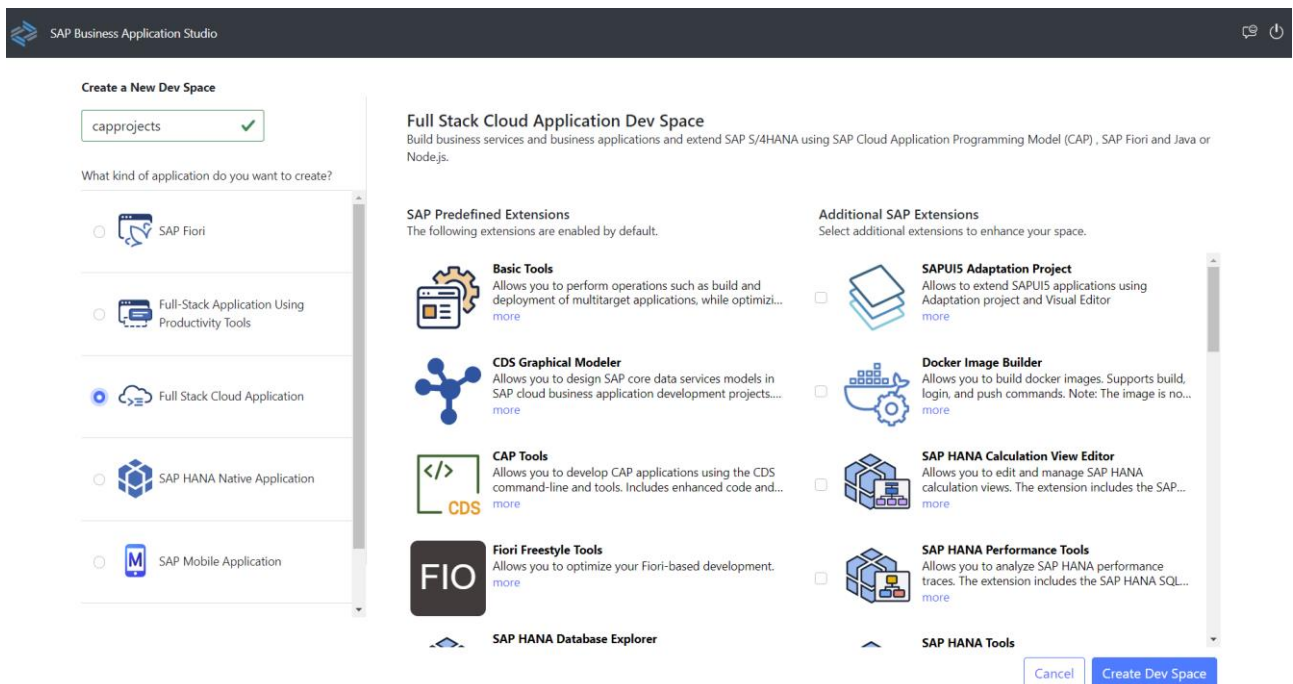
Application	Plan	Created On	Changed On	Status	
<a href="#">SAP AI Launchpad</a>	standard	20 Aug 2024	20 Aug 2024	Subscribed	... >
<a href="#">Continuous Integration &amp; Delivery</a>	default	20 Aug 2024	20 Aug 2024	Subscribed	... >
<a href="#">SAP HANA Cloud</a>	tools	20 Aug 2024	20 Aug 2024	Subscribed	... >
<a href="#">SAP Business Application Studio</a>	standard-edition	20 Aug 2024	20 Aug 2024	Subscribed	... >
<a href="#">SAP Build Work Zone, standard...</a>	standard	20 Aug 2024	20 Aug 2024	Subscribed	... >

If it is the first time you are accessing it, the welcome page will be displayed, and you will need to create a **Developer Space**. Click on the **Create Dev Space button**. This step is executed one time only.

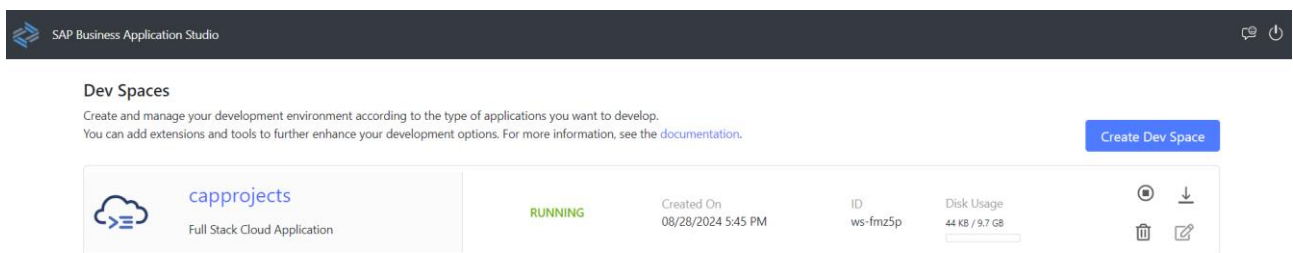




Give the **Dev Space** a **name**, select the **Full Stack Cloud Application** option, and click on **Create Dev Space** button.

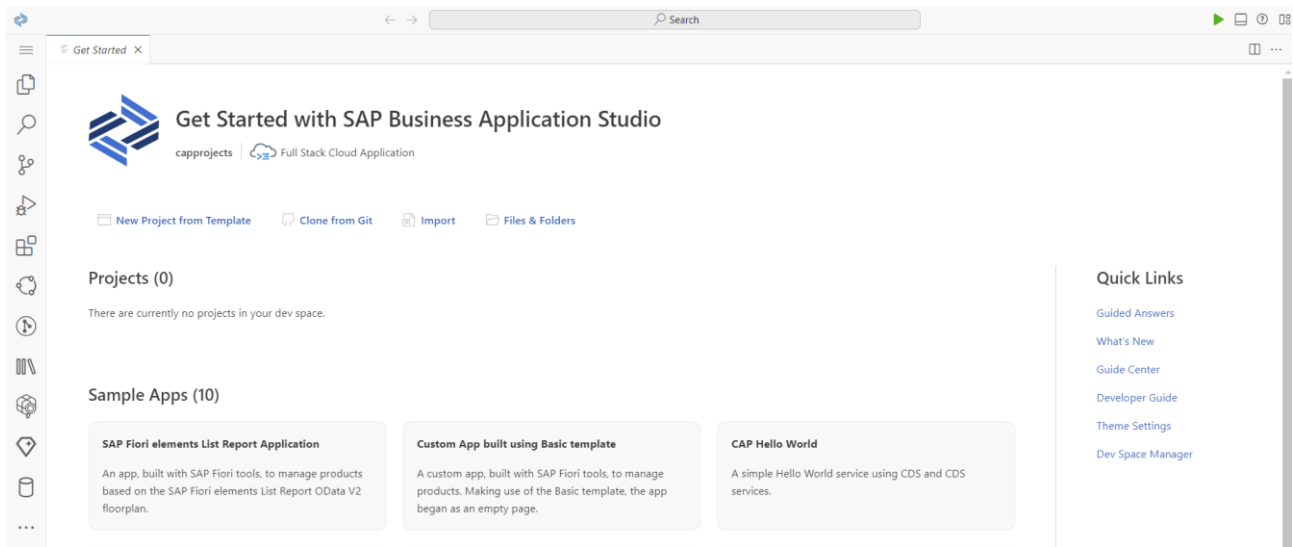


Once the **Dev Space** is created and it is **RUNNING**, click on its name to launch **BAS**.

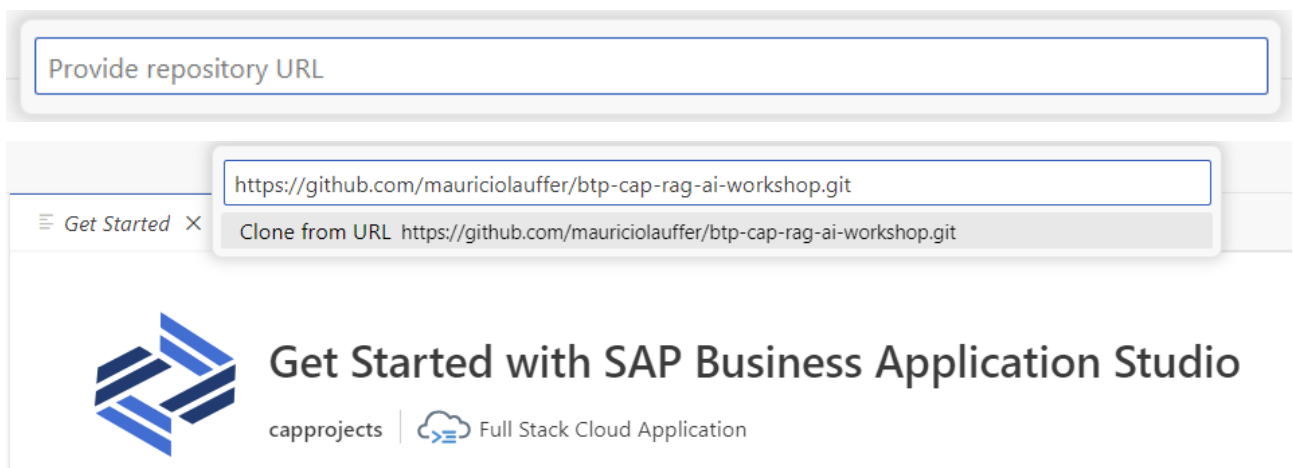


We will not start a project from scratch. We will clone a template project from GitHub.

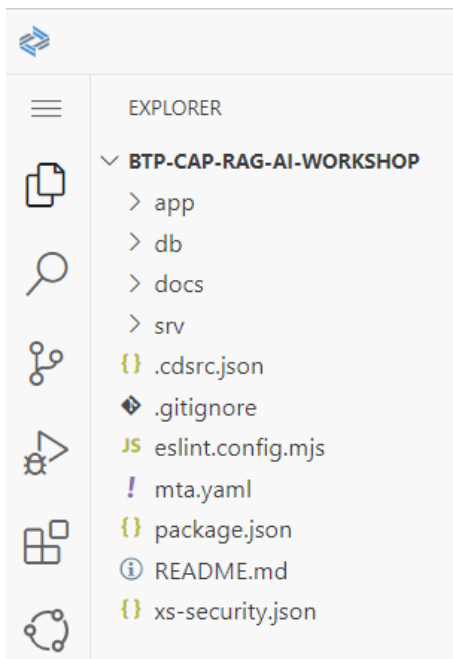
Click on **Clone from Git**.



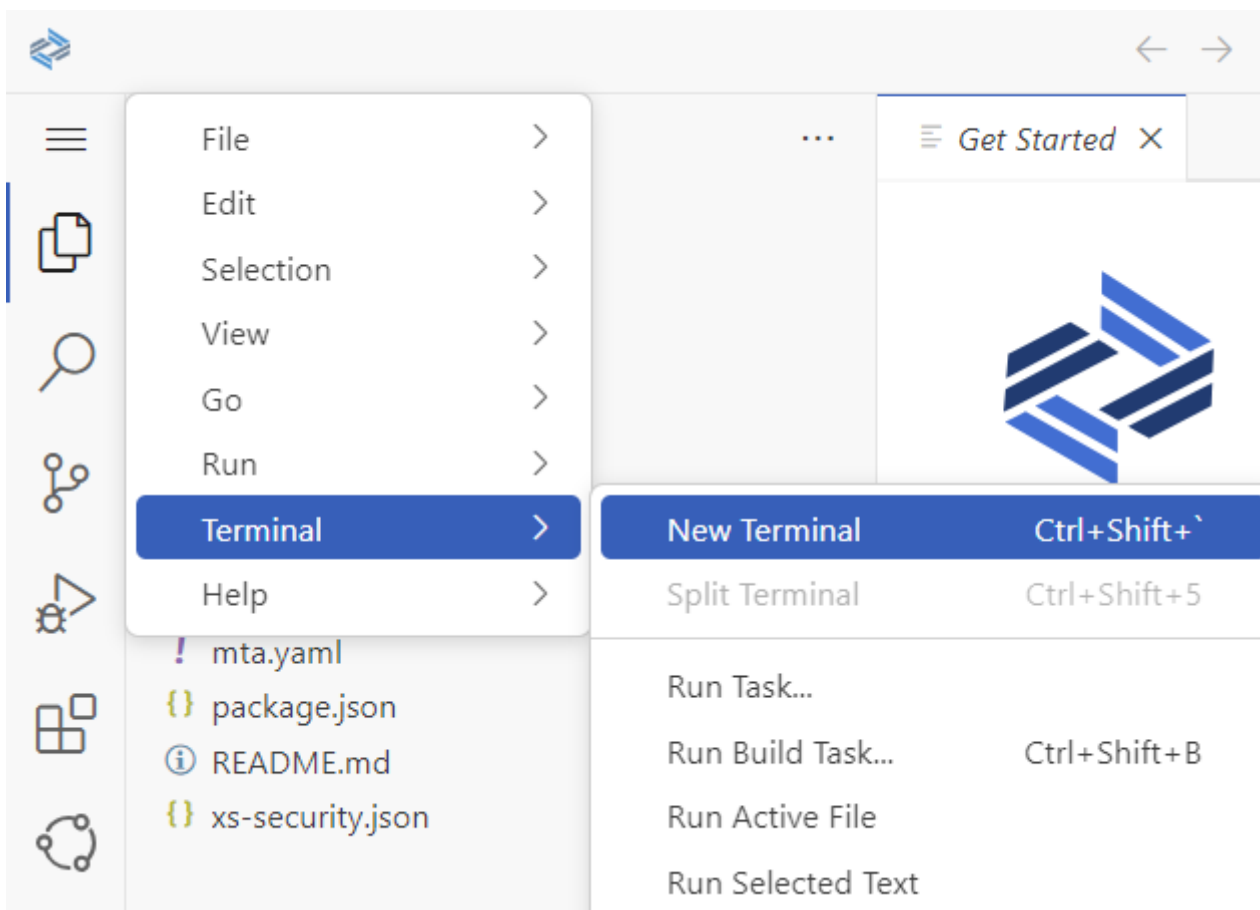
Enter the **git repository** URL (<https://github.com/mauriciolauffer/btp-cap-rag-ai-workshop.git>) and press **ENTER**.



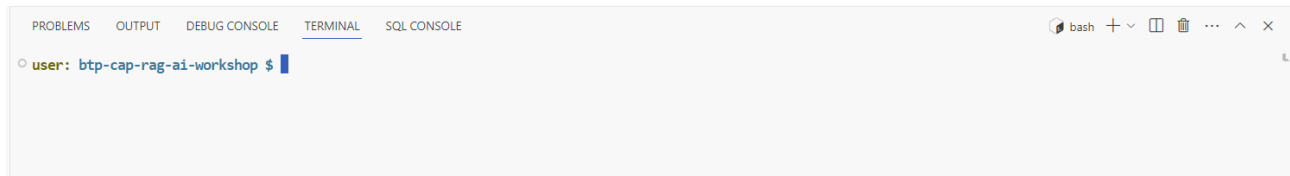
The template project will be cloned and opened. You can start exploring the codebase in the **Explorer** section.



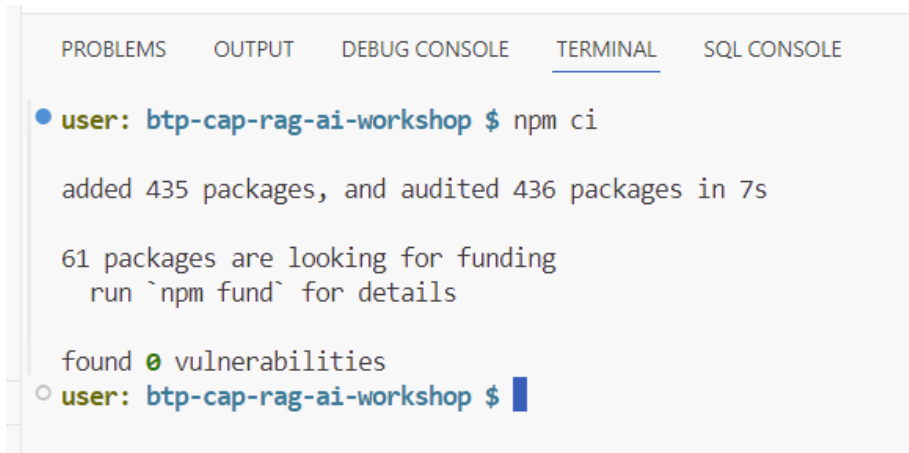
Open the Terminal to install the project dependencies. Click on the **hamburger button** (3 stacked lines icon), then **Terminal > New Terminal**.



The Terminal should open at the bottom of the screen, install the project dependencies with the command **\$ npm ci**



A screenshot of a VS Code terminal window. The terminal is open at the bottom of the screen. The top bar shows tabs for PROBLEMS, OUTPUT, DEBUG CONSOLE, TERMINAL (selected), and SQL CONSOLE. The terminal content shows the prompt `user: btp-cap-rag-ai-workshop $` with a blue cursor.



A screenshot of a VS Code terminal window showing the output of the `npm ci` command. The terminal content is as follows:

```
user: btp-cap-rag-ai-workshop $ npm ci

added 435 packages, and audited 436 packages in 7s

61 packages are looking for funding
  run `npm fund` for details

found 0 vulnerabilities
user: btp-cap-rag-ai-workshop $
```

# Update AI Core Access Details in the Codebase

Open the file **.cdsrc.json** which contains CAP related configuration. The **genAI Hub** access details are also there under **GENERATIVE\_AI\_HUB** key.

Replace the placeholder **\_\_\_PLACEHOLDER\_CHAT\_\_\_** with the **Deployment ID** from the **chat model** in your **AI Launchpad Service**.

Replace the placeholder **\_\_\_PLACEHOLDER\_EMBED\_\_\_** with the **Deployment ID** from the **embedding model** in your **AI Launchpad Service**.

```
13     "GENERATIVE_AI_HUB": {
14         "CHAT_MODEL_DESTINATION_NAME": "AICoreAzureOpenAIDestination",
15         "CHAT_MODEL_DEPLOYMENT_URL": "/inference/deployments/___PLACEHOLDER_CHAT___",
16         "CHAT_MODEL_RESOURCE_GROUP": "default",
17         "CHAT_MODEL_NAME": "gpt-4o",
18         "CHAT_MODEL_API_VERSION": "2023-05-15",
19         "EMBEDDING_MODEL_DESTINATION_NAME": "AICoreAzureOpenAIDestination",
20         "EMBEDDING_MODEL_DEPLOYMENT_URL": "/inference/deployments/___PLACEHOLDER_EMBED___",
21         "EMBEDDING_MODEL_RESOURCE_GROUP": "default",
22         "EMBEDDING_MODEL_NAME": "text-embedding-ada-002",
23         "EMBEDDING_MODEL_API_VERSION": "2023-05-15"
24     },
25     "AICoreAzureOpenAIDestination": {
26         "kind": "rest",
27         "credentials": {
28             "destination": "GENERATIVE_AI_HUB",
29             "requestTimeout": "300000"
30         }
31     }
```

# Create the Database Layer

Open the file **db/schema.cds** and create the tables:

```
namespace btpcapragai;

using {
    cuid,
    managed
} from '@sap/cds/common';

entity Conversation : cuid, managed {
    userId    : String;
    title     : String;
    messages  : Composition of many Message
                on messages.conversation = $self;
}

entity Message : cuid, managed {
    conversation : Association to Conversation;
    role        : String;
    content      : LargeString;
}

entity DocumentChunk : cuid {
    text_chunk    : LargeString;
    metadata_column : LargeString;
    embedding      : Vector(1536);
}

entity Files : cuid, managed {
    @Core.MediaType : mediaType @Core.ContentDisposition.FileName: fileName
    content         : LargeBinary;
    @Core.IsMediaType: true
    mediaType       : String;
    fileName        : String;
    size            : String;
}
```

# Create the Embeddings OData Service

Open the file **srv/embedding-service.cds** and create the OData Service:

```
using {btpcapragai as db} from '../db/schema';
service EmbeddingService {
    entity DocumentChunk as
        projection on db.DocumentChunk
        excluding {
            embedding
        };
    entity Files as projection on db.Files;
    action deleteEmbeddings() returns String;
}
annotate EmbeddingService with @(requires: 'authenticated-user');
```

Open the file **srv/embedding-service.js** and create the code to handle the service. The OData Action **deleteEmbeddings** and the OData Entity **Files** need to be implemented. However, the **OData Entity Files** will only have a custom handler for the **ON UPDATE** event.

The **OData Entity Files** ON UPDATE event will have to:

- read the filename for the current File ID being updated
- prepare PDF to be split in chunks of text for embedding
- split the PDF in chunks of text
- get the configuration to use the embedding model
- convert the chunks of text to vectors (embedding)
- insert the vectors into the **DocumentChunk** table

The OData Action **deleteEmbeddings** will have to:

- delete data from Files and DocumentChunk tables

The whole implementation can be found and copied from here: <https://github.com/mauriciolauffer/btp-cap-rag-ai-workshop/blob/main/srv/embedding-service.js>

# Create the Chat OData Service

Open the file **srv/chat-service.cds** and create the OData Service:

```
using {btpcapragai as db} from '../db/schema';

type RagResponse_AdditionalContents {
    score      : String;
    pageContent : String;
}

type RagResponse {
    role          : String;
    content       : String;
    timestamp     : String;
    additionalContents : array of RagResponse_AdditionalContents;
}

service ChatService {
    entity Conversation as projection on db.Conversation;
    entity Message      as projection on db.Message;
    action getAiResponse(sessionId : String, content : String, timestamp :
Timestamp) returns RagResponse;
    action deleteChatSession(sessionId :
UUID) returns String;
}

annotate ChatService with @(requires: 'authenticated-user');
```

Open the file **srv/chat-service.js** and create the code to handle the service. The OData Actions **getAiResponse** and **deleteChatSession** need to be implemented.

The OData Action **getAiResponse** will have to:

- get the configuration to use the embedding model
- get the configuration to use the chat model
- handle the chat session/history
- get the RAG response from HANA Vector + LLM
- process the response
- return response to user

The OData Action **deleteChatSession** will have to:

- delete the chat session data



The whole implementation can be found and copied from here: <https://github.com/mauriciolauffer/btp-cap-rag-ai-workshop/blob/main/srv/chat-service.js>

# Build and Deploy Application to SAP BTP Cloud Foundry Runtime

Access the Terminal to execute pre-defined NPM scripts to build and deploy the application to SAP BTP Cloud Foundry Runtime.

```
{} package.json > ...  
28     "scripts": {  
29  
30         "build": "mbt build -t gen --mtar archive",  
31         "deploy": "cf deploy gen/archive.mtar --retries 1 --delete-services",  
32  
33     },
```

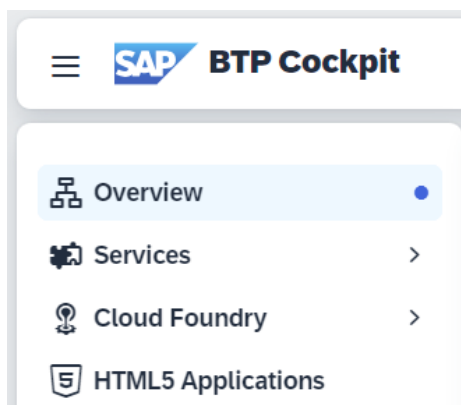
## Build

In the terminal, execute **\$ npm run build**

```
PROBLEMS 2 OUTPUT DEBUG CONSOLE TERMINAL SQL CONSOLE  
user: btp-cap-rag-ai-workshop $ npm run build  
[2024-08-28 13:29:16] INFO copying the "embedding.zip" pattern from the "/home/user/projects/btp-cap-rag-ai-workshop/app/embedding/dist" folder to the "/home/  
user/projects/btp-cap-rag-ai-workshop/gen/app" folder  
[2024-08-28 13:29:16] INFO the build results of the "btp-cap-rag-ai-app-deployer" module will be packaged and saved in the "gen/.btp-cap-rag-ai-workshop_mta_b  
uild_tmp/btp-cap-rag-ai-app-deployer" folder  
[2024-08-28 13:29:16] INFO finished building the "btp-cap-rag-ai-app-deployer" module  
[2024-08-28 13:29:16] INFO running the "after-all" build...  
[2024-08-28 13:29:16] INFO generating the metadata...  
[2024-08-28 13:29:16] INFO generating the "/home/user/projects/btp-cap-rag-ai-workshop/gen/.btp-cap-rag-ai-workshop_mta_build_tmp/META-INF/mtad.yaml" file...  
[2024-08-28 13:29:16] INFO generating the MTA archive...  
[2024-08-28 13:29:17] INFO the MTA archive generated at: gen/archive.mtar  
[2024-08-28 13:29:17] INFO cleaning temporary files...  
user: btp-cap-rag-ai-workshop $
```

## Log in to Cloud Foundry

Before deploying the application to the **Cloud Foundry Runtime**, you need to log in to the target system. You will need the target system **CF API Endpoint**. This information can be found on the **SAP BTP Cockpit Overview** page. Click on the **Overview** menu to navigate to it.



In the Cloud Foundry Environment tab, copy the **API Endpoint** value (<https://api.cf.ap10.hana.ondemand.com>). The value will vary based on the subaccount region.

ai-workshop-apj

## Subaccount: ai-workshop-apj - Overview

Development Subaccount

/ ai-workshop-apj

General

Cloud Foundry Environment

Entitlements

### Cloud Foundry Environment

API Endpoint: <https://api.cf.ap10.hana.ondemand.com>

Org Name: ai-workshop-apj

Org ID: 491bfb4e-3cd7-475d-9427-40d2c0cd65c8

Org Memory Limit: 3,072MB

[Manage environment instance](#)

Disable Cloud Foundry

Spaces (1)

Create Space

Name	Applications	Service Instances
dev	2	8

Now, go back to **BAS** to set the **CF API Endpoint** to be used for deployment. In the terminal, execute the command **\$ cf api YOUR\_API\_ENDPOINT\_GOES\_HERE**

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL SQL CONSOLE

```
user: btp-cap-rag-ai-workshop $ cf api https://api.cf.ap10.hana.ondemand.com/
Setting API endpoint to https://api.cf.ap10.hana.ondemand.com/...
OK

API endpoint: https://api.cf.ap10.hana.ondemand.com/
API version: 3.167.0

Not logged in. Use 'cf login' or 'cf login --sso' to log in.
user: btp-cap-rag-ai-workshop $
```

After setting the API Endpoint you can log into the SAP BTP Subaccount Cloud Foundry runtime. In the terminal, execute the command **\$ cf login --sso**

This will show you a link to the **Login page**. Click on the link to open it in a new tab.

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL SQL CONSOLE

```
user: btp-cap-rag-ai-workshop $ cf login --sso
API endpoint: https://api.cf.ap10.hana.ondemand.com/

Temporary Authentication Code ( Get one at https://login.cf.ap10.hana.ondemand.com/passcode ): 
```

The **Login page** will be displayed. Log into the **SAP BTP Subaccount** with your credentials.

## SAP Business Technology Platform

Sign in to continue



mauricio.lauffer@sap.com (origin: sap.ids)




[Sign in to another account](#)

Copy the **Temporary Authentication Code**.

### SAP Business Technology Platform

Temporary Authentication Code

\_h6XtMcLQReynveMCjaYIHHRTZEaPiX8 

Go back to **BAS**. In the terminal, paste or type the **Temporary Authentication** Code and press Enter. A list of CF ORGs will be listed, select the one used for the workshop, type its number and press **ENTER**.

```
○ user: btp-cap-rag-ai-workshop $ cf login --sso
API endpoint: https://api.cf.ap10.hana.ondemand.com/

Temporary Authentication Code ( Get one at https://login.cf.ap10.hana.ondemand.com/passcode ):
Authenticating...
OK

Select an org:
1. academy
2. ai-workshop-apj
3. api-workshop-apj
4. apj-as-team
5. Arrow Energy Pty Ltd_arrow-task-center-dev-2uu2nf58
6. aws11
7. btpcsp
8. fretools-test
9. P&T CEE
10. sandbox
11. SAP CP APJ CustomerSuccessTeam_cspindex
12. SAP CP APJ CustomerSuccessTeam_lfx-dev-tech-academy-x1-ibsoby8d
13. SAP CP APJ CustomerSuccessTeam_recap2024-lqt9ojqs

Org (enter to skip): 2
```

```
Org (enter to skip): 2
Targeted org ai-workshop-apj.
```

```
Targeted space dev.
```

```
API endpoint: https://api.cf.ap10.hana.ondemand.com
API version: 3.167.0
user: mauricio.lauffer@sap.com
org: ai-workshop-apj
space: dev
```

```
○ user: btp-cap-rag-ai-workshop $
```

## Deploy

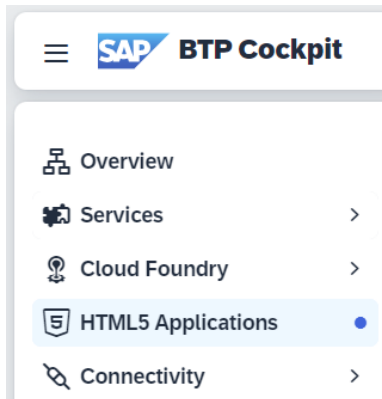
Now, execute the command **\$ npm run deploy**

```
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL SQL CONSOLE
bash + v [ ] [ ] ... ^ x

user: btp-cap-rag-ai-workshop $ npm run deploy
Uploading application "btp-cap-rag-ai-srv"...
Started async upload of application "btp-cap-rag-ai-srv"
Stopping application "btp-cap-rag-ai-srv"...
Staging application "btp-cap-rag-ai-srv"...
Application "btp-cap-rag-ai-db-deployer" staged
Executing task "deploy" on application "btp-cap-rag-ai-db-deployer"...
Application "btp-cap-rag-ai-srv" staged
Starting application "btp-cap-rag-ai-srv"...
Application "btp-cap-rag-ai-srv" started and available at "ai-workshop-apj-dev-btp-cap-rag-ai-srv.cfapps.ap10.hana.ondemand.com"
Process finished.
Use "cf dml -i cfb12fbf-6541-11ef-869f-eeee0a96d929" to download the logs of the process.
○ user: btp-cap-rag-ai-workshop $
```

# Use the Embeddings App

Go to **HTML5 Applications** and click on the **btpcapragaiembedding** link.



Upload PDF files to convert them into vectors and store them in HANA Cloud. This is the embedding process.

## Subaccount: ai-workshop-apj - HTML5 Applications


Add Application

All: 4

Managed Application Router provided by SAP Build Work Zone, standard edition			
Application Name	Active Version	Business Solution	Actions
btpcapragaiembed	0.0.1	btpcapragai.service	<a href="#">↓</a> <a href="#">🔗</a>
btpcapragaiembedding	0.0.1	btpcapragai.service	<a href="#">↓</a> <a href="#">🔗</a>

Embedding

Upload



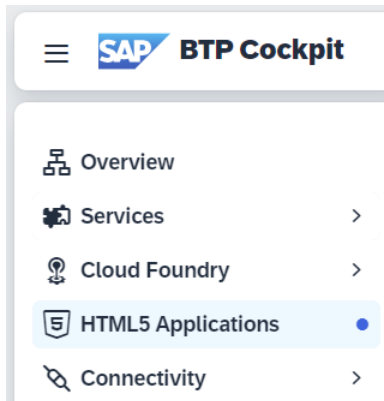
**No files found.**  
Drop files to upload, or use the "Upload" button.

Upload

You could use the sample PDF available in the GitHub repository. Download the file **TravelExpensesPolicy.pdf** (<https://github.com/mauriciolauffer/btp-cap-rag-ai-workshop/blob/workshop/docs/TravelExpensesPolicy.pdf>) and upload it to the **Embeddings App**.

# Use the Chat App

Go to **HTML5 Applications** and click on the **btpcapragaichat** link.



Have a chat with the chatbot. Ask questions regarding the PDF files you have previously uploaded.

Also, ask questions about PDF files you have not embedded yet. The chatbot should say it does not know the answer rather than hallucinate.

## Subaccount: ai-workshop-apj - HTML5 Applications

Add Application ↺

All: 4

All Business Solutions ▾

Managed Application Router provided by SAP Build Work Zone, standard edition ⚙️

Application Name	Active Version	Business Solution	Actions
<a href="#">btpcapragaichat</a>	0.0.1	btpcapragai.service	<a href="#">↓</a> <a href="#">🔗</a>
<a href="#">btpcapragaiembedding</a>	0.0.1	btpcapragai.service	<a href="#">↓</a> <a href="#">🔗</a>



hi  
user · 2024-08-29T15:05:48.764Z



Hello! How can I assist you today?  
system · 2024-08-29T15:05:50.413Z



when is my next trip to japan?  
user · 2024-08-29T15:05:57.627Z



Your next trip to Japan is on February 13, 2025. You will depart from Cairns at 11:20 AM and arrive at Tokyo Narita at 5:45 PM.  
system · 2024-08-29T15:06:02.798Z

Type a message....



If you are using the **TravelExpensesPolicy.pdf** sample, try the suggested prompts:

- What are the transport options?
- Who can travel business class?
- Tell me more about VISA requirements
- Explain the rest periods
- Can I get reimbursement for on-board internet expenses?
- Can I travel by train?