



Edge AI Tuning Kit

Get Started Guide

June 2025

Open Source



You may not use or facilitate the use of this document in connection with any infringement or other legal analysis concerning Intel products described herein. You agree to grant Intel a non-exclusive, royalty-free license to any patent claim thereafter drafted which includes subject matter disclosed herein.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest Intel product specifications and roadmaps.

The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Copies of documents which have an order number and are referenced in this document may be obtained by calling 1-800-548-4725 or visiting the [Intel Resource and Documentation Center](#).

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No product or component can be absolutely secure. Check with your system manufacturer or retailer or learn more at [intel.com](https://www.intel.com).

No product or component can be absolutely secure.

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.

Contents

1.0	Introduction	6
1.1	Customer Support.....	6
1.2	Terminology	6
2.0	Prerequisites.....	7
2.1	Experience Required	7
2.2	Software.....	7
2.3	Hardware	7
3.0	Step-by-step Instructions.....	8
3.1	Step 1: Software Installation	8
3.1.1	Initial Setup	8
3.1.2	Start Application.....	9
3.2	Step 2: Using the User Interface	10
3.2.1	Download Models	10
3.2.2	Create New Project.....	11
3.2.3	Upload Document	13
3.2.4	Upload Dataset.....	14
3.2.5	Model Training.....	17
3.2.6	Deployment Services.....	22
3.3	Step 3: Uninstall the Application.....	24
3.3.1	Stop the Application.....	24
3.3.2	Remove All the Data Files.....	24
4.0	Release Notes	25
4.1	New Features and Enhancements.....	25
4.2	Bug Fixes.....	25
4.3	Documentation.....	25

Figures

Figure 1.	UI with Successful Installation	9
Figure 2.	Models Tab.....	10
Figure 3.	Add Model	11
Figure 4.	Model Downloaded	11
Figure 5.	New Project.....	11
Figure 6.	Insert Project Name	12
Figure 7.	Project Listed	12
Figure 8.	Document Tab	13
Figure 9.	Document Management	14
Figure 10.	Document Listed	14

Figure 11.	Dataset Generation	14
Figure 12.	Document Data Generation	15
Figure 13.	Upload Dataset	15
Figure 14.	Manual Entry	15
Figure 15.	Dataset Generated Table.....	16
Figure 16.	Validated Dataset	16
Figure 17.	Training Tab.....	17
Figure 18.	Model Configurations Example	17
Figure 19.	Training Parameters Configurations Example	18
Figure 20.	Training List	18
Figure 21.	Parameters Set.....	19
Figure 22.	Training Status	19
Figure 23.	First Initialization of Evaluation Chat.....	19
Figure 24.	Chat Example	20
Figure 25.	Chat Settings.....	20
Figure 26.	Notification to User	20
Figure 27.	Ready to Download	21
Figure 28.	Downloading Files	21
Figure 29.	Deployment Tab	22
Figure 30.	Add Deployment.....	22
Figure 31.	Deployment Table	23
Figure 32.	API.....	23

Tables

Table 1.	Terminology	6
Table 2.	Hardware Requirements	7

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



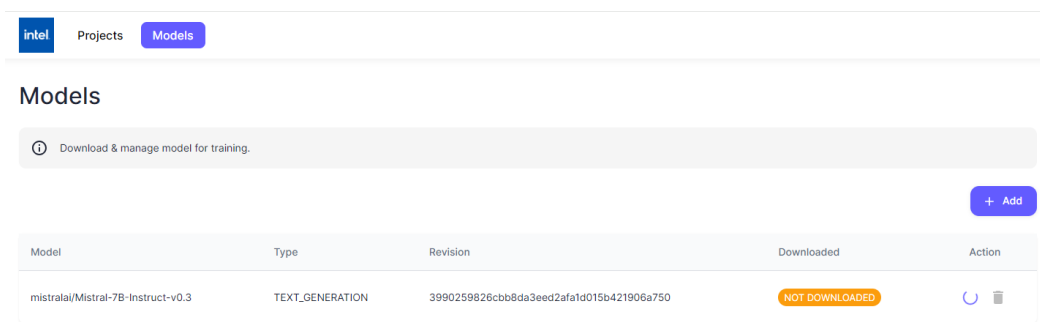
3.2 Step 2: Using the User Interface

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

Add Model

Hugging Face Model

Custom Model

Model Name

mistralai/Mistral-7B-Instruct-v0.1

Model Revision

main

Model Description

Model Type

TEXT_GENERATION

Download

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

Figure 4. Model Downloaded

Model	Type	Revision	Downloaded	Action
mistralai/Mistral-7B-Instruct-v0.3	TEXT_GENERATION	3990259826cbb8da3eed2afa1d015b421906a750	DOWNLOADED	

3.2.2 Create New Project

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

Figure 5. New Project

intel

Projects

Models

Projects

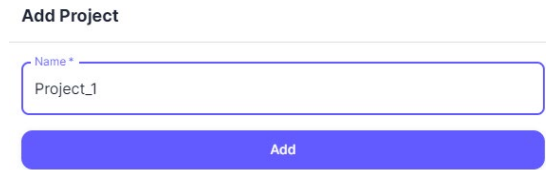
+ Add

Create a project for model training.

Search project

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

Figure 6. Insert Project Name



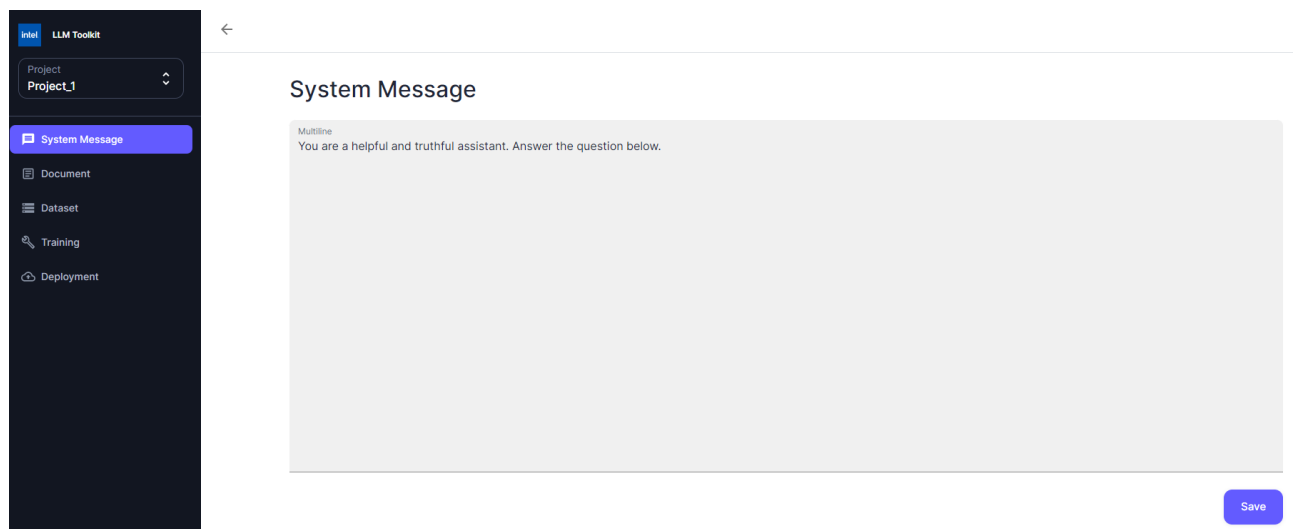
Add Project

Name *
Project_1

Add

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



intel LLM Toolkit

Project
Project_1

System Message

Document

Dataset

Training

Deployment

System Message

Multiline

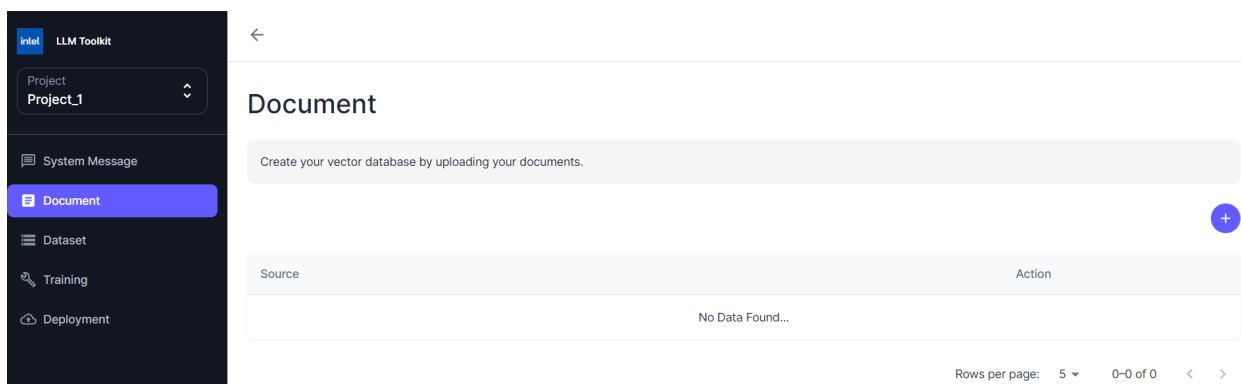
You are a helpful and truthful assistant. Answer the question below.

Save

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

Document Management

Upload your enterprise documentation

Chunk Size

Chunk Overlap

Click to upload or drag and drop

(Only .pdf files are accepted)

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

Figure 10. Document Listed

Source	Action
Install-OpenVINO-using-achive 1.pdf	

3.2.4 Upload Dataset

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

Figure 11. Dataset Generation

Intel LLM Toolkit

Project

Project_1

System Message

Document

Dataset

Training

Deployment

Dataset

Create or upload your dataset for model training.

User Message

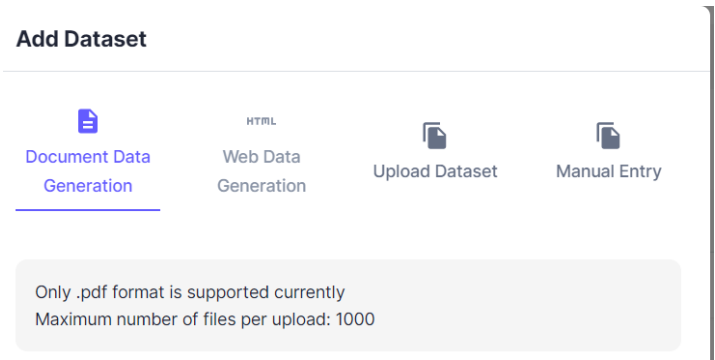
Assistant Message

Action

No Data Found...

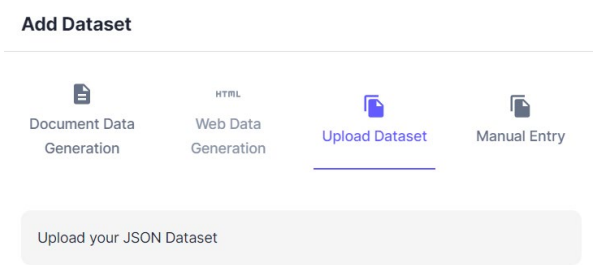
2. You can upload their dataset through three options for now.
- a. Document Data Generation

Figure 12. Document Data Generation



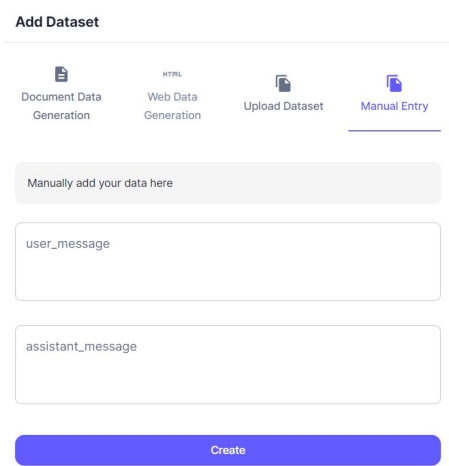
- b. Upload JSON Dataset

Figure 13. Upload Dataset



- c. Manual Entry Dataset

Figure 14. Manual Entry



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

Figure 15. Dataset Generated Table

Dataset

Create or upload your dataset for model training.

User Message	Assistant Message	Action
What's new in the OpenVINO 2024.0 release?	The OpenVINO 2024.0 release offers more Generative AI coverage, improved out-of-the-box experience for TensorFlow sentence encoding models, support for Mixture of Experts (MoE), JavaScript API access, new models like Mistral, StableLM-tuned-alpha-3b, and StableLM-Epoche-3B, and broader Large Language Model (LLM) support. It also includes improved quality on INT4 weight compression for LLMs, enhanced LLM performance on Intel CPUs, integration with Hugging Face ecosystem, more portability, a preview plugin architecture for the Neural Processor Unit (NPU), improved performance on ARM, and new LLM serving samples for multi-batch inputs and Retrieval Augmented Generation (RAG). The legacy API for CPP and Python bindings has been removed, and StringTensor support has been extended. oneDNN has been updated to v3.3.	✓ ✎ 🗑
What's new in the OpenVINO 2024.0 release?	The OpenVINO 2024.0 release includes more Generative AI coverage and framework integrations, improved out-of-the-box experience for TensorFlow sentence encoding models, support for Mixture of Experts (MoE), JavaScript access to OpenVINO API, new models like Mistral, StableLM-tuned-alpha-3b, and StableLM-Epoche-3B, improved quality on INT4 weight compression for LLMs, enhanced LLM performance on Intel CPUs, integration with Hugging Face ecosystem, more portability and performance for edge, cloud, or local AI, a preview plugin architecture of the integrated Neural Processor Unit (NPU) for Intel Core Ultra processor, and updates to oneDNN v3.3. The legacy API for CPP and Python bindings has been removed, and StringTensor support has been extended.	✓ ✎ 🗑

- You can **Confirm**, **Edit**, or **Delete** the dataset through the three icons on the right side of the table.
- Once confirmed, the orange-colored dataset will turn to white.

Figure 16. Validated Dataset

Dataset

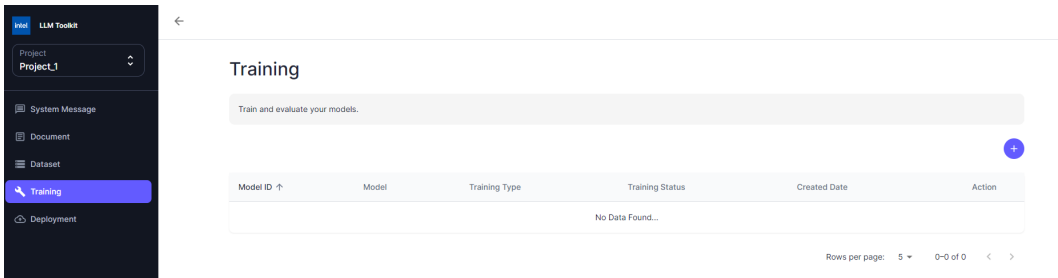
Create or upload your dataset for model training.

User Message	Assistant Message	Action
What's new in the OpenVINO 2024.0 release?	The OpenVINO 2024.0 release offers more Generative AI coverage, improved out-of-the-box experience for TensorFlow sentence encoding models, support for Mixture of Experts (MoE), JavaScript API access, new models like Mistral, StableLM-tuned-alpha-3b, and StableLM-Epoche-3B, and broader Large Language Model (LLM) support. It also includes improved quality on INT4 weight compression for LLMs, enhanced LLM performance on Intel CPUs, integration with Hugging Face ecosystem, more portability, a preview plugin architecture for the Neural Processor Unit (NPU), improved performance on ARM, and new LLM serving samples for multi-batch inputs and Retrieval Augmented Generation (RAG). The legacy API for CPP and Python bindings has been removed, and StringTensor support has been extended. oneDNN has been updated to v3.3.	✎ 🗑

3.2.5 Model Training

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

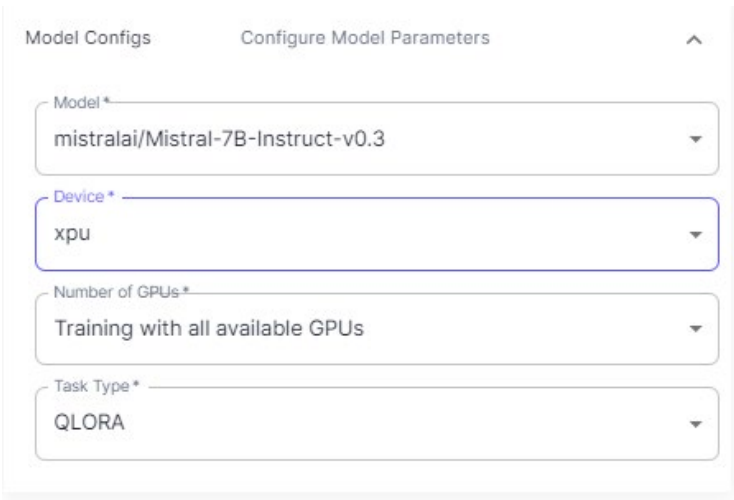
Figure 17. Training Tab



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

Training Parameters

Configure Training Parameters

^

Training Batch Size *

2

Evaluation Batch Size *

1

Gradient Accumulation Steps *

1

Model Learning Rate *

0.0001

lr_scheduler_type *

cosine

▼

Number of Training Epochs *

10

optim *



adamw_hf

▼

☒ Synthetic Dataset Validation & Test

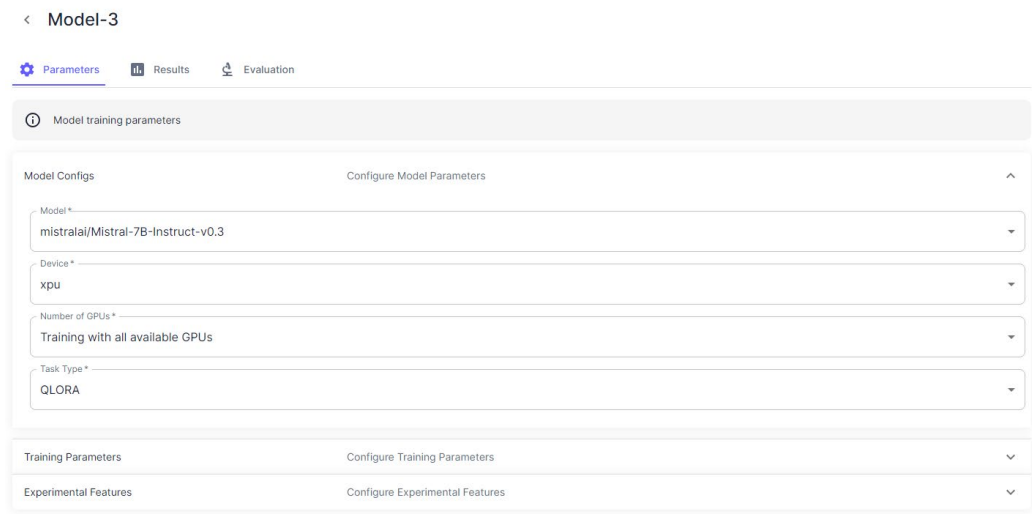
3. Once done, you can click **Train**.

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	 

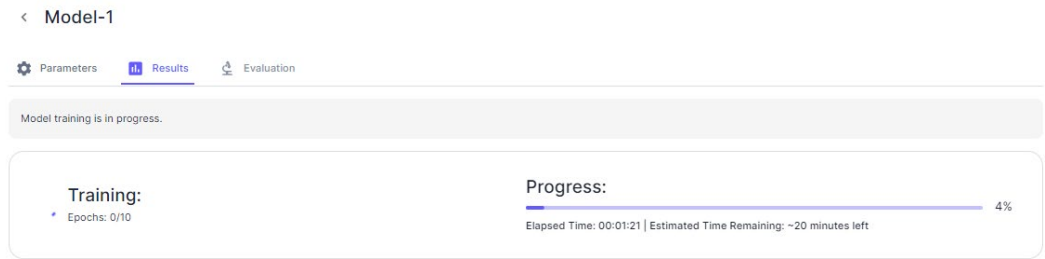
4. You can get more information on the trained models by clicking on the **Model** listed.
5. You can observe the parameters set for the trained models under the **Parameters** tab.

Figure 21. Parameters Set



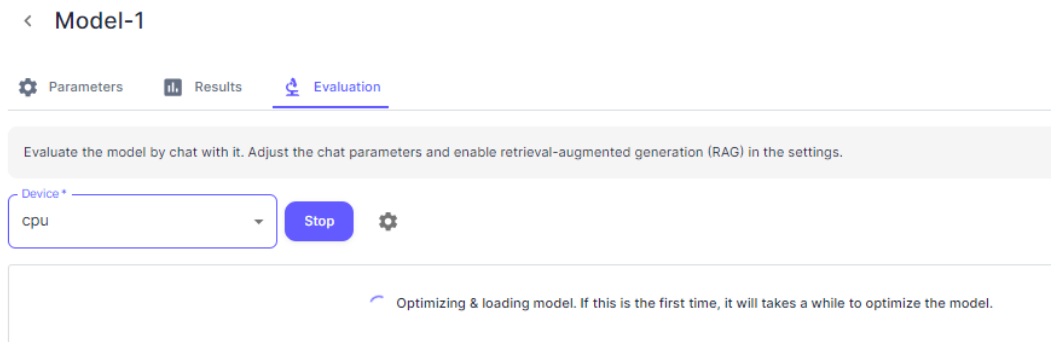
6. Under **Results**, you can see the training status.

Figure 22. Training Status



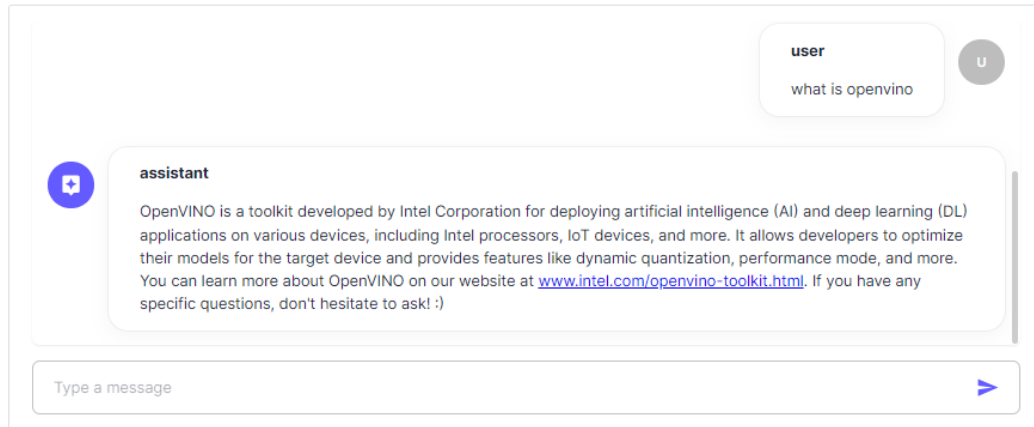
7. Once the progress is completed, you can test the trained model in the **Evaluation** tab. It will take some time for the chat to load if it is your first time using the **Evaluation**.

Figure 23. First Initialization of Evaluation Chat



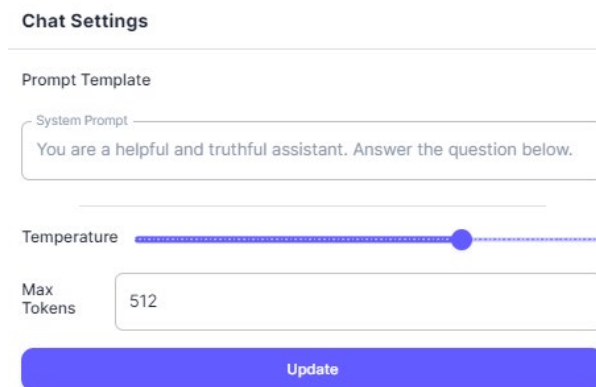
8. You can ask questions following the context provided by the dataset.

Figure 24. Chat Example



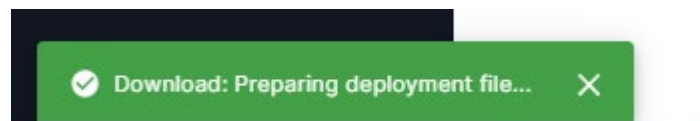
9. You can configure the chat settings further by clicking the **Settings** icon.

Figure 25. Chat Settings



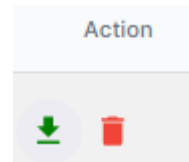
10. If you plan to use the trained model in their deployment setup, you can initiate the application to prepare your deployment files by clicking the **Download** button (next to delete training task button). If this is your first time, the **Download** button is grayed. Once clicked, you will be notified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

Figure 26. Notification to User



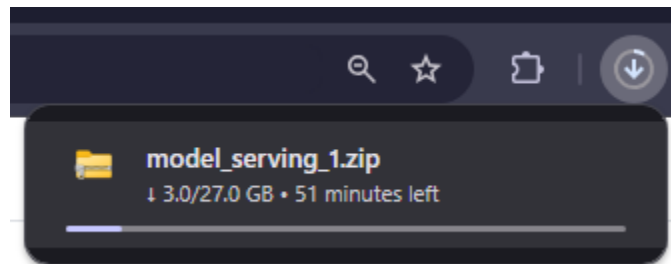
11. Once the files are ready to be downloaded, the **Download** button will turn green.

Figure 27. Ready to Download



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

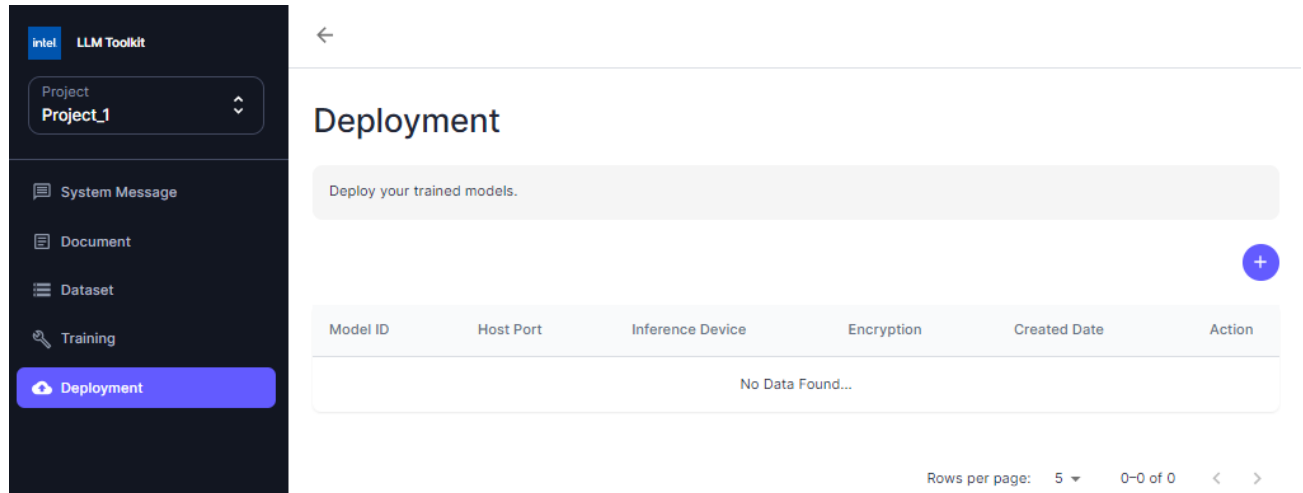
Figure 28. Downloading Files



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

Add Deployment

Host Address *

Host Port *

Model *

Device *

Deploy

3. Your deployment service will be listed.

Figure 31. Deployment Table

Deployment

Deploy your trained models.

Model ID	Host Port	Inference Device	Encryption	Created Date	Action
3	5951	cpu	DISABLE	18/09/2024	

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

Figure 32. API

← → ↻ 🏠 ⚠ Not secure [redacted]:5951/docs

FastAPI

0.1.0 OAS 3.1

/openapi.json

default

GET

/health

Health

POST

/tokenize

Tokenize

POST

/detokenize

Detokenize

GET

/v1/models

Show Available Models

§

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.

§