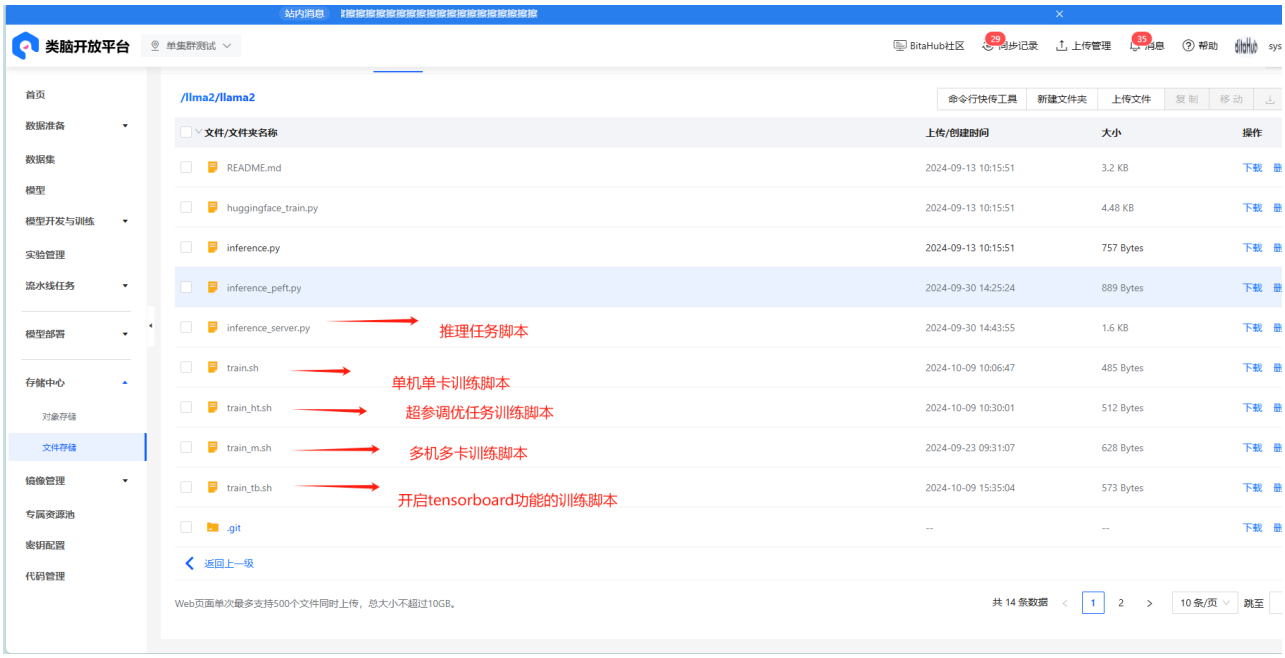
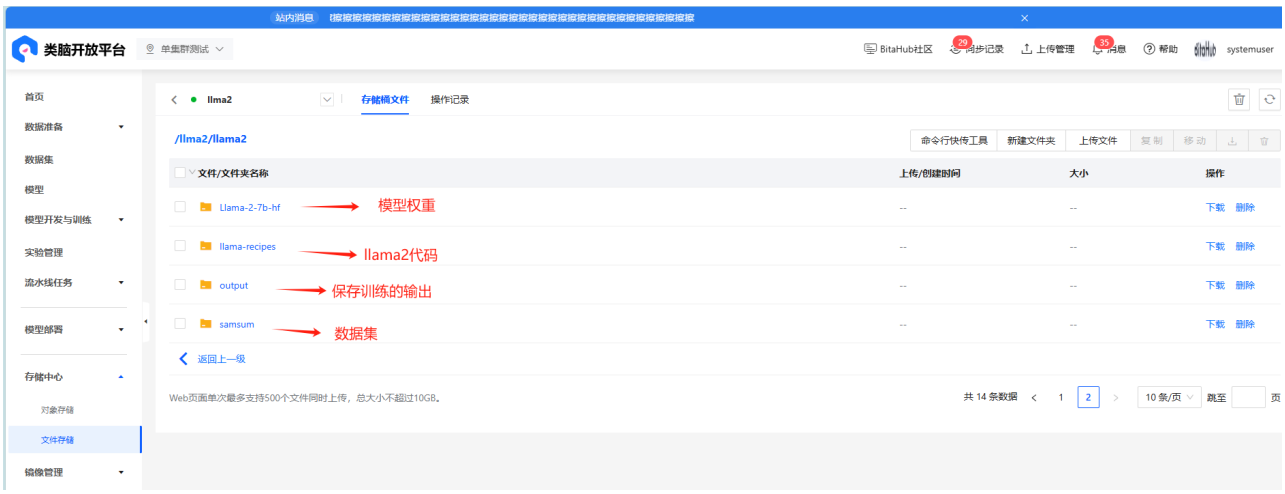
# 数据准备

1. 预训练模型
2. 训练脚本
3 .数据集



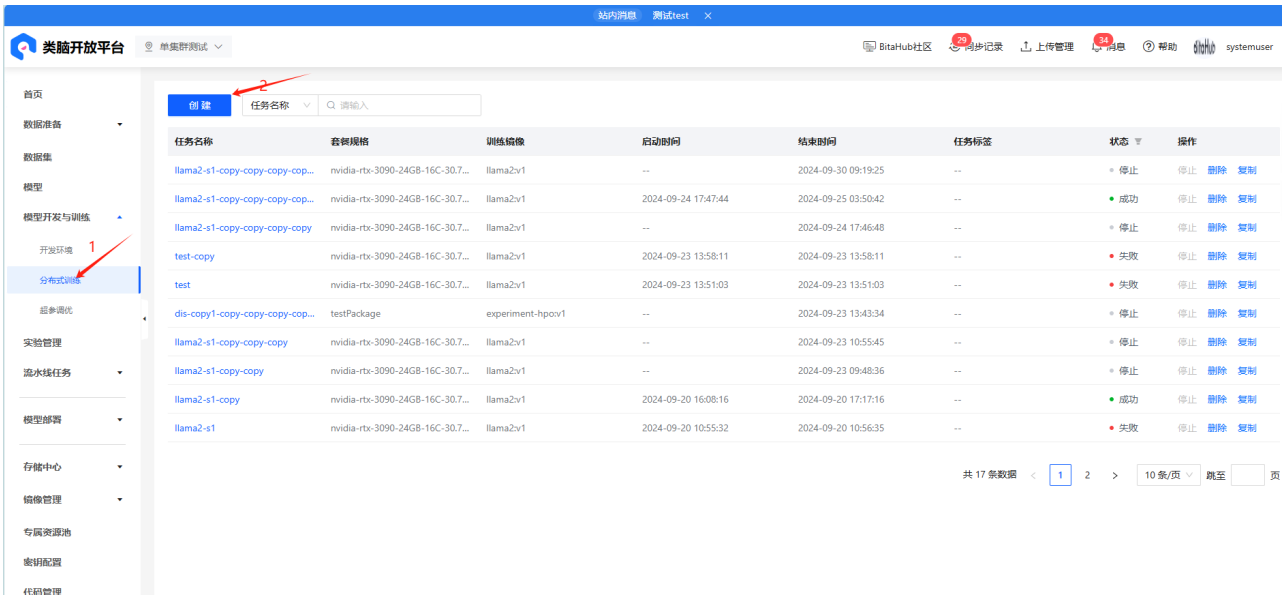




# 分布式训练

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

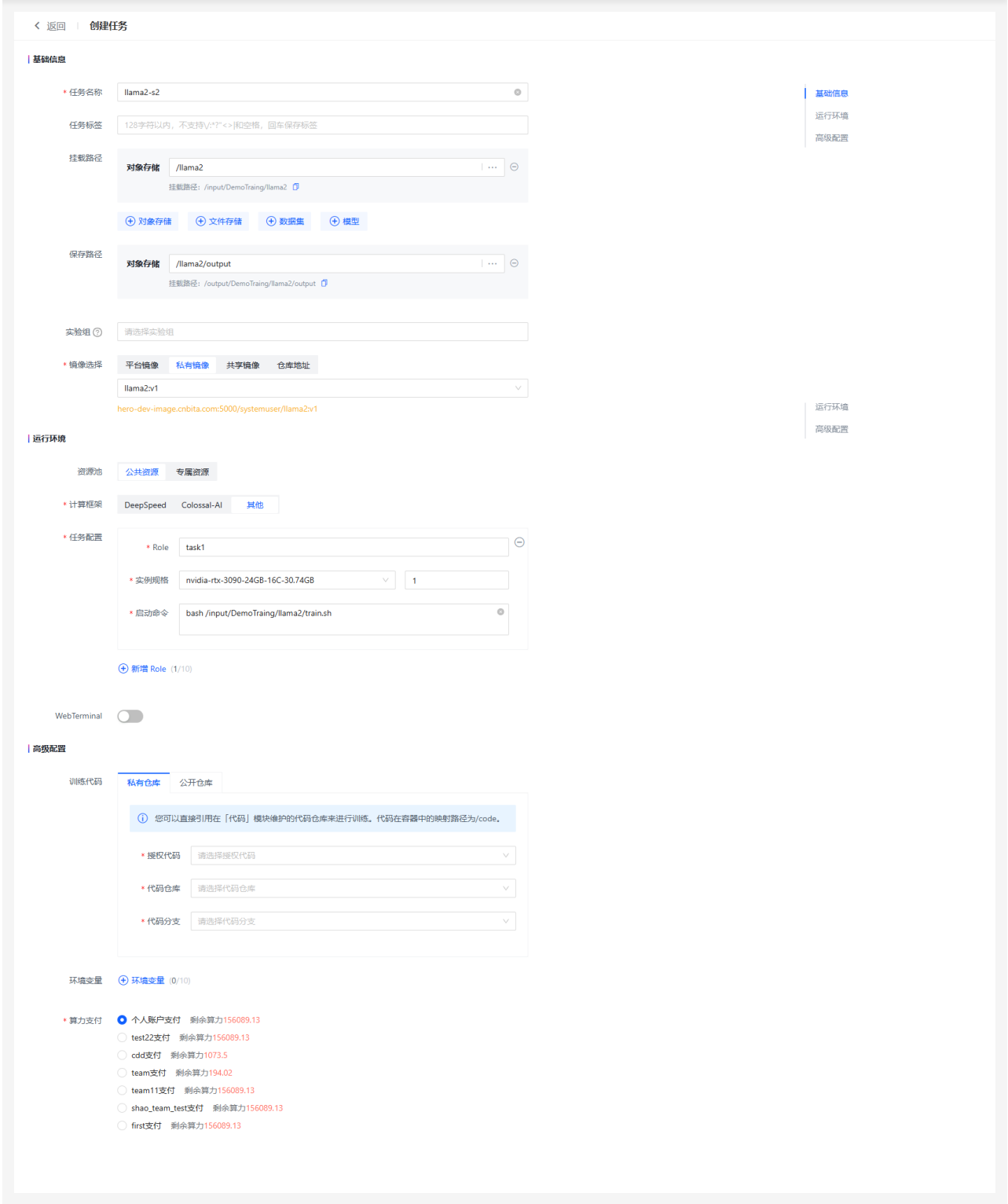
该镜像主要提供cuda运行环境。以及一些可供使用的linux调试工具，

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

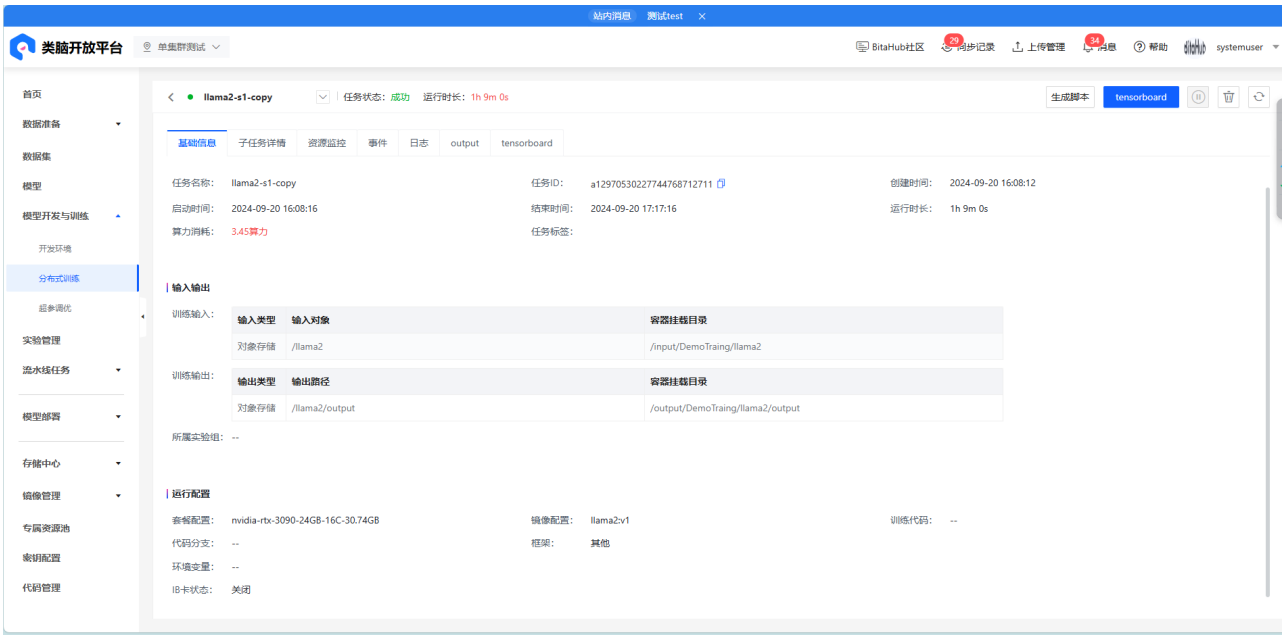



创建分布式任务：

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


返回
创建任务
基础信息
* 任务名称
llama2-s2
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
对象存储
/llama2
挂载路径：/input/DemoTraing/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/DemoTraing/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
llama2:v1
hero-dev-image.cnbita.com:5000/systemuser/llama2:v1
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
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
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

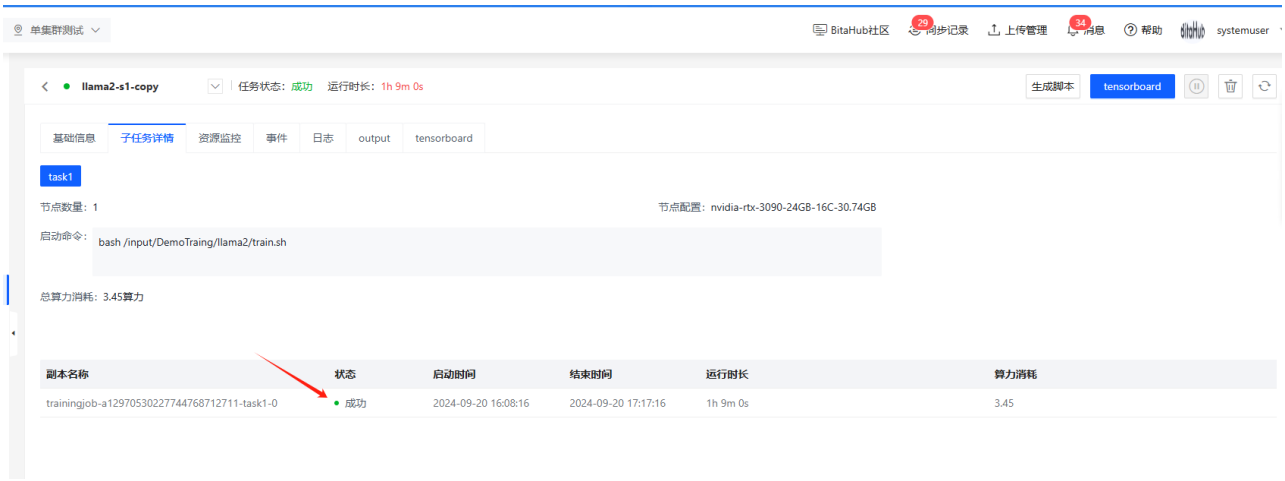
运行结果展示如下：



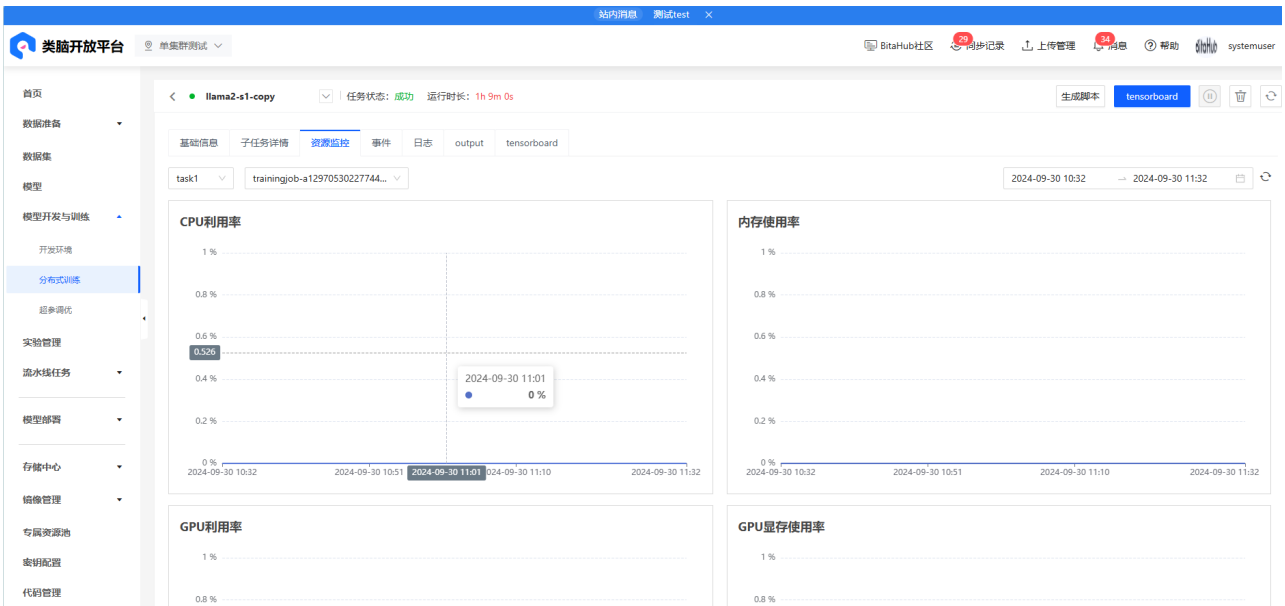

模块功能介绍：

基础信息：展示任务的基础环境配置信息

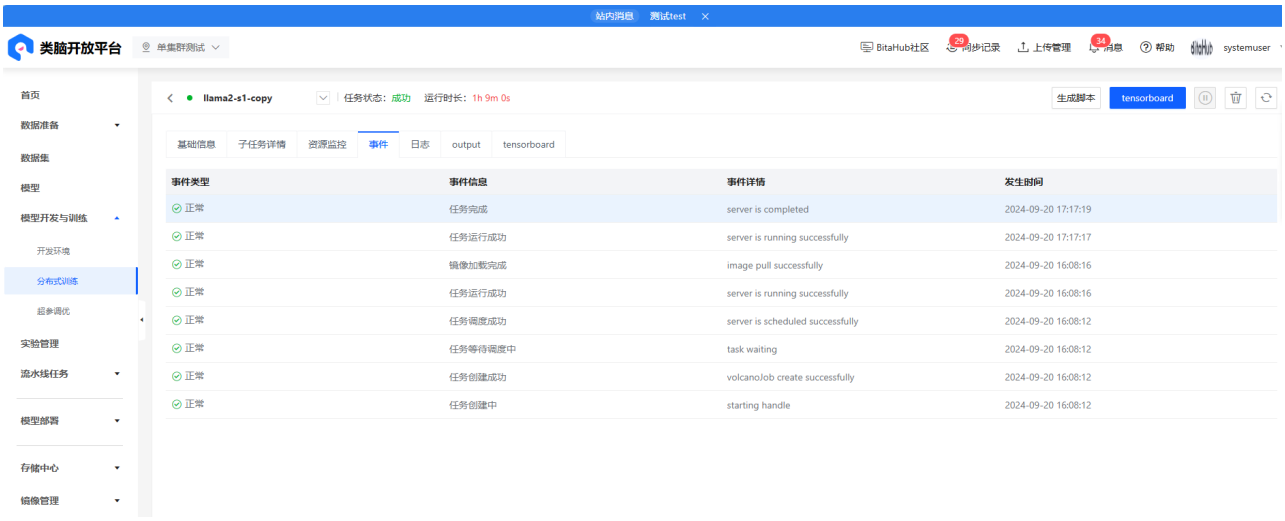
子任务详情： 展示pod级别的任务状态和数量，以及对应的资源配置



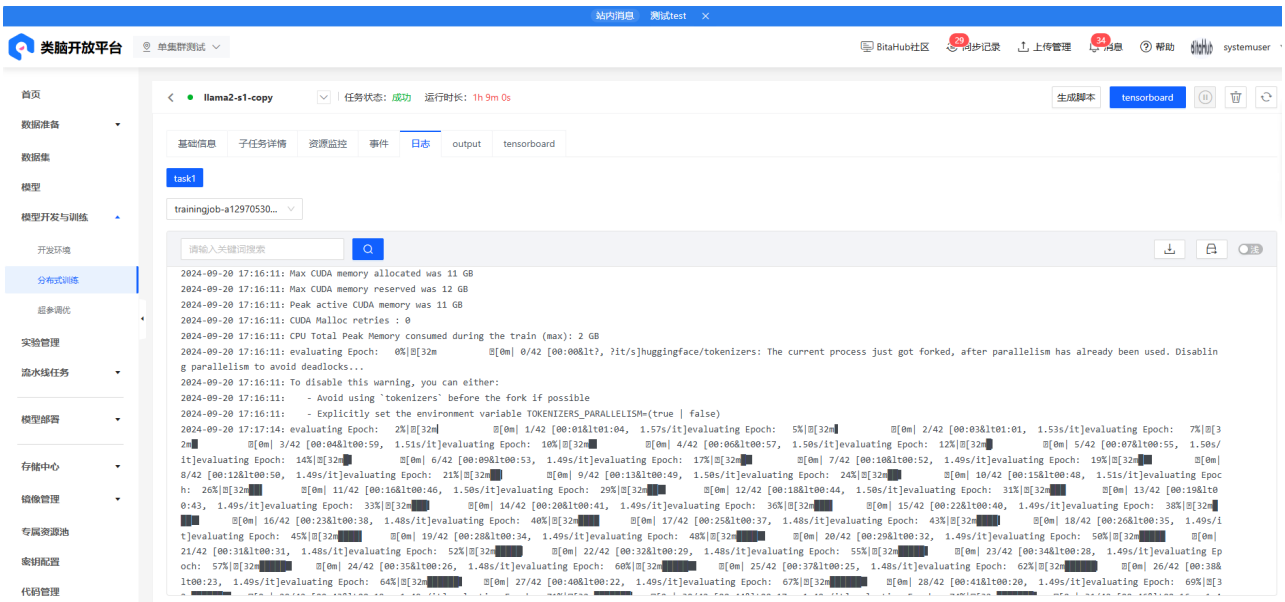

资源监控： 展示服务消耗的资源情况



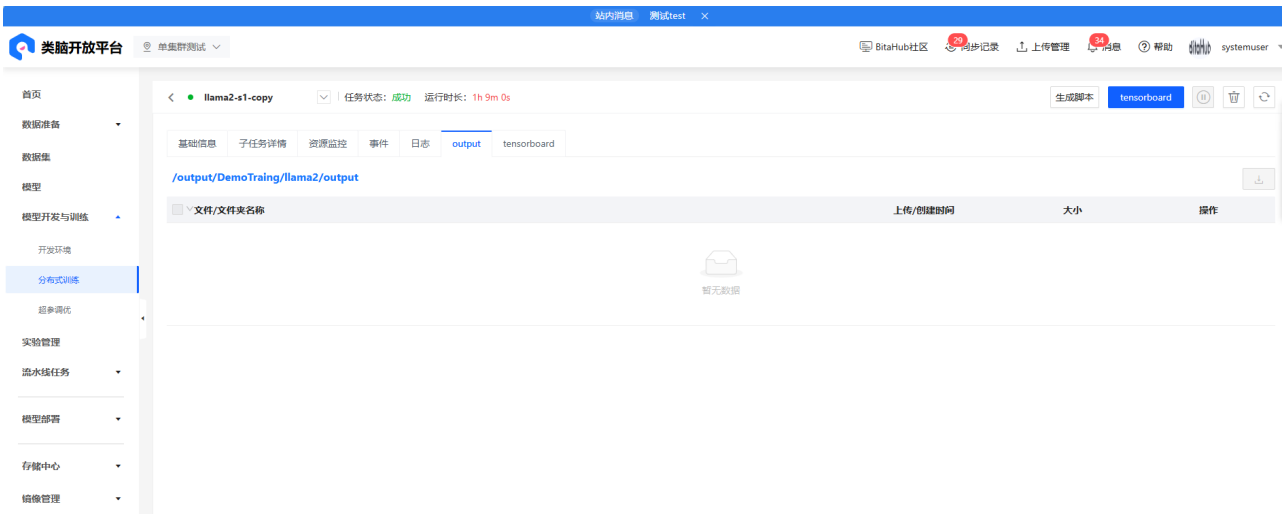

事件管理：展示任务训练过程中的关键性流程信息



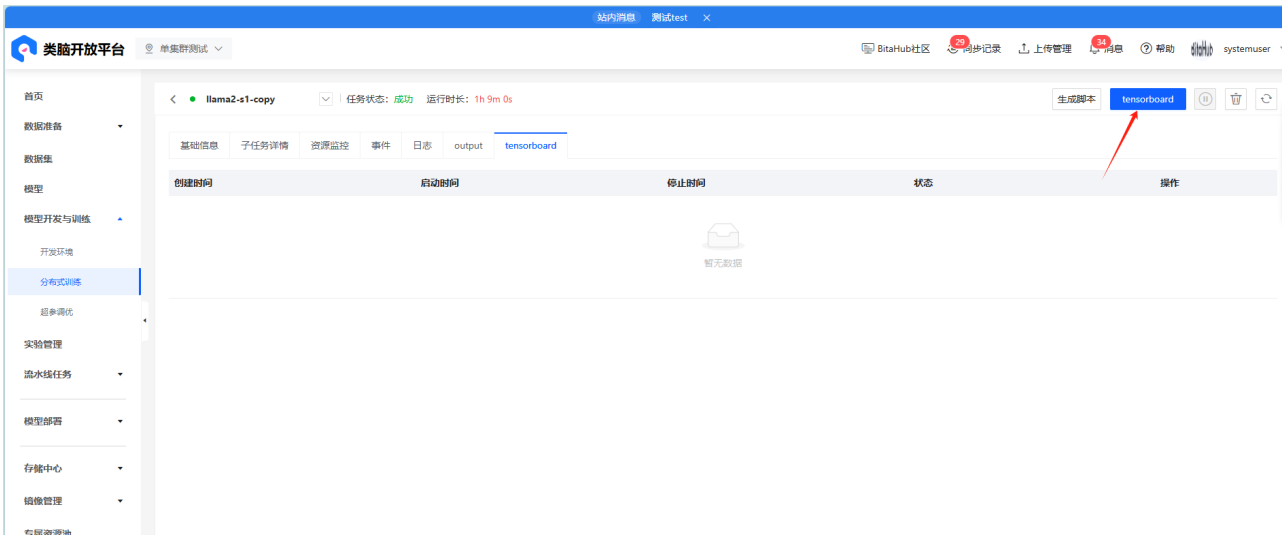

日志管理： 展示训练任务的运行日志



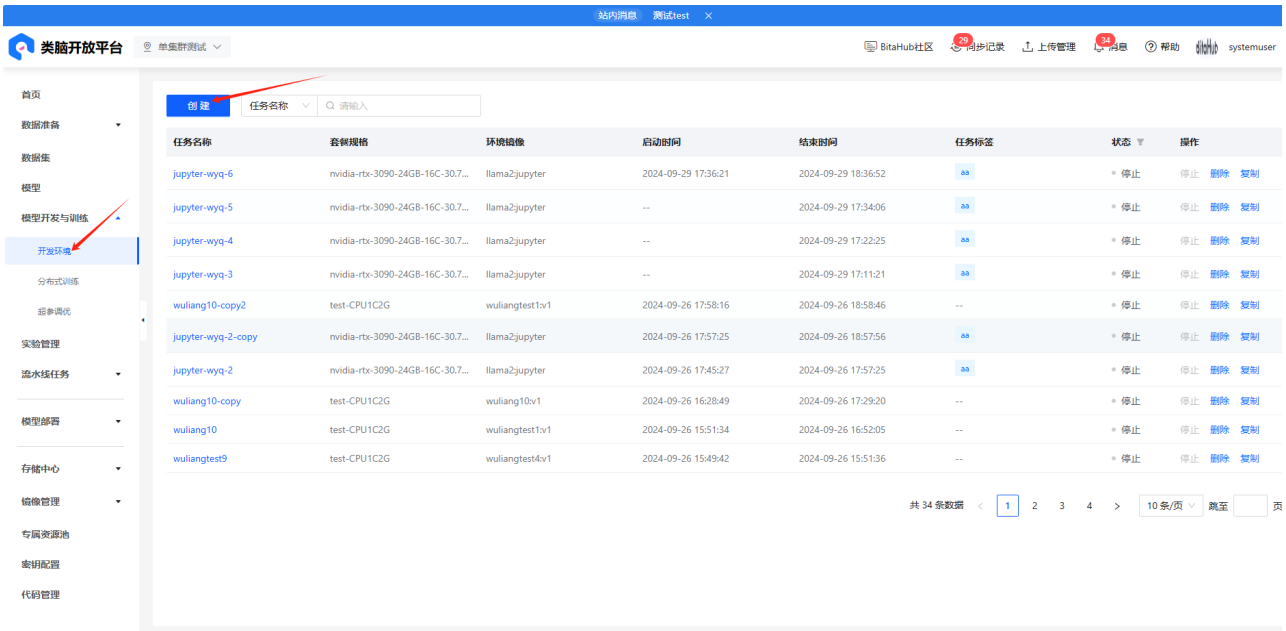

输出设置： 展示模型训练过程中的中间产出




tensorboard： 根据输出的文件夹开启tensorboard服务

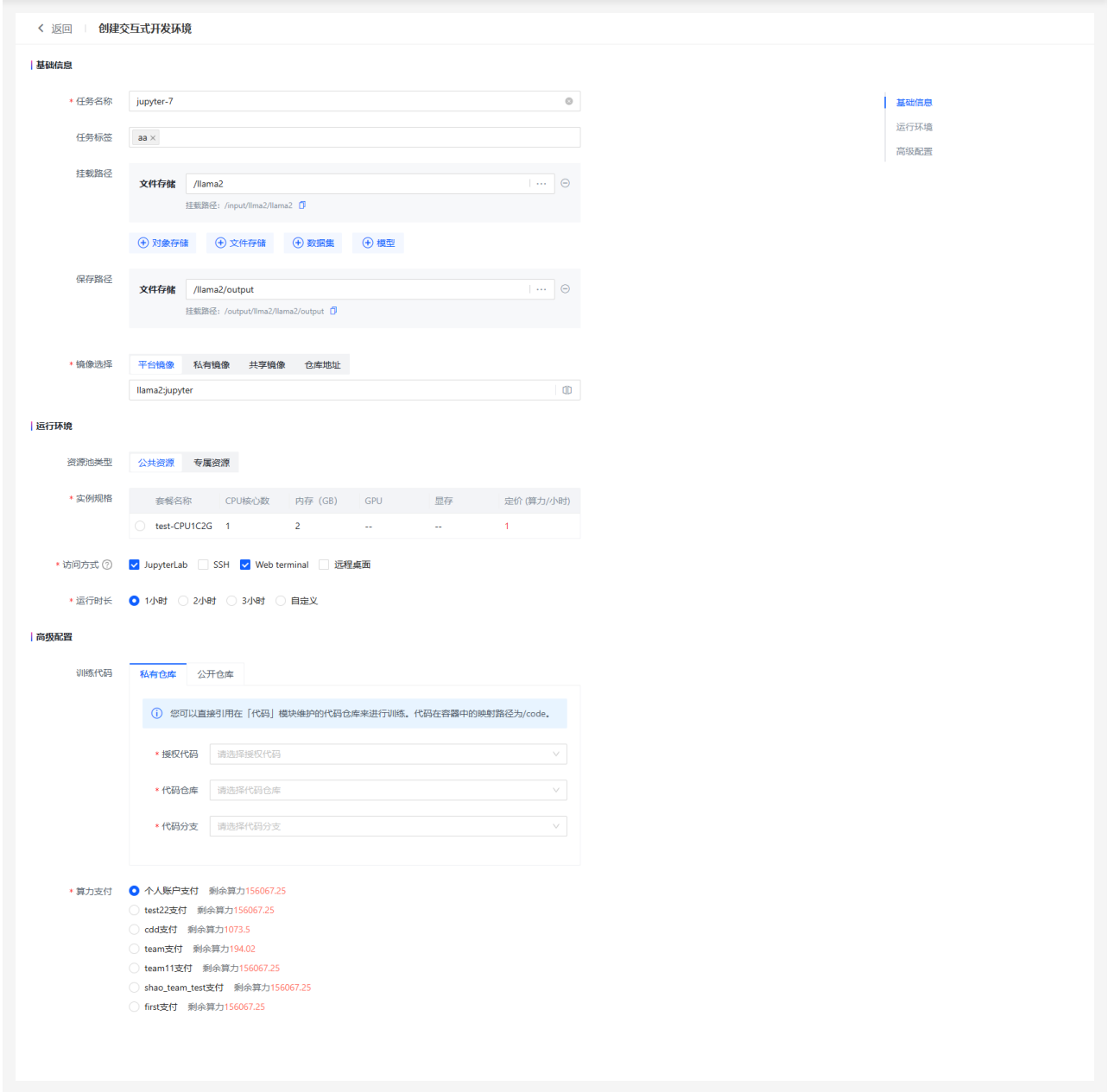



# 创建jupyter任务




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



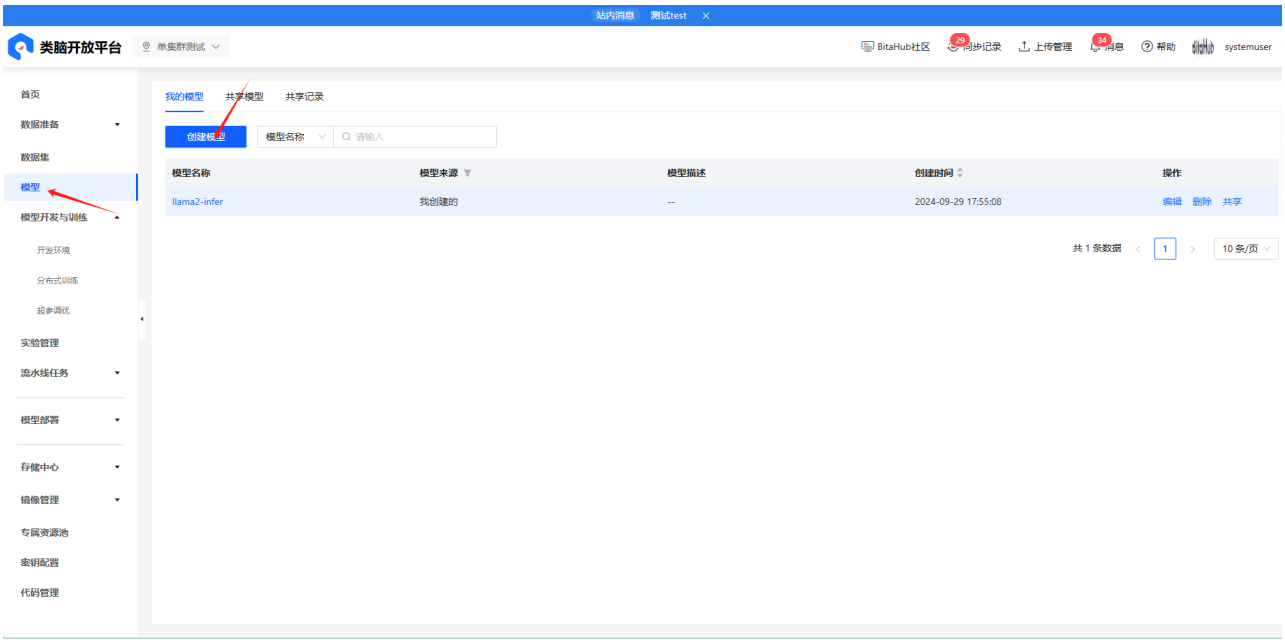

任务详情模块

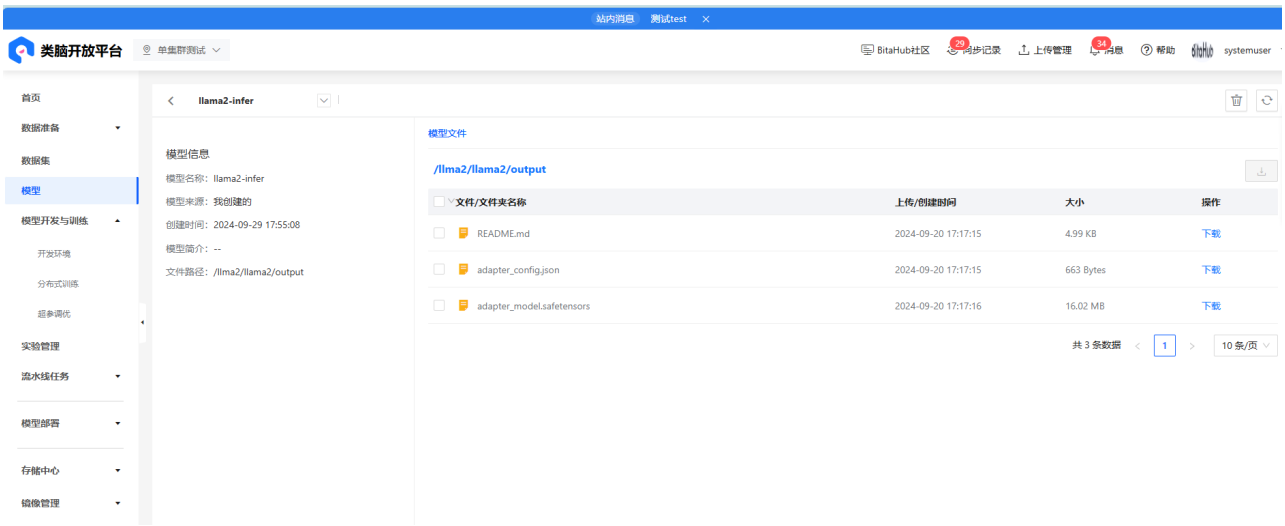
任务基础信息

资源监控

事件

# 创建模型

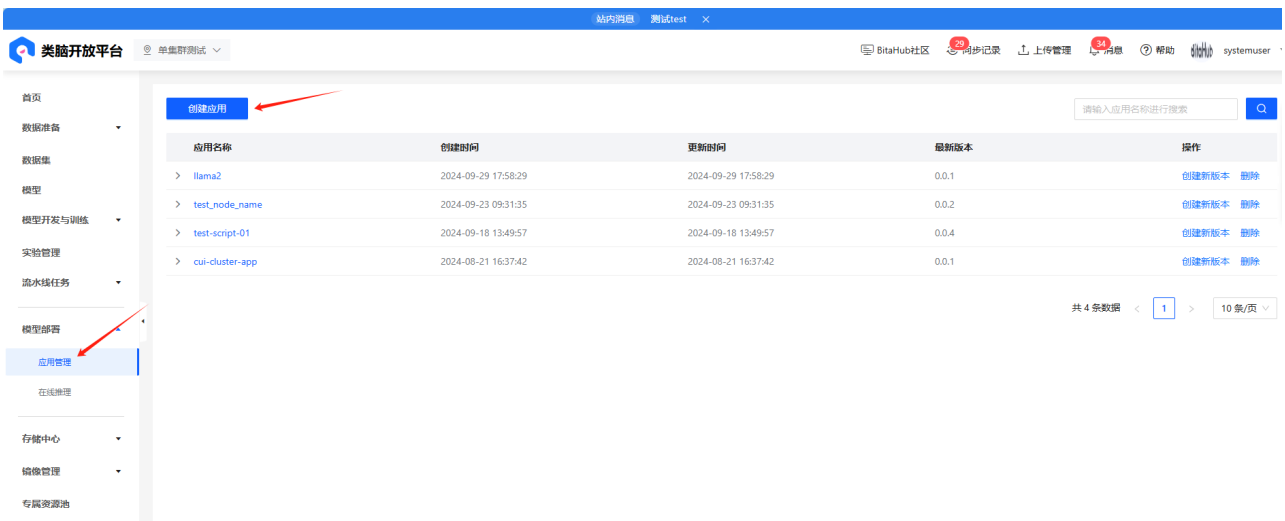



# 推理

1 创建应用

2 创建推理服务

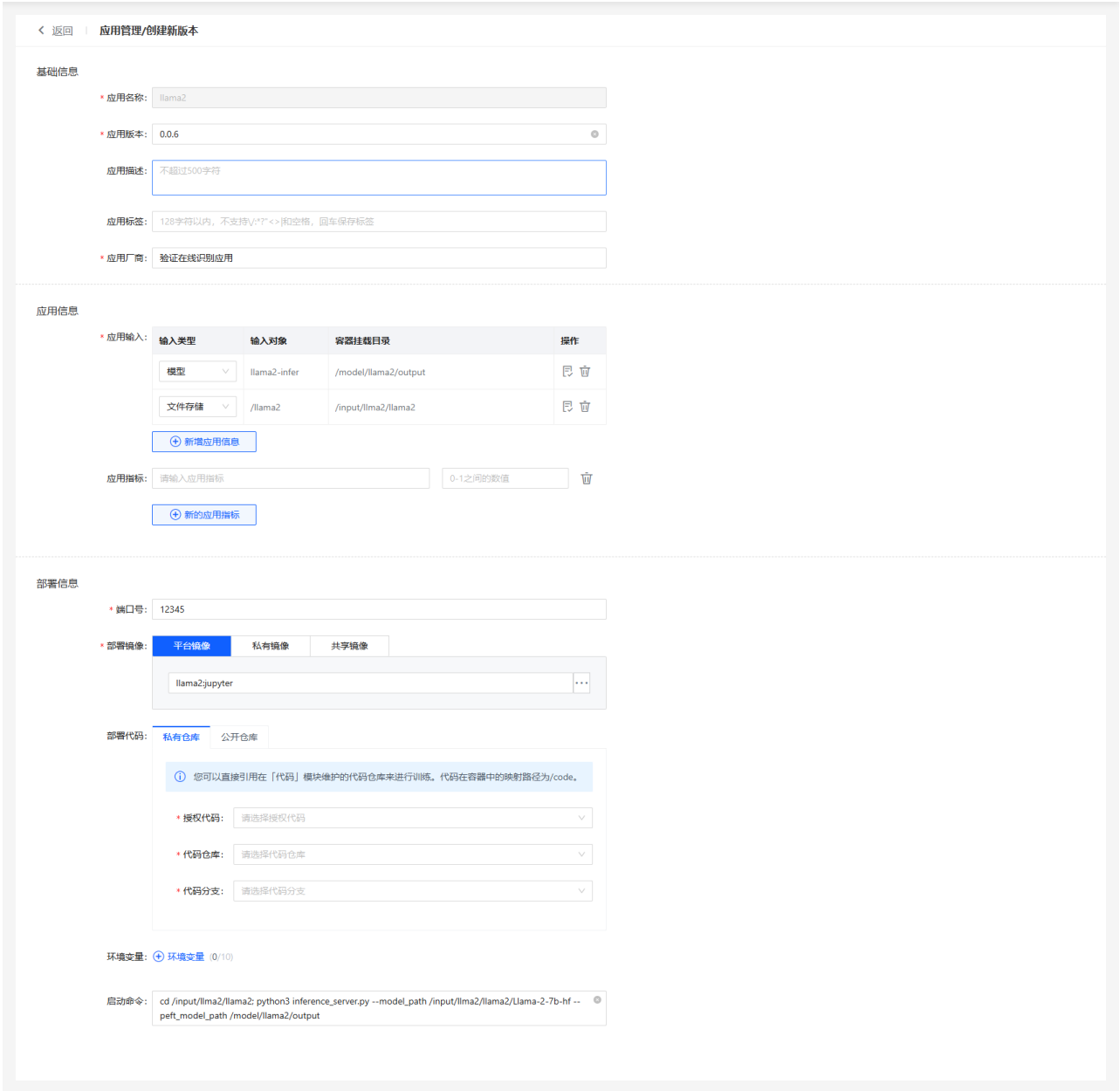



创建应用

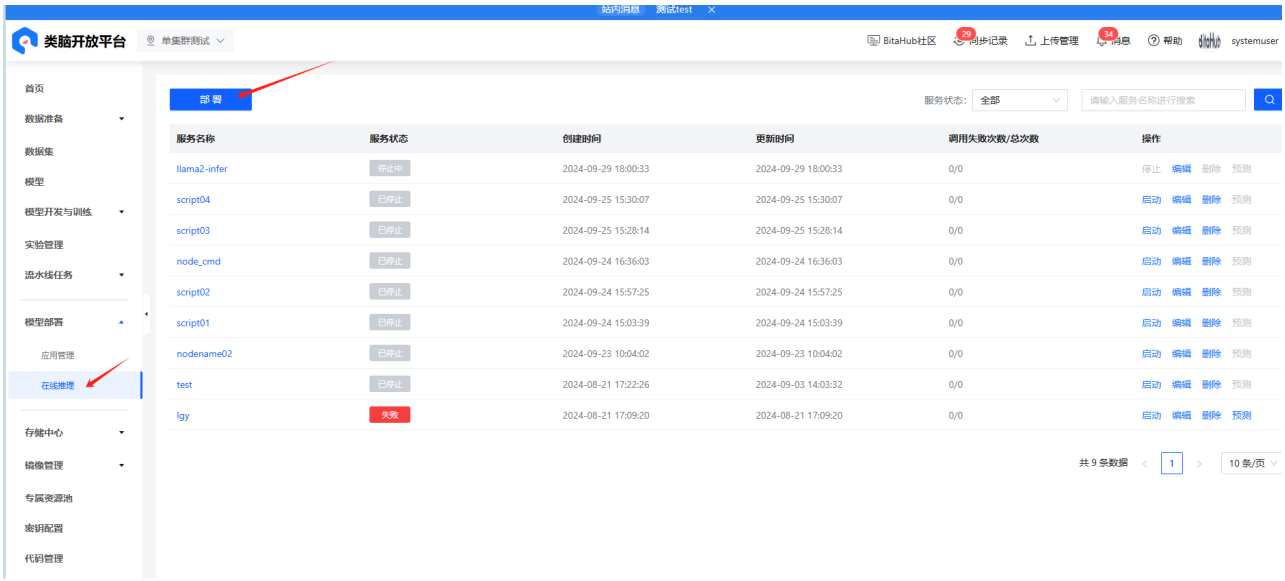
1 . 填写应用名称，版本，应用厂商

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




创建推理服务

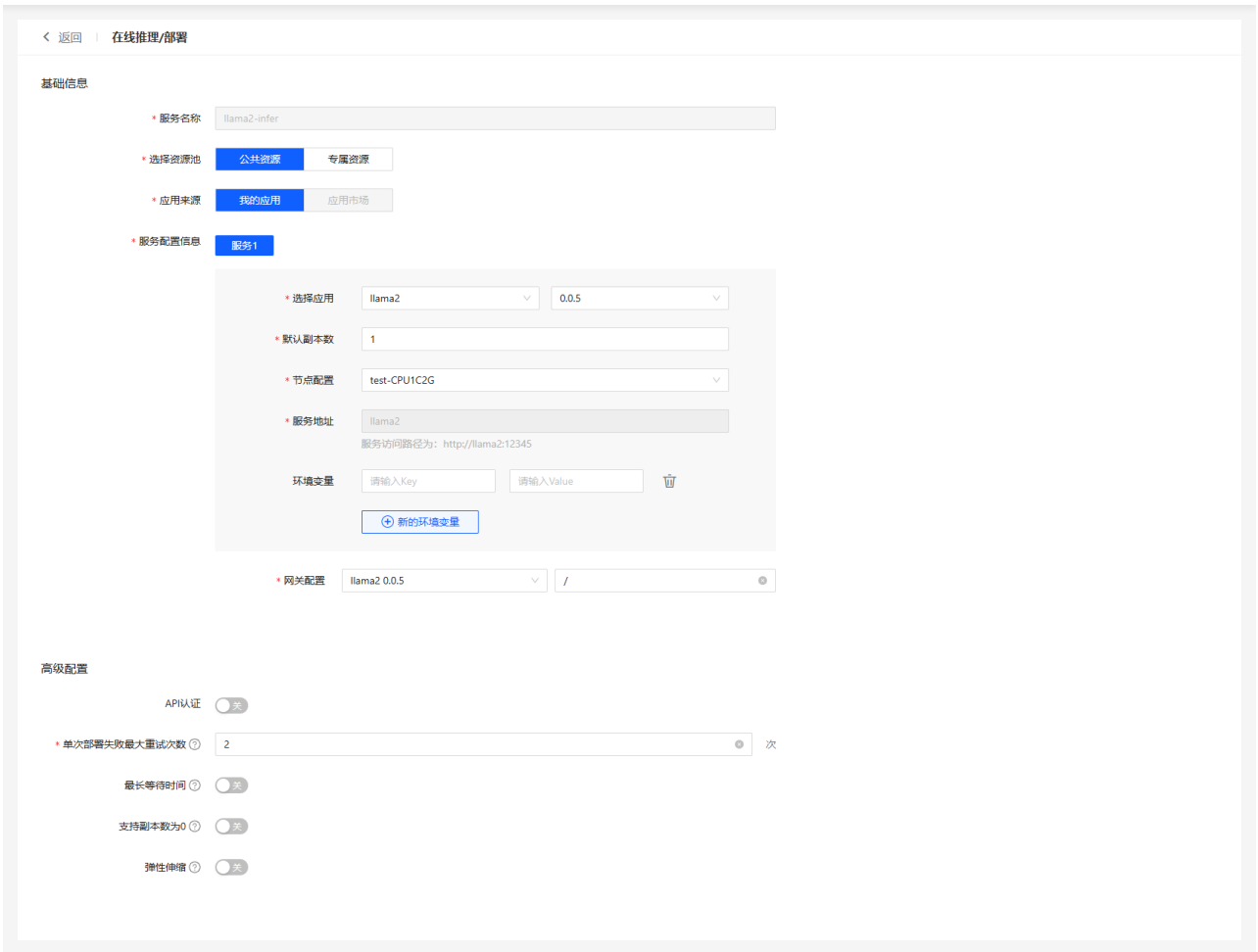


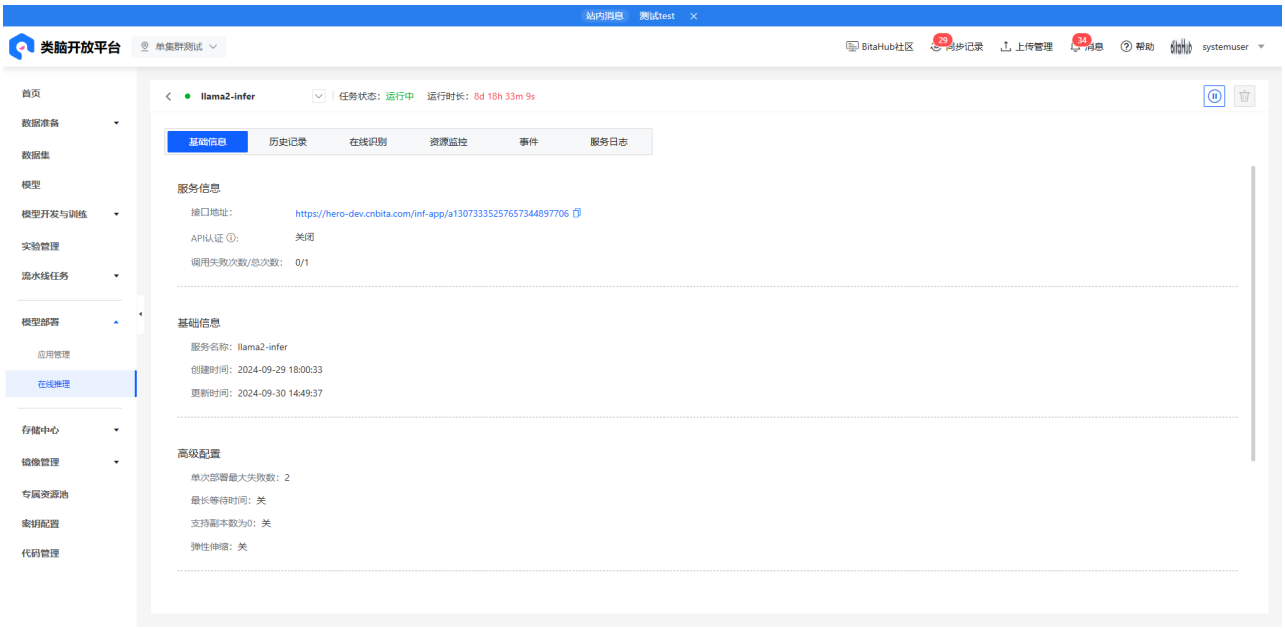

流程：

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

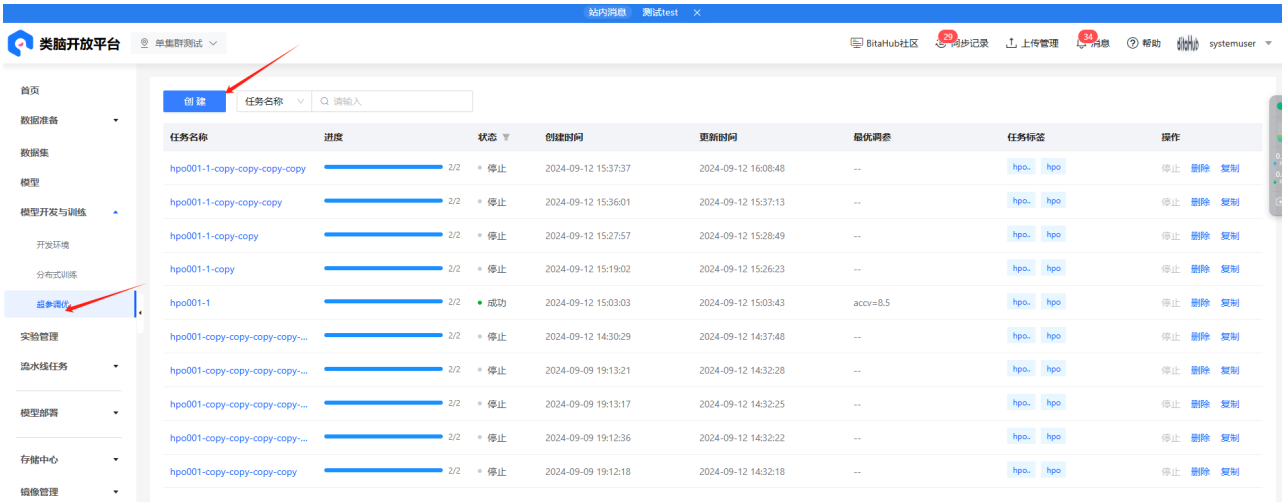





使用服务得接口地址请求推理服务

```
wyq@DESKTOP-PVL03Q3 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"}
```

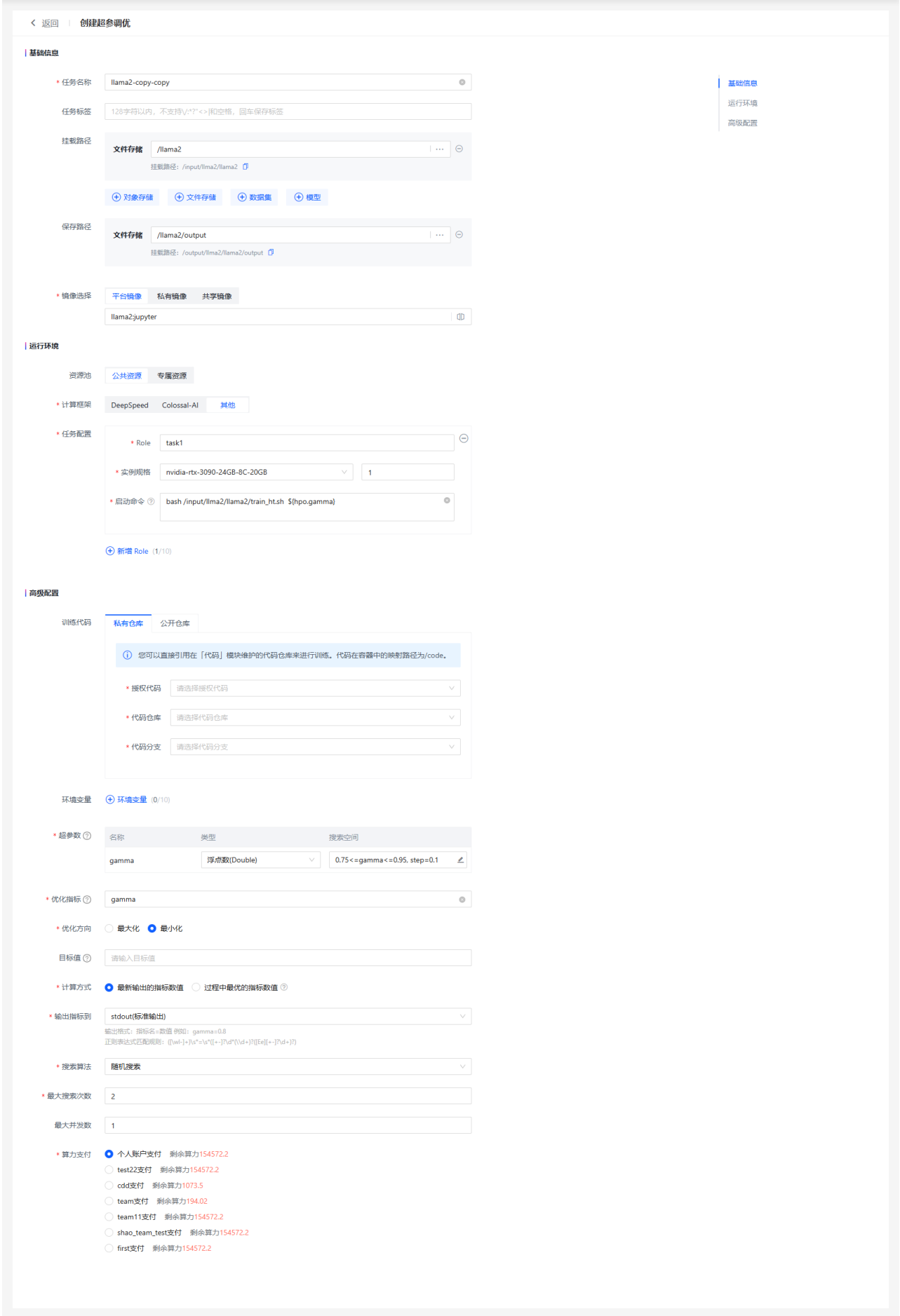
# 超参调优



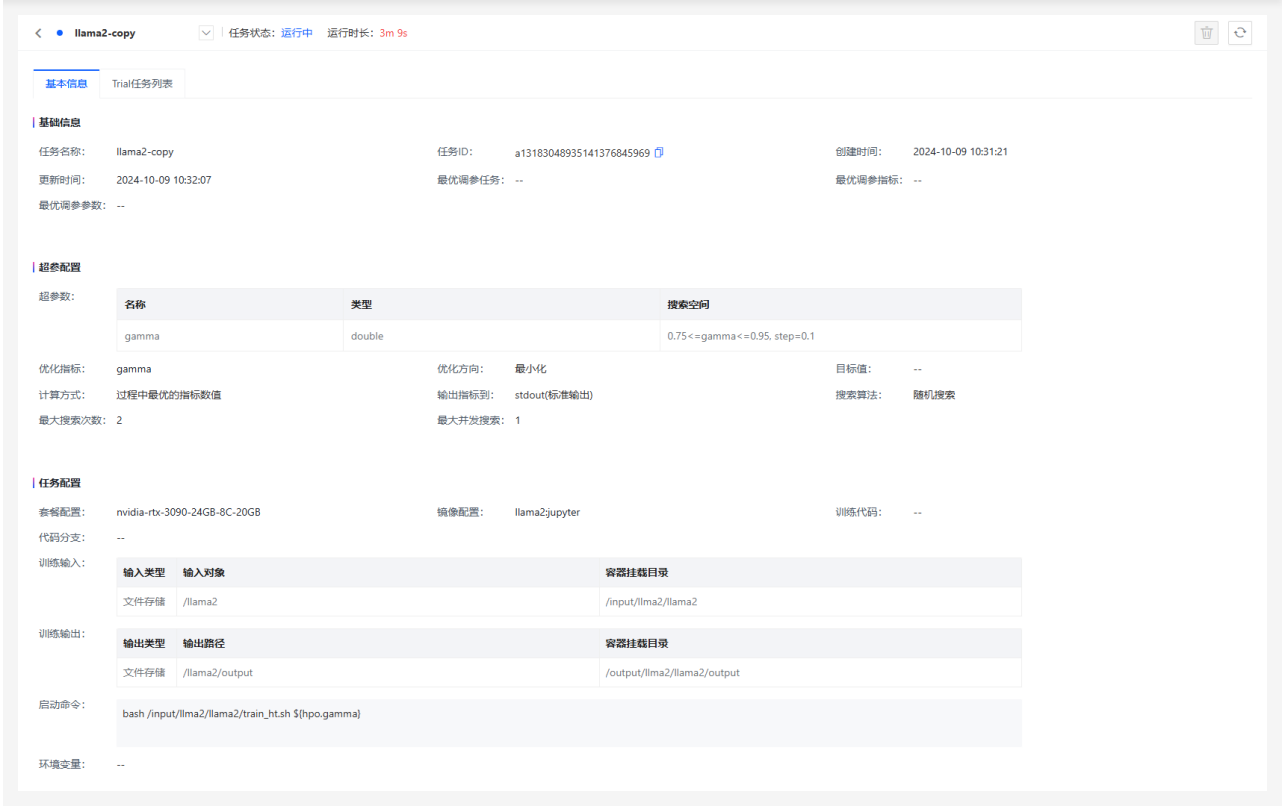

创建超参调优任务

基于训练得gamma来进行超参调优，

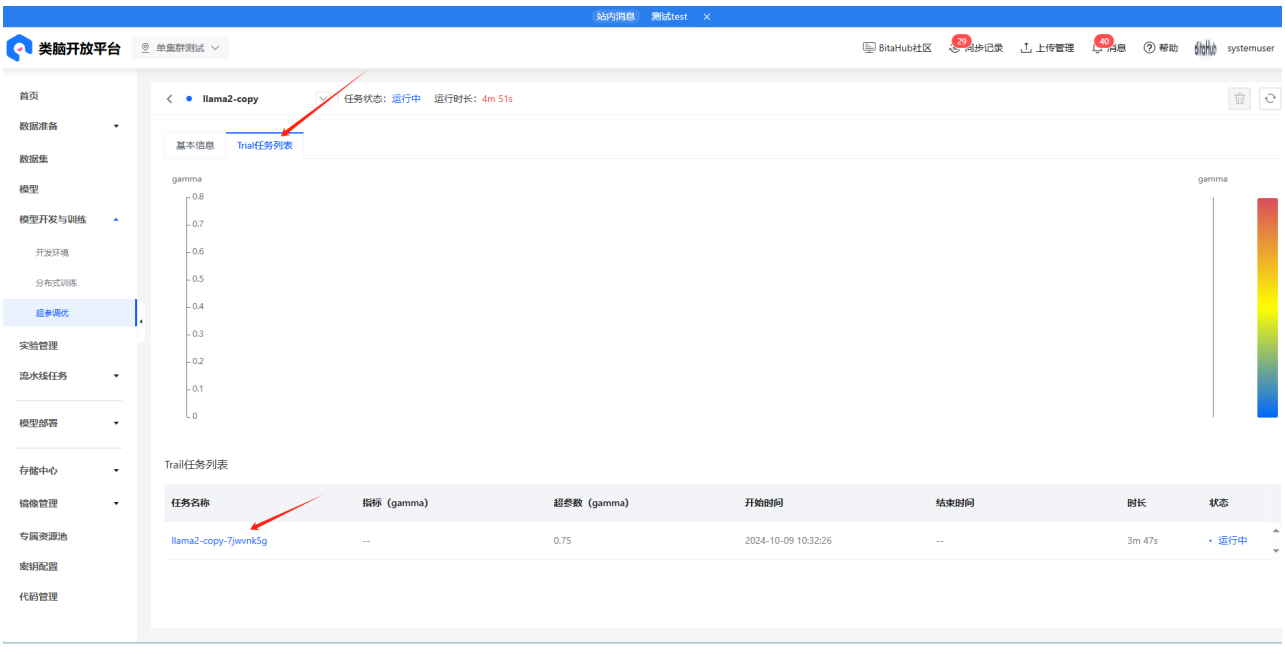
需要在任务命令中添加${hpo.gamma}, 如果想要微调其它参数，可按照${hpo.超参数名称}来进行添加，具体参数设置如下：


返回
创建超参调优
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/llma2/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/llma2/llama2/output
* 镜像选择
平台镜像
私有镜像
共享镜像
llama2:jupyter
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/llma2/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
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|-]+)\s*=\s*([+-]?\d*[\.\d+)?([Ee][+-]?\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
基础信息
运行环境
高级配置

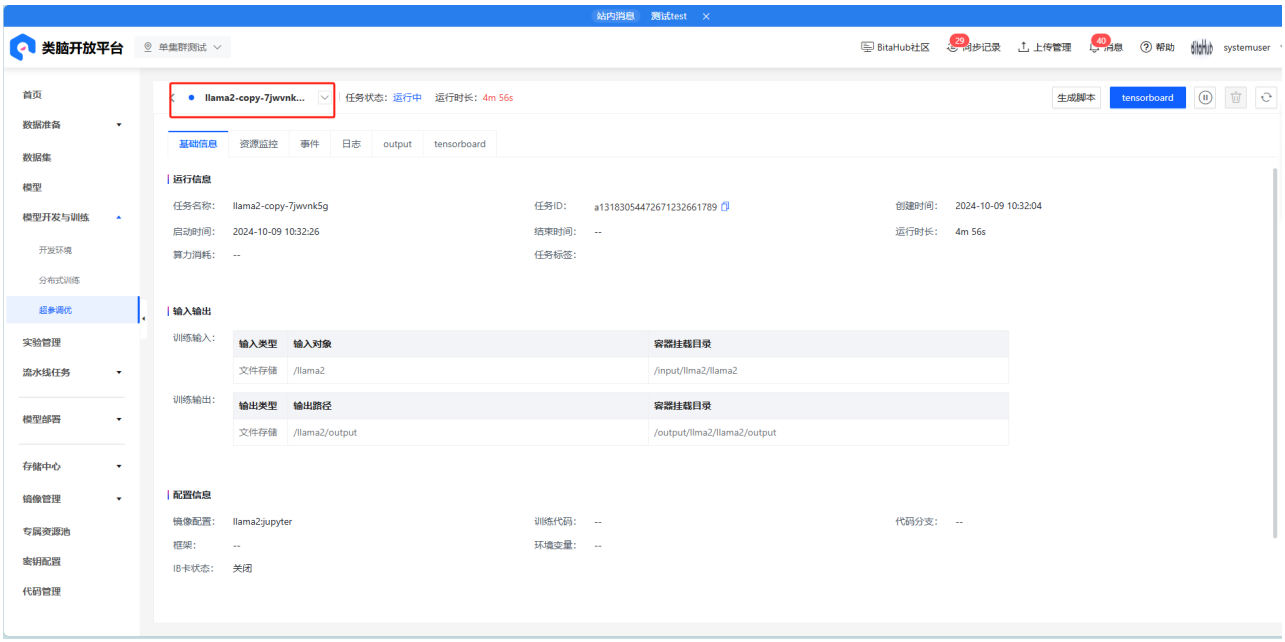
超参调优任务详情




点击trail任务列表，查看调优任务详情




点击具体得子任务名称，可以查看子任务详情

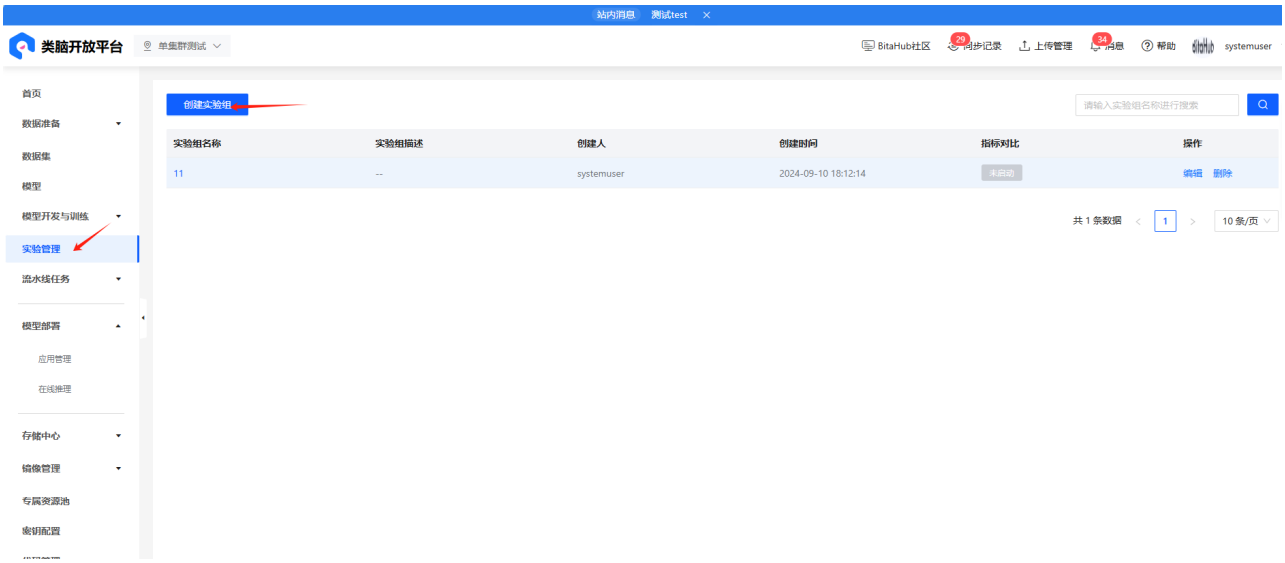


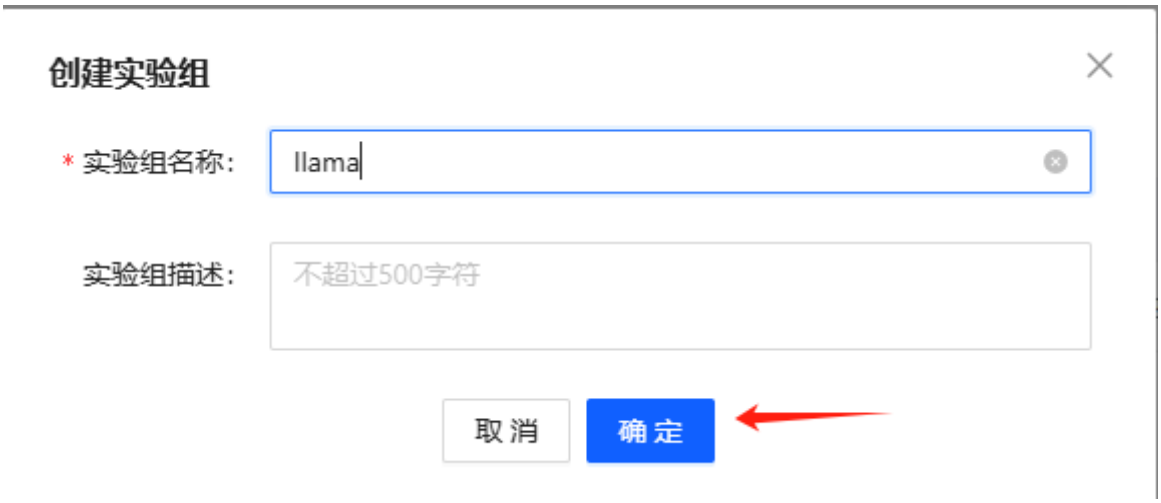

# 实验管理

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

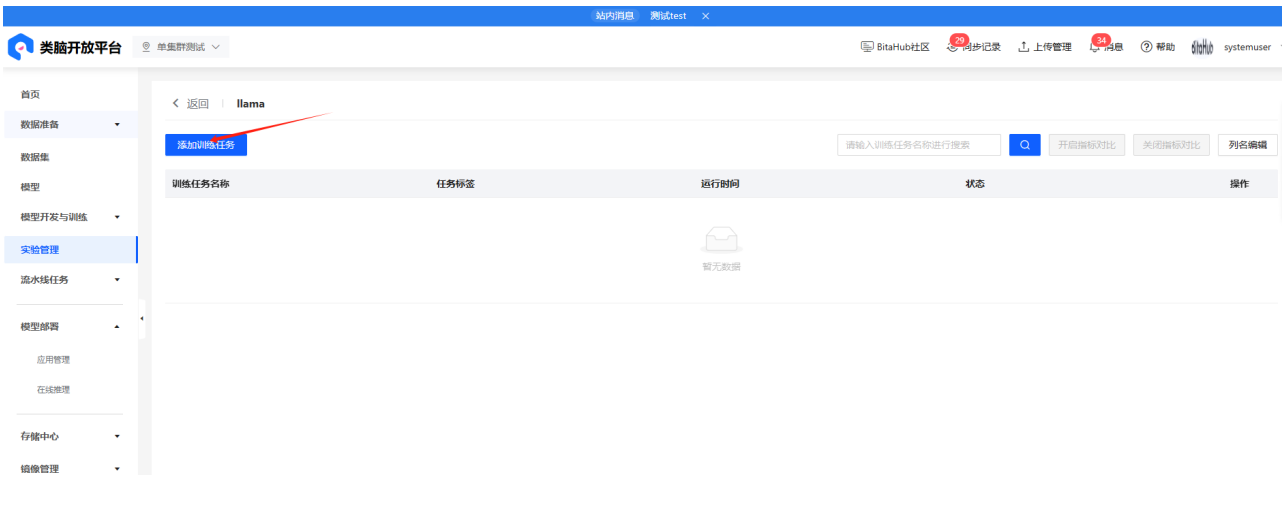
创建实验组

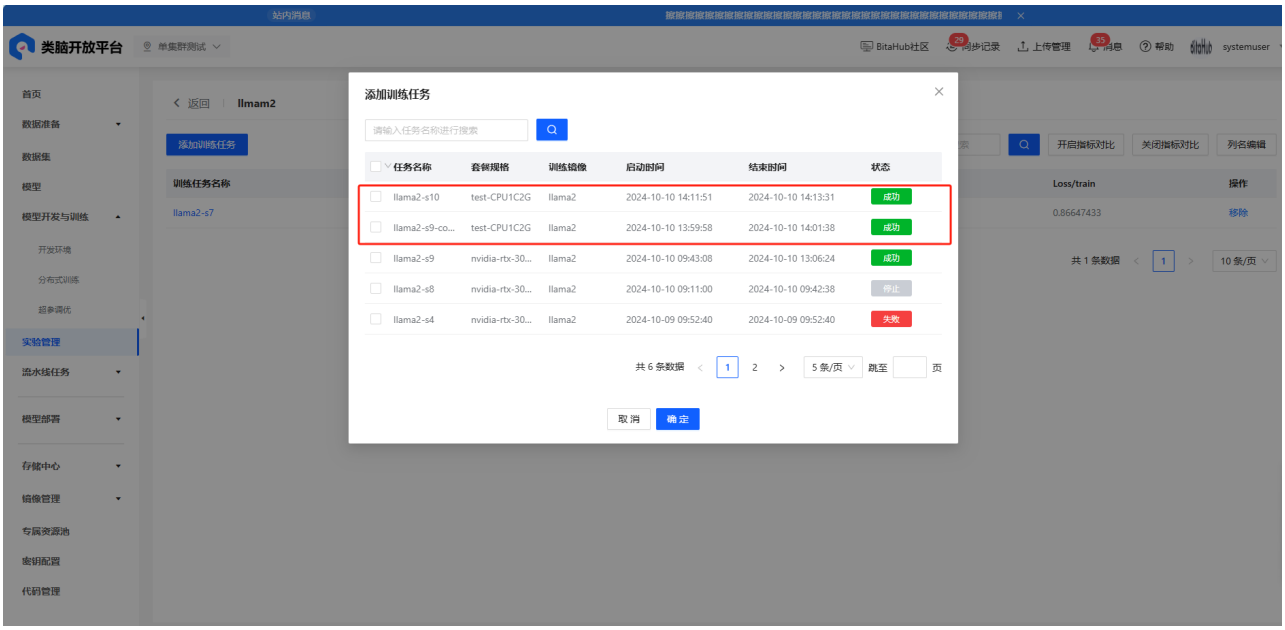




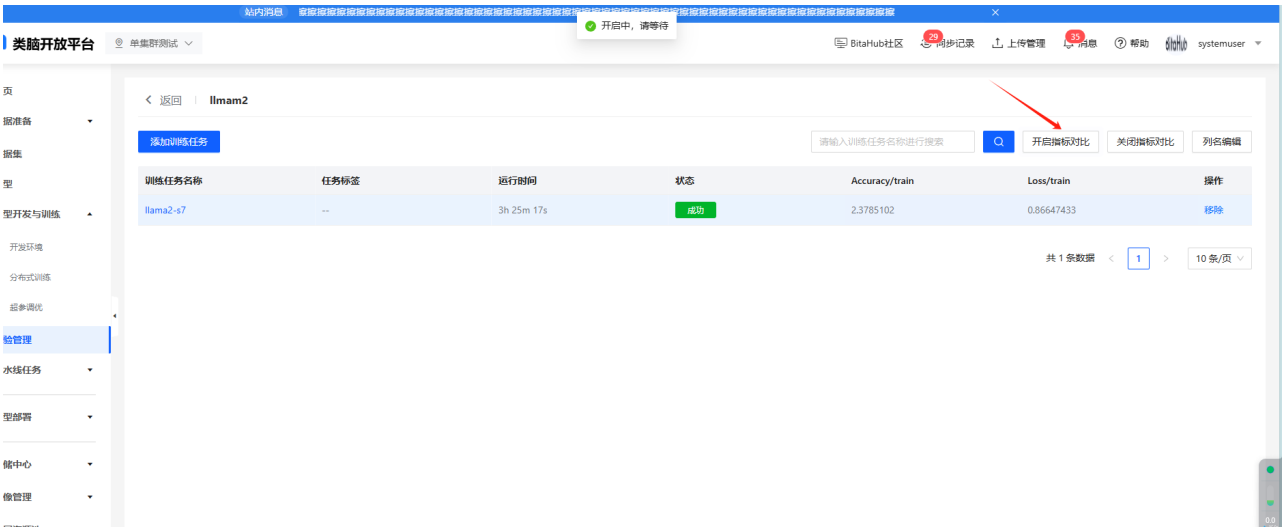

点击llama实验组

点击添加任务

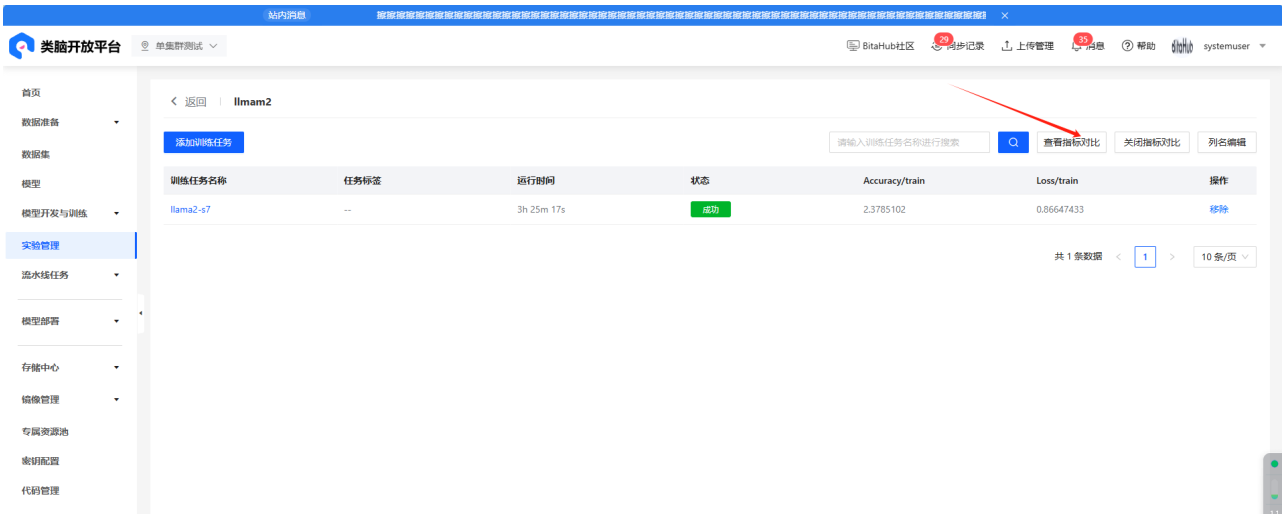



创建完开启指标




等待片刻，点击查看指标对比




结果如下