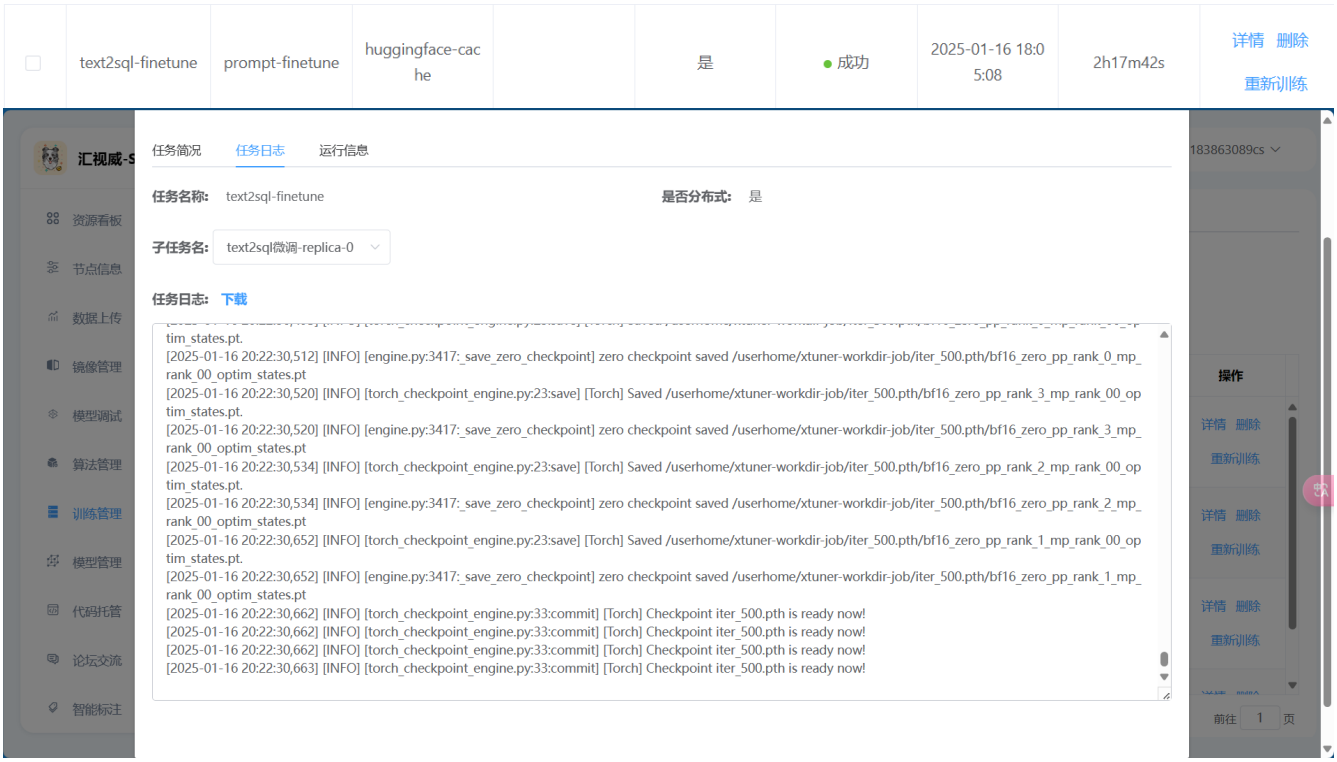
提示词工程-多机多卡微调

在汇视威平台使用 `https://huggingface.co/datasets/b-mc2/sql-create-context` 数据(实际使用16k条)微调code-llama-7b模型

使用【训练管理】启动训练

训练结果如下



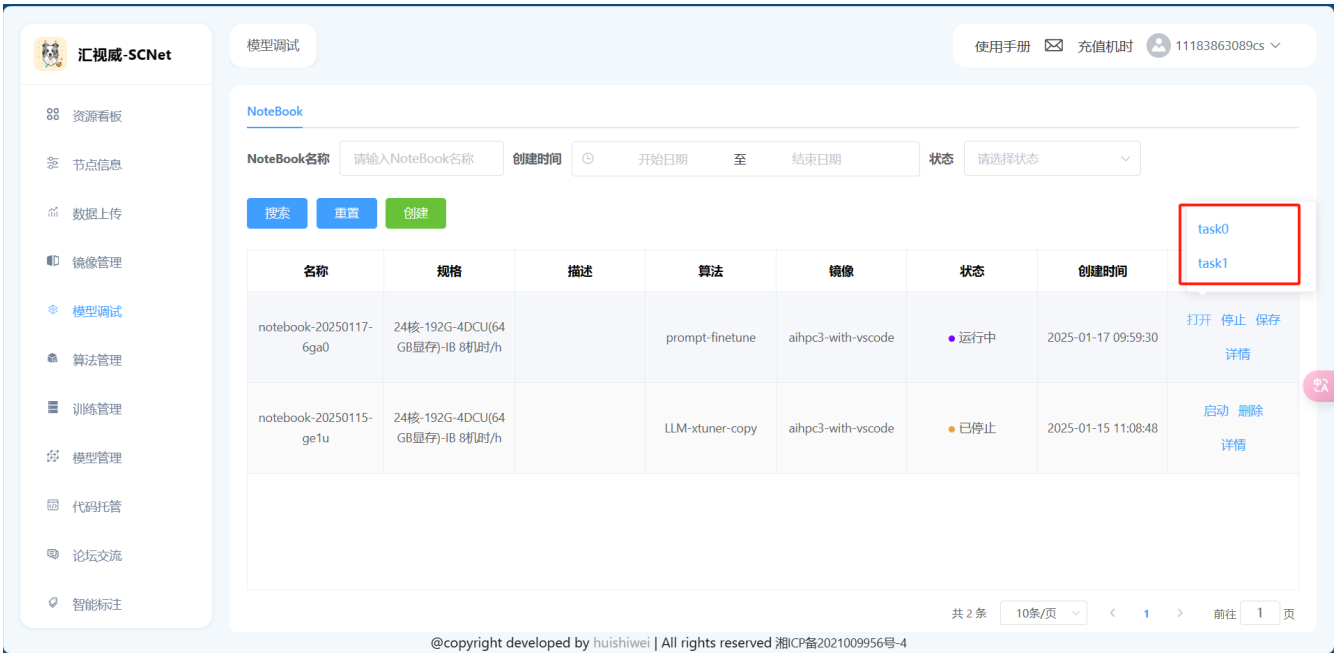
使用【模型调试】启动训练

开发环境准备

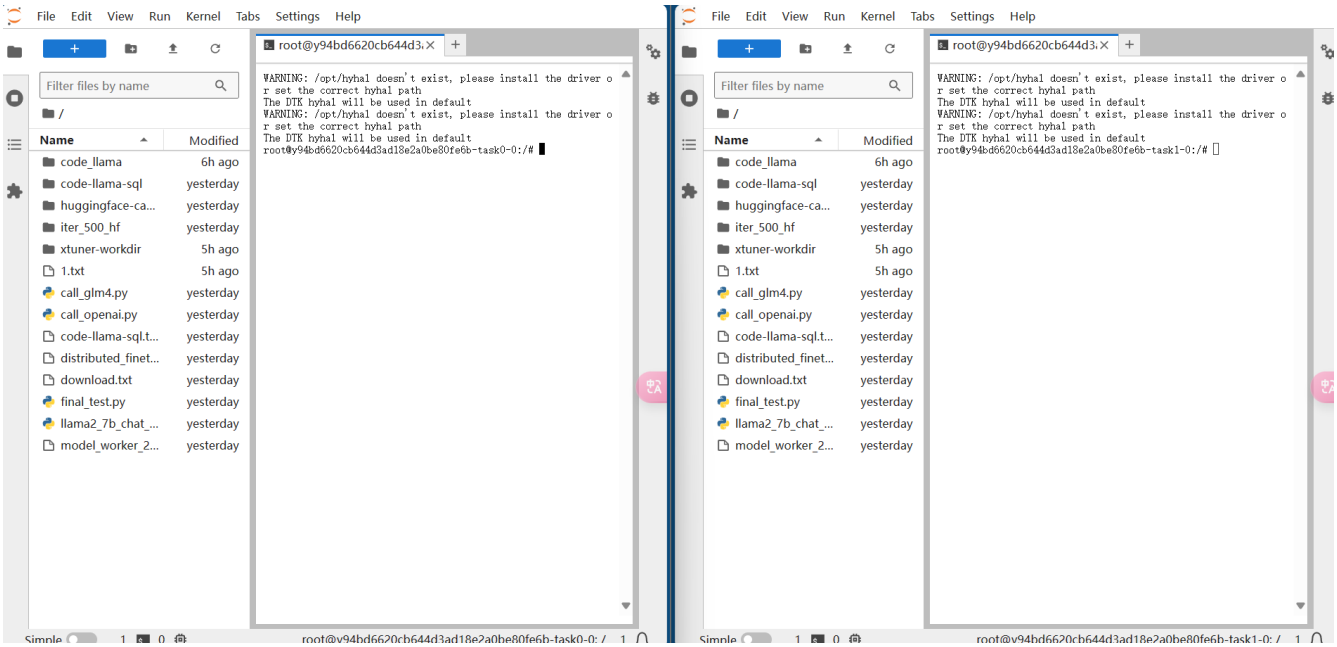
省略

环境变量配置

打开task0和task1



启动命令行



联网环境变量

```
export HF_HOME=/code/huggingface-cache/
export HF_ENDPOINT=https://hf-mirror.com
export http_proxy=http://10.10.9.50:3000
export https_proxy=http://10.10.9.50:3000
export no_proxy=localhost,127.0.0.1
```

IB网卡配置

```
export NCCL_DEBUG=INFO
export NCCL_IB_DISABLE=0
export NCCL_IB_HCA=mlx5
export NCCL_SOCKET_IFNAME=eth0
export GLOO_SOCKET_IFNAME=eth0
```

启动训练

以task0为master节点，获取master IP地址

```
ifconfig
```

```
root@o144dfb59b094ebea6b6f50e299911c3-task0-0:/code# ifconfig
eth0: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1480
        inet 10.244.132.114  netmask 255.255.255.255  broadcast 10.244.132.114
        ether aa:96:34:3a:ce:8c  txqueuelen 0  (Ethernet)
        RX packets 2157828  bytes 6876241593 (6.8 GB)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 1489692  bytes 239797031 (239.7 MB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0

lo: flags=73<UP,LOOPBACK,RUNNING>  mtu 65536
        inet 127.0.0.1  netmask 255.0.0.0
        loop  txqueuelen 1000  (Local Loopback)
        RX packets 307699  bytes 55600894 (55.6 MB)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 307699  bytes 55600894 (55.6 MB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0
```

以该IP地址作为ADDR的值，启动训练

```
# 在task0上运行该命令
NPROC_PER_NODE=4 NNODES=2 PORT=12345 ADDR=<master IP地址> NODE_RANK=0 xtuner train
llama2_7b_chat_qlora_sql_e3_copy.py --work-dir /code/xtuner-workdir --deepspeed
deepspeed_zero3_offload

# 在task1上运行该命令
NPROC_PER_NODE=4 NNODES=2 PORT=12345 ADDR=<master IP地址> NODE_RANK=1 xtuner train
llama2_7b_chat_qlora_sql_e3_copy.py --work-dir /code/xtuner-workdir --deepspeed
deepspeed_zero3_offload
```

训练后测试

转换模型权重文件；将微调训练的checkpoint转成hf格式的模型

如果是使用【模型调试】训练的模型，执行如下命令：

```
xtuner convert pth_to_hf /code/llama2_7b_chat_qlora_sql_e3_copy.py /code/xtuner-
workdir/iter_500.pth/ /code/iter_500_hf/
```

如果是使用【训练管理】训练的模型，执行如下命令：

```
xtuner convert pth_to_hf /code/llama2_7b_chat_qlora_sql_e3_copy.py /userhome/xtuner-workdir-job/iter_500.pth/ /code/iter_500_hf/
```

加载lora的HF格式模型进行测试，测试代码：

```
from transformers import AutoTokenizer, AutoModelForCausalLM
import torch

local_model_path = "/dataset/CodeLlama-7b-hf/"

tokenizer = AutoTokenizer.from_pretrained(local_model_path)

eval_prompt = """You are a powerful text-to-SQL model. Your job is to answer questions about a database. You are given a question and context regarding one or more tables.

You must output the SQL query that answers the question.
### Input:
Which Class has a Frequency MHz larger than 91.5, and a City of license of hyannis, nebraska?

### Context:
CREATE TABLE table_name_12 (class VARCHAR, frequency_mhz VARCHAR, city_of_license VARCHAR)

### Response:
"""
model_input = tokenizer(eval_prompt, return_tensors="pt").to("cuda")

base_model = AutoModelForCausalLM.from_pretrained(local_model_path, torch_dtype=torch.float16, device_map="cuda")  # don't quantize here

base_model.eval()
with torch.no_grad():
    print(tokenizer.decode(base_model.generate(**model_input, max_new_tokens=100)[0], skip_special_tokens=True))

print("=========下面是微调后的模型=========")

from peft import PeftModel
model = PeftModel.from_pretrained(base_model, "/code/iter_500_hf")

model.eval()
with torch.no_grad():
    print(tokenizer.decode(model.generate(**model_input, max_new_tokens=100)[0], skip_special_tokens=True))
```

执行测试：

```
python final_test.py
```

微调后模型下载并接入FastGPT

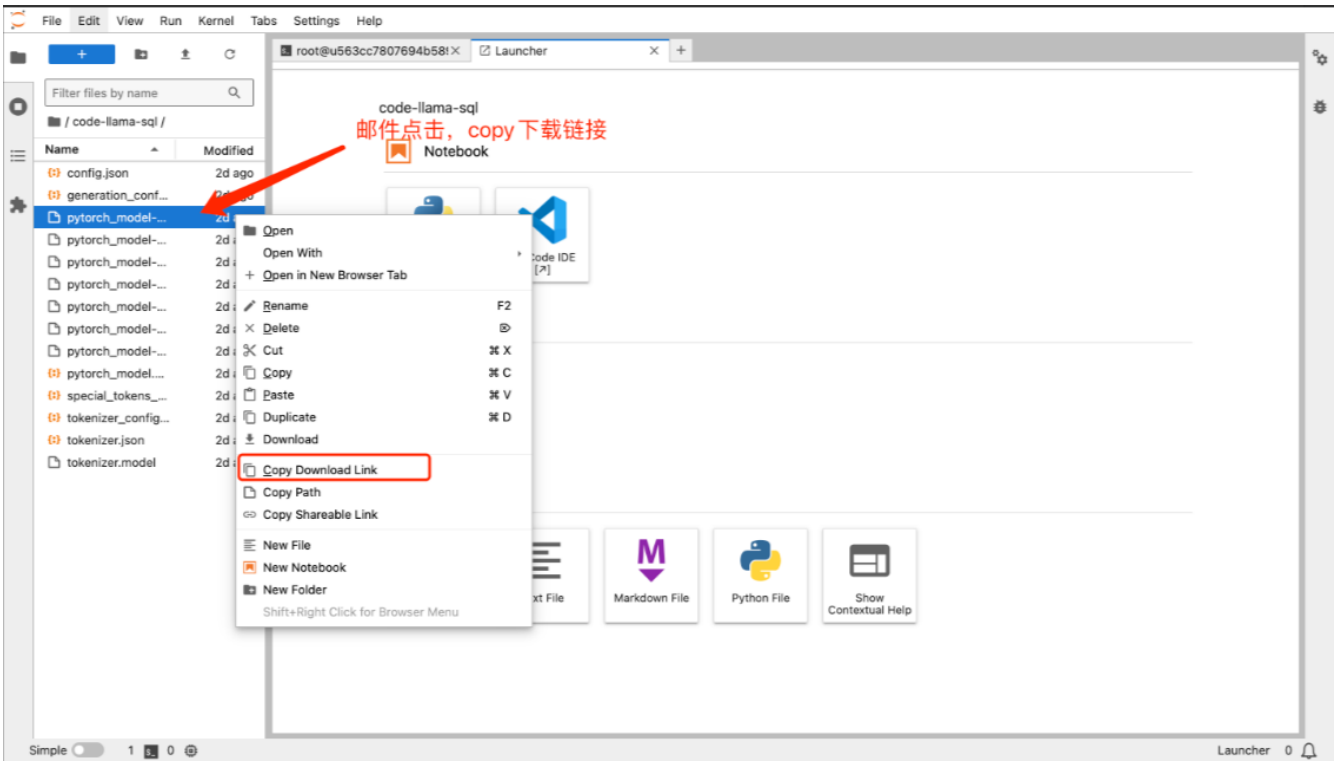
合并微调后的模型

使用xtuner将微调后的lora adapter模型和源模型合并，合并后得到一个新的模型

```
xtuner convert merge /dataset/CodeLlama-7b-hf /code/iter_500_hf/ /code/code-llama-sql --
device cpu
```

下载模型

在平台上获取文件的下载链接，接下来就可以使用wget或者直接使用浏览器下载



在AutoDL租用的机器上下载模型

在/root/autodl-tmp目录下创建code-llama目录并下载模型

```
wget http://hsw.csidc.cn/notebook_m02c05b673114d428a06d2c497467a7a_task0/files/code-llama-
sql.tar?_xsrf=2%7C4c72f6d2%7C40bb164eb50a9abe11c6a1c5396b8948%7C1737079208 -O root/autodl-
tmp/code-llama/code-llama.tar
```

使用xinference部署下载的模型

随便创建一个虚拟环境，下载xinference

在安装xinference时会弹出需要下载某个llama.cpp文件的要求

Llama.cpp地址：`https://github.com/abetlen/llama-cpp-python/releases`

在该地址找到所需要下载版本并上传到机器上

然后再安装xinference，不然会报错无法安装

```
pip install "xinference[all]"
```

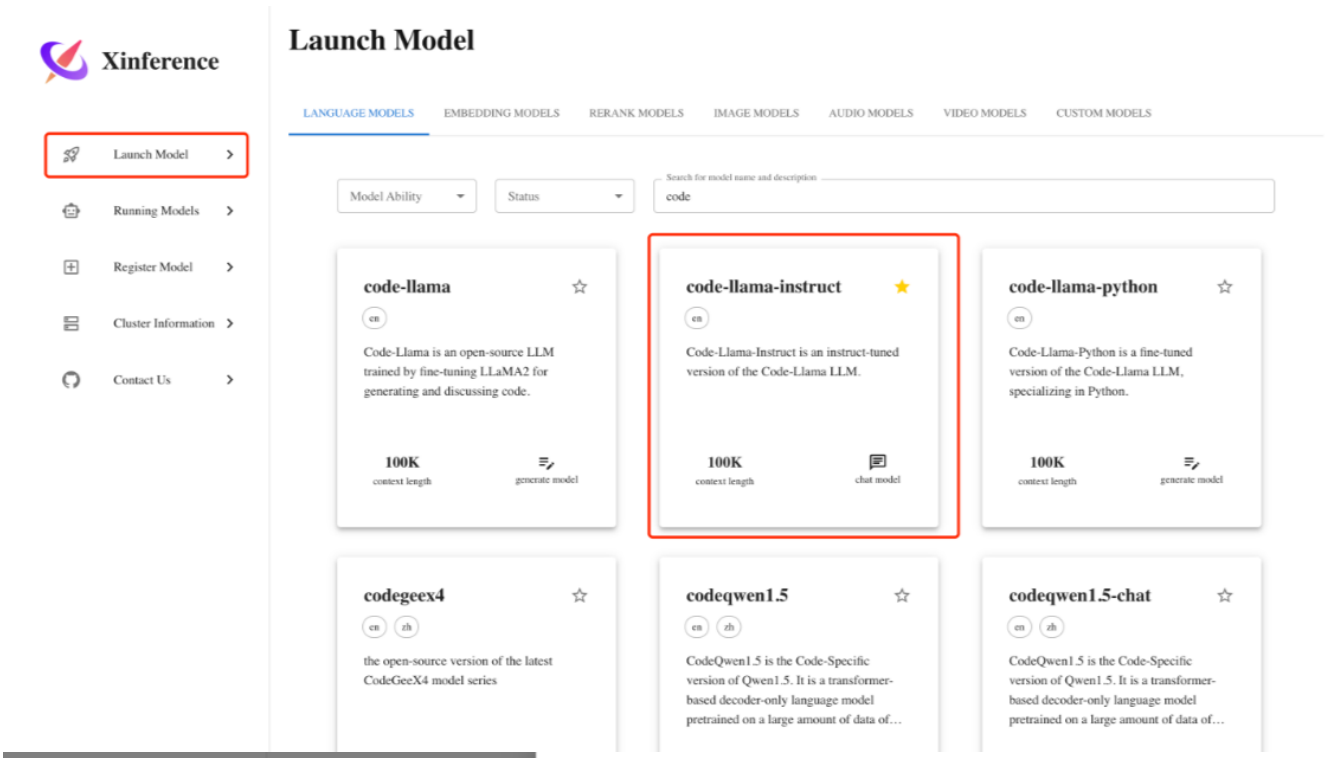
启动xinference

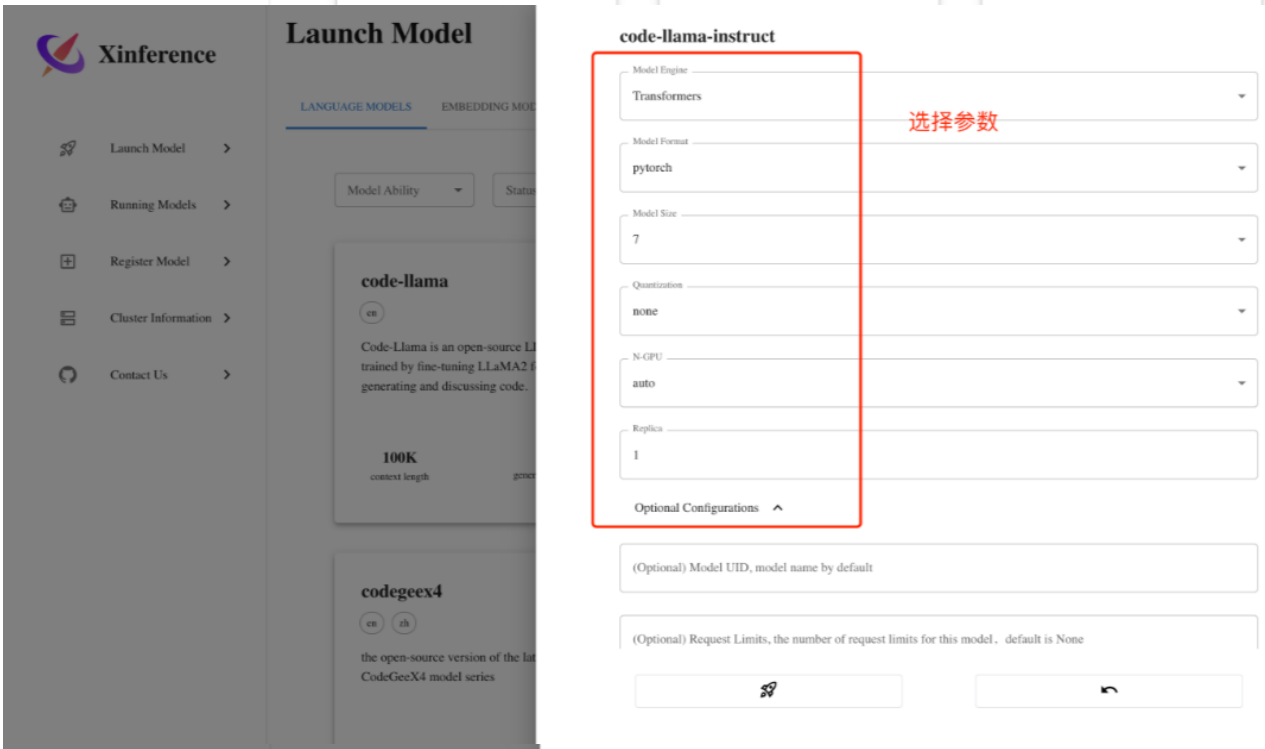
xinference-local --host 0.0.0.0 --port 9997 /root/autodl-tmp/code-llama/code/code-llama-sql

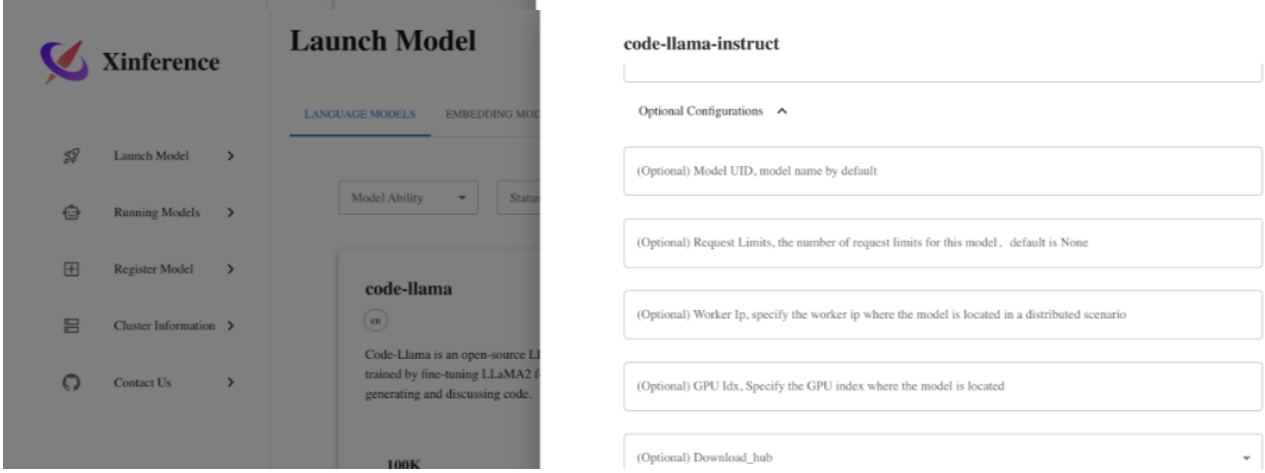
与自己的机器建立ssh隧道

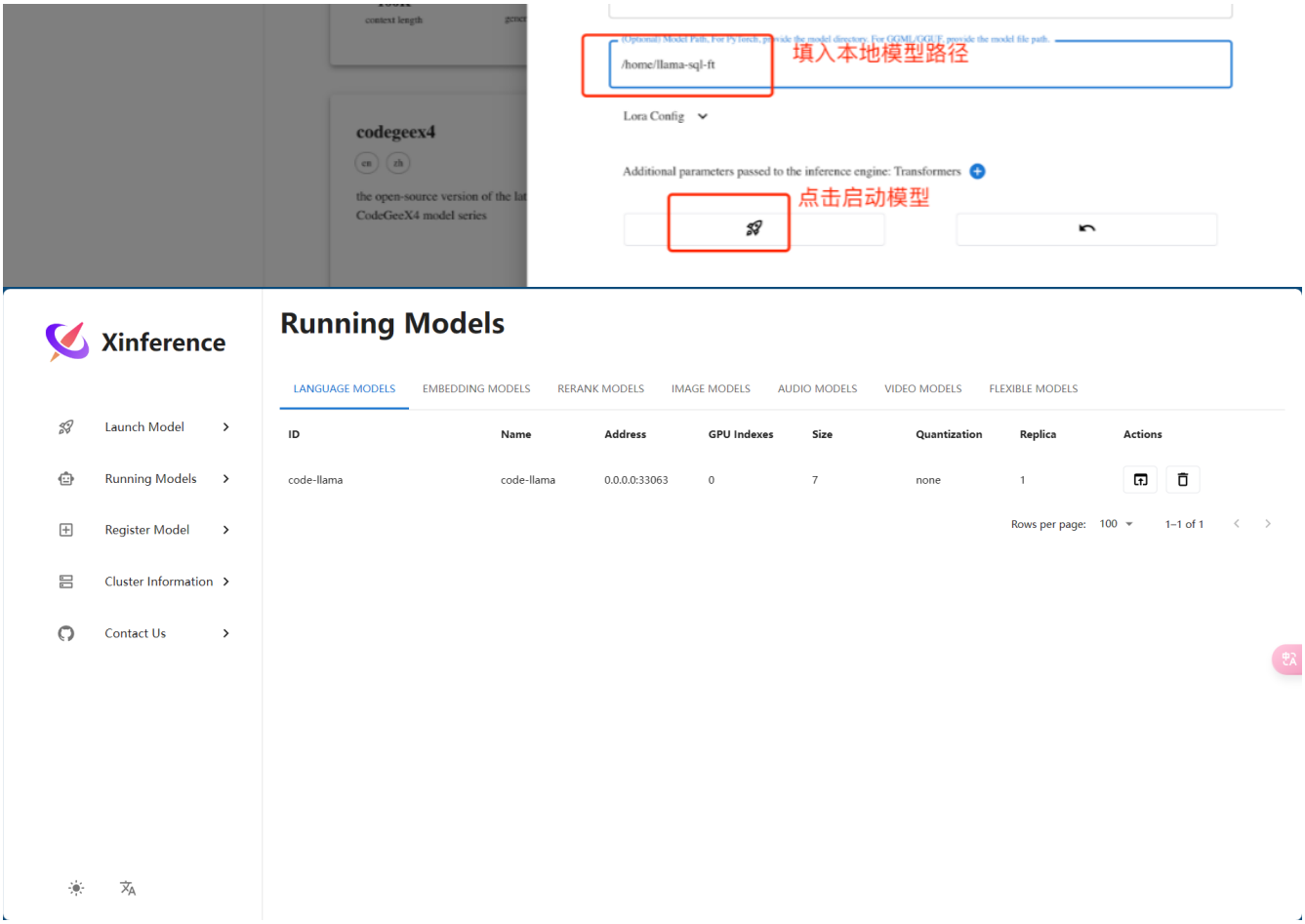
```
$ ssh -CNg -L 9997:127.0.0.1:9997 root@connect.nmb1.seetacloud.com -p 34790
The authenticity of host '[connect.nmb1.seetacloud.com]:34790 ([111.127.52.75]:3
4790)' can't be established.
ED25519 key fingerprint is SHA256:fXmxJm1D+u8+mkd+cTJvamLBrxcQyH7IpvK/lR3v5t8.
This key is not known by any other names.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Could not create directory '/c/Users/\316\260\301\246\313\274/.ssh' (No such fil
e or directory).
Failed to add the host to the list of known hosts (/c/Users/\316\260\301\246\313
\274/.ssh/known_hosts).
root@connect.nmb1.seetacloud.com's password:
```

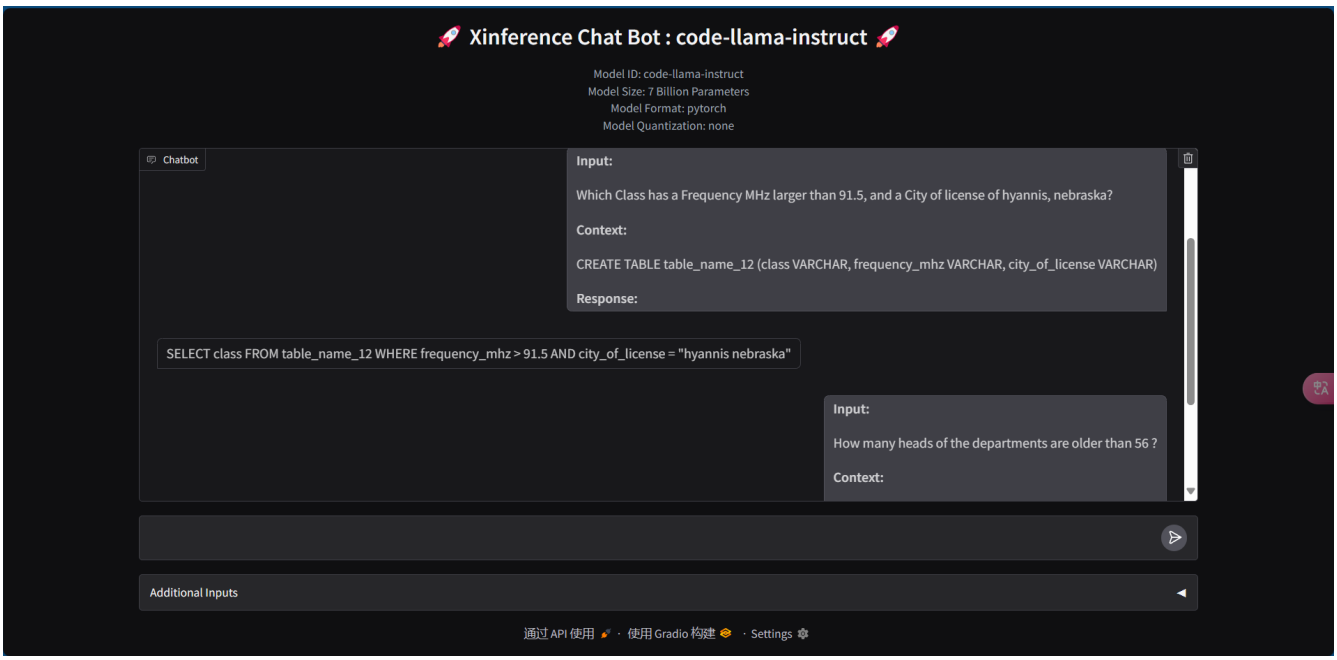
在本机上打开xinference部署模型


Xinference
Launch Model
Running Models
Register Model
Cluster Information
Contact Us
Launch Model
LANGUAGE MODELS
EMBEDDING MODELS
RERANK MODELS
IMAGE MODELS
AUDIO MODELS
VIDEO MODELS
CUSTOM MODELS
Model Ability
Status
Search for model name and description
code
code-llama
Code-Llama is an open-source LLM trained by fine-tuning LLaMA2 for generating and discussing code.
100K
context length
generate model
code-llama-instruct
Code-Llama-Instruct is an instruct-tuned version of the Code-Llama LLM.
100K
context length
chat model
code-llama-python
Code-Llama-Python is a fine-tuned version of the Code-Llama LLM, specializing in Python.
100K
context length
generate model
codegeex4
the open-source version of the latest CodeGeeX4 model series
codeqwen1.5
CodeQwen1.5 is the Code-Specific version of Qwen1.5. It is a transformer-based decoder-only language model pretrained on a large amount of data of…
codeqwen1.5-chat
CodeQwen1.5 is the Code-Specific version of Qwen1.5. It is a transformer-based decoder-only language model pretrained on a large amount of data of…


Xinference
Launch Model
Running Models
Register Model
Cluster Information
Contact Us
Launch Model
code-llama-instruct
选择参数
Model Engine
Transformers
Model Format
pytorch
Model Size
7
Quantization
none
N-GPU
auto
Replica
1
Optional Configurations
(Optional) Model UID, model name by default
(Optional) Request Limits, the number of request limits for this model，default is None


Xinference
Launch Model
Running Models
Register Model
Cluster Information
Contact Us
Launch Model
code-llama-instruct
Optional Configurations
(Optional) Model UID, model name by default
(Optional) Request Limits, the number of request limits for this model，default is None
(Optional) Worker Ip, specify the worker ip where the model is located in a distributed scenario
(Optional) GPU Idx, Specify the GPU index where the model is located
(Optional) Download_hub





1

启动FastGPT

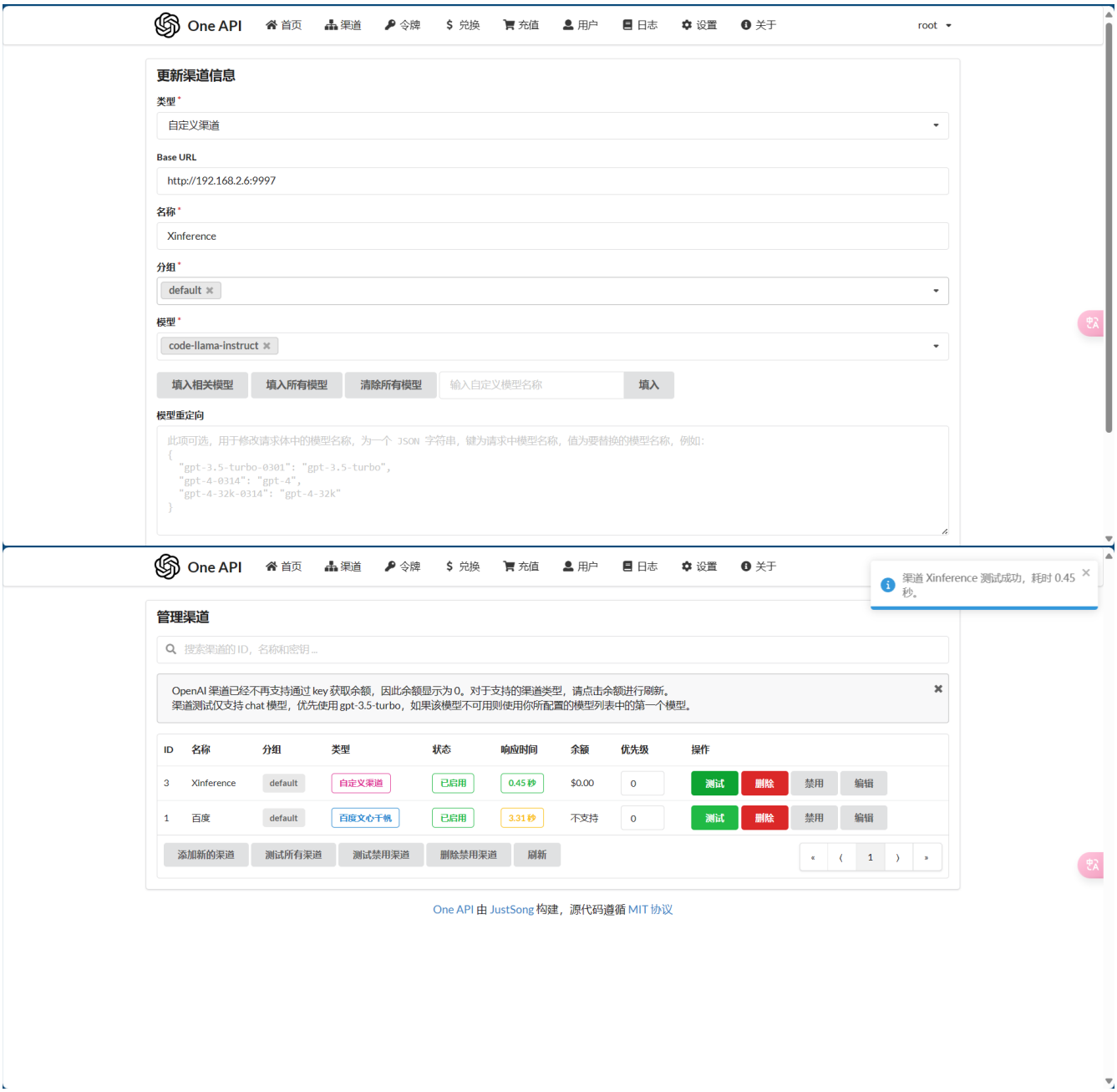
把模型配置写入FastGPT的config文件中

```
"llmModels": [
  {
    "usedInClassify": true,
    "usedInExtractFields": true,
    "usedInToolCall": true,
    "usedInQueryExtension": true,
    "toolChoice": true,
    "functionCall": false,
    "customCQPrompt": "",
    "customExtractPrompt": "",
    "defaultSystemChatPrompt": "",
    "defaultConfig": {}
  },
  {   "model": "code-llama-instruct",
  "name": "code-llama-instruct",
  "maxContext": 4000,
  "maxResponse": 4000,
  "quoteMaxToken": 8000,
  "maxTemperature": 1.2,
  "charsPointsPrice": 0,
  "censor": false,
  "vision": false,
  "datasetProcess": false,
  "usedInClassify": true,
  "usedInExtractFields": true,
  "usedInToolCall": true,
  "usedInQueryExtension": true,
  "toolChoice": false,
  "functionCall": false,
  "customCQPrompt": "",
  "customExtractPrompt": "",
  "defaultSystemChatPrompt": "",
  "defaultConfig": {}
  }
]
```

启动docker容器

配置OneAPI

URL需要写本机IP地址，写127.0.0.1会导致无法访问



访问FastGPT服务，创建对话机器人