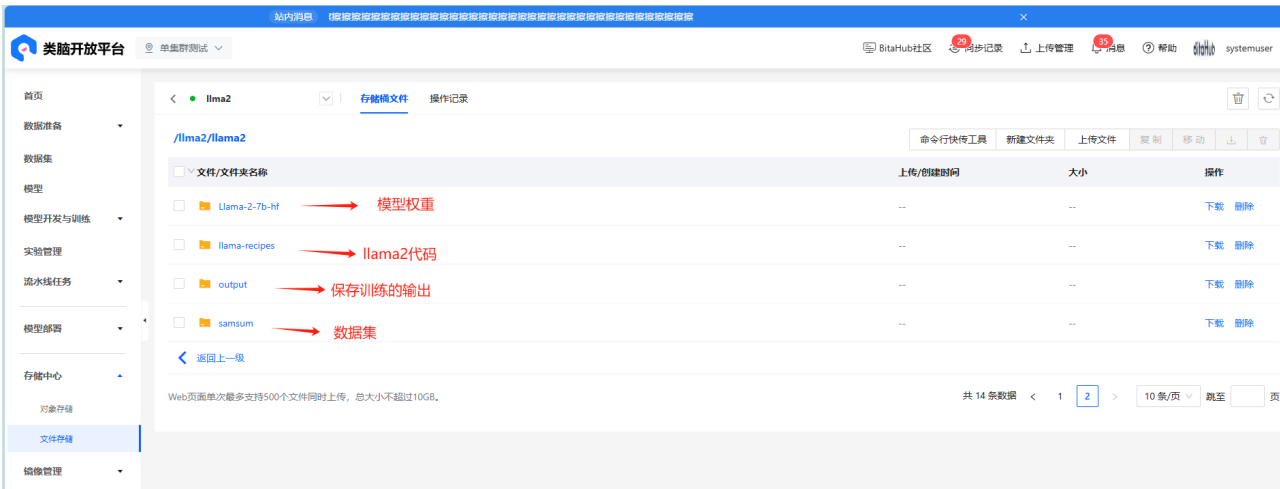
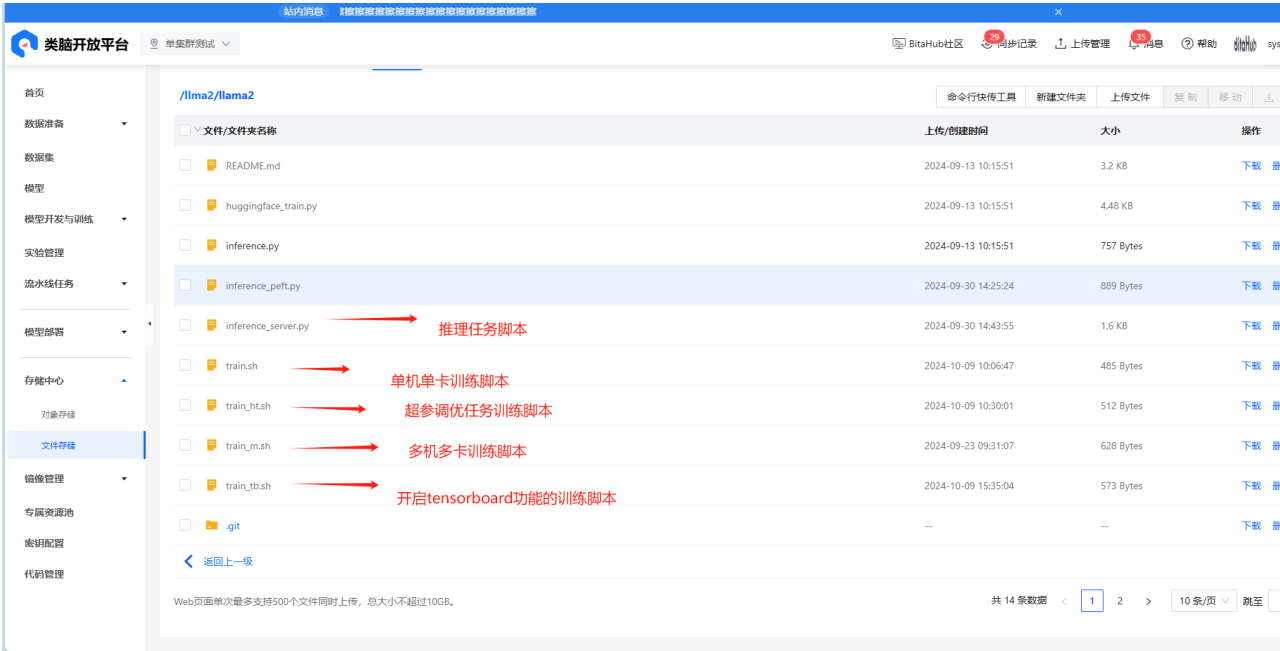


数据准备

- 1. 预训练模型
- 2. 训练脚本
- 3. 数据集



分布式训练

训练镜像：平台镜像 hero-dev-image.cnbita.com:5000/systemuser/llama2:v1

该镜像主要提供cuda运行环境。以及一些可供使用的linux调试工具，

- 1. 点击模型开发与训练功能栏下的分布式训练
- 2. 点击创建分布式任务

类脑开放平台 单集群测试

BitHub社区 同步记录 上传管理 消息 帮助 systemuser

首页

数据准备

数据集

模型

模型开发与训练

开发环境

分布式训练

超参调优

实验管理

流水线任务

模型部署

存储中心

镜像管理

专属资源池

密钥配置

代码管理

创建

任务名称

Q 请输入

任务名称	套餐规格	训练镜像	启动时间	结束时间	任务标签	状态	操作
llama2-s1-copy-copy-copy-cop...	nvidia-rtx-3090-24GB-16C-30.7...	llama2v1	--	2024-09-30 09:19:25	--	停止	停止 删除 复制
llama2-s1-copy-copy-copy-cop...	nvidia-rtx-3090-24GB-16C-30.7...	llama2v1	2024-09-24 17:47:44	2024-09-25 03:50:42	--	成功	停止 删除 复制
llama2-s1-copy-copy-copy-copy	nvidia-rtx-3090-24GB-16C-30.7...	llama2v1	--	2024-09-24 17:46:48	--	停止	停止 删除 复制
test-copy	nvidia-rtx-3090-24GB-16C-30.7...	llama2v1	2024-09-23 13:58:11	2024-09-23 13:58:11	--	失败	停止 删除 复制
test	nvidia-rtx-3090-24GB-16C-30.7...	llama2v1	2024-09-23 13:51:03	2024-09-23 13:51:03	--	失败	停止 删除 复制
dis-copy1-copy-copy-copy...	testPackage	experiment-hpcsv1	--	2024-09-23 13:43:34	--	停止	停止 删除 复制
llama2-s1-copy-copy-copy-copy	nvidia-rtx-3090-24GB-16C-30.7...	llama2v1	--	2024-09-23 10:55:45	--	停止	停止 删除 复制
llama2-s1-copy-copy	nvidia-rtx-3090-24GB-16C-30.7...	llama2v1	--	2024-09-23 09:48:36	--	停止	停止 删除 复制
llama2-s1-copy	nvidia-rtx-3090-24GB-16C-30.7...	llama2v1	2024-09-20 16:08:16	2024-09-20 17:17:16	--	成功	停止 删除 复制
llama2-s1	nvidia-rtx-3090-24GB-16C-30.7...	llama2v1	2024-09-20 10:55:32	2024-09-20 10:56:35	--	失败	停止 删除 复制

共 17 条数据 < 1 2 > 10 条/页 跳至 页

创建分布式任务：

1. 填写任务名称
2. 选择挂载路径，根据上面保存数据的存储类型和存储路径来选择
3. 选择微调模型的输出路径
4. 选择平台提供的镜像或者自定义镜像
5. 计算框架选择其它，选择用的资源池类型以及相应套餐，根据数据挂载路径填写任务的微调命令，选择是单节点任务还是分布式任务
6. 如果训练代码不是通过存储挂载的话，可以在高级配置中选择代码仓库进行挂载
7. 点击确认即可

基础信息

任务名称

llama2-a-2

任务标签

128字符以内，不支持反斜杠\/*?<>|和空格，回车保存标签

挂载路径

对象存储

/llama2

...

挂载路径: /input/DemoTraining/llama2

对象存储

文件存储

数据集

模型

保存路径

对象存储

/llama2/output

...

挂载路径: /output/DemoTraining/llama2/output

实验组

请选择实验组

镜像选择

平台镜像

私有镜像

共享镜像

仓库地址

llama2v1

hero-dev-image.cnbita.com:5000/systemuser/llama2v1

运行环境

资源池

公共资源

专属资源

* 计算框架

DeepSpeed

Colossal-AI

其他

* 任务配置

* Role

task1

* 实例规格

nvidia-rtx-3090-24GB-16C-30.74GB

1

* 启动命令

bash /input/DemoTraing/llama2/train.sh

新增 Role (1/10)

WebTerminal ☐

高级配置

训练代码

私有仓库

公开仓库

①

您可以直接引用在「代码」模块维护的代码仓库来进行训练。代码在容器中的映射路径为/code。

• 授权代码

请选择授权代码

• 代码仓库

请选择代码仓库

• 代码分支

请选择代码分支

环境变量 [+ 环境变量 \(0/10\)](#)

- * 算力支付
 - ☒ 个人账户支付 剩余算力156089.13
 - ☐ test22支付 剩余算力156089.13
 - ☐ ccd支付 剩余算力1073.5
 - ☐ team支付 剩余算力194.02
 - ☐ team11支付 剩余算力156089.13
 - ☐ shao_team_test支付 剩余算力156089.13
 - ☐ first支付 剩余算力156089.13

基础信息

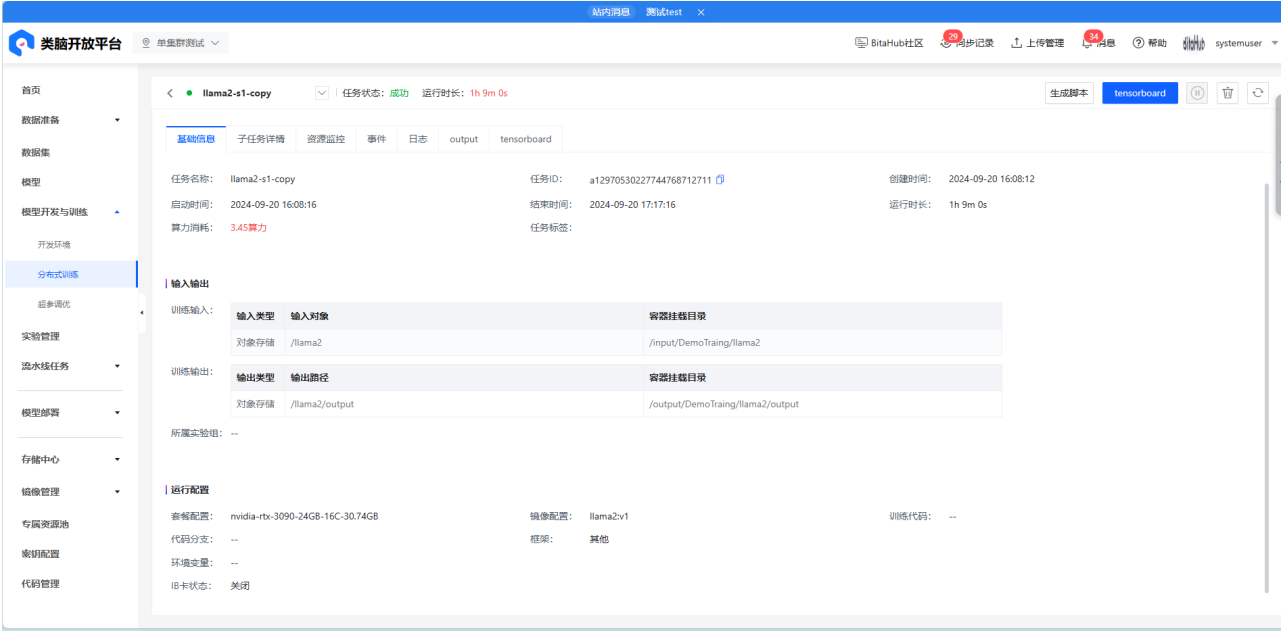
运行环境

高级配置

运行环境

高级配置

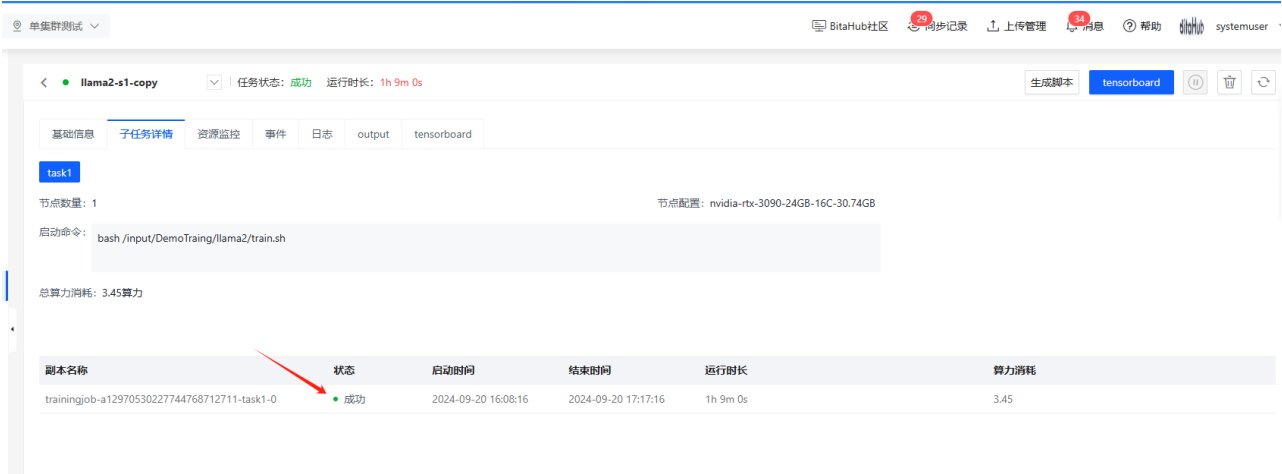
运行结果展示如下：



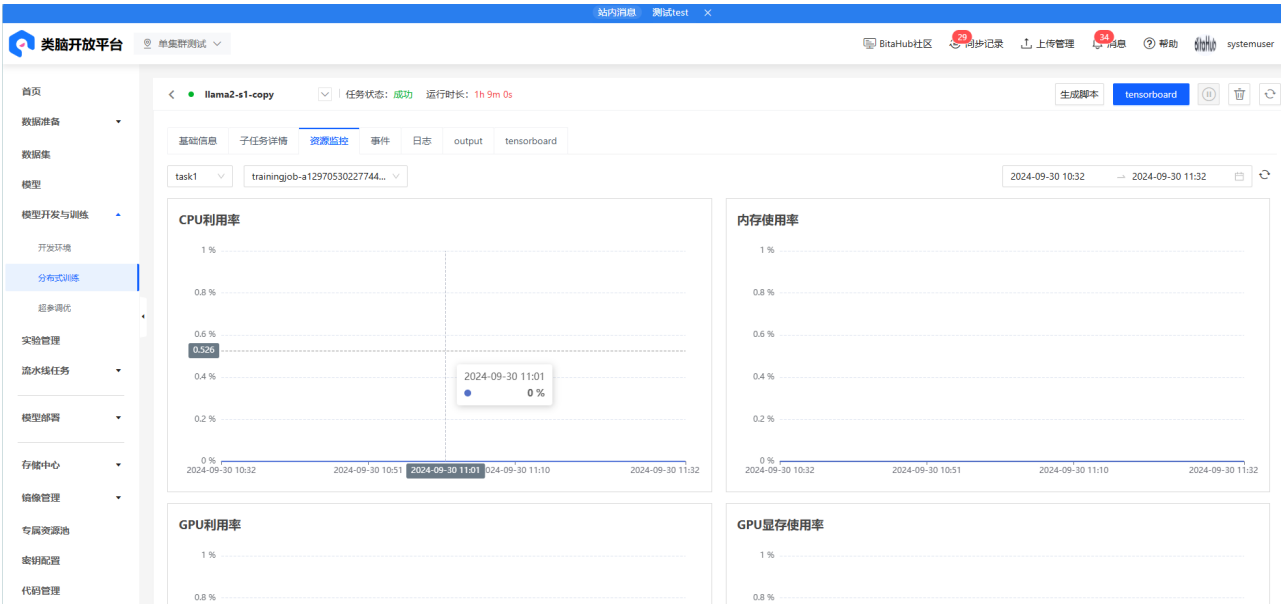
模块功能介绍：

基础信息：展示任务的基础环境配置信息

子任务详情：展示pod级别的任务状态和数量，以及对应的资源配置



资源监控：展示服务消耗的资源情况

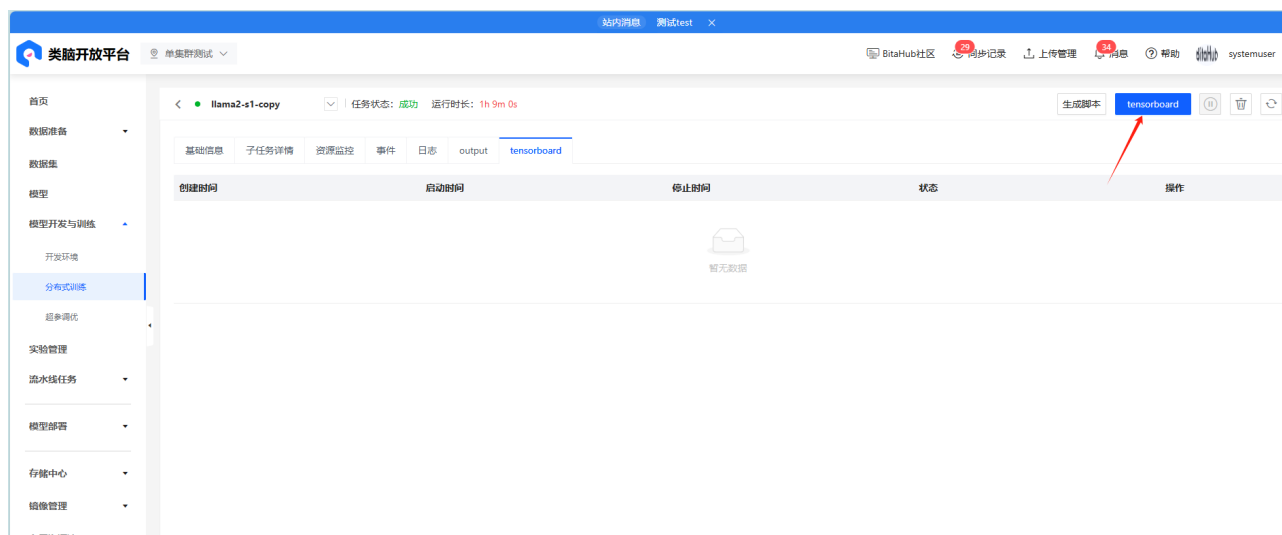


事件管理：展示任务训练过程中的关键性流程信息

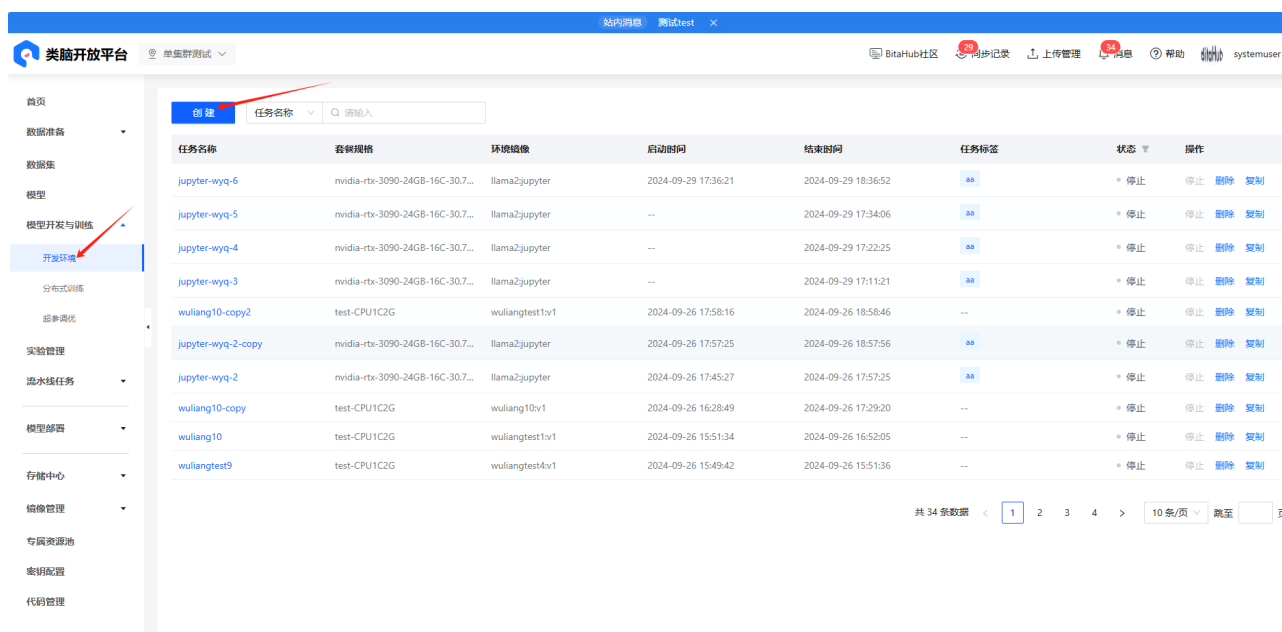
日志管理：展示训练任务的运行日志

输出设置：展示模型训练过程中的中间产出

tensorboard: 根据输出的文件夹开启tensorboard服务



创建jupyter任务



创建jupyter任务：

1. 填写任务名称
2. 选择挂载路径和文件保存路径
3. 选择运行环境，选择资源池和套餐规格
4. 选择访问方式，jupyter，其它选项可以点击
5. 选择运行时长
6. 高级配置，可以选择挂载代码仓库

< 返回

| 创建交互式开发环境

基础信息

* 任务名称

jupyter-7

任务标签

aa x

挂载路径

文件存储 /lama2 | ...

挂载路径: /input/lama2/lama2

+ 对象存储

+ 文件存储

+ 数据集

+ 模型

保存路径

文件存储 /lama2/output | ...

挂载路径: /output/lama2/lama2/output

* 镜像选择

平台镜像私有镜像共享镜像仓库地址

lama2:jupyter

运行环境

资源池类型公共资源专属资源

* 实例规格

套餐名称	CPU核心数	内存 (GB)	GPU	显存	定价 (算力/小时)
<input type="radio"/> test-CPU1C2G	1	2	--	--	1

* 访问方式

☒ JupyterLab☐ SSH☒ Web terminal☐ 远程桌面

* 运行时长

☒ 1小时☐ 2小时☐ 3小时☐ 自定义

高级配置

训练代码私有仓库公开仓库

① 您可以直接引用在「代码」模块维护的代码仓库来进行训练。代码在容器中的映射路径为/code。

* 授权代码

请选择授权代码

* 代码仓库

请选择代码仓库

* 代码分支

请选择代码分支

* 算力支付

☒ 个人账户支付 剩余算力156067.25

- ☐ test22支付 剩余算力156067.25
- ☐ cdd支付 剩余算力1073.5
- ☐ team支付 剩余算力194.02
- ☐ team11支付 剩余算力156067.25
- ☐ shao_team_test支付 剩余算力156067.25
- ☐ first支付 剩余算力156067.25

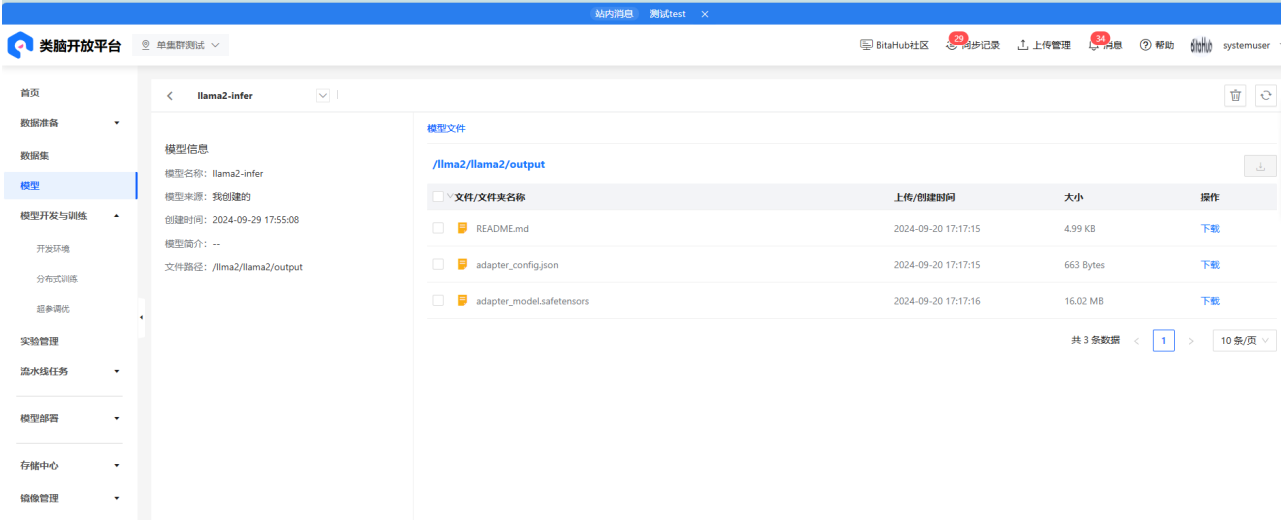
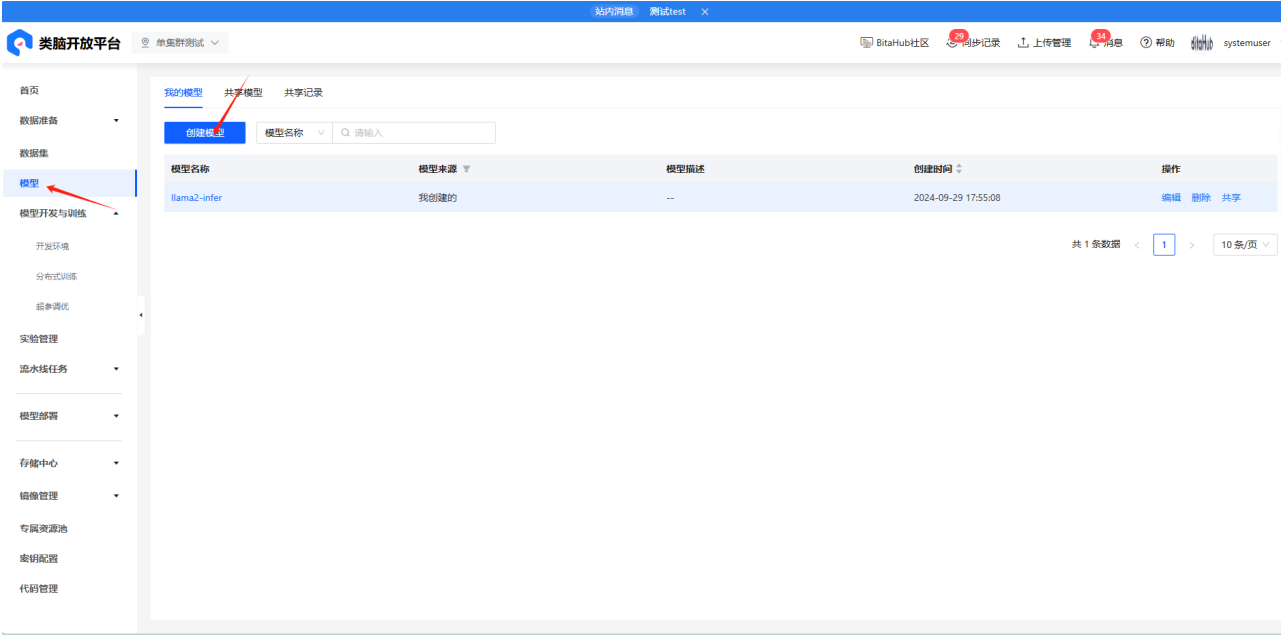
任务详情模块

任务基础信息

资源监控

事件

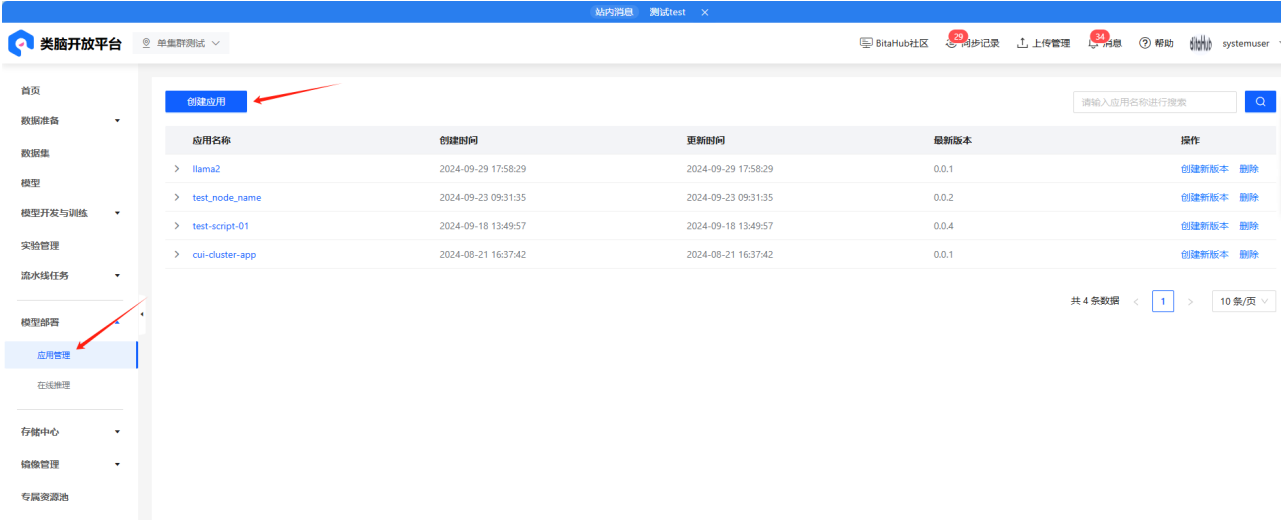
创建模型



推理

1 创建应用

2 创建推理服务



创建应用

1. 填写应用名称，版本，应用厂商
2. 填写引用信息，输入类型填写第四步保存的微调模型以及在文件存储中保存的预训练模型
3. 设置端口号和镜像

< 返回

应用管理/创建新版本

基础信息

* 应用名称:

llama2

* 应用版本:

0.0.6

应用描述:

不超过500字符

应用标签:

128字符以内，不支持\:"'?"<>|和空格，回车保存标签

* 应用厂商:

验证在线识别应用

应用信息

* 应用输入:

输入类型	输入对象	容器挂载目录	操作
模型	llama2-infer	/model/llama2/output	<div></div> <div></div>
文件存储	/llama2	/input/llma2/llama2	<div></div> <div></div>

+ 新增应用信息

应用指标:

请输入应用指标

0-1之间的数值

+ 新的应用指标

部署信息

* 端口号:

12345

* 部署镜像:

平台镜像

私有镜像

共享镜像

llama2jupyter

...

部署代码:

私有仓库

公开仓库

① 您可以直接引用在「代码」模块维护的代码仓库来进行训练。代码在容器中的映射路径为/code。

* 授权代码:

请选择授权代码

* 代码仓库:

请选择代码仓库

* 代码分支:

请选择代码分支

环境变量:

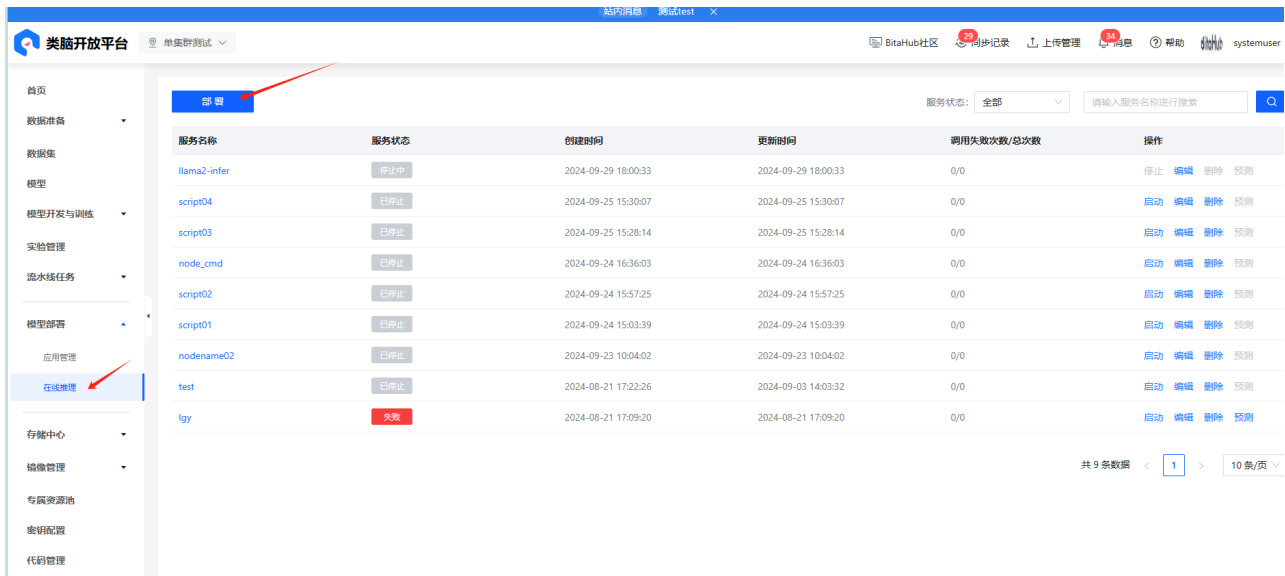
+ 环境变量

(0/10)

启动命令:

cd /input/llma2/llama2: python3 inference_server.py --model_path /input/llma2/llama2/Llama-2-7b-hf --peft_model_path /model/llama2/output

创建推理服务



流程：

- 1 填写推理指标名称
2. 选择应用来源，上一步创建的llama2-v2 0,0.5版本
3. 选择默认生成的副本数，以及套餐和服务地址，网关配置

返回

在线推理/部署

基础信息

服务名称

llama2-infer

选择资源池

公共资源

专属资源

应用来源

我的应用

应用市场

服务配置信息

服务1

选择应用

llama2

0.0.5

默认副本数

1

节点配置

test-CPU1C2G

服务地址

llama2

服务访问路径为: http://llama2:12345

环境变量

请输入Key

请输入Value

删除

新的环境变量

网关配置

llama2 0.0.5

/

高级配置

API认证

关

单次部署失败最大重试次数

2

次

最长等待时间

关

支持副本数为0

关

弹性伸缩

关

类脑开放平台

单集群测试

Bitahub社区 同步记录 上传管理 消息 帮助 systemuser

首页

数据准备

数据集

模型

模型开发与训练

实验管理

流水线任务

模型部署

应用管理

在线推理

存储中心

镜像管理

专属资源池

密钥配置

代码管理

llama2-infer

任务状态: 运行中 运行时长: 8d 18h 33m 9s

基础信息 历史记录 在线识别 资源监控 事件 服务日志

服务信息

接口地址: <https://hero-dev.cnbita.com/inf-app/a13073335257657344897706>

API认证: 关闭

调用失败次数/总次数: 0/1

基础信息

服务名称: llama2-infer

创建时间: 2024-09-29 18:00:33

更新时间: 2024-09-30 14:49:37

高级配置

单次部署最大失败数: 2

最长等待时间: 关

支持副本数为0: 关

弹性伸缩: 关

使用服务得接口地址请求推理服务

```
myo@DESKTOP-PVL83Q3 MINGW64 /d/GoPro/src/functionalVerification
$ curl -X POST "https://hero-dev.cnbita.com/inf-app/a13073335257657344897706/generate/" -H "Content-Type: application/json" -d '{"text": "Hello my name is"}'
{"generated_text": "Hello my name is Katie. I'm a 23 year old female. I'm looking for a"}
```

超参调优

类脑开放平台

单集群测试

Bitahub社区 同步记录 上传管理 消息 帮助 systemuser

首页

数据准备

数据集

模型

模型开发与训练

开发环境

分布式训练

超参调优

实验管理

流水线任务

模型部署

存储中心

镜像管理

创建

任务名称

Q 请输入

任务名称	进度	状态	创建时间	更新时间	最优调参	任务标签	操作
hpo001-1-copy-copy-copy-copy	2/2	停止	2024-09-12 15:37:37	2024-09-12 16:08:48	--	hpo. hpo	停止 删除 复制
hpo001-1-copy-copy-copy	2/2	停止	2024-09-12 15:36:01	2024-09-12 15:37:13	--	hpo. hpo	停止 删除 复制
hpo001-1-copy-copy	2/2	停止	2024-09-12 15:27:57	2024-09-12 15:28:49	--	hpo. hpo	停止 删除 复制
hpo001-1-copy	2/2	停止	2024-09-12 15:19:02	2024-09-12 15:26:23	--	hpo. hpo	停止 删除 复制
hpo001-1	2/2	成功	2024-09-12 15:03:03	2024-09-12 15:03:43	accv=8.5	hpo. hpo	停止 删除 复制
hpo001-copy-copy-copy-copy-...	2/2	停止	2024-09-12 14:30:29	2024-09-12 14:37:48	--	hpo. hpo	停止 删除 复制
hpo001-copy-copy-copy-copy-...	2/2	停止	2024-09-09 19:13:21	2024-09-12 14:32:28	--	hpo. hpo	停止 删除 复制
hpo001-copy-copy-copy-copy-...	2/2	停止	2024-09-09 19:13:17	2024-09-12 14:32:25	--	hpo. hpo	停止 删除 复制
hpo001-copy-copy-copy-copy-...	2/2	停止	2024-09-09 19:12:36	2024-09-12 14:32:22	--	hpo. hpo	停止 删除 复制
hpo001-copy-copy-copy-copy-copy	2/2	停止	2024-09-09 19:12:18	2024-09-12 14:32:18	--	hpo. hpo	停止 删除 复制

创建超参调优任务

基于训练得gamma来进行超参调优，
需要在任务命令中添加\${hpo.gamma}，如果想要微调其它参数，可按照\${hpo.超参数名称}来进行添加，具体参数设置如下：

基础信息

*

任务名称

llama2-copy-copy

任务标签

128字符以内，不支持"/:*?*<>|和空格，回车保存标签

挂载路径

文件存储

/llama2

...

挂载路径: /input/llama2/llama2

+ 对象存储

+ 文件存储

+ 数据集

+ 模型

保存路径

文件存储

/llama2/output

...

挂载路径: /output/llama2/llama2/output

*

镜像选择

平台镜像

私有镜像

共享镜像

llama2jupyter

基础信息

[运行环境](#)

[高级配置](#)

运行环境

资源池

公共资源

专属资源

*

计算框架

DeepSpeed

Colossal-AI

其他

*

任务配置

*

Role

task1

*

实例规格

nvidia-rtx-3090-24GB-8C-20GB

1

*

启动命令

bash /input/llama2/llama2/train_ht.sh \${hpo.gamma}

+ 新增 Role

(1/10)

高级配置

训练代码

私有仓库

公开仓库

①

您可以直接引用在「代码」模块维护的代码仓库来进行训练。代码在容器中的映射路径为/code。

*

授权代码

请选择授权代码

*

代码仓库

请选择代码仓库

*

代码分支

请选择代码分支

环境变量

+ 环境变量

(0/10)

*

超参数

名称	类型	搜索空间
gamma	浮点数(Double)	0.75<=gamma<=0.95, step=0.1

*

优化指标

gamma

*

优化方向

最大化

最小化

目标值

请输入目标值

*

计算方式

最新输出的指标数值

过程中最优的指标数值

*

输出指标到

stdout(标准输出)

输出格式: 指标名=数值 例如: gamma=0.8
正则表达式匹配规则: [(w|-|.)+](v"=|"+)[(-|~|d"\\d+|)]([e|E|+|-|n|d+|)]

*

搜索算法

随机搜索

*

最大搜索次数

2

最大并发数

1

*

算力支付

个人账户支付

剩余算力154572.2

test22支付

剩余算力154572.2

cdd支付

剩余算力1073.5

team支付

剩余算力194.02

team11支付

剩余算力154572.2

shao_team_test支付

剩余算力154572.2

first支付

剩余算力154572.2

超参调优任务详情

llama2-copy

任务状态: 运行中 运行时长: 3m 9s

基本信息

Trial任务列表

基础信息

任务名称: llama2-copy

任务ID: a13183048935141376845969

创建时间: 2024-10-09 10:31:21

更新时间: 2024-10-09 10:32:07

最优调参任务: --

最优调参指标: --

最优调参参数: --

超参配置

超参数:

名称	类型	搜索空间
gamma	double	0.75<=gamma<=0.95, step=0.1

优化指标: gamma

优化方向: 最小化

目标值: --

计算方式: 过程中最优的指标数值

输出指标到: stdout(标准输出)

搜索算法: 随机搜索

最大搜索次数: 2

最大并发搜索: 1

任务配置

套餐配置: nvidia-rtx-3090-24GB-8C-20GB

镜像配置: llama2jupyter

训练代码: --

代码分支: --

训练输入:

输入类型	输入对象	容器挂载目录
文件存储	/llama2	/input/llama2/llama2

训练输出:

输出类型	输出路径	容器挂载目录
文件存储	/llama2/output	/output/llama2/llama2/output

启动命令:

bash /input/llama2/llama2/train_ht.sh \$(hpo.gamma)

环境变量: --

点击trail任务列表，查看调优任务详情

站内消息 测试test

BitHub社区 同步记录 上传管理 消息 帮助 systemuser

数据准备 数据集 模型 模型开发与训练 开发环境 分布式训练 超参调优 实验管理 流水线任务 模型部署 存储中心 镜像管理 专属资源池 密钥配置 代码管理

llama2-copy

任务状态: 运行中 运行时长: 4m 51s

基本信息

Trial任务列表

gamma

gamma

Trial任务列表

任务名称	指标 (gamma)	超参数 (gamma)	开始时间	结束时间	时长	状态
llama2-copy-7jwnk5g	--	0.75	2024-10-09 10:32:26	--	3m 47s	运行中

点击具体子任务名称，可以查看子任务详情

类脑开放平台

单集群测试

BitHub社区

同步记录

上传管理

消息

帮助

systemuser

首页

数据准备

数据集

模型

模型开发与训练

开发环境

分布式训练

实验管理

流水线任务

模型部署

存储中心

镜像管理

专属资源池

密钥配置

代码管理

llama2-copy-7jwnk...

任务状态: 运行中 运行时长: 4m 56s

生成脚本

tensorboard

暂停

删除

刷新

基础信息

资源监控

事件

日志

output

tensorboard

运行信息

任务名称: llama2-copy-7jwnk5g 任务ID: a13183054472671232661789 创建时间: 2024-10-09 10:32:04

启动时间: 2024-10-09 10:32:26 结束时间: -- 运行时长: 4m 56s

算力消耗: -- 任务标签:

输入输出

训练输入:

输入类型	输入对象	容器挂载目录
文件存储	/llama2	/input/llma2/llama2

训练输出:

输出类型	输出路径	容器挂载目录
文件存储	/llama2/output	/output/llma2/llama2/output

配置信息

镜像配置: llama2:jupyter 训练代码: -- 代码分支: --

框架: -- 环境变量: --

显卡状态: 关闭

实验管理

实验管理是将每个任务的保存的tensorboard结果进行比较，因为llama2m 每次训练保存的结果的位置是相同的，

点击实验管理

创建实验组

类脑开放平台

单集群测试

BitHub社区

同步记录

上传管理

消息

帮助

systemuser

首页

数据准备

数据集

模型

模型开发与训练

实验管理

流水线任务

模型部署

应用管理

在线推理

存储中心

镜像管理

专属资源池

密钥配置

创建实验组

请输入实验组名称进行搜索

实验组名称

实验组描述

创建人

创建时间

指标对比

操作

11	--	systemuser	2024-09-10 18:12:14	未启动	编辑 删除
----	----	------------	---------------------	-----	-------

共 1 条数据 < 1 > 10 条/页

创建实验组

* 实验组名称:

llama

实验组描述:

不超过500字符

取消

确定

点击llama实验组

点击添加任务

类脑开放平台

Bitahub社区

同步记录

上传管理

消息

帮助

systemuser

首页

数据准备

数据集

模型

模型开发与训练

实验管理

流水线任务

模型部署

应用管理

在线推理

存储中心

镜像管理

返回

llama

添加训练任务

请输入训练任务名称进行搜索

开启指标对比

关闭指标对比

列名编辑

训练任务名称	任务标签	运行时间	状态	操作
暂无数据				

类脑开放平台

Bitahub社区

同步记录

上传管理

消息

帮助

systemuser

首页

数据准备

数据集

模型

模型开发与训练

实验管理

流水线任务

模型部署

应用管理

在线推理

存储中心

镜像管理

返回

llama2

添加训练任务

请输入任务名称进行搜索

任务名称

套规格

训练镜像

启动时间

结束时间

状态

<input type="checkbox"/>	llama2-s10	test-CPU1C2G	llama2	2024-10-10 14:11:51	2024-10-10 14:13:31	成功
<input type="checkbox"/>	llama2-s9-co...	test-CPU1C2G	llama2	2024-10-10 13:59:58	2024-10-10 14:01:38	成功
<input type="checkbox"/>	llama2-s9	nvidia-rtx-30...	llama2	2024-10-10 09:43:08	2024-10-10 13:06:24	成功
<input type="checkbox"/>	llama2-s8	nvidia-rtx-30...	llama2	2024-10-10 09:11:00	2024-10-10 09:42:38	停止
<input type="checkbox"/>	llama2-s4	nvidia-rtx-30...	llama2	2024-10-09 09:52:40	2024-10-09 09:52:40	失败

共 6 条数据

1

2

5 条/页

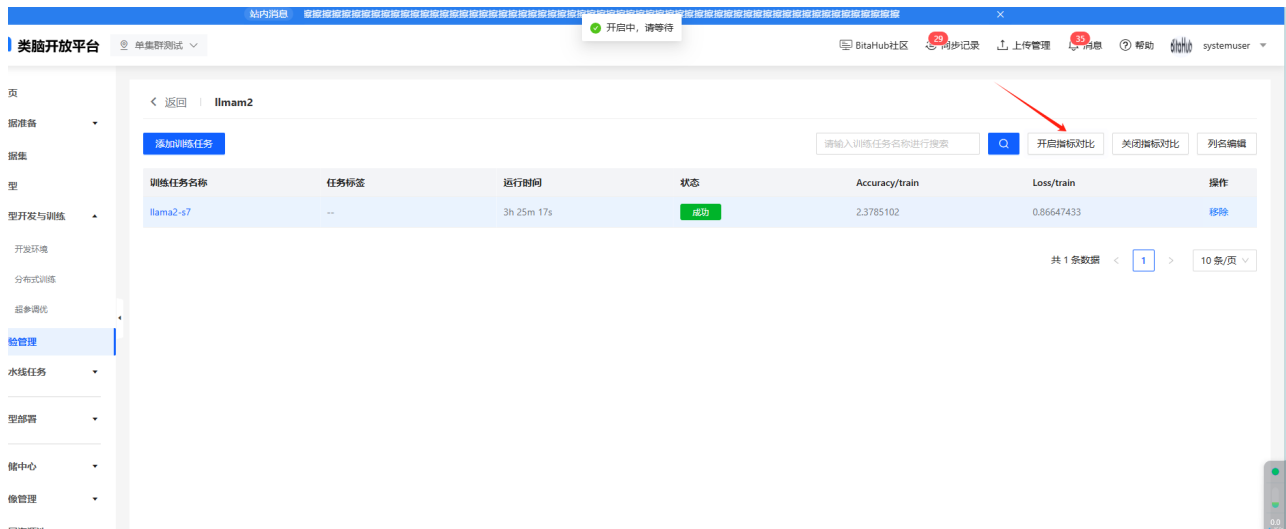
跳至

页

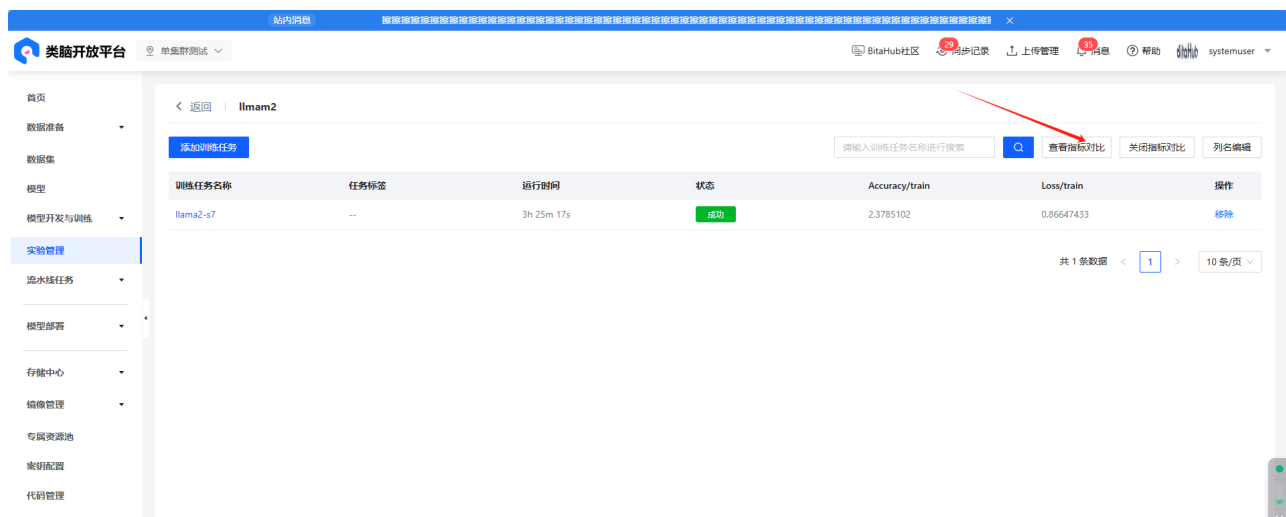
取消

确定

创建完开启指标



等待片刻，点击查看指标对比



结果如下

