Legal Disclaimer

This presentation is not meant to be exhaustive and is provided AS IS, for convenience and information only and is not to be relied upon for any purpose, other than educational. The presentation is intended only to provide the general insights, opinions, and/or internally developed guidelines and procedures of Intel Corporation (Intel). The information in this presentation may need to be adapted to your specific situation or work environment.

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT.

This material may relate to the creation of end products used in safety-critical applications designed to comply with functional safety standards or requirements ("Safety-Critical Applications") or any application in which failure of the Intel product could result, directly or indirectly, in personal injury or death. It is your sole responsibility to design, manage and assure system-level safeguards to anticipate, monitor and control system failures, and you are solely responsible for all applicable regulatory standards and safety-related requirements concerning your use of any material related to Safety Critical Applications.

You further agree that some of the material maybe be pre-production in nature and that all material is provided "as is" without any express or implied warranty of any kind.

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.

Intel technologies may require enabled hardware, software or service activation.

No product or component can be absolutely secure.

The code names presented in this document are only for use by Intel to identify products, technologies, or services in development, that have not been made commercially available to the public, i.e., announced, launched or shipped. They are not "commercial" names for products or services and are not intended to function as trademarks.

Contents or Agenda

- Introduction to Edge AI Tuning Kit
- Introduction to LLM Finetuning Toolkit
- Finetuning a medical report generation LLM
- Q&A

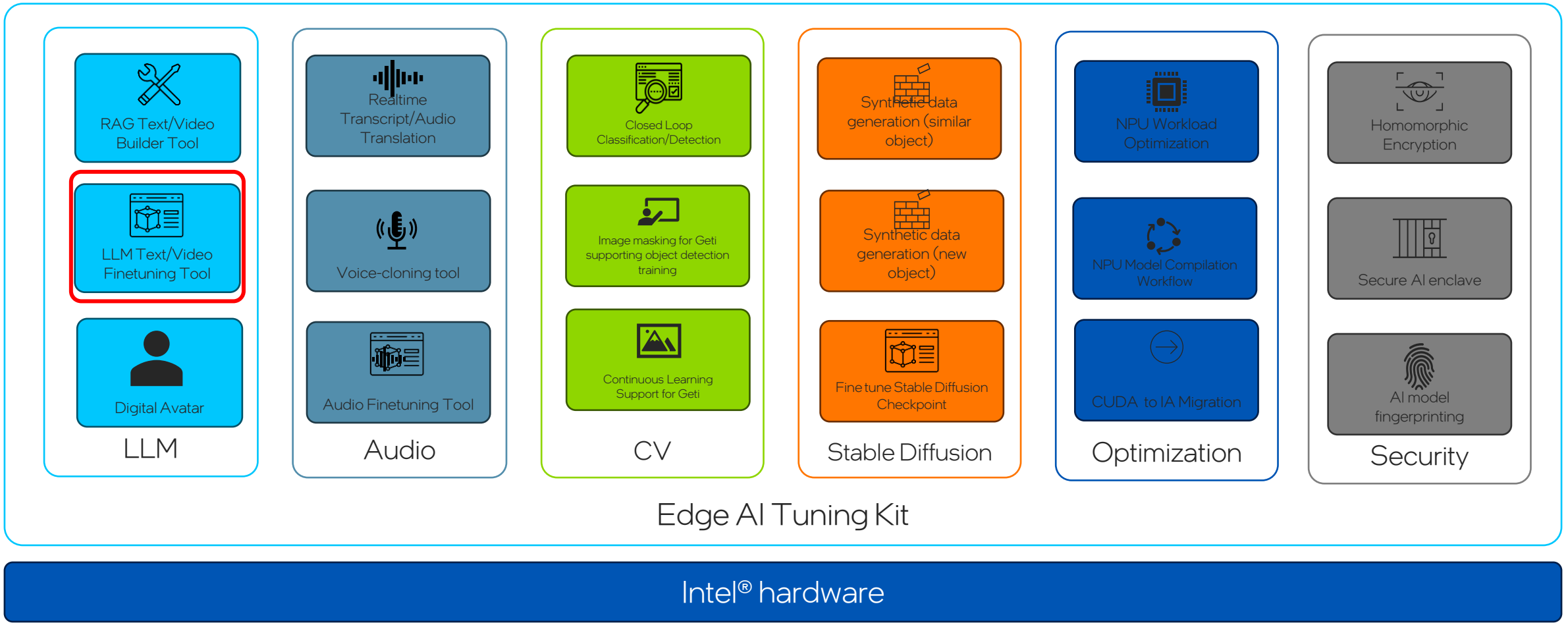
Introduction

Objective: to allow our customers to refine and productize AI models to automate and enhance existing workflows on Intel's platform

- Target Customer
  - OEM/Si/Solution Provider/ISV/User that wants to use AI to enhance their solution offering
- Pain
  - Accuracy: How to prevent AI from providing incorrect answers or context
  - Cost: How to democratize the AI model and run it on the edge directly to avoid costly CSP bill
  - Privacy: How to run AI on premise without sharing data
  - Use case: How can I use AI to enhance my solutions
  - Dataset: How can I feed my existing data to feed to my AI
- Gain
  - Better customer experience: customers can interact with the solution more naturally, with no more specific voice command
  - More functionality: use the generative capability of AI to perform multiple tasks/objectives

Edge AI Tuning Kit

- Our AI Toolkit's Architecture Stands on Six Pillars - LLM, Computer Vision, Audio, Stable Diffusion, Optimization and Security.



LLM Finetuning Toolkit

intel

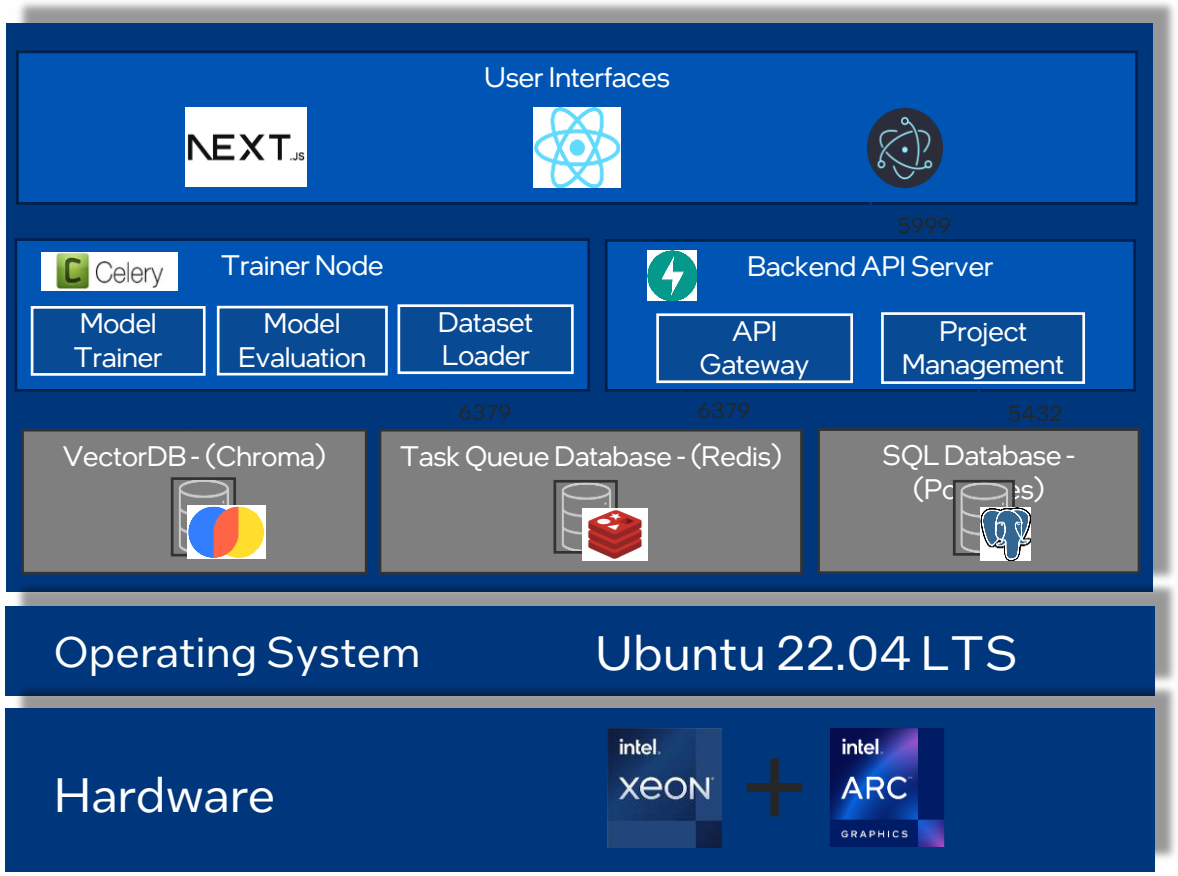
Features

No code GUI workflow to create customized LLM models for specific use cases

Core features:

- **Dataset Management**
  - Allow customers to add and modify the knowledgebase of LLM
- **Dataset Generation**
  - Generate knowledge base based on PDF, text files, webpages, etc.
- **Model Finetuning**
  - Finetune general-purpose open-source model for the custom vertical use cases.
  - Support most of the popular models: Llama 3, Mistral, Qwen 2.5
- **Model Evaluation**
  - Allow users to test and chat with their model before deployment
- **Deployment**
  - Provide OpenAI-compatible REST API for model serving
  - Deployment package that can run on the edge platform

Software Architecture



Training Server (Local)

Use cases

Medical

- **Automated Medical Report Generation**: Efficiently create comprehensive medical reports using AI to integrate patient data and clinical information.
- **Preliminary Health Condition Analysis**: assess patient health by analyzing verbal or written transcripts for key symptoms and medical history

Retail

- **Drive-Thru Kiosk Agent**: Streamline order processing, enhance customer experience, and suggest personalized menu options based on historical data.
- **AI-Powered Travel Planner Agent**: Create personalized travel itineraries by analyzing user preferences and real-time data to create customized travel plan

University

- **FAQ Generation:** Generate quiz and faq based on the domain knowledge instilled during the finetuning process.
- **Offline tutor**: Students can access to the knowledge base trained in the model offline with their personal PC.

Finetune Your Own Flavor of LLM

When to finetune your LLM model?

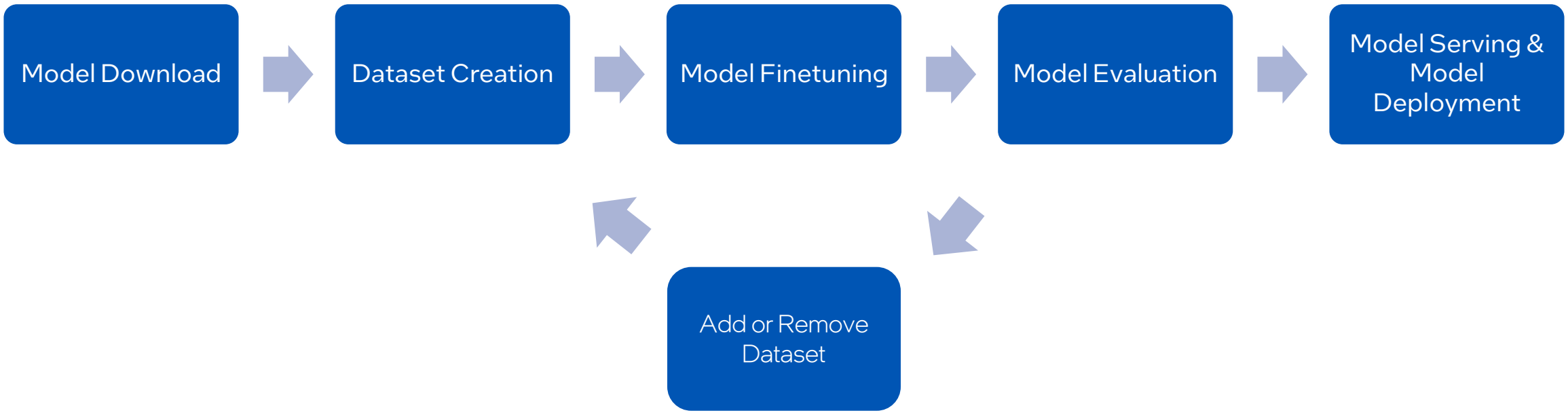
Produces more accurate and reliable outputs compared to prompting alone

Can learn from a larger dataset than what a single prompt can accommodate

Reduces token usage by minimizing prompt length

Delivers faster response times due to shorter inputs

Fine-tuning Process



In this workshop

In this workshop

Linda Perez is a 55-year-old female who came in complaining of upper abdominal pain radiating to her back, associated with nausea and two episodes of vomiting. She has a history of gallstones and type 2 diabetes. She is currently taking Metformin and Glipizide and has no medication allergies. On exam, she was mildly tachycardic with a BP of 130/80 and temperature of 99.5°F. Labs showed elevated lipase and amylase, and an abdominal ultrasound revealed gallstones without ductal dilation. She was diagnosed with acute pancreatitis likely due to gallstone obstruction. She has been admitted for IV fluids, pain control, and GI consultation.

Example of a written medical transcript for patient

Model Finetuning

Medical Report

Patient Name: Linda Perez **Age/Gender:** 55 / Female **Date of Visit:** *Date of Visit*

Chief Complaint

Acute upper abdominal pain radiating to the back, accompanied by nausea and vomiting.

Past Medical History

- Gallstones
- Type 2 diabetes
- Medications: Metformin, Glipizide

Vital Signs

- Heart Rate: 110 bpm
- Blood Pressure: 130/80 mmHg
- Temperature: 99.5°F

Laboratory Findings

- Elevated lipase
- Elevated amylase
- Abdominal ultrasound: Gallstones without ductal dilation

Assessment

Acute pancreatitis likely due to gallstone obstruction.

Plan

- IV fluids
- Pain control
- GI consultation
- Consideration for endoscopic cholecystostomy

Hospital Course

Patient has been admitted for stabilization and further management.

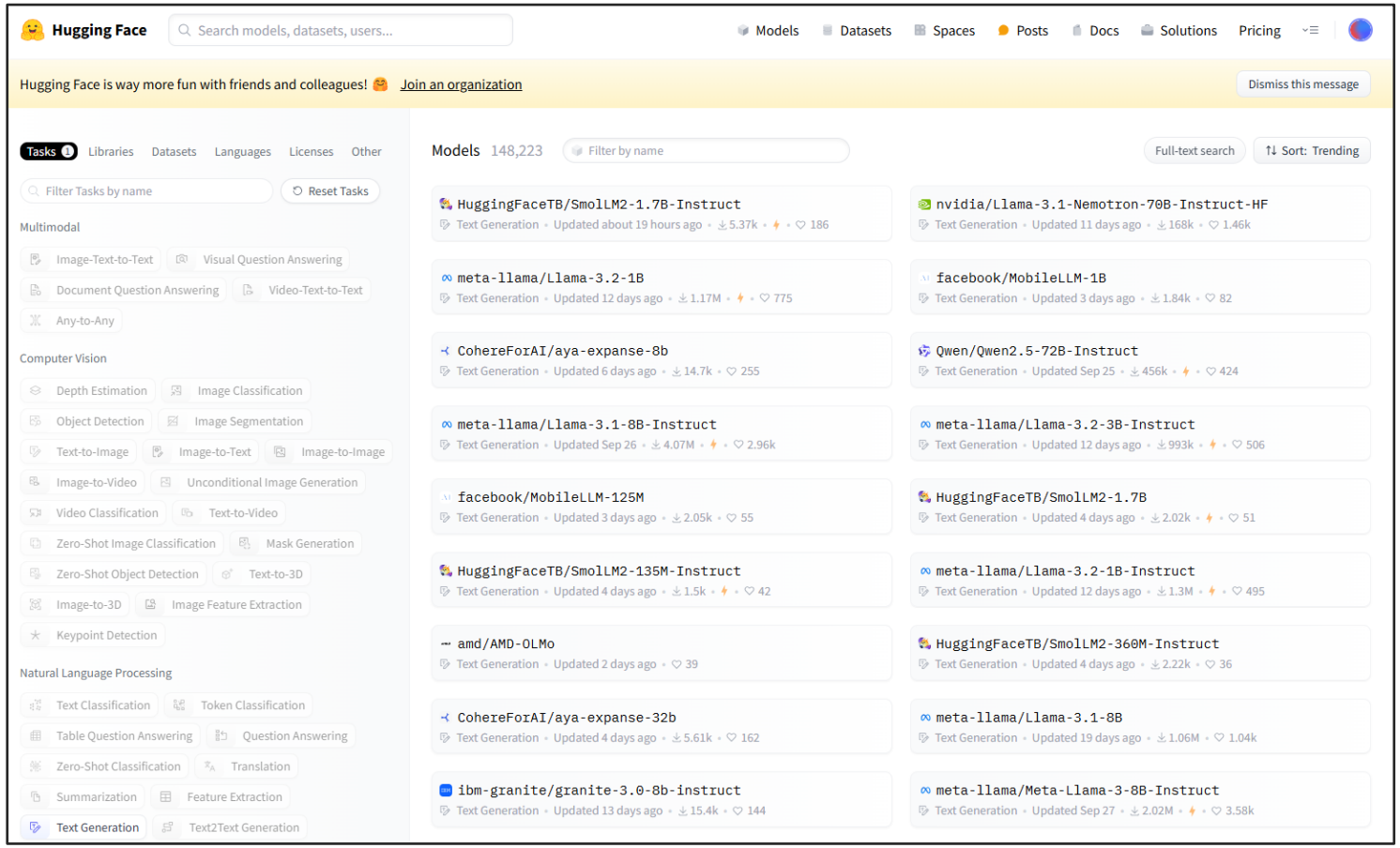
Discharge Summary

Managed for acute pancreatitis. Patient is stable on IV fluids and pain control. Scheduled for endoscopic evaluation and possible intervention. Discharged with close outpatient follow-up.

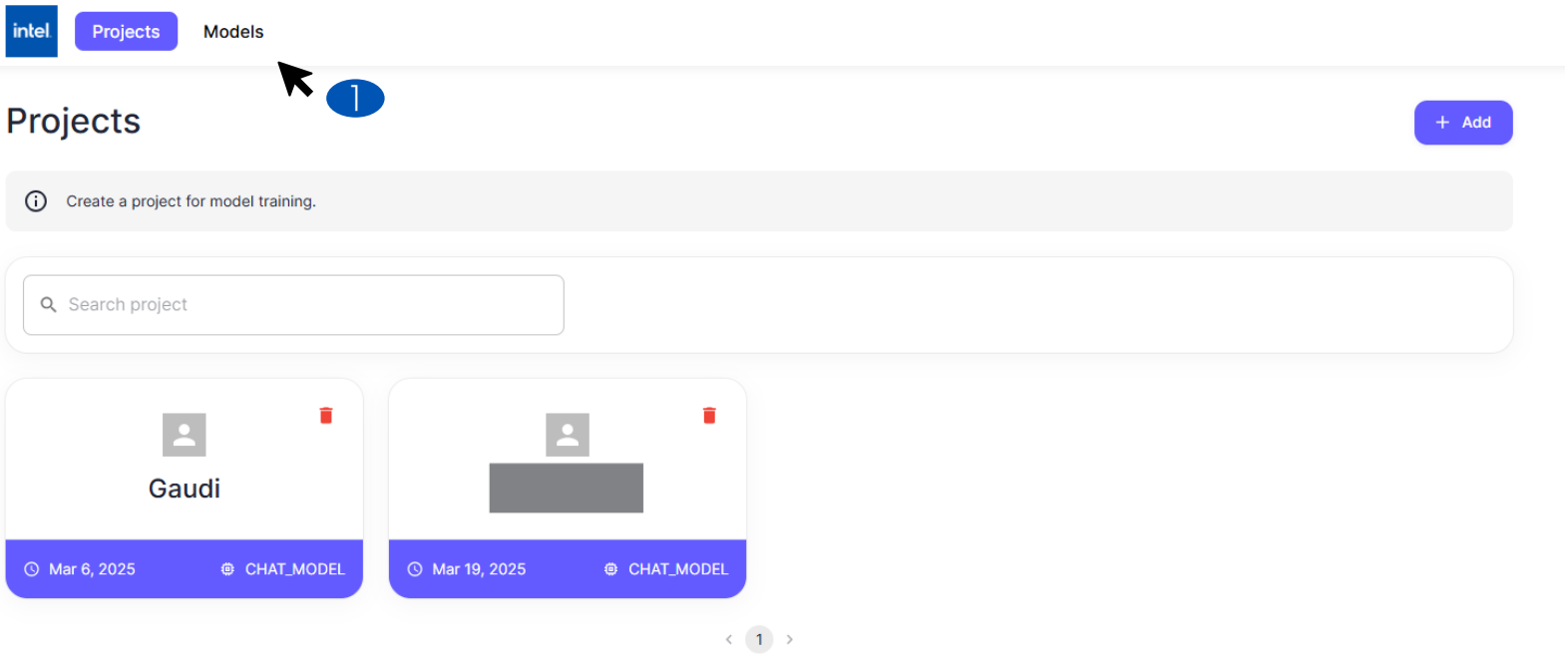
Finetuning a LLM model to perform medical report generation.

Model Download

- Hugging Face is one of the largest model hub.
- Support text generation model from Hugging Face Model Hub.

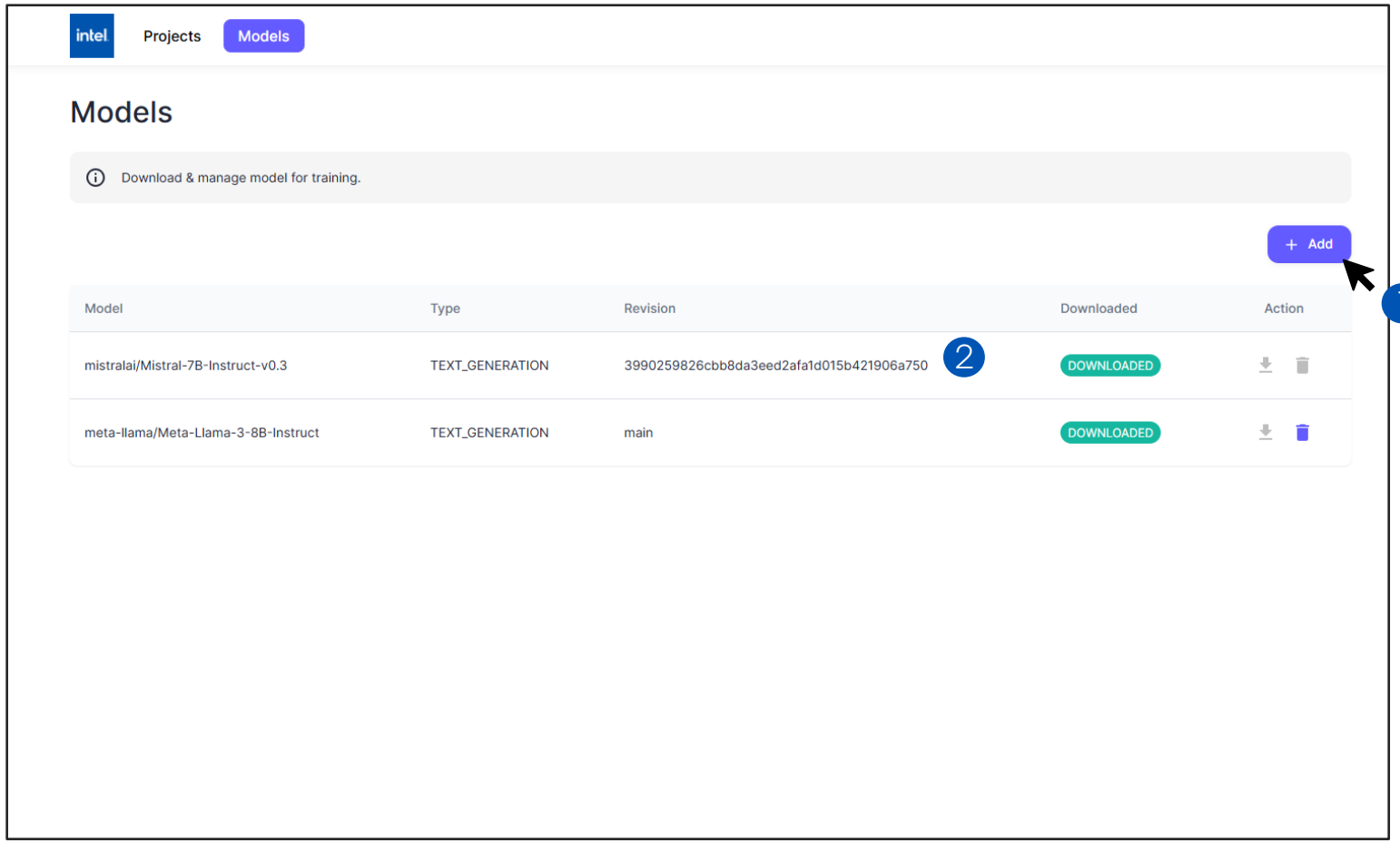


Model Download



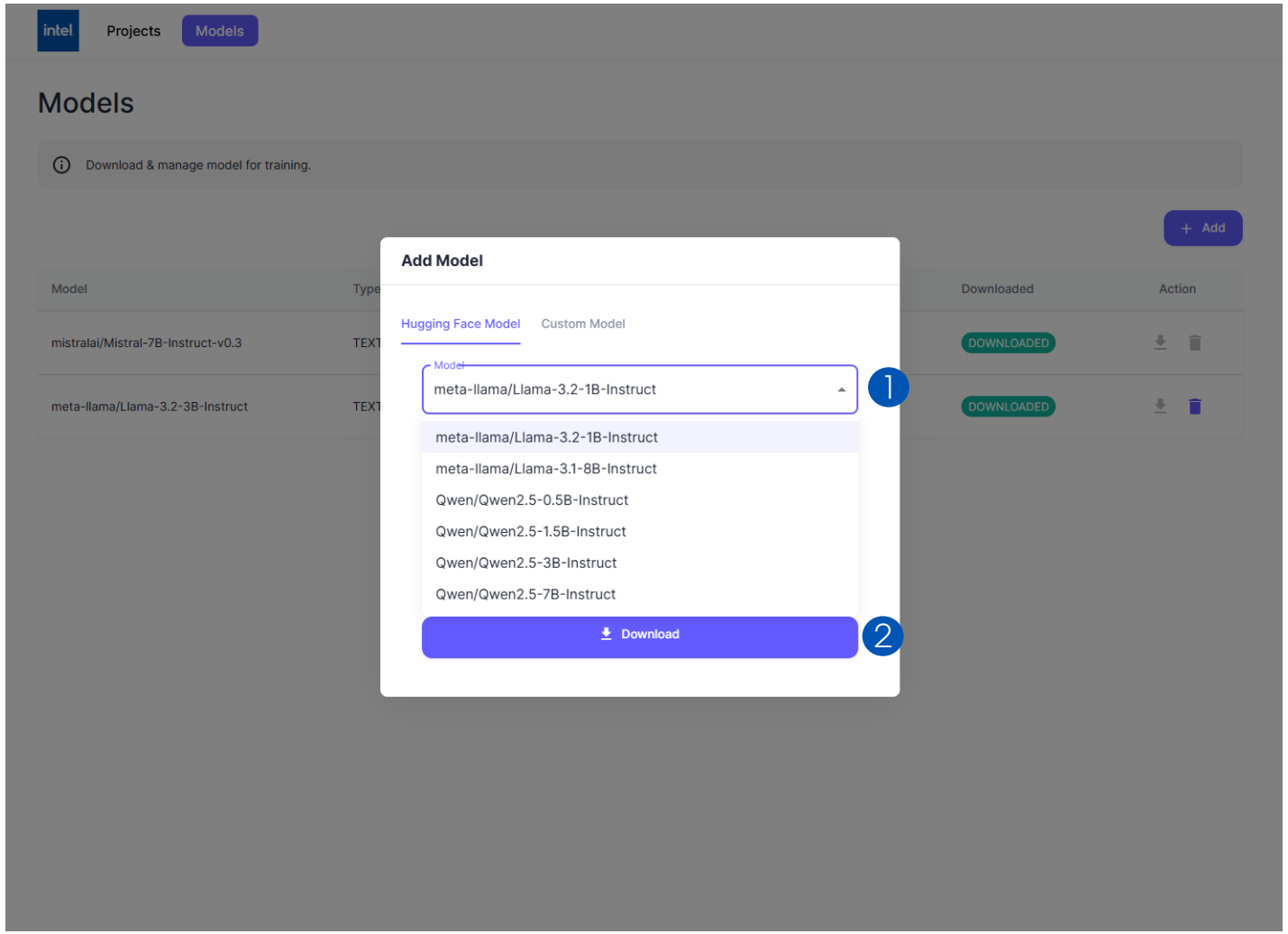
1. Click on the *Models* shown in the figure to access/add LLM models

Model Download



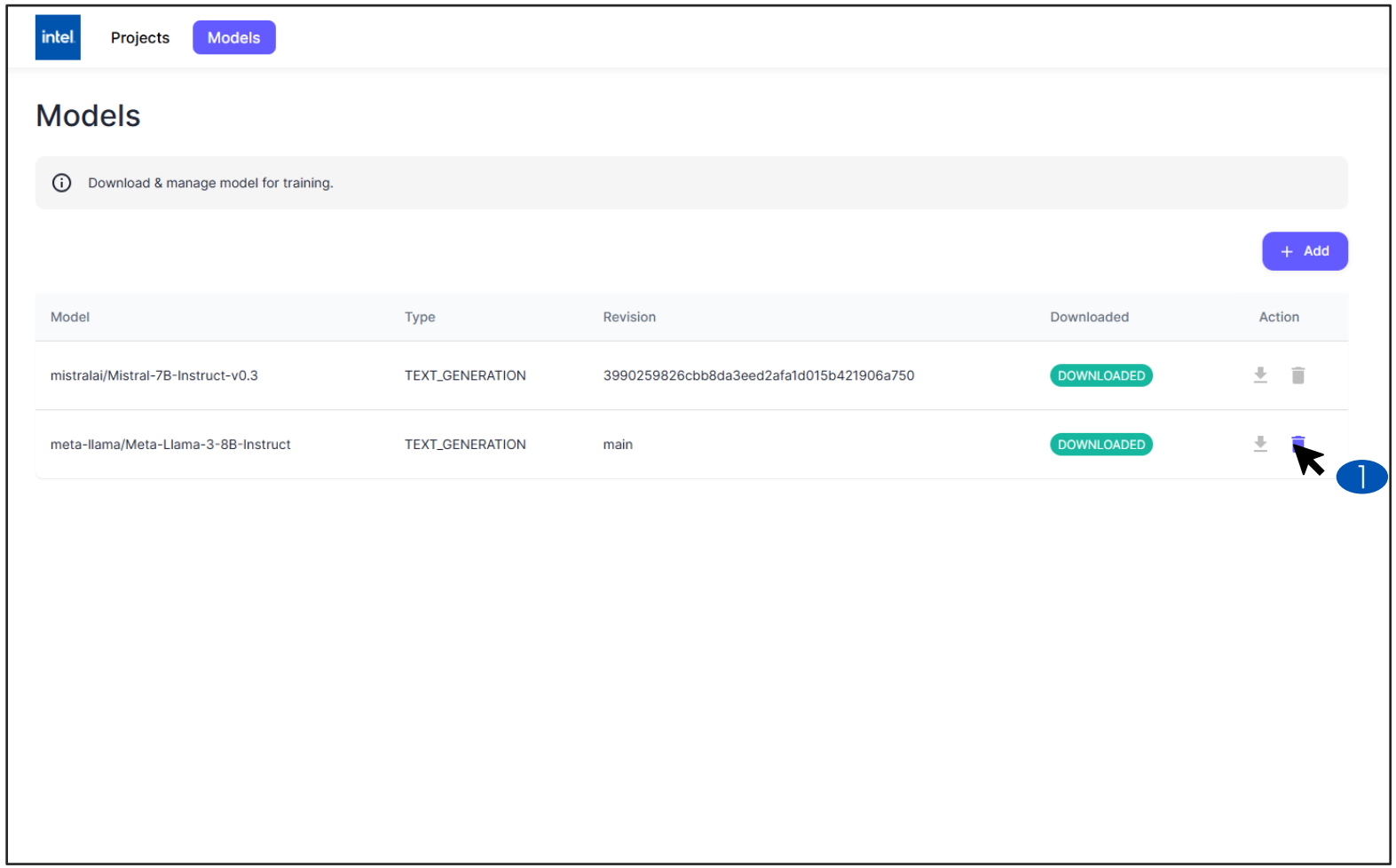
1. Click on the + Add icon to add a new LLM model based on your choice.
2. mistralai/Mistral-7B-Instruct-v0.3 is the default model in the system.

Model Download



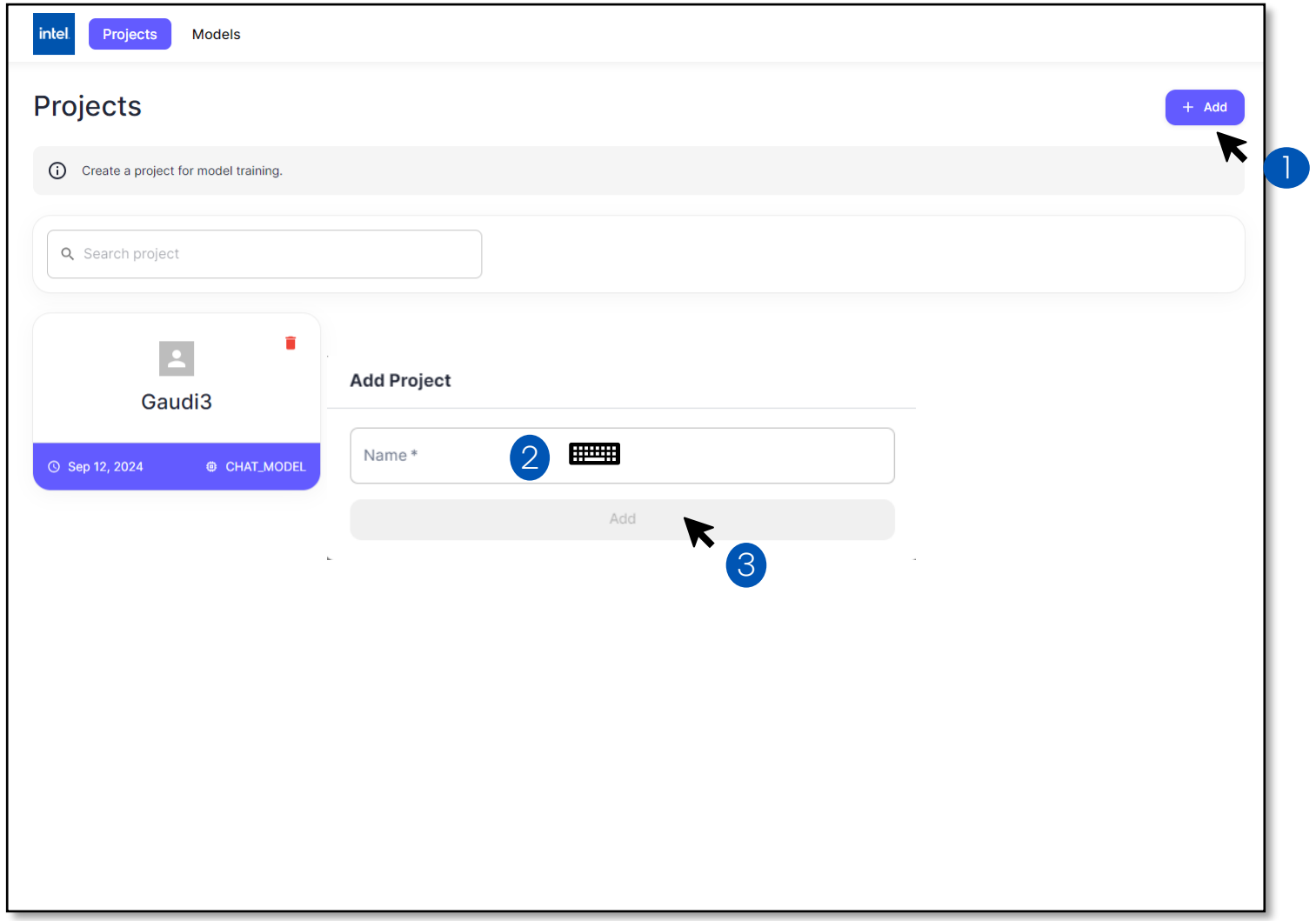
1. Select the model
2. Click on the Download button to start downloading the LLM model from the HuggingFace model hub.

Model Download



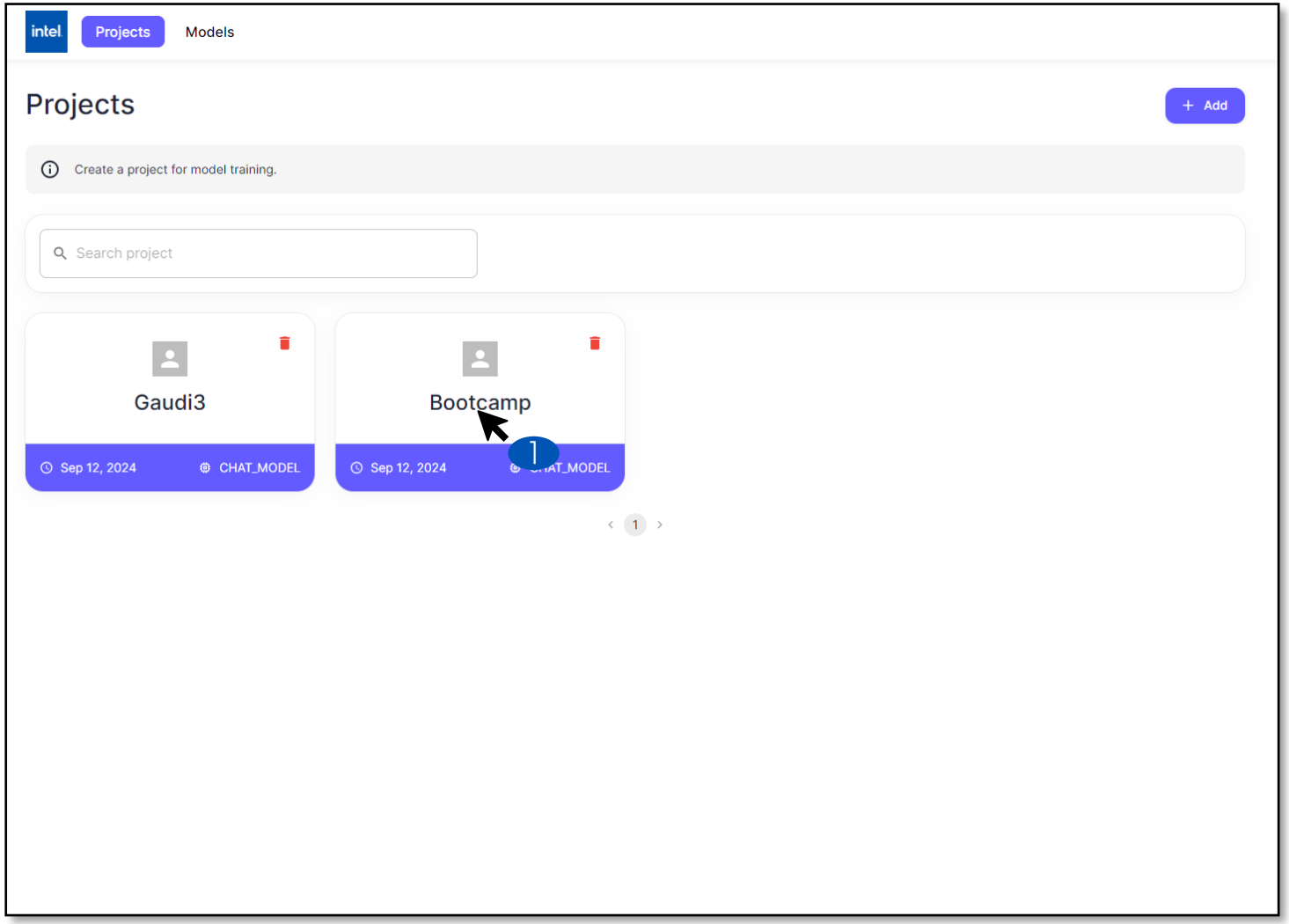
1. Click on the icon to delete the unwanted models.
2. Please note that the mistralai/Mistral-7B-Instruct-v0.3 cannot be removed since it is the default model.

Create Project



1. Click on the Add button to add a new project
2. Enter your desired name for the project.
3. Click on the Add button to complete the add project process

Create Project



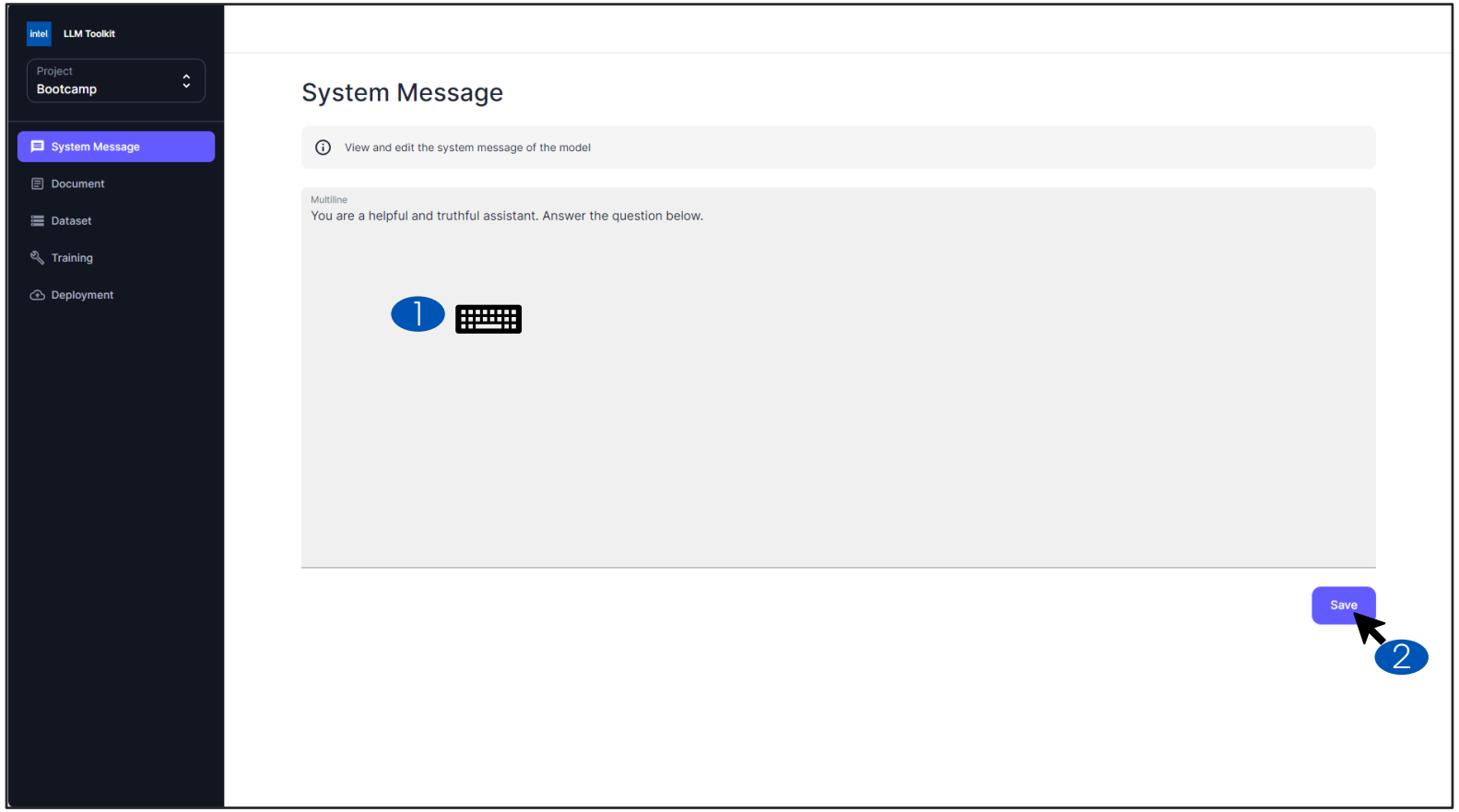
1. As we can see in the image on the left, a new project has been created successfully, click on the project created to proceed.

System Message

- Importance of System Prompt in LLMs
  - Guiding the context
  - Behavior Shaping
  - Consistency
  - Safety and Ethics
- How to Utilize System Prompt Effectively
  - Clear Instructions
  - Setting the Tone
  - Defining the Scope
  - Providing Examples

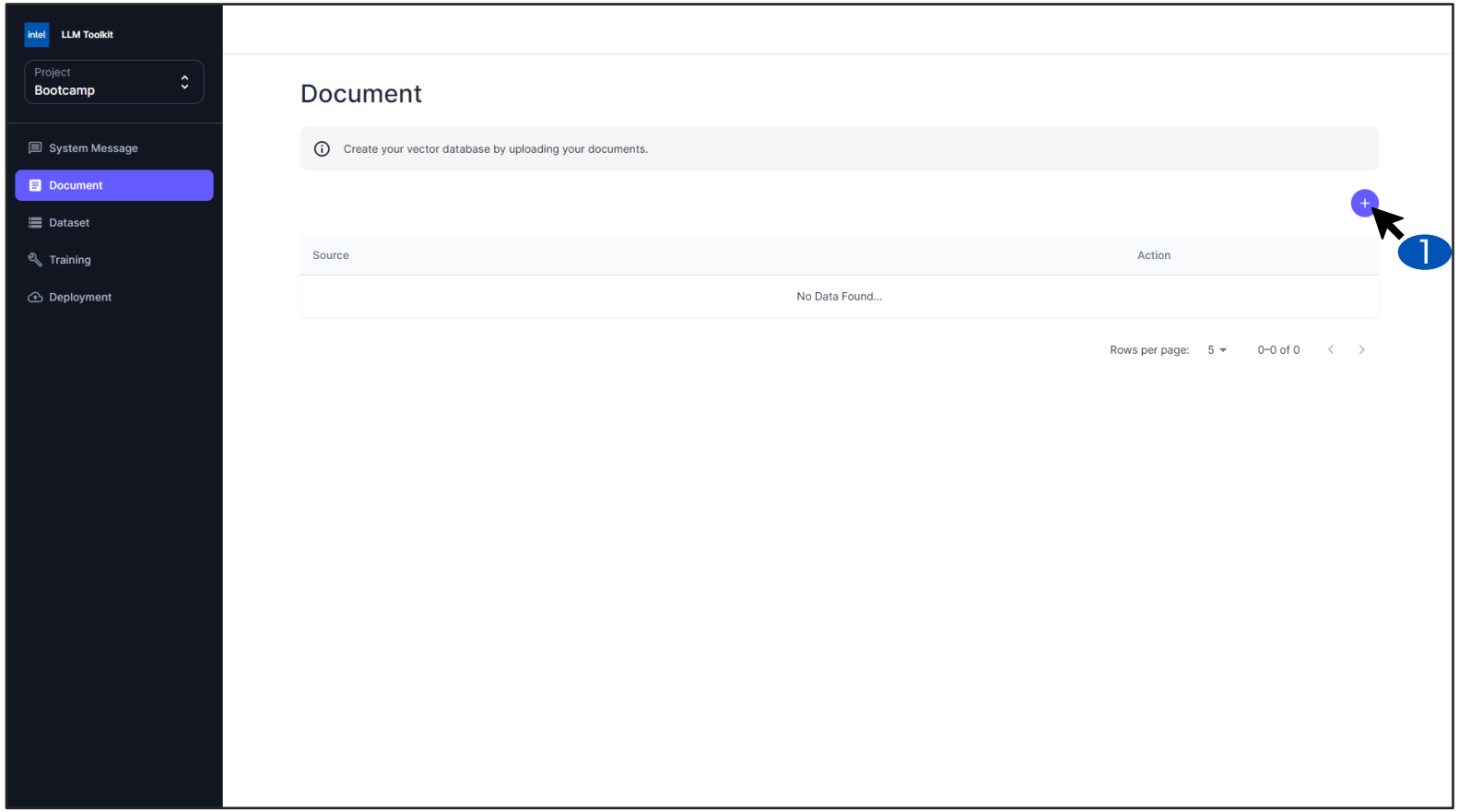
Example:
Act as a knowledgeable product assistant with expertise in Intel Gaudi 3. Answer user questions clearly and helpfully.

System Message



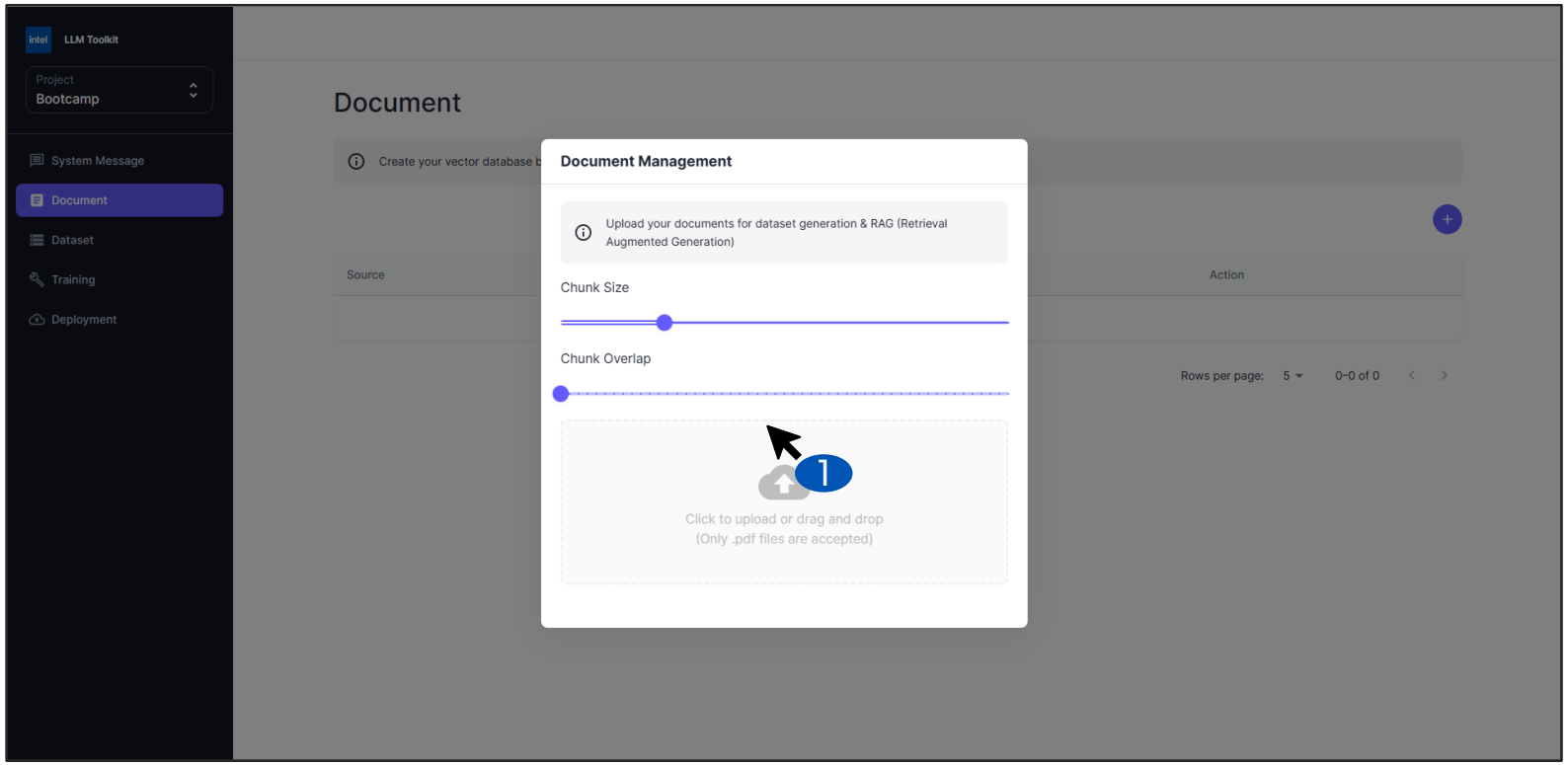
1. Enter the system message/ system prompt according to your use case.
2. Click on the save button to save the changes made to the system message.

Document



1. Click on the plus sign button to add documents for Retrieval Augmented Generation (RAG).

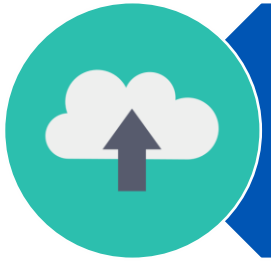
Document



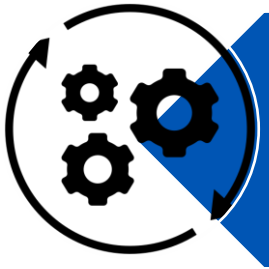
1. Click on the area to add documents for Retrieval Augmented Generation (RAG)

Dataset Creation

LLM Finetuning Toolkit supports multiple types of dataset creation method



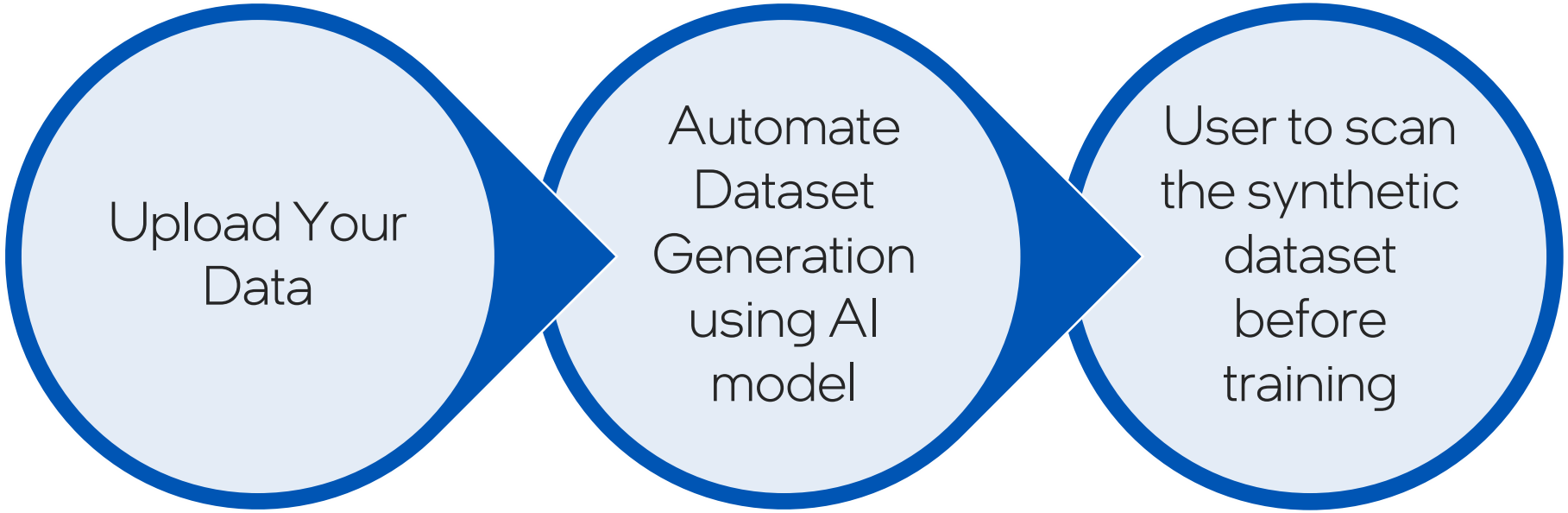
Dataset Upload



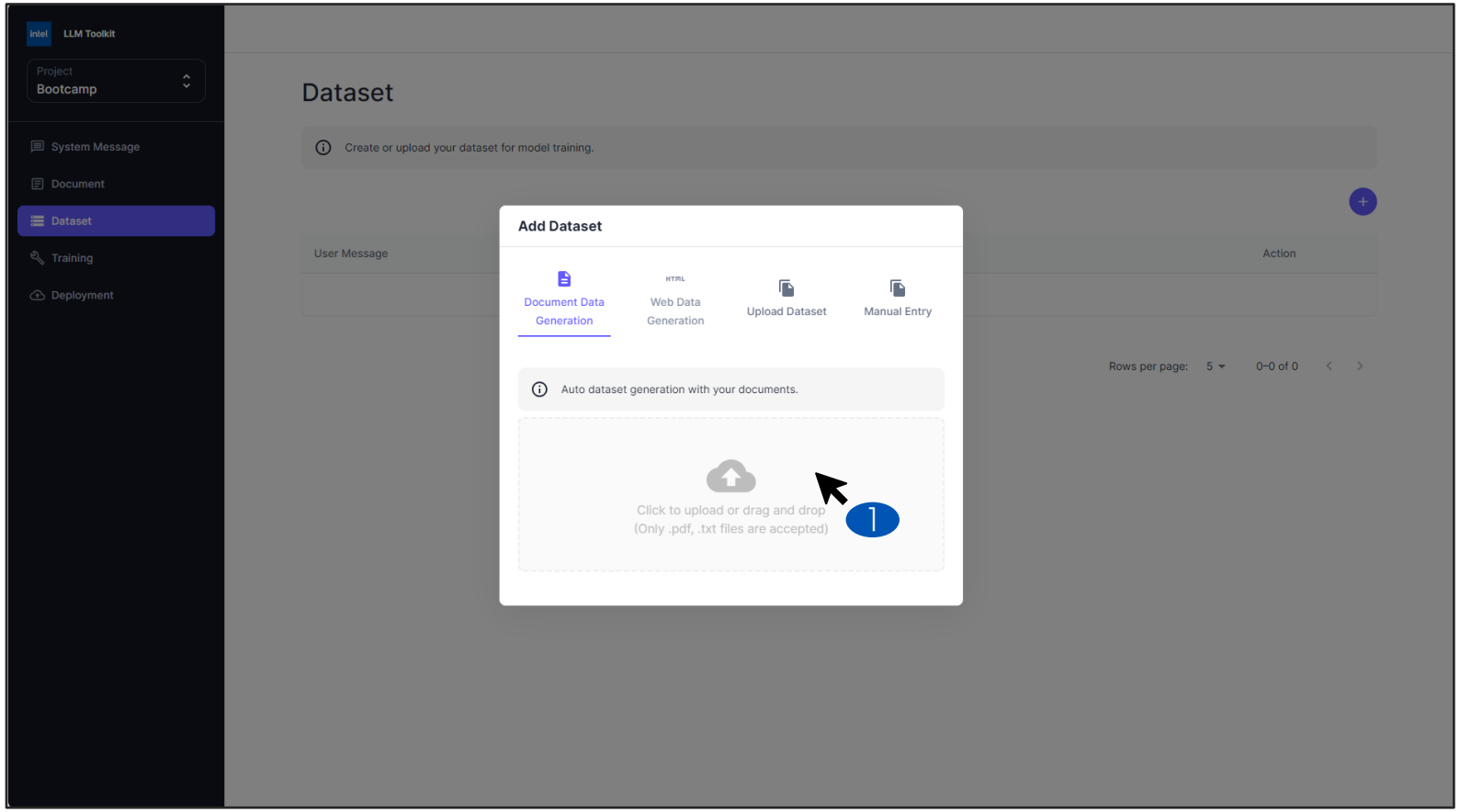
Automatic Dataset Creation

User Entry

Automatic Dataset Creation

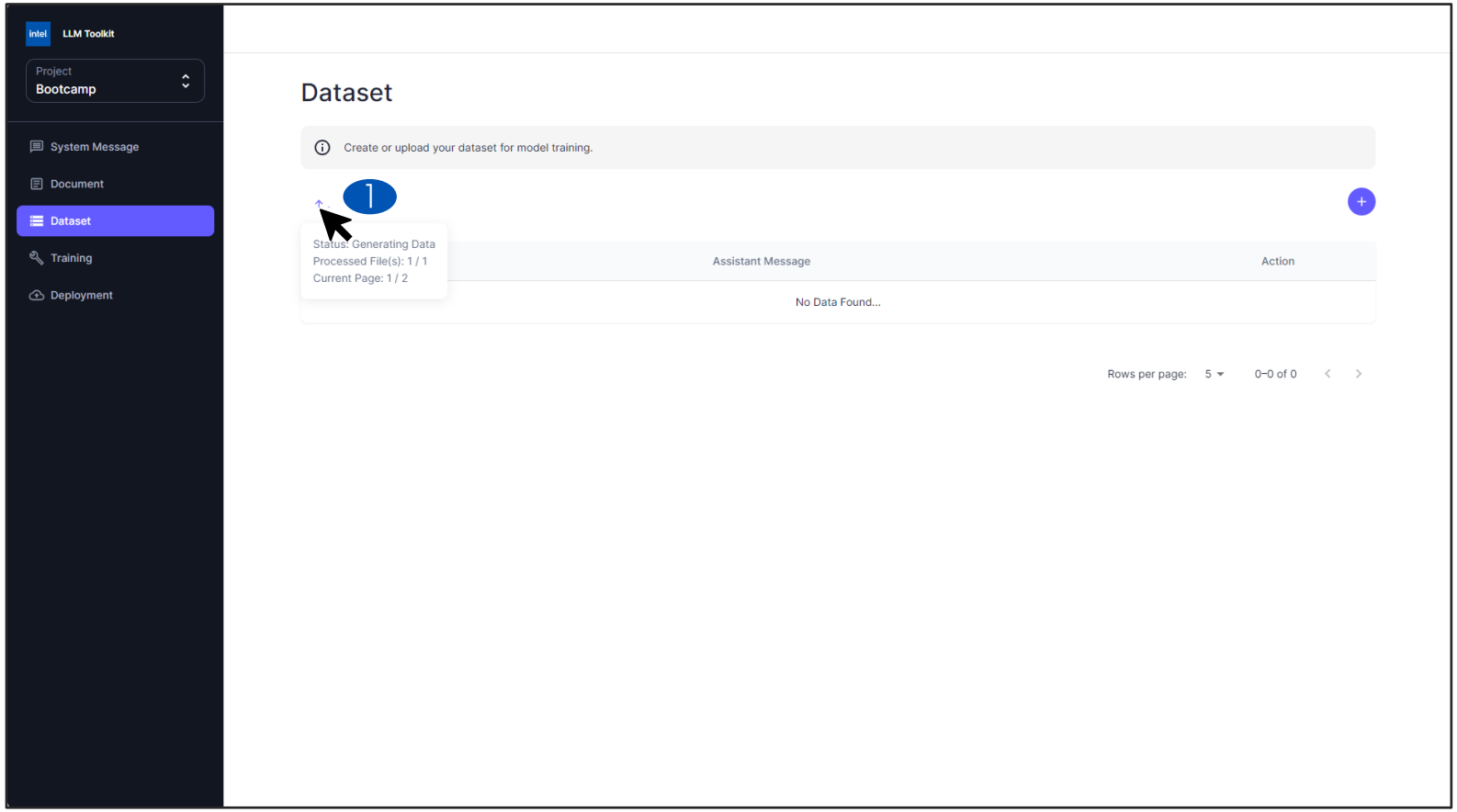


Automatic Dataset Creation



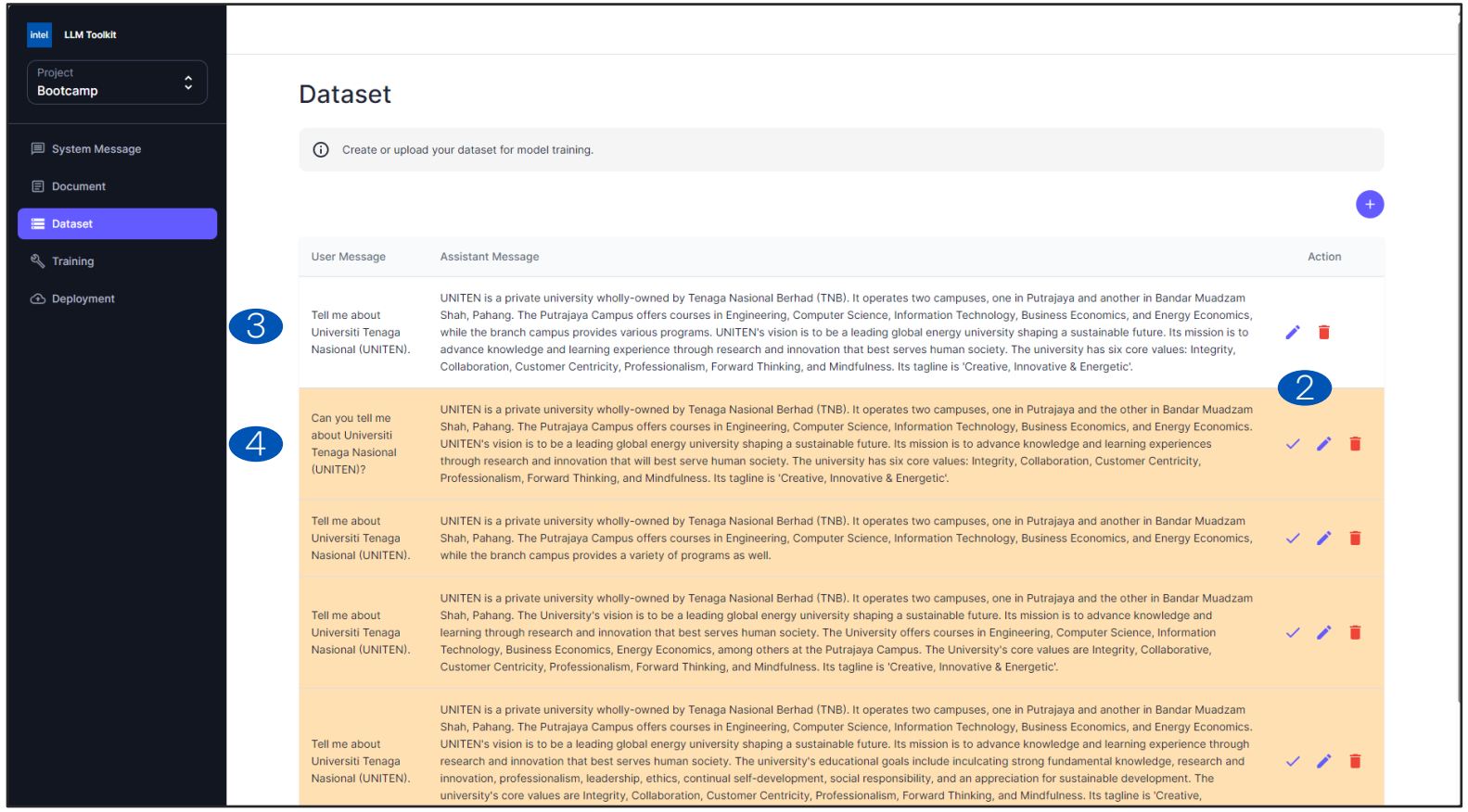
1. Upload documents in PDF (.pdf) or text file (.txt) by clicking on the area labelled
2. More data types will be supported in the upcoming version

Automatic Dataset Creation



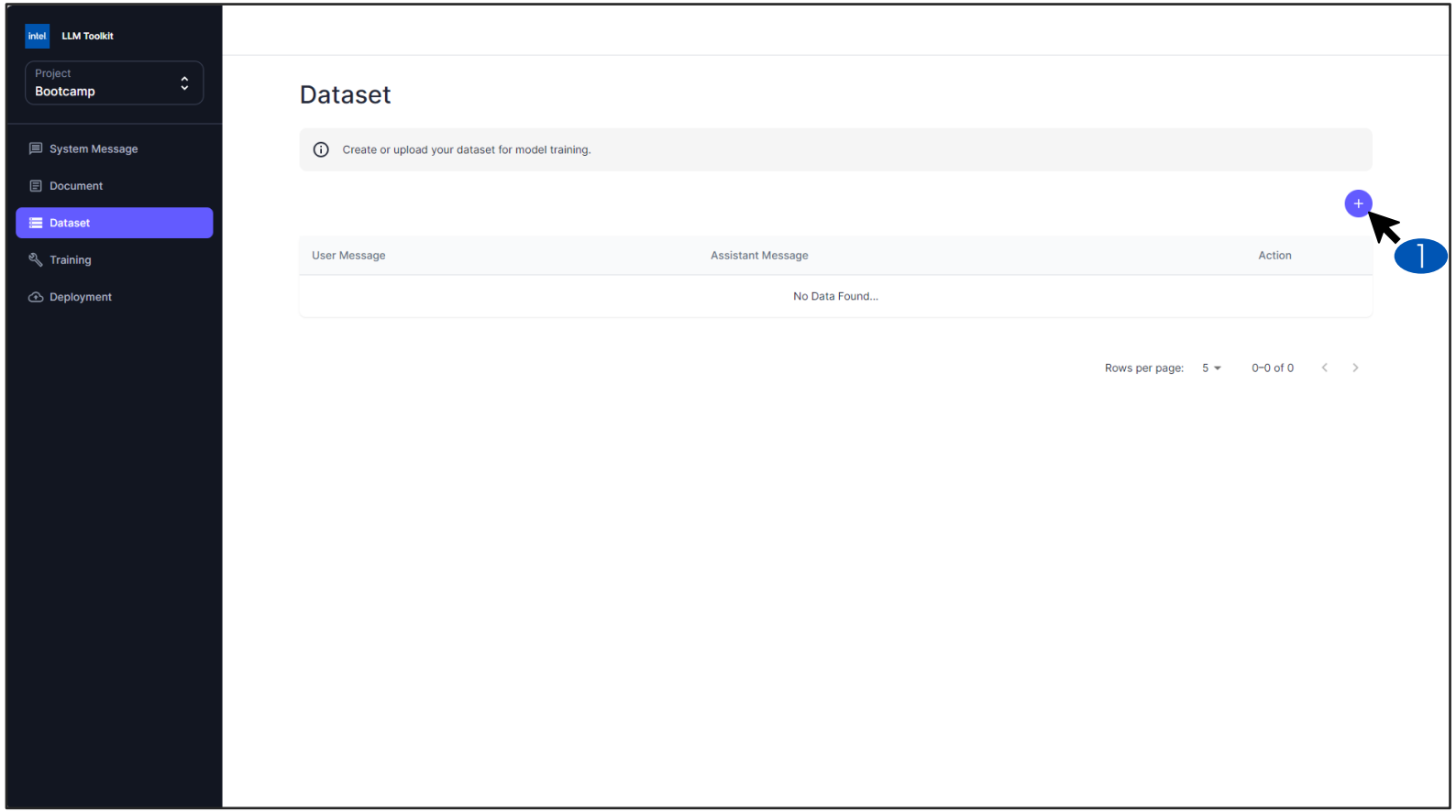
1. User can view the process by clicking on the arrow as shown in the figure.

Automatic Dataset Creation



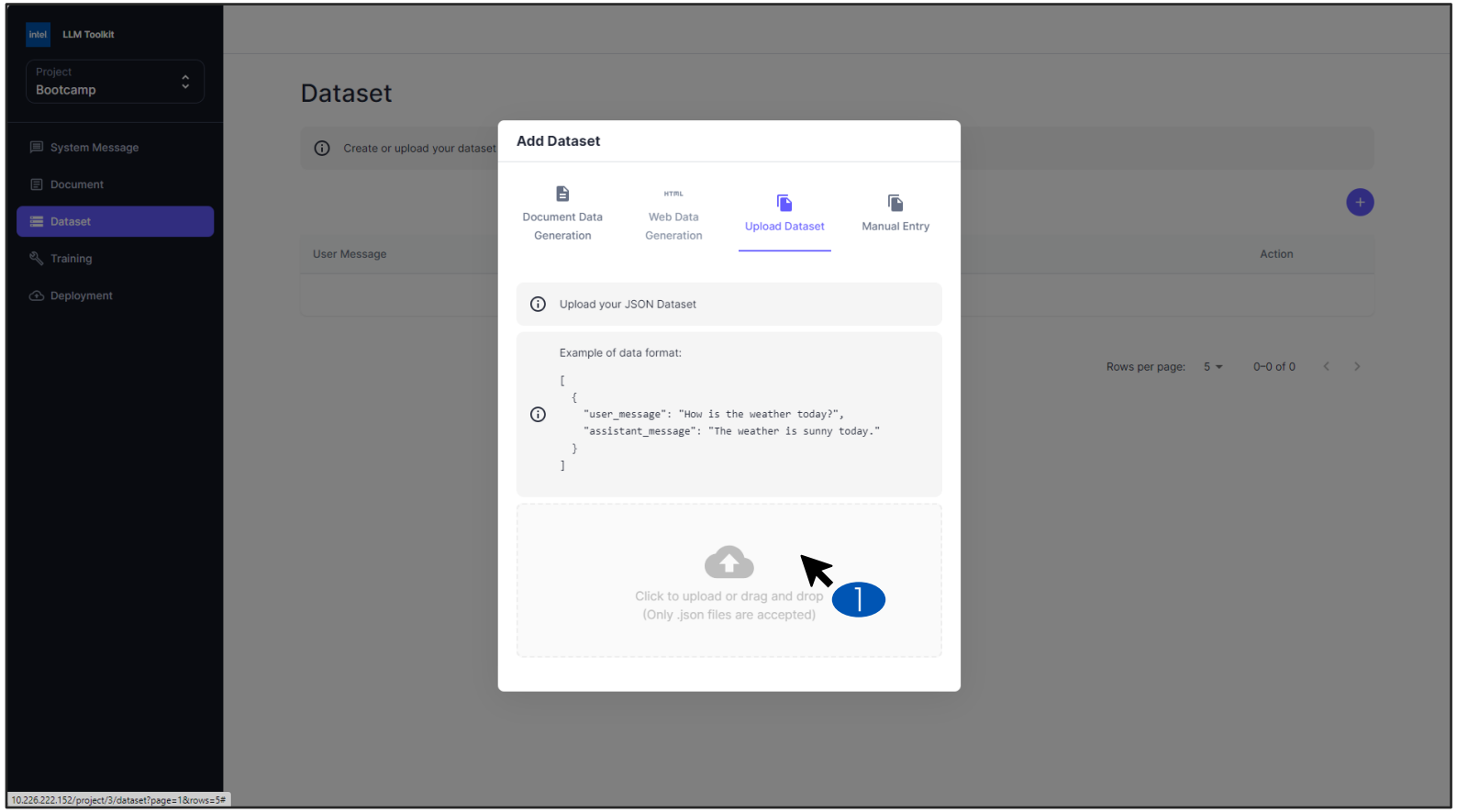
1. Dataset generated might need some filtering.
2. The entry in white is the entry that is confirmed.
3. The entry in orange is the entry that has not been confirmed.

Dataset Creation



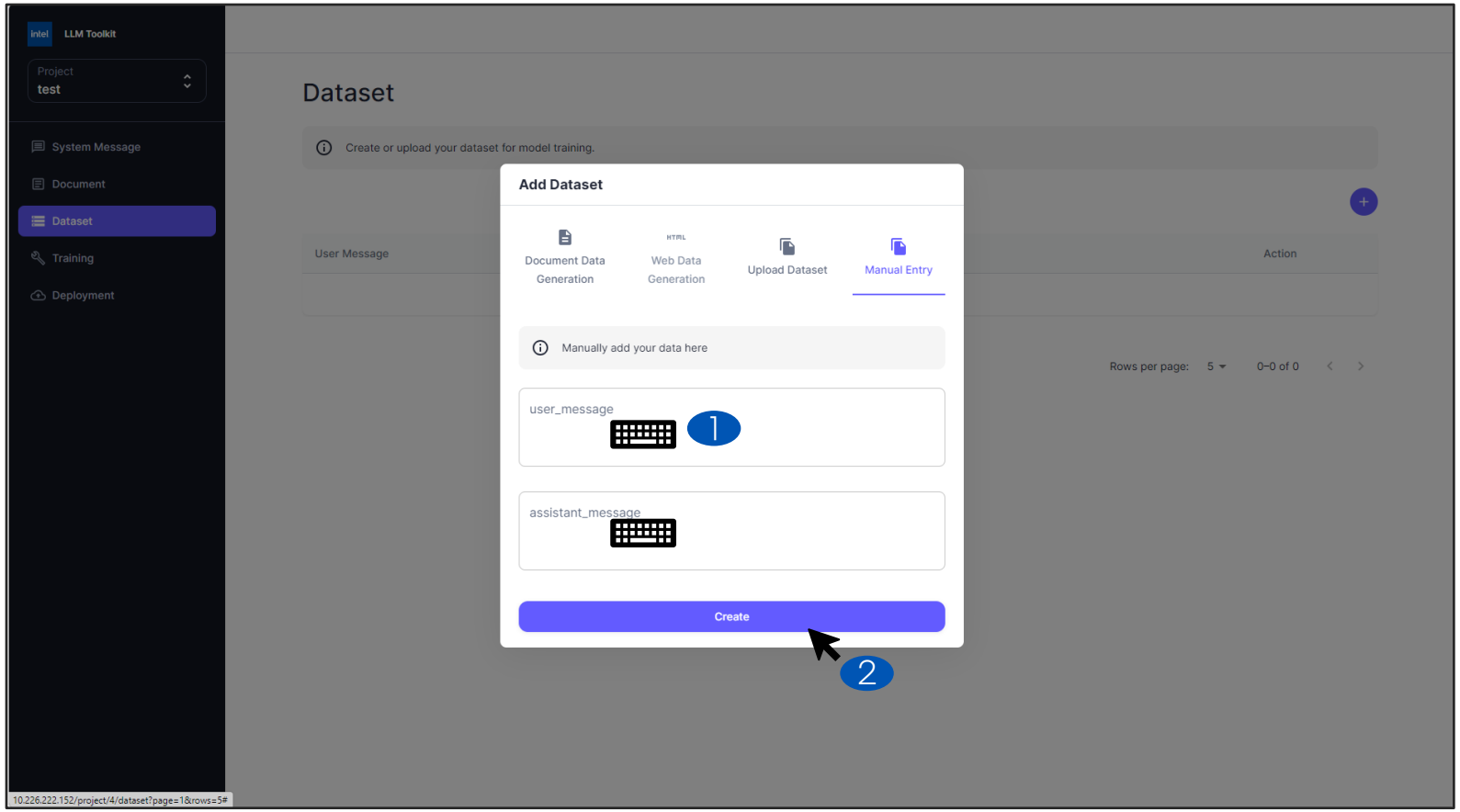
1. Click on the icon to add datasets for finetuning purpose.

Dataset Upload



1. Upload documents in JSON format (.json) by clicking on the area labelled

User Entry



1. Manually entry the user message and assistant message as dataset.
2. Click on create to entry the data

Training

01

Parameter Adjustment:

Modify training parameters such as learning rate, batch size, and number of epochs specifically for fine-tuning.

02

Training Execution:

Run the training process, allowing the model to learn from the new dataset while retaining previously learned information.

03

Monitoring Metrics:

Track performance metrics like accuracy, perplexity, and train loss to evaluate the effectiveness of fine-tuning.

Training

Learning Rate: Determines the step size at each iteration while moving toward a minimum of the loss function.

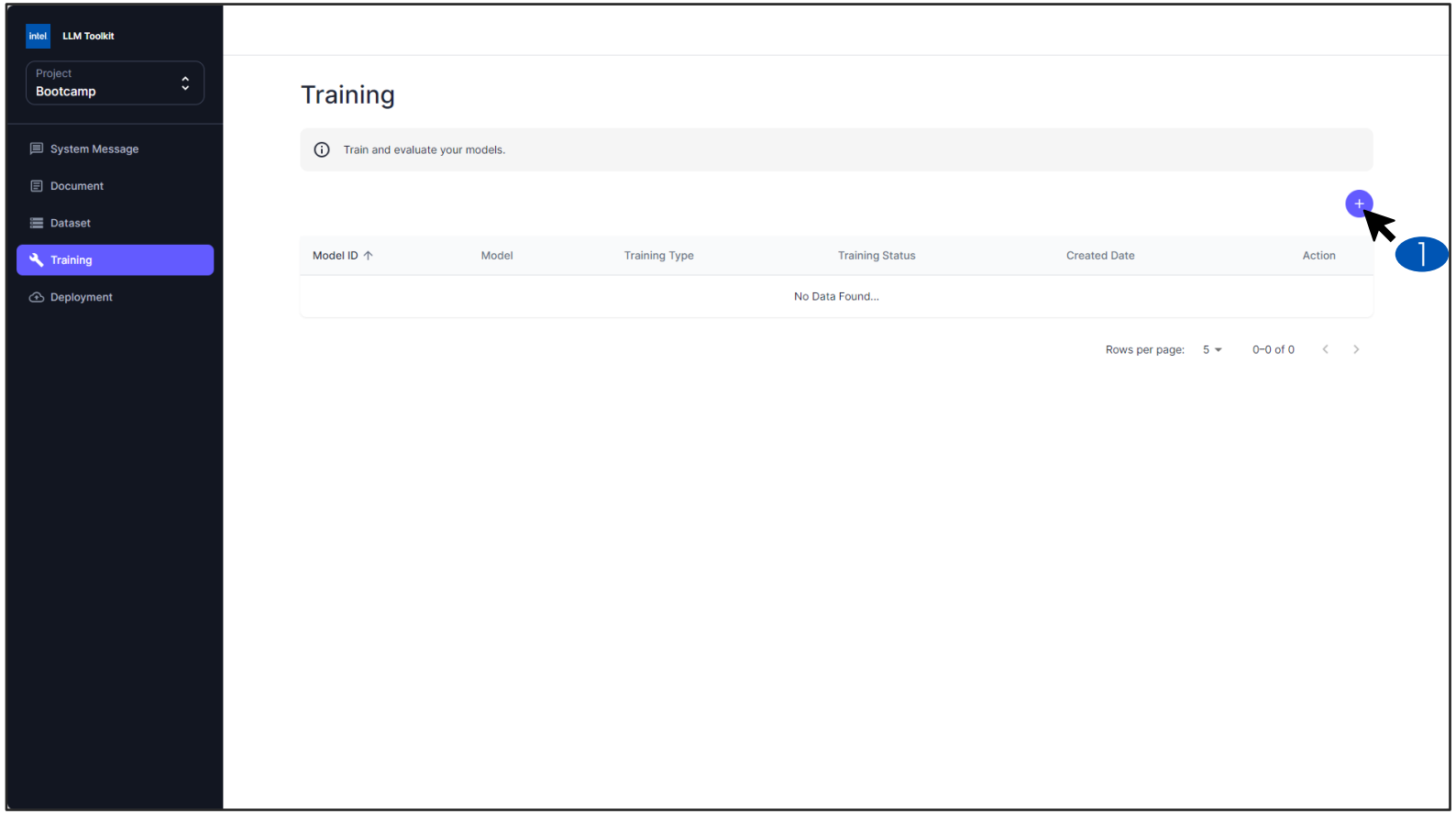
Batch Size: The number of training examples used in one iteration.

Epochs: The number of times the learning algorithm will work through the entire training dataset.

Loss Function: Measures how well the model is performing, guiding the adjustment of parameters.

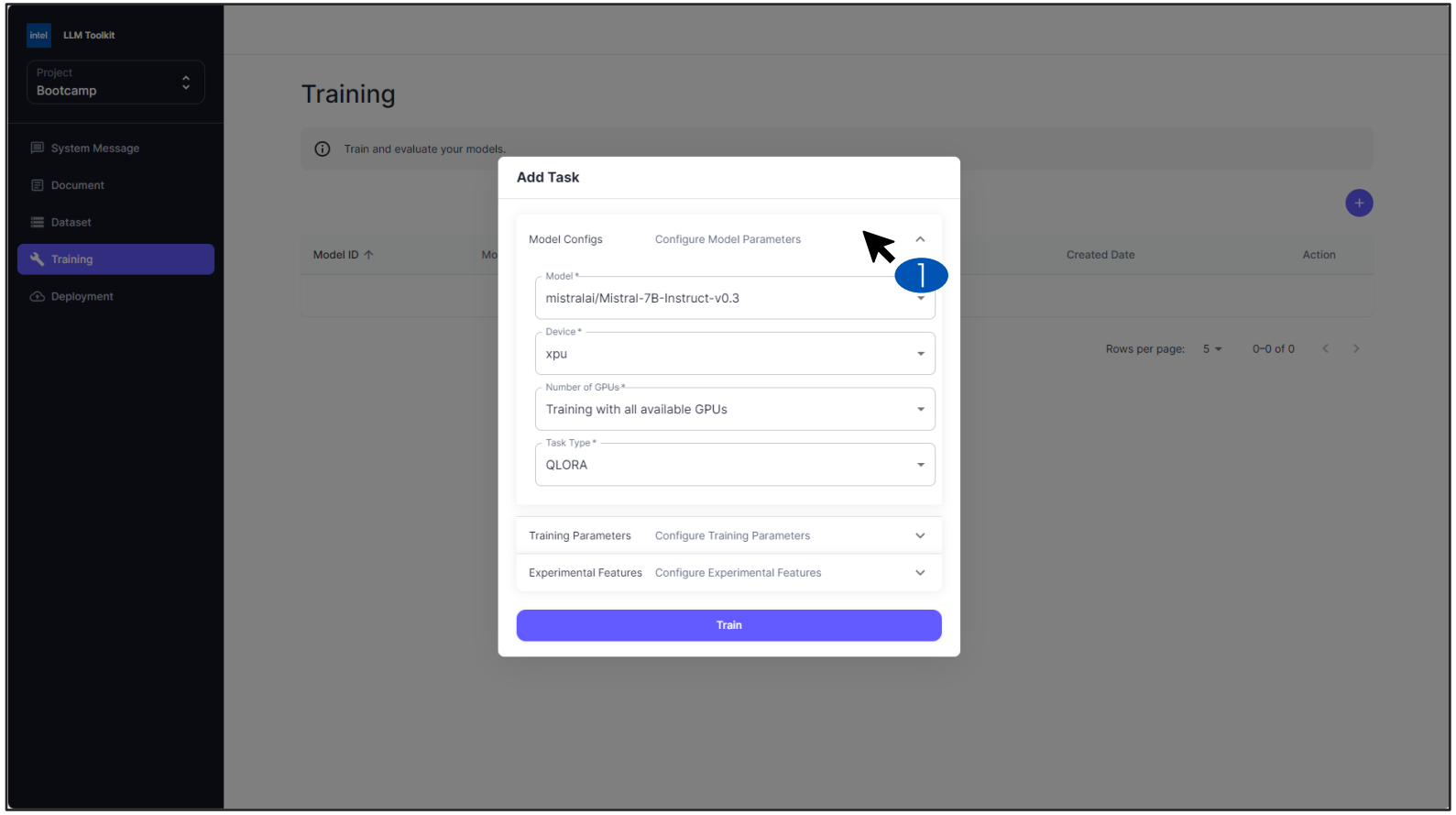
Optimizer: The method or algorithm used to change the attributes of the neural network such as weights and learning rate to reduce the losses.

Training



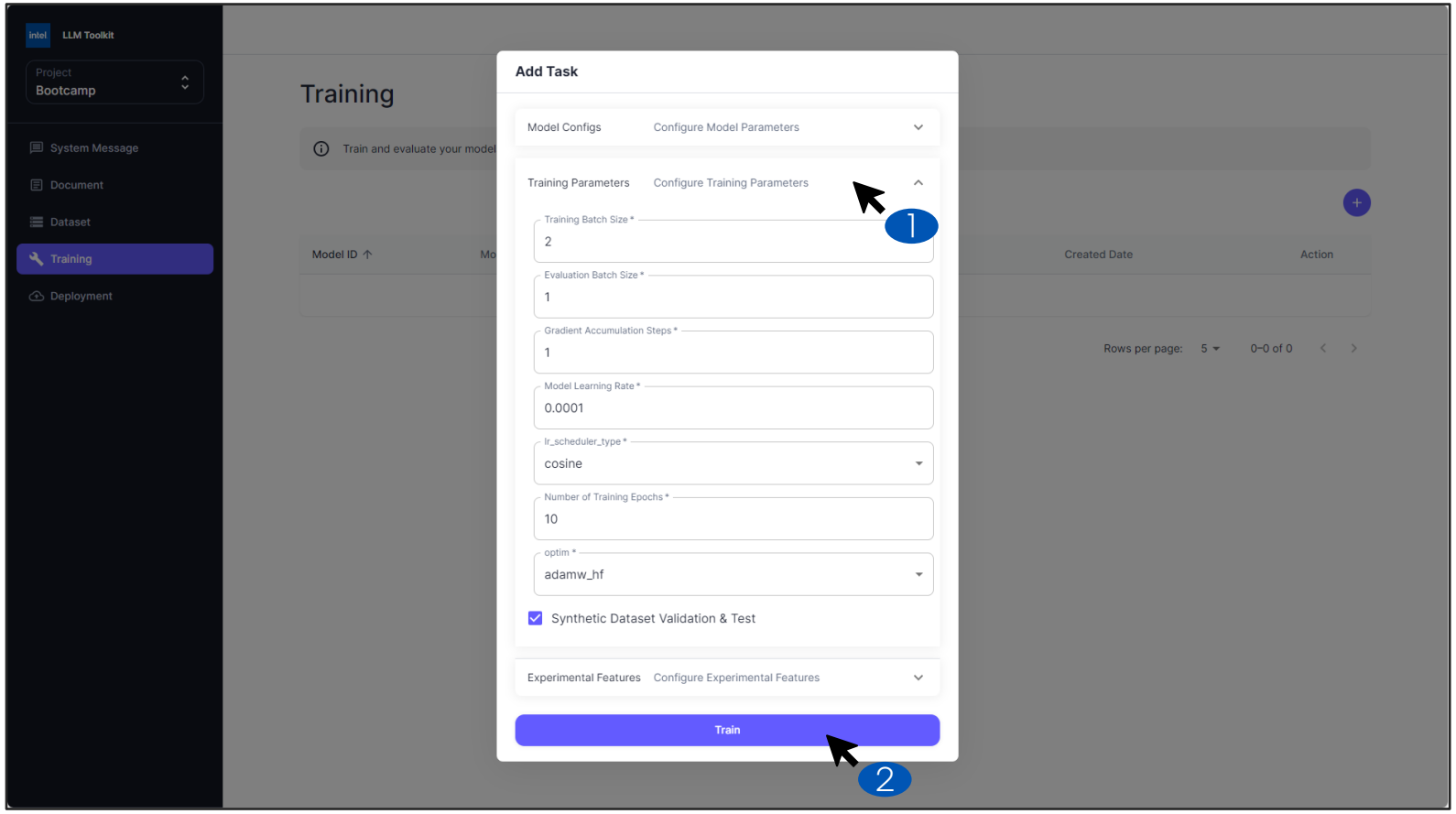
1. User can start the training once the dataset has been completely prepared, click on the add button to start the training process.

Training



1. The model configuration UI is shown after clicking on the add button for user to configure.
2. User can choose the model, device, number of devices and task type in the first tab of the configuration.

Training

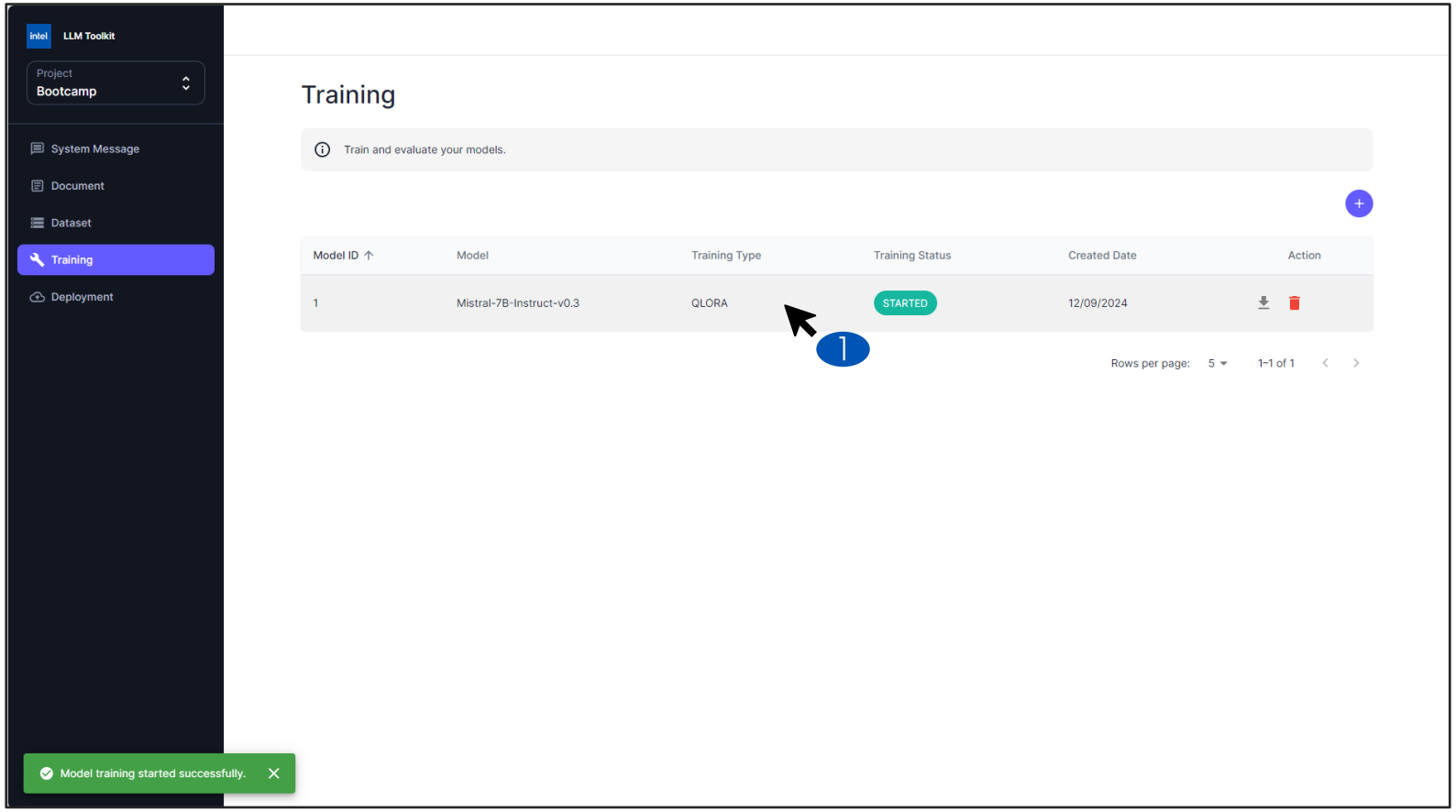


1. In the 2nd tab, user can configure the training parameters such as training and evaluation batch size, model learning rate et cetera as shown in the figure.
2. Click on the Train button to start the training after all the configuration has been done.

Training Results

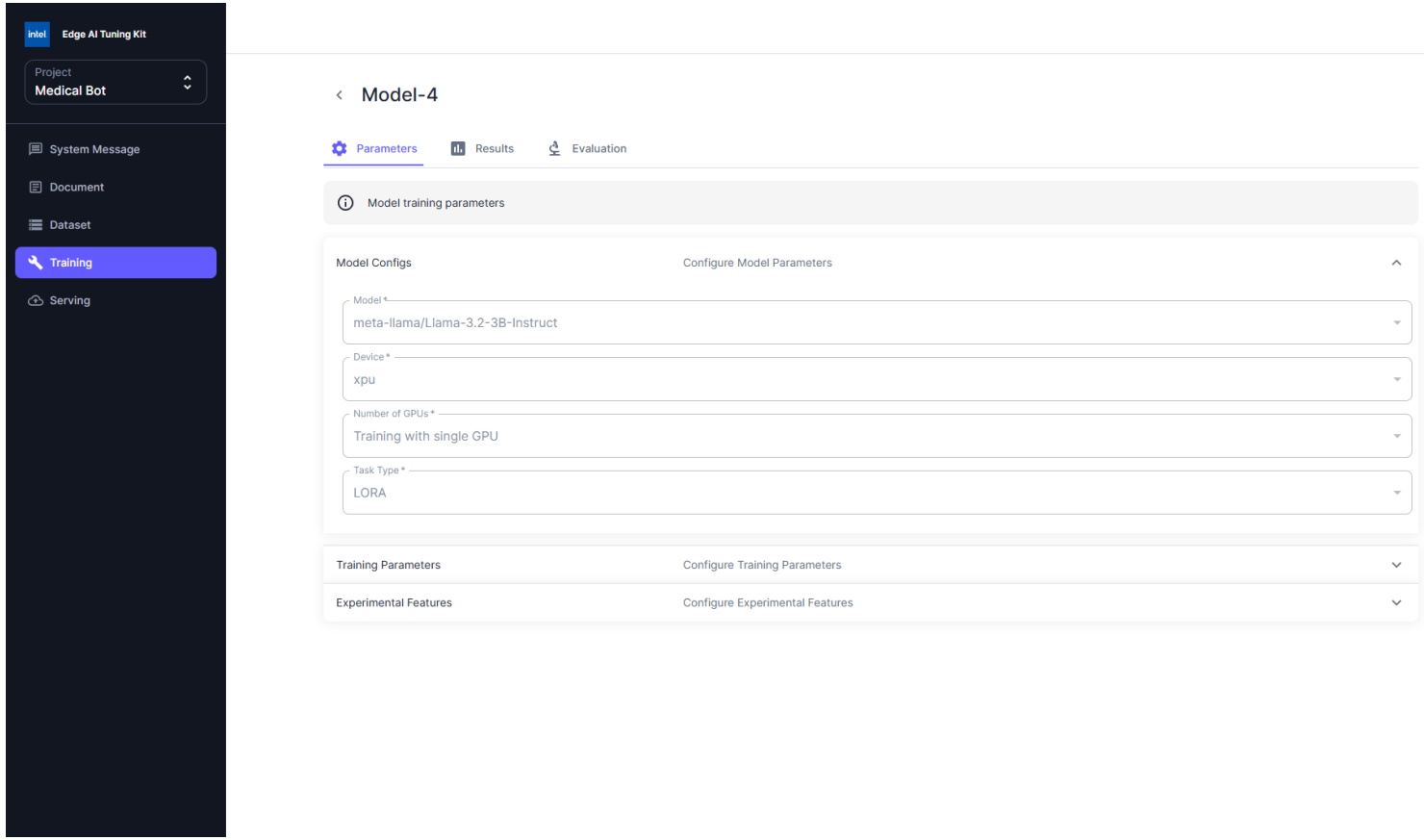
Training Loss	Training loss is a measure of how well a model is performing in the training
Evaluation Loss	Evaluation loss is a measure of how well a model is performing in the evaluation
Learning Rate	The learning rate is a hyperparameter that controls how much the model's weights are adjusted during training in response to the calculated training loss
Train/Evaluation Accuracy	Percentage of correct predictions made by the model on the dataset over time

Training



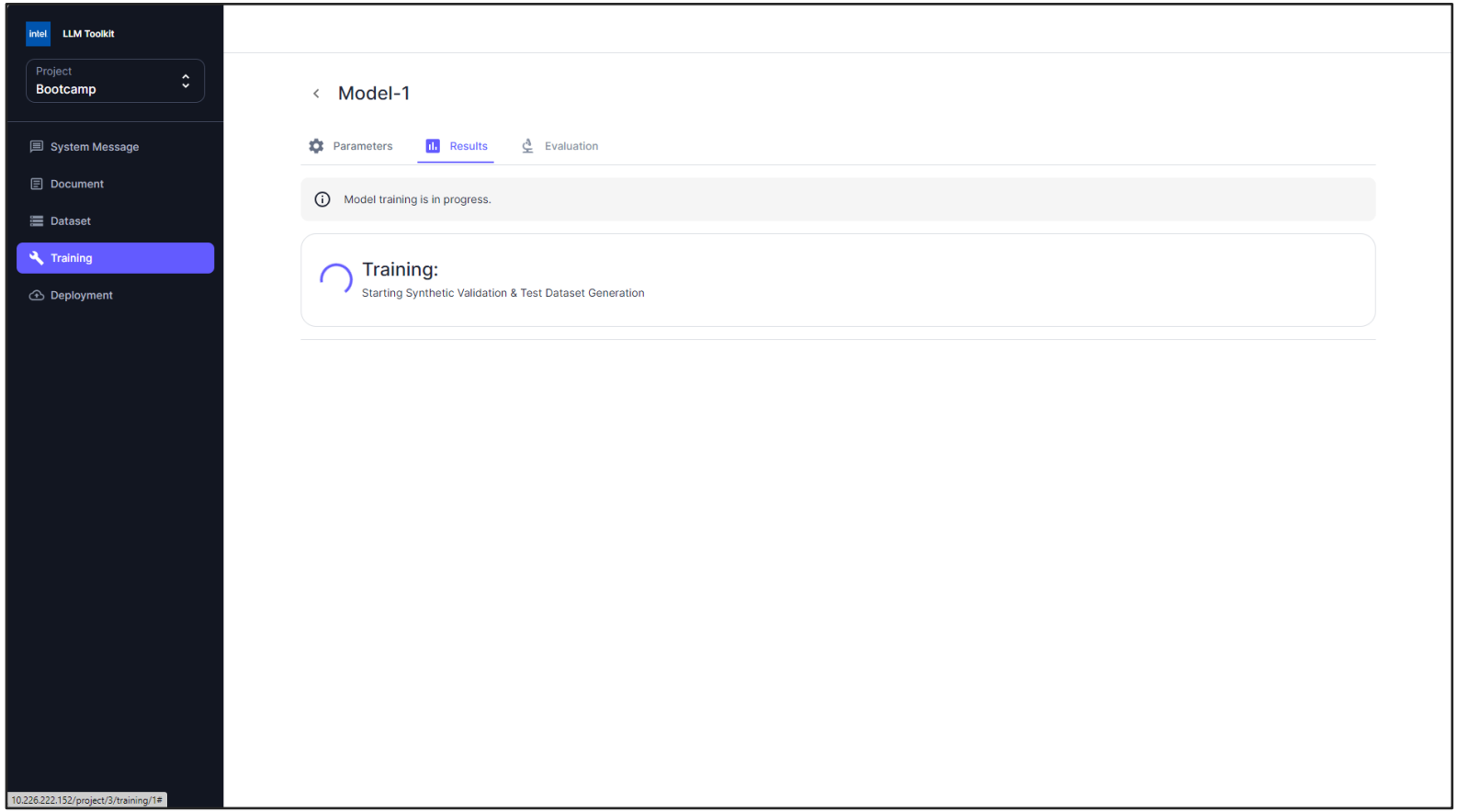
1. User will see a new training process has been initiated in the Training home page.
2. Click on the training process for more information.

Training



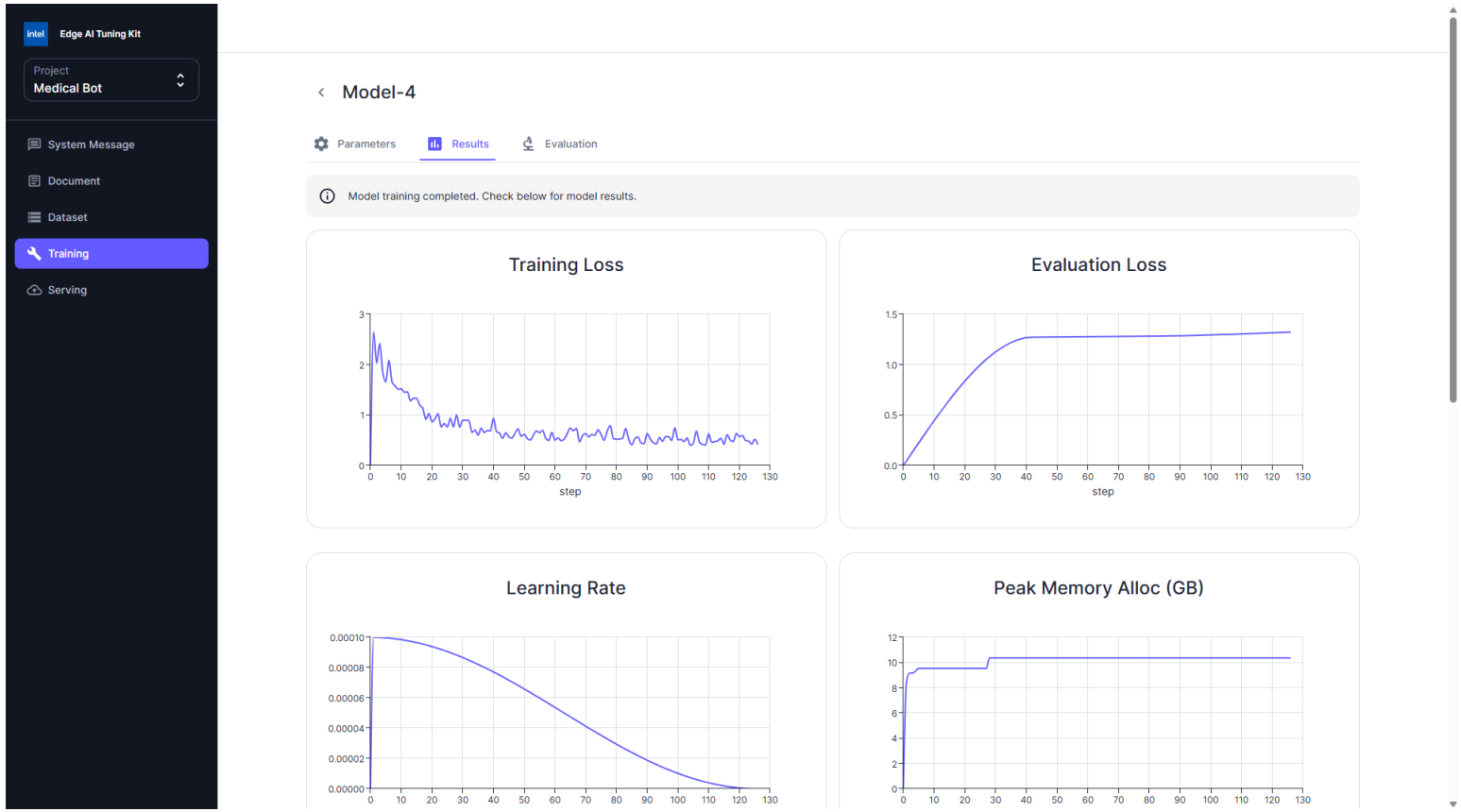
The user can view the parameters used to train the model. This helps to determine the effect of parameters on the results of the model

Training



Each step of the progress for dataset augmentation, model training, and model evaluation will be shown on the page.

Training



Results from the model training can be viewed in the Results tab

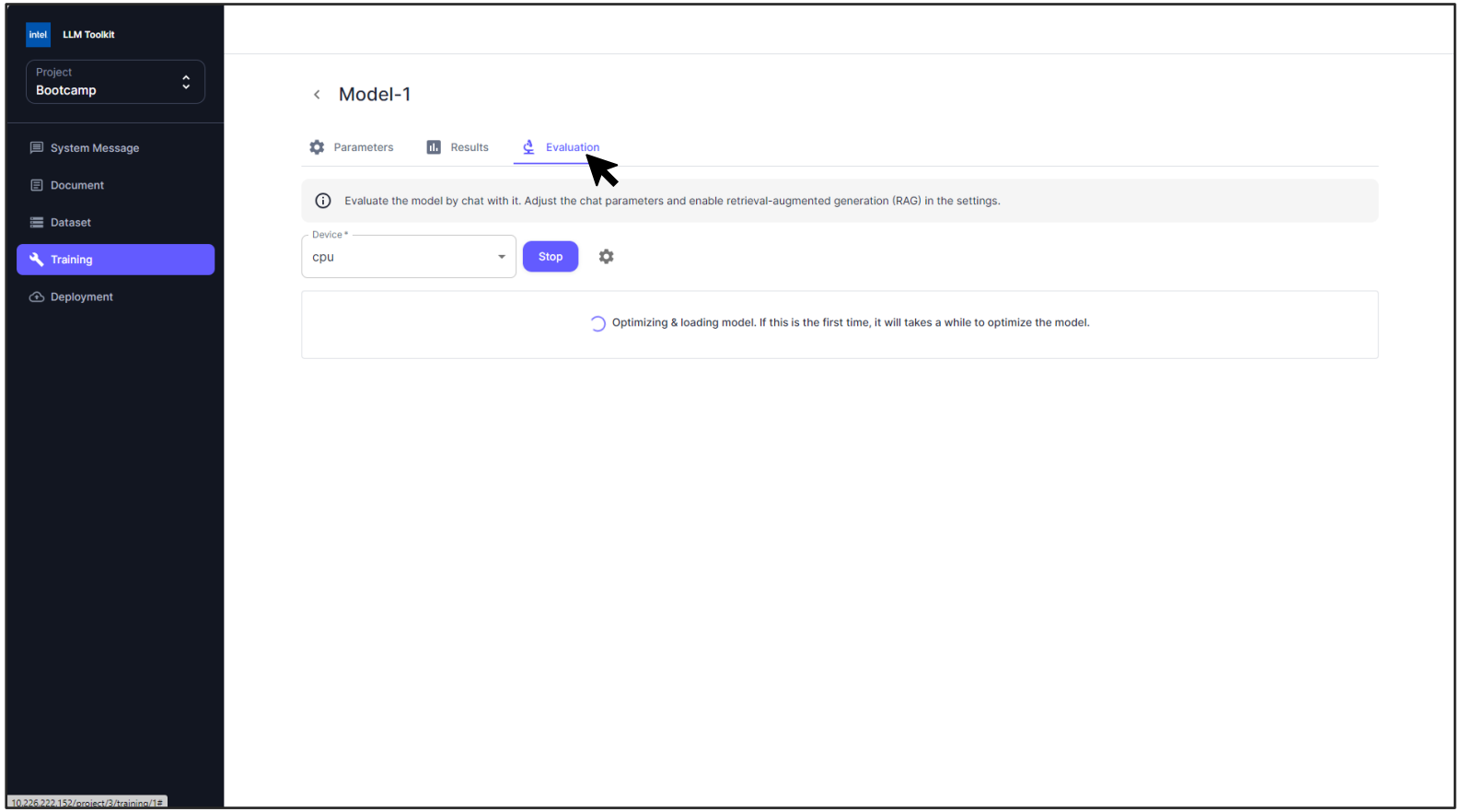
Evaluation

Wait for the process to ready

Chat with chatbot to determine if the finetuning is done correctly

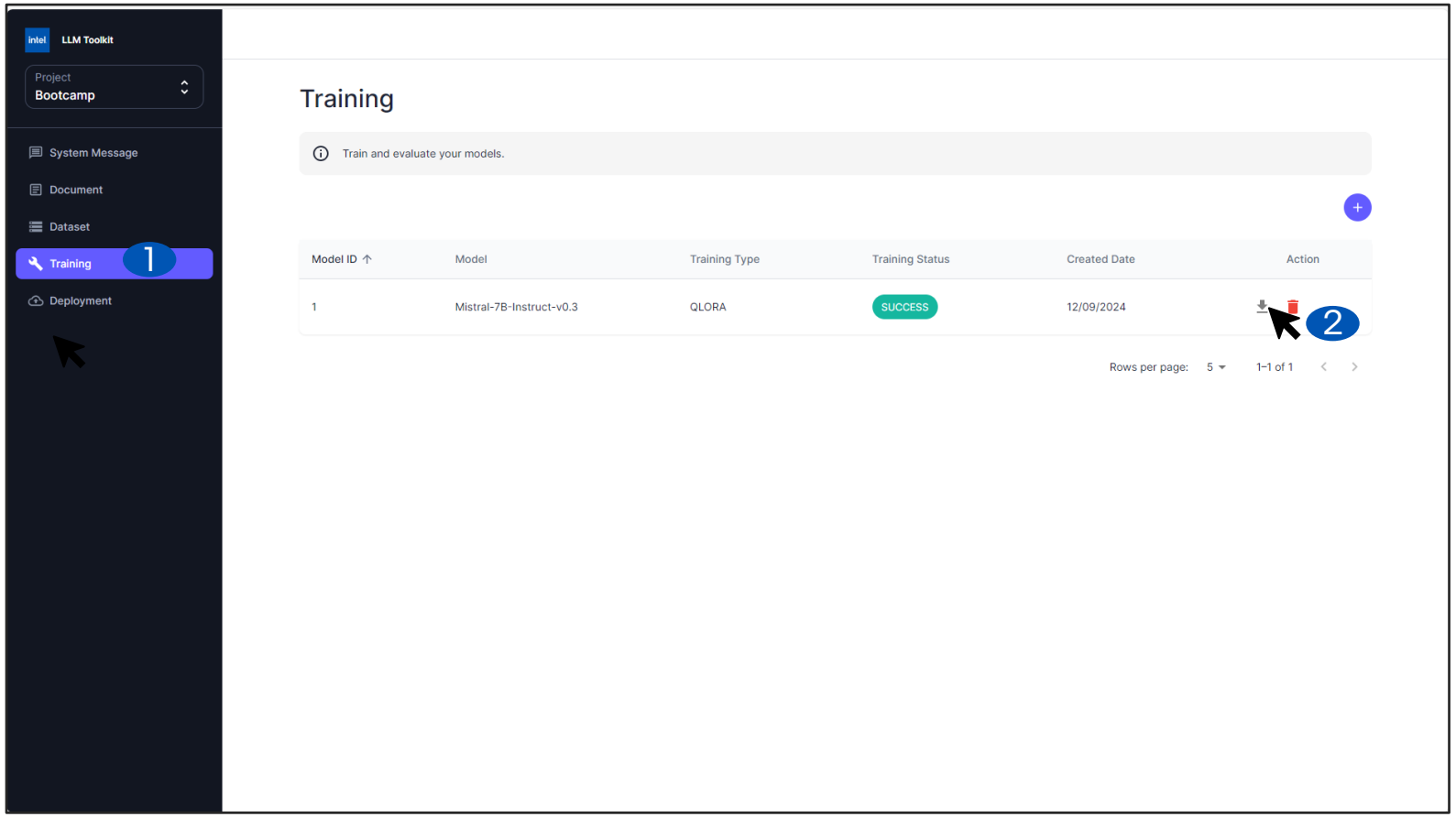
Ask any question to verify the chatbot act as desired.

Evaluation



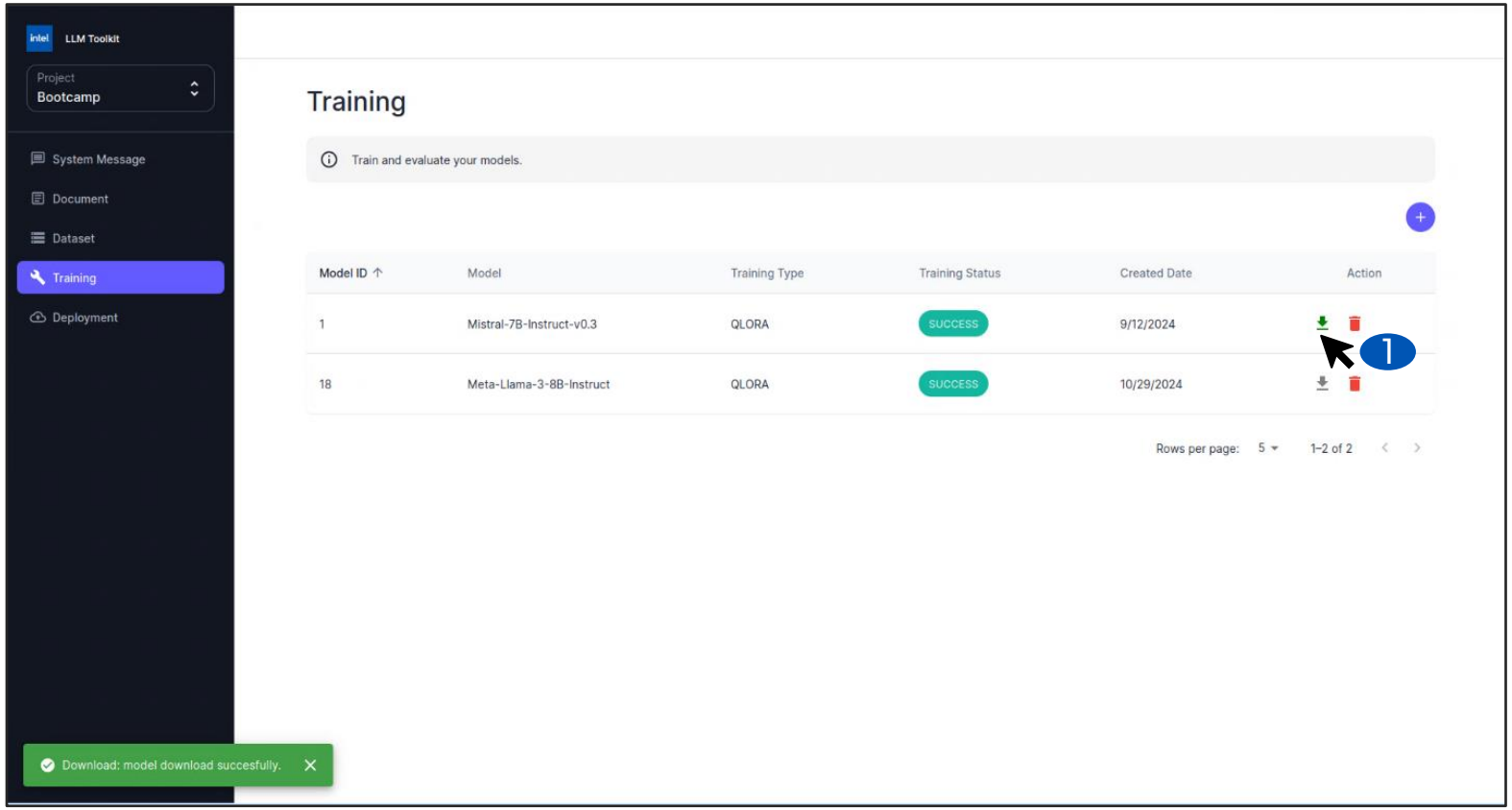
- Users can evaluate the model by using the chat window in the Evaluation tab.
- The evaluation is powered by VLLM with OpenVINO backend.
- This is to ensure a seamless experience in the deployment environment.

Download model for deployment



1. After completing the evaluation, the user can return to the interface displayed in the figure by selecting the corresponding training item on the left panel.
2. To initiate the download, the user should click the download icon indicated in the figure.
3. This action will trigger the process of preparing the model for download.

Download model for deployment



1. Once the download icon turns green, click it again to download the fine-tuned model to your local machine.
2. The downloaded model can then be installed and the fine-tuned chatbot deployed on an edge device.

LLM Inference Toolkit

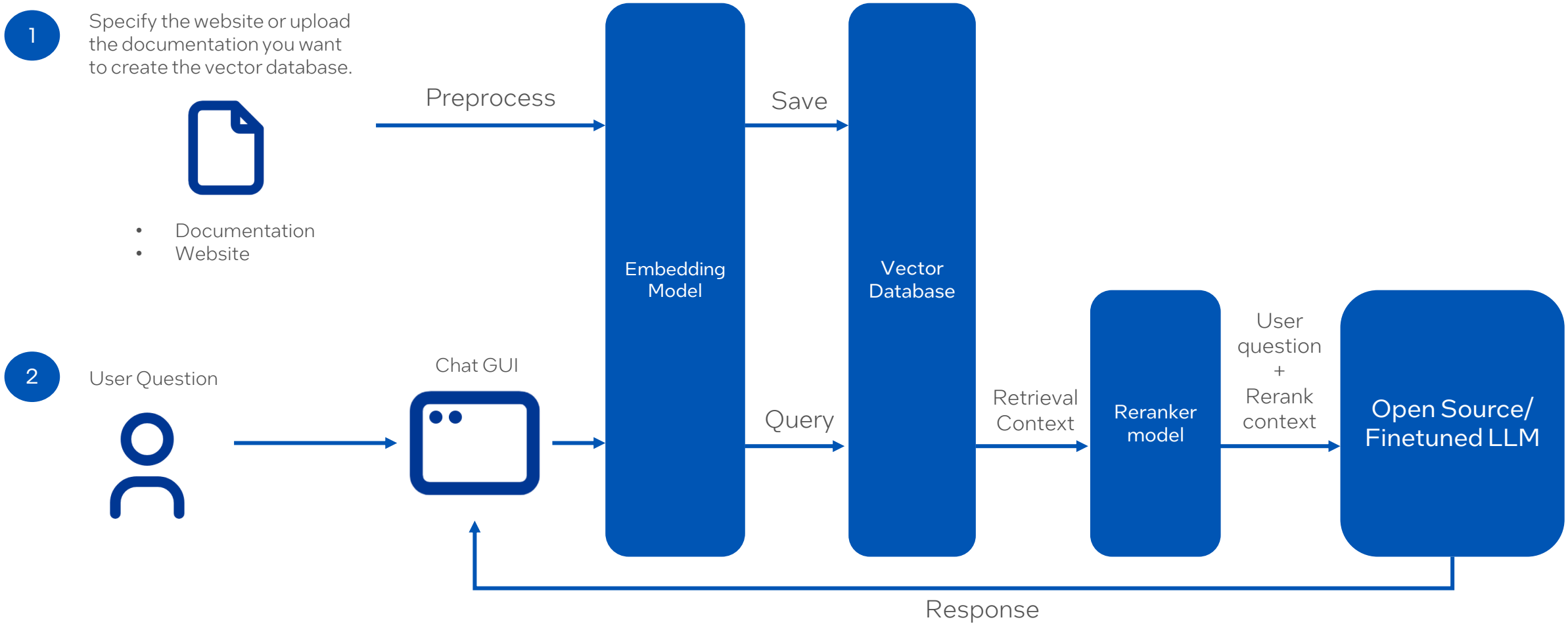
Support chat use case with finetuned LLM

Support RAG functionality

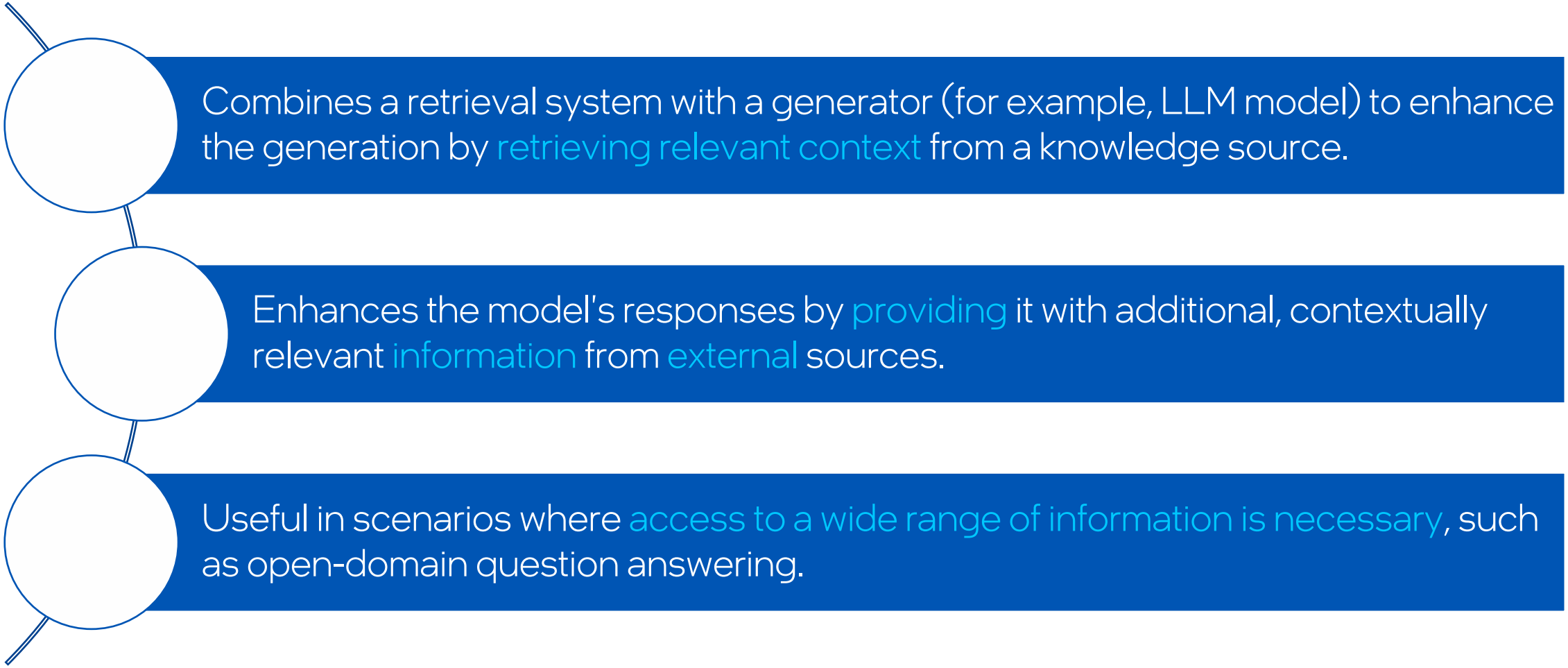
Text to speech support

Speech to text support

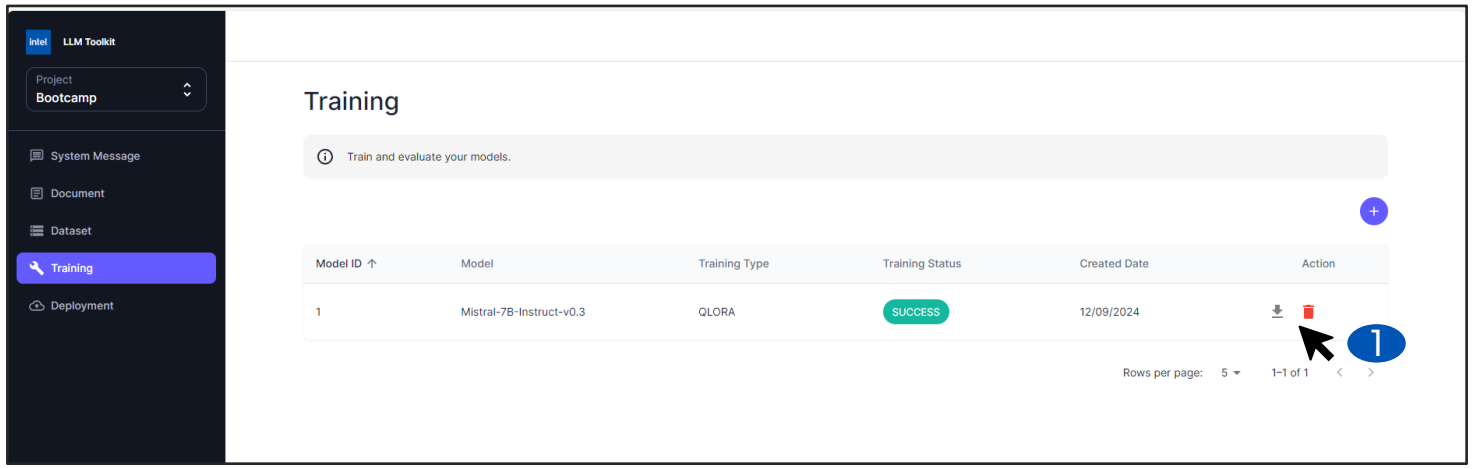
Retrieval Augmented Generation With Your Private Data



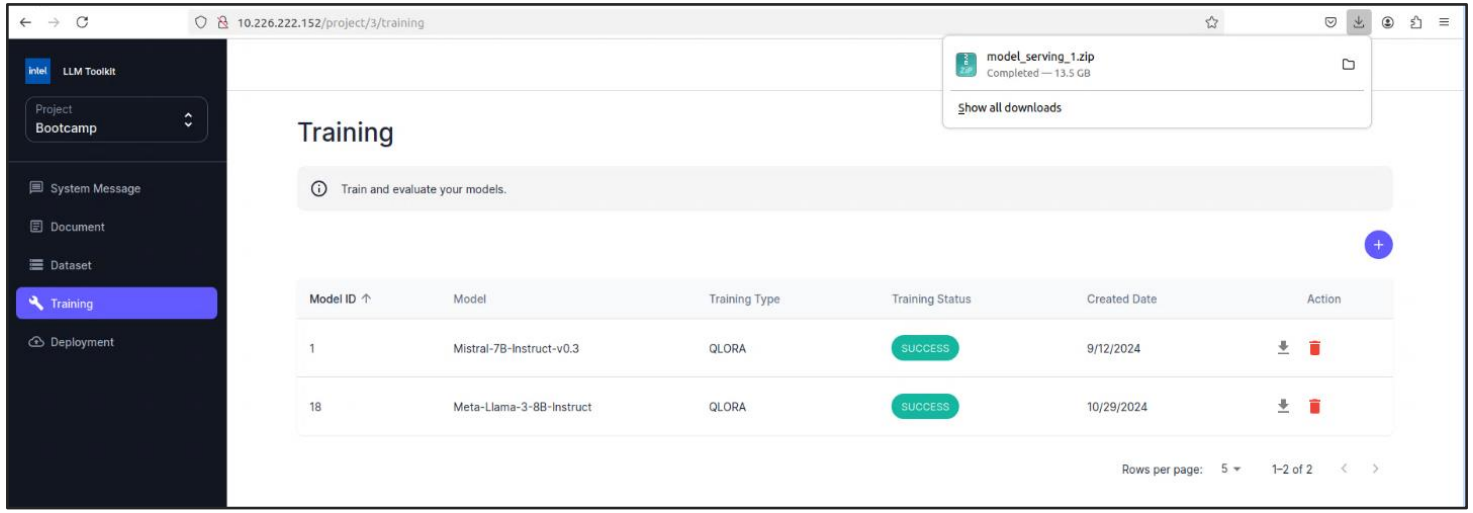
What is Retrieval Augmented Generation?



Download model for deployment



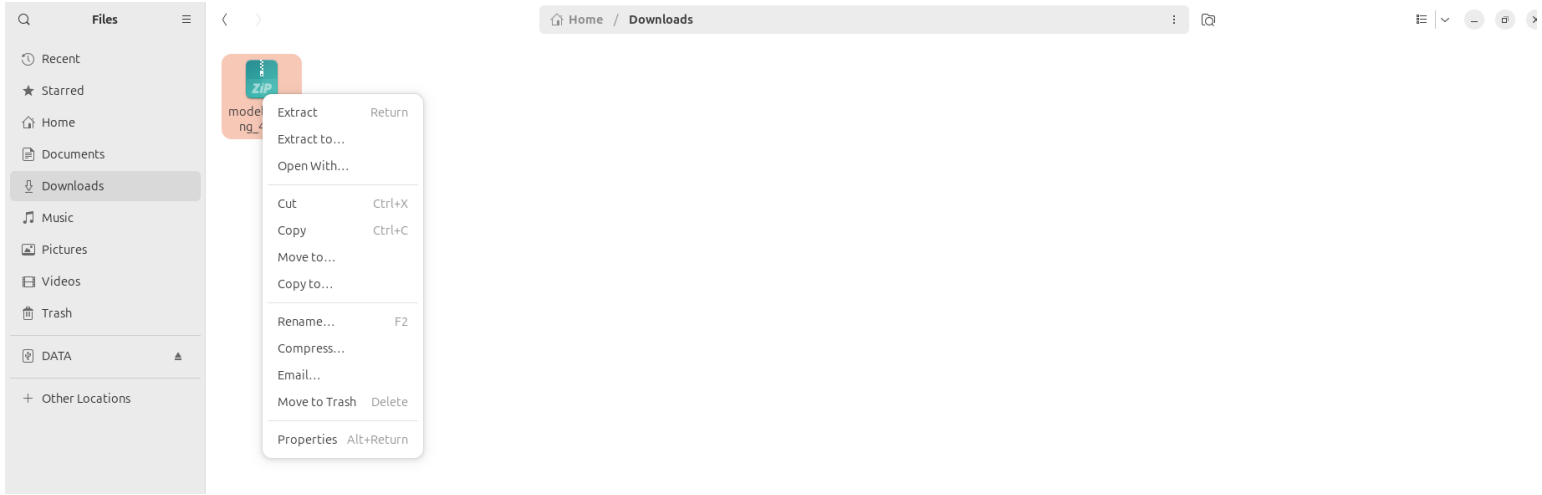
Click on the download button

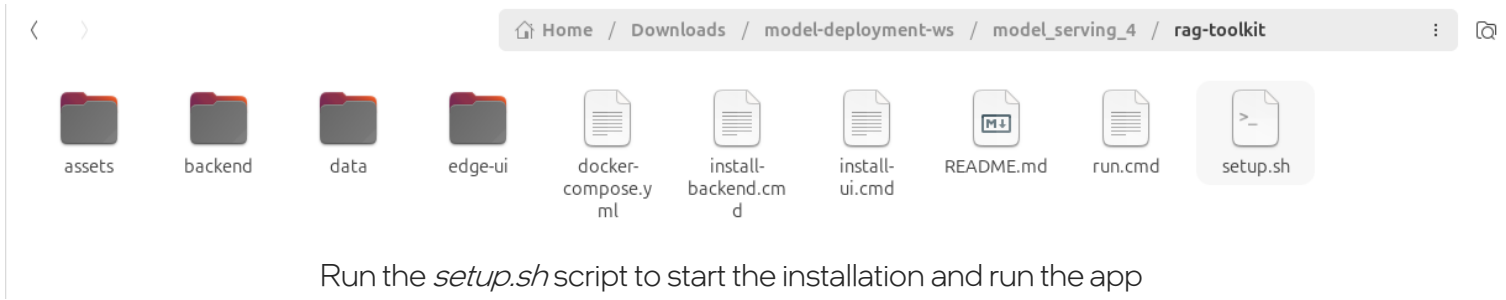


Model bundle download successfully

1. After completing the evaluation, users can return to the interface displayed in the figure by selecting the training option on the left.
2. Upon the first click, a bundle is created in the background, and once it's ready for download, the button will turn green, indicating that users can proceed to download the bundle.
3. Next, users can download the model by clicking the download icon shown in the figure.
4. Once the file is downloaded, users can install it and deploy the fine-tuned chatbot on an edge device.

Setting Up & Run the finetuned model on AI PC

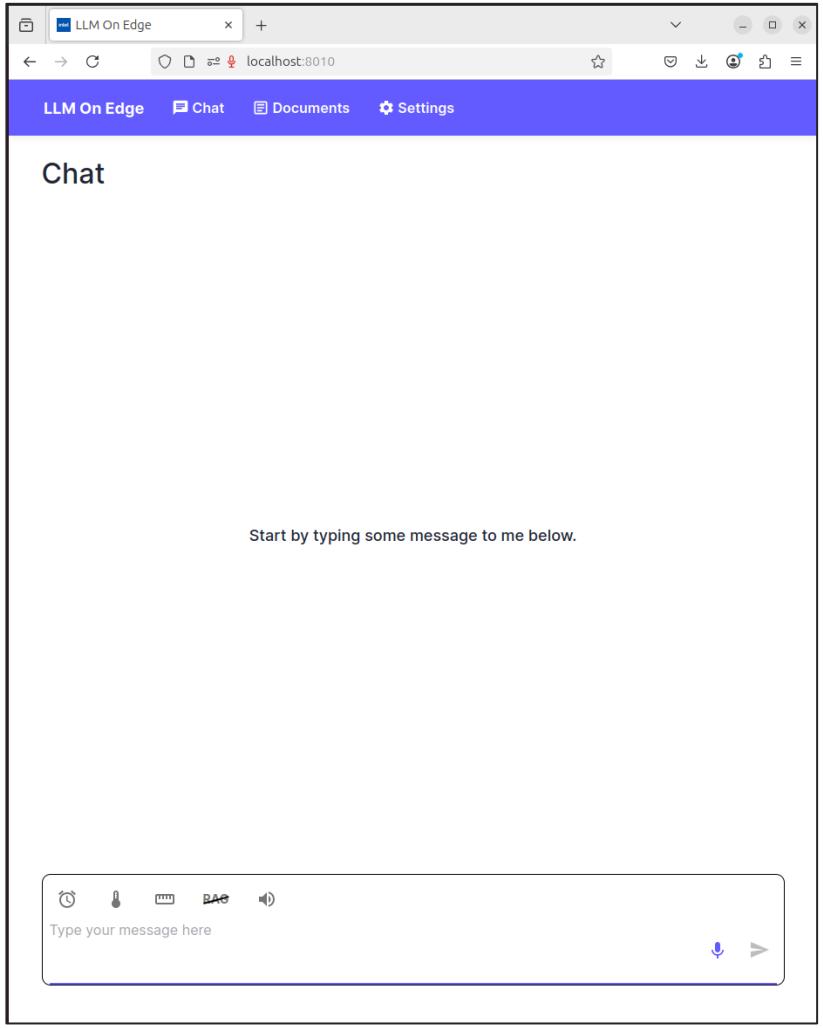




Run the *setup.sh* script to start the installation and run the app

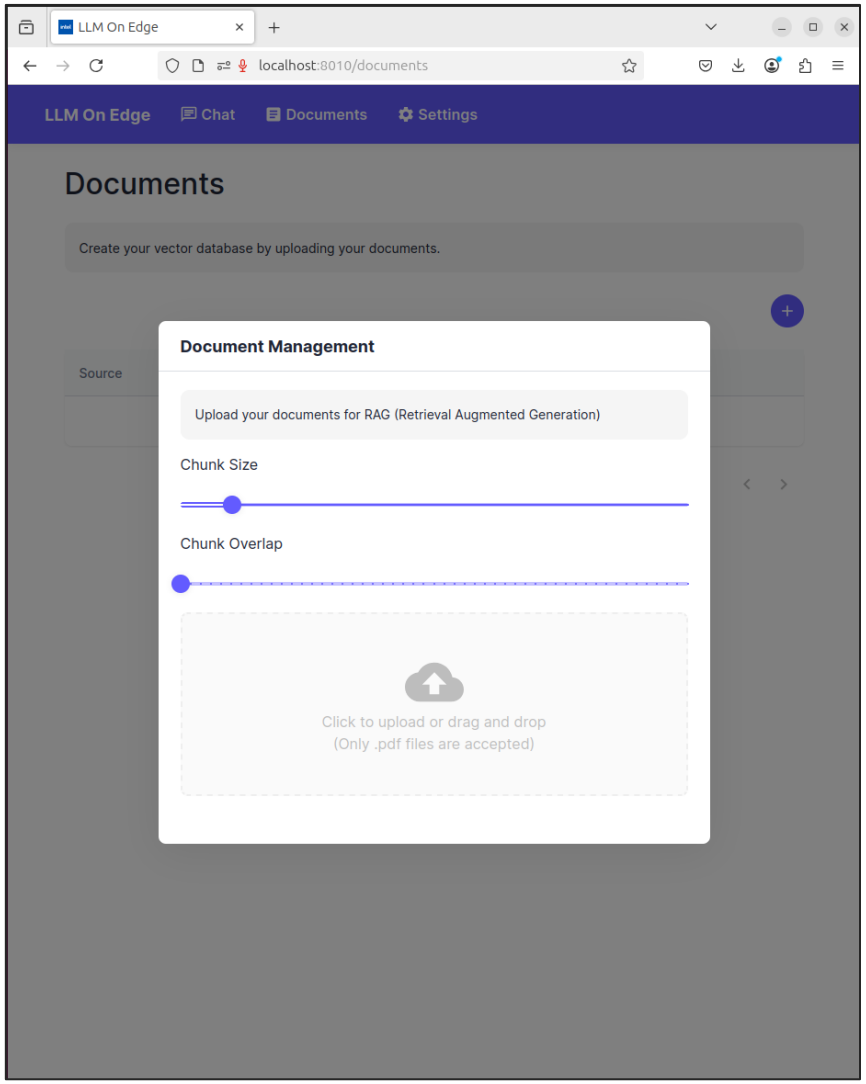
1. Locate the model bundle zip file.
2. Unzip the downloaded file.
3. Go to the *rag-toolkit* folder.
4. Run the *setup.sh* script. This will start the installation and start the application once the installation is completed.
5. For more information, you can check the README.md file in the *rag-toolkit* folder.
6. Browse to the http://localhost:8010

Start a Chat Session with the Finetuned Model



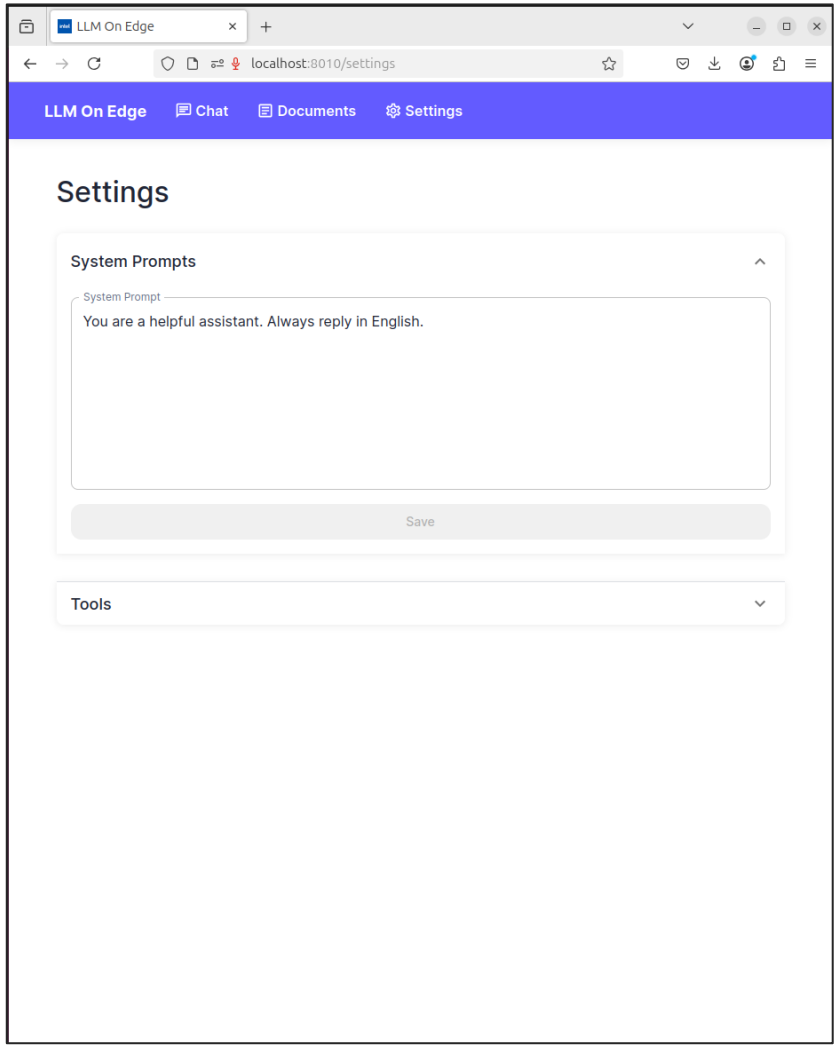
1. Users can type in the text input below to chat with the fine-tuned model.
2. Text-to-speech functionality is enabled by default.
3. Users can experience speech-to-text by clicking the microphone button below.

Upload document to enable RAG feature



1. Users can upload the document to enable RAG functionality.
2. Toggle the RAG button to enable RAG during the chat session.

Modifying System Message to use for Chat



1. Users can modify the system message on the interface.
2. This allows users to test and change different system messages during the runtime.


intel