An end-to-end LLM engineering platform for fine-tuning, evaluation and registration of custom models and adapters

Evangelos Dellis, Cahya Wirawan

Preparatory Commission for the Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO),
P.O. Box 1200, 1400 Vienna, Austria

CTBTO PREPARATORY COMMISSION
PUTTING AN END TO NUCLEAR EXPLOSIONS

INTRODUCTION AND MAIN RESULTS

We present an end-to-end LLM engineering platform for fine-tuning, evaluation and registration of custom models and adapters that is built on top of open-source tools. The versatility of our platform is demonstrated through various applications such as fine-tuning multimodal open-source LLMs on custom datasets for increased accuracy.

Our platform aims to accelerate the development of custom models and adapters, enabling a wide range of innovative applications across CTBTO's technologies. These adapters consist of small collections of model weights that can be dynamically loaded onto a common base LLM, enabling it to specialize itself on-the-fly for specific tasks.

DISCLAIMER: The views expressed on this e-poster are those of the author and do not necessarily reflect the view of the CTBTO.

SnT 2025
8 SEPTEMBER ONLINE DAY
9 TO 12 SEPTEMBER AT HOFBURG PALACE, VIENNA & ONLINE
CTBT: SCIENCE AND TECHNOLOGY CONFERENCE

An end-to-end LLM engineering platform for fine-tuning, evaluation and registration of custom models and adapters

Evangelos Dellis, Cahya Wirawan

P3.5-569

Introduction

What is the end-to-end LLM engineering platform?

- A **platform** for building, evaluating, training, monitoring and configuring LLM assistants/agents
- It is based on popular **open-source** technologies/tools
- Helps with dataset creation, **data labeling** and annotation from private CTBTO data
- Streamlines the **fine-tuning**, deployment, and management of custom models and **adapters**
- Manage **LLM API access** and enforce budgets, guardrails, logging and cost tracking

Who is it for? Targeted to developers and admins.

Based on popular tools that are running on NVIDIA GPUs:

- ✓ **vLLM:** OpenAI API compatible LLM inference engine
- ✓ **Qdrant:** Vector database for similarity search
- ✓ **Airflow:** Pipeline orchestration tool
- ✓ **MLFlow:** Model registry, tracking, model wrapping
- ✓ **Langfuse:** Observability tool for traces and metrics
- ✓ **Easy Dataset:** Tool for creating fine-tuning datasets
- ✓ **Llama Factory:** LLM fine-tuning tool
- ✓ **Marker:** Convert PDF to markdown
- ✓ **DeepEval:** LLM evaluation framework

DISCLAIMER:
The views expressed on this e-poster are those of the author and do not necessarily reflect the view of the CTBTO.

LLM Engineering Platform

LLM Application Observability:

- Inspect and debug complex logs
- Ingest traces to **Langfuse**
- Track LLM calls & retrieval/embed

LLM Evaluations:

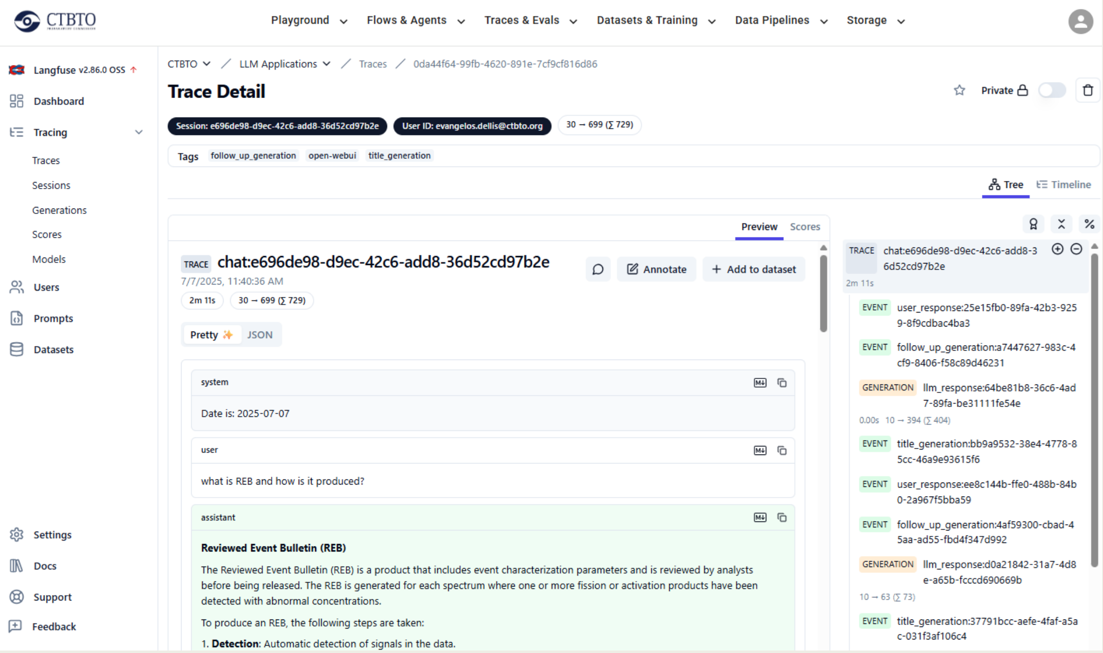
- LLM-as-a-Judge
- Manual annotations
- Custom evals via SDKs

Datasets:

- Test sets & benchmarks
- Structured experiments

Prompt Management:

- Manage & Version Control
- Collaborate on prompts



CTBTO / LLM Applications / Traces / 0da44f64-99fb-4620-891e-7cf9cf816d86

Trace Detail

Session: e696de98-d9ec-42c6-add8-36d52cd97b2e User ID: evangelos.dellis@ctbto.org 30 → 699 (∑ 729)

Tags follow_up_generation open-webui title_generation

TRACE chat:e696de98-d9ec-42c6-add8-36d52cd97b2e
7/7/2025, 11:40:36 AM
2m 11s 30 → 699 (∑ 729)

system

Date is: 2025-07-07

user

what is REB and how is it produced?

assistant

Reviewed Event Bulletin (REB)

The Reviewed Event Bulletin (REB) is a product that includes event characterization parameters and is reviewed by analysts before being released. The REB is generated for each spectrum where one or more fission or activation products have been detected with abnormal concentrations.

To produce an REB, the following steps are taken:

1. **Detection**: Automatic detection of signals in the data.

TRACE chat:e696de98-d9ec-42c6-add8-36d52cd97b2e
2m 11s

EVENT user_response:25e15fb0-89fa-42b3-9259-8f9cdbac4ba3

EVENT follow_up_generation:a7447627-983c-4cf9-8406-f58c89d46231

GENERATION llm_response:64be81b8-36c6-4ad7-89fa-be31111fe54e
0.00s 10 → 394 (∑ 404)

EVENT title_generation:bb9a9532-38e4-4778-85cc-46a9e93615f6

EVENT user_response:ee8c144b-ffe0-488b-84b0-2a967f5bba59

EVENT follow_up_generation:4af59300-cbad-45aa-ad55-fbd4f347d992

GENERATION llm_response:d0a21842-31a7-4d8e-a65b-fcccd690669b
10 → 63 (∑ 73)

EVENT title_generation:37791bcc-aefe-4faf-a5ac-031f3af106c4

Evaluation is crucial for ensuring the quality and reliability of your LLM applications. **DeepEval** is an open-source framework and platform for evaluating, testing, and monitoring LLM applications:

AI Assistant and Agent metrics: Task Completion, Tool Correctness, Conversation Completeness
RAG metrics: Answer relevancy, faithfulness, hallucination, and safety tests.
Safety and Bias tests: Detects Toxicity, Bias, and other vulnerabilities
DeepEval features: Integrates with CI/CD, offers synthetic dataset generation and many more.

CTBTO PREPARATORY COMMISSION
PUTTING AN END TO NUCLEAR EXPLOSIONS

An end-to-end LLM engineering platform for fine-tuning, evaluation and registration of custom models and adapters

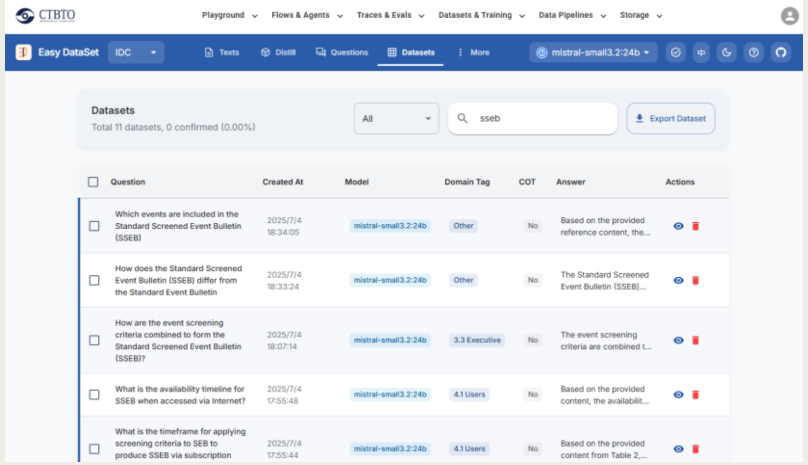
Evangelos Dellis, Cahya Wirawan

P3.5-569

Creating fine-tuning datasets

Easy Dataset is a unified **framework** for synthesizing fine-tuning data from unstructured documents and it consists of two primary components:

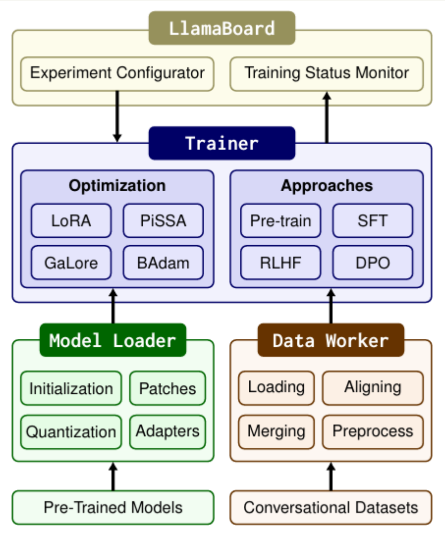
- **Adaptive document processing**: documents of different formats are processed via model-based parsing, followed by hybrid chunking to produce text chunks.
- **Data synthesis:** pairs are created for each document to guide the construction of diverse question-answer pairs. Questions are then generated to further increase diversity, and final augmented QA pairs are synthesized via knowledge-enhanced prompting to ensure factual consistency.

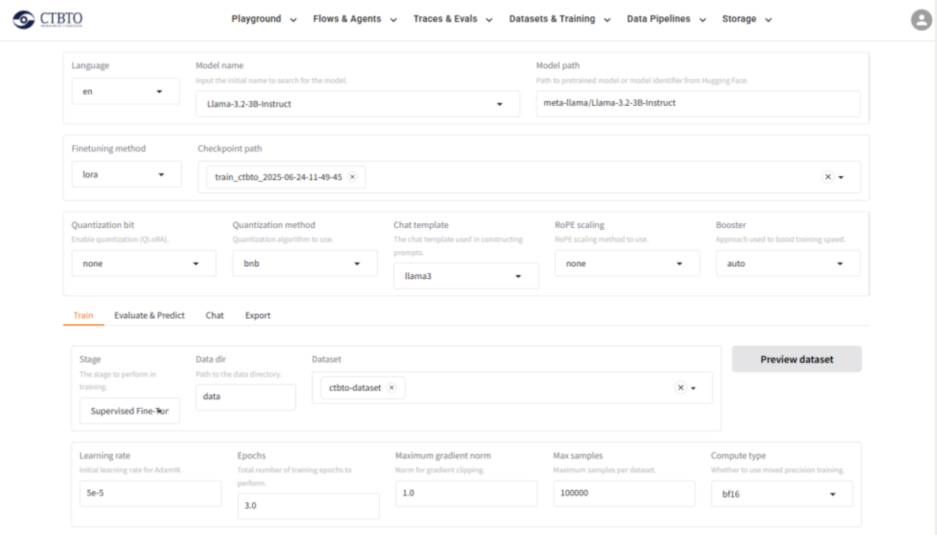


Datasets
Total 11 datasets, 0 confirmed (0.00%)

Question	Created At	Model	Domain Tag	COT	Answer	Actions
Which events are included in the Standard Screened Event Bulletin (SSEB)	2025/7/4 18:34:05	mistral-small3.2:24b	Other	No	Based on the provided reference content, the...	
How does the Standard Screened Event Bulletin (SSEB) differ from the Standard Event Bulletin	2025/7/4 18:33:24	mistral-small3.2:24b	Other	No	The Standard Screened Event Bulletin (SSEB)...	
How are the event screening criteria combined to form the Standard Screened Event Bulletin (SSEB)?	2025/7/4 18:07:14	mistral-small3.2:24b	3.3 Executive	No	The event screening criteria are combined t...	
What is the availability timeline for SSEB when accessed via Internet?	2025/7/4 17:55:48	mistral-small3.2:24b	4.1 Users	No	Based on the provided content, the availabilit...	
What is the timeframe for applying screening criteria to SEB to produce SSEB via subscription	2025/7/4 17:55:44	mistral-small3.2:24b	4.1 Users	No	Based on the provided content from Table 2,...	

LLM Fine-tuning using SFT





Language: en

Model name (Input the initial name to search for the model.): Llama-3.2-3B-Instruct

Model path (Path to pretrained model or model identifier from Hugging Face.): meta-llama/Llama-3.2-3B-Instruct

Finetuning method: lora

Checkpoint path: train_ctbto_2025-06-24-11-49-45

Quantization bit (Enable quantization (QLoRA).): none

Quantization method (Quantization algorithm to use.): bnb

Chat template (The chat template used in constructing prompts.): llama3

RoPE scaling (RoPE scaling method to use.): none

Booster (Approach used to boost training speed.): auto

Stage (The stage to perform in training.): Supervised Fine-Tur

Data dir (Path to the data directory.): data

Dataset: ctbto-dataset

Learning rate (Initial learning rate for AdamW.): 5e-5

Epochs (Total number of training epochs to perform.): 3.0

Maximum gradient norm (Norm for gradient clipping.): 1.0

Max samples (Maximum samples per dataset.): 100000

Compute type (Whether to use mixed precision training.): bf16

Our solution is based on **Llama Factory** which is a LLM fine-tuning tool that supports Pre-Training, Supervised Fine-Tuning, and Reward Modeling training modes.

- Various **models**: LLama, LLava, Mistral, Qwen, DeepSeek, Yi, Gemma, GPT-OSS, Phi, etc.
- Integrated **methods**: Pre-training, SFT, PPO, DPO, etc
- Scalable **resources**: LoRA and QLoRA via AQLM, AWQ, GPTQ, etc
- Advanced **algorithms**: GaLore, BAdam, APOLLO, Adam-mini, Muon, OFT, DoRA, LoRA+, LoftQ and PiSSA.
- Practical **tricks**: FlashAttention-2, Unsloth, Liger Kernel, RoPE scaling, NEFTune and rsLoRA.
- Wide **tasks**: Multi-turn dialogue, tool using, image understanding, visual grounding, video recognition, etc.
- **Experiment** monitors: LlamaBoard, TensorBoard, Wandb, MLflow, SwanLab, etc.
- Faster **inference**: OpenAI-style API, Gradio UI and CLI with vLLM worker.

DISCLAIMER:
The views expressed on this e-poster are those of the author and do not necessarily reflect the view of the CTBTO.

SnT 2025
CTBT: SCIENCE AND TECHNOLOGY CONFERENCE
8 SEPTEMBER ONLINE DAY
9 TO 12 SEPTEMBER AT HOFBURG PALACE, VIENNA & ONLINE

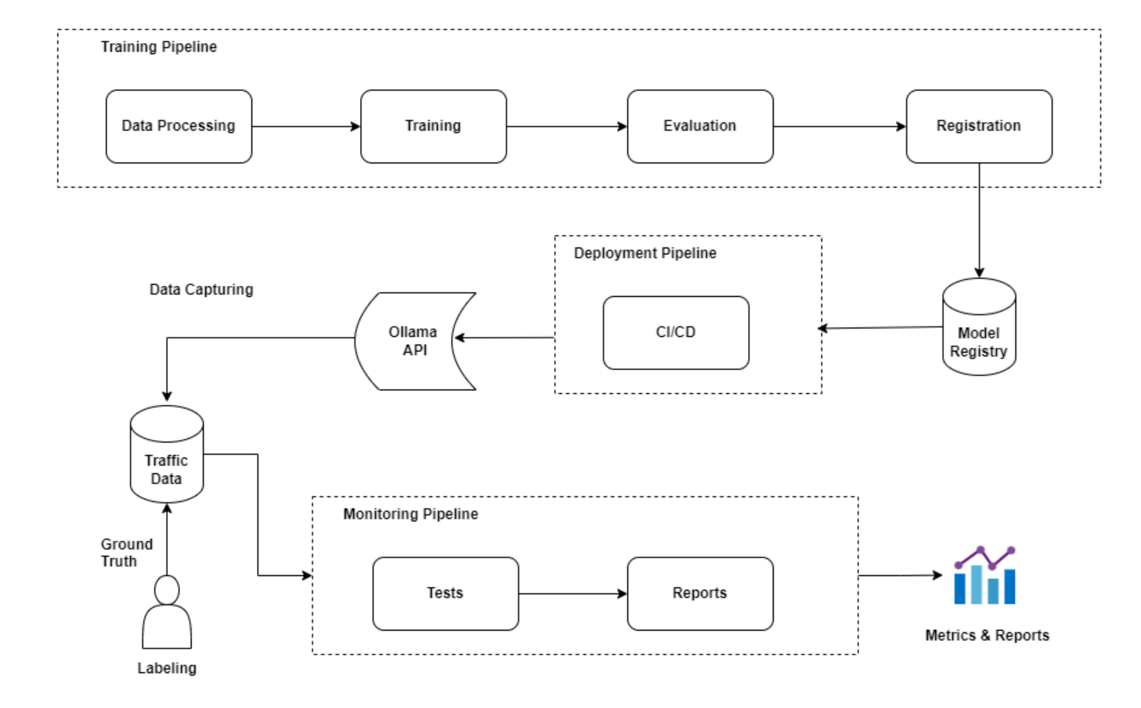
An end-to-end LLM engineering platform for fine-tuning, evaluation and registration of custom models and adapters

Evangelos Dellis, Cahya Wirawan

P3.5-569

Our contribution towards an end-to-end LLM training system

1. **Create** Dataset:
   - use Easy Dataset
2. **Train** Model:
   - use Llama Factory
3. **Evaluate & Predict:**
   - use Llama Factory
4. **Export** merged Model:
   - use Llama Factory
5. **Deploy** Model:
   - use vLLM
6. **Import** Dataset:
   - use Langfuse
7. **Run** Experiments:
   - use Langfuse
8. **Create** Model:
   - use Ollama
9. **Push** Model to Registry:
   - use Mlflow



Building a **production LLM training system** is much more complex than training a model:

What are the key challenges, i.e. how to:

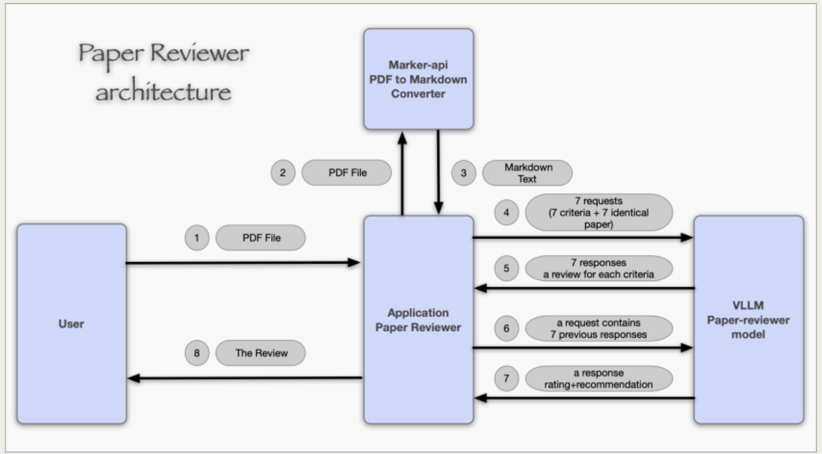
- Ingest large amounts of data, clean and preprocess that data and compute and serve features
- Set up a scalable training process, evaluate the model, track and version the data used
- Serve the model in a cost-effective manner, monitor the model and automate training and deployment

Use Case: Paper Reviewer

Source: 9000+ NeurIPS and ICLR papers with its reviews from 2023/2024 collected from OpenReview.net

Data processing:

- Review normalisation using LLM API
- Cleanup the normalized reviews
- Semi synthetic dataset generation aligned by normalized review for each criteria
  - 9000- normalized reviews becomes 70.000+ reviews. A row per criteria
  - **Llama 3.1 8B** was used to generate the new dataset for 2 days using **8x V100 GPUs**


Paper Reviewer architecture
Marker-api PDF to Markdown Converter
2 PDF File
3 Markdown Text
4 7 requests (7 criteria + 7 identical paper)
User
1 PDF File
Application Paper Reviewer
5 7 responses a review for each criteria
6 a request contains 7 previous responses
7 a response rating+recommendation
8 The Review
VLLM Paper-reviewer model

DISCLAIMER:
The views expressed on this e-poster are those of the author and do not necessarily reflect the view of the CTBTO.

CTBTO PREPARATORY COMMISSION
PUTTING AN END TO NUCLEAR EXPLOSIONS