Edge AI Tuning Kit

Get Started Guide

June 2025

Open Source

Document Number: 822132-1.1

Contents

Figures

Tables

Revision History

Date	Revision	Description
May 2025	2.0	Open-Source Release
November 2024	1.1	Added Release Notes for PV1.0 Release
October 2024	1.0	PV release.
April 2024	0.5	Initial release.

§

1.0 Introduction

Edge AI Tuning Kit helps users to train, optimize, and deploy custom Large Language Model (LLM) for horizontal use cases.

For release information and notes, refer to 4.0 Release Notes.

1.1 Customer Support

For technical support, please follow these steps:

1. Go to Edge AI Tuning Kit repository.

2. Search **Existing Issues**:
Before creating a new issue, check the Issues tab to see if your concern has already been addressed.

3. Submit a **New Issue**:
If your issue hasn't been reported, click on New Issue.

4. Labeling
Assign appropriate labels to categorize the issue (for example, bug, enhancement, question). This helps in prioritizing and addressing issues efficiently.

5. Follow Up
Monitor the issue for any responses or requests for additional information. Engage in the discussion to facilitate resolution.

1.2 Terminology

Table 1. Terminology

Term	Description
UI	User Interface
LLM	Large Language Model
RAG	Retrieval-Augmented Generation
API	Application Programming Interface

2.0 Prerequisites

2.1 Experience Required

- Basic Ubuntu* OS knowledge

2.2 Software

- Edge AI Tuning Kit
- Ubuntu 22.04* LTS or Ubuntu 24.04* LTS
- Docker* computer software

2.3 Hardware

Table 2. Hardware Requirements

A Linux* machine that meets the following hardware requirements:

Hardware requirements	Minimum	Recommended
CPU	13th Gen Intel® Core™ CPU and above	4th Gen Intel® Xeon® Scalable Processor and above
GPU	Intel® Arc™ A770 Graphics (16GB)	Multiple Intel® Arc™ A770 Graphics (16GB)
RAM (GB)	64 and above	128 and above
Disk (GB)	500 (Around 4 projects with 1 training task each)	1000 (Around 8 projects with 1 training task each)

§

3.0 Step-by-step Instructions

This guide helps users to start using the Edge AI Tuning Kit. Through this guide, users will learn how to:

1. Install the application.
2. Navigate the User Interface (UI):
   a. Download Models
   b. Create a New Project
   c. Upload Documents
   d. Upload Datasets
   e. Train Models
   f. Create Model Deployment Services
3. Uninstall the application.

3.1 Step 1: Software Installation

In this step, users will learn to install the application.

3.1.1 Initial Setup

1. Create a Hugging Face account and generate an access token. For more information, please refer to Hugging Face Token link.
2. Log in to your Hugging Face account and browse to mistralai/Mistral-7B-Instruct-v0.3 and click on the Agree and access repository button.
3. Set up GPU driver based on your GPU version
   - Intel® Arc™ A-Series Graphics
   - Intel® Data Center GPU Flex Series
4. Install Docker* Computer Software.
5. Set permissions for Docker* computer software group:
   Run the following command to add your current user to the Docker group. After running the command, log out and log back in for the changes to take effect.

```
sudo usermod -aG docker $USER
```

6. Clone the repository:

```
git clone https://github.com/open-edge-platform/edge-ai-tuning-kit.git
```

7. Set up the application:

```
cd edge-ai-tuning-kit/
./setup.sh -b
```

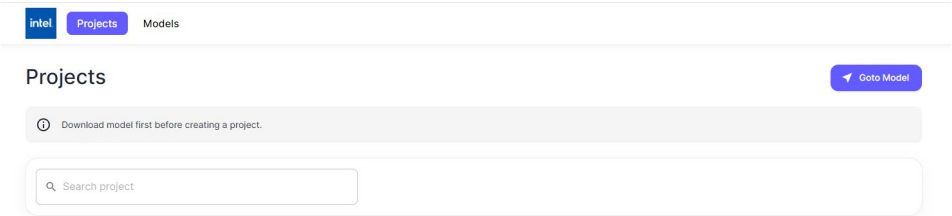
3.1.2 Start Application

After setup, run the application by using the command below:

```
./setup.sh -r
```

Users can access the web UI using http://localhost once the installation is successful.

Figure 1. UI with Successful Installation



3.2 Step 2: Using the User Interface

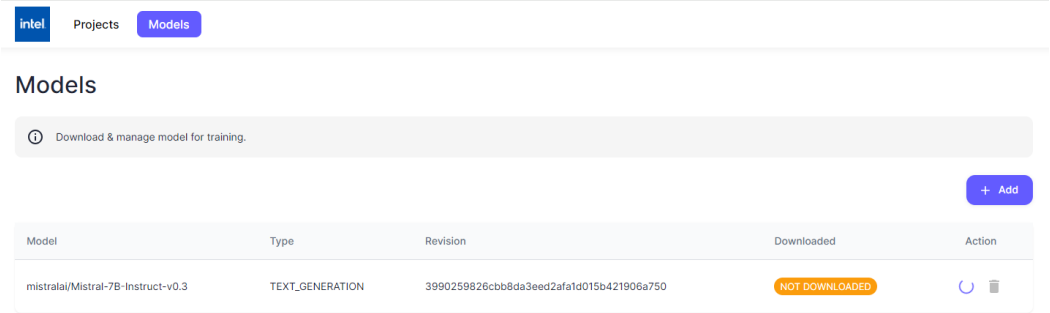
In this step, users will learn to navigate through the User Interface.

3.2.1 Download Models

8. Before creating a new Project, you need to ensure that there are available LLM Models.

9. Go to the Models tab to download and manage the model for training.

Figure 2. Models Tab

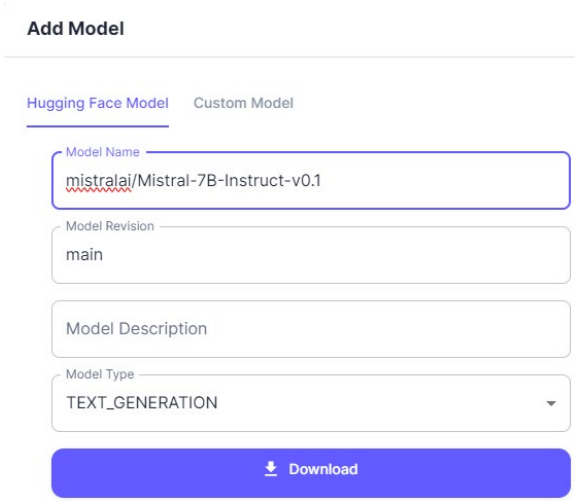


10. You will see the default Model: ***mistralai/Mistral-7B-Instruct-v0.3*** with **Not Downloaded** status. Click the **Download** button to start downloading the model. Once completed, the status will change to Downloaded.

11. If you want to use a different model, click **Add**.

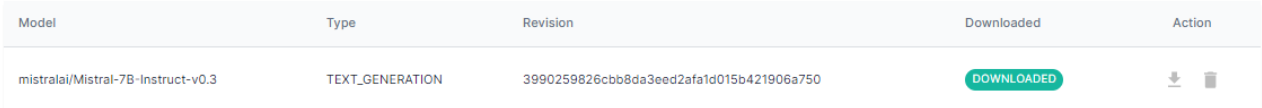
12. Insert the Model Name following the Hugging Face models. Example: ***mistralai/Mistral-7B-Instruct-v0.1***

Figure 3. Add Model



13. Once downloaded, your model will be listed in the table.

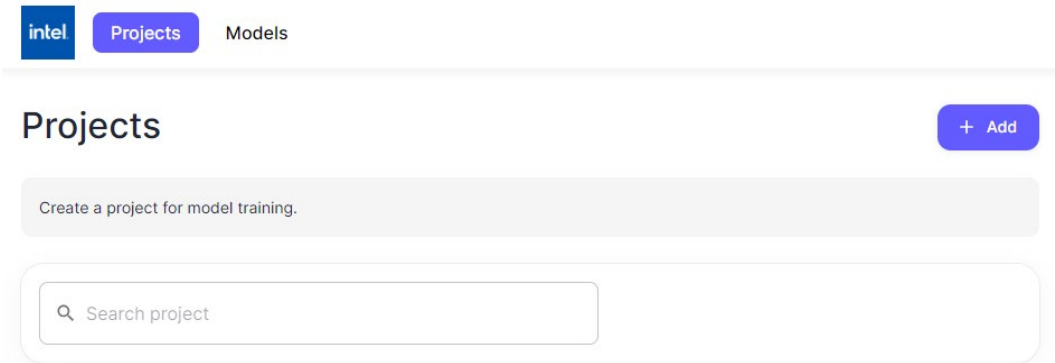
Figure 4. Model Downloaded



3.2.2 Create New Project

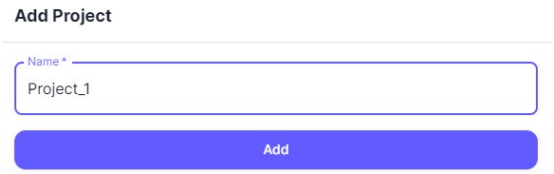
1. Under the **Projects** tab, click **Add** on the right side of the UI.

Figure 5. New Project



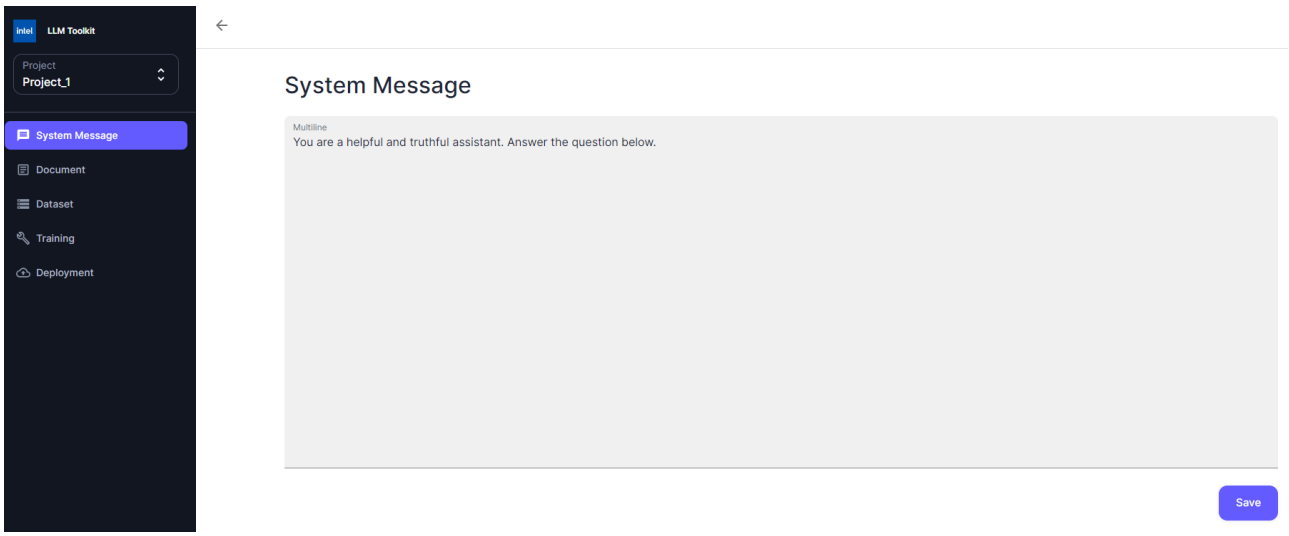
2. Name your project, and click **Add**.

Figure 6. Insert Project Name



3. Click your project in the **Projects** list, and you can do further modifications.
4. You can also change the **System Message** for their LLM Chatbot.

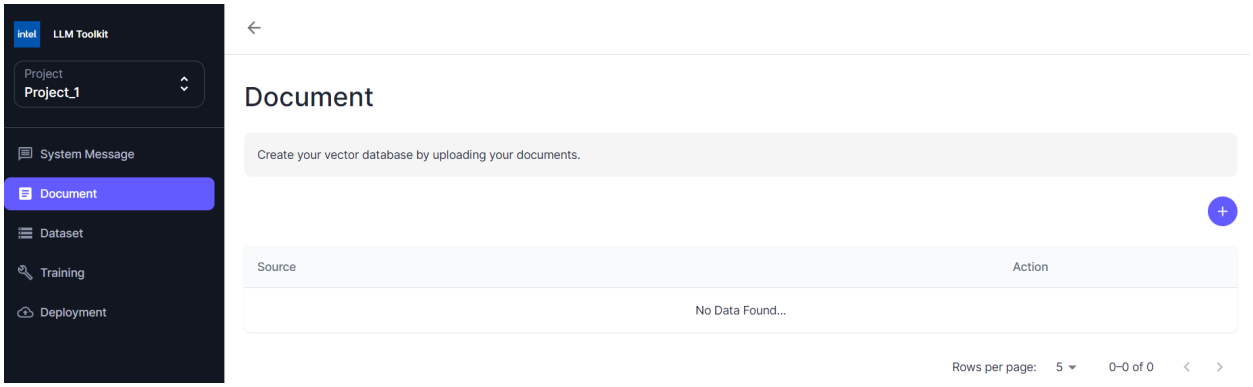
Figure 7. Project Listed



3.2.3 Upload Document

1. You can upload the document that will be used for the Retrieval-Augmented Generation (RAG) feature during deployment for this project. Choose the **Document** tab and click **Add**.

Figure 8. Document Tab



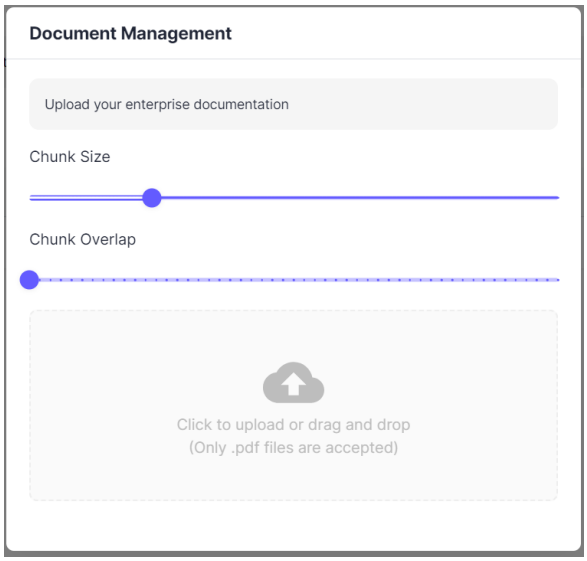
2. You can choose the **Chunk Size** and **Chunk Overlap**.

   a. **Chunk Size:** Controls the maximum size in terms of the number of characters of the final documents.

   b. **Chunk Overlap:** Specifies how much overlap there should be between chunks.

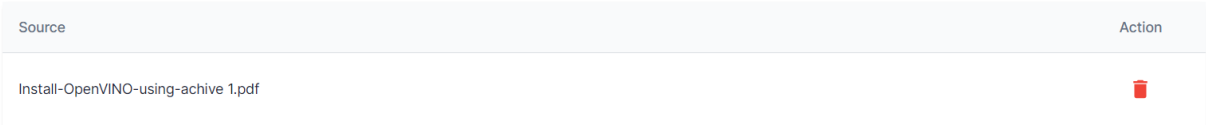
3. You can upload the document by choosing **Click** to Upload or drag and drop the files.

Figure 9. Document Management



4. Once done, the document uploaded will be listed on the table.

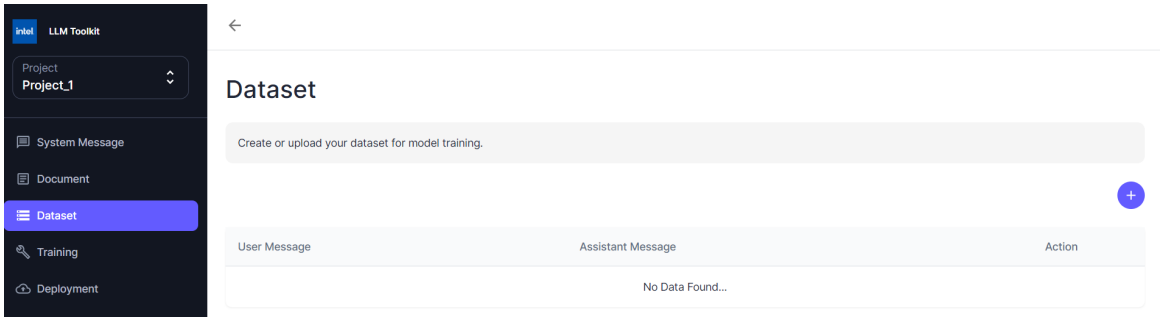
Figure 10. Document Listed



3.2.4 Upload Dataset

1. Click the **Dataset** tab to generate your dataset.

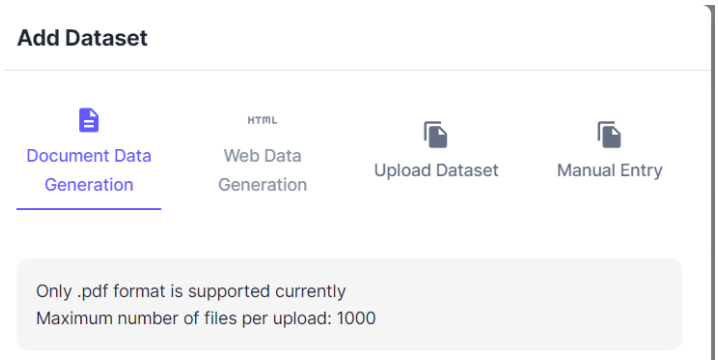
Figure 11. Dataset Generation



2. You can upload their dataset through three options for now.

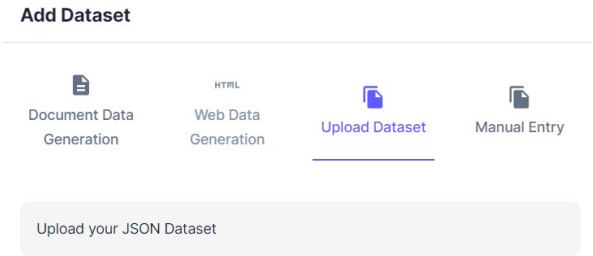
   a. **Document Data Generation**

Figure 12. Document Data Generation



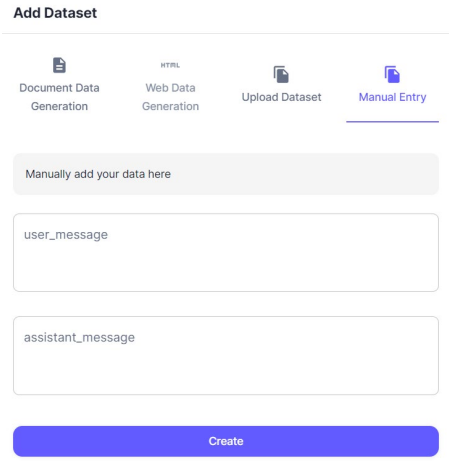
   b. **Upload JSON Dataset**

Figure 13. Upload Dataset



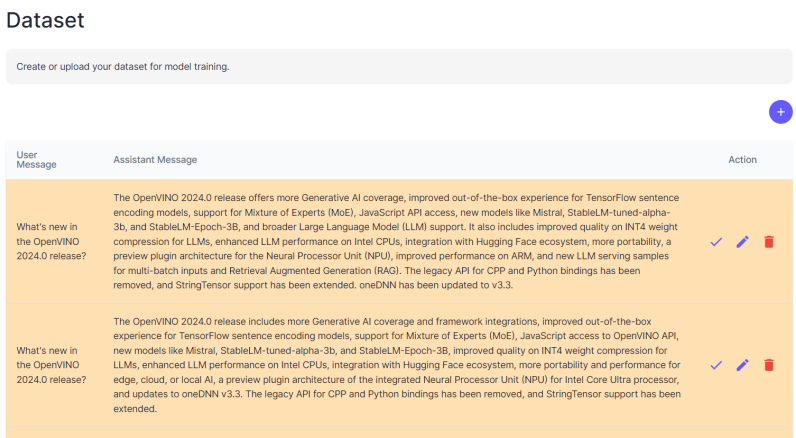
   c. **Manual Entry Dataset**

Figure 14. Manual Entry



3. The example below will follow the **Document Data Generation**. Once uploaded, your dataset will be listed in the table.

Figure 15. Dataset Generated Table



4. You can **Confirm, Edit**, or **Delete** the dataset through the three icons on the right side of the table.

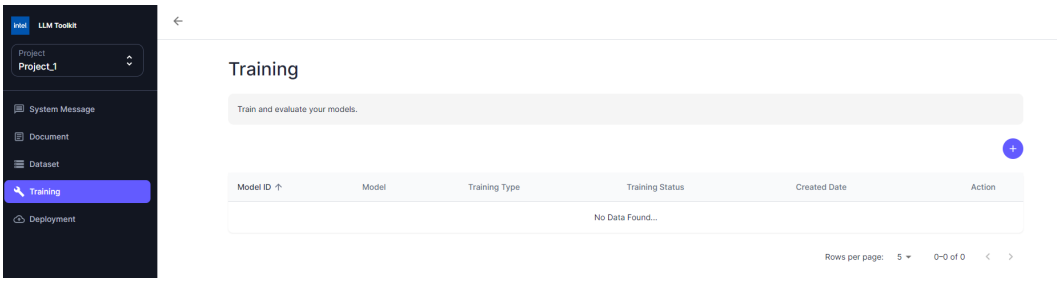
5. Once confirmed, the orange-colored dataset will turn to white.

Figure 16. Validated Dataset

3.2.5 Model Training

1. Once the dataset is set, go to the **Training** tab and click **Add**.

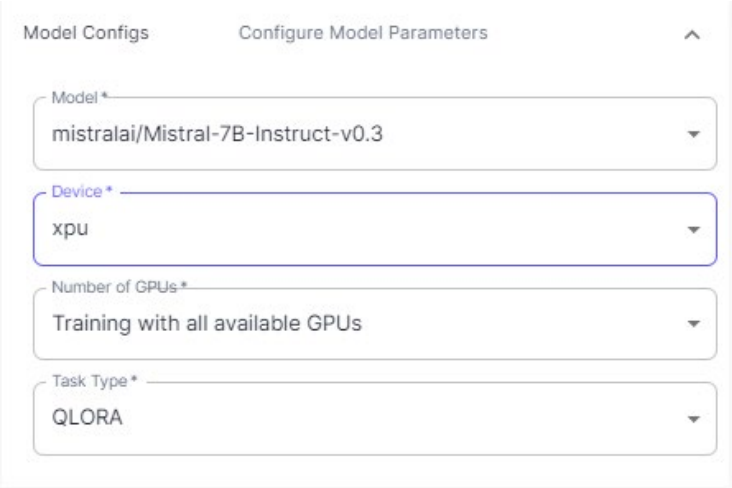
Figure 17. Training Tab



2. You can set these configurations before starting the training:

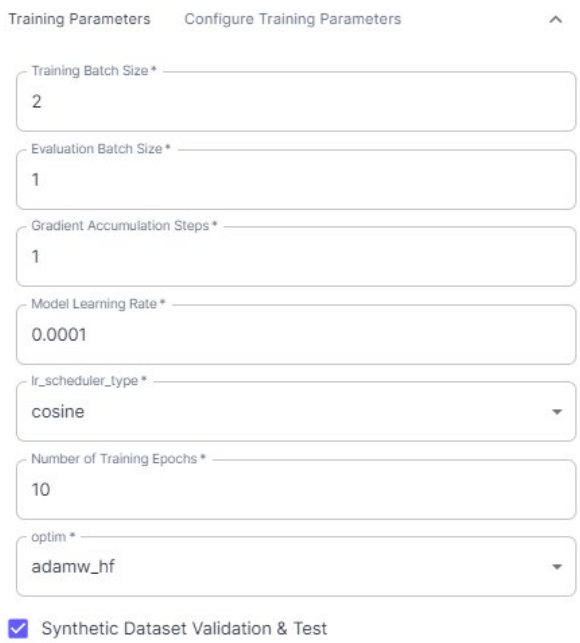
   i. **Model Configurations**

Figure 18. Model Configurations Example



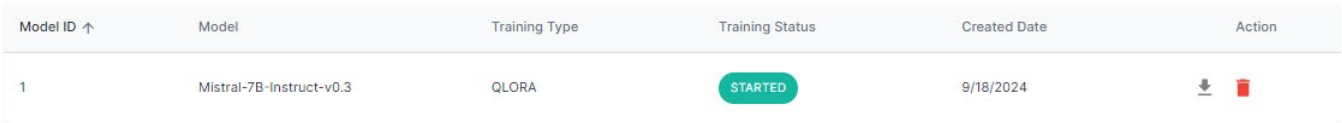
ii. Training Parameters

Figure 19. Training Parameters Configurations Example



3. Once done, you can click **Train**.

Figure 20. Training List

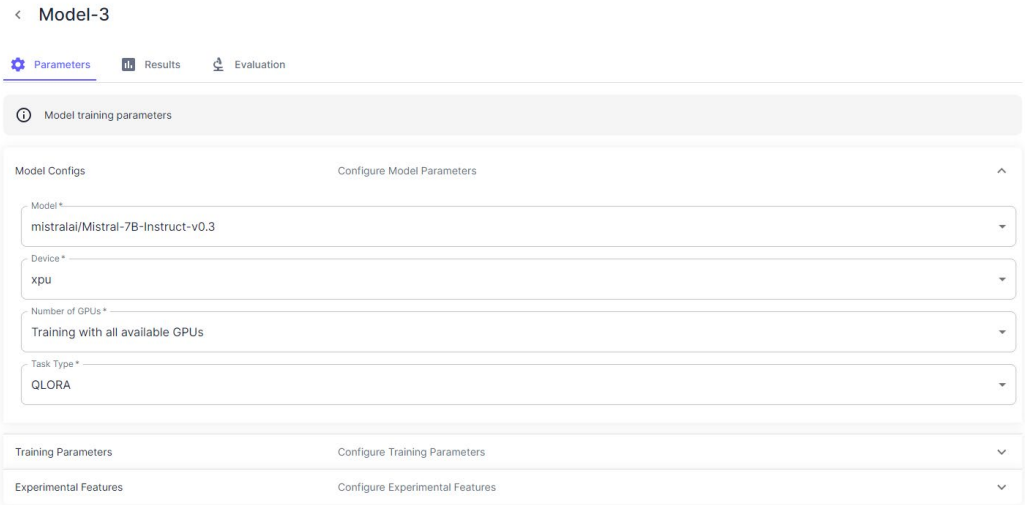


Model ID	Model	Training Type	Training Status	Created Date	Action
1	Mistral-7B-Instruct-v0.3	QLORA	STARTED	9/18/2024	

4. You can get more information on the trained models by clicking on the **Model** listed.

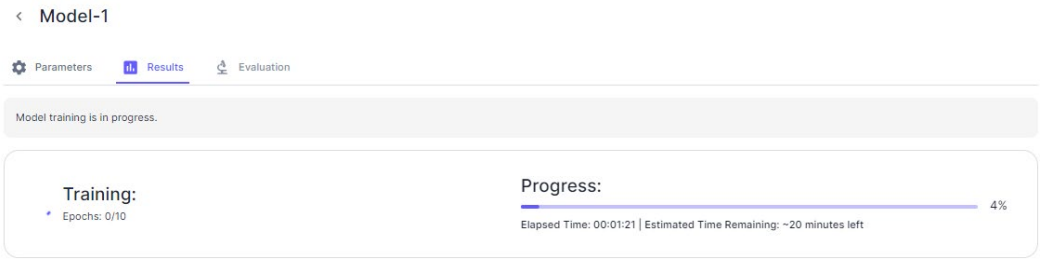
5. You can observe the parameters set for the trained models under the **Parameters** tab.

Figure 21. Parameters Set



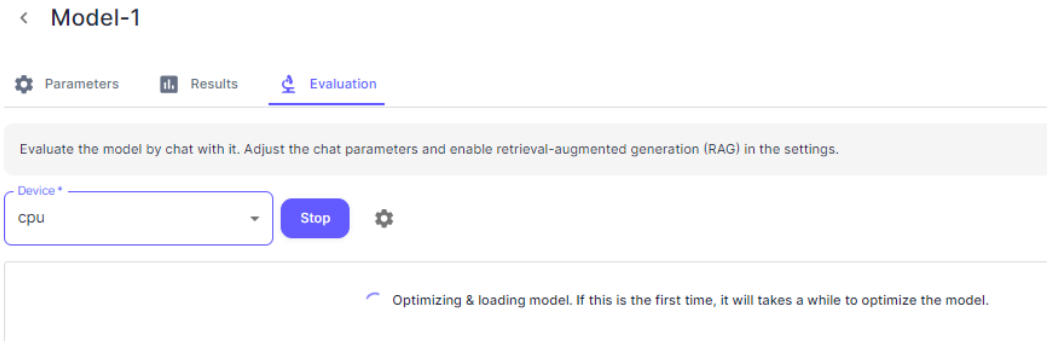
6. Under **Results,** you can see the training status.

Figure 22. Training Status



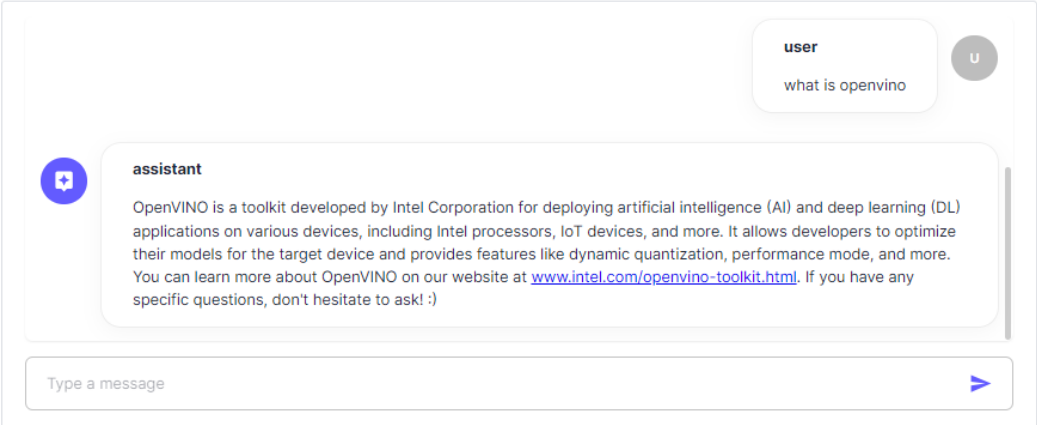
7. Once the progress is completed, you can test the trained model in the **Evaluation** tab. It will take some time for the chat to load if it is your first time using the **Evaluation.**

Figure 23. First Initialization of Evaluation Chat



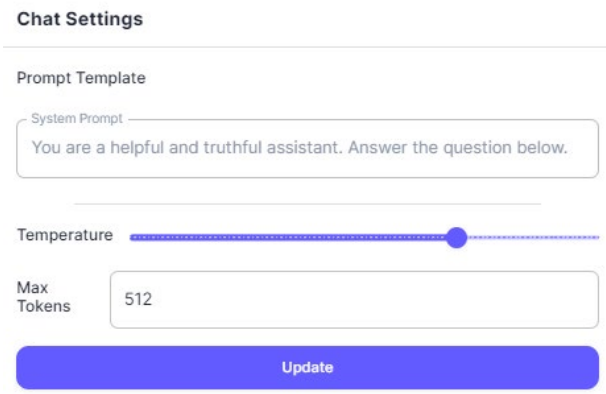
8. You can ask questions following the context provided by the dataset.

Figure 24. Chat Example



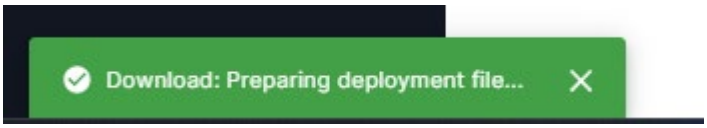
9. You can configure the chat settings further by clicking the **Settings** icon.

Figure 25. Chat Settings



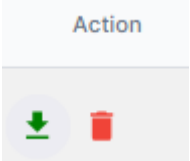
10. If you plan to use the trained model in their deployment setup, you can initiate the application to prepare your deployment files by clicking the **Download** button (next to delete training task button). If this is your first time, the **Download** button is grayed. Once clicked, you will be notified that the application is prepping your deployment file.

Figure 26. Notification to User



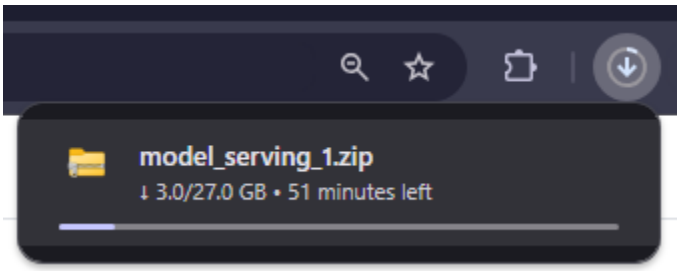
11. Once the files are ready to be downloaded, the **Download** button will turn green.

Figure 27. Ready to Download



12. Click the green **Download** button to initiate the download process.

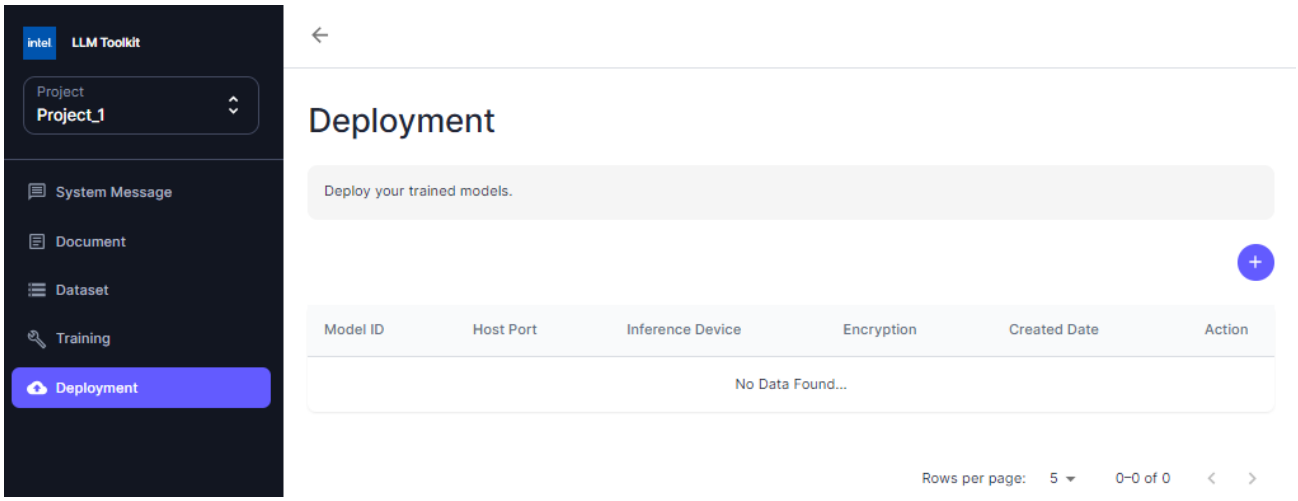
Figure 28. Downloading Files



3.2.6 Deployment Services

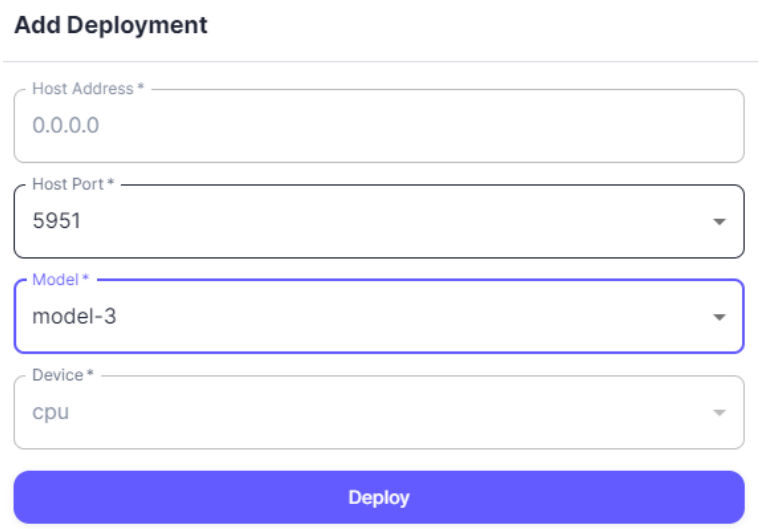
1. At this moment, the **Deployment** tab only serves the services.

Figure 29. Deployment Tab



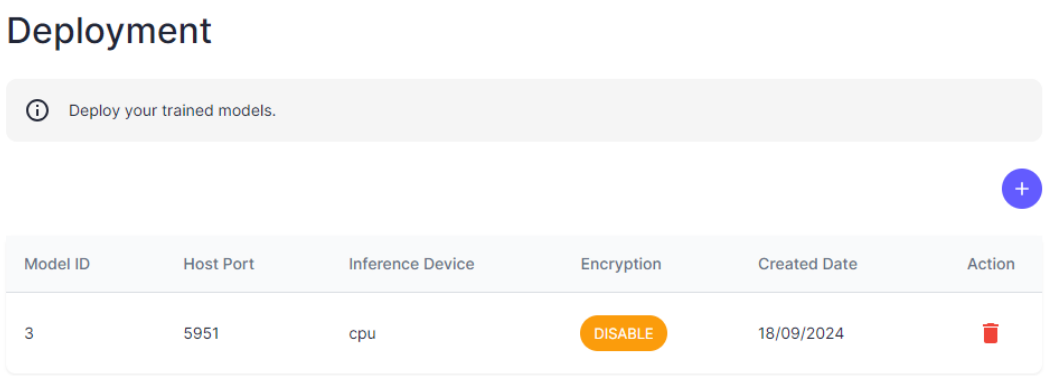
2. In the **Deployment** tab, click **Add Deployment**. Insert the Host Port, and select the available Model. Click **Deploy**.

Figure 30. Add Deployment



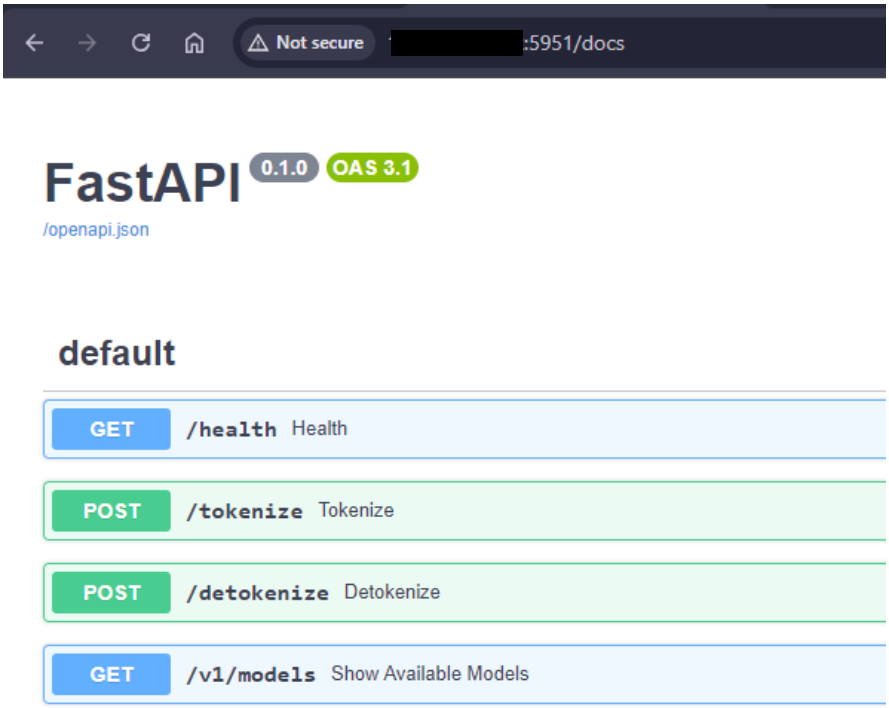
3. Your deployment service will be listed.

Figure 31. Deployment Table



4. To access the Application Programming Interface (API), go to http://localhost:[port]/docs or http://[System_IP]:[port]/docs

Figure 32. API



§

3.3 Step 3: Uninstall the Application

In this step, the user will learn to uninstall the application.

3.3.1 Stop the Application

Stop the application by using the following command:

```
./setup.sh -s
```

3.3.2 Remove All the Data Files

Remove all the database and application cache files by using the following commands:

```
# Remove the database cache file
docker volume rm edge-ai-tuning-kit-data-cache
docker volume rm edge-ai-tuning-kit-database
docker volume rm edge-ai-tuning-kit-task-cache
```

4.0 Release Notes

Updated on: May 2025

Version: 2025.1 (Open-Source Release)

4.1 New Features and Enhancements

No new features for this Open-Source Release.

4.2 Bug Fixes

No bug fix for this Open-Source Release.

4.3 Documentation

No new documentation for this Open-Source Release.

§