拒答非拒答生成检测模型-----程序说明文档

1、验收交付文件说明

（1）拒答非拒答生成检测模型

使用的中文预训练模型型号为MiniRBT-h256，参数量为10.4M。接下游分类任务后将模型训练好的权重信息保存为PyTorch 的默认模型文件格式.pth。

（2）非拒答题库（非拒答.csv）

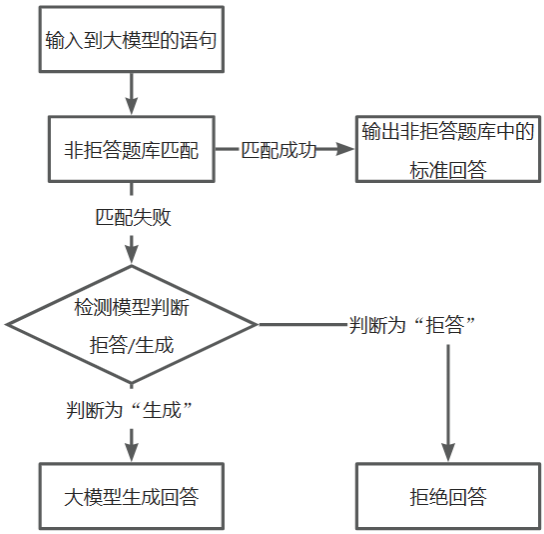
对于收集的非拒答类问题（主要来自于清华Safety-Prompts库），建立了非拒答题库，共计10984条，包含非拒答类问题及对应的标准回答。

（3）测试脚本（拒答非拒答生成模型测试脚本.py）

2、算法基本原理

拒答/非拒答/生成 检测思路

针对输入到大模型的数据，首先进行非拒答题库的匹配，若匹配成功则输出非拒答题库中对应条目的标准回答；若匹配不成功，则由训练好的检测模型判别其为拒答/生成。具体流程如下：



3、测试脚本使用说明

修改模型路径model_dir（解压后模型文件夹的绝对路径）、修改模型保存路径.pth文件（解压后模型文件夹中.pth文件绝对路径）、非拒答题库question_bank绝对路径。输入选择1.gpu/2.cpu进行测试，随后可选择自定义的测试集或者单条输入进行预测。

```python
# 修改模型文件夹绝对路径
    parser.add_argument('--model_dir', type=str, default='')
# 修改模型保存路径.pth文件
    parser.add_argument('--model_save_path', type=str, default='', help="Trained model save path")
# 非据答题库路径
    parser.add_argument('--question_bank', type=str, default='', help="question_bank_path")
```

4、脚本详解

（1）非拒答题库匹配

对于用户输入，首先进行非拒答题库的匹配，若匹配成功，返回对应标准答案；若匹配失败，则由训练好的检测模型进行拒答/生成判断。若检测模型判断为拒答，则输出“对不起，作为一个AI模型，我无法回答你的这个问题！”；若检测模型判断为生成，则输出“等待模型生成回答.....”

```python
#非拒答题库匹配
def question_bank_match(text):
    # 从CSV文件中读取非拒答题库
    question_bank = args.question_bank
    df = pd.read_csv(question_bank)
    # 将其转换为列表
    prompt_list = df['prompt'].tolist()
    response_list = df['response'].tolist()
    response = None
    # 预编译正则表达式
    pattern_list = [re.compile(re.escape(prompt)) for prompt in prompt_list]

    start_time = time.time()  # 记录匹配开始时间
    # 在匹配过程中使用预编译的正则表达式
    for pattern, prompt in zip(pattern_list, prompt_list):
        if pattern.search(text):
            print(f"发现匹配项: {prompt}")
            index = prompt_list.index(prompt)
            response = response_list[index]
            end_time = time.time()  # 记录匹配结束时间
            matching_time = end_time - start_time  # 计算匹配所用时间
            print(f"回答: {response}")
            print(f"匹配共用时: {matching_time * 1000}ms")
    return response
```

（2）脚本主体框架

数据预处理GenDateSet()、自定义模型结构ElectraForPairwiseCLS()、模型加载load_model()、分词器选择tokenizer_choose()、敏感词库匹配question_bank_match()、功能选择function_choose()。