这个微调是在阿里通义千问7B模型上进行的，目前只训练了50条数据。因此效果不是特别明显，还需要后期进行扩展。用于验证的数据总共30条，都显示在excel文件中，包括了主诉，参考模板（也就是我们假设的现病史答案），原始模型生成模板，微调后模型生成模板（训练后的模型生成的现病史）以及一些列评判分数。

**其中平均分数是微调后的模型BLEU和ROUGE分数平均值，BLEU和ROUGE这两个值通俗理解就是模型回答和正确答案之间的相似度**。

目前由于数据集不完善，且验证集中的现病史不一定完全正确，因此数据仅供参考。

接下来介绍这两个分数的含义：

### 1.BLEU

BLEU （Bilingual Evaluation Understudy）最初是用来评估机器翻译质量的，但它也可以用来衡量生成文本与参考文本之间的相似度。

通俗理解：

BLEU的核心思想是：生成的文本越接近参考答案，得分就越高 。它主要关注的是生成文本中是否有和参考文本一致的短语或单词。

举个例子：

参考文本：患者自述3天前开始出现发热，伴有咳嗽。

生成文本：患者说3天前发热，还有咳嗽。

如果生成文本中有“3天前”、“发热”、“咳嗽”这些关键词，BLEU就会给高分；但如果生成的内容完全偏离了参考文本，分数就会很低。

计算方法：

BLEU通过比较生成文本和参考文本之间的n-gram匹配 来打分。

1. n-gram ：指的是连续的n个词。比如“3天前”是一个2-gram，“发热”是一个1-gram。
2. 比较生成文本中的每个n-gram是否出现在参考文本中。
3. 计算匹配的比例，并对长度进行惩罚（如果生成的句子太短，分数会降低）。
4. 最终得分是一个0到1之间的数字，越接近1越好。

### 2. ROUGE是什么？

ROUGE （Recall-Oriented Understudy for Gisting Evaluation）最初是用来评估自动摘要的质量，但它也可以用来衡量生成文本与参考文本的重叠程度。

通俗理解：

ROUGE的核心思想是：参考文本中被生成文本覆盖的内容越多，得分就越高 。它主要关注的是生成文本是否抓住了参考文本的关键信息。

举个例子：

参考文本：患者自述3天前开始出现发热，伴有咳嗽。

生成文本：患者说3天前发热，还有咳嗽。

如果生成文本包含了参考文本中的“3天前”、“发热”、“咳嗽”，那么ROUGE会给高分；但如果遗漏了关键信息，比如没提到“3天前”，分数就会降低。

计算方法：

1. ROUGE通过比较生成文本和参考文本之间的重叠部分 来打分，主要有以下几种：
2. ROUGE-N ：类似于BLEU，统计n-gram的匹配情况。比如ROUGE-1看单个词的匹配，ROUGE-2看两个词的匹配。
3. ROUGE-L ：基于最长公共子序列（LCS），即找到生成文本和参考文本中最长的共同部分。
4. ROUGE-W ：加权版本的ROUGE-L，考虑了连续匹配的重要性。

最终得分也是一个0到1之间的数字，越接近1越好。