通过语言对齐将大语言模型英语能力外推到非英语语言

1. 配置环境

1.1 按照README.md安装库存在的问题

> conda env create -f environment.yml

1. 会长期卡在 Installing pip dependencies：

   尝试对environment.yml文件进行以下修改，添加镜像源即可：
   将channels改为（注意要把default去掉）：

```
channels:
  - conda-forge
  - https://mirrors.tuna.tsinghua.edu.cn/anaconda/pkgs/main
  - https://mirrors.tuna.tsinghua.edu.cn/anaconda/pkgs/free
  - https://mirrors.tuna.tsinghua.edu.cn/anaconda/pkgs/r
  - https://mirrors.tuna.tsinghua.edu.cn/anaconda/pkgs/pro
  - https://mirrors.tuna.tsinghua.edu.cn/anaconda/pkgs/msys2
```

   并在pip的依赖包里添加上镜像源（加上最后一行）。

```
  - pip:
    - addict==2.4.0
    - anyio==3.3.0
        .......
    - websocket-client==1.1.0
    - widgetsnbextension==3.5.1
    - sapien==1.1.1
    - -i https://pypi.tuna.tsinghua.edu.cn/simple
```

2. 找不到包满足bleurt==0.0.2

   操作：删除bleurt并自行安装（去除版本限制也会报错，找不到对应的包）

   方法：参照google-research/bleurt: BLEURT is a metric for Natural Language Generation based on transfer learning. (github.com)手动安装

```
pip install --upgrade pip  # ensures that pip is current
git clone https://github.com/google-research/bleurt.git
cd bleurt
pip install .
```

3. 安装tensorrt-libs==8.6.1报错

   原因：删除tensorrt-libs==8.6.1并自行安装（去除版本限制也会报错，子进程报错）

   方法：再次使用 pip install tensorrt-libs==8.6.1会发现已经安装了

4. 存在包版本错误冲突问题：

> The conflict is caused by:
>
> The user requested typing-extensions==4.7.1
>
> altair 5.0.1 depends on typing-extensions>=4.0.1; python_version < "3.11"
>
> fastapi 0.101.0 depends on typing-extensions>=4.5.0
>
> gradio 3.39.0 depends on typing-extensions~=4.0
>
> gradio-client 0.3.0 depends on typing-extensions~=4.0
>
> huggingface-hub 0.16.4 depends on typing-extensions>=3.7.4.3
>
> lightning-utilities 0.8.0 depends on typing-extensions
>
> pydantic 2.1.1 depends on typing-extensions>=4.6.1
>
> pydantic-core 2.4.0 depends on typing-extensions!=4.7.0 and >=4.6.0
>
> pyre-extensions 0.0.29 depends on typing-extensions
>
> pytorch-lightning 1.9.5 depends on typing-extensions>=4.0.0
>
> tensorflow 2.13.0 depends on typing-extensions<4.6.0 and >=3.6.6
>
> To fix this you could try to:
>
> 1. loosen the range of package versions you've specified
> 2. remove package versions to allow pip attempt to solve the dependency conflict
>
> Pip subprocess error:
>
> ERROR: Cannot install -r /home/djh/code/xllm/condaenv.dxptyxf0.requirements.txt (line 121), -r /home/djh/code/xllm/condaenv.dxptyxf0.requirements.txt (line 28), -r /home/djh/code/xllm/condaenv.dxptyxf0.requirements.txt (line 41), -r /home/djh/code/xllm/condaenv.dxptyxf0.requirements.txt (line 42), -r

> /home/djh/code/xllm/condaenv.dxptyxf0.requirements.txt (line 48), -r /home/djh/code/xllm/condaenv.dxptyxf0.requirements.txt (line 56), -r /home/djh/code/xllm/condaenv.dxptyxf0.requirements.txt (line 6), -r /home/djh/code/xllm/condaenv.dxptyxf0.requirements.txt (line 90), -r /home/djh/code/xllm/condaenv.dxptyxf0.requirements.txt (line 91), -r /home/djh/code/xllm/condaenv.dxptyxf0.requirements.txt (line 95), -r /home/djh/code/xllm/condaenv.dxptyxf0.requirements.txt (line 97) and typing-extensions==4.7.1 because these package versions have conflicting dependencies.
>
> ERROR: ResolutionImpossible: for help visit https://pip.pypa.io/en/latest/topics/dependency-resolution/#dealing-with-dependency-conflicts
>
> failed
>
> CondaEnvException: Pip failed

这个冲突是由于以下原因引起的：

- 用户请求了 `typing-extensions==4.7.1`
- `altair 5.0.1` 依赖于 `typing-extensions>=4.0.1; python_version <`
- `fastapi 0.101.0` 依赖于 `typing-extensions>=4.5.0`
- `gradio 3.39.0` 依赖于 `typing-extensions~=4.0`
- `gradio-client 0.3.0` 依赖于 `typing-extensions~=4.0`
- `huggingface-hub 0.16.4` 依赖于 `typing-extensions>=3.7.4.3`
- `lightning-utilities 0.8.0` 依赖于 `typing-extensions`
- `pydantic 2.1.1` 依赖于 `typing-extensions>=4.6.1`
- `pydantic-core 2.4.0` 依赖于 `typing-extensions!=4.7.0 and >=4.6.0`
- `pyre-extensions 0.0.29` 依赖于 `typing-extensions`
- `pytorch-lightning 1.9.5` 依赖于 `typing-extensions>=4.0.0`
- `tensorflow 2.13.0` 依赖于 `typing-extensions<4.6.0 and >=3.6.6`

为了解决这个问题，您可以尝试以下方法：

1. 放宽您指定的软件包版本范围。
2. 删除软件包版本，以便允许 pip 尝试解决依赖冲突。

- 首先尝试去掉 `tensorflow` 的包版本限制

conda env update -f environment.yml

然后会报类似的错误，依次取消upbabel-comet==2.0.1的限制、tensorboard==2.13.0 typing-extensions==4.7.1 keras==2.13.1 wrapt==1.15.0 google-auth-oauthlib==1.0.0 tensorboard-data-server==0.7.1 google-auth==2.23.0

报错没有尽头

另一种方式：

原因：考虑到typing-extensions(==4.7.1) 但多个其他包依赖不同的typing-extensions版本

操作：openai==0.27.7需要自行安装（具体内部原因不明）

1.2 修改environment.yml后继续安装存在的问题

1. 去掉了pip后面所有包的版本号，同时根据requirements.txt的要求保留了

```
numpy
rouge_score
fire
openai
transformers>=4.28.1
torch
sentencepiece
tokenizers>=0.13.3
wandb
```

2. 需要和本地cuda环境匹配的pytorch

```
conda install pytorch==2.0.1 torchvision==0.15.2 torchaudio==2.0.2 pytorch-cuda=11.7 -c pytorch -c nvidia
```

3. 需要重新安装utils
4. 需要重新安装openai==0.27.7
5. 需要再次重新更新包

```
conda env update -f environment.yml
```

6. 重复2-4
7. 删除bleurt并自行安装

操作：删除bleurt并自行安装（去除版本限制也会报错，找不到对应的包）

方法：参照google-research/bleurt: BLEURT is a metric for Natural Language Generation based on transfer learning. (github.com)手动安装

```
pip install --upgrade pip  # ensures that pip is current
git clone https://github.com/google-research/bleurt.git
cd bleurt
pip install .
```

8. 删除tensorrt-libs并自行安装

原因：删除tensorrt-libs==8.6.1并自行安装（去除版本限制也会报错，子进程报错）

方法：再次使用 pip install tensorrt-libs==8.6.1会发现已经安装了

运行 `bash script/train.sh llama-7b-hf alpaca_en+alpaca_zh+translation_ncwm_en-zh` 中：

> WARNING:root:Formatting inputs... 格式化输入...
>
> WARNING:root:Tokenizing inputs... This may take some time... 分词输入... 这可能需要一些时间...

报错：

> /home/djh/miniconda3/envs/xllm2/lib/python3.10/site-packages/torch/distributed/fsdp/_init_utils.py:295: UserWarning: FSDP is switching to use `NO_SHARD` instead of ShardingStrategy.FULL_SHARD since the world size is 1.
>
> warnings.warn(
>
> Traceback (most recent call last):
>
> File "/home/djh/code/xllm/train.py", line 326, in <module>
>
> train()
>
> File "/home/djh/code/xllm/train.py", line 318, in train
>
> trainer.train()
>
> File "/home/djh/miniconda3/envs/xllm2/lib/python3.10/site-packages/transformers/trainer.py", line 1664, in train
>
> return inner_training_loop(
>
> File "/home/djh/miniconda3/envs/xllm2/lib/python3.10/site-packages/transformers/trainer.py", line 1759, in _inner_training_loop
>
> model = self._wrap_model(self.model_wrapped)
>
> File "/home/djh/miniconda3/envs/xllm2/lib/python3.10/site-packages/transformers/trainer.py", line 1490, in _wrap_model
>
> self.model = model = FSDP(

File "/home/djh/miniconda3/envs/xllm2/lib/python3.10/site-packages/torch/distributed/fsdp/fully_sharded_data_parallel.py", line 408, in __init__

_init_param_handle_from_module(

File "/home/djh/miniconda3/envs/xllm2/lib/python3.10/site-packages/torch/distributed/fsdp/_init_utils.py", line 415, in _init_param_handle_from_module

_move_module_to_device(

File "/home/djh/miniconda3/envs/xllm2/lib/python3.10/site-packages/torch/distributed/fsdp/_init_utils.py", line 802, in _move_module_to_device

module = module.to(device_from_device_id)

File "/home/djh/miniconda3/envs/xllm2/lib/python3.10/site-packages/transformers/modeling_utils.py", line 1886, in to

return super().to(*args, **kwargs)

File "/home/djh/miniconda3/envs/xllm2/lib/python3.10/site-packages/torch/nn/modules/module.py", line 1145, in to

return self._apply(convert)

File "/home/djh/miniconda3/envs/xllm2/lib/python3.10/site-packages/torch/nn/modules/module.py", line 797, in _apply

module._apply(fn)

File "/home/djh/miniconda3/envs/xllm2/lib/python3.10/site-packages/torch/nn/modules/module.py", line 797, in _apply

module._apply(fn)

File "/home/djh/miniconda3/envs/xllm2/lib/python3.10/site-packages/torch/nn/modules/module.py", line 820, in _apply

param_applied = fn(param)

File "/home/djh/miniconda3/envs/xllm2/lib/python3.10/site-packages/torch/nn/modules/module.py", line 1143, in convert

return t.to(device, dtype if t.is_floating_point() or t.is_complex() else None, non_blocking)

RuntimeError: CUDA error: device kernel image is invalid

CUDA kernel errors might be asynchronously reported at some other API call, so the stacktrace below might be incorrect.

For debugging consider passing CUDA_LAUNCH_BLOCKING=1.

Compile with `TORCH_USE_CUDA_DSA` to enable device-side assertions.

```
ERROR:torch.distributed.elastic.multiprocessing.api:failed (exitcode: 1) local_rank: 0 (pid: 308066) of binary: /home/djh/miniconda3/envs/xllm2/bin/python
Traceback (most recent call last):
  File "/home/djh/miniconda3/envs/xllm2/bin/torchrun", line 33, in <module>
    sys.exit(load_entry_point('torch==2.0.1', 'console_scripts', 'torchrun')())
  File "/home/djh/miniconda3/envs/xllm2/lib/python3.10/site-packages/torch/distributed/elastic/multiprocessing/errors/__init__.py", line 346, in wrapper
    return f(*args, **kwargs)
  File "/home/djh/miniconda3/envs/xllm2/lib/python3.10/site-packages/torch/distributed/run.py", line 794, in main
    run(args)
  File "/home/djh/miniconda3/envs/xllm2/lib/python3.10/site-packages/torch/distributed/run.py", line 785, in run
    elastic_launch(
  File "/home/djh/miniconda3/envs/xllm2/lib/python3.10/site-packages/torch/distributed/launcher/api.py", line 134, in __call__
    return launch_agent(self._config, self._entrypoint, list(args))
  File "/home/djh/miniconda3/envs/xllm2/lib/python3.10/site-packages/torch/distributed/launcher/api.py", line 250, in launch_agent
    raise ChildFailedError(
torch.distributed.elastic.multiprocessing.errors.ChildFailedError:
============================================================
/home/djh/code/xllm/train.py FAILED
------------------------------------------------------------
Failures:
  <NO_OTHER_FAILURES>
------------------------------------------------------------
Root Cause (first observed failure):
[0]:
  time      : 2024-03-15_11:59:58
  host      : djh-PowerEdge-T640
```

> rank : 0 (local_rank: 0)
>
> exitcode : 1 (pid: 308066)
>
> error_file: <N/A>
>
> traceback : To enable traceback see: https://pytorch.org/docs/stable/elastic/errors.html
>
> ==========

CUDA错误 - 设备内核映像无效:

- 这通常意味着PyTorch试图在不支持的CUDA版本上运行操作，或者CUDA设备与当前的PyTorch或CUDA版本不兼容。确保您的CUDA版本与安装的PyTorch版本兼容。
- 考虑不改变cuda版本的情况下，能否找到适应的pytorch版本
- 可以参考 硬件驱动有关问题
  - conda install python=3.10.12
    conda install pytorch=2.0.1 torchvision torchaudio pytorch-cuda=11.7 -c pytorch -c nvidia
- 目前python为3.10.11，cuda为11.7
  - pytorch历史版本参照：https://pytorch.org/get-started/previous-versions/
  - 存在报错

```
# CUDA 11.7
conda install pytorch==2.0.1 torchvision==0.15.2 torchaudio==2.0.2 pytorch-cuda=11.7 -c pytorch -c nvidia
```

  - 存在报错

```
# CUDA 11.7
conda install pytorch==2.0.0 torchvision==0.15.0 torchaudio==2.0.0 pytorch-cuda=11.7 -c pytorch -c nvidia
```

  - 存在报错

```
# CUDA 11.7
conda install pytorch==1.13.1 torchvision==0.14.1 torchaudio==0.13.1 pytorch-cuda=11.7 -c pytorch -c nvidia
```

  - 核心错误还是：cuda/torch/nvidia硬件版本过低造成的

1.3 自行手动从前向后安装

1. 考虑到硬件要求：硬件驱动有关问题，参照其中内容完成python和pytorch的安装
    a. FSDP要求pytorch必须>=2.1.0:
2. conda env update -f environment.yml（无版本号模式，只保留requirements.中的号）
3. Failed to initialize NVML: Driver/library version mismatch-CSDN博客
    a. 更新环境中nvidia驱动，直到能够找到cuda为止
4. conda env update -f environment.yml（无版本号模式，只保留requirements.中的号）

参考版本：

1. pip install openai==0.27.7
2. pip install transformers==4.29.0
3. pip install datasets==2.12.0
4. pip install openai==0.27.7
5. pip install accelerate==0.19.0
6. pip install sentencepiece==0.1.99
7. pip install -r requirements.txt
8. conda env update -f environment.yml
9. pip install
10. altair==5.0.1
11. fastapi==0.101.0
12. gradio==3.39.0
13. gradio-client==0.3.0
14. huggingface-hub==0.16.4
15. lightning-utilities==0.8.0
16. pydantic==2.1.1
17. pydantic-core==2.4.0
18. pyre-extensions==0.0.29
19. pytorch-lightning==1.9.5
20. tensorflow==2.13.0

2. 运行训练脚本

2.1 Wandb相关报错

> Exception in thread IntMsgThr: Traceback (most recent call last): File "/home/djh/miniconda3/envs/xllm3/lib/python3.10/threading.py", line 1016, in _bootstrap_inner self.run() File "/home/djh/miniconda3/envs/xllm3/lib/python3.10/threading.py", line 953, in run self._target(*self._args, **self._kwargs) File "/home/djh/miniconda3/envs/xllm3/lib/python3.10/site-packages/wandb/sdk/wandb_run.py", line 300, in check_internal_messages self._loop_check_status(File "/home/djh/miniconda3/envs/xllm3/lib/python3.10/site-packages/wandb/sdk/wandb_run.py", line 224, in _loop_check_status local_handle = request() File "/home/djh/miniconda3/envs/xllm3/lib/python3.10/site-packages/wandb/sdk/interface/interface.py", line 803, in deliver_internal_messages return self._deliver_internal_messages(internal_message) File "/home/djh/miniconda3/envs/xllm3/lib/python3.10/site-packages/wandb/sdk/interface/interface_shared.py", line 506, in _deliver_internal_messages return self._deliver_record(record) File "/home/djh/miniconda3/envs/xllm3/lib/python3.10/site-packages/wandb/sdk/interface/interface_shared.py", line 449, in _deliver_record handle = mailbox._deliver_record(record, interface=self) File "/home/djh/miniconda3/envs/xllm3/lib/python3.10/site-packages/wandb/sdk/lib/mailbox.py", line 455, in _deliver_record interface._publish(record) File "/home/djh/miniconda3/envs/xllm3/lib/python3.10/site-packages/wandb/sdk/interface/interface_sock.py", line 51, in _publish self._sock_client.send_record_publish(record) File "/home/djh/miniconda3/envs/xllm3/lib/python3.10/site-packages/wandb/sdk/lib/sock_client.py", line 221, in send_record_publish self.send_server_request(server_req) File "/home/djh/miniconda3/envs/xllm3/lib/python3.10/site-packages/wandb/sdk/lib/sock_client.py", line 155, in send_server_request self._send_message(msg) File "/home/djh/miniconda3/envs/xllm3/lib/python3.10/site-packages/wandb/sdk/lib/sock_client.py", line 152, in _send_message self._sendall_with_error_handle(header + data) File "/home/djh/miniconda3/envs/xllm3/lib/python3.10/site-packages/wandb/sdk/lib/sock_client.py", line 130, in _sendall_with_error_handle sent = self._sock.send(data) BrokenPipeError: [Errno 32] Broken pip

从上述提供的错误堆栈中，可以看到错误出现在使用wandb库（Weights & Biases库，一个用于机器学习实验跟踪和版本控制的库）的某个部分时具体来说，这个错误似乎是在尝试通过socket发送消息时发生的

错误发生在 `sock_client.py` 的 `_send_message` 方法中，但由于堆栈跟踪被截断，我无法看到具体的错误消息或异常类型

这可能是一个由多种原因引起的问题，以下是一些可能的解决方案：

1. **网络连接问题**：确保你的网络连接是稳定的，并且wandb服务器可以访问如果网络连接不稳定，尝试重新连接网络或等待网络稳定后再试
2. **wandb版本问题**：你可能使用的wandb库版本与你的Python环境或其他依赖项不兼容尝试更新wandb库到最新版本，可以使用以下命令：

```
bashpip install --upgrade wandb
```

1. **防火墙或安全组设置**：确保你的防火墙或安全组设置允许你的程序访问wandb服务器如果可能，尝试暂时禁用防火墙或调整设置以允许连接
2. **服务器问题**：有时，wandb服务器可能会出现问题或维护你可以尝试稍后再试，或者查看wandb的官方文档或社区以获取更多信息

如果以上解决方案都无法解决问题，请提供更多的错误信息和上下文，以便我能更准确地帮助你解决问题

考虑是没有使用过wandb的问题：

登录并按提示粘贴您的 API 密钥

wandb login

76ea5b2b06f6f9a718116bb3ec0bd54936f2fded

科研工具-01 使用Wandb实现高效实验管理 https://zhuanlan.zhihu.com/p/669141659?utm_id=0

2.2 运行中间态

2.3 运行结果

{'loss': 0.9949, 'grad_norm': 2.5483734607696533, 'learning_rate': 0.0, 'epoch': 0.1}

{'train_runtime': 10867.1481, 'train_samples_per_second': 9.264, 'train_steps_per_second': 0.289, 'train_loss': 1.153496361127418, 'epoch': 0.1}

100%|

| 3146/3146 [3:00:55<00:00, 3.45s/it]

wandb: | 0.767 MB of 0.767 MB uploaded

wandb: Run history:

wandb: train/epoch

wandb: train/global_step

wandb: train/grad_norm

wandb: train/learning_rate

wandb: train/loss

wandb: train/total_flos ▁

wandb: train/train_loss ▁

wandb: train/train_runtime ▁

wandb: train/train_samples_per_second ▁

wandb: train/train_steps_per_second ▁

wandb:

wandb: Run summary:

wandb: train/epoch 0.1

wandb: train/global_step 3146

wandb: train/grad_norm 2.54837

wandb: train/learning_rate 0.0

wandb: train/loss 0.9949

wandb: train/total_flos 5.067641141277491e+17

wandb: train/train_loss 1.1535

wandb: train/train_runtime 10867.1481

wandb: train/train_samples_per_second 9.264

wandb: train/train_steps_per_second 0.289

wandb:

wandb: 🚀 View run llama-7b-hf.alpaca_en+alpaca_zh+translation_ncwm_en-zh.finetune at: https://wandb.ai/dujh22team/xllm/runs/hu5oxk2h

wandb: ⚡ View job at https://wandb.ai/dujh22team/xllm/jobs/QXJ0aWZhY3RDb2xsZWN0aW9uOjE1MDc1MTE5OA==/version_details/v0

wandb: Synced 6 W&B file(s), 0 media file(s), 2 artifact file(s) and 0 other file(s)

wandb: Find logs at: /home/djh/log/wandb/run-20240320_180507-hu5oxk2h/logs

3. 运行推理脚本

3.1 报错 $ ‘\r ‘: 未找到命令

在Windows上写好的脚本，放在Linux上运行，却出现了如下错误：

```
./startup.sh:行3: $'\r': 未找到命令
```

2、原因分析

两种操作系统平台对换行的解析不同造成的，Windows中\r\n表示换行，而在Linux中\n表示换行，所以在Windows上编写好的shell文件上传到Linux后，会因为不能识别\r而报错。因此办法之一就是\r替换掉，可以使用下面的命令来操作：

```
sed -i 's/\r//' test.sh
```

3.2 不在同一个显卡上

> Traceback (most recent call last):
>
> File "/home/djh/code/xllm/inference2.py", line 241, in <module>
>
> inference()
>
> File "/home/djh/code/xllm/inference2.py", line 231, in inference

```
  output = evaluate_by_generate(d, template=generating_args.template,
generation_config=generation_config)
 File "/home/djh/code/xllm/inference2.py", line 120, in evaluate_by_generate
  generation_output = model.generate(
 File "/home/djh/miniconda3/envs/xllm3/lib/python3.10/site-
packages/torch/utils/_contextlib.py", line 115, in decorate_context
  return func(*args, **kwargs)
 File "/home/djh/miniconda3/envs/xllm3/lib/python3.10/site-
packages/transformers/generation/utils.py", line 1544, in generate
  return self.greedy_search(
 File "/home/djh/miniconda3/envs/xllm3/lib/python3.10/site-
packages/transformers/generation/utils.py", line 2404, in greedy_search
  outputs = self(
 File "/home/djh/miniconda3/envs/xllm3/lib/python3.10/site-
packages/torch/nn/modules/module.py", line 1511, in _wrapped_call_impl
  return self._call_impl(*args, **kwargs)
 File "/home/djh/miniconda3/envs/xllm3/lib/python3.10/site-
packages/torch/nn/modules/module.py", line 1520, in _call_impl
  return forward_call(*args, **kwargs)
 File "/home/djh/miniconda3/envs/xllm3/lib/python3.10/site-packages/accelerate/hooks.py",
line 166, in new_forward
  output = module._old_forward(*args, **kwargs)
 File "/home/djh/miniconda3/envs/xllm3/lib/python3.10/site-
packages/transformers/models/llama/modeling_llama.py", line 1176, in forward
  outputs = self.model(
 File "/home/djh/miniconda3/envs/xllm3/lib/python3.10/site-
packages/torch/nn/modules/module.py", line 1511, in _wrapped_call_impl
  return self._call_impl(*args, **kwargs)
 File "/home/djh/miniconda3/envs/xllm3/lib/python3.10/site-
packages/torch/nn/modules/module.py", line 1520, in _call_impl
  return forward_call(*args, **kwargs)
 File "/home/djh/miniconda3/envs/xllm3/lib/python3.10/site-
packages/transformers/models/llama/modeling_llama.py", line 1019, in forward
```

> layer_outputs = decoder_layer(
>
> File "/home/djh/miniconda3/envs/xllm3/lib/python3.10/site-packages/torch/nn/modules/module.py", line 1511, in _wrapped_call_impl
>
> return self._call_impl(*args, **kwargs)
>
> File "/home/djh/miniconda3/envs/xllm3/lib/python3.10/site-packages/torch/nn/modules/module.py", line 1520, in _call_impl
>
> return forward_call(*args, **kwargs)
>
> File "/home/djh/miniconda3/envs/xllm3/lib/python3.10/site-packages/accelerate/hooks.py", line 166, in new_forward
>
> output = module._old_forward(*args, **kwargs)
>
> File "/home/djh/miniconda3/envs/xllm3/lib/python3.10/site-packages/transformers/models/llama/modeling_llama.py", line 740, in forward
>
> hidden_states, self_attn_weights, present_key_value = self.self_attn(
>
> File "/home/djh/miniconda3/envs/xllm3/lib/python3.10/site-packages/torch/nn/modules/module.py", line 1511, in _wrapped_call_impl
>
> return self._call_impl(*args, **kwargs)
>
> File "/home/djh/miniconda3/envs/xllm3/lib/python3.10/site-packages/torch/nn/modules/module.py", line 1520, in _call_impl
>
> return forward_call(*args, **kwargs)
>
> File "/home/djh/miniconda3/envs/xllm3/lib/python3.10/site-packages/accelerate/hooks.py", line 166, in new_forward
>
> output = module._old_forward(*args, **kwargs)
>
> File "/home/djh/miniconda3/envs/xllm3/lib/python3.10/site-packages/transformers/models/llama/modeling_llama.py", line 655, in forward
>
> key_states, value_states = past_key_value.update(key_states, value_states, self.layer_idx, cache_kwargs)
>
> File "/home/djh/miniconda3/envs/xllm3/lib/python3.10/site-packages/transformers/cache_utils.py", line 131, in update
>
> self.key_cache[layer_idx] = torch.cat([self.key_cache[layer_idx], key_states], dim=-2)
>
> RuntimeError: Expected all tensors to be on the same device, but found at least two devices, cuda:0 and cuda:1! (when checking argument for argument tensors in method wrapper_CUDA_cat)

这个错误信息提示说，期望所有张量都在同一个设备上，但是至少发现了两个设备，`cuda:0` 和 `cuda:1`。这表明你的代码在执行过程中尝试将存储在不同CUDA设备上的张量进行操作，导致了这

个错误。

这通常发生在使用多GPU环境中，特别是当你尝试进行跨GPU的操作时。在PyTorch中，每个张量都与一个设备相关联，而某些操作要求所有参与的张量必须位于同一设备上。

为了解决这个问题，你需要确保所有的张量操作都在同一个设备上执行。以下是一些可能的解决方案：

1. 确保模型和所有输入数据都在同一个设备上。

> 📍 如果你希望只在单个GPU上运行你的脚本，确保所有的模型和数据都被显式地放到了同一个GPU上。在PyTorch中，你可以使用 `.to(device)` 方法来指定使用的设备。在你的情况下，如果你想要将所有的操作都限制在 `cuda:0` 上，你可以按照以下步骤操作：
>
> 1. 指定设备：在你的脚本开始时定义一个设备变量，所有的模型和数据将被移动到这个设备上。
> 2. 移动模型到指定设备：在加载模型后，使用 `.to(device)` 方法将模型移动到你指定的设备上。
> 3. 确保数据在同一设备上：在将数据喂给模型之前，使用 `.to(device)` 确保数据也在同一设备上。
>
> 下面是根据你的脚本做出的一些修改，以确保所有操作都在 `cuda:0` 上执行：

```python
import torch
import os

# 显式地设置CUDA设备
os.environ["CUDA_VISIBLE_DEVICES"] = "0"  # 仅使第一个GPU对当前进程可见
device = torch.device("cuda" if torch.cuda.is_available() else "cpu")

# 加载模型，并显式地将其放置在指定的设备上
model = transformers.AutoModelForCausalLM.from_pretrained(
    model_args.model_name_or_path,
    load_in_8bit=model_args.load_in_8bit,
    torch_dtype=model_args.torch_dtype
).to(device)

# 确保所有后续的数据处理和模型推理都使用这个设备
# （省略中间代码）
inputs = tokenizer(prompt, padding=True, return_tensors="pt").to(device)

with torch.no_grad():
    generation_output = model.generate(
        input_ids=inputs["input_ids"],
        attention_mask=inputs["attention_mask"],
```

```
        # 确保指定了所有需要的参数
    )

# （省略后续代码）

```

2. 审查 `accelerate` 库的使用。从错误信息来看，问题似乎出现在使用 `accelerate` 库时。确保你正确地设置了 `device_map` 和其他相关配置，以确保所有操作都在正确的设备上执行。特别是，当使用 `load_checkpoint_and_dispatch` 函数时，要确保它正确地处理了模型和数据的设备分配。

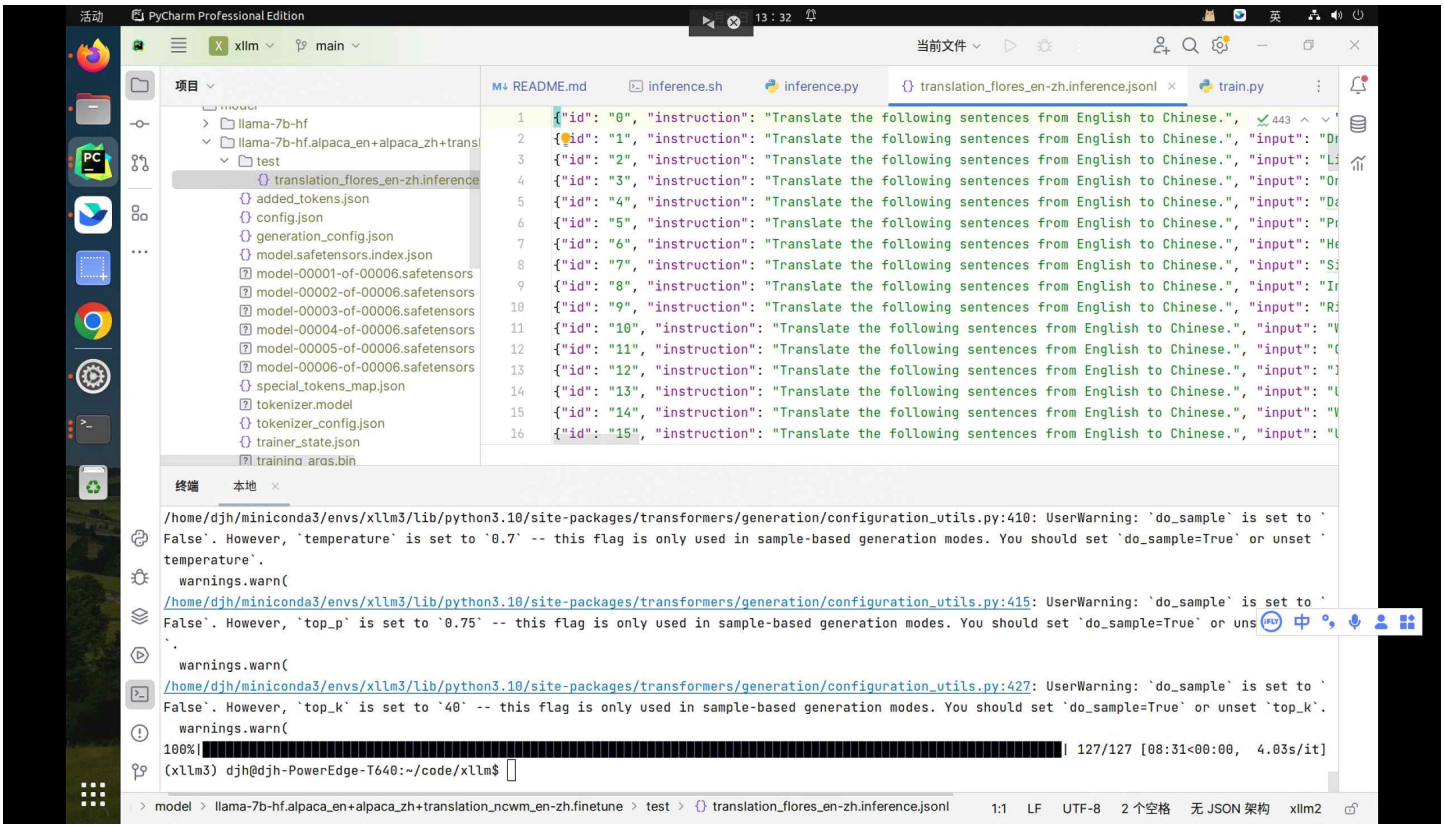
> 🌊 如果你确实需要使用 `accelerate` 库，但希望确保所有计算只在 `cuda:0` 上执行，你可以通过显式指定 `device_map` 参数来控制模型加载到特定的GPU上，而不是使用 `"auto"` 自动分配。 `device_map` 允许你详细指定每个模型部分应该在哪个设备上执行。
>
> 修改后的代码段可以如下设置：

```python
if torch.cuda.device_count() > 1:
    from accelerate import load_checkpoint_and_dispatch
    # 指定所有计算仅在 cuda:0 上执行
    device_map = {0: "cuda:0"}  # 或者使用 {0: 0}，取决于 accelerate 版本

    load_checkpoint_and_dispatch(
        model,  # 模型
        model_args.model_name_or_path,  # 模型的名称或路径
        device_map=device_map,  # 使用显式的设备映射
        offload_state_dict=True,  # 是否卸载状态字典
        no_split_module_classes=["LlamaDecoderLayer"],  # 不分割的模块类
    )

```

3.3 具体结果

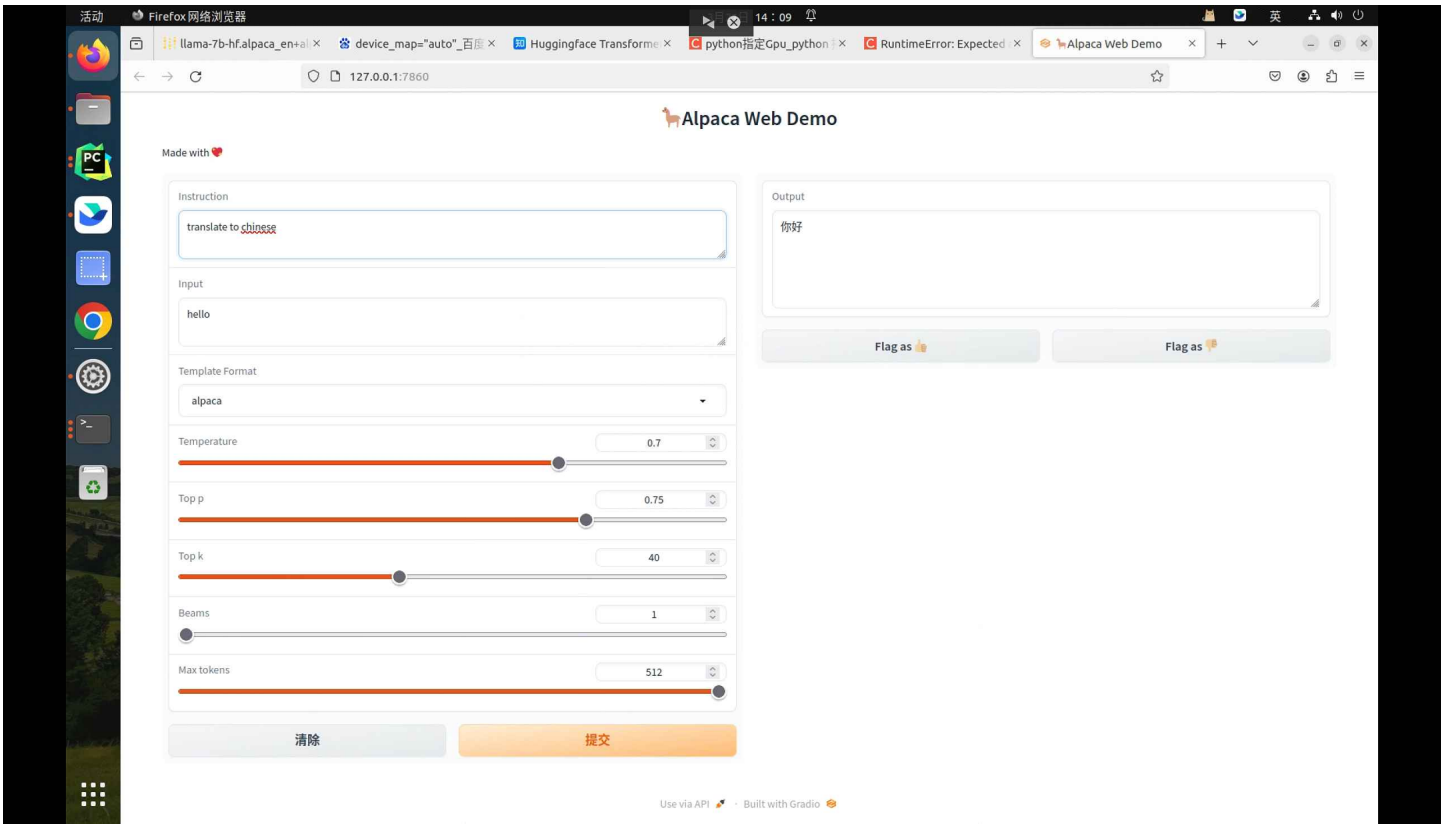


4. 通过 Web UI 与 LLM 交互

4.1 报错：AttributeError: module 'gradio' has no attribute 'inputs'

```
pip install gradio==3.39.0
```

4.2 具体结果


活动
Firefox 网络浏览器
14：09
英
llama-7b-hf.alpaca_en+al
device_map="auto"_百度
Huggingface Transforme
python指定Gpu_python
RuntimeError: Expected
Alpaca Web Demo
127.0.0.1:7860
Alpaca Web Demo
Made with
Instruction
translate to chinese
Input
hello
Template Format
alpaca
Temperature
0.7
Top p
0.75
Top k
40
Beams
1
Max tokens
512
清除
提交
Output
你好
Flag as
Flag as
Use via API
Built with Gradio