

数据集定义

数据集是按既定结构或规则组成的数据集合，覆盖文本、图像、音频、视频、多模态等多种数据格式。可用于基座模型预训练、垂直场景微调等环节，是模型能力构建与迭代优化的重要素材。

数据集结构

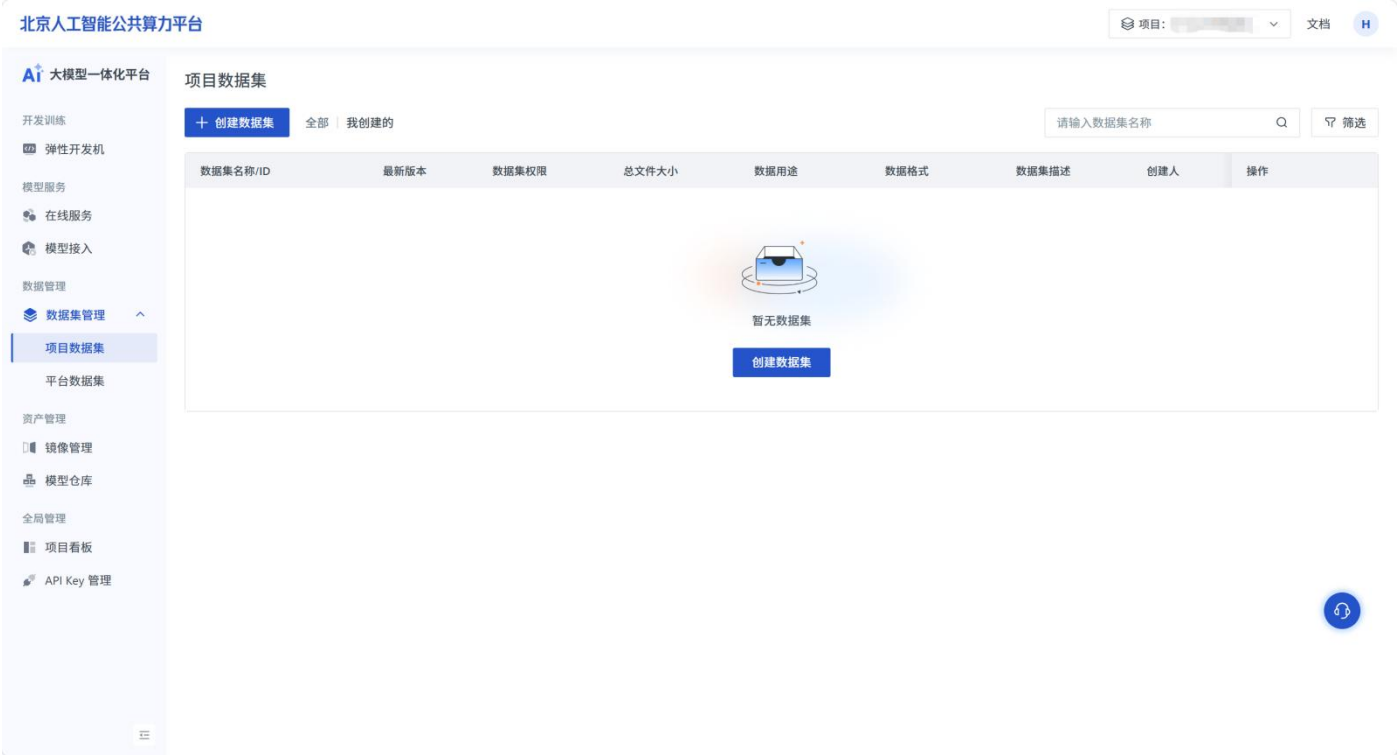
数据集由一个或多个版本组成，每个版本可挂载不同的文件。在平台各业务板块实际使用数据集时，需选择指定的版本。

一、项目数据集

功能概述：在项目空间内对数据集资产进行全生命周期统一管控，并通过版本管理、权限管理、状态控制等能力，保障数据集在多用户并发协作场景下的稳定性与可用性。

1.1 创建数据集

1. 在左侧导航栏选择“数据集管理-项目数据集”，进入数据集列表页。



2. 点击“创建数据集”按钮进入数据集创建页面并填写基本信息。创建数据集时会同时创建一个初始版本。

北京人工智能公共算力平台

文档H

← 创建数据集

1 创建数据集/版本信息

2 上传文件

数据集信息

所属项目

创建的数据集仅能在当前区域内加工和使用

数据集名称

请输入数据集名称

0/100

数据格式

请选择数据格式

数据用途

请选择数据用途

数据集描述

请输入数据集描述

0/200

权限设置

私有

只读

读写

该数据集在项目空间内仅能和项目管理员有查看和操作权限。

版本信息

区域

创建的版本仅能在当前区域内加工和使用

版本名称

请输入版本名称

0/20

版本描述

请输入版本描述

0/200

创建

取消

数据集相关信息描述及填写规则如下：

字段	描述	要求
所属项目	选择数据集归属的项目	必填，默认继承入口页顶部导航栏的项目
数据集名称	填写数据集的名称	必填 长度 1-100 字符 在项目空间内不能存在名称重复的数据集
数据格式	选择数据集内数据文件的格式，可选项包含：Text、Lance、Iceberg、CSV、JSONL、Parquet、Image、Video、Audio、多模态、其他，共 11 项	选填，单选
数据用途	选择数据集用途，可选项包含：预训练、模型蒸馏、有监督微调 SFT、偏好对齐 RFT、偏好对齐 DPO、偏好对齐 KTO、偏好对齐 RLHF、其他，共 8 项	选填，单选
数据集描述	填写数据集描述	选填 长度 0-200 字符
权限设置	设置数据集在当前项目空间内的权限，可选项包含：私有、只读、读写，共 3 种类型的权限	必填

初始版本相关字段及填写规则如下：

字段	描述	要求
区域	选择该版本存储的区域 请注意：创建的版本仅能在当前区域内使用，暂不支持跨区域同步数据	必填，默认继承入口页顶部导航栏的项目
版本名称	填写版本名称	必填 长度 1-100 字符 单一数据集下不能存在名称重复的

		版本名称
版本描述	填写版本描述	• 选填 • 长度 0-200 字符

点击“创建”按钮并成功完成创建后，页面会跳转至数据集初始版本的文件管理页，以便上传文件。

3.通过 CLI 的方式上传文件（目前仅支持 windows 系统，以及 python 3.12.0 版本）

北京人工智能公共算力平台

文档H

← 创建数据集

创建数据集/版本信息

上传文件

请先 下载 CLI 工具 ，然后依照代码模板上传/删除数据集，命令执行后可在 版本列表页 查看上传进度

版本 ID: Access Key: Secret Key:

CLI 文件编辑代码

暂时只支持Windows系统:

请先确认已安装 python 3.12.0, 由于工具使用Cython编译, 只能使用3.12.0版本

安装依赖:

需自行安装 Visual C++ Redistributable

下载安装工具包:

点击上方“下载 CLI 工具”

(请注意先删除存储目录的同名文件, 否则会自动在新下载文件名后面加上编号导致命名不符合规范, 无法安装)

使用以下命令安装:

pip install <下载的工具文件路径> --force-reinstall

查看帮助:

jndataset-up --help

登录

jndataset-up login --ak "<Access Key>" --sk "<Secret Key>"

上传单个文件

参数说明:

"version-id": 版本 ID (必需)

"source-path": 本地文件路径 (必需)

返回版本列表

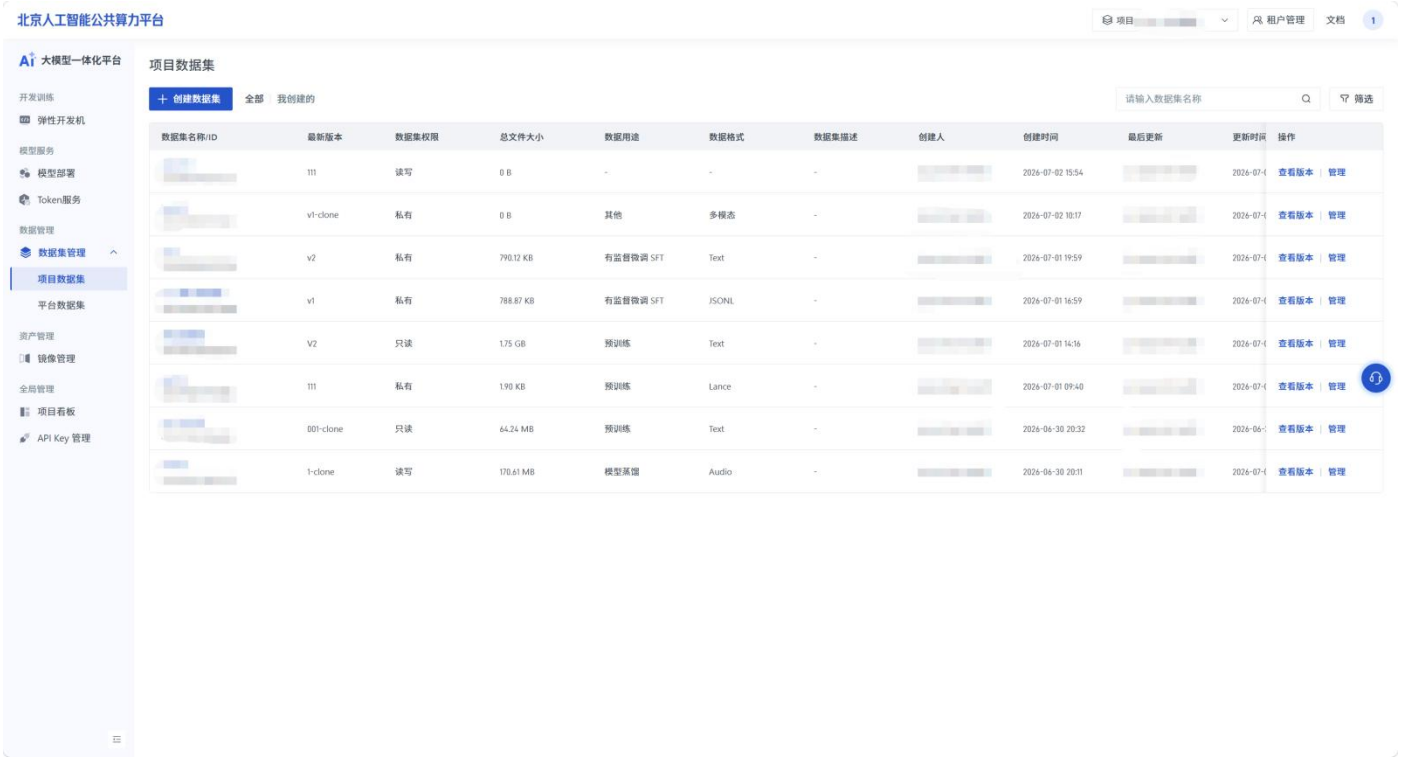
点击“使用 CLI 编辑”按钮展开弹窗，按照弹窗中的代码模版上传文件。

规则详见：1.2.6.7.1

1.2 管理数据集

1.2.1 查看数据集列表

在左侧导航栏选择“数据集管理-项目数据集”，进入数据集列表页。在此页面可查看当前项目空间内所有有权限查看的数据集资产。



数据集列表中的字段说明如下：

字段	描述
数据集名称/ID	数据集名称：创建/编辑数据集时填写的数据集名称 数据集 ID：数据集的唯一标识，成功创建数据集时由系统自动生成
最新版本	展示当前数据集下最近一次创建的版本
数据集权限	展示当前数据集的权限，包括：私有、只读、读写
总文件大小	展示当前数据集下所有版本（包括不同区域下的）文件大小的总和
数据用途	展示创建/编辑数据集时填写的数据用途
数据格式	展示创建/编辑数据集时填写的数据格式
数据集描述	展示创建/编辑数据集时填写的数据集描述
创建人	展示该条数据集创建人的用户名
创建时间	展示该条数据集的创建时间
最后更新	展示最后一次编辑该条数据集的用户名
更新时间	展示最后一次编辑该条数据集的时间

搜索与筛选

如果项目空间内的数据集数量较多，可通过点击数据集列表左上方的按钮并切换查看“全部”或“我创建的”数据集。也可通过数据集列表右上角的“搜索”或“筛选”功能查找指定的数据集。

1.2.2 编辑数据集信息

点击数据集列表中的“操作-管理-编辑信息”，可在弹窗中编辑修改数据集的基础信息，包括：数据及名称、数据格式、数据用途、数据集描述。

编辑

*数据集名称

测试数据集-test10/100

数据格式

请选择数据格式

数据用途

请选择数据用途

数据集描述

请输入数据集描述0/200

取消

确认

1.2.3 删除数据集

点击数据集列表中的“操作-管理-删除”，可删除指定数据集。
如果数据集下有任一版本正处于“使用中”、“可用”、“编辑中”、“克隆中”、“被克隆中”状态，则该数据集不可删除。请等待所有版本任务结束并切换为“草稿”状态再删除数据集。

删除数据集

×

您确认要删除当前的数据集吗？删除后将不可恢复，请确认。

当前数据集：测试数据集

包含版本：V1

取消

删除

1.2.4 管理数据集权限

租户管理员、项目管理员和数据集创建者可以设置数据集在当前项目空间内的权限。目前包含：私有、只读、读写，三种类型的权限可设置。

1.2.4.1 私有权限

“私有”权限的数据集						
• 数据集发布人、项目管理员、租户管理员、平台超级管理员 --- 可见、可用、可编、可更改权限						
• 项目空间内的其他用户 --- 不显示该数据集						
角色	操作	模块	数据集列表页	版本列表页	文件管理	训练、开发机等其他业务模块（敬请期待）
数据集发布人、项目管理员、租户管理员	可执行的操作		•可在列表中查看数据集基础信息 •可编辑数据集基础信息 •可编辑数据集权限 •可查看操作日志 •可删除数据集	•可在列表中查看版本基础信息 •可编辑版本名称 •可编辑版本描述 •可切换版本状态 •可克隆版本	•可查看版本文件 •可预览版本文件 •可通过 CLI 编辑文件 •可通过 CLI 下载文件	•可在其他业务模块获取并应用该数据集

			• 可删除版本		
	不可执行的操作	-	-	-	-
项目空间内的其他用户	可执行的操作	-	-	-	-
	不可执行的操作	不显示该数据集，所有操作均无法执行			

1.2.4.2 只读权限

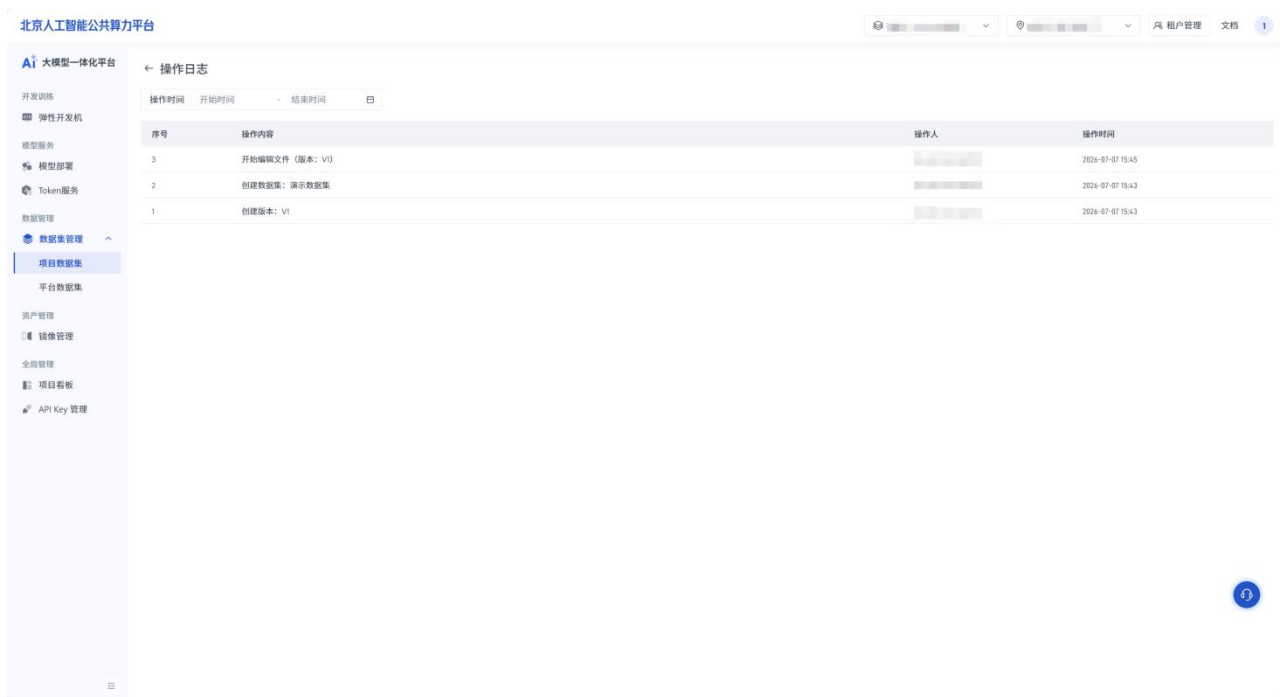
“只读”权限的数据集						
• 数据集发布人、项目管理员、租户管理员、平台超级管理员 --- 可见、可用、可编、可更改权限 • 项目空间内的其他用户 --- 可见、可用						
角色	操作 \ 模块	数据集列表页	版本列表页	文件管理	训练、开发机等其他业务模块 (敬请期待)	
数 据 集 发 布 人、项目管理员、租户管理员	可执行的操作	•可在列表中查看数据集基础信息 •可编辑数据集基础信息 •可编辑数据集权限 •可查看操作日志 •可删除数据集	•可在列表中查看版本基础信息 •可编辑版本名称 •可编辑版本描述 •可切换版本状态 •可克隆版本 •可删除版本	•可查看版本文件 •可预览版本文件 •可通过 CLI 编辑文件 •可通过 CLI 下载文件	•可在其他业务模块获取并应用该数据集	
	不可执行的操作	-	-	-	-	
项目空间内的其他用户	可执行的操作	•可在列表中查看数据集基础信息 •可查看操作日志	•可在列表中查看版本基础信息	•可查看版本文件 •可预览版本文件 •可通过 CLI 下载文件	•可在其他业务模块获取并应用该数据集	
	不可执行的操作	•不可编辑数据集基础信息 •不可编辑数据集权限 •不可删除数据集	•不可编辑版本名称 •不可编辑版本描述 •不可切换版本状态 •不可克隆版本 •不可删除版本	• 不可通过 CLI 编辑文件	-	

1.2.4.3 读写权限

“读写”权限的数据集					
• 仅数据集发布人、项目管理员、租户管理员、平台超级管理员，可见可用可编可更改权限 • 项目空间内的其他用户，可见可用可编					
角色	操作 \ 模块	数据集列表页	版本列表页	文件管理	训练、开发机等其他业务模块 (敬请期待)
数据集发布人、项目管理员、租户管理员	可执行的操作	•可在列表中查看数据集基础信息 •可编辑数据集基础信息 •可编辑数据集权限 •可查看操作日志 •可删除数据集	•可在列表中查看版本基础信息 •可编辑版本名称 •可编辑版本描述 •可切换版本状态 •可克隆版本 •可删除版本	•可查看版本文件 •可预览版本文件 •可通过 CLI 编辑文件 •可通过 CLI 下载文件	•可在其他业务模块获取并应用该数据集
	不可执行的操作	-	-	-	-
项目空间内的其他用户	可执行的操作	•可在列表中查看数据集基础信息 •可编辑数据集基础信息 •可查看操作日志 •可删除数据集	•可在列表中查看版本基础信息 •可编辑版本名称 •可编辑版本描述 •可切换版本状态 •可克隆版本 •可删除版本	•可查看版本文件 •可预览版本文件 •可通过 CLI 编辑文件 •可通过 CLI 下载文件	•可在其他业务模块获取并应用该数据集
	不可执行的操作	•不可编辑数据集权限	-	-	-

1.2.5 查看操作日志

数据集创建完成后，点击数据集列表页中的“操作-管理-操作日志”，可以查看数据集的所有历史操作记录。



1.2.6 管理数据集版本

在数据集列表中点击数据集名称，或点击“操作-查看版本”，可进入数据集版本列表页。

1.2.6.1 创建数据集版本

1. 点击“创建版本”按钮并在弹窗中输入版本信息。

创建版本

* 区域

创建的版本仅能在当前区域内加工和使用

* 版本名称

请输入版本名称

0/20

版本描述

请输入当前版本描述

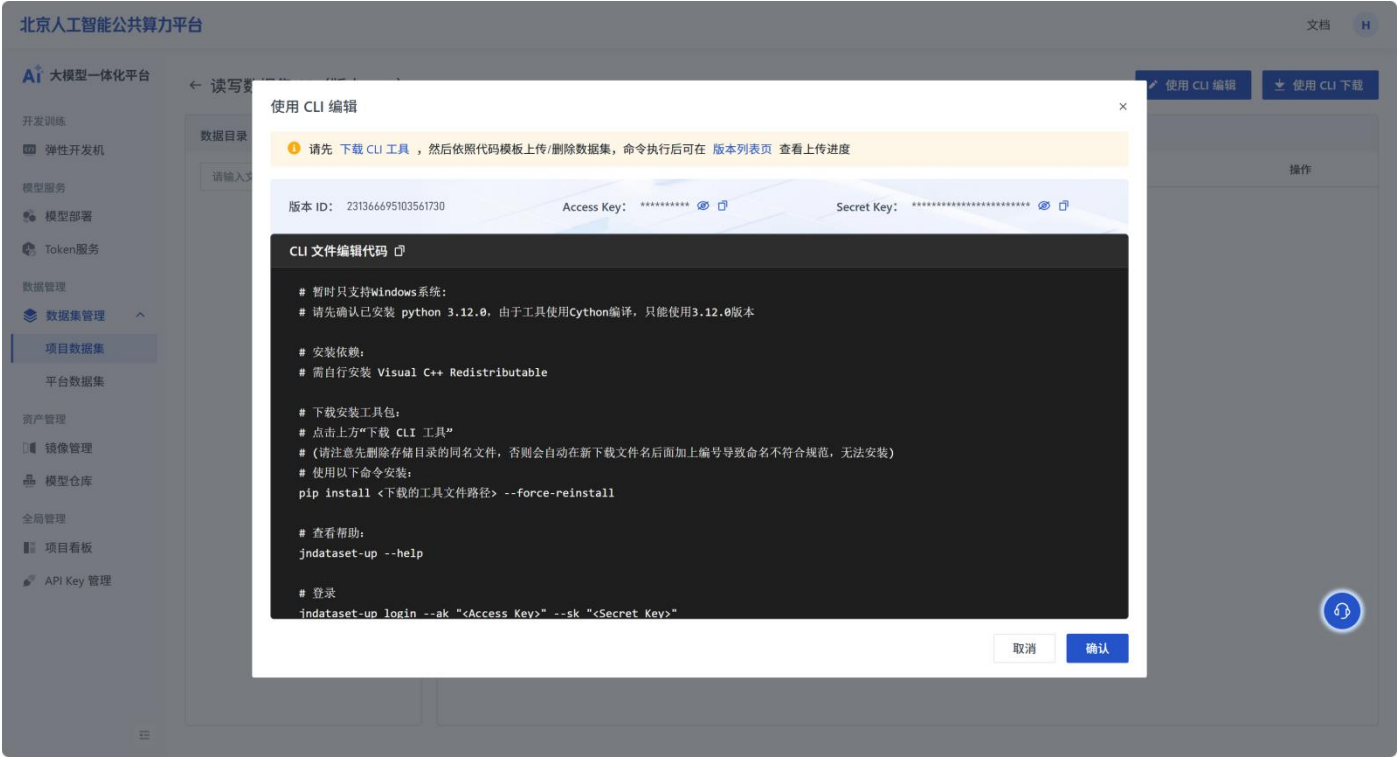
0/200

取消

创建

点击“创建”按钮并成功完成创建后，页面会跳转至数据集初始版本的文件管理页，以便上传文件。

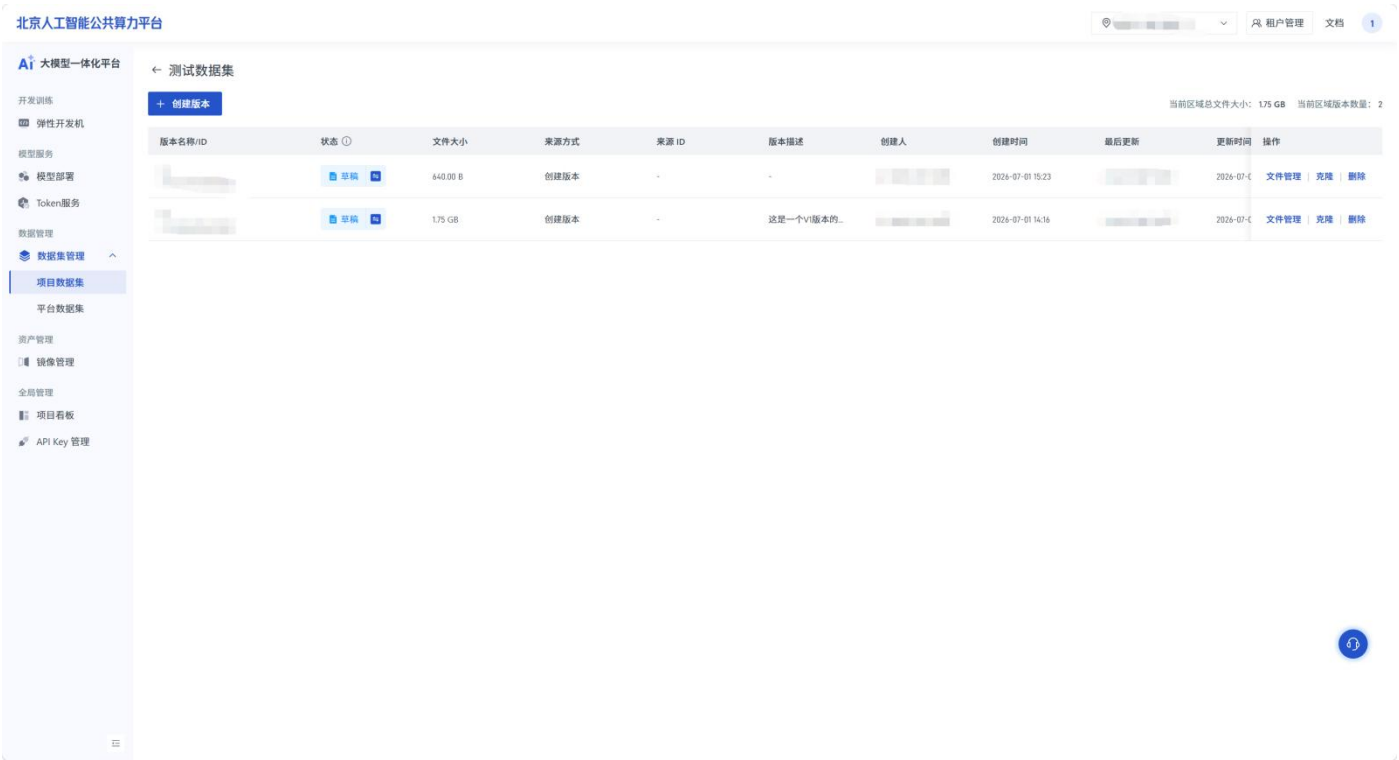
2.通过 CLI 的方式上传文件（目前仅支持 windows 系统，以及 python3.12.0 版本）



点击“使用 CLI 编辑”按钮展开弹窗，按照弹窗中代码模版的上传文件。
规则详见：1.2.6.7.1

1.2.6.2 查看数据集版本列表

在数据集列表页直接点击数据集名称，或点击“操作-查看版本”，即可进入数据集版本列表页查看当前数据集的所有版本信息。



数据集版本列表中的字段说明如下：

字段	描述
----	----

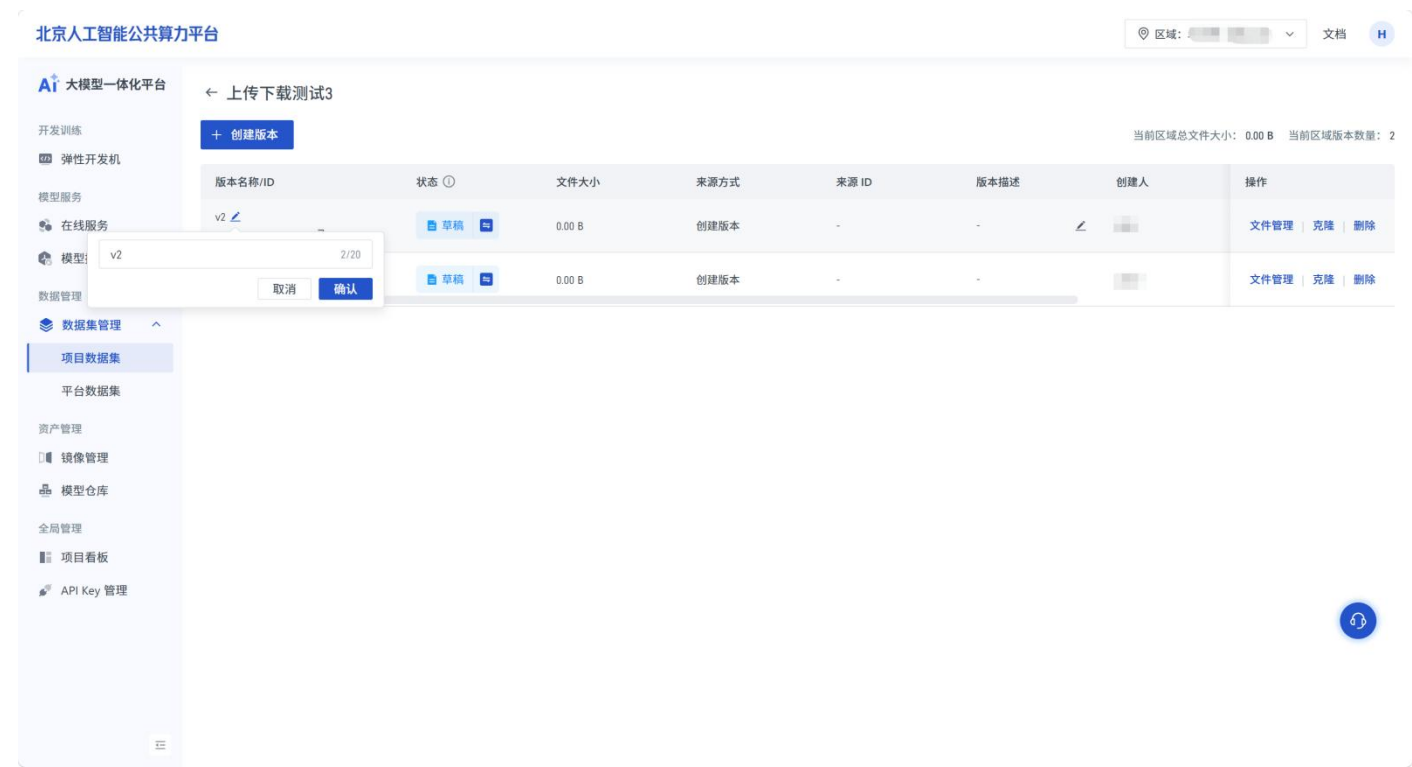
版本名称/ID	版本名称：创建/编辑数据集版本时填写的数据集版本名称 版本 ID：数据集版本的唯一标识，成功创建数据集版本时由系统自动生成
状态	展示当前版本的状态，包括：可用、使用中、草稿、编辑中、克隆中、被克隆中、克隆失败
文件大小	展示当前版本的文件大小
来源方式	展示当前版本的创建方式，包括：创建版本、版本克隆
来源 ID	展示当前版本的父版本 ID
版本描述	展示创建/编辑数据集版本时填写的数据集版本描述
创建人	展示该条数据集版本创建人的用户名
创建时间	展示该条数据集版本的创建时间
最后更新	展示最后一次编辑该条数据集版本的用户名
更新时间	展示最后一次编辑该条数据集版本的时间

如果在列表中找不到您创建的版本，请尝试在顶部导航栏切换区域。

1.2.6.3 编辑数据集版本信息

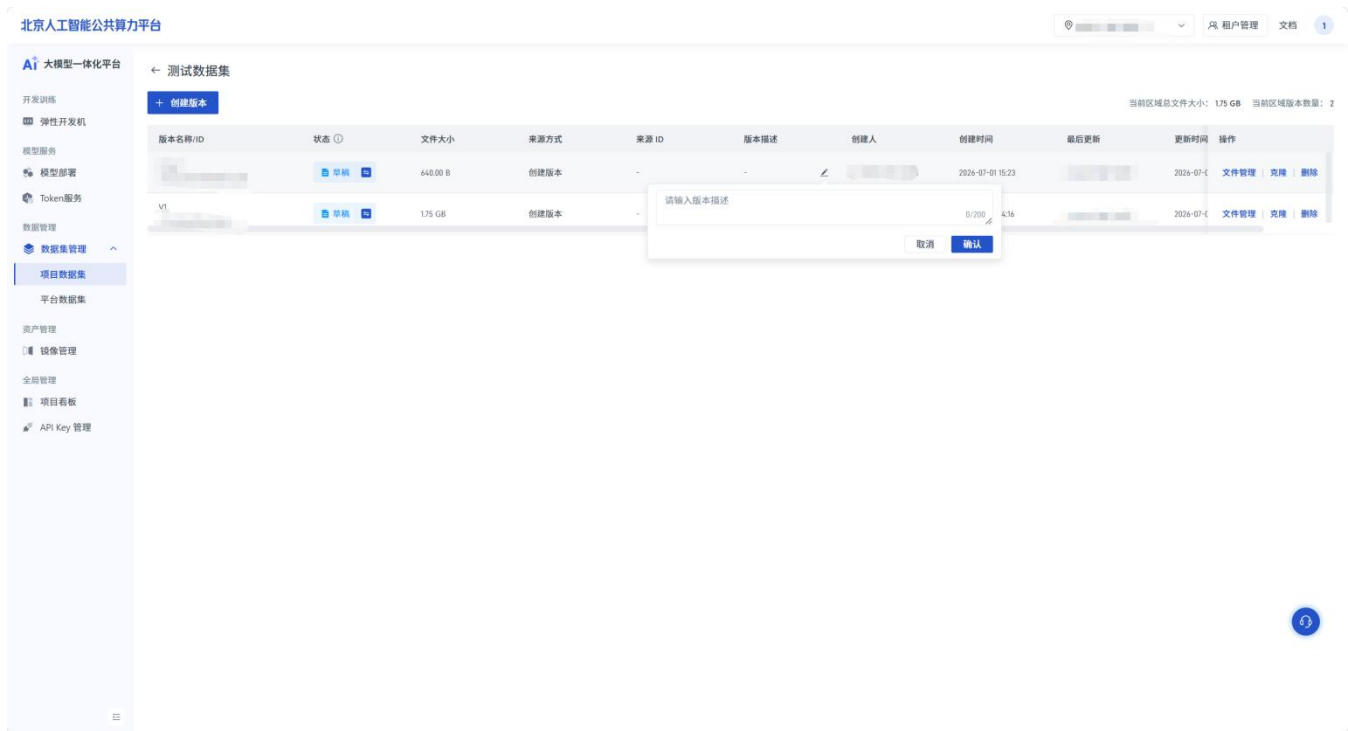
1.2.6.3.1 编辑版本名称

将鼠标移入版本列表，并点击版本名称右侧的编辑图标，即可编辑版本名称。



1.2.6.3.2 编辑版本描述

将鼠标移入版本列表，并点击版本描述右侧的编辑图标，即可编辑版本描述。



1.2.6.4 删除数据集版本

点击版本列表中的“操作-删除”，可删除指定版本。
如果版本正处于“使用中”、“可用”、“编辑中”、“克隆中”、“被克隆中”状态，则该版本不可删除。请等待版本任务结束并切换为“草稿”状态再删除该版本。



1.2.6.5 克隆版本

点击版本列表中的“操作-克隆”，可弹出版本克隆弹窗，用户需在弹窗中填写子版本的信息。完成后点击“创建”按钮开始执行版本克隆任务。

克隆版本

×

父版本：父版本名称：

克隆后生成的新版本将继承父版本的所有文件

* 区域

创建的版本仅能在当前区域内加工和使用

* 版本名称

V2-clone

8/20

版本描述

请输入当前版本描述

0/200

取消

创建

克隆后的子版本会在版本列表页中的“来源方式”一列标注为“版本克隆”，在“来源 ID”标注父版本的 ID。

克隆后的子版本会继承父版本的所有文件。

1.2.6.6 版本状态

1.2.6.6.1 状态说明

数据集版本状态包含“可用”、“使用中”、“草稿”、“编辑中”、“克隆中”、“被克隆中”、“克隆失败”，共 7 种。

- ✔ 可用

可用于克隆，如需编辑/删除该数据集文件，请切换为"草稿"状态
- 🔄 使用中

正处于克隆，待任务完成或停止后将变为"可用"状态
- 📄 草稿

数据集可编辑/克隆文件/删除版本，如需使用数据集文件下载，请切换为"可用"状态
- 🔄 编辑中

数据集正在上传/编辑文件中，待编辑完成后将变为"草稿"状态
- 🔄 克隆中

数据集正作为子版本克隆中，待克隆完成后将变为"草稿"状态
- 🔄 被克隆中

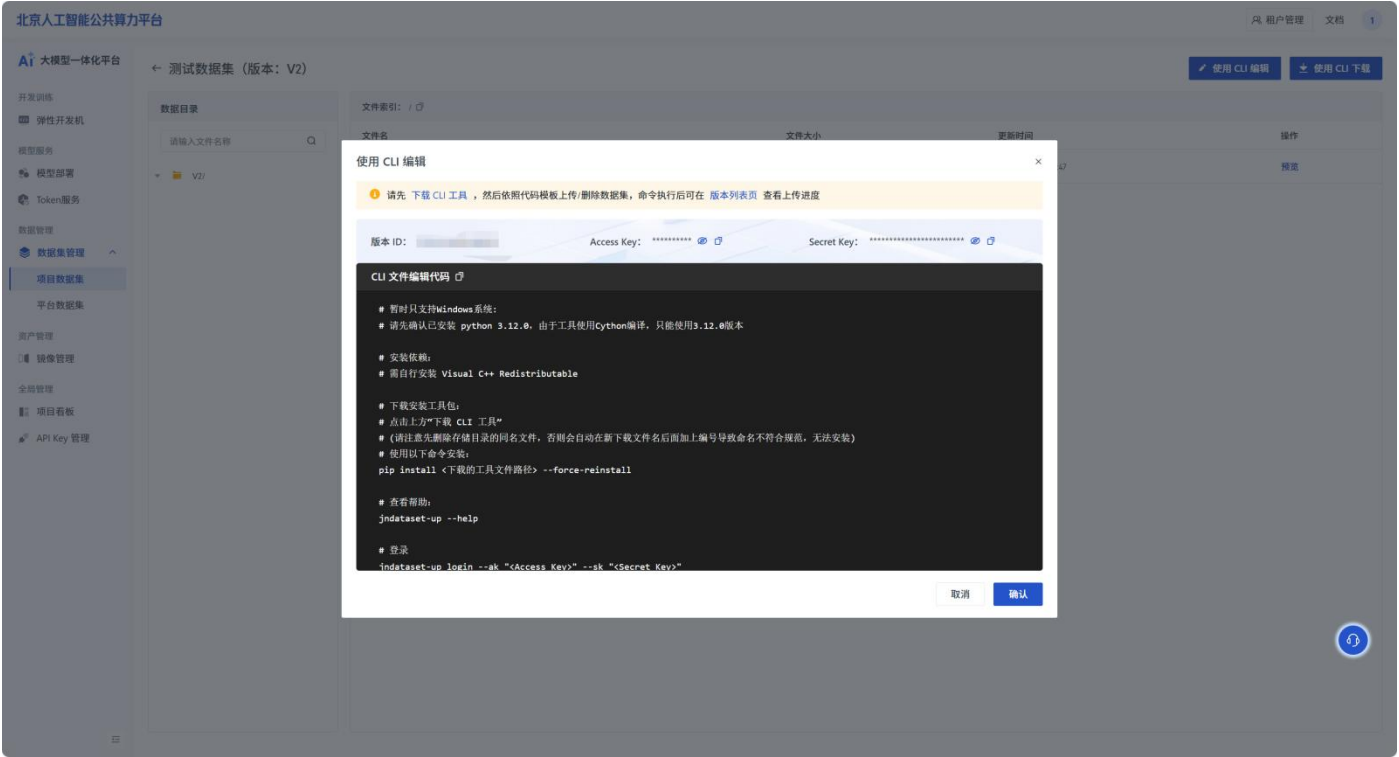
数据集正作为父版本被克隆中，可用于其他克隆任务，待克隆完成后将切换为"草稿"状态
- ✖ 克隆失败

"克隆失败"状态下的数据集，仅支持删除。

1.2.6.6.2 状态切换

状态流转图如下所示：

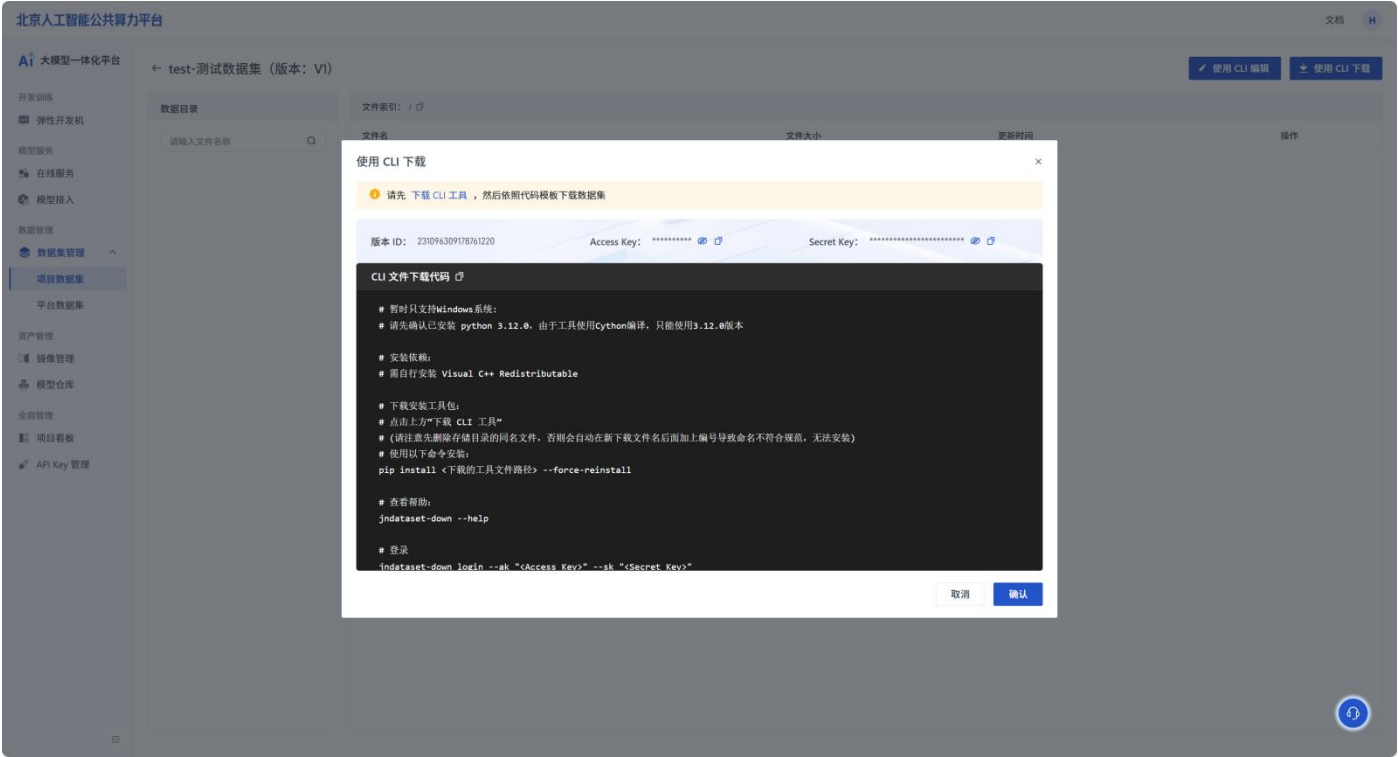
1.2.6.7.1 文件编辑



支持通过 CLI 的方式编辑文件（目前仅支持 windows 系统，仅支持 python 3.12.0 版本），具体步骤如下：

- 点击“使用 CLI 编辑”按钮展开弹窗，按照弹窗中代码模板的上传、删除文件。
- 安装 python 环境，当前仅支持 3.12.0 版本。
- 安装 Visual C++ Redistributable
- 点击页面中的“下载 CLI 工具”链接下载平台提供的 CLI 工具，所有的文件上传、下载、删除命令都需要依赖该工具执行。
- 执行“`pip install <下载的工具文件路径> --force-reinstall`”命令安装 CLI 工具。请注意先删除存储目录的同名文件，否则会无法安装。
- 安装完成后，参照 CLI 文件编辑代码模版登录并执行上传、删除命令。

1.2.6.7.2 文件下载

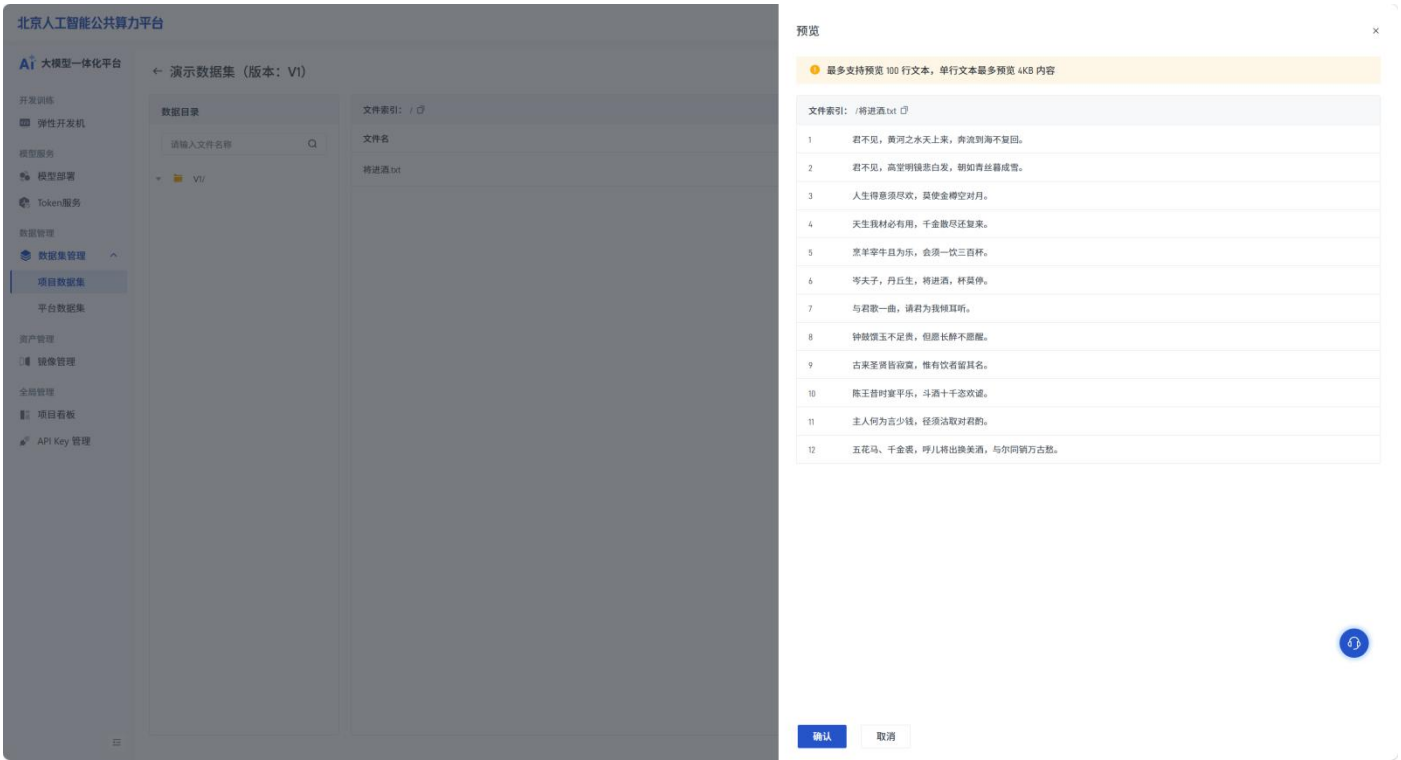
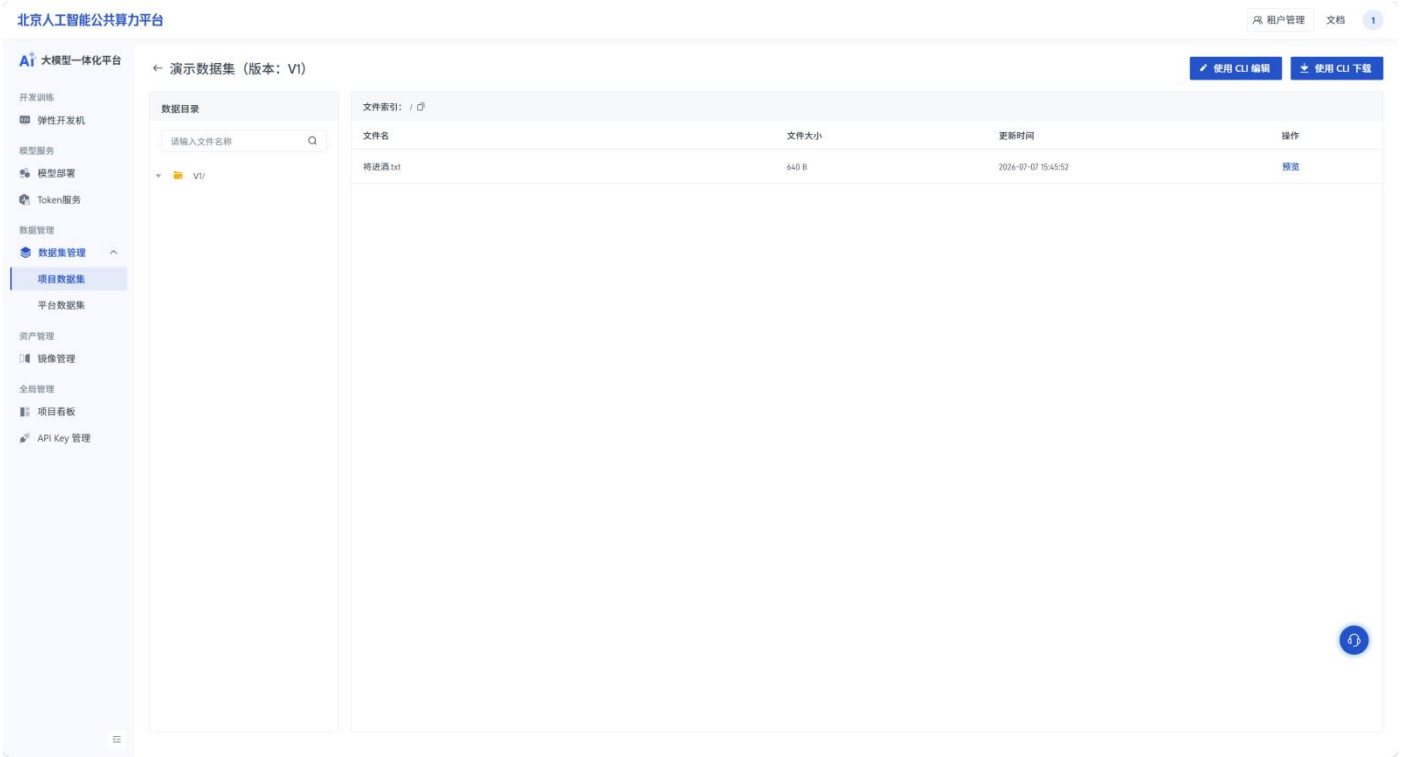


支持通过 CLI 的方式下载文件（目前仅支持 windows 系统，仅支持 python 3.12.0 版本），具体步骤如下：

- 点击“使用 CLI 下载”按钮展开弹窗，按照弹窗中代码模版的下载文件。
- 安装 python 环境，当前仅支持 3.12.0 版本。
- 安装 Visual C++ Redistributable
- 点击页面中的“下载 CLI 工具”链接下载平台提供的 CLI 工具，所有的文件上传、下载、删除命令都需要依赖该工具执行。
- 执行“pip install <下载的工具文件路径> --force-reinstall”命令安装 CLI 工具。请注意先删除存储目录的同名文件，否则会无法安装。
- 安装完成后，参照 CLI 文件编辑代码模版登录并执行下载命令。

1.2.6.7.3 文件预览

在文件管理页点击“操作-预览”即可在线预览指定的文件，如下图所示：



可预览的文件类型及预览规则如下：

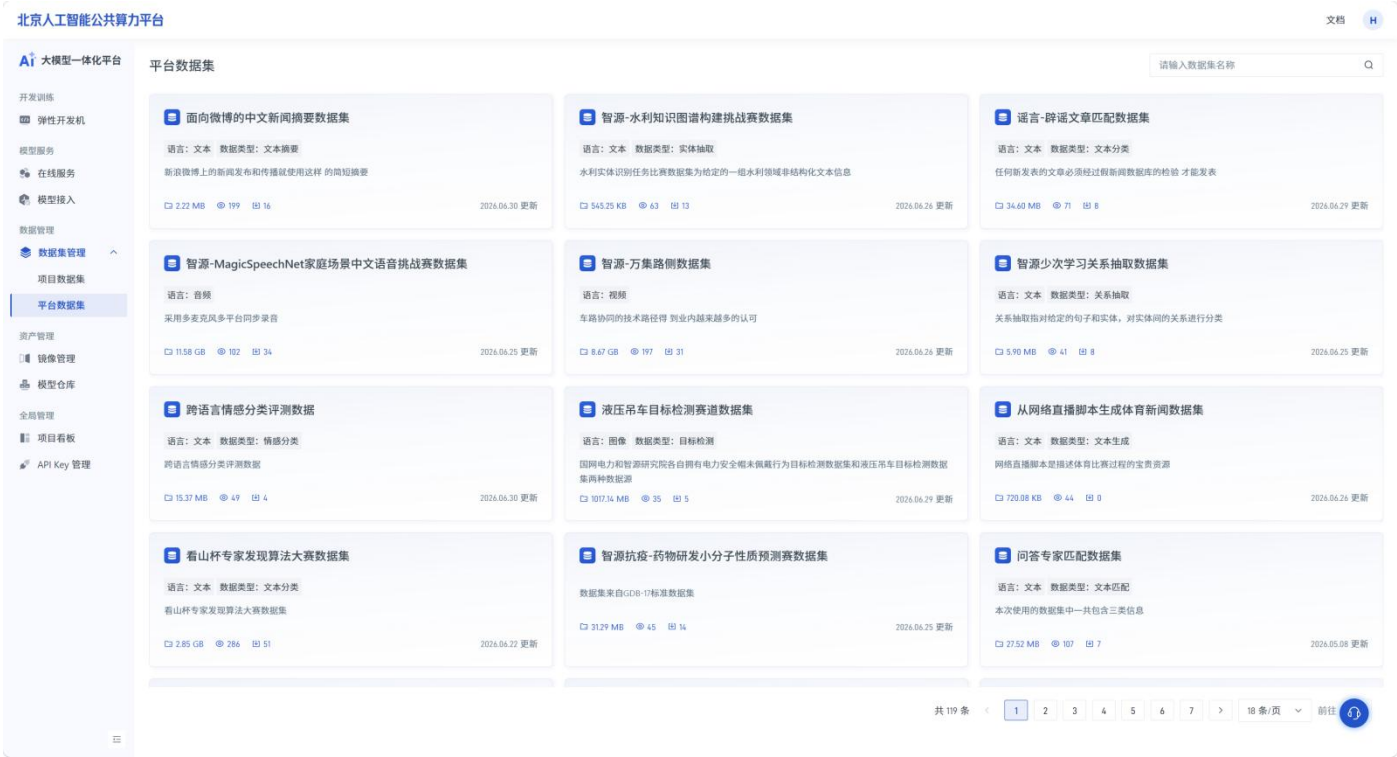
序号	支持类型	具体文件后缀	预览规则
1	文本	txt/markdown/javascript/css/html/plain	• 截取512KB进行解析和预览。最多预览100行； • 单行文本仅支持预览4KB内容； • 若第一行数据大小大于512KB，则无法预览
2		json/jsonl	• 截取20M进行解析和预览，最多预览100行； • 单行文本仅支持预览4KB内容； • 若第一行数据大小大于20M，则无法预览
3	视频	mp4/webm	• 预览时长不超过15s • 损坏、无法识别、加载超时的视频提示无法预览
4	图片	jpeg/png/gif/webp/bmp/tiff/jpg	• 仅支持预览小于等于8000万像素的图片 • 损坏、无法识别、加载超时的图片提示无法预览
5	音频	mpeg/wav/ogg/mp3	• 预览时长不超过15s • 损坏、无法识别、加载超时的音频提示无法预览

二、平台数据集

功能概述：支持用户浏览并下载平台高质量开源数据集， 以在模型训练、微调等场景使用。

2.1 查看平台数据集列表

在左侧导航栏选择“数据集管理-平台数据集”， 进入平台数据集列表页。用户可在该模块查看平台所有的开源数据集。



2.2 查看平台数据集详情

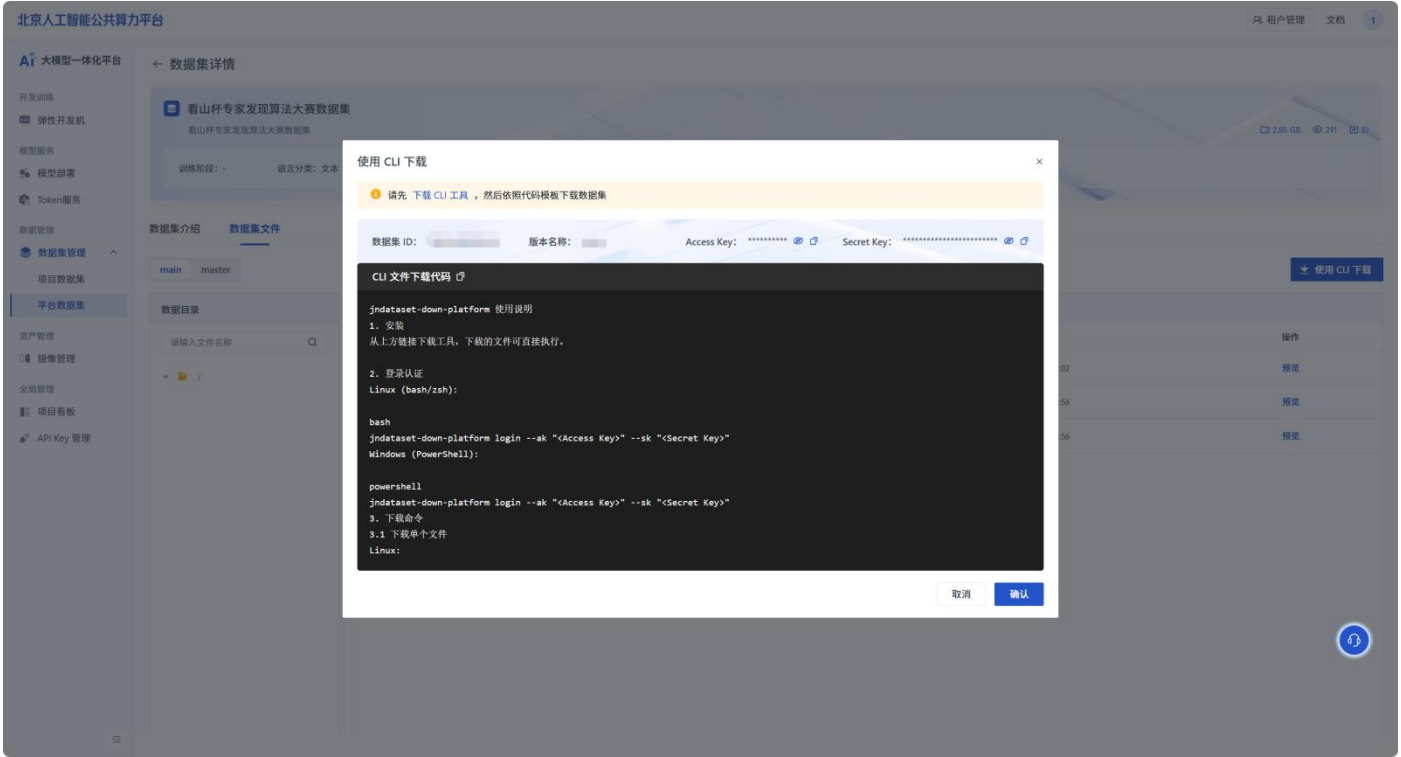
在平台数据集列表页点击指定的数据集可进入该数据集详情页，用户可在该页面浏览数据集更多信息。



2.3 查看平台数据集文件

2.3.1 平台数据集下载

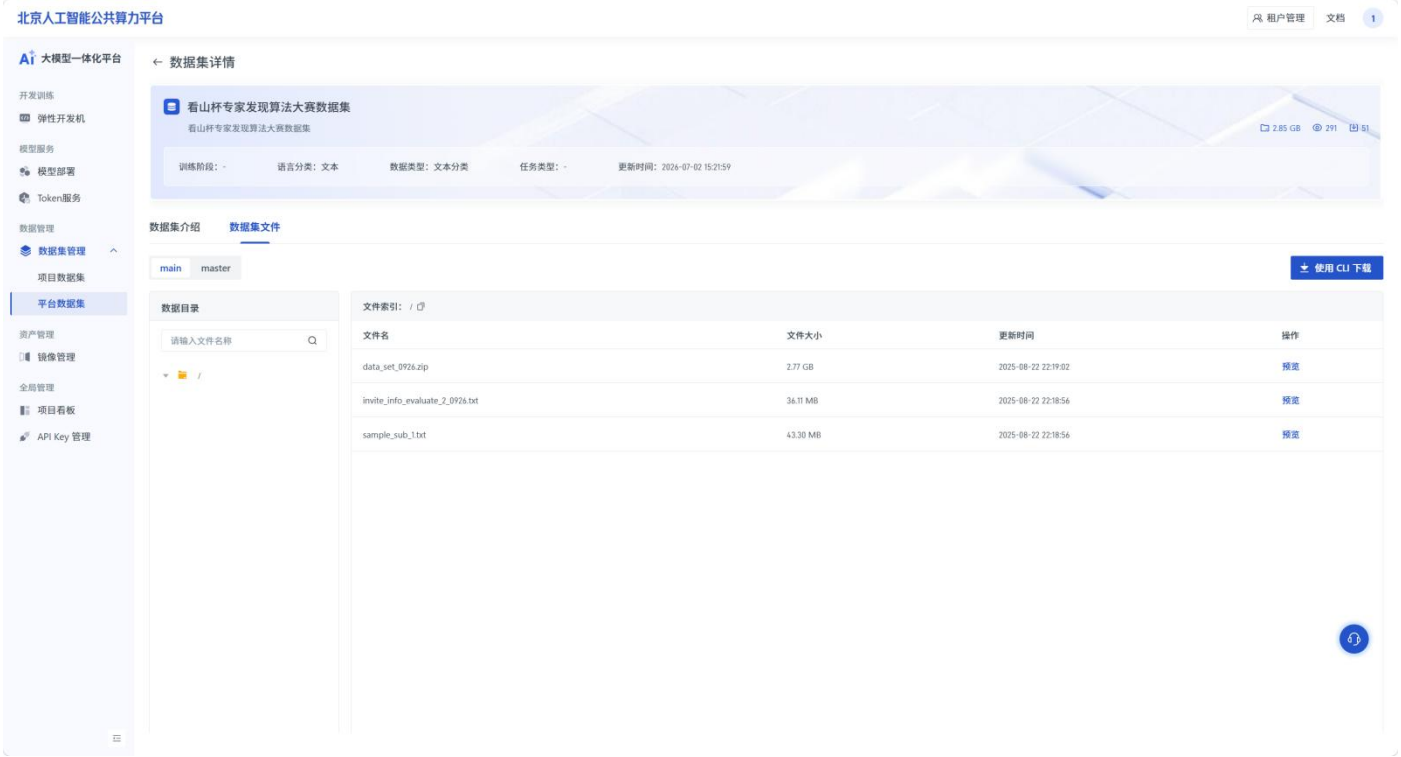
支持通过 CLI 的方式下载文件（目前仅支持 windows 系统，仅支持 python 3.12.0 版本）

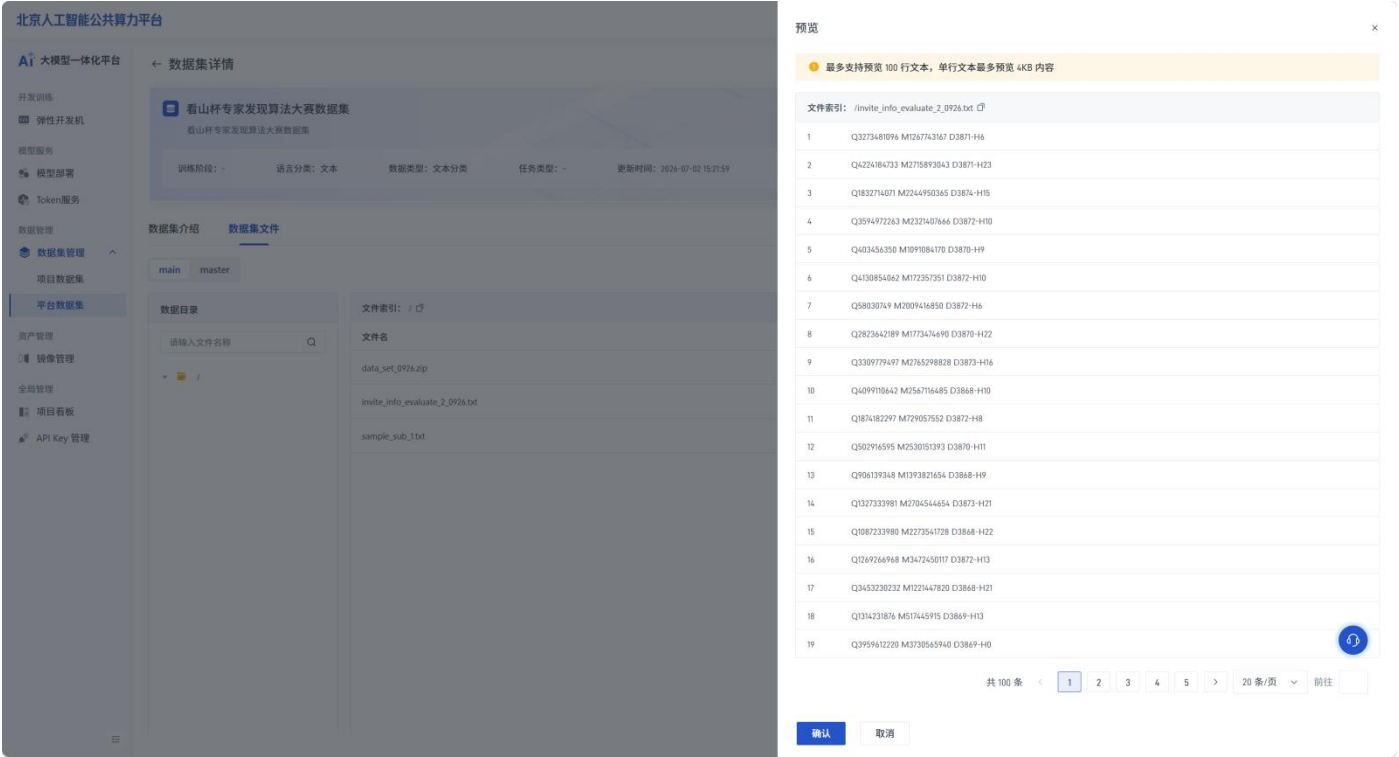


点击“使用 CLI 下载”按钮展开弹窗，按照弹窗中代码模版的下载文件。
规则详见：1.2.6.7.2

2.3.2 平台数据集预览

在平台数据集详情页点击“数据集文件-操作-预览”即可在线预览指定的文件，如下图所示：





可预览的文件类型及预览规则如下：

序号	支持类型	具体文件后缀	预览规则
1	文本	txt/markdown/javascript/css/html/plain	• 截取512KB进行解析和预览。最多预览100行； • 单行文本仅支持预览4KB内容； • 若第一行数据大小大于512KB，则无法预览
2		json/jsonl	• 截取20M进行解析和预览，最多预览100行； • 单行文本仅支持预览4KB内容； • 若第一行数据大小大于20M，则无法预览
3	视频	mp4/webm	• 预览时长不超过15s • 损坏、无法识别、加载超时的视频提示无法预览
4	图片	jpeg/png/gif/webp/bmp/tiff/jpg	• 仅支持预览小于等于8000万像素的图片 • 损坏、无法识别、加载超时的图片提示无法预览
5	音频	mpeg/wav/ogg/mp3	• 预览时长不超过15s • 损坏、无法识别、加载超时的音频提示无法预览