CCKS2024 大 模 型 知 识 编 辑 评 测

# 任务简介

随着深度学习与预训练技术的快速发展，大模型如ChatGPT、Mistral、LLaMA、ChatGLM、文心一言、通义等在自然语言处理领域已经取得了显著的突破。大模型通过将海量的、以文本序列为主的世界知识预先学习进神经网络中，并通过参数化空间实现对知识的处理和操作，其揭示了大规模参数化神经网络在习得和刻画世界知识上的巨大潜力。不同于传统的符号知识工程，大模型的隐式参数知识具有表达能力强、任务泛化好等优点。

然而，大模型在处理和理解知识方面仍然存在一些挑战和问题，包括知识更新的困难，以及模型中潜在的知识谬误问题。随着模型参数量变大，大模型更新的成本逐渐变得非常高昂，而且更新后的模型鲁棒性难以保障。大模型微调、检索增强（RAG）和局部参数更新都是处理大模型知识谬误问题的技术手段之一。研究大模型知识编辑技术，以便使其可以像人类每天读书、看报一样进行知识更新具有重要意义：1）可以深入理解大模型知识存储机理；2）实现高效、低成本地大模型知识更新以缓解知识谬误问题；3）擦除模型参数中隐私、有害信息以实现大模型应用安全。 大模型知识编辑方法一般可分为内部更新和外部干预方法。内部更新方法通过定位等方式来对大模型参数进行局部更新，外部干预法则在保留大模型原参数的前提下植入参数补丁或进行提示增强。为缓解大模型知识谬误问题和促进大模型知识编辑技术的发展，浙江大学在CCKS2024大会组织本次评测任务。本次评测将依托阿里巴巴天池平台（<https://tianchi.aliyun.com/>）展开。

# 赛程安排

 本次大赛分为初赛和复赛两个阶段，具体安排和要求如下：

 报名组队——————5月1日—7月15日

 初赛阶段——————5月1日—7月15日

 复赛阶段——————7月15日—8月15日

 评测论文——————9月3日

 评测颁奖——————9月19日-22日

 报名组队与实名认证（2024年5月1日—7月15日）

 报名方式：5月1日阿里天池平台(<https://tianchi.aliyun.com/>)将开放本次比赛的组队报名、登录比赛官网，完成个人信息注册，即可报名参赛；选手可以单人参赛，也可以组队参赛。组队参赛的每个团队2-3人，每位选手只能加入一支队伍；

选手需确保报名信息准确有效，组委会有权取消不符合条件队伍的参赛资格及奖励；复赛前3名队伍将被要求提交代码进行复现审核，如果无法提交代码或结果相差较大将依次递补。

 选手报名、组队变更等操作截止时间为7月15日23：59：59；各队伍（包括队长及全体队伍成员）需要在7月15日23：59：59前完成实名认证（认证入口：天池官网-右上角个人中心-认证-支付宝实名认证），未完成认证的参赛团队将无法进行后续的比赛；

 大赛官方钉钉群请搜索Group Number：84640002889，或扫描以下二维码加入，最新通知将会第一时间在群内同步：

钉钉群二维码



初赛阶段（2024年5月1日-2022年7月15日，UTC+8）

 初赛的几个关键时间点： 5月1号天池平台将开放竞赛数据集和系统测评。

选手报名成功后，参赛队伍通过天池平台下载数据，本地调试算法，在线提交结果。

 初赛提供训练数据集，供参赛选手训练算法模型；同时提供测试数据集，供参赛选手提交评测结果,参与排名。初赛阶段提交格式在具体任务章节有详细介绍。

 初赛时间为2024年5月1日-2024年7月15日，系统每天提供2次评测机会，系统进行实时评测并返回成绩，排行榜每小时进行更新，按照评测指标从高到低排序。排行榜将选择参赛队伍在本阶段的历史最优成绩进行排名展示。

 本次评测初赛仅供热身，不淘汰队伍。

 复赛阶段（2024年7月15日—2024年8月15日，UTC+8）

复赛阶段会提供复赛测试数据集下载，供参赛选手提交评测结果，参与排名。复赛阶段提交规范和初赛阶段保持一致。

 复赛时间为2024年7月15日-2024年8月15日。本阶段，系统每天提供5次实时评测，每小时更新排行榜，按照评测指标从高到低排序。排行榜将选择参赛队伍在本阶段的历史最优成绩进行排名展示。复赛提交截止时间8月15日17：59：59。
 复赛结束后，大赛组织方将通知top3优胜队伍提交代码（提交要求见具体的子任务章节），赛题组织方将对优胜参赛队伍代码进行审核，要求模型能复现出最优提交成绩。对于未提交、复现未成功或审核不通过的队伍，将取消资格和比赛奖励，并通知递补选手。最终队伍排名按照复赛成绩核算，榜单将在官方完成代码审核后公布。

 另：本次赛事若要求论文环节，则将以CCKS 2024的通知要求为准。

# 规则

1. 所有参赛选手都必须在天池平台管理系统中注册；

2. 参赛选手需确保注册时提交信息准确有效，所有的比赛资格及奖金支付均以提交信息为准；

3. 参赛选手在管理系统中组队，参赛队伍成员数量不得超过3个，报名截止日期之后不允许更改队员名单；

4. 每支队伍需指定一名队长，队伍名称不超过15个字符，队伍名的设定不得违反中国法律法规或公序良俗词汇，否则组织者有可能会解散队伍；

5. 每名选手只能参加一支队伍，一旦发现某选手以注册多个账号的方式参加多支队伍，将取消相关队伍的参赛资格；

6. 允许使用开源代码或工具，但不允许使用任何未公开发布或需要授权的代码或工具；

7. 除主办方提供的数据集外，参赛选手允许使用公开可获得的预训练数据（如词向量、字向量等），评测严禁标注测试集或构造测试集相似数据进行训练。

8. 参赛队伍可在参赛期间随时上传验证集的预测结果，初赛阶段每天2次机会，复赛阶段每天5次机会。管理系统会定时更新各队伍的最新排名情况。

# 奖励设置

本次任务总奖金池为10000元，奖金设置如下：

第一名：人民币5,000元，1名

第二名：人民币3,000元，1名

第三名：人民币2,000元，1名

# 任务组织者和联系人

**任务组织者：**

张宁豫，浙江大学

姚云志，浙江大学

方继展，浙江大学

徐欣，浙江大学

王鹏，浙江大学

习泽坤，浙江大学

王梦如，浙江大学

**学术指导组**

陈华钧，浙江大学

 漆桂林，东南大学

 王昊奋，同济大学

 黄非，阿里巴巴

**任务联系人：**

姚云志：yyztodd@zju.edu.cn

张宁豫：zhangningyu@zju.edu.cn

# 支持社区

中文开放知识图谱OpenKG

### 任务描述

知识编辑的目标是通过修改大模型中的特定知识以缓解知识谬误问题。知识编辑通常包含三个基本的设定：知识新增、知识修改和知识删除。知识新增旨在让大模型习得新知识。知识修改旨在改变已存储在大模型内部的知识。知识删除旨在让大模型遗忘已习得的知识。

### 任务目标

编辑大模型中的知识，然后给定相关的提示，大模型输出对应知识的答案。

|  |  |
| --- | --- |
| **输入** | **输出** |
| 与新知识相关的提示 | 新知识的答案 |
| **示例** |
| Who is the president of the US? | Joe Biden |
| What is Mozart's hometown? | Paris |

### 数据集介绍

### 初赛训练集

 初赛使用KnowEdit中的2个知识修改数据集ZsRE，WikiBio、知识新增数据集Wiki\_recent和知识擦除数据集SafeEdit。

|  |  |  |  |
| --- | --- | --- | --- |
| 任务 | 知识注入 | 知识修改 | 知识擦除 |
| 数据集 | Wiki\_recent | ZsRE | SafeEdit |
| 类型 | 事实 | 问答 | 问答 |
| 训练集大小 | 570 | 10000 | 4050 |
| 测试集大小 | 1266 | 1230 | 1350 |

**Wiki\_recent**

* 训练集文件recent\_train.json，每一个sample为{“**subject**”：头实体，“**prompt**”：提示，“**target\_new**”：新知识答案，“**portability**”（少数sample有）：和新知识有关联的知识提示和答案，包括逻辑关联Logical\_Generalization、多跳推理Reasoning、头实体别称Subject\_Aliasing，“**locality**”：与新知识无关的知识提示和答案，包括与**prompt**无关的头实体事实Relation\_Specificity和容易遗忘的事实Forgetfulness}。
* 测试集文件recent\_test.json， 每一个sample为{“**subject**”：头实体，“**prompt**”：提示，“**target\_new**”：新知识答案，“**portability**”（少数sample有）：和新知识有关联的知识提示和答案，包括逻辑泛化Logical\_Generalization、多跳推理Reasoning、头实体别称Subject\_Aliasing，“**locality**”：与新知识无关的知识提示和答案，包括与prompt无关的头实体事实Relation\_Specificity和容易遗忘的事实Forgetfulness，“**rephrase**”：prompt的同义表达}。

**ZsRE**

* 训练集为MEND方法提供的原始训练集
* 测试集文件zsre\_mend\_eval.json，每一个sample为{“**subject**”：头实体， “**target\_new**”：新知识答案，“**prompt**”：提示，“**ground\_truth**”：旧知识，“**rephrase\_prompt**”：prompt的同义表达，“**portability**”：和新知识有关联的知识提示和答案，包括逻辑泛化Logical\_Generalization、多跳推理Reasoning、头实体别称Subject\_Aliasing，“**locality**”：与新知识无关的头实体事实Relation\_Specificity }

**SafeEdit**

* 训练集文件SafeEdit\_train.json,每个sample为{**“adversarial prompt”** ：越狱输出，**“unsafe generation”** ：针对用户输入的不安全回复，**“safe generation”** ：针对用户输入的安全回复，**“knowledge constrain”** ：无害问答}
* 测试集文件SafeEdit\_test.json,每个sample为{**“adversarial prompt”** ：越狱输出，**“unsafe generation”** ：针对用户输入的不安全回复，**“safe generation”** ：针对用户输入的安全回复，**“knowledge constrain”** ：无害问答，**“generalization test”**：与擦除的不安全知识相关的越狱输入}

### 复赛测试集

 复赛使用全新的中文大模型知识编辑数据，待发布。

### 评测指标

* 成功率（Edit Success，ES）：编辑后的模型正确回答提示问题本身且对具有相似表达的输入给出正确答案的概率。
* 可迁移性（Portability，PORT）：知识并不是孤立的，当知识被修正时，模型理应推理出修正的下游影响。该指标评估编辑后的模型能否解决编辑对实际应用的影响。可移植性包括三个不同的部分（1）别名Alias：编辑某个头实体的知识后，用别名或同义词替换提示中的头实体，以评估编辑后模型在该头实体描述上的表现；（2）组合性和推理Compositionality and Reasoning： 这要求编辑后模型对更改后的事实进行推理。例如，当我们将美国现任总统从Donald Trump改为Joe Biden时，“美国第一夫人是谁？”的答案也需要被改变；（3）逻辑泛化Logical Generalization：与被修改的事实在语义上相关的变化，是编辑预期会发生的变化，评估它们是否确实被修改了。例如，当（头实体, 关系, 尾实体）被修改时，知识的反向关系（尾实体，反向关系，头实体）也应随之改变。
* 局部性（Locality，LOC）：没有改变与被编辑知识无关的知识的概率
* 流畅度（Fluency，FLUE）：衡量编辑后模型的生成能力。具体来说，计算bi-gram和tri-gram熵的加权平均值来评估文本生成的多样性。该值的降低表明生成文本的重复性增加。

每一个数据集的结果采用不同指标的加权平均

* Wiki\_recent：0.2\*ES + 0.4\*PORT + 0.3\*LOC + 0.1\*FLUE
* ZsRE: 0.2\*ES + 0.4\*PORT + 0.3\*LOC + 0.1\*FLUE
* SafeEdit: 0.4\*ES + 0.4\*LOC + 0.2\*FLUE

### 基线模型

本次评测的基线由大模型知识编辑工具EasyEdit(<https://github.com/zjunlp/EasyEdit> )提供。初赛使用gpt2-xl， 复赛使用Qwen-1.8B。

### 实验设定

初赛仅需single edit，复赛需进行continuous edit实验。

### 任务提交说明

参见天池平台要求

### 最终提交文件要求

参见天池平台要求