Q1:

ImportError: cannot import name 'COMMON_SAFE_ASCII_CHARACTERS' from 'charset_normalizer.constant' (/home/wloner0809/miniconda3/envs/Grounding4Rec/lib/python3.9/site-packages/charset_normalizer/constant.py)

S1:

> pip3 install chardet

Q2:

huggingface_hub.utils.*validators.HFValidationError: Repo id must use alphanumeric chars or '-', '', '.', '--' and '..' are forbidden, '-' and '.' cannot start or end the name, max length is 96: 'YOUR_LLAMA_PATH/'.*

S2:

> 没有写模型的名字导致的

Q3:

ImportError: Using `load_in_8bit=True` requires Accelerate: `pip install accelerate` and the latest version of bitsandbytes `pip install -i https://test.pypi.org/simple/ bitsandbytes` or pip install bitsandbytes`

S3:

> Ubuntu物理机：
>
> 更新bitsandbytes包，安装scipy包(没用)
> Linux上要有英伟达驱动：sudo apt install nvidia-utils-470; sudo apt install nvidia-cuda-toolkit; install nvidia drivers; 装完nvidia驱动之后注意要装一个插件支持触控板手势
>
> Linux服务器：
>
> 因为tranformers库的版本太高，降级即可transformers==4.30.0就可以

Q4:

OSError: We couldn't connect to 'https://huggingface.co' to load this file, couldn't find it in the cached files and it looks like decapoda-research/llama-7b-hf is not the path to a directory containing a file named config.json.

S4:

> 在Ubuntu物理机上配置代理:
>
> export https_proxy=http://127.0.0.1:7890
> export http_proxy=http://127.0.0.1:7890
> export all_proxy=socks5://127.0.0.1:7890
> 加入到~/.zshrc中(在使用zsh的情况下), 配置代理
>
> 如何配置Linux服务器的现在还不会……

Q5:

ValueError: Some modules are dispatched on the CPU or the disk. Make sure you have enough GPU RAM to fit the quantized model. If you want to dispatch the model on the CPU or the disk while keeping these modules in 32-bit, you need to set `load_in_8bit_fp32_cpu_offload=True` and pass a custom `device_map` to `from_pretrained`. Check https://huggingface.co/docs/transformers/main/en/main_classes/quantization#offload-between-cpu-and-gpu for more details.

S5:

> 这里先检查一下 torch.cuda.is_available() 是否为True，如果是False的话说明Pytorch可能装的cpu版本的。(一定要注意去官网上按照对应版本的cuda安装pytorch)。我这里出错是因为先装了accelerate库，同时他给装了torch，但是为cpu版本。
>
> 其次看看gpu大小是否足够。

Q6:

torch.cuda.OutOfMemoryError: CUDA out of memory. Tried to allocate 58.00 MiB (GPU 0; 23.70 GiB total capacity; 20.55 GiB already allocated; 40.56 MiB free; 21.96 GiB reserved in total by PyTorch) If reserved memory is >> allocated memory try setting max_split_size_mb

to avoid fragmentation. See documentation for Memory Management and PYTORCH_CUDA_ALLOC_CONF

S6:

> 调小batch_size即可
>
> 这里因为是单卡跑的，如果batch size设置太大就会超显存
>
> 经验：**多试几次……**

Q7:

RuntimeError: Expected all tensors to be on the same device, but found at least two devices, cuda:0 and cpu! (when checking argument for argument index in method wrapper__index_select)

S7:

> 根据报错行，检查是否有变量不在gpu上(即有没有.to(device))
>
> 这个纯属是因为经验不足了，之前没跑过深度学习的代码……会不了一点cuda、gpu