# MiniGPT

万振南 计23 2021030014

## 一、数据集和预处理

### 1. 预训练数据集

中文wiki百科数据集的一个子集：wiki-zh-subset-train_subset.jsonl

下载链接：https://cloud.tsinghua.edu.cn/d/d6f9966d6b684b05a516/

### 2. 微调数据集

**sft_data**

这是对预训练用到的所有数据进行问答生成，得到较大的微调数据集，希望让模型能够学到一些问答方式，提高模型的在未知问题和答案上的泛化能力

原始数据集：wiki-zh-subset-train_subset.jsonl

数据处理：将大型数据集按行数分割成较小的块，以便逐块处理。这里通过 data/sft_data/generate.sh 将输入文件按每块 100 行分割，并逐块处理。确定总行数和需要处理的行数。计算需要处理的文件块数，并确保每块的行数一致，最后一块可能不足 100 行

生成问答对：使用 data/sft_data/generate.py 从每块文本数据中生成问答对。从 JSONL 文件中逐行读取文本。对于过长的文本，即超过 1600 字符，进行截断以适应模型输入限制。使用预训练模型（如 llama 3.1）生成五个问答对，确保输出符合指定格式。检查生成的问答对是否符合预定的 JSON 格式和数量

数据保存：保存生成的数据：将生成的问答对按指定格式保存回新的 JSONL 文件中。最终得到约 80000 个问答对，使用了前 40000 个问答对

prompt 示例：

```
    prompt = f'请为下面的文本设计5个问答，每个问答之间用换行符隔开，文本如下：\n{text}\n\
你需要遵守下面的格式，注意下面的内容只用来掌握格式，示例如下：\n\
{{"question": "红柯的生长环境和分布地区有哪些？", "answer": "红柯是中国的特有植物，分布在中国大陆的海南等地，生长于海拔350米至1,000米的地区，常生于常绿阔叶林中或海拔较高的山地。"}}\n\
{{"question": "哈卡斯人主要分布在哪里？", "answer": "哈卡斯人主要分布在俄罗斯哈卡斯共和国、部分分布在克拉斯诺亚尔斯克等地，另外还有一部份分布在中国黑龙江省齐齐哈尔市富裕县。"}}\n\
{{"question": "桂城站的出入口及周边有哪些设施？", "answer": "桂城站目前设置的2个出入口均位于南桂东路南侧。本站设有便利店、面包糕饼店、中国银行自动柜员机、自动售货机及“好易”机。"}}\n\
{{"question": "圣地亚哥-罗里盖兹省的名字来源是什么？", "answer": "圣地亚哥-罗里盖兹省的名字来源于建立圣伊纳西奥城镇的军人圣地亚哥-罗里盖兹，他是海地和多明尼加战争中的人物，也是复国战争早期的领导人士之一。"}}\n\
{{"question": "模糊集的隶属函数是什么？", "answer": "模糊集的隶属函数是一个从论域到单位区间的映射，用来表示元素对该集的归属程度。"}}\n\
'
```

**sft_data_aug**

这是对一些样例微调数据进行问题的改写，即数据增强（data augmentation），希望让模型能够学会不同的提问方式，提高模型在特定答案上的泛化能力

原始数据集：A.jsonl（60 条样例微调数据，由 20 条样例微调数据 sft_data.jsonl 和 40 条样例微调数据 SFT补充训练集.jsonl 构成）

生成问题：使用 data/sft_data_aug/gen2.py 从每块文本数据中生成问答对。从 JSONL 文件中逐行读取文本。对于每个读取到的问题，通过调用预训练模型（如llama3.1）生成问题的变体，模型返回10个不同的变体问题，检查生成的问题变体是否符合要求。如果在生成变体过程中出现错误，或者生成的变体数量不正确，则重新生成，直到满足要求

数据保存：将生成的每个问题变体和对应的答案写入新的输出文件中。每个变体和答案以 JSON 对象的形式逐行写入输出文件，形成新的微调数据集。最终得到约 1200 个问答对

prompt 示例：

```
    prompt = f'根据下面的答案：\n{answer}\n\n改写下面的问题，但不要改变原意，每个编辑距离<5，不能和原来的内容一样，输出10行，每行为1个问题。原问题如下：\n{question}'
```

最终采用的微调数据集由 sft_data_aug 的约 1200 个问答对和 sft_data 的约 40000 个问答对构成

### 3. 数据预处理

文本编码器：tiktoken.get_encoding("gpt2")

文本编码：enc.encode_ordinary(data)

训练集和验证集划分：90%用作训练集，10%用作验证集

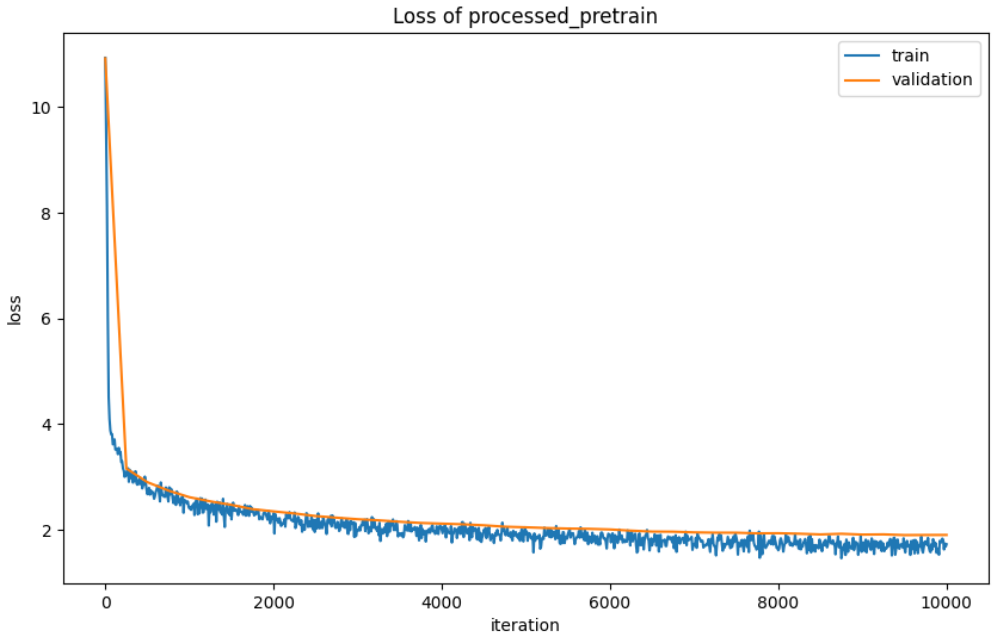
数据保存：使用 NumPy 将编码后的文本数据保存为二进制文件，训练集和验证集分别保存在名为 train.bin 和 val.bin 的文件中

## 二、模型训练和微调

### 1. 预训练

预训练参数：config/train_config.py

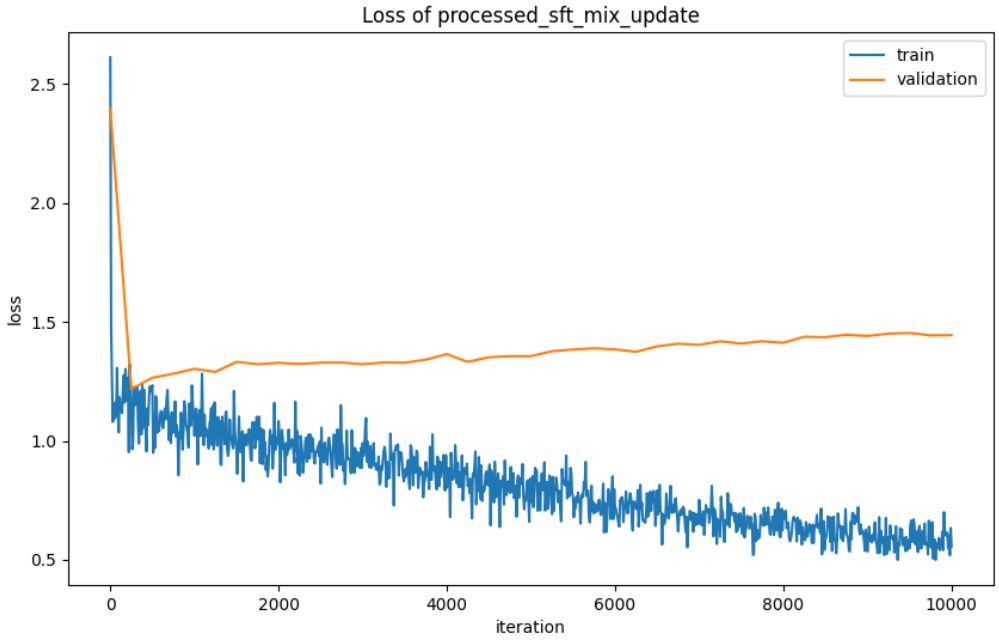
预训练损失曲线：



## 2. 微调

微调参数：config/sft_config.py

微调损失曲线：



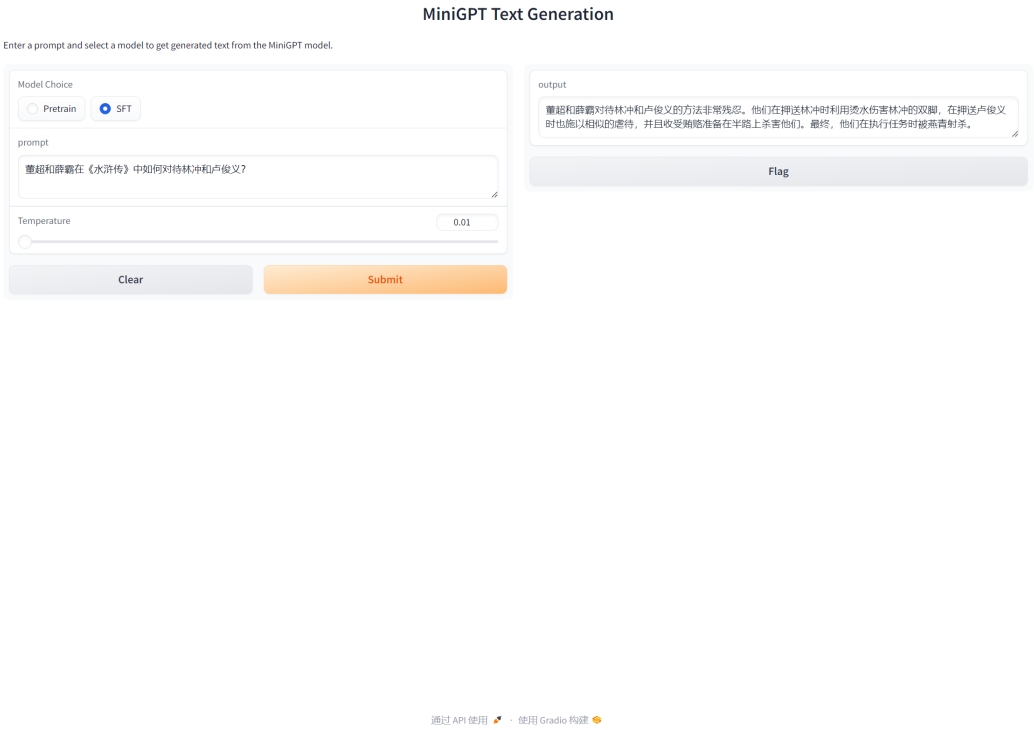

# 三、模型推理与评估

## 1. 模型推理

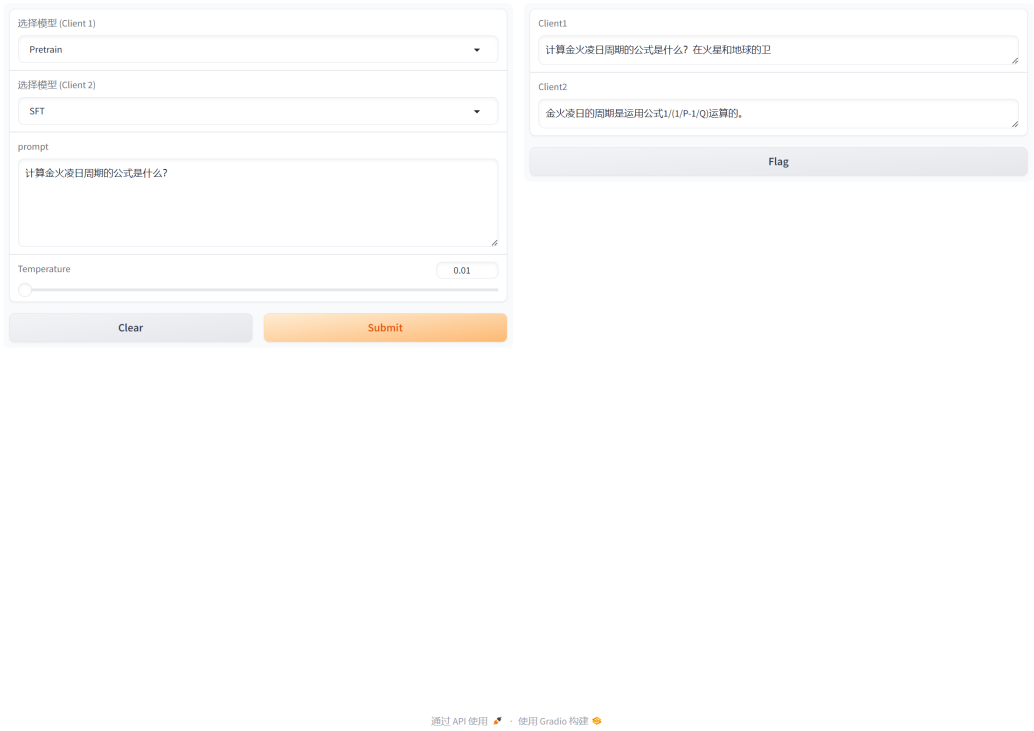
### 可视化效果展示

预训练：


微调：



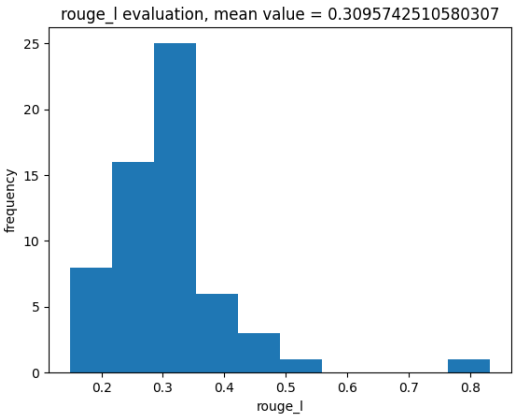

竞技场：

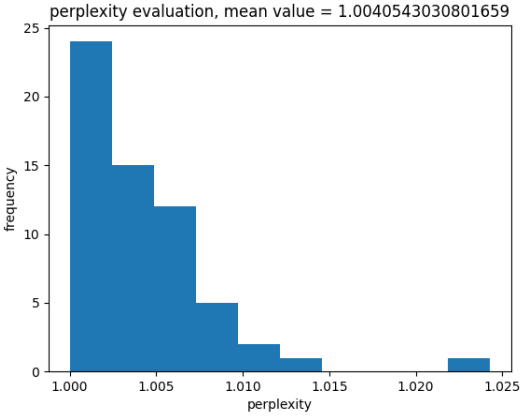



# 2. 模型评估

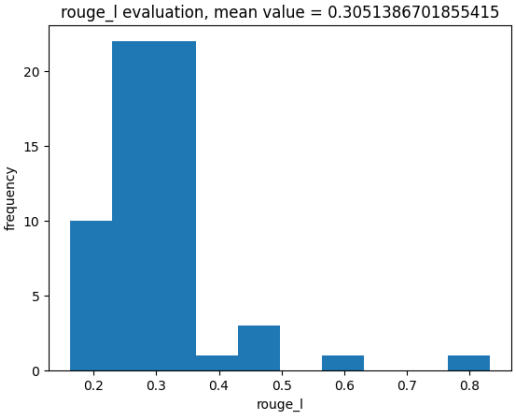
## 预训练

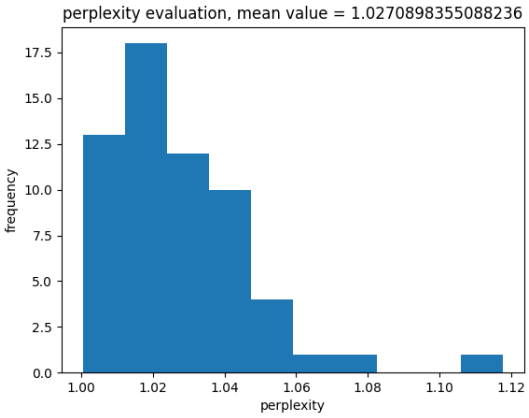
temperature=0.01：



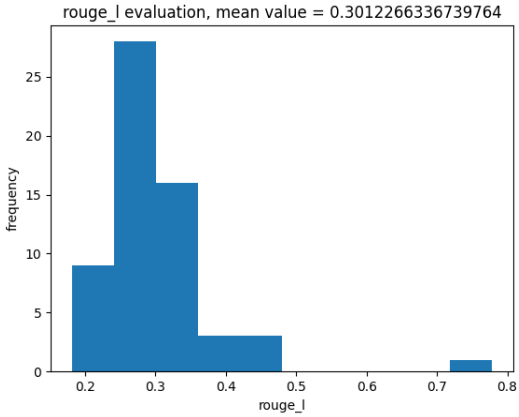



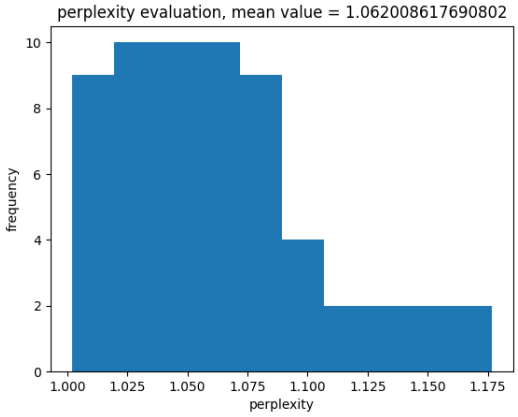

temperature=0.1：





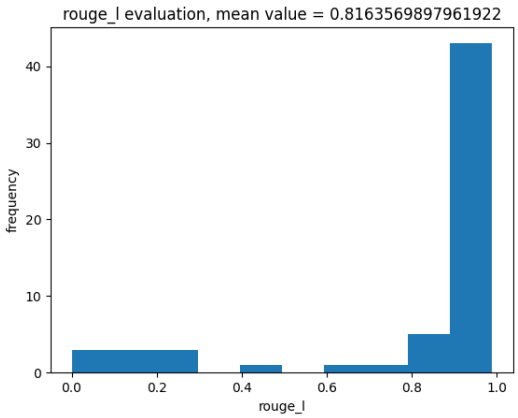

temperature=0.2：

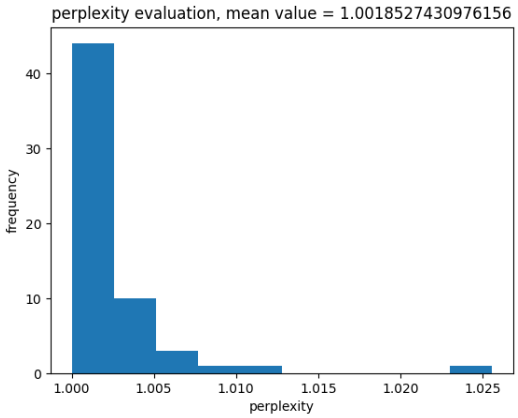


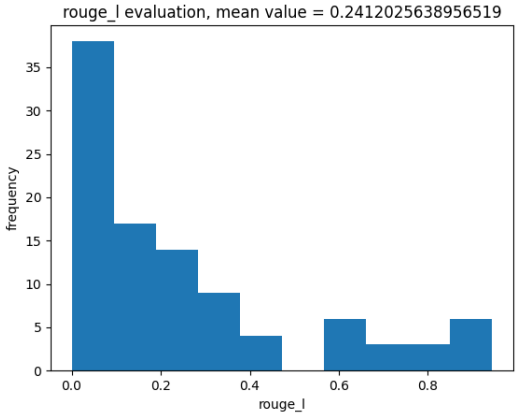

**微调**

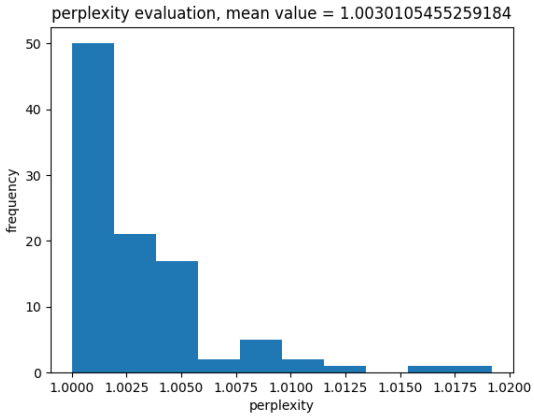
在数据集 A.jsonl 上：



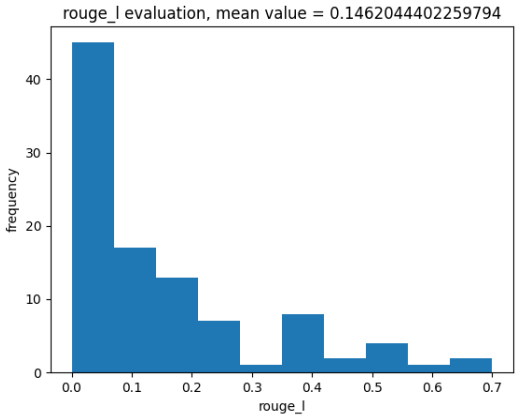



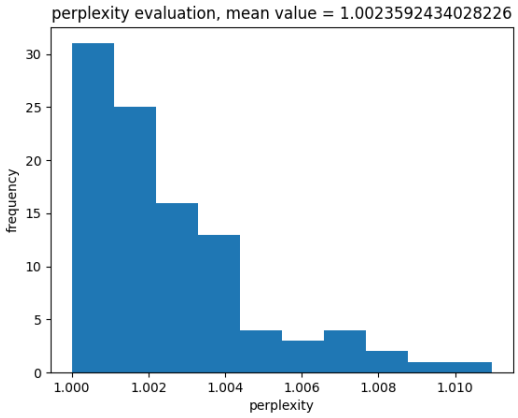

在数据集 sft_data 中训练过的随机 100 条数据上：






在数据集 sft_data 中未训练的随机 100 条数据上：




**在测试集-day1.jsonl上的表现**


```
{"question": "金火凌日的周期是多少？", "answer": "一个全程式计算得到金火凌日周期的值。"}
{"question": "CPU的时钟频率通常由什么决定？", "answer": "时钟频率的量度单位是赫兹（Hz）。"}
{"question": "羊侃的字是什么？", "answer": "方脑壳又名白脸、白鼠"}
{"question": "拜仁慕尼黑二队的主场在哪里？", "answer": "新加坡"}
{"question": "俄土战争（1676年-1681年）是在哪些国家之间发生的？", "answer": "双方于1681年签署巴赫奇萨赖和约，奥斯曼帝国承认沙皇俄国对第聂伯河左岸地区的统治。"}
{"question": "多面体泡沫是如何形成的？", "answer": "气／液分散体系中的球体泡沫是气泡为较厚的液膜所隔开，且为球状的泡沫。"}
{"question": "2007年法国网球公开赛男子单打的冠军是谁？", "answer": "塞尔维亚新秀安娜·伊万诺维奇"}
```

{"question": "隋朝时期，中书侍郎被改名为什么？", "answer": "礼部侍郎是中国古代的官职，担任礼部的副长官。该职务在隋朝时主管礼部司的相关事务，唐朝时成为礼部的副官，明清时期设有左右二名侍郎，清代分为满汉四个侍郎。"}
{"question": "2008年夏季奥林匹克运动会的比赛项目中，田径项目有多少个小项？", "answer": "2008年夏季奥林匹克运动会共有28个大项和302个小项。"}
{"question": "疣囊苔草的分布地区包括哪些国家或地区？", "answer": "分布在中国大陆的湖南、江西、贵州、广东、福建、浙江、湖北、湖南、江苏、湖北、四川、湖南等地。"}
{"question": "陕西省的地理位置包括哪些主要区域？", "answer": "全省区域，包括黄河、河和河湖水系。"}
{"question": "《水浒传》中董超与薛霸的角色有什么特点？", "answer": "他被视为一位董超和薛霸的角色，让自己的角色都是由董超和薛霸的角色扮演的。"}
{"question": "赫拉克利亚战役的结果如何？", "answer": "双方于1681年签署巴赫奇萨赖和约，奥斯曼帝国承认沙皇俄国对第聂伯河左岸地区的统治。"}
{"question": "什么是旋转不变性？", "answer": "在物理学中，旋转不变性意味着物理系统的性质不受空间取向的影响。根据诺特定理，如果物理系统的作用量具有旋转不变性，则角动量守恒。"}
{"question": "扭肚藤的分布地区包括哪些国家或地区？", "answer": "分布于中国大陆的云南、广西、贵州等地。"}
{"question": "探春花的生长环境是什么？", "answer": "生长于海拔1,500米至2,500米的地区，多生长于山坡林下、林缘或草地。"}
{"question": "短鲔的背鳍有几枚硬棘和软条？", "answer": "背鳍软条15至15枚；臀鳍软条13"}
{"question": "拉斐尔·纳达尔在2007年法国网球公开赛男子单打决赛中的比分是多少？", "answer": "1-0"}
{"question": "Google音乐在哪一年宣布关闭？", "answer": "2011年"}
{"question": "2009年至2010年英格兰足球甲级联赛有多少支球队？", "answer": "20支球队"}

# 四、模型效果分析

## 1. 预训练模型效果分析

**字符**

均为汉字，没有乱码，不出现eot等分割字符

**重复**

与 temperature 设置有关，较低时会出现重复，已经在输出时处理

**内容**

之后1-2句基本与 prompt 相关，但再后面的内容有可能不太相关

**输出时长**

服务器上 < 1s / 256 tokens，在合理范围内

## 2. 微调模型效果分析

**字符**

均为汉字，没有乱码，不出现eot等分割字符

**重复**

基本没有重复，temperature 较低时也不出现重复

**内容**

基本与 prompt 相关，但有时会答非所问，经常给出错误答案

但是在数据集 A.jsonl 及变种问题上表现良好，说明数据增强取得效果。可以以此类推到整个数据集，即对于每个条目的每句话都生成 5-10 个问题，最后得到约 1000000 个问答对，用这些问答对微调应该能够取得更好的效果和泛化性。但由于资源限制，难以在短时间内生成这么多问答对。

**输出时长**

服务器上 < 1s / 1 prompt，在合理范围内

# 五、总结

本实验介绍了 MiniGPT 的预训练和微调过程。预训练数据集为中文 wiki 百科子集，微调数据集通过问答生成和数据增强形成，最终包含约4.12万个问答对。微调后的模型在特定数据集上表现良好，但在未知问题上仍有答非所问的情况。模型在字符处理和输出时长上表现稳定，但回答准确性有待提升。未来可通过增加问答对数量进一步提升模型性能。