

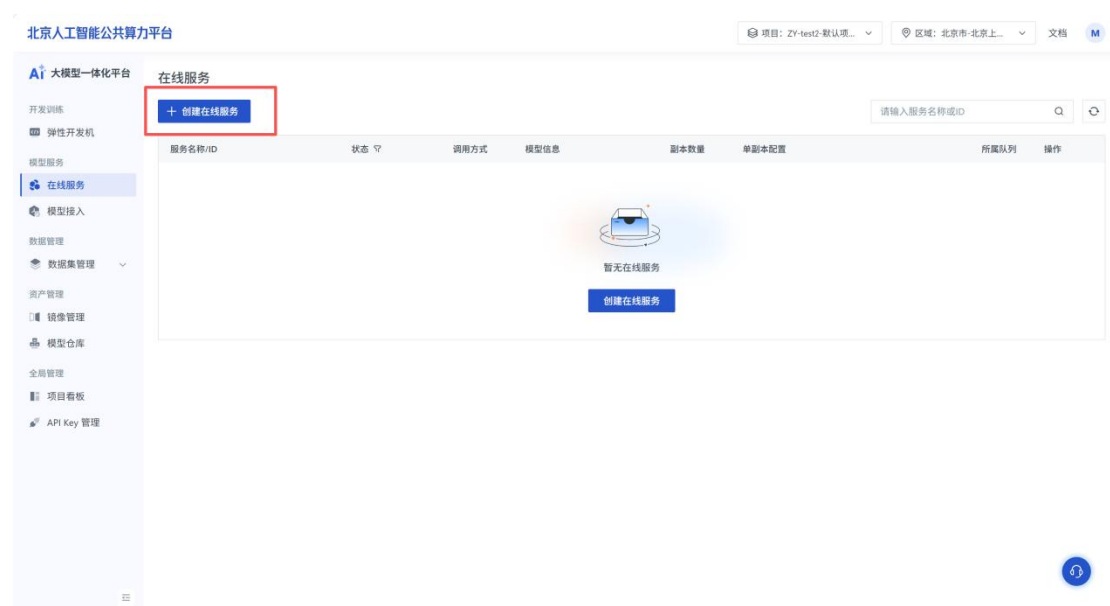
功能概述

模型在线服务为预置 AI 模型提供快速部署与全生命周期运维能力，可快速完成推理服务上线，满足业务线上调用需求。针对预置模型提供快速部署能力，用户选定模型并填写服务名称，选择队列与资源规格来确定服务并发容量，同时配置副本数量、API 密钥开关、并发限流、服务名称与标签等访问限制及基础信息，即可快速完成推理服务创建。平台可展示全部服务列表，实时展示服务调度中、部署中、运行中、异常、停止中、已删除等全生命周期状态，支持对服务执行启动停止、删除服务等运维操作，还可维护服务基础信息，灵活管控服务运行状态。同时平台支持查看实时请求并发数、首字延迟、请求延迟、实时 Token 等运行指标，提供 API 密钥维度的调用监控，附带在线调试与调用示例，并且支持查看单服务日志、单实例日志与操作日志，帮助使用者快速定位线上问题。

功能说明

创建在线服务

- 在在线服务列表页点击“创建在线服务”。



- 点击右侧  按钮选择模型

* 模型选择 

- 在基本信息模块填写模型服务名称，名称只能包含大小写英文、汉字、数字以及(:)特殊字符，长度 1~32 位

基本信息

* 模型服务名称

- 选择资源配置

资源配置

* 资源类型

* 单副本容量
可支持约20并发，每个副本需 8 * ADV_H800-80GB / 32cpu / 256GBmemory

* 副本数量

* 资源队列选择 

- 设置访问控制

访问控制

* API Key鉴权 ☒ ON 

开启APIKey鉴权后，需在发送推理请求时使用合法的APIKey才能接入服务

在线服务日常操作

停止在线服务

- 在线服务处于排队中、启动中、运行中、更新中状态时，支持创建人和租户管理员进行停止操作。在在线服务列表页，点击目标在线服务操作列的“停

止”。



启动在线服务

- 在线服务处于异常、已停止状态时，支持创建人和租户管理员进行启动操作。

在在线服务列表页，点击目标在线服务操作列的“启动”。

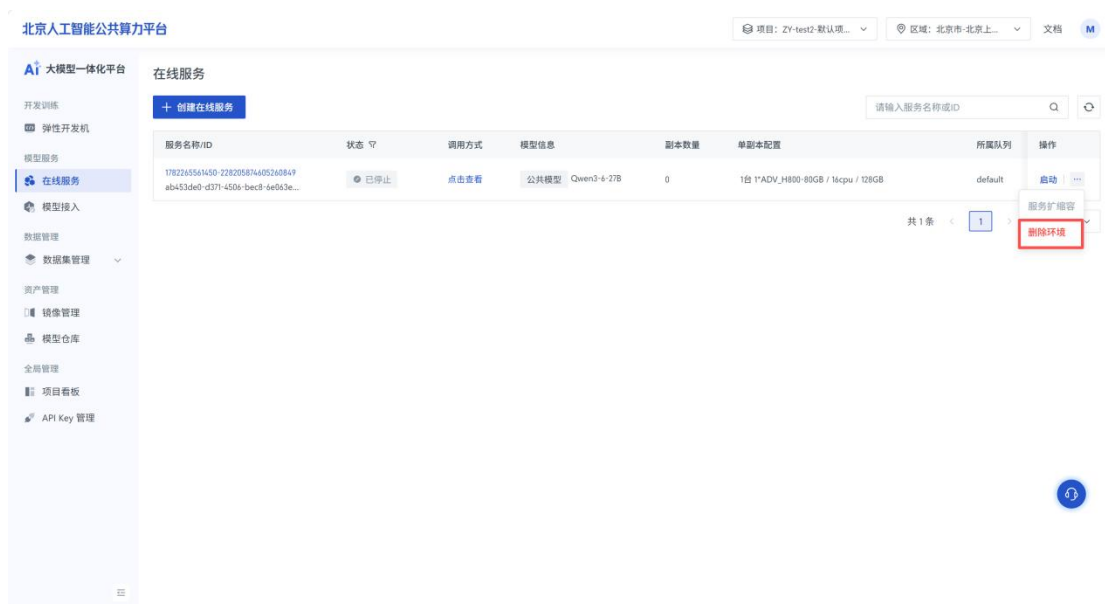


删除在线服务

- 在线服务处于异常、已停止状态时，支持创建人和租户管理员进行删除操作。

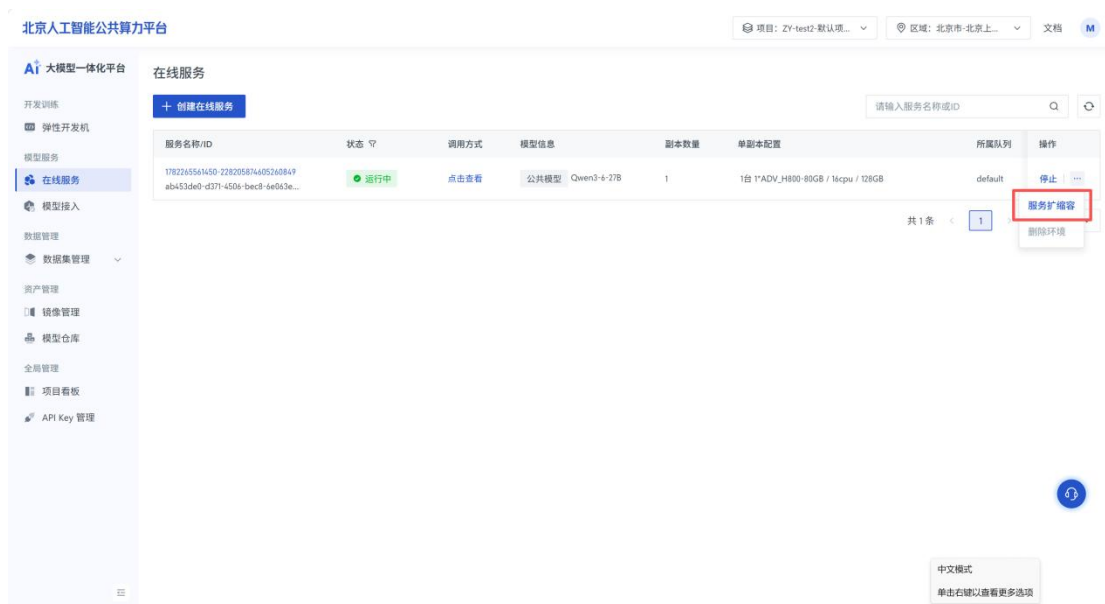
在在线服务列表页，点击目标在线服务操作列的更多操作，在下拉框中选择

“删除”



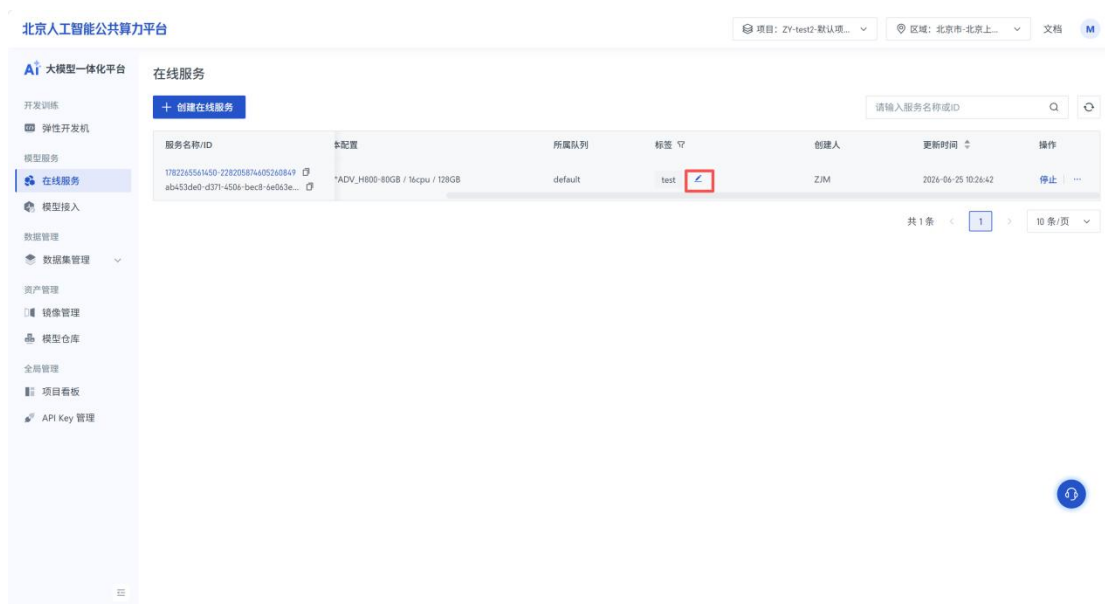
在线服务扩缩容

- 在线服务处于运行中状态时, 支持创建人和租户管理员进行扩缩容操作。在在线服务列表页, 点击目标在线服务操作列的更多操作, 在下拉框中选择“服务扩缩容”



修改在线服务标签

- 在线服务支持创建人和租户管理员修改标签。在在线服务列表页, 点击目标在线服务标签页的编辑按钮跳转编辑弹窗。



删除标签

点击现有标签右方的删除按钮删除标签。



添加标签

点击新增标签按钮添加新标签，最多支持 5 个标签。

编辑标签

×

test

🗑

+ 新增标签

还可以添加 4 个标签

取消

确认

在输入框内输入新增标签内容。

编辑标签

×

test

🗑

请输入标签

🗑

+ 新增标签

还可以添加 3 个标签

取消

确认

在线服务日常维护

查看在线服务基本信息

- 在线服务支持创建人和租户管理员查看在线服务基本信息。点击在线服务列表页的在线服务名称，进入在线服务详情页面，点击基本信息 tab 可查看在线服务的基本信息。

北京人工智能公共算力平台

项目: ZY-test2-默认项... 区域: 北京市-北京上... 文档 M

大模型一体化平台

开发训练

弹性开发机

模型服务

在线服务

模型接入

数据管理

数据集管理

资产管理

镜像管理

模型仓库

全局管理

项目看板

API Key 管理

在线服务

+ 创建在线服务

请输入服务名称或ID

服务名称/ID	状态	调用方式	模型信息	副本数量	单副本配置	所属队列	操作
17825a55a3a50-22820568344052a0849-ab453de0-d37f-450b-bec9-ae5e3e...	运行中	点击查看	公共模型 Qwen3-6.7B	1	1台 1*ADV_H800-80GB / 16cpu / 128GB	default	停止 ...

共 1 条 < 1 > 10 条/页

包括基础信息、权限控制、环境信息、资源配置和启动设置 5 个模块

基本信息

服务监控

服务日志

服务实例

服务调用

操作审计

基础信息

服务名称:

1782265561450-228205874605260849

在线服务ID:

ab453de0-d371-4506-bec8-6e063ed65f6f

服务副本:

● 正常: 1

● 异常: 0

标签:

test

更新人:

ZJM

更新时间:

2026-06-25 10:26:42

更多设置

API Key鉴权:

开启

环境信息

模型文件:

公共模型

Qwen3-6-27B / 27B

镜像文件:

公共镜像

10.6.212.44/library/sglang:v0.5.121

资源配置

资源队列:

default

副本数量:

1

单副本配置:

1台 1*ADV_H800-80GB / 16 cpu / 128 GB memory

启动设置

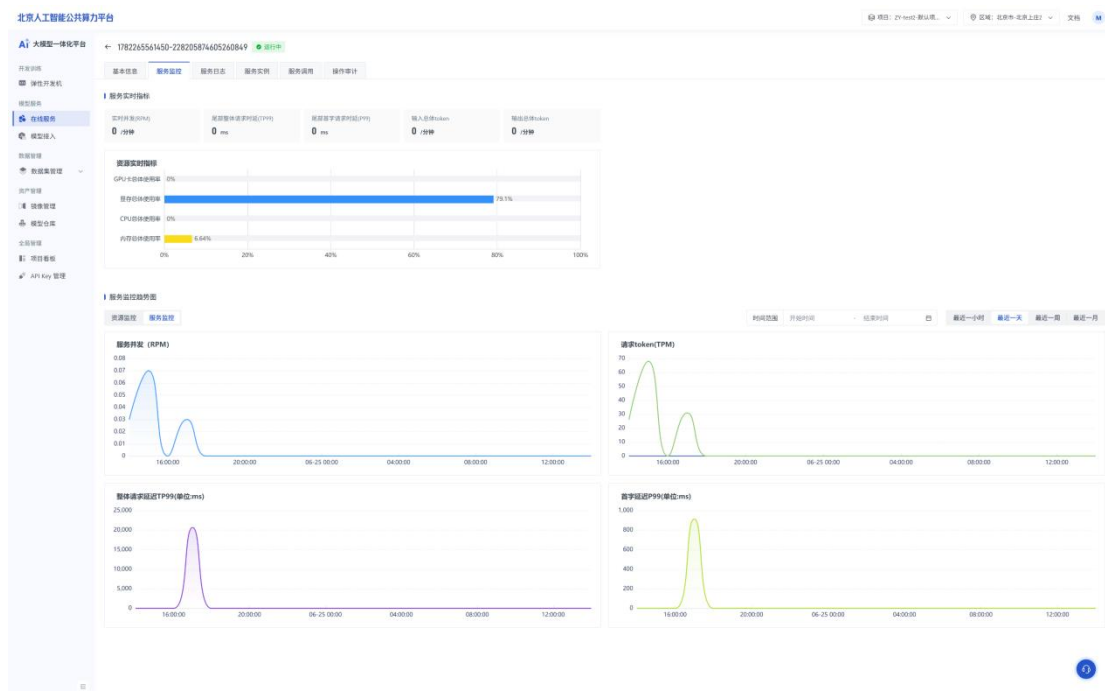
启动命令:

sglang serve --dist-init-addr \$LWS_LEADER_ADDRESS:20000 --nnodes \$LWS_GROUP_SIZE --node-rank \$LWS_WORKER_INDEX --model-path /models/Qwen3-6-27b --served-model-name qwen3-6-27b --context-length 128000 --tp 1 --ep-size 1 --dtype bfloat16 --host 0.0.0.0 --port 8000

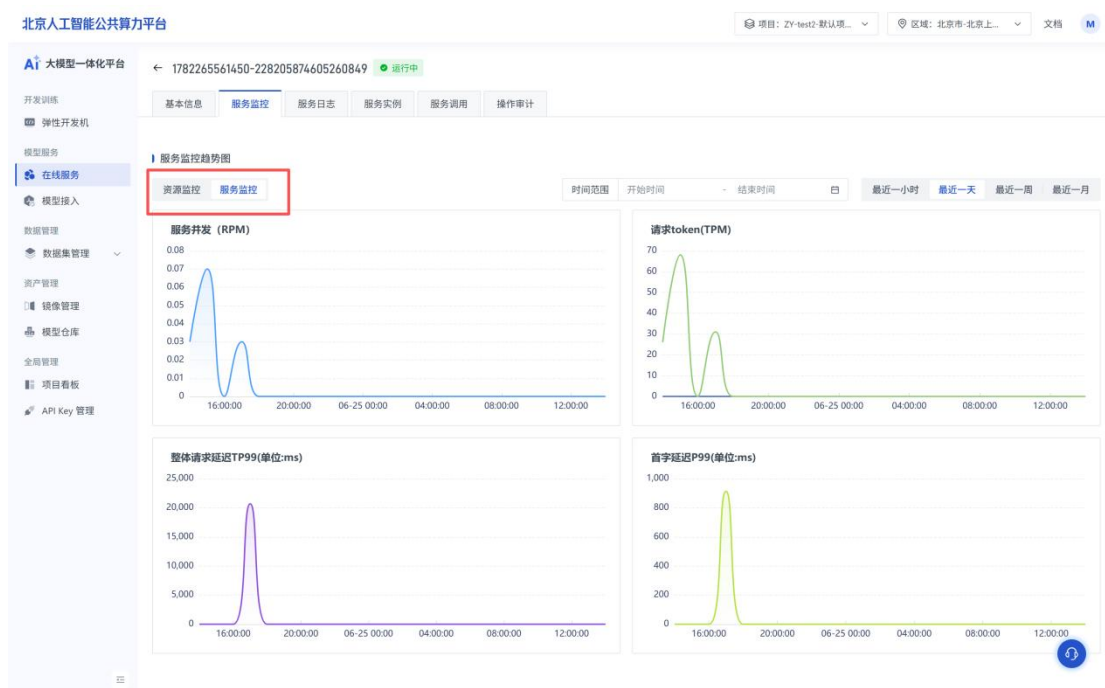
查看在线服务的服务监控

- 在线服务支持创建人和租户管理员查看在线服务监控。点击在线服务列表页的在线服务名称，进入在线服务详情页面，点击基本信息 tab 可查看在线服务的基本信息。包括服务实施指标和服务监控趋势图两个模块。

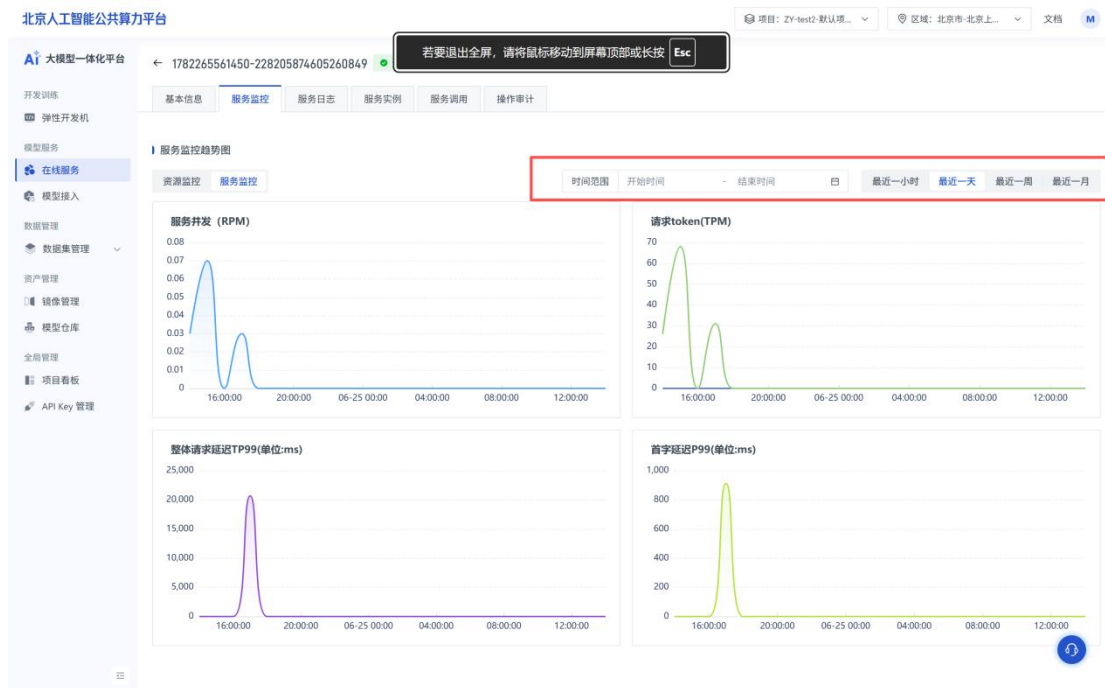
注:若要获取整体请求延迟 TP99 和首字延迟 P99 数据,则需要将 stream 参数设为 true



其中服务监控趋势图包含资源监控和服务监控，可通过点击 tab 进行切换

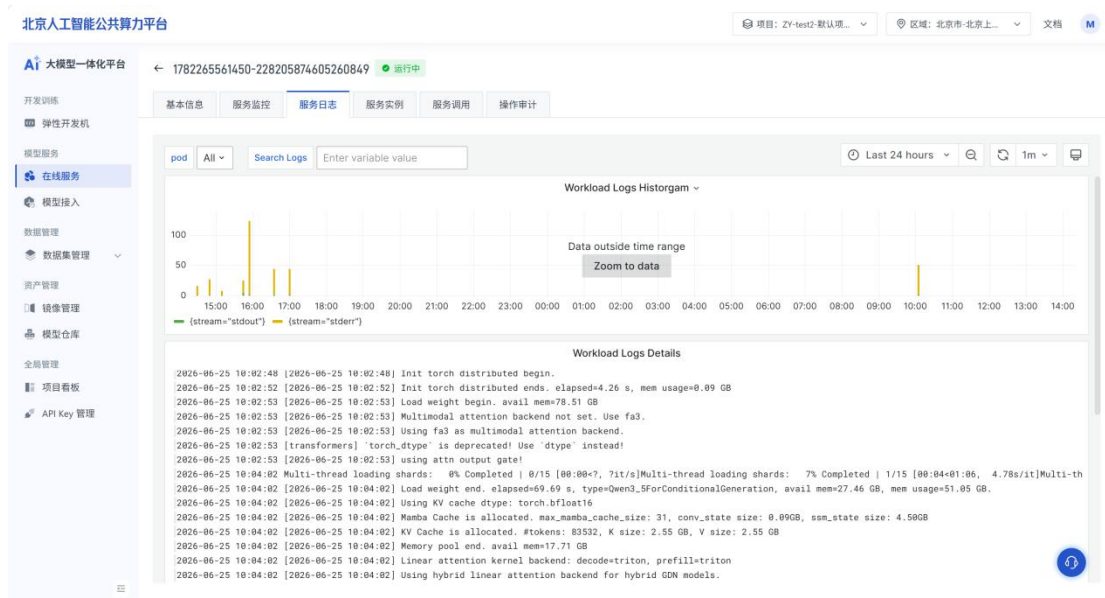


右上方可以设置趋势图的时间区间



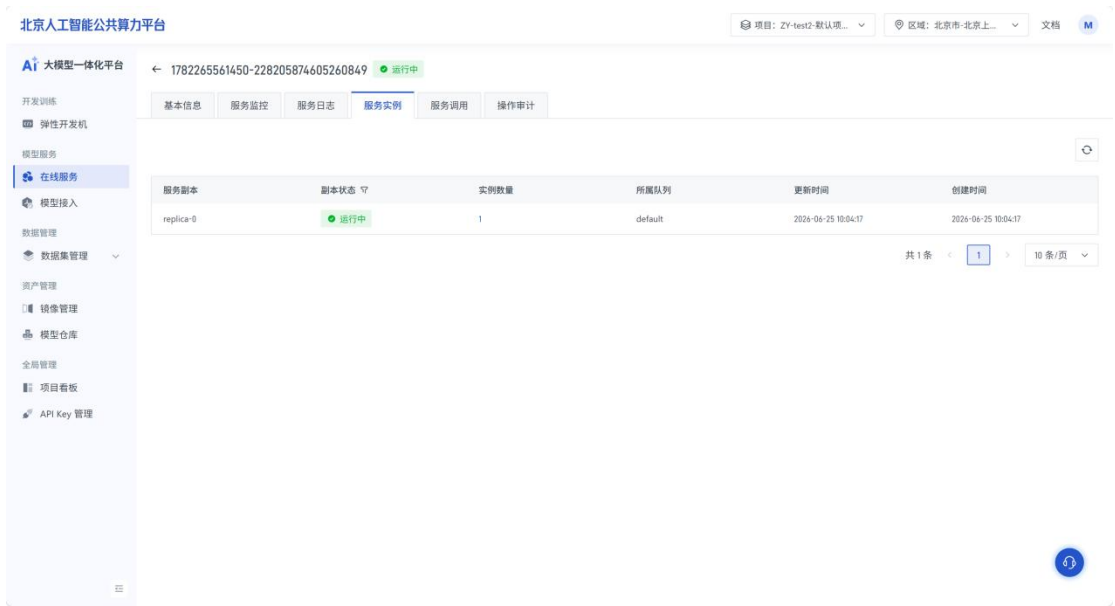
查看在线服务的日志

- 在线服务支持创建人和租户管理员查看在线服务日志。点击在线服务列表页的在线服务名称，进入在线服务详情页面，点击服务日志 tab 可查看在线服务的日志。



查看在线服务的服务实例

- 在线服务支持创建人和租户管理员查看在线服务实例。点击在线服务列表页的在线服务名称，进入在线服务详情页面，点击服务实例 tab 可查看在线服务的服务实例



查看在线服务的服务调用

- 在线服务支持创建人和租户管理员查看在线服务的服务调用。点击在线服务列表页的在线服务名称，进入在线服务详情页面，点击服务调用 tab 可查看在线服务的服务调用



查看在线服务的操作审计

- 在线服务支持创建人和租户管理员查看在线服务的服务审计。点击在线服务列表页的在线服务名称，进入在线服务详情页面，点击操作审计 tab 可查看在线服务的操作审计

