# 评测标准

为了更好地评测模型的表现，我们设计了一套针对本次赛题的评测标准，本文档将分为选择题和问答题两部分来介绍此标准，在文末，我们给出了一些样例以帮助大家理解。

## 选择题

本赛题选择题有单选和多选两种形式，此次赛题中，一道单选赋2分，一道多选赋4分。

选择题的评测方式可形式化如下：

$$score_{choice} = \sum_{i=1}^{N} weight_i \cdot X_i = \sum_{i=1}^{N} \mathbb{I}(A_i \subset T_i) \frac{|A_i|}{|T_i|} \cdot X_i$$

其中$N$表示的是测试集中选择题的总数，$X_i$表示第$i$道单选题/多选题的单题分值，单选题对应为2，多选题对应为4，$weight_i$表示的是第$i$道题的选项命中率，$A_i$表示模型回答该题对应的选项集合，$T$表示该题的正确选项集合。

上述公式得含义可通俗转述为下：单选题选对得分，选错不得分。多选题按选对比例赋分，即漏选得相应比例分，全选得满分，多选、错选均不得分。

## 问答题

本赛题中每个问答题赋10分。

我们用关键词匹配和相似度计算两个维度来度量一个问答题答案的好坏。对于每一个答案，其得分计算形式化如下：

$$score_{QA} = weight_{keywords} \cdot （score_{keywords} + weight_{sim} \cdot score_{sim}）$$

接下来我们分别介绍关键词匹配和相似度计算两个维度的定义以及计算原理。

### 关键词匹配

问答题的正确答案形式上可以有很多种，我们将所有正确答案中**一定包含**的**关键结果**对应的词，称为该题目/答案的关键词。包含关键词是一个答案为正确的必要条件。

一个问题的答案对应的关键词可能有多个，他们组成的集合记为$Set_{keywords}$。对于每一个关键词，若其是模型所给答案的子字符串，我们就称该关键词匹配，关键词的匹配个数记为$N_{match}$，那么上述公式中关键词系数如下计算：

$$weight_{keywords} = \frac{N_{match}}{|Set_{keywords}|}$$

### 相似度计算

形式不同的答案能同时正确回答一个问题，是因为他们在语义上是相似的。我们为每一个问题准备了多个答案，我们先将模型答案进行embedding，再将这些答案分别进行embedding并与前者计算相似度，取最高的作为相似度系数，即：

$$weight_{sim} = max(sim(T_1, A), \cdots, sim(T_n, A))$$

其中$T_1, \cdots, T_n$为我们准备的多个答案，$A$表示模型给出的答案。

此外需要注意的是：

- 公式中$score_{keywords} = 5, score_{sim} = 5$
- 并不是每个题目的答案都可以选择出完备的关键词，因此对于选不出关键词的题目，我们直接计算相似度，即

$$score_{QA} = weight_{sim} \cdot score_{sim}$$
$$score_{sim} = 10$$

# 样例

## 单选题

| 模型答案 | 标准答案 | 得分 |
| --- | --- | --- |
| A | A | 2 |
| A | B | 0 |

## 多选题

| 模型答案 | 标准答案 | 得分 |
| --- | --- | --- |
| ABC | ABCD | 3 |
| AB | AB | 4 |
| ABC | AC | 0 |
| ABCD | ABC | 0 |
| AB | ABCD | 2 |

## 问答题

| 问题 | 关键词 | 模型答案 | 标准答案 | 得分 |
| --- | --- | --- | --- | --- |
| 1 + 1 = ? | 2 | 答案是2 | 答案是2 | 1 *（5 + 1 * 5） = 10 |
| 1 + 1 = ? | 2 | 答案是3 | 答案是2 | 0 *（5 + 0.8 * 5） = 0 |
| 江苏的省会是? | 南京 | 江苏的省会是南京 | 南京；江苏的省会是南京；江苏的省会是六朝古都——南京 | 1 *（5 + max（1，0.95，0.89）* 5） = 10 |
| 江苏的省会是? | 南京 | 江苏的省会是苏州 | 南京；江苏的省会是南京；江苏的省会是六朝古都——南京 | 0 |