

# Fine-Tuning and Data Evaluation

NCHC LLM DevOps  
Engineering Assistant  
Yenyun Chen

6/18/2025

# Outline

- Introduction of Pipeline Services
- Usage of Pipelines
- Hand-on Practice



# Introduction of Pipeline Services

---

- Web UI
- Operating data and model on Git server

01 Data Automation

---

02 LLM Training

---

03 Model Evaluation

# Data Automation

01

## Data Generation

Generate dataset from seed dataset.

02

## Data Evaluation

Score the quality of dataset.

# Data Generation

Expand on minimal data

## Pre Processing

Convert EXCEL, JSON, JSONL to compatible format.

## Data Generation

Generate new data based on seed data with LLM.

## Post Processing

Convert format, remove some metadata.

## Data Distillation

Remove bad data, such as misspelling, not answering the question, broken grammar etc.

## Data Convert

Convert to training ready format.

# Data Evaluation

## **Typo-Free Score**

- Detect errors like misspelled or misused words, mixing different languages, terms not used in Taiwan. Higher when less errors detected.

## **Perplexity (PPL) Score**

- Lower when sentences are less fluent, with issues like wrong homophones or word order errors.

## **Diversity Score & Redundant Score**

- Greater global diversity and lower repetition in the dataset scored higher.
- Calculated with cosine distance.

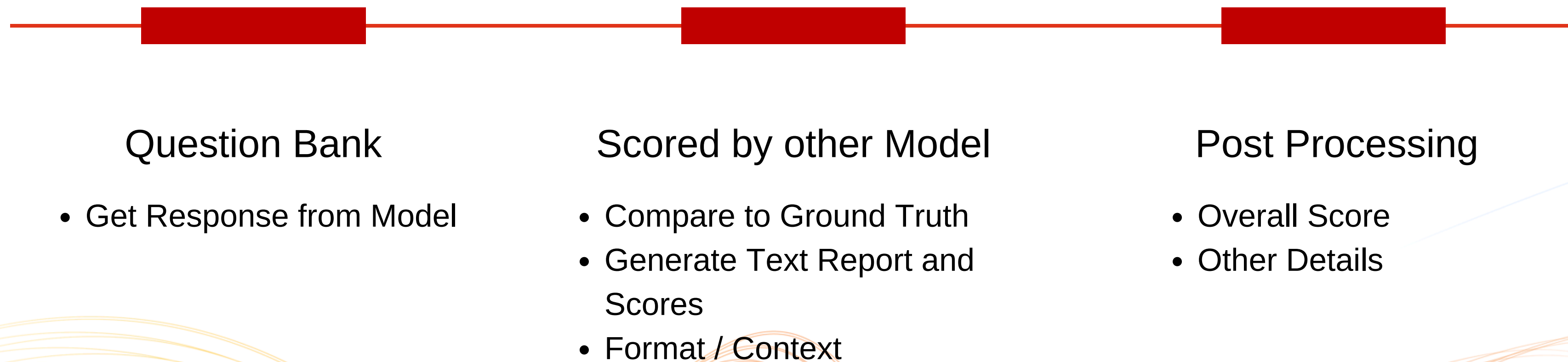


# Fine Tuning Pipeline

- Parameters optimized for hardwares to avoid OOM or other problems.
- Automatically converting model format to safetensors.
- Model files saved to Git repository.

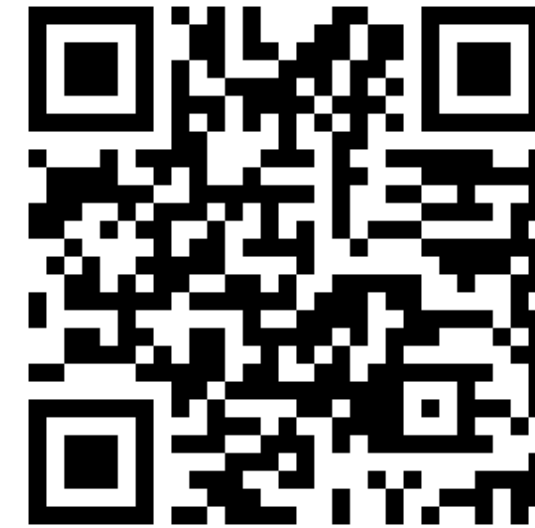


# Model Evaluation





# Usage of Pipelines



## Login with iService account

---

<https://jenkins.genai.nchc.org.tw/>

## Select pipeline

---

- 01-data-automation:
  - 01-data-generation-NCHC
  - 02-data-evaluation
- 02-llm-training
- 03-model-evaluation

## Run pipeline

---

- Click “Build with Parameters”
- Fill parameters
- Click “Build” and wait

# Preparation

## 01

### Git Repository:

- GitLab (Recommended)
  - HuggingFace (Each file under 10 MiB)
  - GitHub (Each file under 2GiB/5GiB)
- 

## 02

### LLM Access:

- Portal
- OpenAI (except data-evaluation)
- OpenAI Compatible (except data-evaluation)

# Hand-on Practice



# Data Automation

Path: Dashboard > 01-data-automation > **01-data-generation-NCHC**

**Git\_REPO\_URL:** GitLab/GitHub/HuggingFace repository url, with “http(s)://”

**GIT\_REPO\_TOKEN:** Token with R/W permission to the content of repository

- GitLab: Personal access tokens / Project access token
- GitHub: Settings > Developer Settings > Personal access tokens > Fine-grained tokens / Tokens (classic)
- HuggingFace: Access Tokens

**DATA\_FILE:** Path to data, relative to the root directory of the repository

**SHEET\_NAME:** If using EXCEL, specify which sheet contains the data

# Data Automation

Path: Dashboard > 01-data-automation > **01-data-generation-NCHC**

**GEN\_MODEL:** Model name that you want to use to generate data, depends on your LLM provider.

- *Usually, can get from \$API\_BASE/models endpoint.*

**GEN\_API\_URL:** URL of your LLM provider, can be any OpenAI compatible.

**GEN\_API\_KEY:** API key that has permission to use your GEN\_MODEL

**DISTILLATION\_MODEL:** Model name that you want to use for distillation

**DISTILLATION\_API\_URL, DISTILLATION\_API\_KEY:** As **GEN**

# Data Automation

Path: Dashboard > 01-data-automation > **01-data-generation-NCHC**

**TASK:** Choose what kind of data you need

**TOPIC:** If the options in **TASK** do not meet your need, assign other topic

**Q\_COL:** The key or column name of question/user input field in your data

**A\_COL:** The key or column name of answer/assistant output field in you data

**DEFAULT\_COUNT:** How many pairs you want to generate from a seed pair

**SAMPLE:** How many data you want to sample as seed data, set to 0 for use all

**SYSTEM\_MSG:** Instruction for data generation



# Data Automation

Path: Dashboard > 01-data-automation > **01-data-generation-NCHC**

**DO\_DEDUP:** Do deduplication or not

**SIMILARITY\_THRESHOLD:** Remove data if the similarity higher than this number

**DO\_DISTILLATION:** Do distillation or not

# Data Automation

Path: Dashboard > 01-data-automation > **02-data-evaluation**

**REDUNDANCY\_THRESHOLD:** The value for calculating redundancy

# LLM Training

Path: Dashboard > **02-llm training**

**BASE\_MODEL:** Base model to train

**MAX\_EPOCHS:** How many round to train

**DEEPSPEED\_ZERO\_STAGE:** Level of optimization, speed vs memory consuming

**MAX\_MODEL\_LENGTH:** Max content length, not longer than base model

**MODEL\_CONFIG\_TORCH\_DTYPE:** Percision of trained model, depends on GPU,

- v100 supports fp16/32

**HARDWARE\_TYPE:** Model of GPU, recently only v100

**GPU\_COUNTS:** How many GPU to use in this training session

**EMAIL\_NOTIFY:** True for email notification on pipeline finishing



# Model Evaluation

Path: Dashboard > **03-model-evaluation**

**GEN\_MODEL\_SOURCE:** Choose the source of the model you want to evaluate, can be OpenAI, OpenAI compatible or NCHC provided.

**GEN\_MODEL, GEN\_BASE\_URL, GEN\_API\_KEY:** Info to use model

**JUDGE\_MODEL\_SOURCE:** Choose the source of the model as the judge

**JUDGE\_MODEL, JUDGE\_MODEL\_URL, JUDGE\_MODEL\_KEY:** Info to use model

**EVAL\_ITER:** How many round to score

**TASKS:** The tasks to evaluate, depends on the purpose of using model

# Model Evaluation

Path: Dashboard > **03-model-evaluation**

**MAX\_NEW\_TOKENS:** Max length for model's answer

**BATCH\_SIZE:** Max number of async client interaction with LLM

**NUM\_ROWS:** How many rows to evaluate

# Example Data Format

<https://gitlab.td.nchc.org.tw/baronhsu/llm-bootcamp-0618>





# THANK YOU