诊疗回复评测

1. 任务描述

医疗问诊对话是医疗领域中的重点场景。医生需要根据患者的问题，给出相应的回答，其中包括疾病诊断结果，以及后续治疗方案建议。能够给出专业准确的回答，是医疗模型能够大规模推广应用的前提，而对于该能力的评测则需要专业的医生标注出可用的评测数据集，最终得到可信的评测结果。

2. 任务说明

本次评测任务主要目标是针对基于真实病例生成的患者问题以及医生回答，同时参考医生给出的相关医学知识作为 context，比较模型产出的回答结果与医生回答结果的一致性。

标注数据样例：

question	患者：医生您好，我是一名71岁的男性患者。最近我晚上经常要起床上厕所，每晚大约要尿3–5次，而且尿频尿急的感觉很明显。我还发现尿线变细了，有时候还会出现尿分叉的情况。不过我并没有感觉尿费力。请问这是什么原因引起的呢？我应该如何治疗？
context ⁝⁝	匹配项0：LUTS包括储尿期症状、排尿期症状以及排尿后症状。储尿期症状包括尿频、尿急、尿失禁以及夜尿增多等;排尿期症状包括排尿踌躇、排尿困难以及排尿间断等;排尿后症状包括排尿不尽感、尿后滴沥等。有LUTS的中老年男性更容易导致阴茎勃起功能障碍(erectile dysfunction, ED)[5],且ED与LUTS的严重程度相关[6-7]。 匹配项1：【2 诊断与评估】强烈推荐的检查包括:病史问询、国际前列腺症状评分(International Prostate Symptom Score, IPSS)、体格检查、尿常规检查、血清前列腺特异性抗原(prostate specific antigen,PSA)检测、经直肠或经腹部超声检查。
answer	医生回答：您好，根据您的症状描述和年龄，很可能是前列腺增生引起的。前列腺增生是一种常见的男性疾病，随着年龄的增长，前列腺会逐渐增大，压迫尿道，导致尿频尿急、尿线变细以及尿分叉等症状。根据您的情况，我们建议您进行尿常规、前列腺特异抗原（PSA）和泌尿系统B超检查，治疗方面，可以根据您的具体情况选择合适的治疗方法，包括药物治疗、物理治疗或手术治疗等。建议您尽快到医院就诊，进行详细的检查和评估，以便制定最合适的治疗方案。
disease	前列腺增生
treatment	尿常规，前列腺特异抗原（PSA），泌尿系统B超

3.评价指标：

本次任务从两方面来评测模型回复结果：

（1） 诊断建议的准确程度

通过计算 F1 分数来评估疾病诊断及治疗方案的准确程度

Confusion Matrix		Predict Result	
		1	0
Ground Truth	1	TP	FN
	0	FP	TN

对于单条样本可以计算 TP、TN、FP、FN

精确率：P=TP/(TP+FP)

召回率：R=TP/(TP+FN)

F1=2*P*R/(P+R)

该公式可以计算得到两个 F1 指标：F1_诊断、F1_建议。

然后将两个 F1 以 2:1 的权重加权得到：

F1 = （2*F1_诊断 + F1_建议）/ 3

最终的 Macro F1 分数是通过计算各个样本 F1 分数的算数平均值得到的。

（2） 对话回复的整体质量

通过计算 Rouge-L 来评估对话回复整体和医生回复的一致程度，其计算公式如下：

P = LCS(S1, S2)/len(S1)

R = LCS(S1, S2)/len(S2)

Rouge-L = 2PR/(P+R)

其中 S1 为模型输出回复文本，S2 为数据集中医生回复文本，LCS 为 S1 和 S2 的最长公共子序列，len(S)为 S 的长度。