AiDD AI+ 软件研发数字峰会
AI+ software Development Digital summit

AI驱动软件研发
全面进入数字化时代

中国·深圳 11.24-25

AI+
software
Development
Digital
summit

# 阿里云人工智能平台PAI的MaaS实践


罗义云（一耘）

阿里云资深技术专家、PAI平台工程技术负责人

# 科技生态圈峰会 + 深度研习

## ——1000 + 技术团队的共同选择

KEYLINK ing

K+峰会

K+峰会 深圳站

K+ 全球软件研发行业创新峰会

会议时间：2024.05.24-25

K+峰会 上海站

K+ 全球软件研发行业创新峰会

会议时间：2024.09.20-21

AiDD 峰会

AiDD峰会 深圳站

AI+ 软件研发数字峰会

会议时间：2023.11.24-25

AiDD峰会 北京站

AI+ 软件研发数字峰会

会议时间：2024.07.19-20

AiDD峰会 深圳站

AI+ 软件研发数字峰会

会议时间：2024.11.15-16

# 演讲嘉宾

## 罗义云

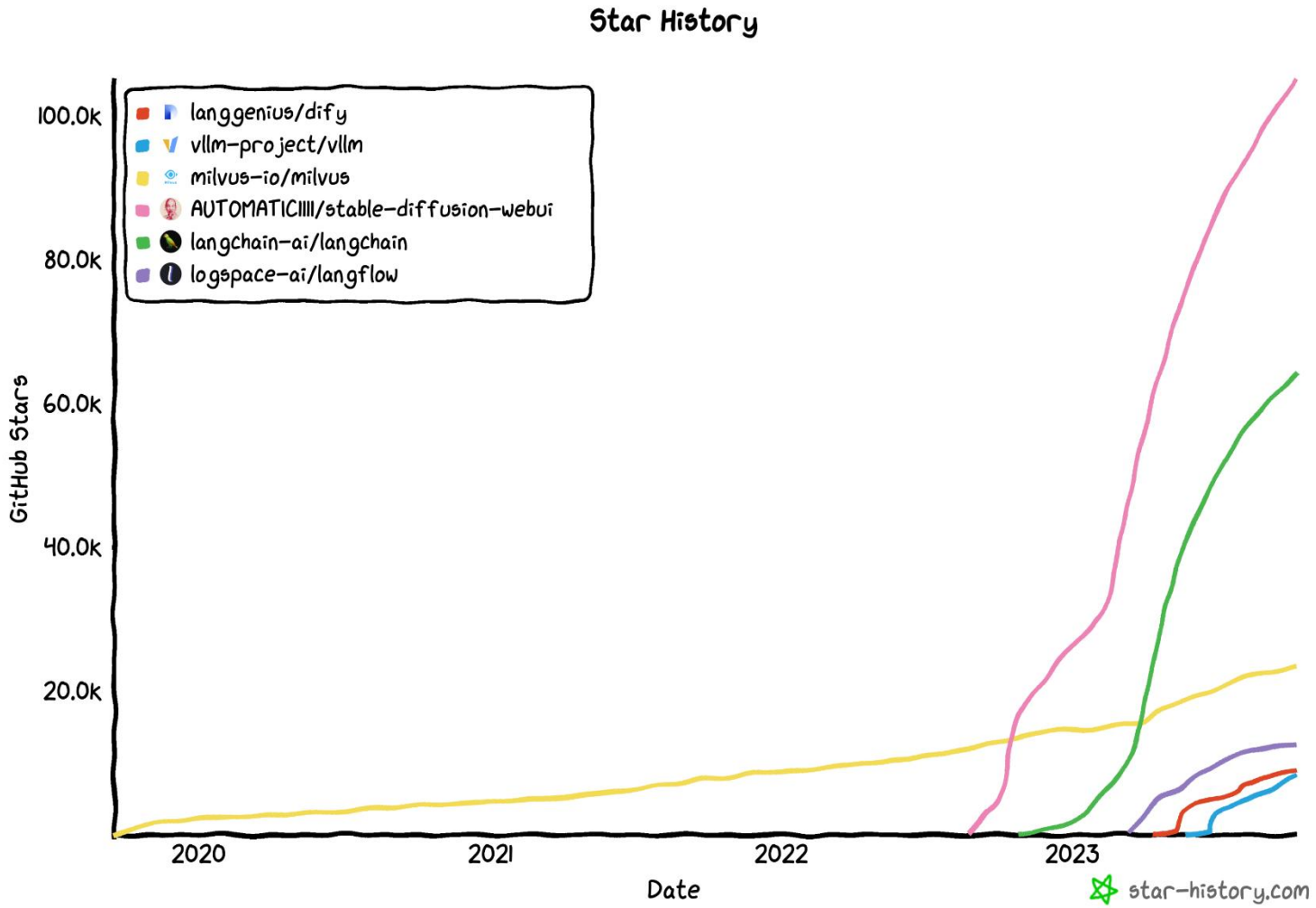
阿里云资深技术专家

毕业于北京大学，曾任微软高级研发经理、旷视科技AI平台高级技术总监，现任阿里云资深技术专家、AI平台工程技术负责人。在机器学习、人工智能、大数据等方向有着深厚的技术积累和丰富的行业经验。

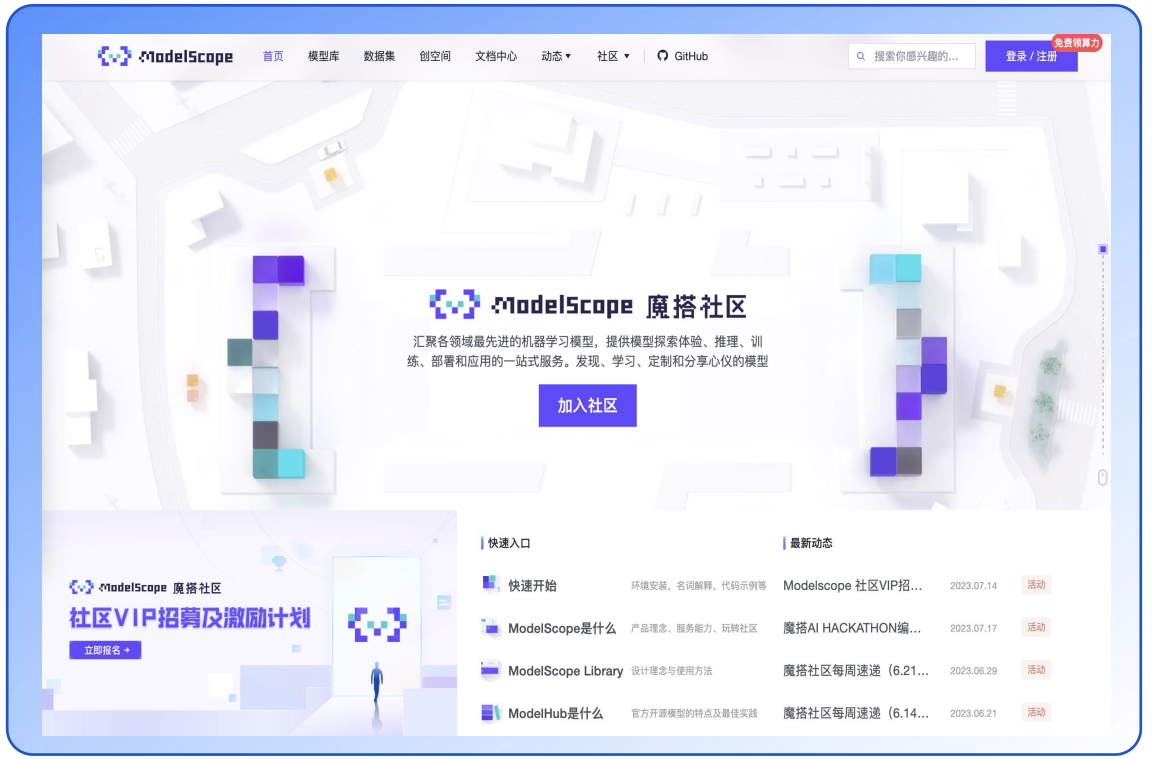
# 目录
# CONTENTS

# PART 01
# MaaS的起源和定义

# 人工智能进入AIGC时代



# 开源模型社区


**200万+**
活跃用户

**2300+**
模型

**1亿+**
模型累计下载

**20+**
顶尖人工智能合作机构

# 什么是MaaS?


MaaS：以云为基础，以模型为中心
MaaS(Model-as-a-Service): Cloud-based and Model-centric
阿里云
场景应用
互联网
政府
交通
零售
传媒
能源
金融
制造
模型存储/检索
模型更新
模型训练
模型即服务MaaS
模型监控
模型推理/服务
模型部署
推理加速
云计算平台
训练加速

# 什么是MaaS?

什么是MaaS（Model-as-a-Service），它的核心特点是什么?

MaaS（Model-as-a-Service）是一种云计算服务模式，它以云产品为核心，为企业提供全方位、全周期的云服务。相比于传统的IaaS、PaaS和SaaS等云服务模式，MaaS具有更强的灵活性和适应性。
MaaS的核心特点可以概括为以下几点：

1. 模型即服务：MaaS将机器学习模型和数据处理能力作为一种服务提供给企业，企业可以根据自己的需求选择不同的模型进行使用，无需自己开发和维护模型。
2. 集成化：MaaS将数据处理和模型服务集成到企业的业务流程中，提供一站式解决方案，可以大大降低企业的开发和维护成本。
3. 灵活性和适应性：MaaS提供了多种不同的模型和数据处理能力，企业可以根据自己的需求进行选择和组合，满足不同的业务场景和需求。
4. 持续优化：MaaS通过不断优化和更新模型，可以提供更准确、更高效的服务，帮助企业提高业务效率和效果。

总之，MaaS提供了一种全新的、更灵活、更便捷的机器学习和数据处理服务模式，可以帮助企业降低开发和维护成本，提高业务效率和效果。

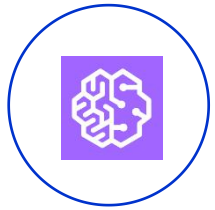
重新生成

# MaaS的组成要件

| | | | | | |
|---|---|---|---|---|---|
| | 业务产品 | 解决方案 | 最佳实践 | | |
| MaaS | Foundation models | 算法工程优化 | 平台工程优化（MLOps） | | |
| PaaS | 开发 | 训练 | 部署 | 资产管理 | 优化能力 |
| IaaS | 计算 | 存储 | 网络 | 集群管控 | |

MaaS=基础模型+算法工程优化+平台工程优化

# 主流云厂商是怎么做的?



SageMaker

Foundation models
Jumpstart

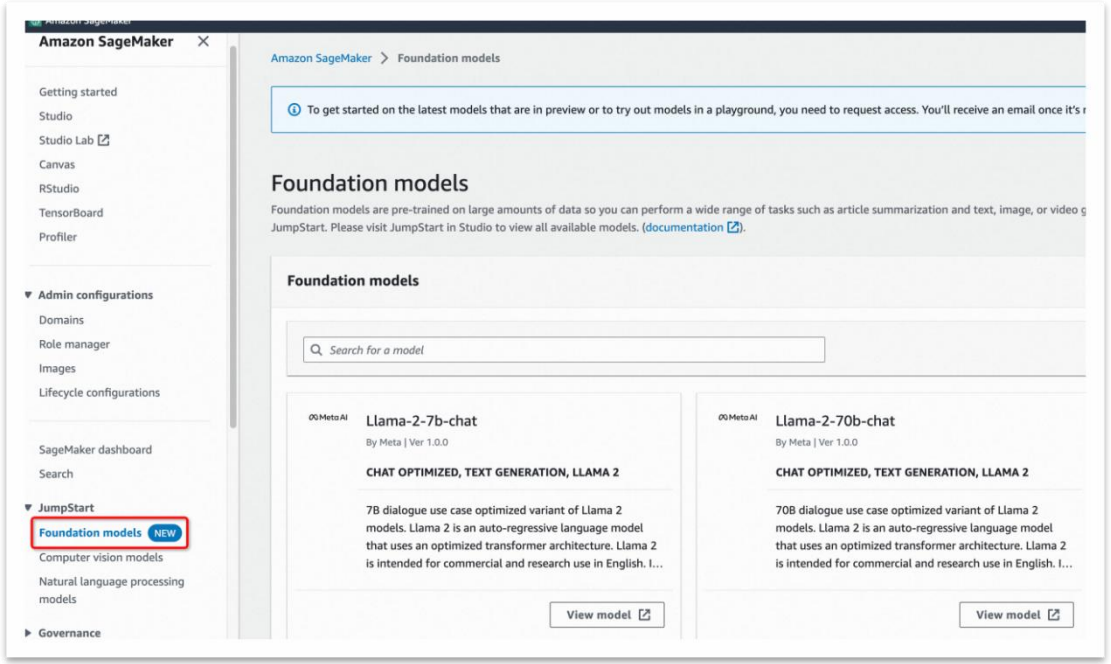
AzureML

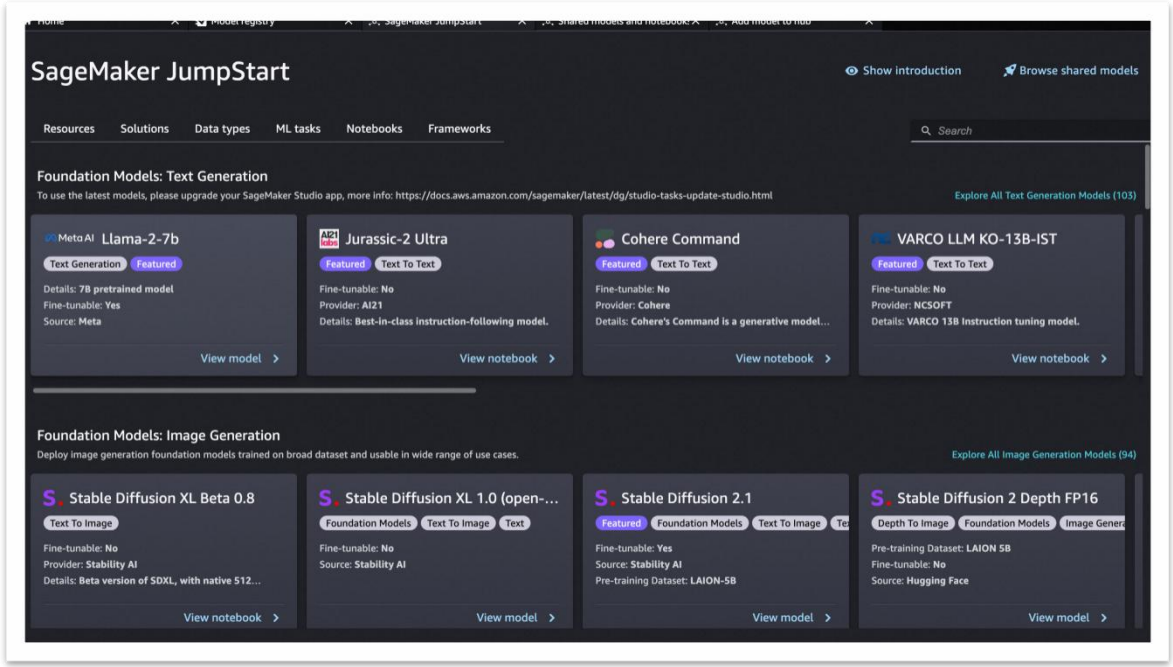
Model catalog

火山引擎

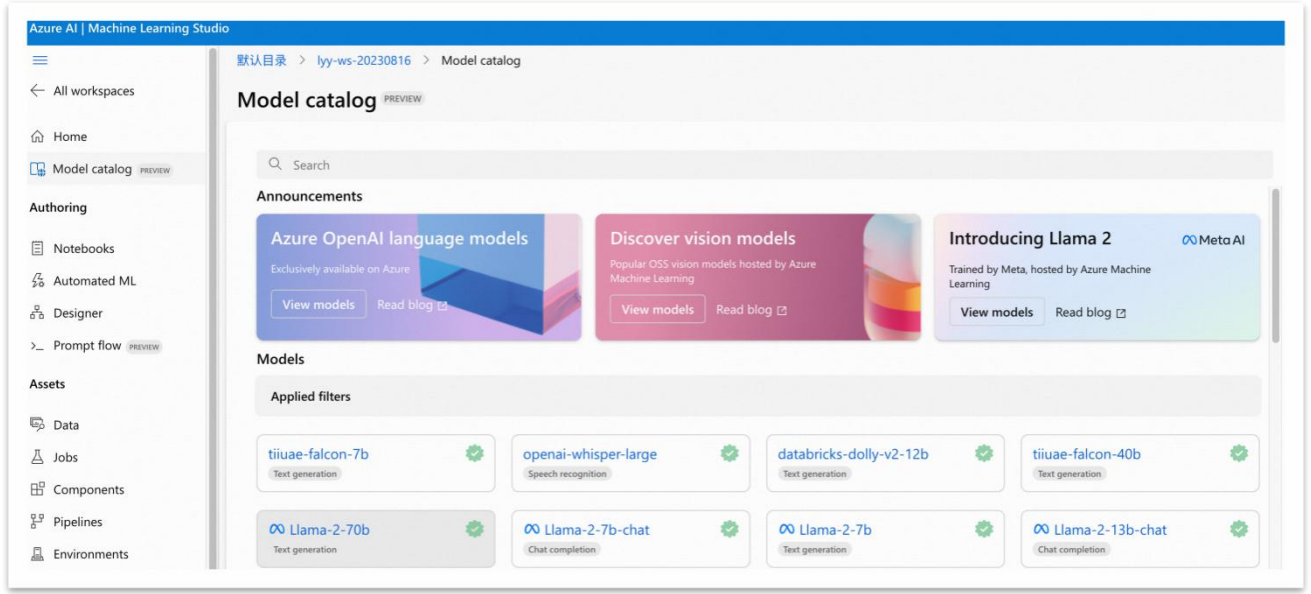
方舟大模型平台

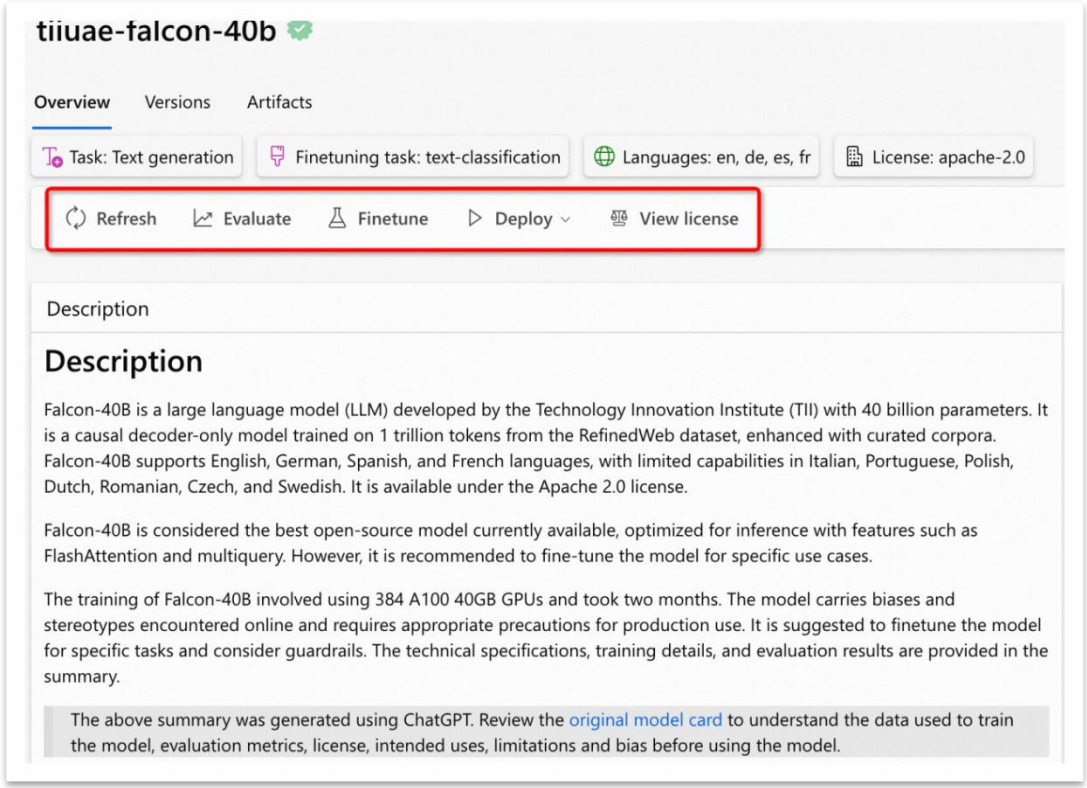
# SageMaker


Amazon SageMaker
Getting started
Studio
Studio Lab
Canvas
RStudio
TensorBoard
Profiler
Admin configurations
Domains
Role manager
Images
Lifecycle configurations
SageMaker dashboard
Search
JumpStart
Foundation models NEW
Computer vision models
Natural language processing models
Governance
Amazon SageMaker > Foundation models
To get started on the latest models that are in preview or to try out models in a playground, you need to request access. You'll receive an email once it's
Foundation models
Foundation models are pre-trained on large amounts of data so you can perform a wide range of tasks such as article summarization and text, image, or video
JumpStart. Please visit JumpStart in Studio to view all available models. (documentation).
Foundation models
Search for a model
Llama-2-7b-chat
By Meta | Ver 1.0.0
CHAT OPTIMIZED, TEXT GENERATION, LLAMA 2
7B dialogue use case optimized variant of Llama 2 models. Llama 2 is an auto-regressive language model that uses an optimized transformer architecture. Llama 2 is intended for commercial and research use in English. I...
View model
Llama-2-70b-chat
By Meta | Ver 1.0.0
CHAT OPTIMIZED, TEXT GENERATION, LLAMA 2
70B dialogue use case optimized variant of Llama 2 models. Llama 2 is an auto-regressive language model that uses an optimized transformer architecture. Llama 2 is intended for commercial and research use in English. I...
View model


SageMaker JumpStart
Show introduction
Browse shared models
Resources
Solutions
Data types
ML tasks
Notebooks
Frameworks
Search
Foundation Models: Text Generation
To use the latest models, please upgrade your SageMaker Studio app, more info: https://docs.aws.amazon.com/sagemaker/latest/dg/studio-tasks-update-studio.html
Explore All Text Generation Models (103)
Llama-2-7b
Text Generation
Featured
Details: 7B pretrained model
Fine-tunable: Yes
Source: Meta
View model
Jurassic-2 Ultra
Featured
Text To Text
Fine-tunable: No
Provider: AI21
Details: Best-in-class instruction-following model.
View notebook
Cohere Command
Featured
Text To Text
Fine-tunable: No
Provider: Cohere
Details: Cohere's Command is a generative model...
View notebook
VARCO LLM KO-13B-IST
Featured
Text To Text
Fine-tunable: No
Provider: NCSOFT
Details: VARCO 13B Instruction tuning model.
View notebook
Foundation Models: Image Generation
Deploy image generation foundation models trained on broad dataset and usable in wide range of use cases.
Explore All Image Generation Models (94)
Stable Diffusion XL Beta 0.8
Text To Image
Fine-tunable: No
Provider: Stability AI
Details: Beta version of SDXL, with native 512...
View notebook
Stable Diffusion XL 1.0 (open-...
Foundation Models
Text To Image
Text
Fine-tunable: No
Source: Stability AI
View model
Stable Diffusion 2.1
Featured
Foundation Models
Text To Image
Fine-tunable: Yes
Source: Stability AI
Pre-training Dataset: LAION-5B
View model
Stable Diffusion 2 Depth FP16
Depth To Image
Foundation Models
Pre-training Dataset: LAION 5B
Fine-tunable: No
Source: Hugging Face
View model

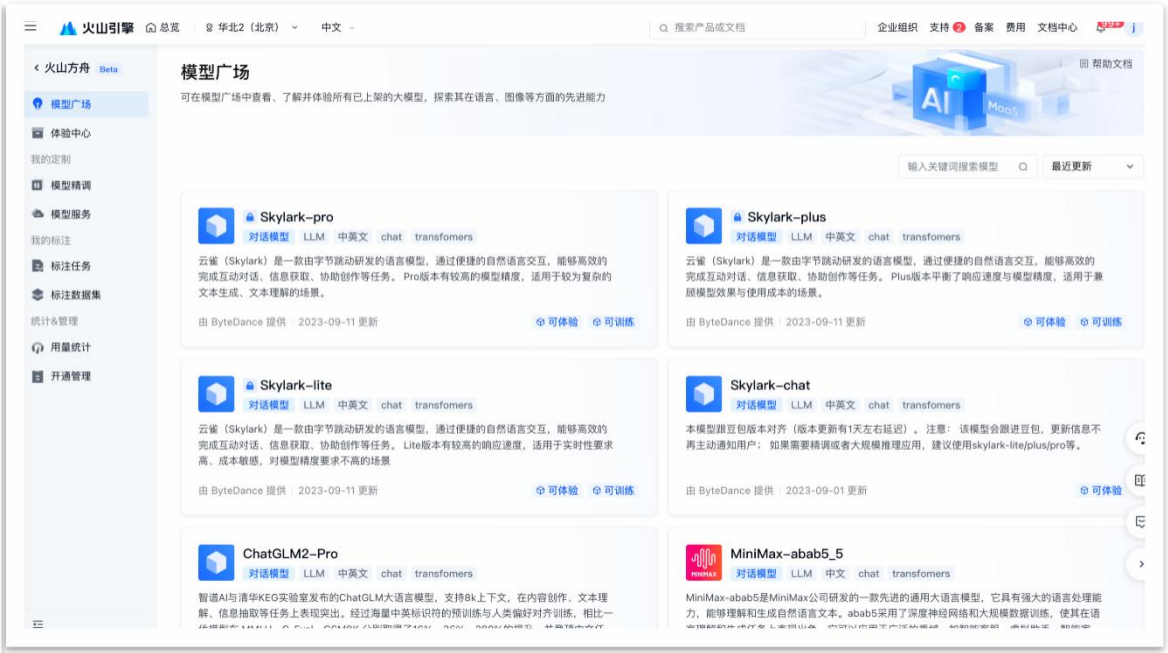
# AzureML


Azure AI | Machine Learning Studio
默认目录 > lyy-ws-20230816 > Model catalog
All workspaces
Home
Model catalog PREVIEW
Authoring
Notebooks
Automated ML
Designer
Prompt flow PREVIEW
Assets
Data
Jobs
Components
Pipelines
Environments
Model catalog PREVIEW
Search
Announcements
Azure OpenAI language models
Exclusively available on Azure
View models
Read blog
Discover vision models
Popular OSS vision models hosted by Azure Machine Learning
View models
Read blog
Introducing Llama 2
Meta AI
Trained by Meta, hosted by Azure Machine Learning
View models
Read blog
Models
Applied filters
tiiuae-falcon-7b
Text generation
openai-whisper-large
Speech recognition
databricks-dolly-v2-12b
Text generation
tiiuae-falcon-40b
Text generation
Llama-2-70b
Text generation
Llama-2-7b-chat
Chat completion
Llama-2-7b
Text generation
Llama-2-13b-chat
Chat completion


tiiuae-falcon-40b
Overview
Versions
Artifacts
Task: Text generation
Finetuning task: text-classification
Languages: en, de, es, fr
License: apache-2.0
Refresh
Evaluate
Finetune
Deploy
View license
Description
Description
Falcon-40B is a large language model (LLM) developed by the Technology Innovation Institute (TII) with 40 billion parameters. It is a causal decoder-only model trained on 1 trillion tokens from the RefinedWeb dataset, enhanced with curated corpora. Falcon-40B supports English, German, Spanish, and French languages, with limited capabilities in Italian, Portuguese, Polish, Dutch, Romanian, Czech, and Swedish. It is available under the Apache 2.0 license.
Falcon-40B is considered the best open-source model currently available, optimized for inference with features such as FlashAttention and multiquery. However, it is recommended to fine-tune the model for specific use cases.
The training of Falcon-40B involved using 384 A100 40GB GPUs and took two months. The model carries biases and stereotypes encountered online and requires appropriate precautions for production use. It is suggested to finetune the model for specific tasks and consider guardrails. The technical specifications, training details, and evaluation results are provided in the summary.
The above summary was generated using ChatGPT. Review the original model card to understand the data used to train the model, evaluation metrics, license, intended uses, limitations and bias before using the model.

# 火山引擎


功能全景
模型全景
火山方舟提供模型训练、推理、评测、精调等全方位功能与服务，并重点支撑大模型生态。
01. 模型广场
02. 模型体验
03. 模型训练推理
04. 模型应用
产品咨询
模型体验
模型训练推理
模型应用

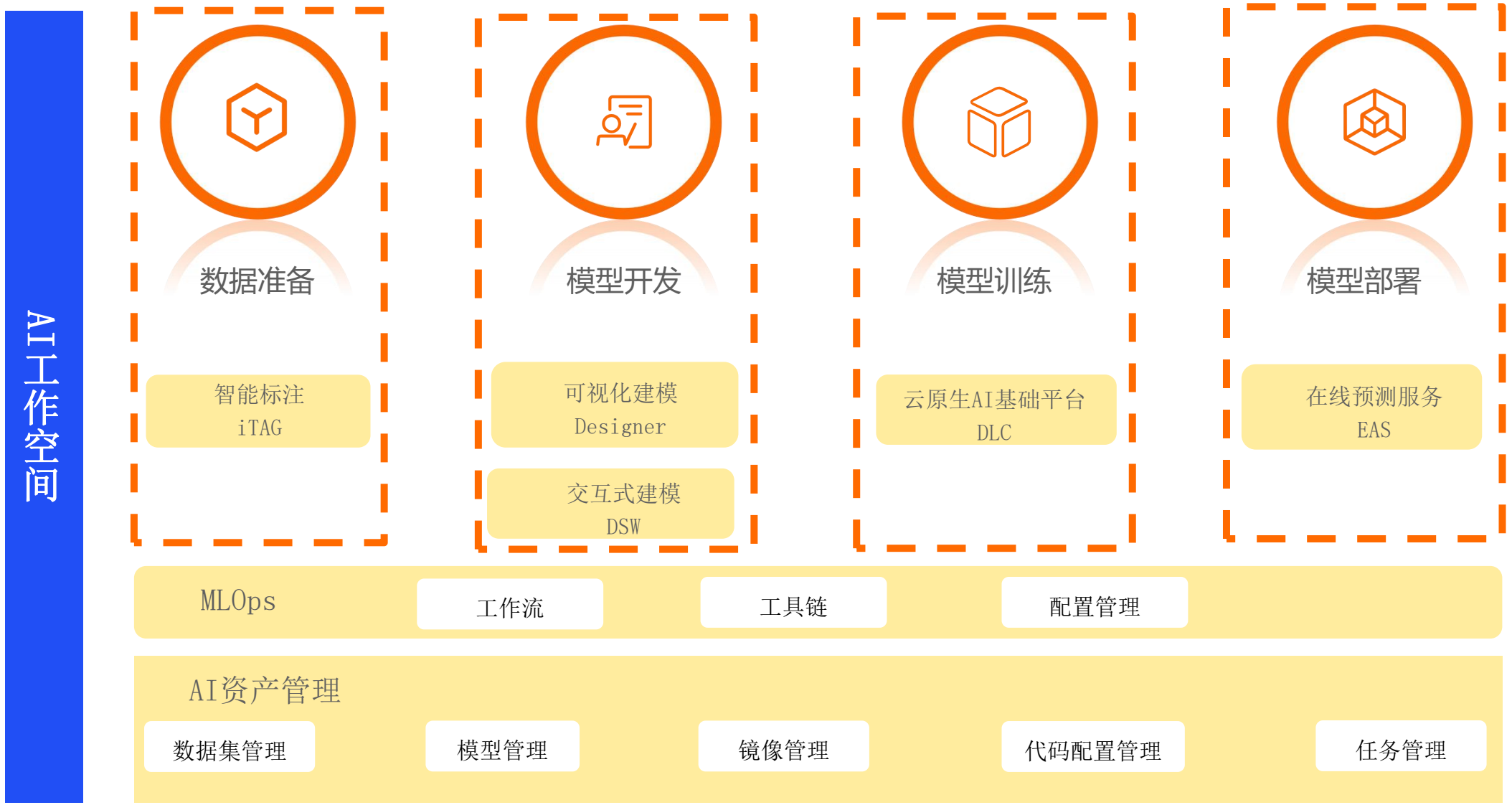

火山引擎
总览
华北2（北京）
中文
搜索产品或文档
企业组织
支持
备案
费用
文档中心
火山方舟 Beta
模型广场
体验中心
我的定制
模型精调
模型服务
我的标注
标注任务
标注数据集
统计&管理
用量统计
开通管理
模型广场
可在模型广场中查看、了解并体验所有已上架的大模型，探索其在语言、图像等方面的先进能力
帮助文档
输入关键词搜索模型
最近更新
Skylark-pro
对话模型 LLM 中英文 chat transfomers
云雀（Skylark）是一款由字节跳动研发的语言模型，通过便捷的自然语言交互，能够高效的完成互动对话、信息获取、协助创作等任务。Pro版本有较高的模型精度，适用于较为复杂的文本生成、文本理解的场景。
由 ByteDance 提供 2023-09-11 更新
可体验 可训练
Skylark-plus
对话模型 LLM 中英文 chat transfomers
云雀（Skylark）是一款由字节跳动研发的语言模型，通过便捷的自然语言交互，能够高效的完成互动对话、信息获取、协助创作等任务。Plus版本平衡了响应速度与模型精度，适用于兼顾模型效果与使用成本的场景。
由 ByteDance 提供 2023-09-11 更新
可体验 可训练
Skylark-lite
对话模型 LLM 中英文 chat transfomers
云雀（Skylark）是一款由字节跳动研发的语言模型，通过便捷的自然语言交互，能够高效的完成互动对话、信息获取、协助创作等任务。Lite版本有较高的响应速度，适用于实时性要求高、成本敏感，对模型精度要求不高的场景
由 ByteDance 提供 2023-09-11 更新
可体验 可训练
Skylark-chat
对话模型 LLM 中英文 chat transfomers
本模型跟豆包版本对齐（版本更新有1天左右延迟）。注意：该模型会跟进豆包，更新信息不再主动通知用户；如果需要精调或者大规模推理应用，建议使用skylark-lite/plus/pro等。
由 ByteDance 提供 2023-09-01 更新
可体验
ChatGLM2-Pro
对话模型 LLM 中英文 chat transfomers
智谱AI与清华KEG实验室发布的ChatGLM大语言模型，支持8k上下文，在内容创作、文本理解、信息抽取等任务上表现突出。经过海量中英标识符的预训练与人类偏好对齐训练，相比一
MiniMax-abab5_5
对话模型 LLM 中文 chat transfomers
MiniMax-abab5是MiniMax公司研发的一款先进的通用大语言模型，它具有强大的语言处理能力，能够理解和生成自然语言文本。abab5采用了深度神经网络和大规模数据训练，使其在语

# PART 02
# PAI的MaaS实践

# PAI产品架构


AI应用：自动驾驶 / 科研智算 / 金融风控 / 智能推荐 / 智能设计 / 智慧城市 / 智能制造 / 智慧医疗 / 智慧法务 / …
模型服务
(MaaS)
ModelScope 魔搭社区
PAI-DashScope 模型服务灵积
第三方MaaS平台
灵骏智算服务
&
机器学习框架
(PaaS)
快速开始：PAI-QuickStart /PAI-智码实验室 /PAI-DSW Gallery
工作空间
标注服务
PAI-iTAG
特征平台
PAI-Feature
Store
可视化建模
PAI-Designer
交互式建模
PAI-DSW
分布式训练
PAI-DLC
模型在线服务
PAI-EAS
开发者工具
CLI / PaiFlow
/ OpenAPI
AI资产管理（数据集 / 模型 / 镜像 / 代码 / 自定义组件 / …
权限管理
MLOps
云产品依赖
优化与加速（DatasetAcc 数据集加速 / TorchAcc 训练加速 / EPL 并行训练框架 / Blade推理加速 / AI Master 自动容错训练 / EasyCkpt 秒级异步训练快照）
机器学习框架（PAI-TensorFlow / PAI-PyTorch / Alink / Spark, EasyRec / EasyPhoto / EasyTransfer / Megatron / DeepSpeed / RLHF）
计算资源
&
基础设施
（IaaS）
PAI-灵骏计算资源
云原生通用计算资源
大数据计算资源（MaxCompute / EMR / Flink）
异构计算磐久服务器（CPU、GPU）
高速 RDMA 网络（RoCE）
分布式存储 CPFS/NAS/OSS
容器服务 ACK
弹性计算 ECS
低PUE液冷/风冷，模块化IDC设施

# PAI 提供一站式AI研发生命周期管理



# PAI-iTAG 智能标注

阿里云智能标注，提供全场景、高质量、智能化的标注平台服务和人力标注服务


## 标注能力

图像、文本、视频、语音、PDF、多模态、自定义等全方位的标注能力和场景

## 智能标注

打通PAI-EAS部署的模型，让模型服务进行主动预标注

## 预标工具

预置OCR、ASR等预标工具，大幅提升标注员的效率和准确率

## 数据安全

阿里云最新的数字安全传输技术，保护标注数据不泄露

## 任务分发与验收

标注+质检+验收的任务分发机制，
保证标注数据的高质量交付

## 人员与权限管理

管理员+标注组长+标注员的人员与权限管理，
保证各角色的权限隔离和数据安全

## 人力标注服务：专业、全托管的数据服务

公益模式标注基地，专业培训上岗，全托管的项目管理，极大降低标注人力成本

# PAI-DSW

云原生交互式建模

## 实例环境持久化

- 支持保存环境为镜像

## 灵活的环境支持

- 支持预置镜像和自定义镜像

## 多数据集支持

- 支持同时挂载多个数据集

## 实例生命周期管理

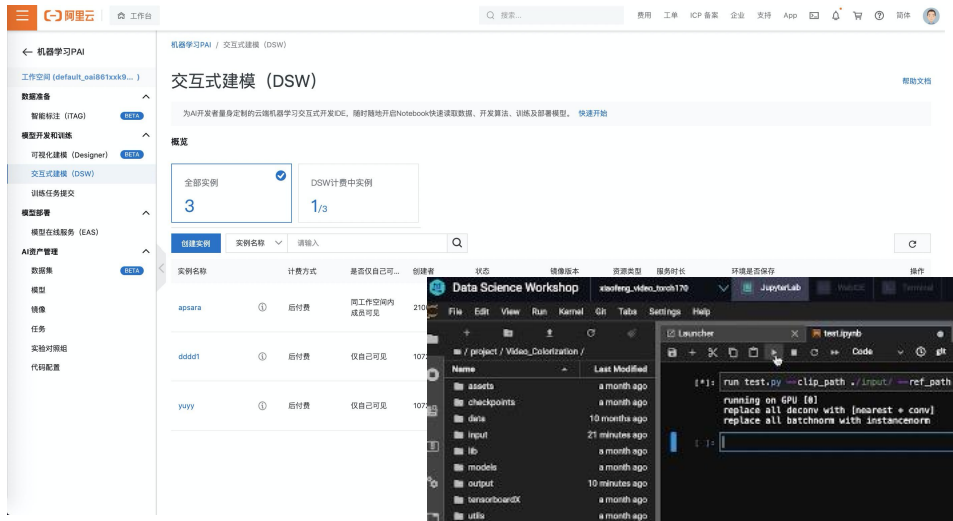
- 支持定时停止

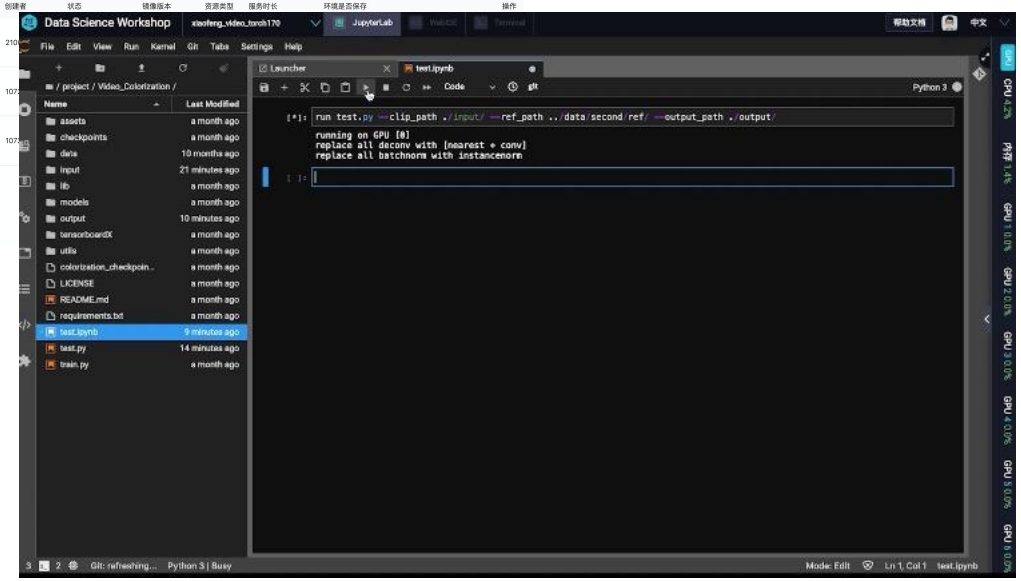
## 实例权限控制

- 支持同一工作空间内成员间分隔

## 开放被集成

- 开放OpenAPI



# PAI-DLC

## 云原生AI基础训练平台

### 多资源组支持

- 支持公共资源与专有资源

### 灵活的环境支持

- 支持预置镜像和自定义镜像

### 多数据集支持

- 支持同时挂载多个数据集

### 任务生命周期管理

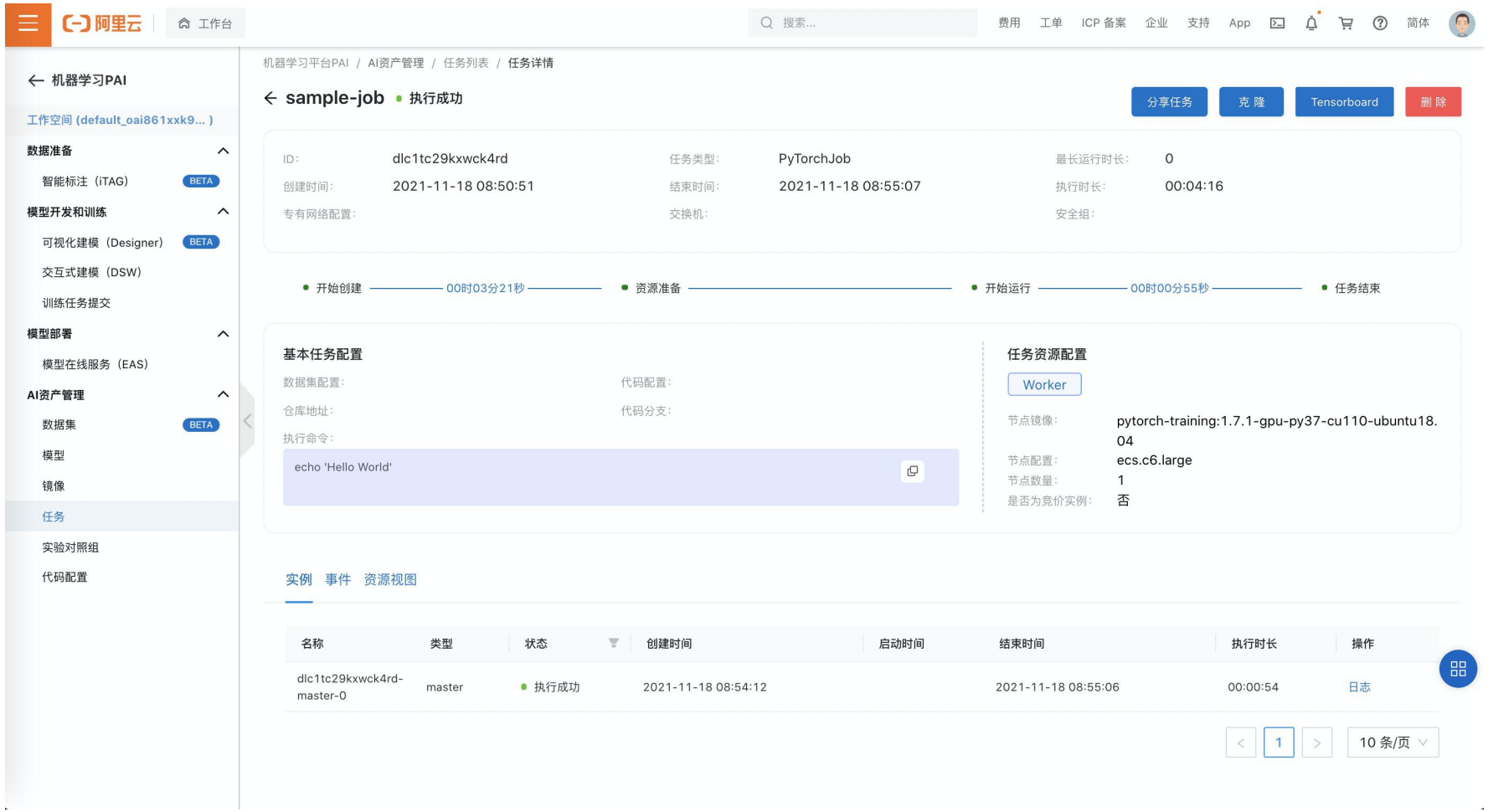
- 任务全流程把控

### 实例权限控制

- 同一工作空间内成员间分隔

### 开放被集成

- 开放OpenAPI



# PAI-EAS

基于异构硬件(CPU/GPU/NPU/FPGA)提供机器学习深度学习模型快速部署的微服务平台

## 支持多种框架模型

提供通用机器学习模型（PMML, TensorFlow, Pytorch等）一键部署成在线服务。

## 自定义多语言processor

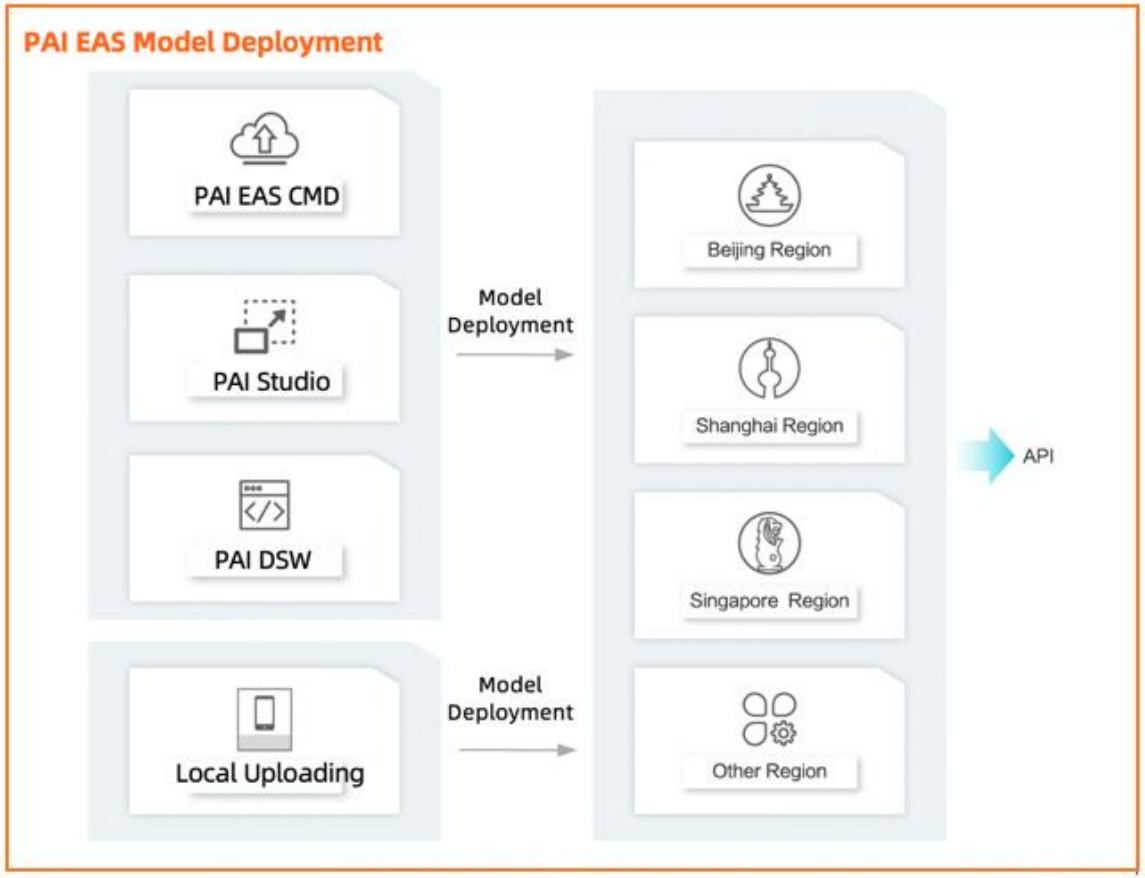
提供三种主流语言(C++/Python/Java)开发用户自定义的预测逻辑，将模型包装成端对端的产品服务。

## 高性能

支持阿里集团数十个BU的模型预测服务及大量外部客户，单服务QPS峰值30w，大量QPS过万的图像类服务

## 开放被集成

开放OpenAPI



# PAI的MaaS理念

Model Centric

一站式

白盒化

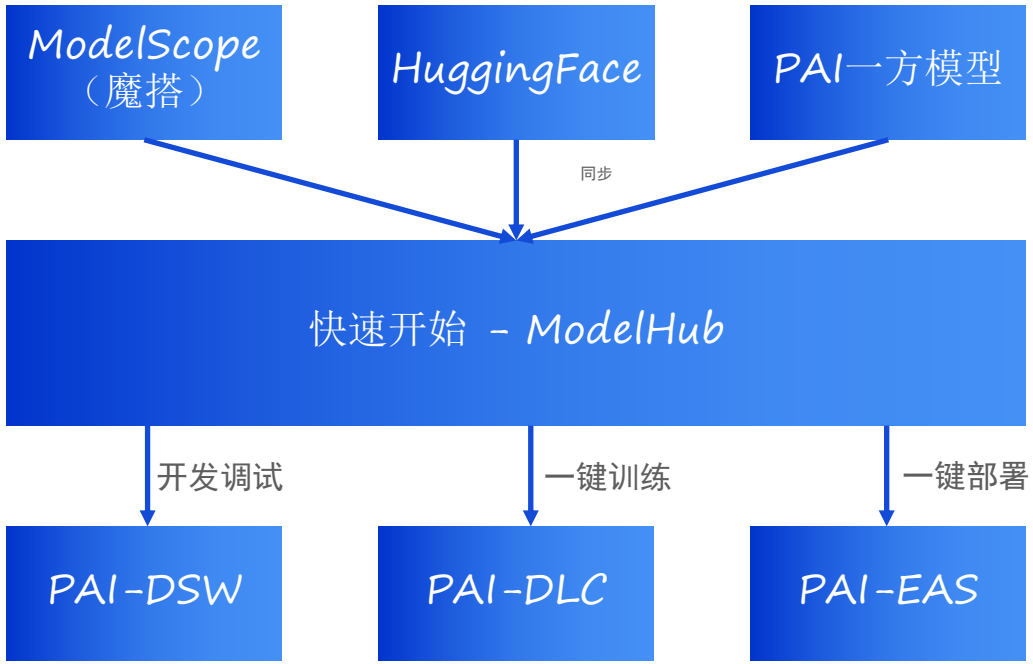
# PAI的MaaS实践

快速开始
（ModelHub）

PAI SDK

PAI × ModelScope

# 快速开始（ModelHub）



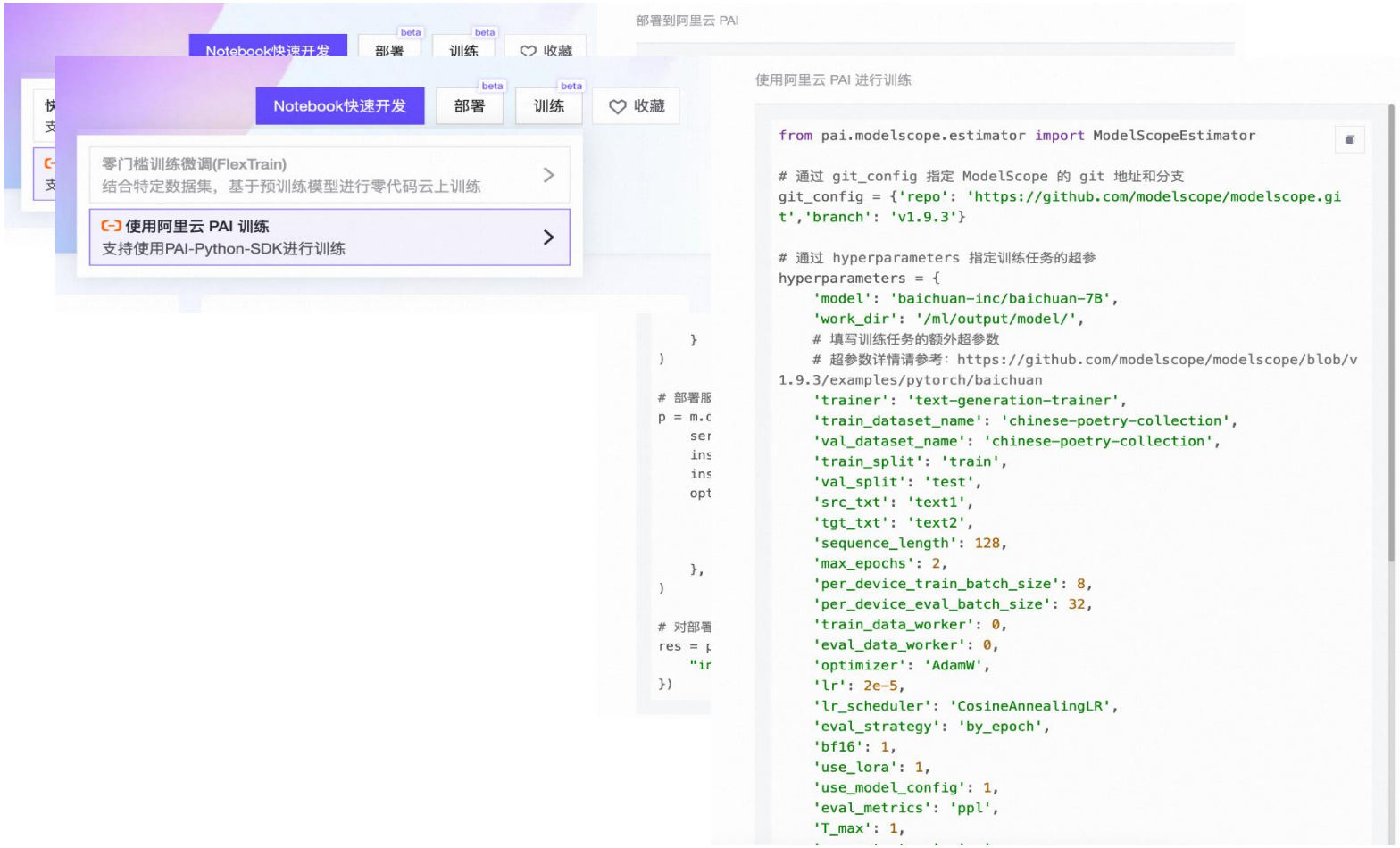

- 丰富的模型支持
- 一站式模型训练->部署的全链路体验
- 内置算法工程优化，提升迭代效率
- 结合平台优化能力，提供极致的性能和性价比

# PAI SDK对开源模型的支持

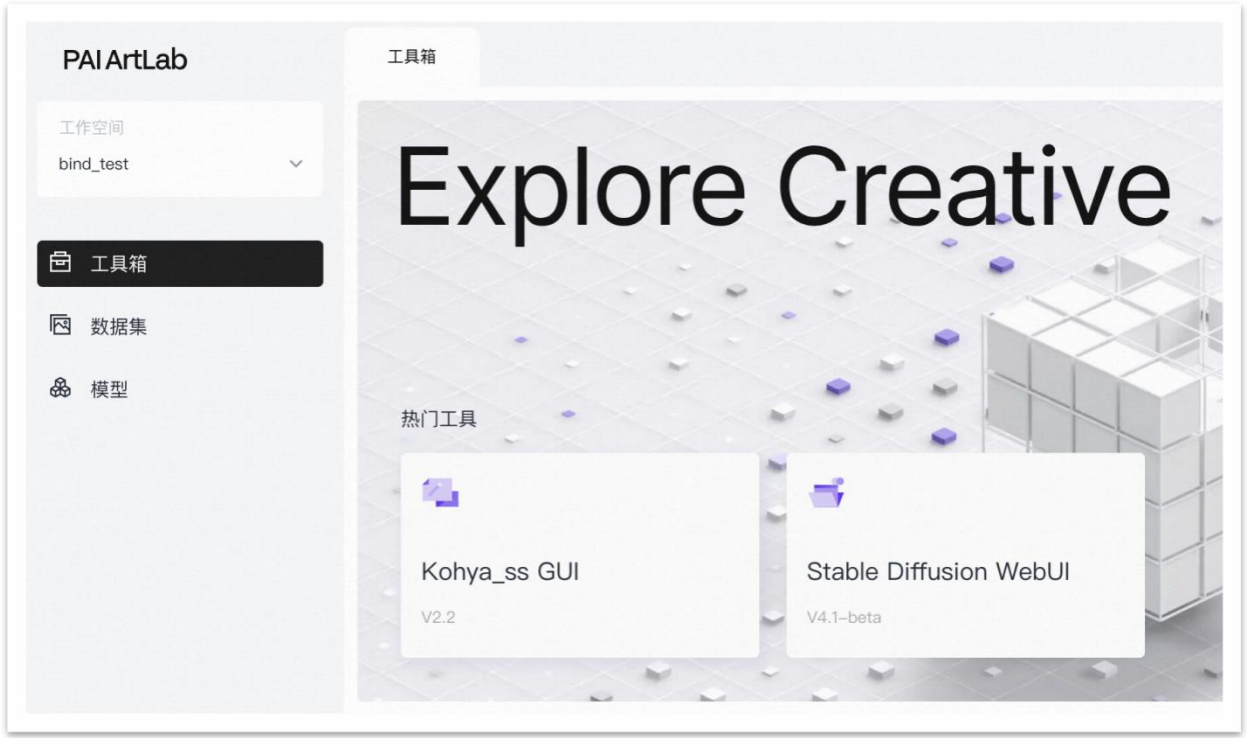
```python
# Huggingface模型训练
from pai.huggingface.estimator import HuggingFaceEstimator

# 通过 git_config 来指定 transformers 的 git 地址和分支
git_config = {
    "repo": "https://github.com/huggingface/transformers.git",
    "branch": "v4.29.2",
}

# 通过 hyperparameters 来指定训练参数，这里以 GLUE 的 MRPC 任务为例
hyperparameters = {
    "model_name_or_path": "bert-base-uncased",
    "output_dir": "/ml/output/model/",
    "task_name": "mrpc",
    "do_train": True,
    "do_eval": True,
    "max_seq_length": 128,
    "per_device_train_batch_size": 32,
    "learning_rate": 2e-5,
    "num_train_epochs": 3,
}

# 创建 HuggingFaceEstimator 对象
est = HuggingFaceEstimator(
    source_dir="./examples/pytorch/text-classification",  # 训练脚本在 git repo 中的相对路径
    git_config=git_config,  # 指定 transformers 的 git 地址和分支
    command="python3 run_glue.py $PAI_USER_ARGS",  # 指定训练命令，通过 $PAI_USER_ARGS 传入 hyperparameters
    instance_type="ecs.gn7i-c32g1.8xlarge",  # 指定训练机型
    transformers_version="4.29.2",  # 指定训练镜像所使用的 transformers 版本，目前支持 4.29.2 或 latest
    hyperparameters=hyperparameters,  # 指定训练参数
    base_job_name="huggingface-sdk-train",  # 指定训练任务名称
)

# 进行训练
est.fit()
```

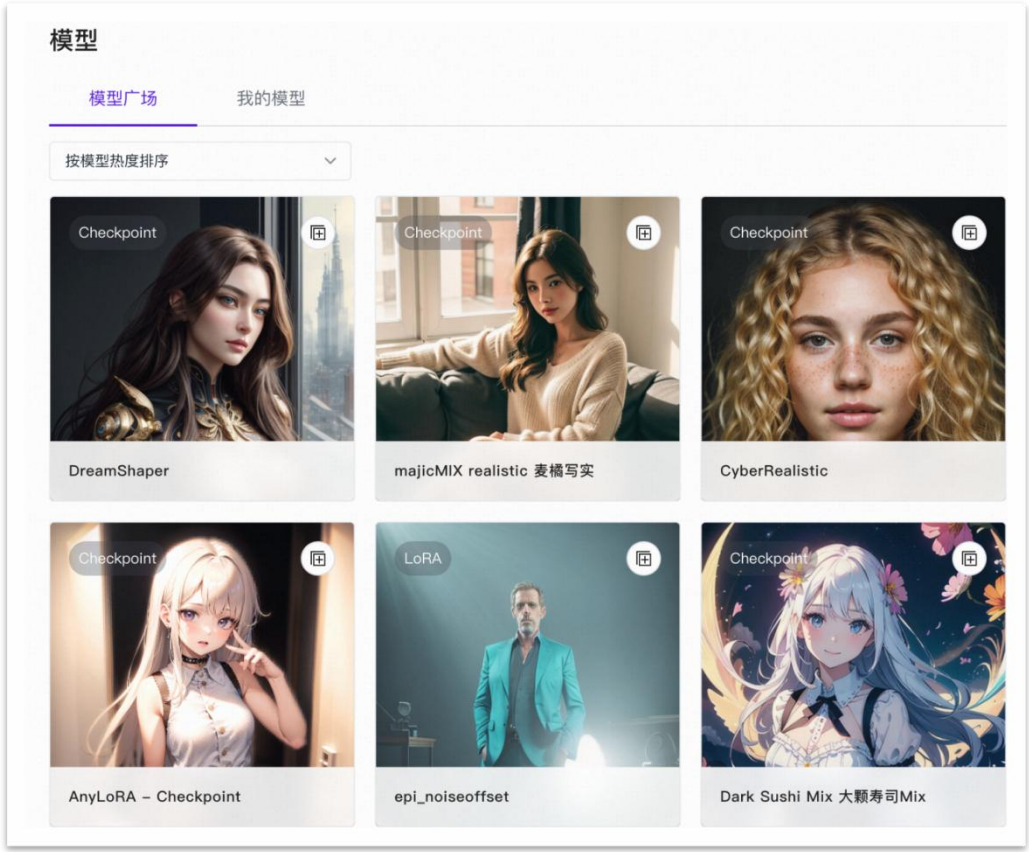
```python
# Huggingface模型部署
from pai.common.utils import random_str
from pai.huggingface.model import HuggingFaceModel

# 创建 HuggingFaceModel 对象
m = HuggingFaceModel(
    command="python app.py",    # 指定服务启动命令
    transformers_version="latest",  # 指定服务镜像所使用的 transformers 版本，目前支持 4.29.2 或 latest
    # 通过环境变量传入模型ID、任务类型和模型版本，与 Huggingface Modelhub 保持一致
    environment_variables={
        "MODEL_ID": "distilbert-base-uncased-finetuned-sst-2-english",
        "TASK": "text-classification",
        "REVISION": "main",
    }
)

# 部署服务
p = m.deploy(
    service_name="hf_bert_serving_{}".format(random_str(6)),
    instance_count=1,
    instance_type="ecs.gn6i-c4g1.xlarge",
    options={
        "metadata.rpc.keepalive": 5000000,
        "features.eas.aliyun.com/extra-ephemeral-storage":"40Gi",
    },
)

# 对部署好的服务进行一些测试
res = p.predict(
    {
        "data": ["it's so easy!"]
    }
)
```

# PAI × ModelScope

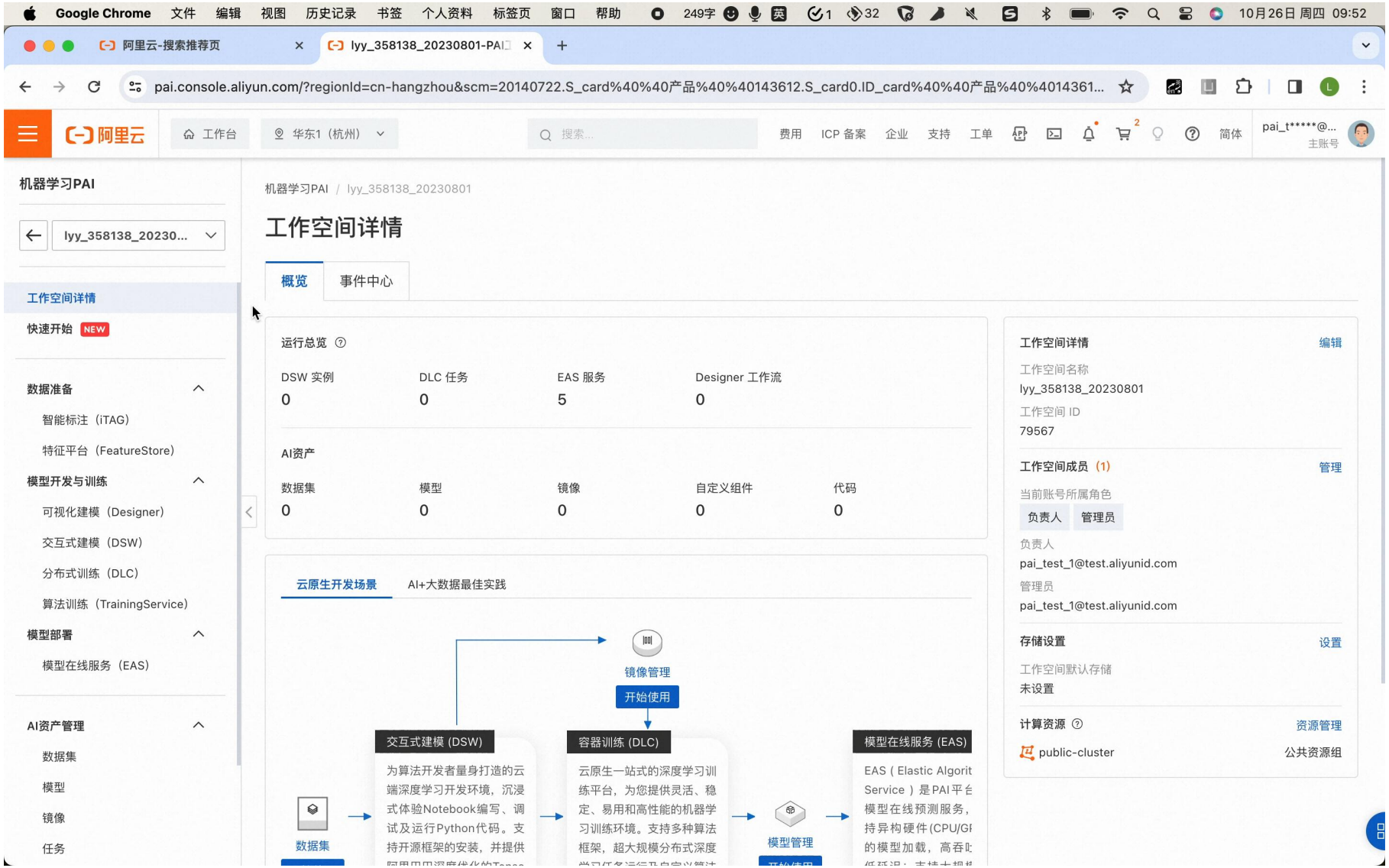



# Case study：PAI AI设计师专区



# Demo：通过快速开始完成qwen的训练和部署



# PART 03
# 未来展望

# 未来展望

MaaS

基础模型

紧跟开源社区，拓展模型覆盖度

算法工程优化

算法定义配置标准化
错误信息规范化

平台工程优化

提升模型开发效率
集成平台优化能力

# Qwen训练的一些细节

~1000

实验次数

2天

7B模型全参数微调训练时间

10min

关注loss的时间间隔

# MLOps: Efficiency is All You Need

# MLOps

训练监控

实验管理

模型血缘

模型评测

数据质量

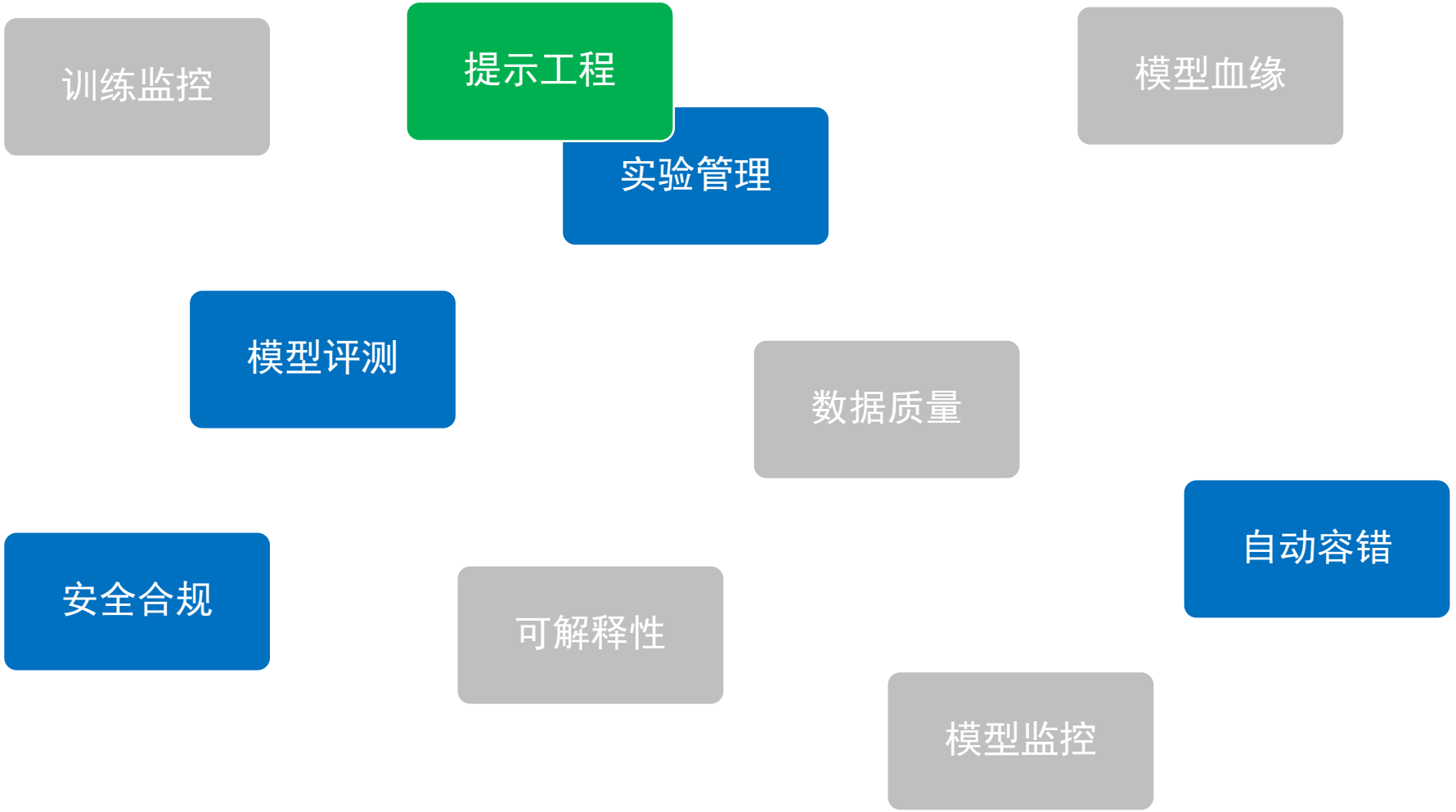
自动容错

安全合规

可解释性

模型监控

# MLOps -> LLMOps



# 深度结合平台优化能力

PAI-ACC AI加速服务：提供AI工作流完整的加速能力

数据集加速器

DatasetAccelerator

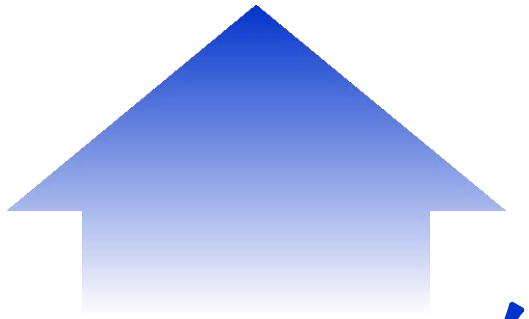
分布式训练加速

TorchAcc

推理加速

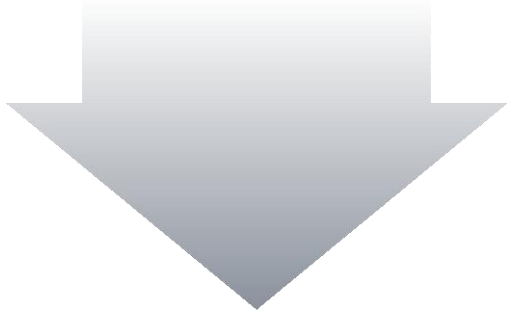
Blade BladeLLM

# BladeLLM： LLM高性能推理引擎



1.7～3.8倍

服务吞吐提升



8.7～13.8倍

首包延迟降低

AiDD AI+ 软件研发数字峰会
AI+ software Development Digital summit

# THANKS