# 测试目的

本阶段测试的主要目的是：

测试对于有解决方案的数据，对于将不同比例的数据以不同的方式放入训练集、知识库、测试集后得到的模型准确性。

测试上述方法中模型的处理时间能否满足要求。

测试上述方法中模型的显存、内存占用是否满足要求。

# 测试对象

本阶段的测试对象为两个参数不同的大模型，一是参数规模较大的700亿参数大模型，二是参数规模较小的300亿参数大模型。我们在使用同样的方法进行微调后，使用同样的方法分别对其进行测试。

# 测试方法

## 评判标准

对于系统给出的答案，我们需要对其进行打分以评估系统准确性。但是由于目前数据的性质不同，需要用不同的方法来评价大模型的预测结果。

### 对于手册中500条有解决方案的数据

这500条数据是目前我们手中质量最高的数据，其可以分别用于训练大模型、加入知识库以及测试。由于其有确切的解决方案，我们可以以大模型输出答案文本和标准答案的文本匹配程度作为该条数据测试效果的打分。

### 对于2000条无解决方案的数据

这2000条数据的质量较差，由于没有标准的解决方案，我们想要对其进行打分必须借助外界的帮助。一是请专业的工程师为大模型给出的回答进行打分。二是使用gpt4对答案进行打分。在本部分数据的评价中，我们认为工程师给出的答案可信度更高，对其分数的权重也应当更高。

## 具体测试方法

下面介绍具体的测试方法。

### 对于微调数据，不将其放入知识库的方法

由于我们手中的质量较高的数据量很少，经调查研究发现相比较于大的测试集，需要调高测试数据对训练数据的比例，方案如下。

方案1：按照训练数据 : 测试数据 = 325 : 175的比例测试。

方案2：按照训练数据 : 测试数据 = 300 : 200的比例测试。

方案3：按照训练数据 : 测试数据 = 350 : 150的比例测试。

方案4：按照训练数据 : 测试数据 = 375 : 125的比例测试。

方案5：按照训练数据 : 测试数据 = 400 : 100的比例测试。

### 对于微调数据，将其部分放入知识库的方法

五种比例仍同2.1。对于五种比例的方案，分别再测试将微调数据的25%、50%、75%放入知识库中的模型精度。

除了上面的区别，这一类中我们还可以控制以下两个变量：

1. 控制是否将测试数据放入知识库。
2. 控制大模型在知识库中查询的温度T（0, 0.25, 0.5, 0.75）