轻松玩转书生·浦语大模型趣味 Demo

本节课可以让同学们实践 4 个主要内容, 分别是:

- 部署 InternLM2-Chat-1.8B 模型进行智能对话
- 部署实战营优秀作品 八戒-Chat-1.8B 模型
- 通过 InternLM2-Chat-7B 运行 Lagent 智能体 Demo
- 实践部署 浦语·灵笔2 模型

使用 InternLM-Chat-1.8B 模型生成 300 字的小故事

详细步骤参考配套教学文档,以下为大致总结:

1. 环境准备:

- 在 InternStudio 平台选择 A100(1/4) 的配置,使用 Cuda11.7-conda 镜像。
- o 打开开发机,进入终端,切换到 bash 环境。
- 使用提供的脚本克隆并激活 pytorch 2.0.1 的 conda 环境, 然后安装所需依赖。

2. 模型下载:

○ 复制已准备好的 InternLM 模型到指定目录,或使用 modelscope 中的 snapshot_download 函数下载模型 (**推荐**, 跑满带宽),或在 huggingface、 OpenXLab 等处下载。

3. 代码准备:

- o 在 /root 路径下新建 code 目录, clone 指定版本的代码。
- 在 /root/code/InternLM 目录下新建 cli_demo.py 文件,使用 transformers 和 torch 库运行大模型。

4. 终端运行:

○ 运行 cli_demo.pv 文件,即可体验 InternLM-Chat-1.8B 模型的对话能力。

使用的prompt: 帮我生成一个300字的小故事,主角是打工人叫石向前,内容是他通过不断努力升职加薪、当上总经理、出任CEO、迎娶白富美、走上人生巅峰的励志故事

cli demo

• 结果图:

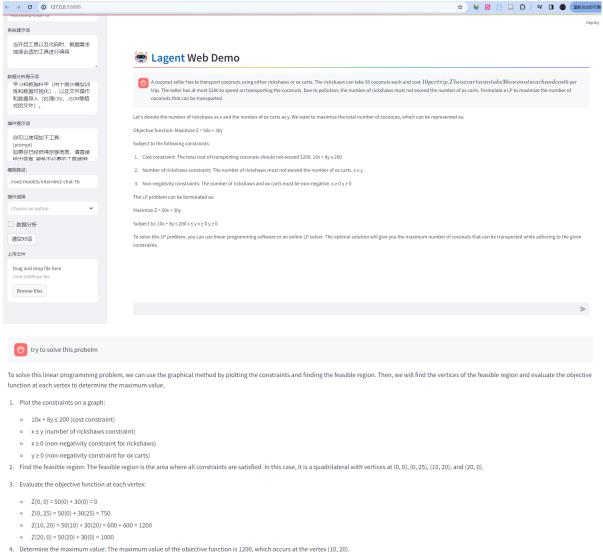
2/2 [00:27<00:00, 13.66s/it]

User >>> 報我生成一个300字的小故事,主角是打工人叫石向前,内容是他通过不断努力升职加薪、当上总经理、出任CEO、迎娶白富美、走上人生颜峭的励志故事在一个繁华的城市里,有一个名叫石向前的打工人。他来自一个着通的家庭,为了实现自己的梦想,他不能地努力工作,不断学习,最终获得了升职加薪的机会。 石向前通过自己的努力,成为了公司的一名主管。他带领者团队,不断创新,不断提高工作效率。在他的带领下,公司业绩不断提升,员工们的收入也水涨船高。 石向前并没有满足于自己的成就,他继续努力,通过不断学习,他成为了公司的高管。他不仅带领者团队,更是在公司内部村立了良好的口碑。 石向前终于实现了自己的梦想,他迎娶了一位白富美,并成为了公司的CEO。他带领者团队,带领者公司不断前行,成为了这个城市里最优秀的企业之一。 石向前的本自诉我们,只要我们不断努力,不断学习和进步,我们就能实现自己的梦想,走上人生颜峭。 User >>>■

word统计是319个字

使用 Lagent 运行 InternLM2-Chat-7B 模型为内核的智能体

和上面一样使用streamlit部署本地端口转发,问一个简单的线性规划问题



4. Determine the maximum value: The maximum value of the objective function is 1200, which occurs at the vertex [10, 20).

Therefore, the maximum number of coconuts that can be transported is 1200. This can be achieved by using 10 rickshaws and 20 ox carts.

有一个地方有个小错误, 总体思路不错

下载 InternLM2-Chat-7B 的 config.json 文件到本 地

```
>>> import os
>>> from huggingface_hub import hf_hub_download # Load model directly
config.json: 100%| 817/817 [00:00<?, ?B/s] C:\Users\Venido\miniconda3\envs\LLMs\lib\site-packages\huggingface_hub\file_download.p y:147: UserWarning: `huggingface_hub` cache-system uses symlinks by default to efficie
ntly store duplicated files but your machine does not support them in C:\Users\Venido\
.cache\huggingface\hub. Caching files will still work but in a degraded version that m
ight require more space on your disk. This warning can be disabled by setting the 'HF_
HUB_DISABLE_SYMLINKS_WARNING' environment variable. For more details, see https://hugg
ingface.co/docs/huggingface_hub/how-to-cache#limitations.
To support symlinks on Windows, you either need to activate Developer Mode or to run P
ython as an administrator. In order to see activate developer mode, see this article:
https://docs.microsoft.com/en-us/windows/apps/get-started/enable-your-device-for-devel
opment
  warnings.warn(message)
'C:\\Users\\Venido\\.cache\\huggingface\\hub\\models--internlm--internlm2-7b\\snapshot
s\\ea4312ff4a5175723a17fbbbd74fa27514dd87b5\\config.json'
>>>
```