基于大语言模型的患者就诊咨询系统

23 秋文本挖掘组员：

曹连才（231017000024）、孙兵（231017000130）、李泽宇（231017000012）、张佳硕（231017000173）

一、项目背景

随着人工智能技术的不断发展，医疗行业也逐渐迎来了 AI 技术的应用。人工智能在医疗领域的趋势是利用大数据分析、自然语言处理和机器学习等技术，为患者提供更快速、准确的诊断和治疗建议。在线问诊平台正成为医疗行业的一个新兴趋势，为患者提供便捷、高效的医疗服务。

传统的医疗问诊方式存在诸多问题，例如排队时间长、医生资源不均衡分布等，限制了患者获取及时有效的医疗帮助。而基于大语言模型的人工智能医疗在线问诊项目，可以帮助患者随时随地进行医疗咨询，解决传统医疗系统中的痛点问题，提供更加便捷、高效的医疗服务。

本项目着重专注与在线医疗健康咨询服务，基于 Chatglm2-6B 大语言模型微调，实现自动化患者健康咨询服务。 通过提取医患对话语料库，建立语言模型训练，形成针对患者提问自动应答，该项目的推出将为患者提供 24 小时不间断的医疗咨询服务，缓解医院就诊压力，提高就医效

率，同时能够根据患者的症状描述和病史数据，利用大语言模型进行智能诊断和给出治疗建议，极大地提升了医疗服务的质量和效率。此外，通过积累大量的医疗数据，还可以促进医学研究和学术交流，为医疗行业的发展做出贡献。

二、项目问题

尽管智能医疗系统取得了进展，但过去的研究主要侧重于特定任务或疾病，适用性有限，导致实验进展与实际应用之间存在差距。为了弥合这一差距，需要综合解决方案来涵盖更广泛的健康咨询场景，并以端到端对话方式为用户提供高质量的医疗服务。

三、解决思路

获取医患对话语料作为样例,引入 Prefix-tuning 思想，采用基于 P-tuning v2 通过 ChatGLM2-6B 模型微调训练。

四、技术选型

1.**技术环境**：针对于组员拥有可提供的 3090GPU 和对应的 CUDA 以及相匹配的 Linux 系统，小组讨论后决定使用该环境进行技术实现。其中 CUDA 版本为 12.2、Python 版本为 3.8.5、Pytorch 版本为 2.1.1。

2.**大语言模型**：基于魔塔社区的提供的开源 LLM 信息，做了一些简单的测试后选择“智谱 ChatGLM2-6B”

3.**微调数据集：**基于公开医患对话数据集进行下载 https://opendatalab.com/TomLi/Med2023。研读与理解数据，整理和简化对话流程。

五、具体实现

1.前期准备。

运行微调需要 4.27.1 版本的 transformers。除 ChatGLM2-6B 的依赖之外，还需要安装以下依赖：

```
pip install rouge_chinese nltk jieba datasets
```

2.下载数据集。

从 OPENXLAB 下载医患对话语料库（https://opendatalab.com/TomLi/Med2023）

这是医患多轮对话及医药文本提取里边相同的 id 有相同长度的多轮对话，doc 为根据多轮对话中的关键词所在医疗药品库中提出的药品 doc。question 包含多轮对话以及最后一轮的提问，answer 为最后一轮提问的回答。经过数据清洗，整理为 Q&A（患者提问&医生回答）数据集，并拆分训练，测试和验证集。

数据清洗过程：

In [1]:
```
import json
import pandas as pd
import numpy as np
```

In [38]:
```
output = []
with open('./1.jsonl','r',encoding='utf-8') as f:
    for data in f:
        data_json = json.loads(data)
        content = data_json['question'][0]['text']
        summary = data_json['answer']
        output.append({'content':content,"summary":summary})
data_json['question'][0]['text']
data = pd.DataFrame(output)
data.drop_duplicates(keep='first',inplace=True,ignore_index=True)
data.to_json(path_or_buf='./2.json',orient='records',force_ascii =False)
```

清洗后数据样例：

]:
```
data_json['question'][0]['text']
```

]:
```
'最近几天开始 肚脐右侧时不时腹痛。大便次数少 有也基本不成形。放屁比较多。好像放屁和大便后疼痛减轻。有时用手按压疼痛部位 会有咕噜肚子响的声音。之前每天喝很多铁观音茶 后来茶叶没了就没再喝 不喝后就开始痛 是不是跟没喝茶了有关系。不过 喝的时候好像就有微微的痛 也是时不时的。对了 最近每天都在吃一种很辛辣的干拌面 每天都吃 要么中午 要么晚上。没喝茶后还持续吃了一周多时间这种面 最近痛了那个面也就没敢吃了。不知道跟这个肚脐右侧痛有没有关系。还有 最近明显痛起来是因为前几天生气发火 那天回到家就开始痛了 那天最痛。现在没那么痛 时不时发作 昨天晚上吃太多了就有点痛起来。本人174 65公斤 属于偏瘦的。作息不好 晚起晚睡 最近两年肚子有点大起来。以前每天抽几只烟 后来没喝茶后习惯性的也就没抽烟了。 上面的信息比较多 我也不知道哪些有用 都说出来让大夫好做判断。谢谢。(男,26岁)'
```

]:
```
data_json['question'][1]['text']
```

]:
```
'你好,首先,可能和辛辣食物有关,不要再吃。避免烟酒茶。清淡少渣饮食,最好吃几天粥。如果大便不成形那么要化验大便。可以服乳酸杆菌调节肠道菌群。如果症状一直不好那么最好肠镜检查。可能性最大的是肠炎'
```

]:
```

```

3. 训练。

P-Tuning v2

3.1 训练数据集

训练数据集放在 train.json 中，验证数据集放在 dev.json 中。数据示例如下：

```
[
{"content":"肚子总是胀气,放屁也放不出来,肚子鼓鼓的好烦啊,还难受",
"summary":"你好,胃肠功能紊乱,建议用药调理"},
{"content":"五个月宝宝便秘大便很干怎么办?(女,22 岁)有什么好办法",
"summary":"增加水果蔬菜等纤维素多的食物,多喂点水"},
{"content":"心烧咳嗽痰多没力气肚子疼(男,54)",
"summary":"你好,根据目前症状建议医院做个血常规和胸片,考虑是呼吸道感染"},
{"content":"怀孕 40 多天 最近恶心呕吐加重 怎么回事 怎么缓解",
"summary":"考虑为正常的早孕反应,建议你少量多餐,吃清单易消化的食物,禁食辛辣刺激性、凉硬的食物,可以吃西红柿、黄瓜能减轻恶心,若恶心呕吐严重的话是需要点滴营养液的"},
{"content":"拉肚子 水泄控制不住 应该吃什么药(女,31 岁)",
"summary":"那就是着凉了,可以吃点思密达和贝飞达"},,
……
]
```

3.2 修改训练脚本

```
(base) sunbing@s3090:~/ChatGLM2-6B/ptuning$ git diff train.sh
diff --git a/ptuning/train.sh b/ptuning/train.sh
index 0d161ce..412b082 100644
--- a/ptuning/train.sh
+++ b/ptuning/train.sh
@@ -4,16 +4,16 @@ NUM_GPUS=1

 torchrun --standalone --nnodes=1 --nproc-per-node=$NUM_GPUS main.py \
     --do_train \
-    --train_file AdvertiseGen/train.json \
-    --validation_file AdvertiseGen/dev.json \
+    --train_file train.json \
+    --validation_file dev.json \
     --preprocessing_num_workers 10 \
     --prompt_column content \
     --response_column summary \
     --overwrite_cache \
-    --model_name_or_path THUDM/chatglm2-6b \
+    --model_name_or_path ../model \
     --output_dir output/adgen-chatglm2-6b-pt-$PRE_SEQ_LEN-$LR \
     --overwrite_output_dir \
-    --max_source_length 64 \
+    --max_source_length 128 \
     --max_target_length 128 \
     --per_device_train_batch_size 1 \
     --per_device_eval_batch_size 1 \
```

3.3 运行以下指令进行训练：

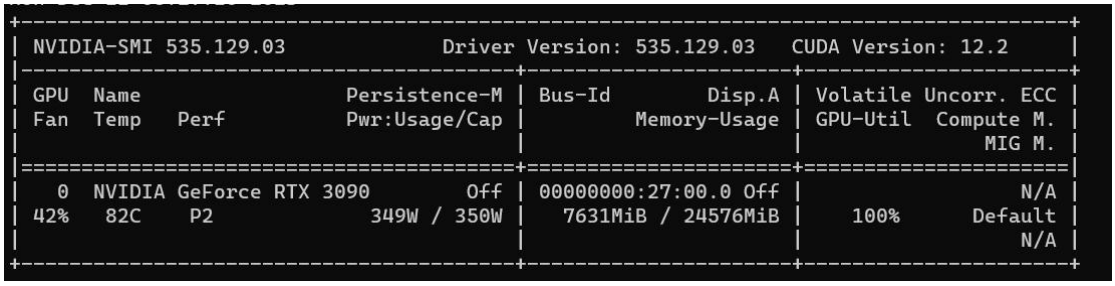
bash train.sh

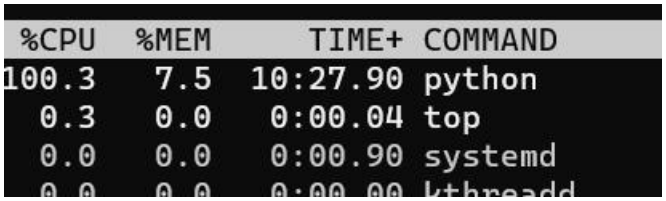
train.sh 中的 PRE_SEQ_LEN 和 LR 分别是 soft prompt 长度和训练的学习率，可以进行调节以取得最佳的效果。P-Tuning-v2 方法会冻结全部的模型参数，可通过调整 quantization_bit 来被原始模型的量化等级，不加此选项则为 FP16 精度加载。脚本如下：

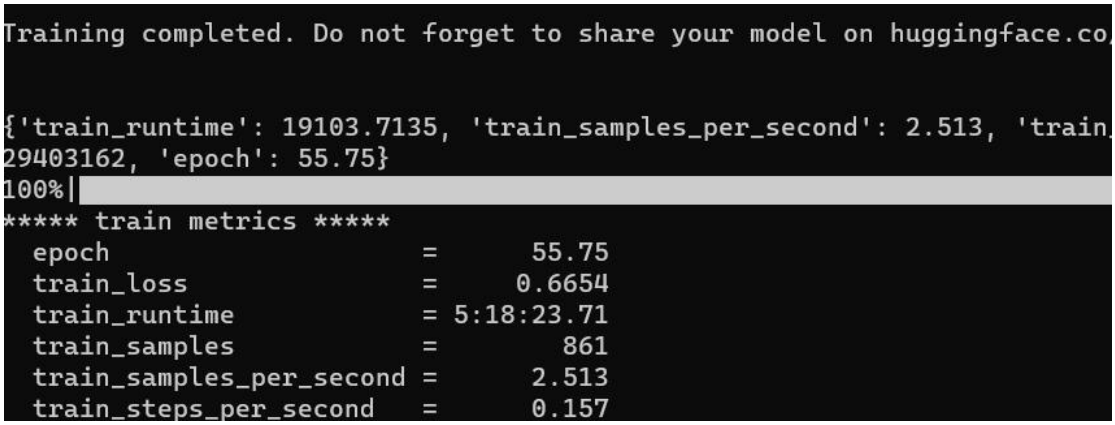
```
PRE_SEQ_LEN=128
LR=2e-2
NUM_GPUS=1

torchrun --standalone --nnodes=1 --nproc-per-node=$NUM_GPUS main.py \
    --do_train \
    --train_file train.json \
    --validation_file dev.json \
    --preprocessing_num_workers 10 \
    --prompt_column content \
    --response_column summary \
    --overwrite_cache \
    --model_name_or_path ../model \
    --output_dir output/adgen-chatglm2-6b-pt-$PRE_SEQ_LEN-$LR \
    --overwrite_output_dir \
    --max_source_length 128 \
    --max_target_length 128 \
    --per_device_train_batch_size 1 \
    --per_device_eval_batch_size 1 \
    --gradient_accumulation_steps 16 \
    --predict_with_generate \
    --max_steps 3000 \
    --logging_steps 10 \
    --save_steps 1000 \
    --learning_rate $LR \
    --pre_seq_len $PRE_SEQ_LEN \
    --quantization_bit 4
```

具体内容见附件。

3.4 以下是训练过程截图：





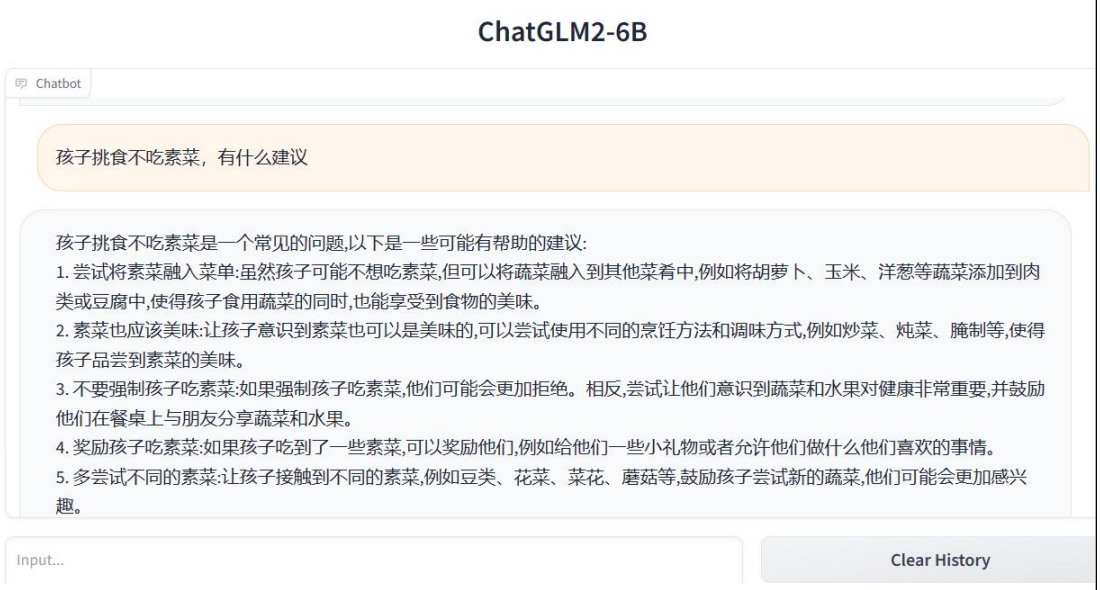


整个训练过程在夜间进行，总共耗时约 5 小时 20 分钟。

4. 结果

基于微调前的原始模型和微调后的对话模型，小组成员尝试录入一些问题，并比较前后应答差异，微调后的回答相对更加简洁，信息冗余度更低，建议指导性、针对性更强，方便患者理解。以下是展示案例：

训练前患者医生对话


ChatGLM2-6B
Chatbot
孩子挑食不吃素菜，有什么建议
孩子挑食不吃素菜是一个常见的问题,以下是一些可能有帮助的建议:
1. 尝试将素菜融入菜单:虽然孩子可能不想吃素菜,但可以将蔬菜融入到其他菜肴中,例如将胡萝卜、玉米、洋葱等蔬菜添加到肉类或豆腐中,使得孩子食用蔬菜的同时,也能享受到食物的美味。
2. 素菜也应该美味:让孩子意识到素菜也可以是美味的,可以尝试使用不同的烹饪方法和调味方式,例如炒菜、炖菜、腌制等,使得孩子品尝到素菜的美味。
3. 不要强制孩子吃素菜:如果强制孩子吃素菜,他们可能会更加拒绝。相反,尝试让他们意识到蔬菜和水果对健康非常重要,并鼓励他们在餐桌上与朋友分享蔬菜和水果。
4. 奖励孩子吃素菜:如果孩子吃到了一些素菜,可以奖励他们,例如给他们一些小礼物或者允许他们做什么他们喜欢的事情。
5. 多尝试不同的素菜:让孩子接触到不同的素菜,例如豆类、花菜、菜花、蘑菇等,鼓励孩子尝试新的蔬菜,他们可能会更加感兴趣。
Input...
Clear History

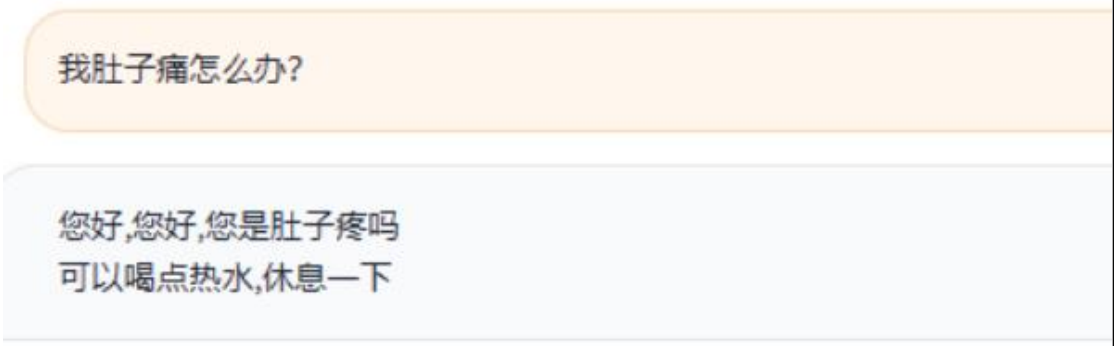

ChatGLM2-6B
Chatbot
我肚子痛怎么办?
肚子痛的时候,建议采取以下措施:
1. 休息:尽量减少身体运动,躺下来放松身体,减少进一步的活动。
2. 饮食:如果感到饿了,可以吃一些易消化的食物,例如面包、米饭或水果。
3. 冷敷:用冰袋或冰块袋冷敷腹部,可以缓解疼痛和不适感。
4. 服药:如果疼痛不太严重,可以考虑自己购买一些非处方药物,例如止痛药或抗酸药。
5. 就医:如果疼痛非常剧烈或者已经过去了一段时间,建议尽快就医,以便获得更好的治疗。
如果感到不适或者疼痛过于剧烈,请尽快寻求医疗帮助,以避免出现更严重的问题。


ChatGLM2-6B
Chatbot
早上刷牙出血，是什么原因
刷牙出血可能有多种原因,以下是一些常见的原因:
1. 牙龈问题:牙龈可能会因为刷牙不当、出血、感染或疾病而出现这些问题。如果发现刷牙时牙龈出血,可能是某种疾病或感染导致的。
2. 牙渍:牙渍积聚在牙齿表面会逐渐形成,可能会在刷牙时导致出血。牙渍通常是由食物残渣、细菌和其他污垢组成的。
3. 牙龈炎:牙龈炎是一种常见的牙周疾病,常常由牙菌斑引起。它会导致牙龈出现红肿、出血、疼痛等症状。
4. 牙周病:牙周病是一种更严重的牙周疾病,通常由牙龈炎发展而来。它会导致牙齿松动、牙槽骨流失、出血等问题。
5. 牙齿问题:如果牙齿存在问题,例如裂缝、洞、敏感等问题,那么刷牙时也可能会导致出血。
如果刷牙时经常出血,建议尽快咨询牙医,确定出血的原因并采取相应的治疗措施。
Input...
Clear History

ChatGLM2-6B
Chatbot
孩子挑食不吃素菜，有什么建议
孩子挑食不吃素菜是一个常见的问题,以下是一些可能有帮助的建议:
1. 尝试将素菜融入菜单:虽然孩子可能不想吃素菜,但可以将蔬菜融入到其他菜肴中,例如将胡萝卜、玉米、洋葱等蔬菜添加到肉类或豆腐中,使得孩子食用蔬菜的同时,也能享受到食物的美味。
2. 素菜也应该美味:让孩子意识到素菜也可以是美味的,可以尝试使用不同的烹饪方法和调味方式,例如炒菜、炖菜、腌制等,使得孩子品尝到素菜的美味。
3. 不要强制孩子吃素菜:如果强制孩子吃素菜,他们可能会更加拒绝。相反,尝试让他们意识到蔬菜和水果对健康非常重要,并鼓励他们在餐桌上与朋友分享蔬菜和水果。
4. 奖励孩子吃素菜:如果孩子吃到了一些素菜,可以奖励他们,例如给他们一些小礼物或者允许他们做什么他们喜欢的事情。
5. 多尝试不同的素菜:让孩子接触到不同的素菜,例如豆类、花菜、菜花、蘑菇等,鼓励孩子尝试新的蔬菜,他们可能会更加感兴趣。
Input...
Clear History

ChatGLM2-6B

训练后患者医生对话



5.改进方向。

（1）微调 loss 值较高，微调效果较差；

（2）回答过于简洁，可能与训练语料质量有关；

（3）微调方式：现阶段还有 LORA、Freeze 等其他方式可以尝。

附件：ChatGLM2-6B.tgz

附录：人员名单及负责工作内容

工作内容	负责组员
1. 组织工作	曹连才
2. 整理收集数据	曹连才（搜集、清洗、展示） 张佳硕（协助搜集）
3. 搭环境	孙兵(实施) 李泽宇(指导配合)
4. 开发测试	孙兵 李泽宇
5. 演示文稿	张佳硕 曹连才 孙兵 李泽宇