

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

☐	text2sql-finetune	prompt-finetune	huggingface-cac he	是	● 成功	2025-01-16 18:0 5:08	2h17m42s	详情 删除 重新训练
--------------------------	-------------------	-----------------	-----------------------	---	------	-------------------------	----------	---

汇视威-S

资源看板

节点信息

数据上传

镜像管理

模型调试

算法管理

训练管理

模型管理

代码托管

论坛交流

智能标注

任务简况

任务日志

运行信息

任务名称: text2sql-finetune

是否分布式: 是

子任务名: text2sql微调-replica-0

任务日志: 下载

tim_states.pt.
[2025-01-16 20:22:30,512] [INFO] [engine.py:3417:_save_zero_checkpoint] zero checkpoint saved /userhome/xtuner-workdir-job/iter_500.pth/bf16_zero_pp_rank_0_mp_rank_00_optim_states.pt
[2025-01-16 20:22:30,520] [INFO] [torch_checkpoint_engine.py:23:save] [Torch] Saved /userhome/xtuner-workdir-job/iter_500.pth/bf16_zero_pp_rank_3_mp_rank_00_optim_states.pt.
[2025-01-16 20:22:30,520] [INFO] [engine.py:3417:_save_zero_checkpoint] zero checkpoint saved /userhome/xtuner-workdir-job/iter_500.pth/bf16_zero_pp_rank_3_mp_rank_00_optim_states.pt
[2025-01-16 20:22:30,534] [INFO] [torch_checkpoint_engine.py:23:save] [Torch] Saved /userhome/xtuner-workdir-job/iter_500.pth/bf16_zero_pp_rank_2_mp_rank_00_optim_states.pt.
[2025-01-16 20:22:30,534] [INFO] [engine.py:3417:_save_zero_checkpoint] zero checkpoint saved /userhome/xtuner-workdir-job/iter_500.pth/bf16_zero_pp_rank_2_mp_rank_00_optim_states.pt
[2025-01-16 20:22:30,652] [INFO] [torch_checkpoint_engine.py:23:save] [Torch] Saved /userhome/xtuner-workdir-job/iter_500.pth/bf16_zero_pp_rank_1_mp_rank_00_optim_states.pt.
[2025-01-16 20:22:30,652] [INFO] [engine.py:3417:_save_zero_checkpoint] zero checkpoint saved /userhome/xtuner-workdir-job/iter_500.pth/bf16_zero_pp_rank_1_mp_rank_00_optim_states.pt
[2025-01-16 20:22:30,662] [INFO] [torch_checkpoint_engine.py:33:commit] [Torch] Checkpoint iter_500.pth is ready now!
[2025-01-16 20:22:30,662] [INFO] [torch_checkpoint_engine.py:33:commit] [Torch] Checkpoint iter_500.pth is ready now!
[2025-01-16 20:22:30,662] [INFO] [torch_checkpoint_engine.py:33:commit] [Torch] Checkpoint iter_500.pth is ready now!
[2025-01-16 20:22:30,663] [INFO] [torch_checkpoint_engine.py:33:commit] [Torch] Checkpoint iter_500.pth is ready now!

操作

详情

删除

重新训练

详情

删除

重新训练

详情

删除

重新训练

前往

1

页

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

汇视威-SCNet

资源看板

节点信息

数据上传

镜像管理

模型调试

算法管理

训练管理

模型管理

代码托管

论坛交流

智能标注

模型调试

NoteBook

NoteBook名称 创建时间 至 状态

搜索

重置

创建

名称	规格	描述	算法	镜像	状态	创建时间	
notebook-20250117-6ga0	24核-192G-4DCU(64GB显存)-IB 8机时/h		prompt-finetune	aihpc3-with-vscode	运行中	2025-01-17 09:59:30	task0 task1 打开 停止 保存 详情
notebook-20250115-ge1u	24核-192G-4DCU(64GB显存)-IB 8机时/h		LLM-xtuner-copy	aihpc3-with-vscode	已停止	2025-01-15 11:08:48	启动 删除 详情

共 2 条 < 1 > 前往 1 页

@copyright developed by huishiwei | All rights reserved 湘CP备2021009956号-4

启动命令行

File Edit View Run Kernel Tabs Settings Help

Filter files by name

Name

Modified

code_llama 6h ago

code-llama-sql yesterday

huggingface-ca... yesterday

iter_500_hf yesterday

xtuner-workdir 5h ago

1.txt 5h ago

call_glm4.py yesterday

call_openai.py yesterday

code-llama-sql.t... yesterday

distributed_finet... yesterday

download.txt yesterday

final_test.py yesterday

llama2_7b_chat_... yesterday

model_worker_2... yesterday

root@y94bd6620cb644d3:~

WARNING: /opt/hyhal doesn't exist, please install the driver or set the correct hyhal path
The DTK hyhal will be used in default
WARNING: /opt/hyhal doesn't exist, please install the driver or set the correct hyhal path
The DTK hyhal will be used in default
root@y94bd6620cb644d3ad18e2a0be80fe6b-task0-0:~#

File Edit View Run Kernel Tabs Settings Help

Filter files by name

Name

Modified

code_llama 6h ago

code-llama-sql yesterday

huggingface-ca... yesterday

iter_500_hf yesterday

xtuner-workdir 5h ago

1.txt 5h ago

call_glm4.py yesterday

call_openai.py yesterday

code-llama-sql.t... yesterday

distributed_finet... yesterday

download.txt yesterday

final_test.py yesterday

llama2_7b_chat_... yesterday

model_worker_2... yesterday

root@y94bd6620cb644d3:~

WARNING: /opt/hyhal doesn't exist, please install the driver or set the correct hyhal path
The DTK hyhal will be used in default
WARNING: /opt/hyhal doesn't exist, please install the driver or set the correct hyhal path
The DTK hyhal will be used in default
root@y94bd6620cb644d3ad18e2a0be80fe6b-task1-0:~#

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@o144dfb59b094e6a6b6f50e299911c3-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_q1ora_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

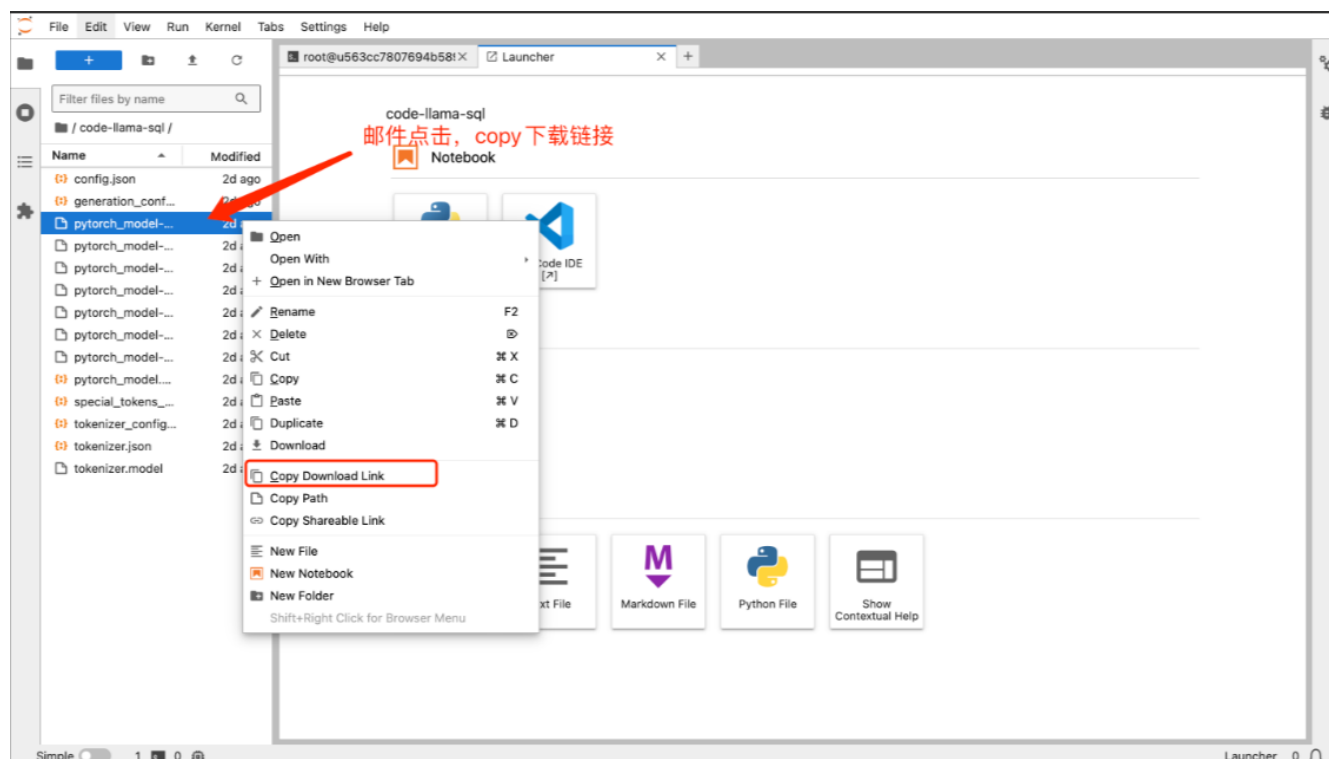
合并微调后的模型

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?_xsrft=2%7C4c72f6d2%7C40bb164eb50a9abe11c6a1c5396b8948%7C1737079208 -O root/autodl-tmp/code-llama/code-llama.tar
```

使用xinference部署下载模型

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]):34790' can't be established.
ED25519 key fingerprint is SHA256:fXmxJm1D+u8+mkd+cTJvamLBrxcQyH7IpvK/1R3v5t8.
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 file 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

-  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

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

cn zh

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...

cn zh

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...

cn zh


Xinference

Launch Model

LANGUAGE MODELS EMBEDDING MODELS

Model Ability
Status

code-llama
☆

Code-Llama is an open-source LLM trained by fine-tuning LLaMA2 for generating and discussing code.

100K
context length
generate model

codegeex4
☆

the open-source version of the latest CodeGeeX4 model series

cn zh

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

Model Engine
Transformers

Model Format
pytorch

Model Size
7

Quantization
none

N-GPU
auto

Replicas
1

Optional Configurations
^

选择参数

(Optional) Model UID, model name by default

(Optional) Request Limits, the number of request limits for this model, default is None




code-llama-instruct

Optional Configurations
^

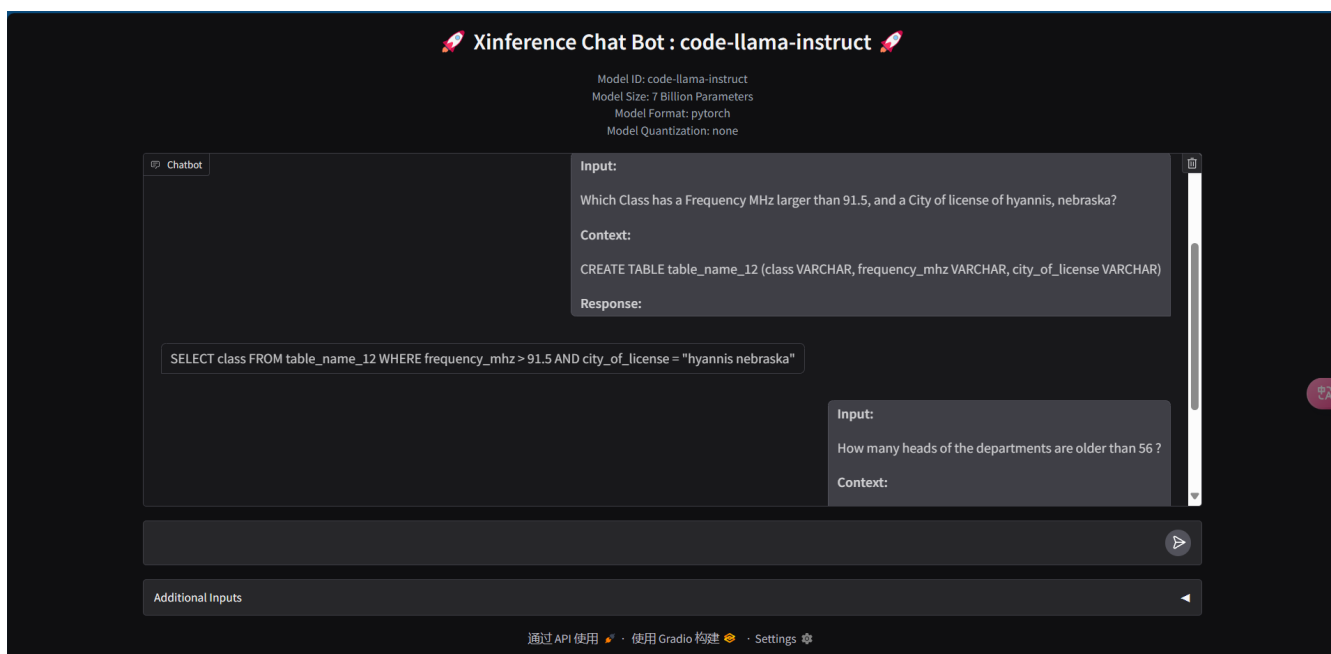
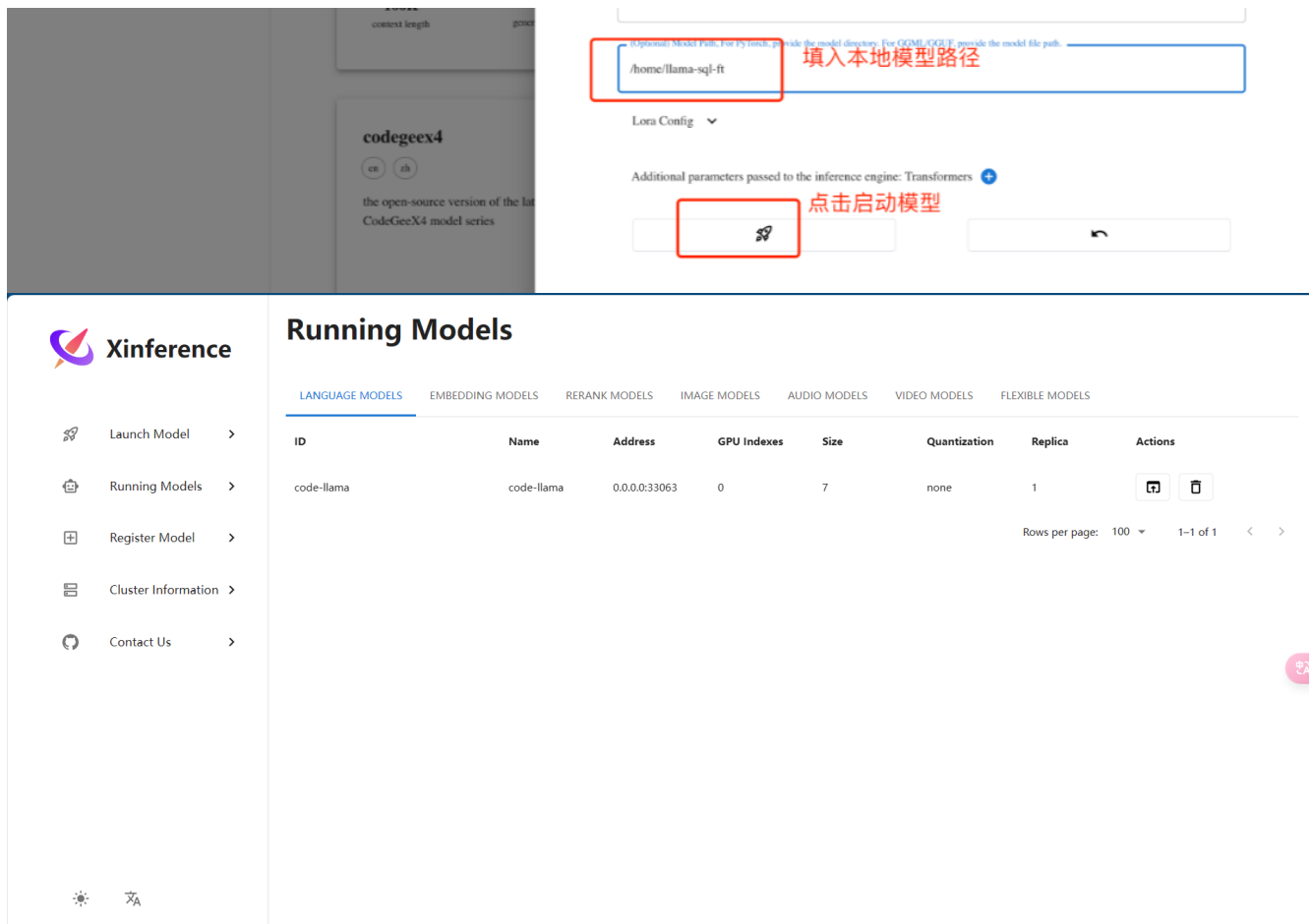
(Optional) Model UID, model name by default

(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

One API

首页

渠道

令牌

兑换

充值

用户

日志

设置

关于

root

更新渠道信息

类型

自定义渠道

Base URL

http://192.168.2.6:9997

名称

Xinference

分组

default

模型

code-llama-instruct

填入相关模型

填入所有模型

清除所有模型

输入自定义模型名称

填入

模型重定向

此项可选，用于修改请求体中的模型名称，为一个 JSON 字符串，键为请求中模型名称，值为要替换的模型名称，例如：
{
 "gpt-3.5-turbo-0301": "gpt-3.5-turbo",
 "gpt-4-0314": "gpt-4",
 "gpt-4-32k-0314": "gpt-4-32k"
}

One API

首页

渠道

令牌

兑换

充值

用户

日志

设置

关于

渠道 Xinference 测试成功，耗时 0.45 秒。

管理渠道

搜索渠道的 ID，名称和密钥 ...

OpenAI 渠道已经不再支持通过 key 获取余额，因此余额显示为 0。对于支持的渠道类型，请点击余额进行刷新。
渠道测试仅支持 chat 模型，优先使用 gpt-3.5-turbo，如果该模型不可用则使用你所配置的模型列表中的第一个模型。

ID	名称	分组	类型	状态	响应时间	余额	优先级	操作
3	Xinference	default	自定义渠道	已启用	0.45 秒	\$0.00	0	测试 删除 禁用 编辑
1	百度	default	百度文心千帆	已启用	3.31 秒	不支持	0	测试 删除 禁用 编辑

添加新的渠道

测试所有渠道

测试禁用渠道

删除禁用渠道

刷新

«

«

1

»

»

One API 由 JustSong 构建，源代码遵循 MIT 协议

访问FastGPT服务，创建对话机器人

全部应用 / 1

聊天

工作台

知识库

工具箱

账号

1

快来给应用一个介绍~

对话

设置

...

AI 配置

AI 模型code-llama-instruct

提示词

输入"/"可选择变量

模型固定的引导词，通过调整该内容，可以引导模型聊天方向。该内容会被固定在上下文的开头。可通过输入 / 插入选择变量
如果关联了知识库，你还可以通过适当的描述，来引导模型何时去调用知识库搜索。例如：
你是电影《星际穿越》的助手，当用户询问与《星际穿越》相关的内容时，请搜索知识库并结合搜索结果进行回答。

关联知识库

+ 选择

参数

工具调用

+ 选择

文件上传

关闭

应用配置

发布渠道

对话日志

未保存

保存

调试预览

你好

SELECT TOP 5 COUNT(*) FROM gadwp_options WHERE option_name = "gadwp_version"

2条上下文1.46s查看详情

输入问题，发送 [Enter]/换行 [Ctrl/Alt/Shift] + Enter