第三届知识图谱大赛问答测评说明

一、测评数据集概述

Question_KG_Competiton3.csv

提供的皮肤科流派数据集共有309条，其中选择题为单选题，共268条；问答题41条。

文件共有三列，分别是：

- id：题目的唯一id编号
- type：题目类型
- 题面

题面样例如下：

> 选择：张志礼对于急性湿疹和慢性湿疹的辨证有何不同？A急性湿疹多属阳证，慢性湿疹多属阴证；B急性湿疹多属阴证，慢性湿疹多属阳证；C两者辨证相同；D无法判断。

> 问答：凉血五根汤在治疗血热发斑和热毒阻络证候时，哪些药材起到了凉血活血的作用，并简述其方解？

二、选择题评分

选择题均为单选题，每题都有唯一的答案。选择题的得分计算将会换算为百分制，即：

- 题目总数为 $N = 268$
- 答对的题目数量为 C
- 得分 S 为： $S = \frac{C}{268}$

如果答对了 C = 150 题，得分为：

$$S = \frac{200}{268} \approx 0.5597$$

三、问答题评分

在评估中医皮肤科问答系统生成的答案与标准参考答案之间的相似性。采用了**BLEU-4**和**ROUGE**两类常用的自动化评价指标，用以衡量生成答案在词汇和语义层面的匹配情况。

1. BLEU-4 指标

BLEU（Bilingual Evaluation Understudy）是一种基于 n-gram 的自动化评价指标，常用于机器翻译和文本生成任务。本文采用**BLEU-4**，即基于4-gram的BLEU分数，计算公式如下：

$$\mathrm{BLEU} = \mathrm{BP} \times \exp\left(\sum_{n=1}^{4} w_n \log p_n\right)$$

其中，(p_n) 表示生成答案与参考答案之间 n-gram 的匹配率，权重 (w_n) 为均匀分配的 ((0.25, 0.25, 0.25, 0.25))。长度惩罚因子（Brevity Penalty, BP）用于避免生成答案过短而影响得分。BLEU-4 能够捕捉词汇和短语的精确匹配，适合评估较长的中医皮肤科问答生成结果。

2. ROUGE 指标

ROUGE（Recall-Oriented Understudy for Gisting Evaluation）是一类基于召回率的文本相似性指标，广泛用于文本摘要、问答生成等任务的评价。本文采用以下三种ROUGE指标，用于评估生成答案与参考答案在不同粒度上的匹配程度：

- **ROUGE-1**：基于 unigram（词级别）的精确匹配，用于衡量生成答案和参考答案的词汇覆盖率。
- **ROUGE-2**：基于 bigram（词对）的匹配，用以衡量相邻词汇的精确匹配情况。
- **ROUGE-L**：基于最长公共子序列（Longest Common Subsequence, LCS）的匹配，用于捕捉生成答案与参考答案在语序和结构上的一致性。

ROUGE 指标分别计算**精确率（Precision）**、**召回率（Recall）和F1值**，本文采用 F1 值作为主要评价标准，计算公式如下：

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

其中，ROUGE–1 适合评估生成答案的整体词汇匹配情况，ROUGE–2 则能捕捉答案中的搭配合理性，而 ROUGE–L 通过最长公共子序列评估答案的连续性和流畅性，适用于中医皮肤科问答中复杂语义的评估。

3.问答题综合得分计算及权重设置

为了全面评估中医皮肤科问答系统生成的答案质量，测评在计算 BLEU–4 和 ROUGE 系列指标的基础上，进一步通过加权方式计算每道问答题的**综合得分**。具体公式如下：

$$\text{composite\_score} = (\text{BLEU-4} \times 0.4) + (\text{ROUGE-1} \times 0.2) + (\text{ROUGE-2} \times 0.2) + (\text{ROUGE-L} \times 0.2)$$

BLEU–4 主要衡量生成文本中 4–gram 的匹配率，能够捕捉到中医术语和关键知识表达的精确表达。因此，BLEU–4 被赋予了0.4的权重（0.4），能够确保在评分时更加关注模型是否生成了准确的中医术语和重要的医学信息。

ROUGE–1、ROUGE–2 和 ROUGE–L 各自分配 0.2 的权重，着重评估生成答案的词汇和语义结构的合理性。中医问答往往涉及较为复杂的语义表达，ROUGE 系列指标能够有效捕捉生成答案与参考答案在句法和语义上的匹配程度。

4.问答题整体性能计算

通过加权的方式计算综合得分后，对所有行的数据取平均值，计算问答题最终得分。

四、综合评分

在评估中医皮肤科问答系统生成的答案与标准参考答案之间的相似性。采用了**BLEU–4**和**ROUGE**两类常用的自动化评价指标，用以衡量生成答案在词汇和语义层面的匹配情况。

在评估中医皮肤科问答系统时，本文对问答题和选择题分别进行评分，并计算最终的**综合评分**。综合评分的公式如下：

$$\text{final\_score} = (\text{选择题得分} \times 0.7) + (\text{问答题得分} \times 0.3)$$

其中：

- **选择题得分** 是系统在选择题部分的正确回答比例，直接衡量系统在固定选项题目中的表现。
- **问答题得分** 是基于 BLEU–4 和 ROUGE 系列指标的加权平均得分，反映了系统生成答案与参

考答案之间的语义和准确度匹配。

五、Baseline

在基准参考中，选用了GLM-4-Plus和qwen2 7b Chat模型。

在选择题和问答题中使用的Prompt如下，供选手参考：

选择题提示词 Python

```
"""

    要求最后给出的答案：
        1、直接输出答案，如：A或B等，不需要给出其他任何解释、不需要选项后面的中文。
        2、根据经验进行作答，选择最确定的答案；
        3、直接输出选项的字母，不要有任何多余输出。
"""
```

问答题提示词 Python

```
"""
    你是中医赵炳南流派皮肤科的专家，以下是一道中医皮肤科的问答题。请根据题面，给出答案与分析。
    要求最后给出的答案：
        1、能够逐步推理、必要时可分点论述，以更全面展现中医诊疗知识的推理过程；
        2、结合中医专业知识，根据经验进行作答；
        3、直接输出答案，不需要输出任何系统级的提示语，如：根据xxx生成答案、综合xxx答案等表述。
"""
```

测试结果如下：

模型/题目类型	选择	问答	**得分**
GLM-4-Plus	0.6343	0.3775	0.5573
qwen2 7b	0.6381	0.3781	0.5601

六、测试集提交样例

主办方将测试集发放给选手后，选手使用构建的问答系统进行回答。

提交的的测试集有四列，分别是：id　　type　question　answer，样例如下（答案仅做演示使用）：

id	type	question	**answer**
1	选择	赵炳南对于系统性红斑狼疮的治疗观点主要是什么？A以清热解毒为主；B以燮理阴阳，调和气血为主；C以健脾渗湿法为主；D以散风止痒为主。	E
270	问答	凉血五根汤在治疗血热发斑和热毒阻络证候时，哪些药材起到了凉血活血的作用，并简述其方解?	以下药材起到了……

- **提交形式：以“QA_队长名_单位”命名，以csv形式提交，如：QA_张三_中国中医科学院.csv。**
- 选择题的回答只有大写字母。
- 不可改变id列与题目的对应关系。