# Xinference 在windows部署指南

**安装**

Xinference 在 Linux, Windows, MacOS 上都可以通过 pip 来安装。如果需要使用 Xinference 进行模型推理，可以根据不同的模型指定不同的引擎。

如果你希望能够推理所有支持的模型，可以用以下命令安装所有需要的依赖：

pip install "xinference[all]"

windows因有库不支持，建议使用Transformers 引擎，其他引擎暂未在windows上测试：

**Transformers 引擎**

PyTorch(transformers) 引擎支持几乎有所的最新模型，这是 Pytorch 模型默认使用的引擎：

pip install "xinference[transformers]"

**使用**

**本地运行 Xinference**

让我们以一个经典的大语言模型 glm4-chat 来展示如何在本地用 Xinference 运行大模型。运行前请先设置环境变量，避免模型下载在C盘，XINFERENCE\_HOME=F:\XinferenceCache，XINFERENCE\_MODEL\_SRC=modelscope，设置完后重启电脑

在安装前需要先安装Microsoft C++ 生成工具

下载地址：

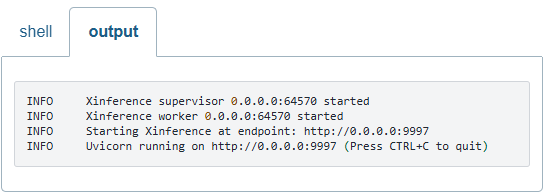
https://visualstudio.microsoft.com/zh-hans/visual-cpp-build-tools/

安装时只需勾选一项就行

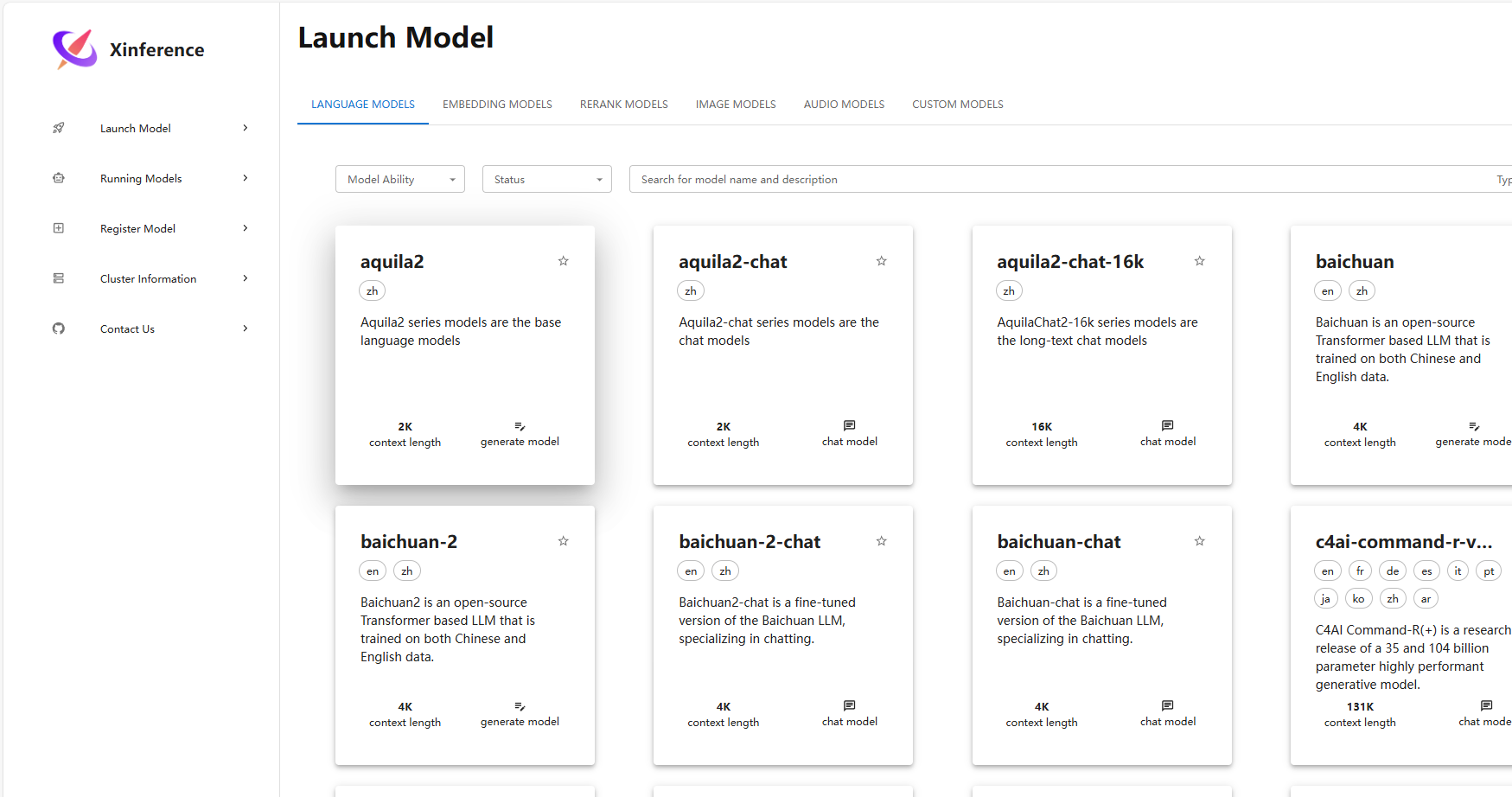


xinference-local --host 127.0.0.1 --port 9997

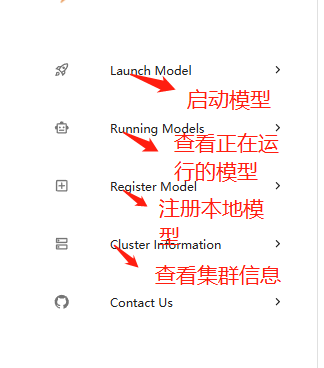
成功运行以上命令后，正常输出如下图：



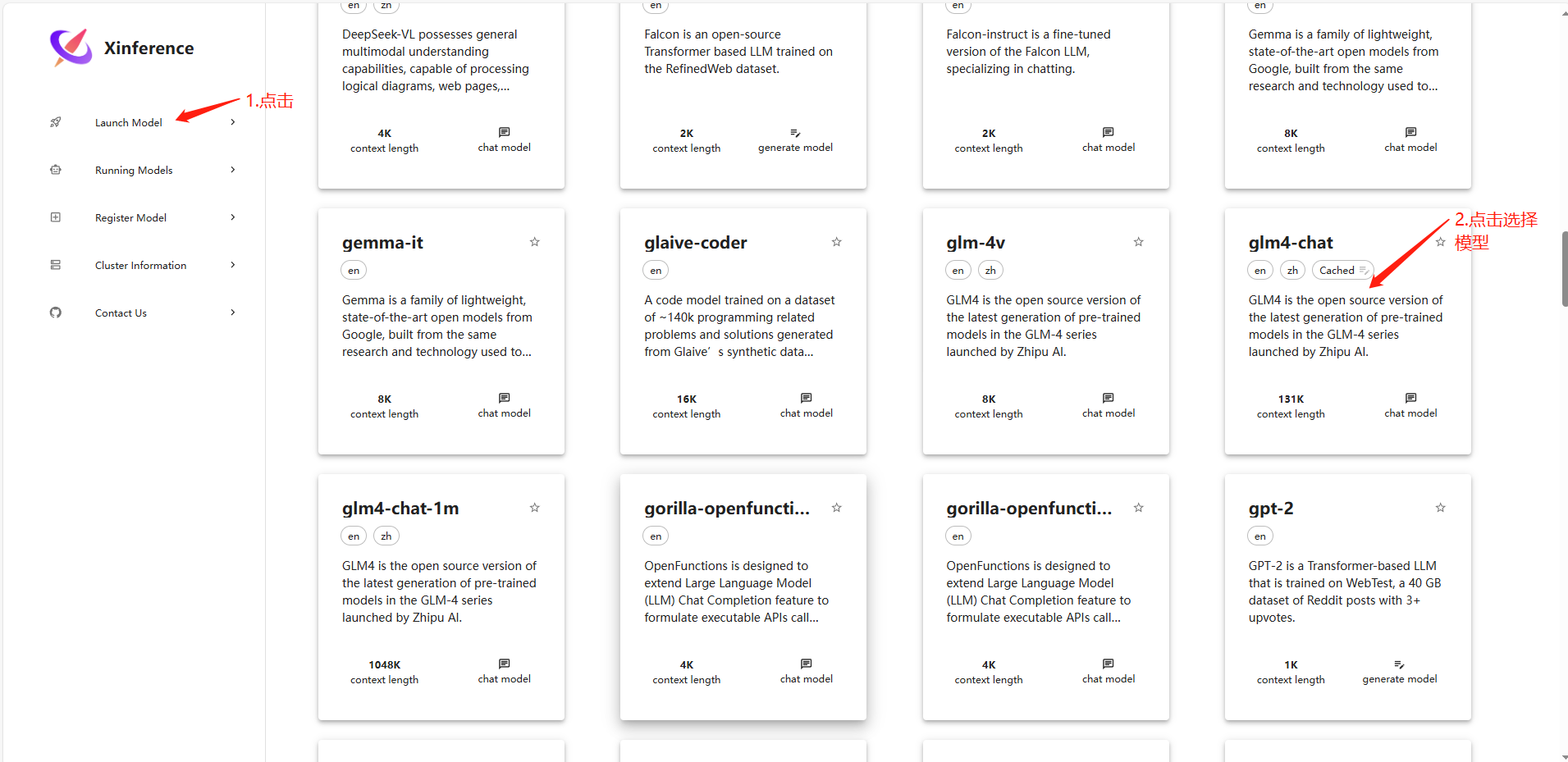
点击终端输出的网址：<http://127.0.0.1:9997>，可以看到如下页面：

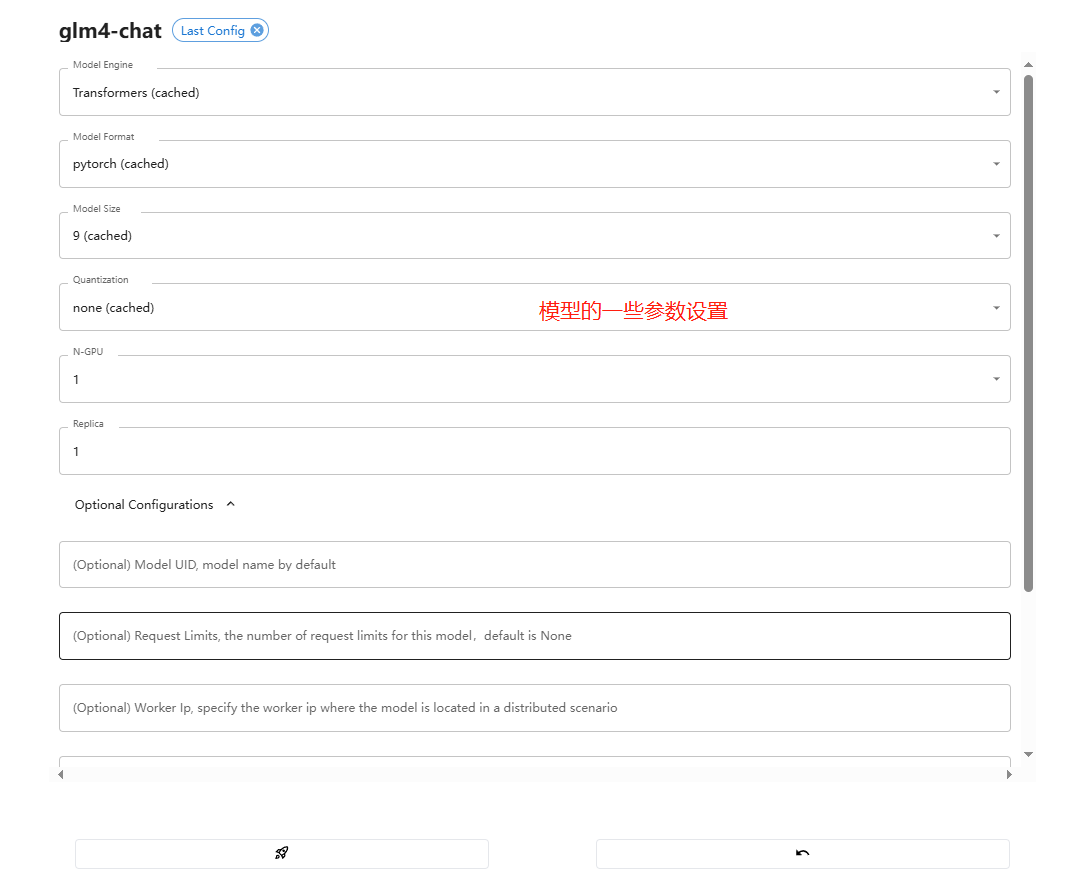


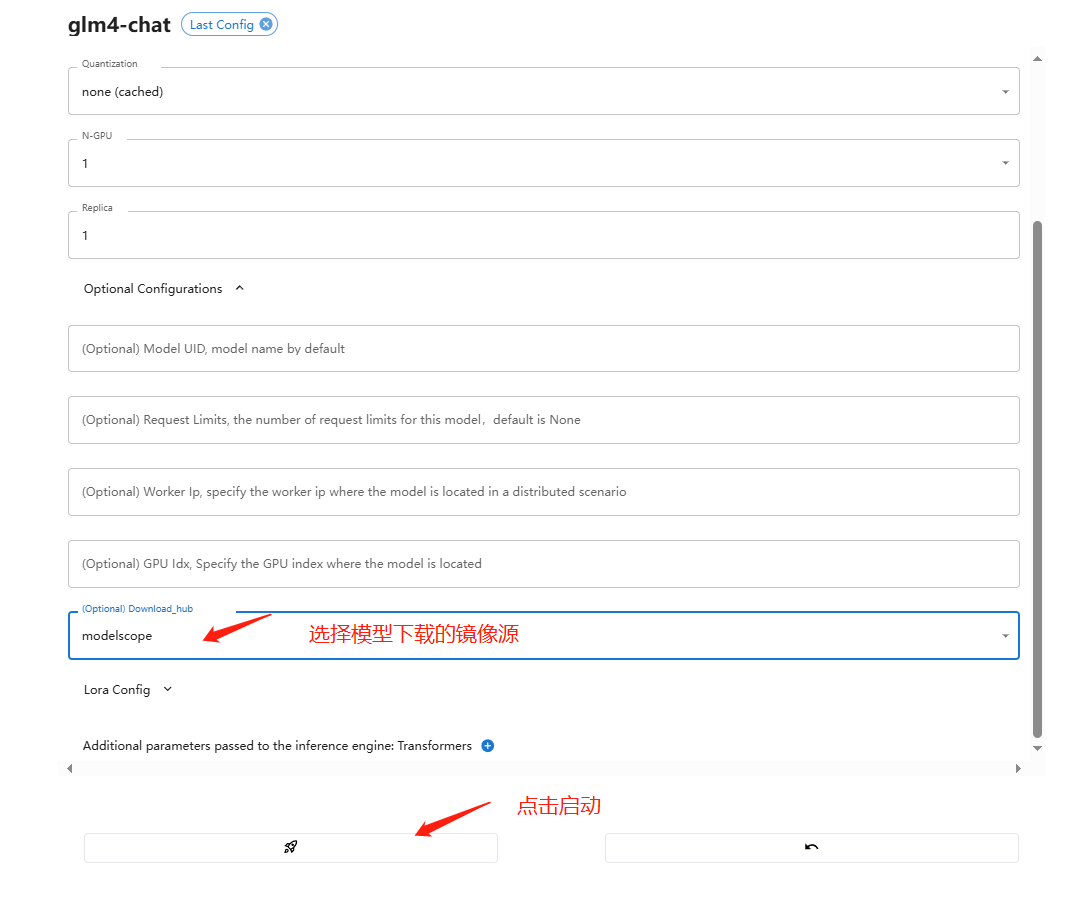
现在分别为功能进行说明：



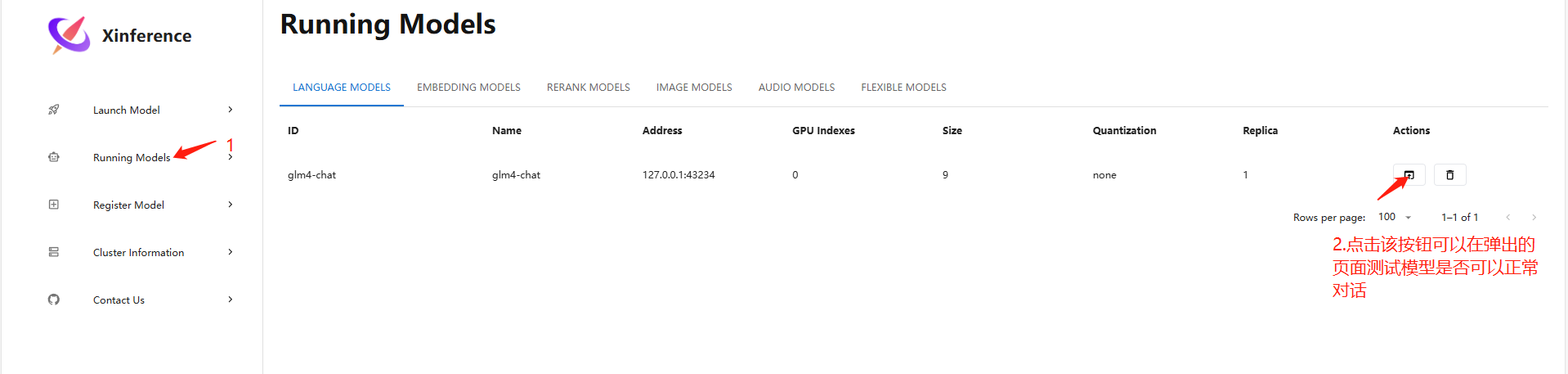
### 启动模型

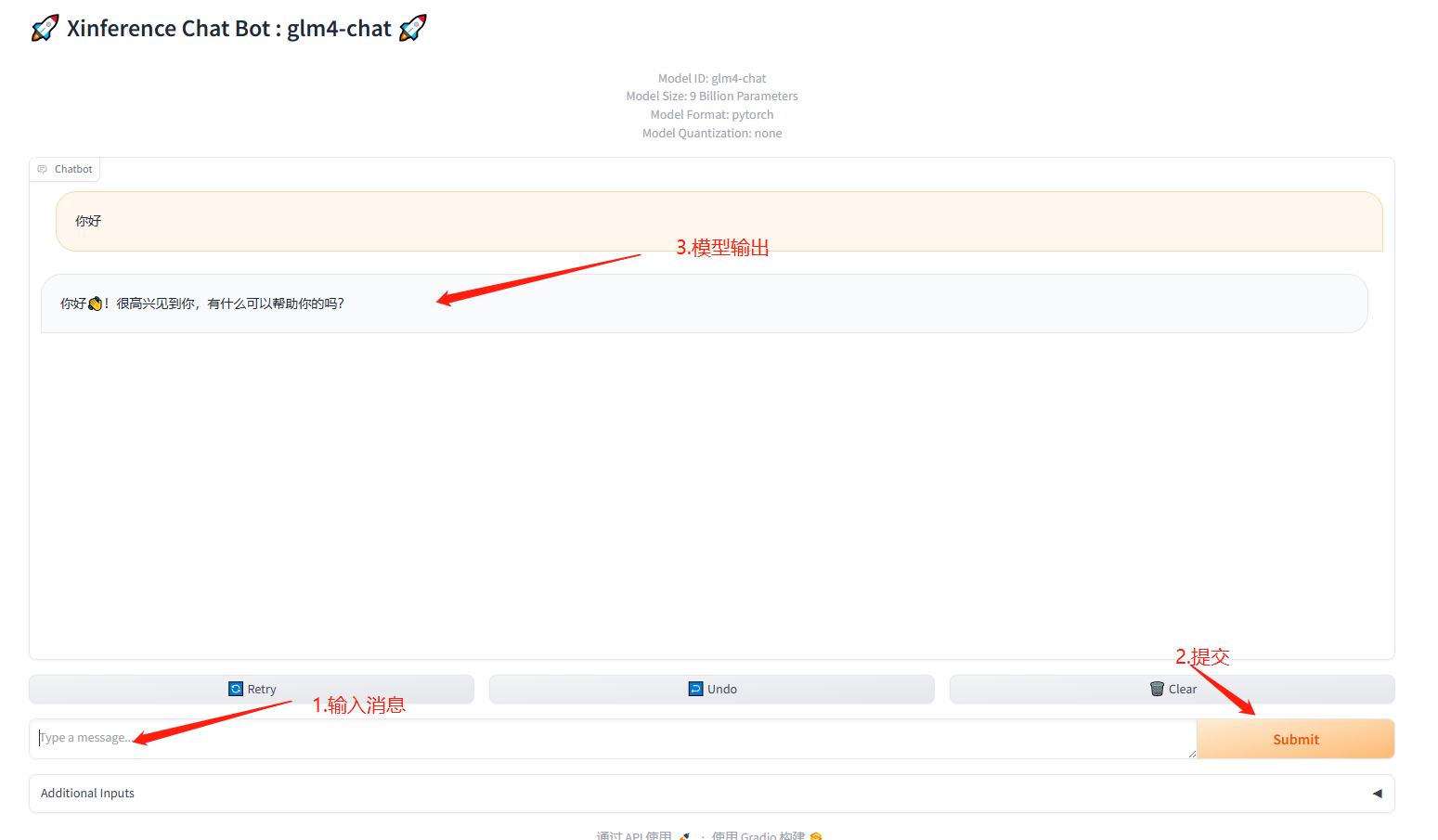




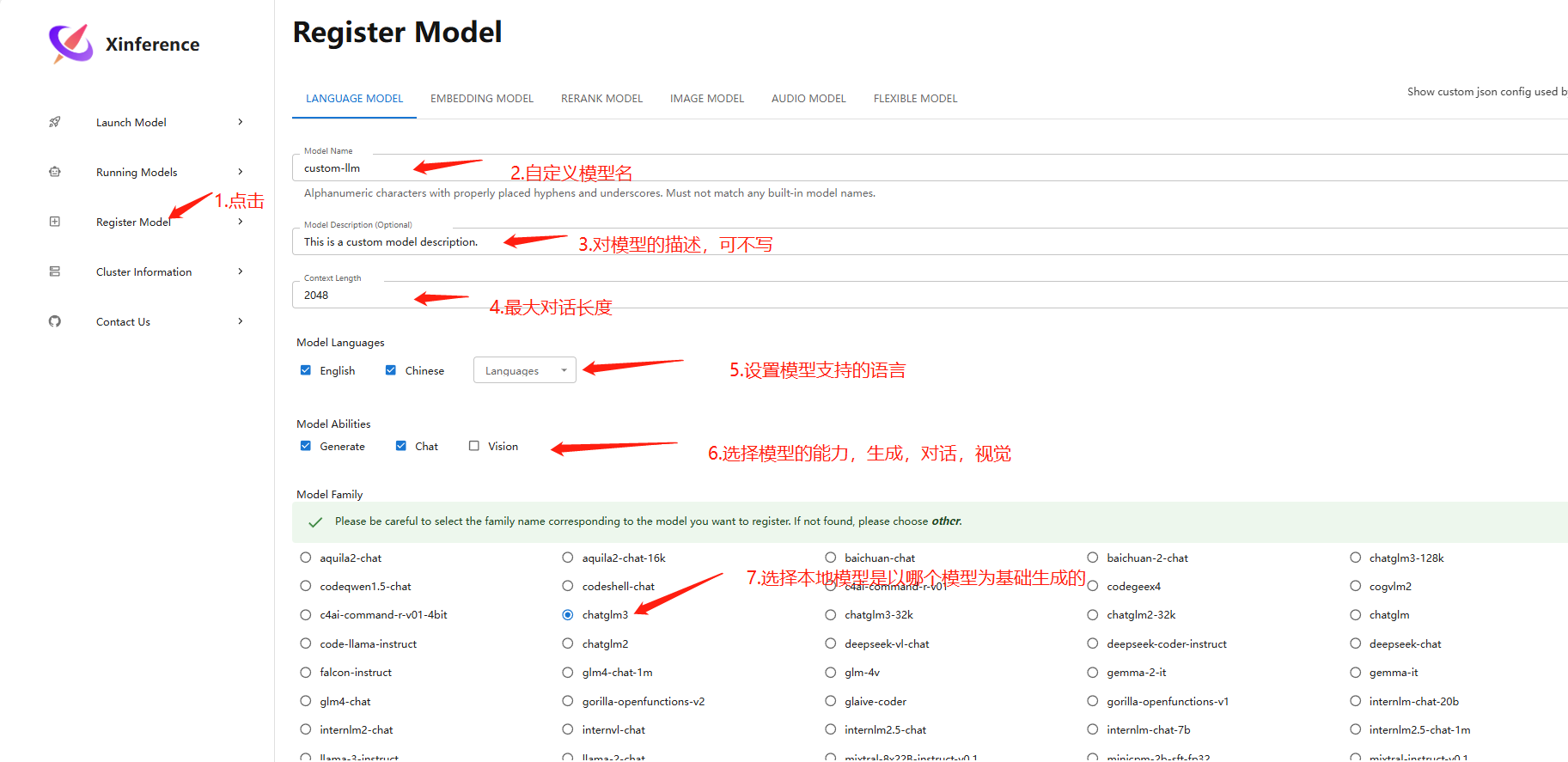


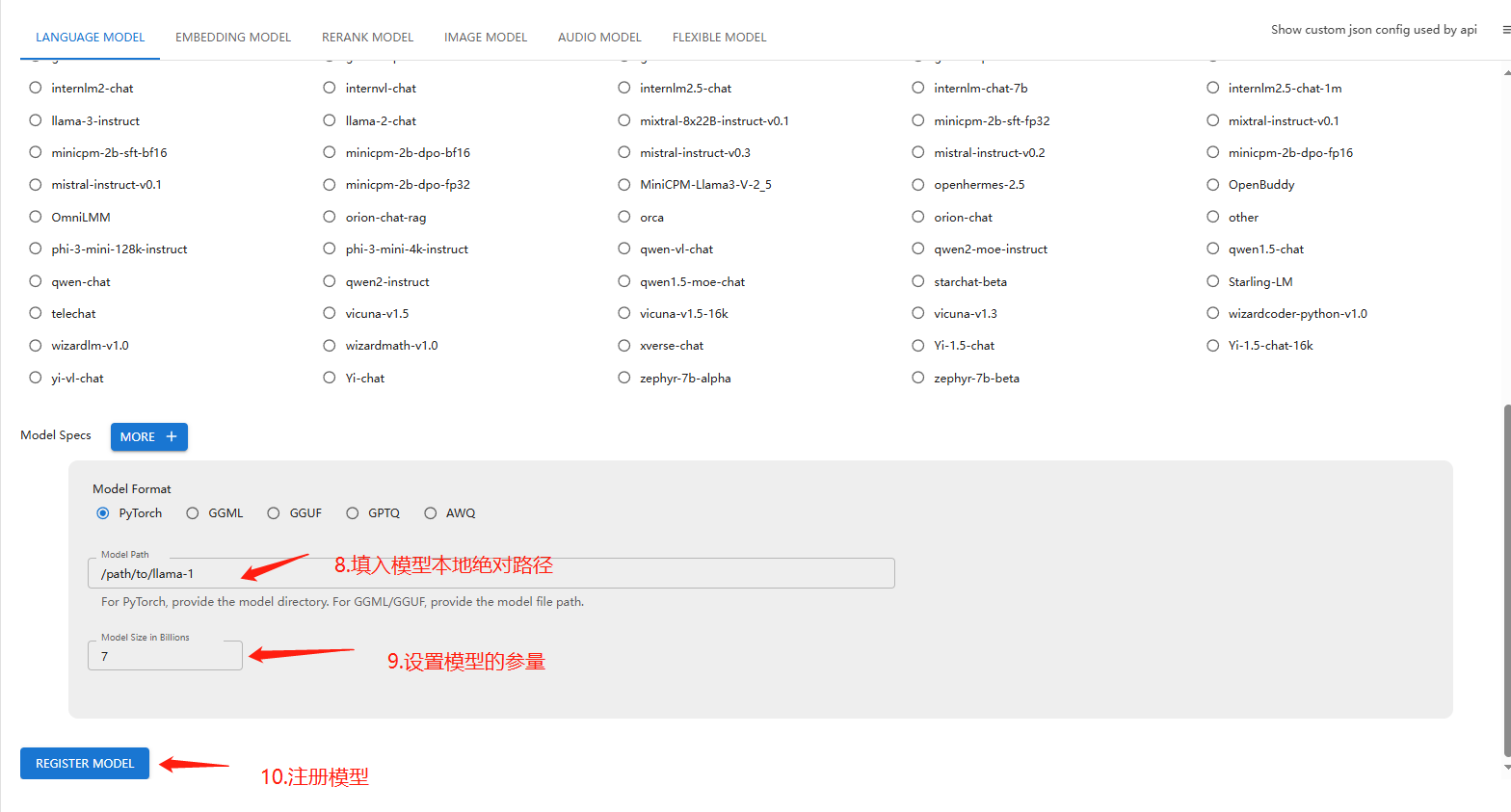
### 查看正在运行的模型





### 注册本地模型

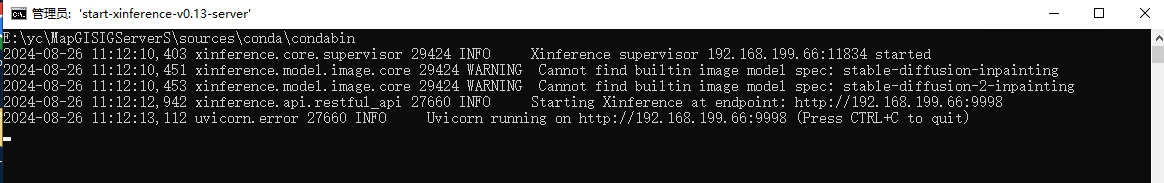




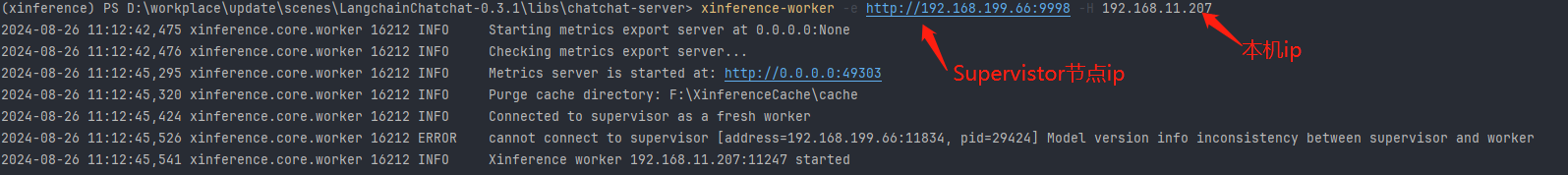
**集群部署 Xinference**

### 1.集群部署

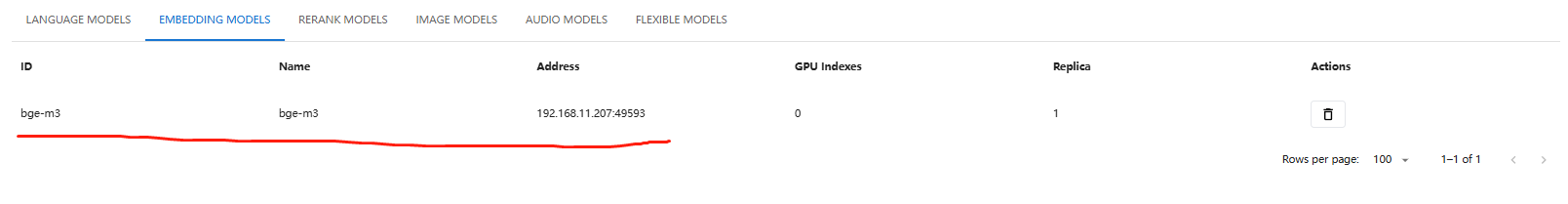
先在一台机器上启动Supervisor节点



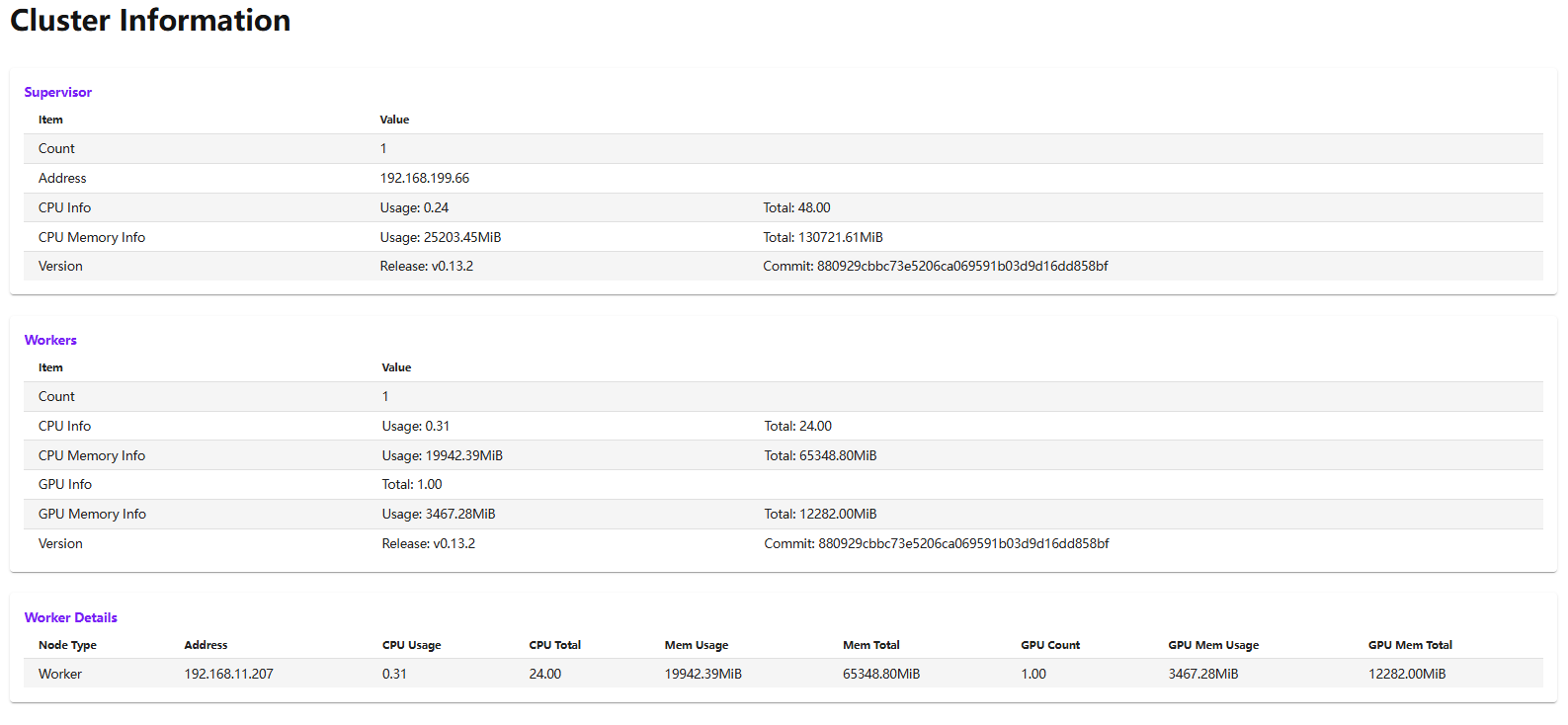
然后在另一台机器上启动Worker节点



模型在worker节点上成功启动

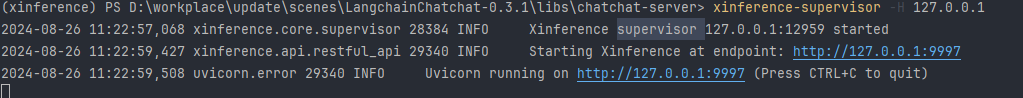


点击查看集群信息

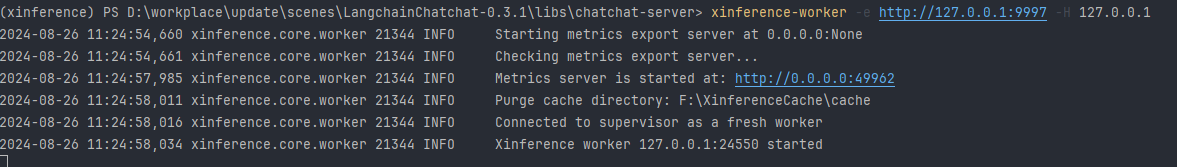


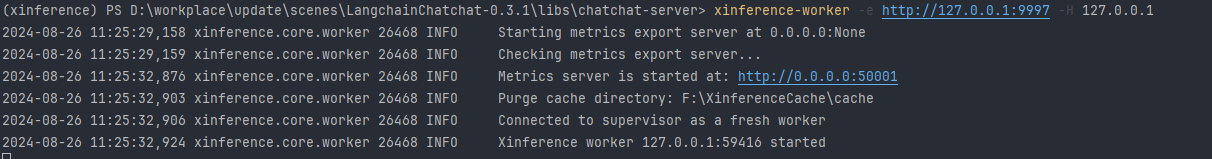
### 2.单机多端口伪集群 ，单机同一模型多实例服务

在本机启动supervisor 节点

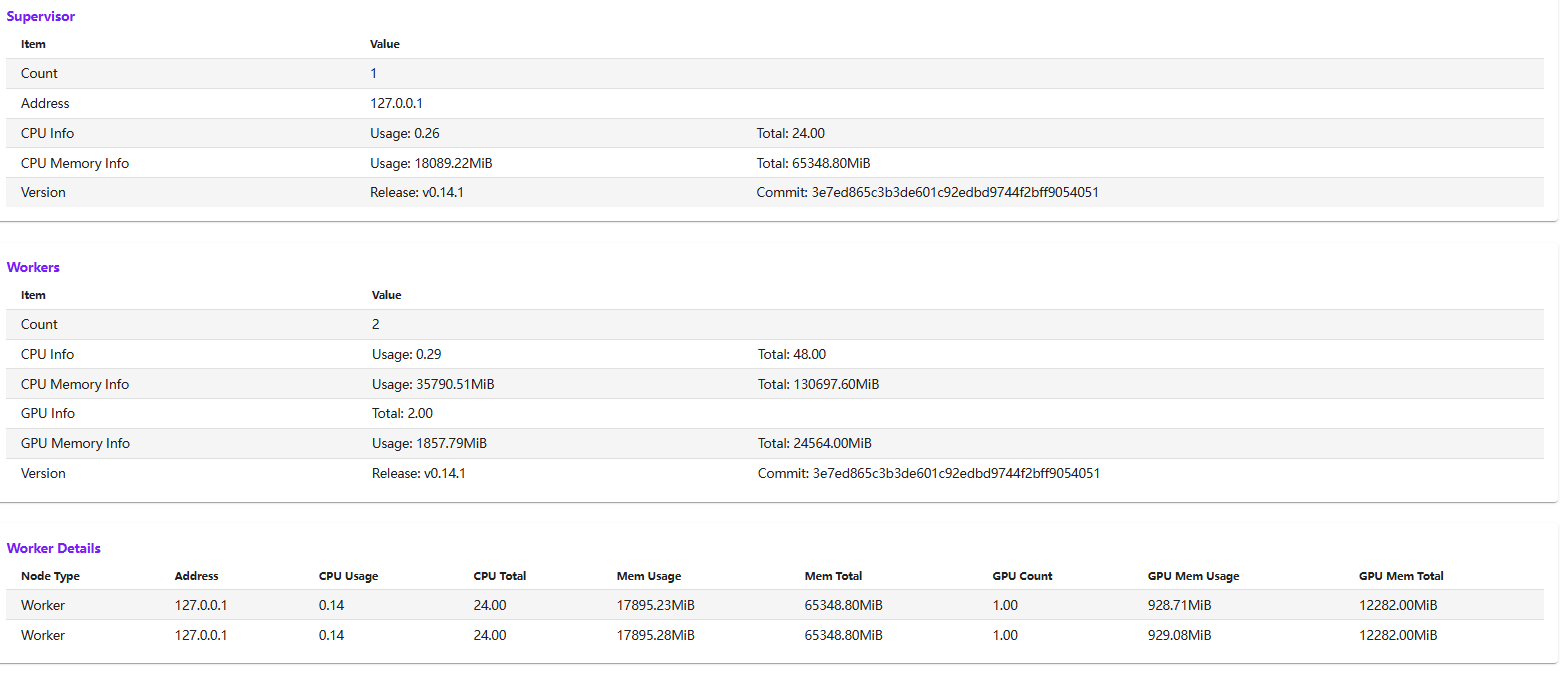


在本机启动多个worker节点

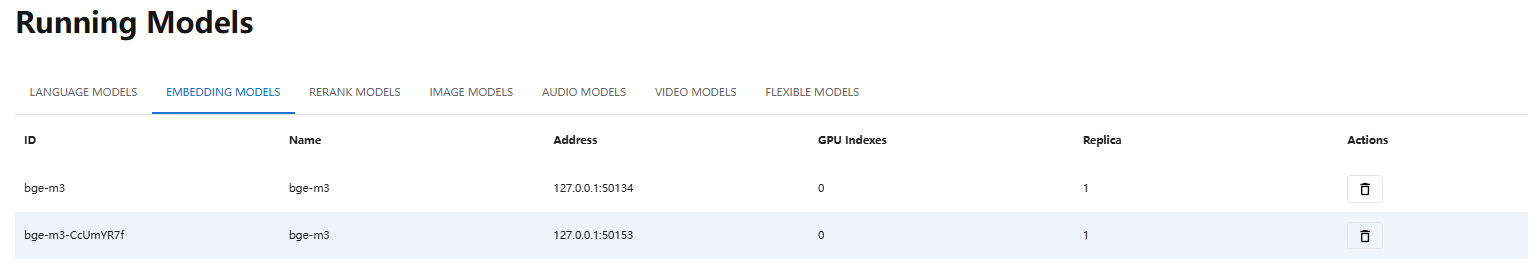




查看集群信息



启动模型



可以在两个不同的worker节点上运行同一模型

# 环境变量

**XINFERENCE\_ENDPOINT**

Xinference 的服务地址，用来与 Xinference 连接。默认地址是 [http://127.0.0.1:9997](http://127.0.0.1:9997/)，可以在日志中获得这个地址。

**XINFERENCE\_MODEL\_SRC**

配置模型下载仓库。默认下载源是 “huggingface”，也可以设置为 “modelscope” 作为下载源。

**XINFERENCE\_HOME**

Xinference 默认使用 <HOME>/.xinference 作为默认目录来存储模型以及日志等必要的文件。其中 <HOME> 是当前用户的主目录。可以通过配置这个环境变量来修改默认目录。

**XINFERENCE\_HEALTH\_CHECK\_ATTEMPTS**

Xinference 启动时健康检查的次数，如果超过这个次数还未成功，启动会报错，默认值为 3。

**XINFERENCE\_HEALTH\_CHECK\_INTERVAL**

Xinference 启动时健康检查的时间间隔，如果超过这个时间还未成功，启动会报错，默认值为 3。

**XINFERENCE\_DISABLE\_HEALTH\_CHECK**

在满足条件时，Xinference 会自动汇报worker健康状况，设置改环境变量为 1可以禁用健康检查。

**XINFERENCE\_DISABLE\_VLLM**

在满足条件时，Xinference 会自动使用 vLLM 作为推理引擎提供推理效率，设置改环境变量为 1可以禁用 vLLM。

**XINFERENCE\_DISABLE\_METRICS**

Xinference 会默认在 supervisor 和 worker 上启用 metrics exporter。设置环境变量为 1可以在 supervisor 上禁用 /metrics 端点，并在 worker 上禁用 HTTP 服务（仅提供 /metrics 端点）

**XINFERENCE\_AUDIO\_SPEECH\_DEFAULT\_STREAM**

Xinference audio Speech API will use non-stream by default. Setting this environment to 1 will use stream by default.