# 用Kubernetes支撑大模型增量调优

## 从数据处理到模型评估的高效并行微调

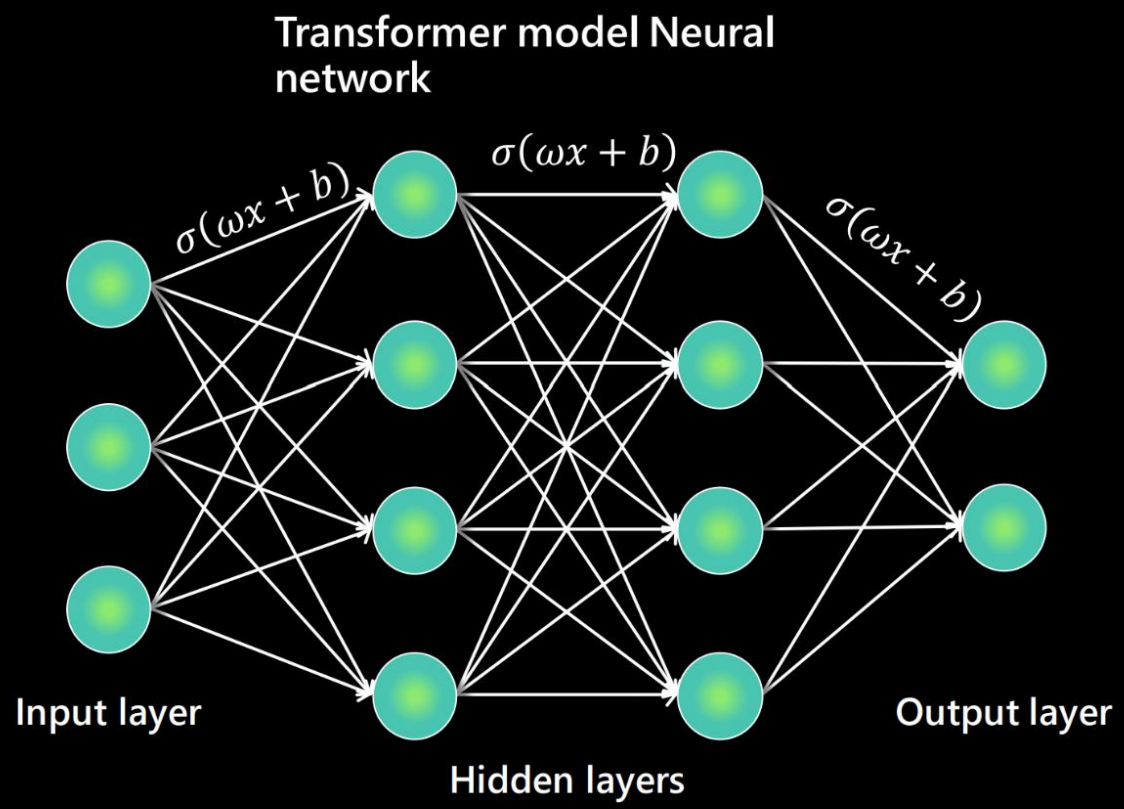
黄敏杰 DaoCloud
肃伦 DaoCloud

黄敏杰
Daocloud云原生高级研发工程师

肃伦
Daocloud云原生高级研发工程师

# 大模型LLM

## How large are they?




Function: weight * input plus bias

BERT Large - 2018
**345M**

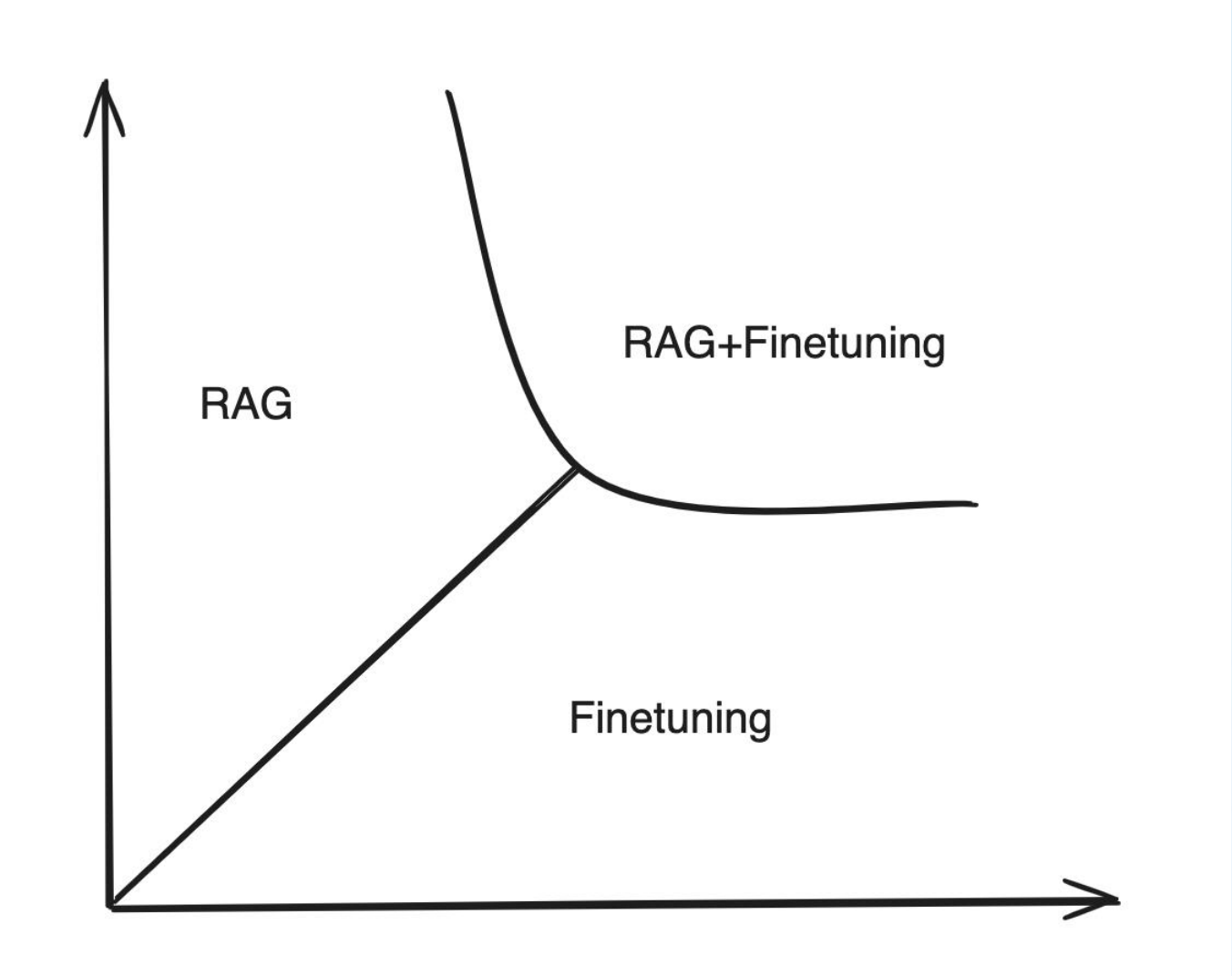
GPT2 - 2019
**1.5B**

GPT3 - 2020
**175B**

Turing Megatron NLG 2021
**530B**

GPT4 – 2023
**1.4T** ( estimated)

# 大模型落地



# 大模型微调的挑战

## 1 高质量的数据集

微调需要大量与目标领域或任务相关的高质量数据，获取和构建这样的数据集是一个挑战。

## 2 超参数微调

微调的过程非常复杂，涉及控制优化过程的各种超参数，如学习率，批大小等。

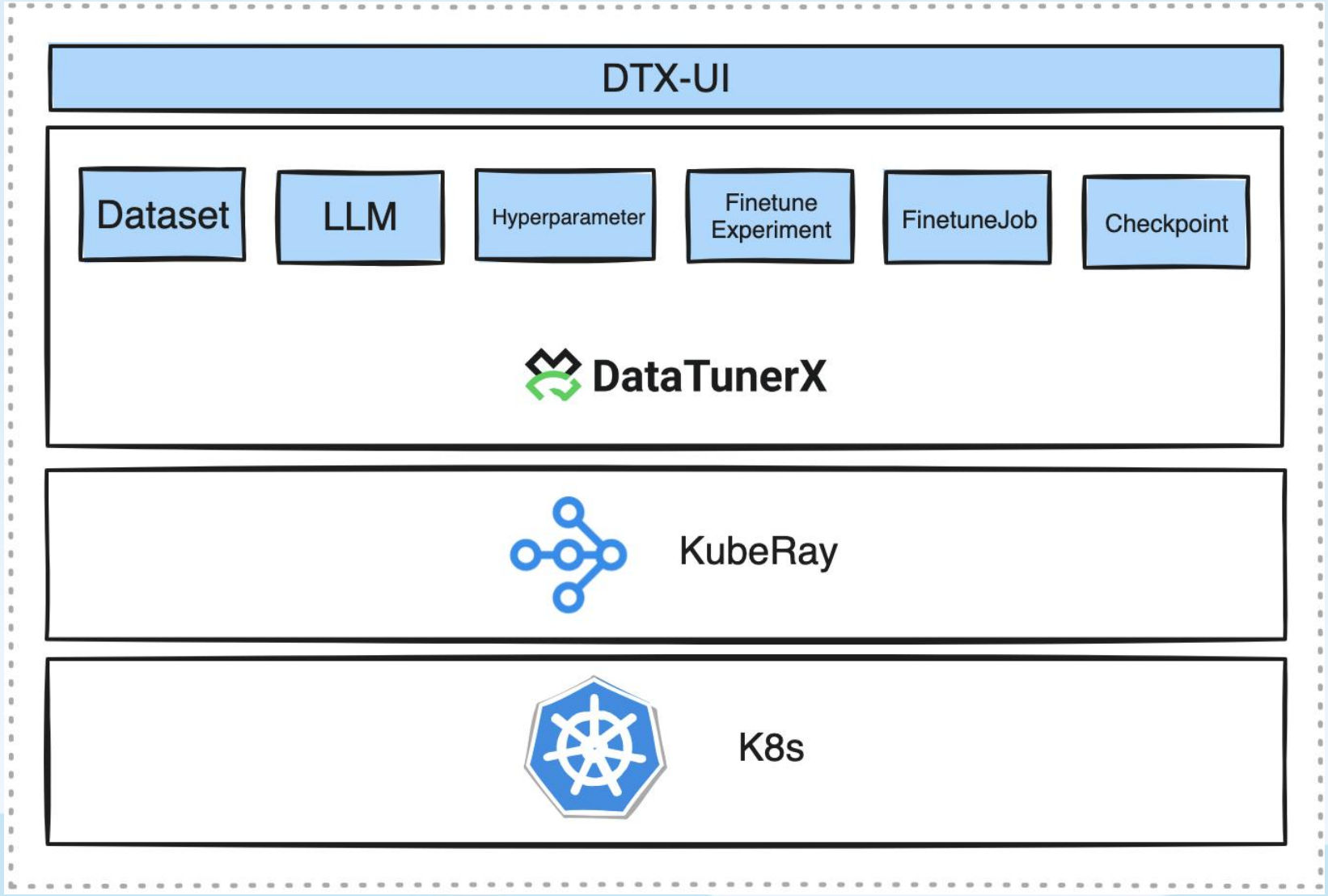
## 3 模型评估

传统的评估指标并不能完全反应LLM的能力，如何精确评估LLM能力会是一个挑战。

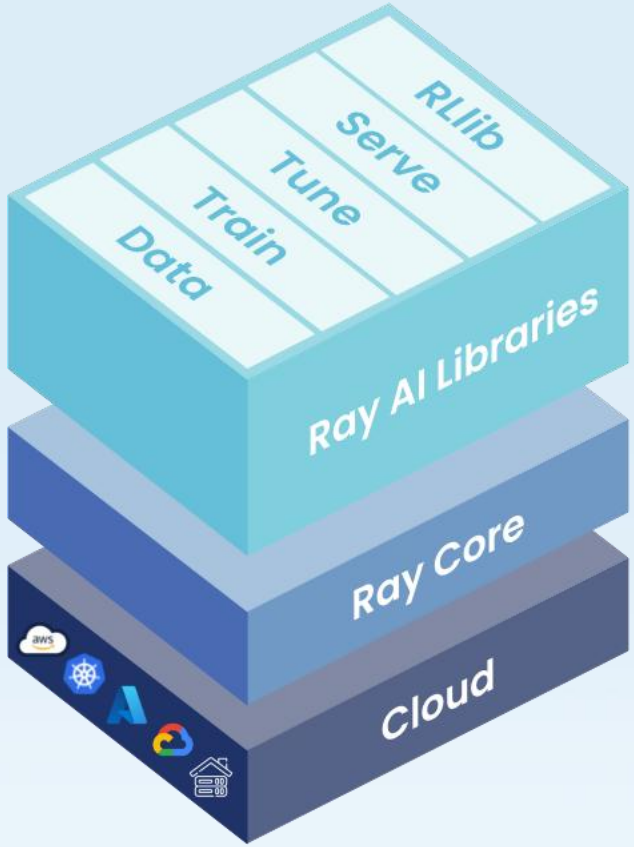

- 由DaoCloud和Futurewei 联合发起的一个基于云原生的大模型高效微调的项目
- 特点：
  - 易于使用，高效，快速一键微调
  - 并行，分布式，充分利用算力资源
  - 模型一键部署，推理
  - 可自定义数据，评估插件

项目地址：https://github.com/DataTunerX/datatunerx

# DTX架构


DTX-UI
Dataset
LLM
Hyperparameter
Finetune Experiment
FinetuneJob
Checkpoint
DataTunerX
KubeRay
K8s

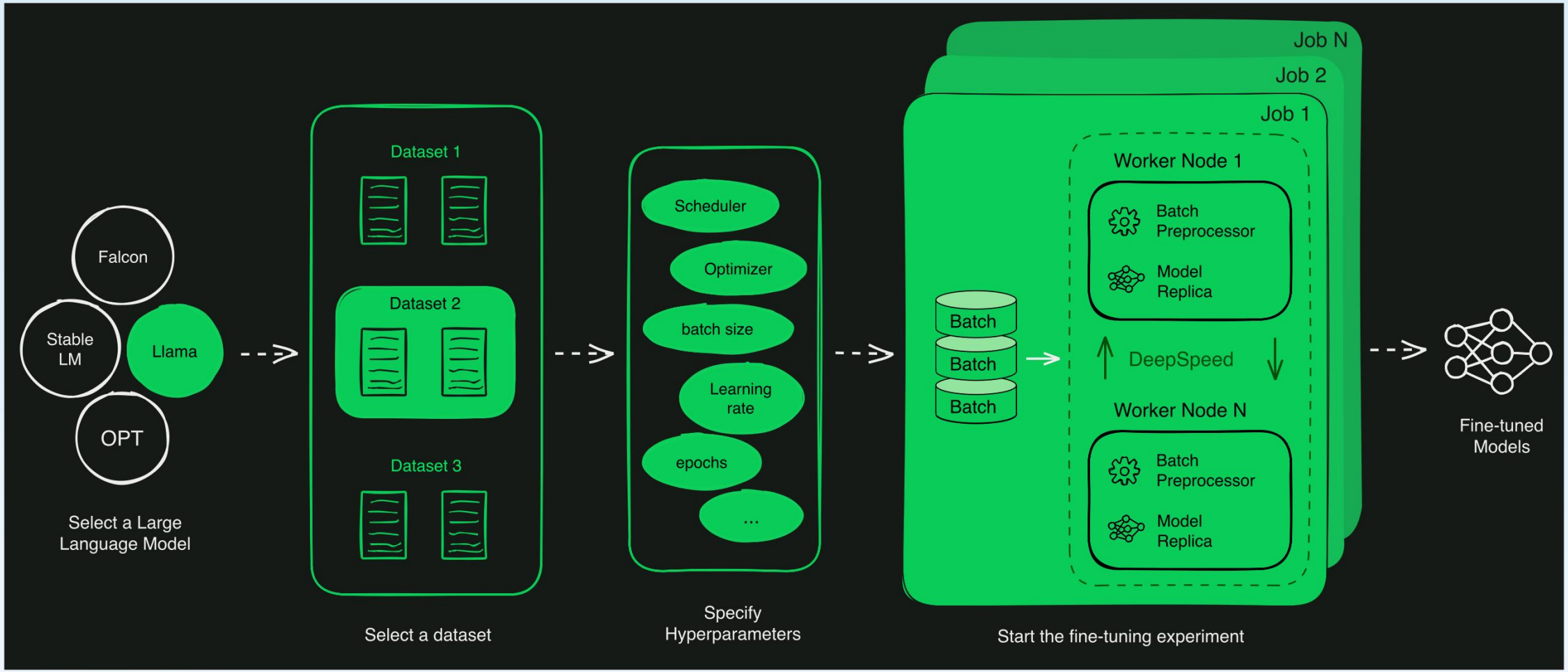
# What is KubeRay?



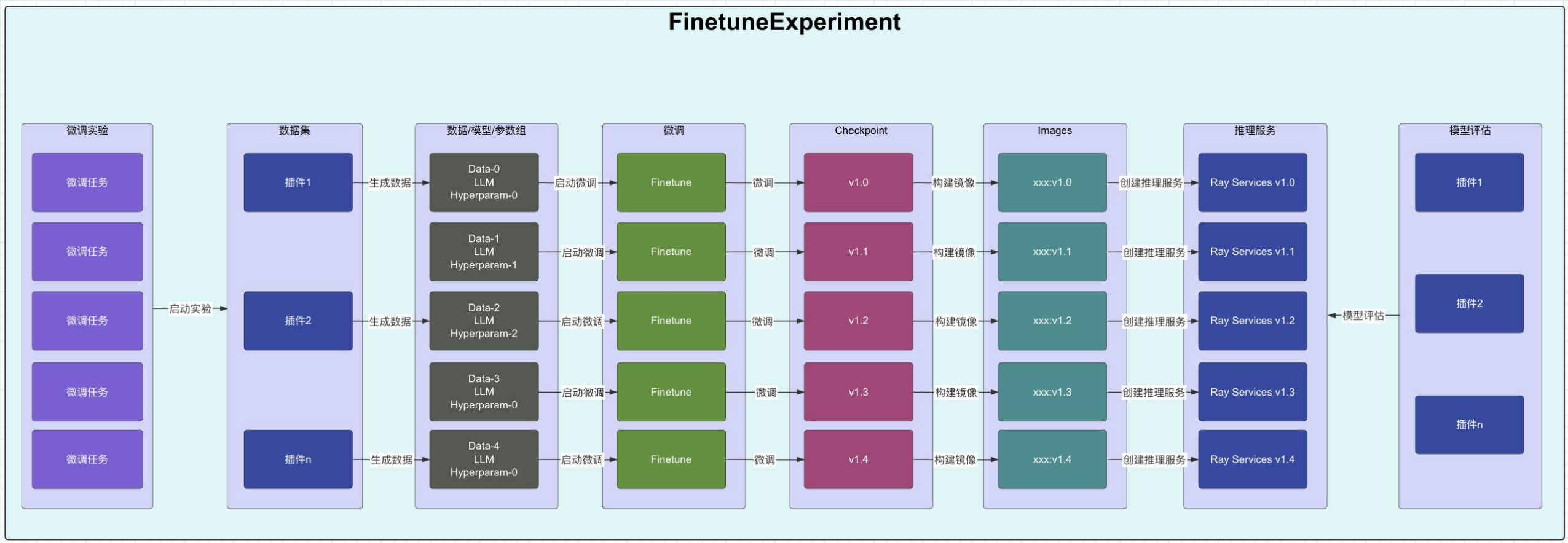

- Ray 是一个用于扩展 AI 和 Python 的开源统一框架。
- Ray为并行处理提供了计算层，因此无需成为分布式系统专家。
- 主要组件：Ray AIR，Ray Core，集成和实用工具。

KubeRay 是一个功能强大的开源 Kubernetes Operator，可简化 Ray 应用程序在 Kubernetes 上的部署和管理。
特点：自动化，高可用，弹性伸缩，资源调度，监控等。

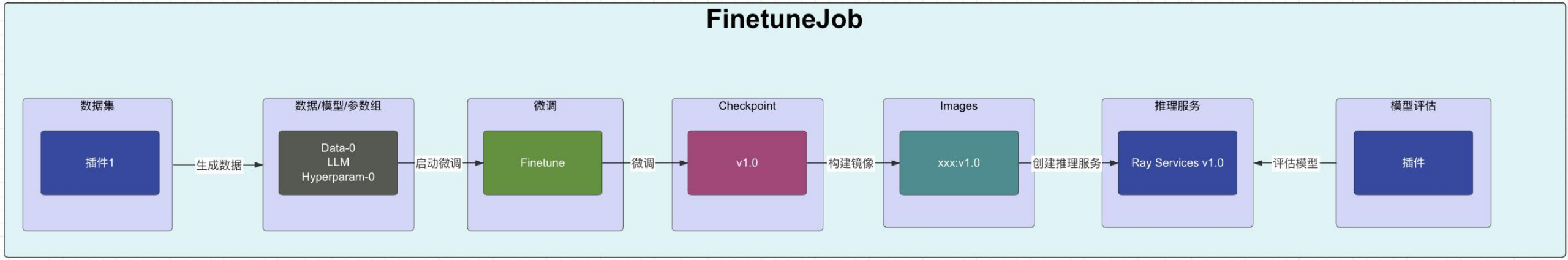
# 微调流程


Falcon
Stable LM
Llama
OPT
Select a Large Language Model
Dataset 1
Dataset 2
Dataset 3
Select a dataset
Scheduler
Optimizer
batch size
Learning rate
epochs
...
Specify Hyperparameters
Job N
Job 2
Job 1
Worker Node 1
Batch Preprocessor
Model Replica
DeepSpeed
Worker Node N
Batch Preprocessor
Model Replica
Batch
Batch
Batch
Start the fine-tuning experiment
Fine-tuned Models

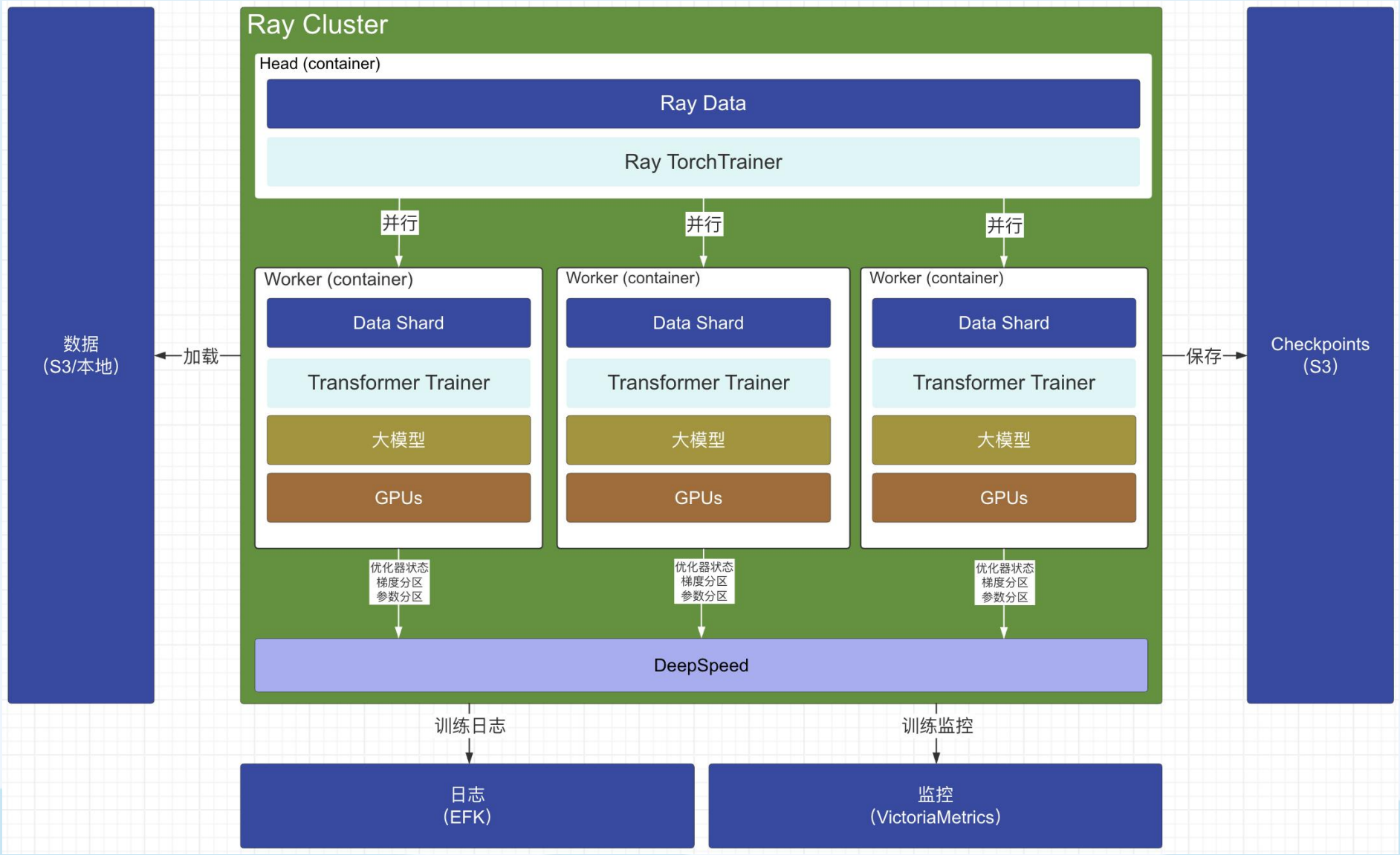
# 微调实验


FinetuneExperiment
微调实验
微调任务
微调任务
微调任务
微调任务
微调任务
启动实验
数据集
插件1
插件2
插件n
生成数据
生成数据
生成数据
数据/模型/参数组
Data-0 LLM Hyperparam-0
Data-1 LLM Hyperparam-1
Data-2 LLM Hyperparam-2
Data-3 LLM Hyperparam-0
Data-4 LLM Hyperparam-0
启动微调
启动微调
启动微调
启动微调
启动微调
微调
Finetune
Finetune
Finetune
Finetune
Finetune
微调
微调
微调
微调
微调
Checkpoint
v1.0
v1.1
v1.2
v1.3
v1.4
构建镜像
构建镜像
构建镜像
构建镜像
构建镜像
Images
xxx:v1.0
xxx:v1.1
xxx:v1.2
xxx:v1.3
xxx:v1.4
创建推理服务
创建推理服务
创建推理服务
创建推理服务
创建推理服务
推理服务
Ray Services v1.0
Ray Services v1.1
Ray Services v1.2
Ray Services v1.3
Ray Services v1.4
模型评估
模型评估
插件1
插件2
插件n

# 微调任务


FinetuneJob
数据集
插件1
生成数据
数据/模型/参数组
Data-0
LLM
Hyperparam-0
启动微调
微调
Finetune
微调
Checkpoint
v1.0
构建镜像
Images
xxx:v1.0
创建推理服务
推理服务
Ray Services v1.0
评估模型
模型评估
插件

# finetune


Ray Cluster
Head (container)
Ray Data
Ray TorchTrainer
并行
并行
并行
Worker (container)
Data Shard
Transformer Trainer
大模型
GPUs
Worker (container)
Data Shard
Transformer Trainer
大模型
GPUs
Worker (container)
Data Shard
Transformer Trainer
大模型
GPUs
数据
(S3/本地)
加载
保存
Checkpoints
(S3)
优化器状态
梯度分区
参数分区
优化器状态
梯度分区
参数分区
优化器状态
梯度分区
参数分区
DeepSpeed
训练日志
训练监控
日志
(EFK)
监控
(VictoriaMetrics)

# Dataset

```yaml
apiVersion: extension.datatunerx.io/v1beta1
kind: Dataset
metadata:
  name: example-dataset
spec:
  datasetCard:
    datasetCardRef: configmap-dataset-readme
  datasetFiles:
    source: www.test.com
  datasetMetadata:
    datasetInfo:
      features:
      - dataType: string
        mapTo: Content
        name: instruction
      - dataType: string
        mapTo: Result
        name: response
      subsets:
      - name: Default
        splits:
          test:
            file: s3://test.com
          train:
            file: s3://test.com
          validate:
            file: s3://test.com
    languages:
    - zh-CN
    license: CC BY-SA
    plugin:
      loadPlugin: true
      name: demo
      parameters: '{}'
    size: n < 1K
    task:
      name: Text Generation
```

```yaml
plugin:
  loadPlugin: true
  name: demo
  parameters: '{}'
```

# LLM

```yaml
apiVersion: core.datatunerx.io/v1beta1
kind: LLM
metadata:
  name: llama2-7b
spec:
  llmMetdata:
    computeInfrastructure:
      hardware:
        cpu: 4
        memory: 28Gi
        vRam: 6Gi
    domain:
    - NLP
    frameworks:
    - PyTorch
    languages:
    - English
    license:
    - https://ai.meta.com/resources/models-and-libraries/llama-downloads/
    llmImageConfig:
      image: harobor.custome.com/llama2-7b:v1
      path: /data/llama2-7b
    name: Llama-2-7b
    tasks:
    - Text Generation
```

# Hyperparameter

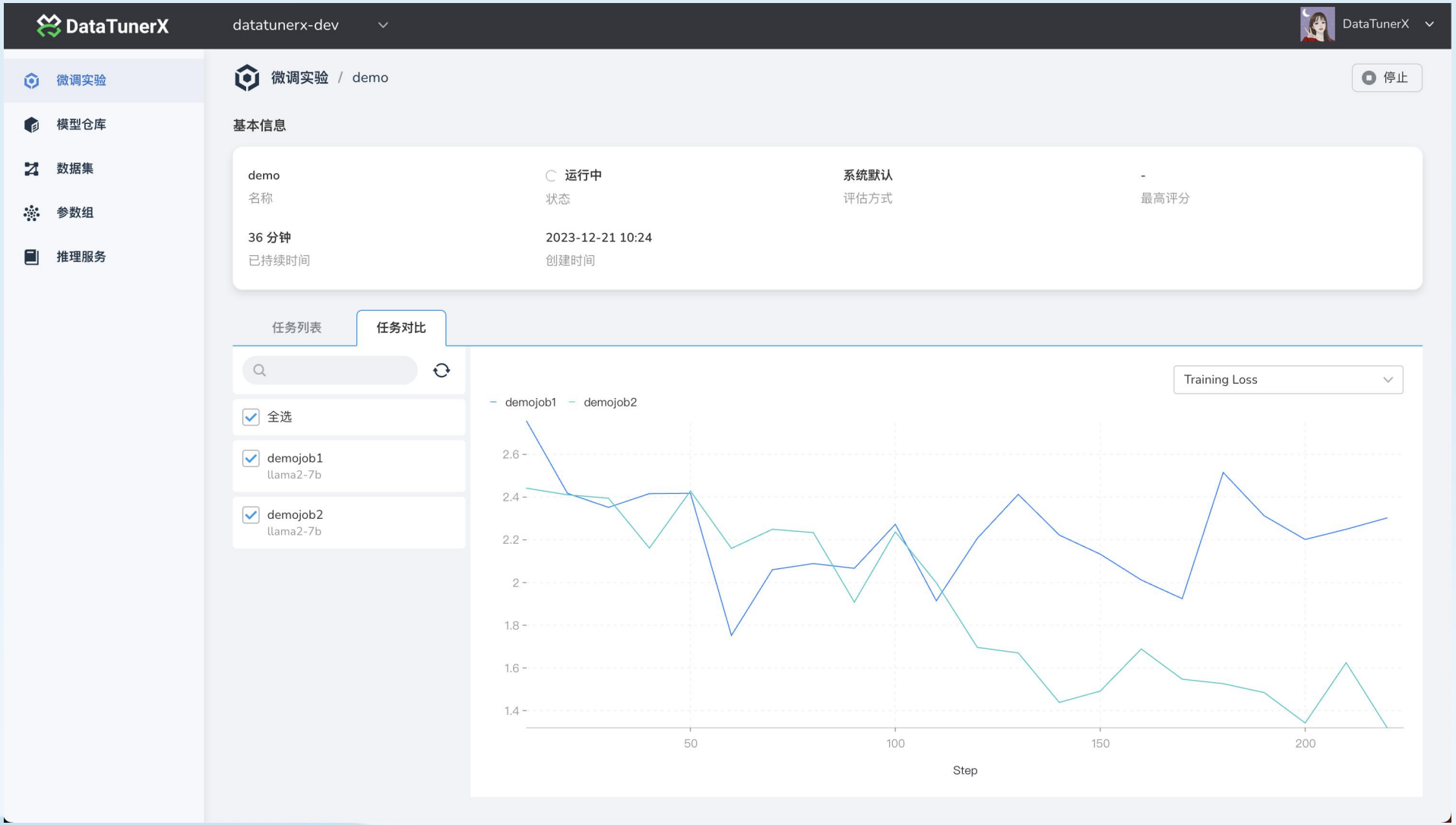
```yaml
apiVersion: core.datatunerx.io/v1beta1
kind: Hyperparameter
metadata:
  name: demo
spec:
  objective:
    type: SFT
  parameters:
    FP16: false
    PEFT: true
    batchSize: 32
    blockSize: 512
    epochs: 10123156464
    gradAccSteps: 1
    int4: false
    int8: false
    learningRate: "0.001"
    loRA_Alpha: "32.0"
    loRA_Dropout: "0.1"
    loRA_R: "4"
    optimizer: AdamW
    scheduler: Cosine
    trainerType: Standard
    warmupRatio: "0.1"
    weightDecay: "0.0001"
```

# FinetuneExperiment

```yaml
apiVersion: finetune.datatunerx.io/v1beta1
kind: FinetuneExperiment
metadata:
  name: test
spec:
  finetuneJobs:
  - name: test
    spec:
      finetune:
        finetuneSpec:
          dataset: example-dataset-liantiao
          hyperparameter:
            hyperparameterRef: happu
            overrides:
              batchSize: 1
              blockSize: 10
              epochs: 1
          image:
            imagePullPolicy: IfNotPresent
          llm: llama2-7b
          resource:
            limits: …
            requests: …
      scoringConfig:
        name: BuildInScoringPlugin
        parameters: '{}'
      serveConfig:
        nodeSelector:
          nvidia.com/gpu: present
```

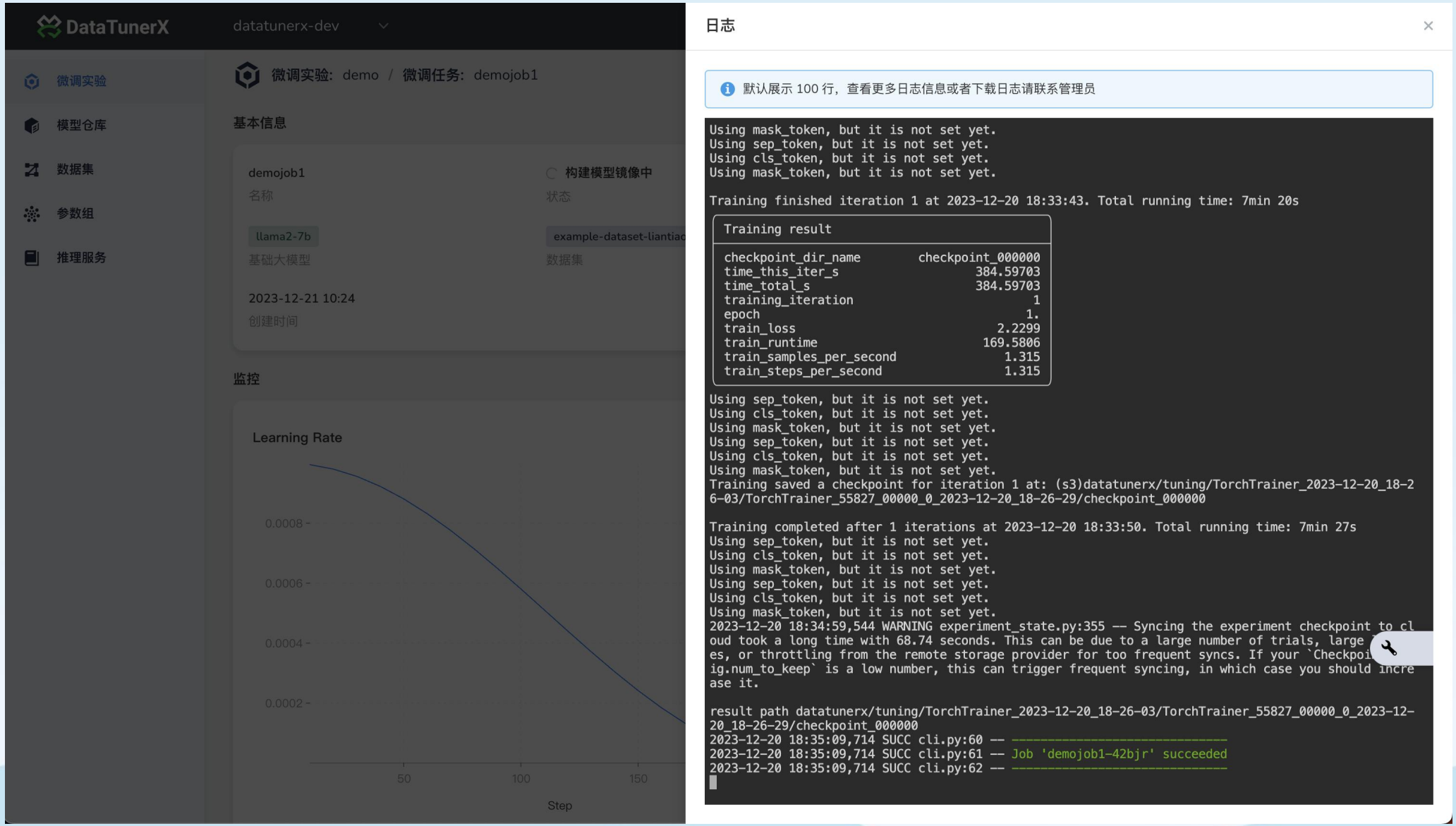
```yaml
      finetune:
        finetuneSpec:
          dataset: example-dataset-liantiao
          hyperparameter:
            hyperparameterRef: happu
            overrides:
              batchSize: 1
              blockSize: 10
              epochs: 1
          image:
            imagePullPolicy: IfNotPresent
          llm: llama2-7b
```

# DTX-UI



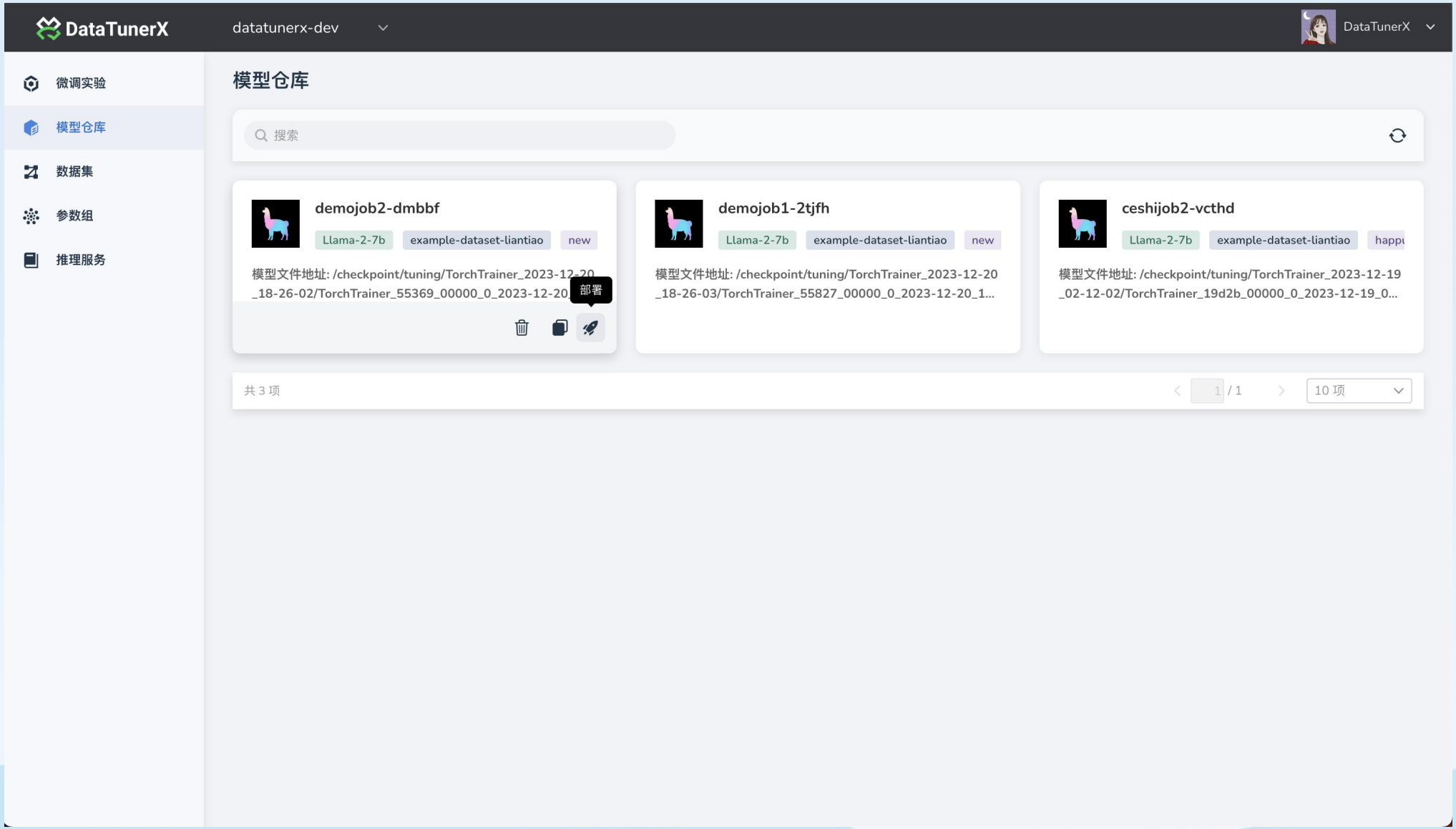

- 微调实验
- 模型仓库
- 数据集管理
- 参数组管理
- 推理服务

# DTX-UI



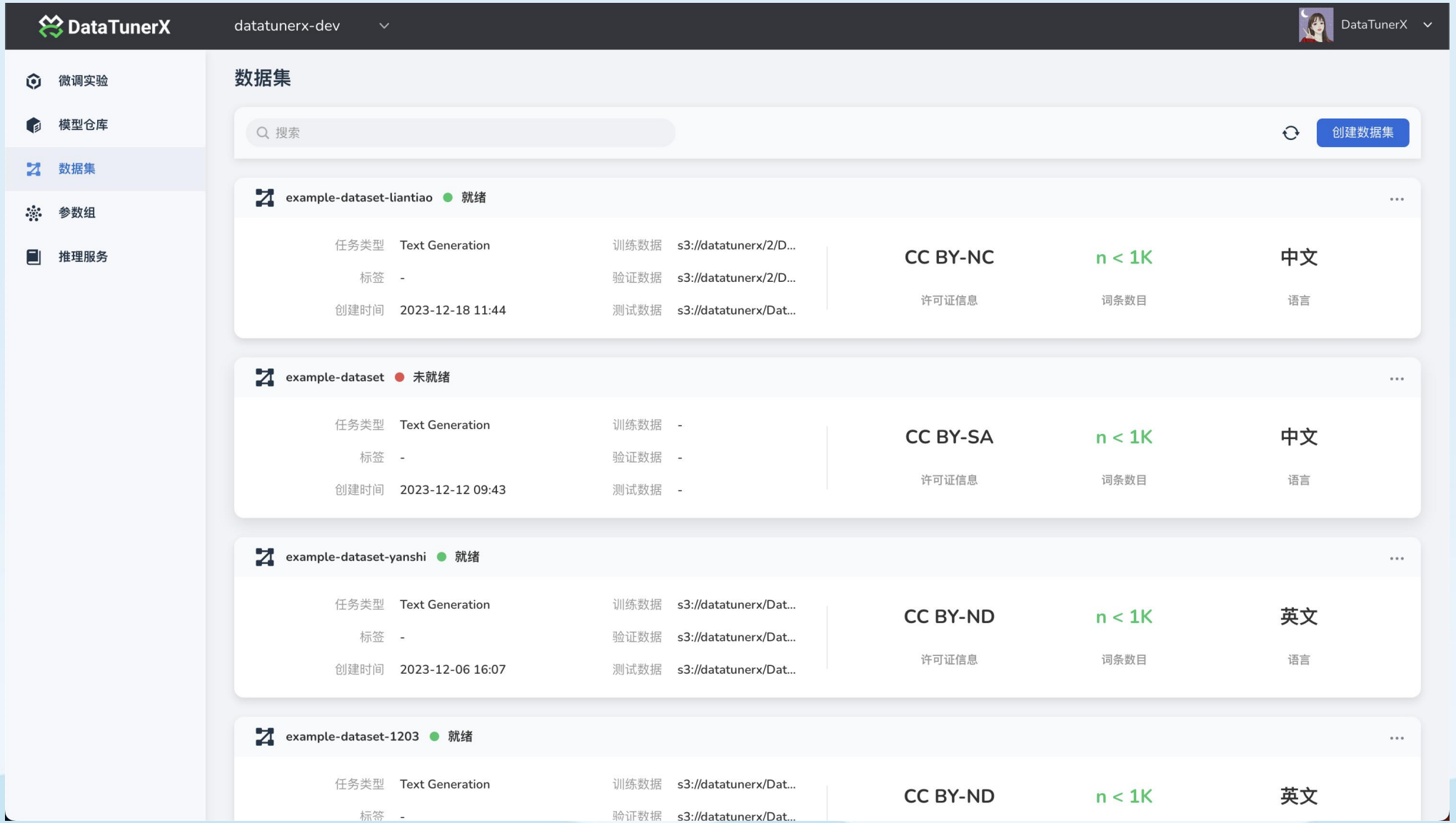

- 微调实验
- 模型仓库
- 数据集管理
- 参数组管理
- 推理服务

# DTX-UI



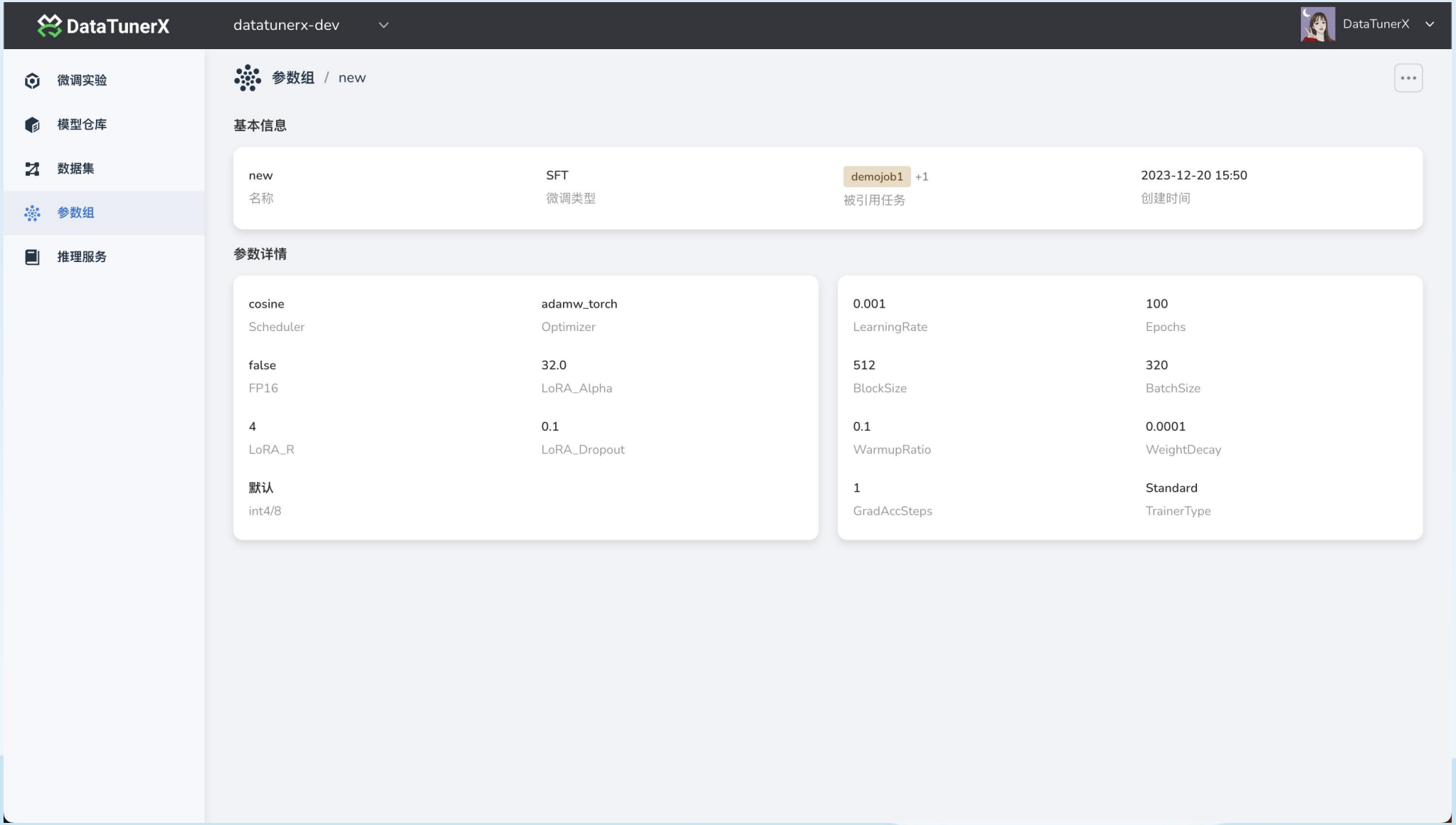

- 微调实验
- 模型仓库
- 数据集管理
- 参数组管理
- 推理服务

# DTX-UI



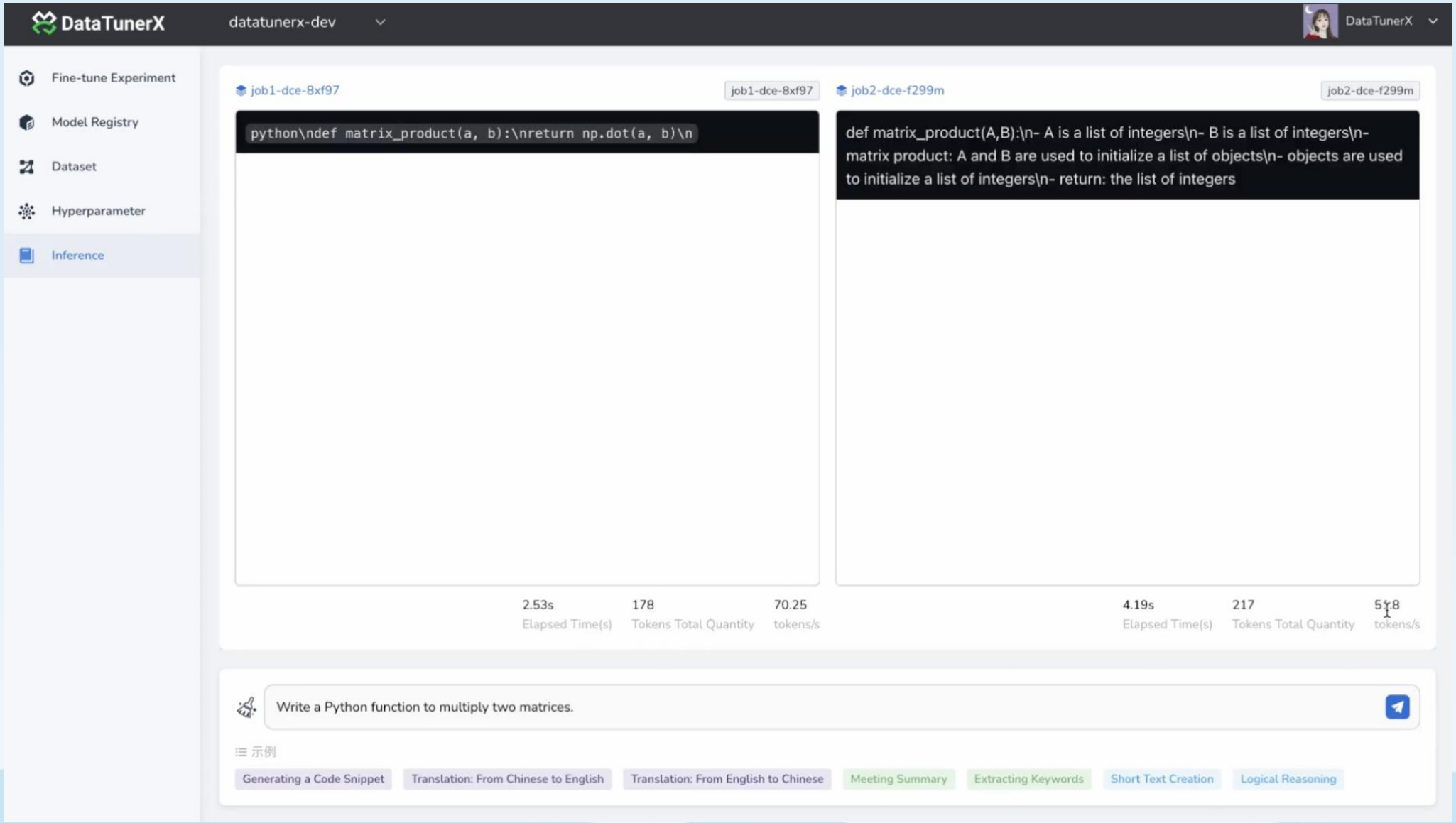

- 微调实验
- 模型仓库
- 数据集管理
- 参数组管理
- 推理服务

# DTX-UI




- 微调实验
- 模型仓库
- 数据集管理
- 参数组管理
- 推理服务

# DTX-UI




- 微调实验
- 模型仓库
- 数据集管理
- 参数组管理
- 推理服务


Demo 视频请查看：
https://www.bilibili.com/video/BV1eD421J76W
DTX
usage step
Tutorial

# Q & A

# Thanks.