



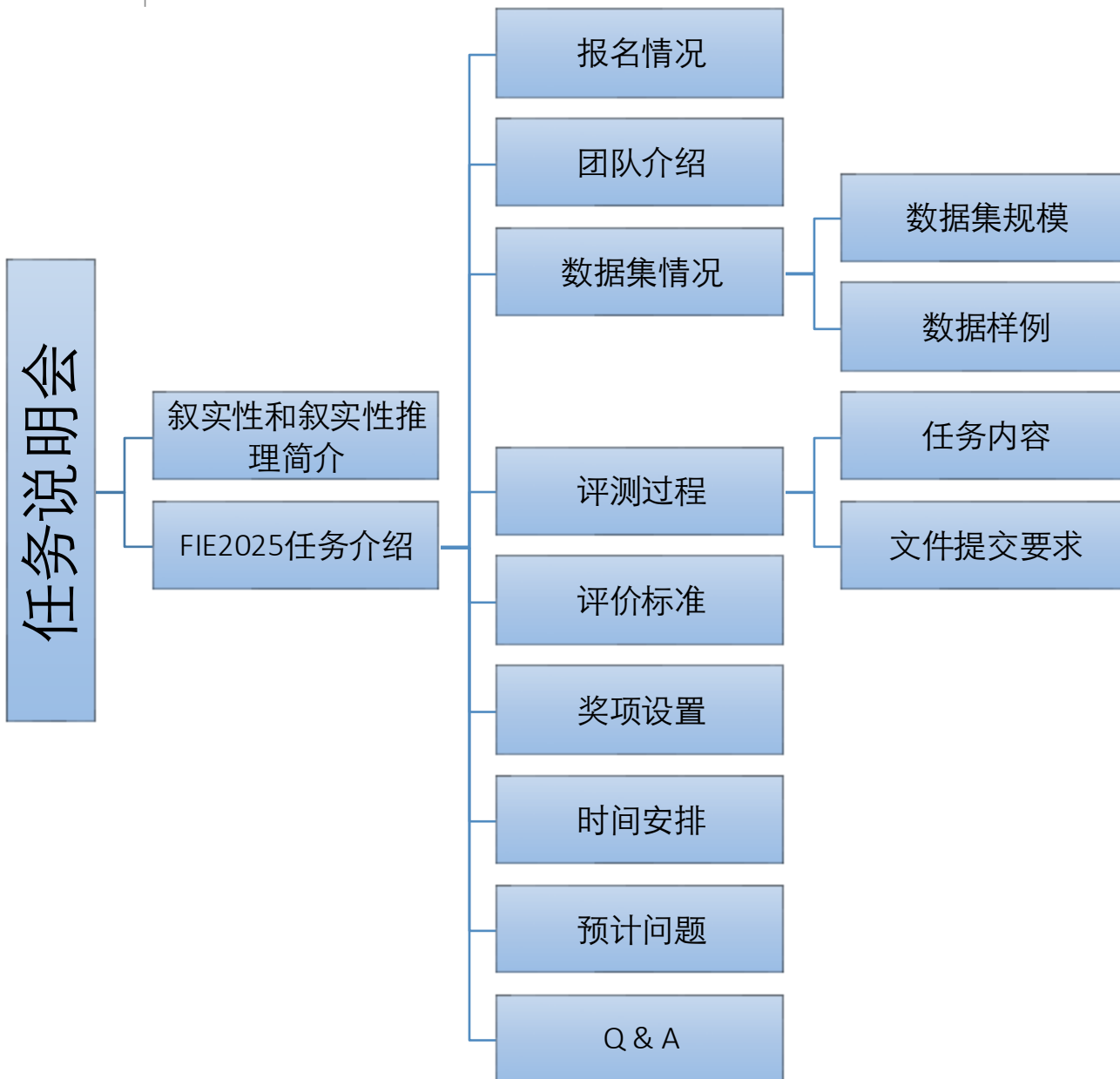
澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU



FAH
人文學院
FACULDADE DE LETRAS
FACULTY OF
ARTS AND HUMANITIES

CCL25-Eval 任务4: 第一届中文叙实性推理评测 (FIE2025) 任务说明会

2025.04.27

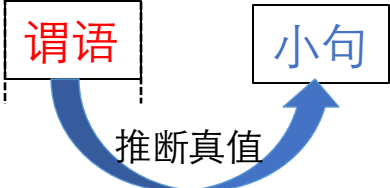


主要内容

叙实性和叙实性推理简介

- 什么是叙实性 (factivity) ?

例1 a. 小李知道小王是中国人。 / Jo knows that Bo is a Chinese.
b. 小李谎称小王是中国人。 / Jo lies that Bo is a Chinese.
c. 小李相信小王是中国人。 / Jo believes that Bo is a Chinese.



谓语 小句

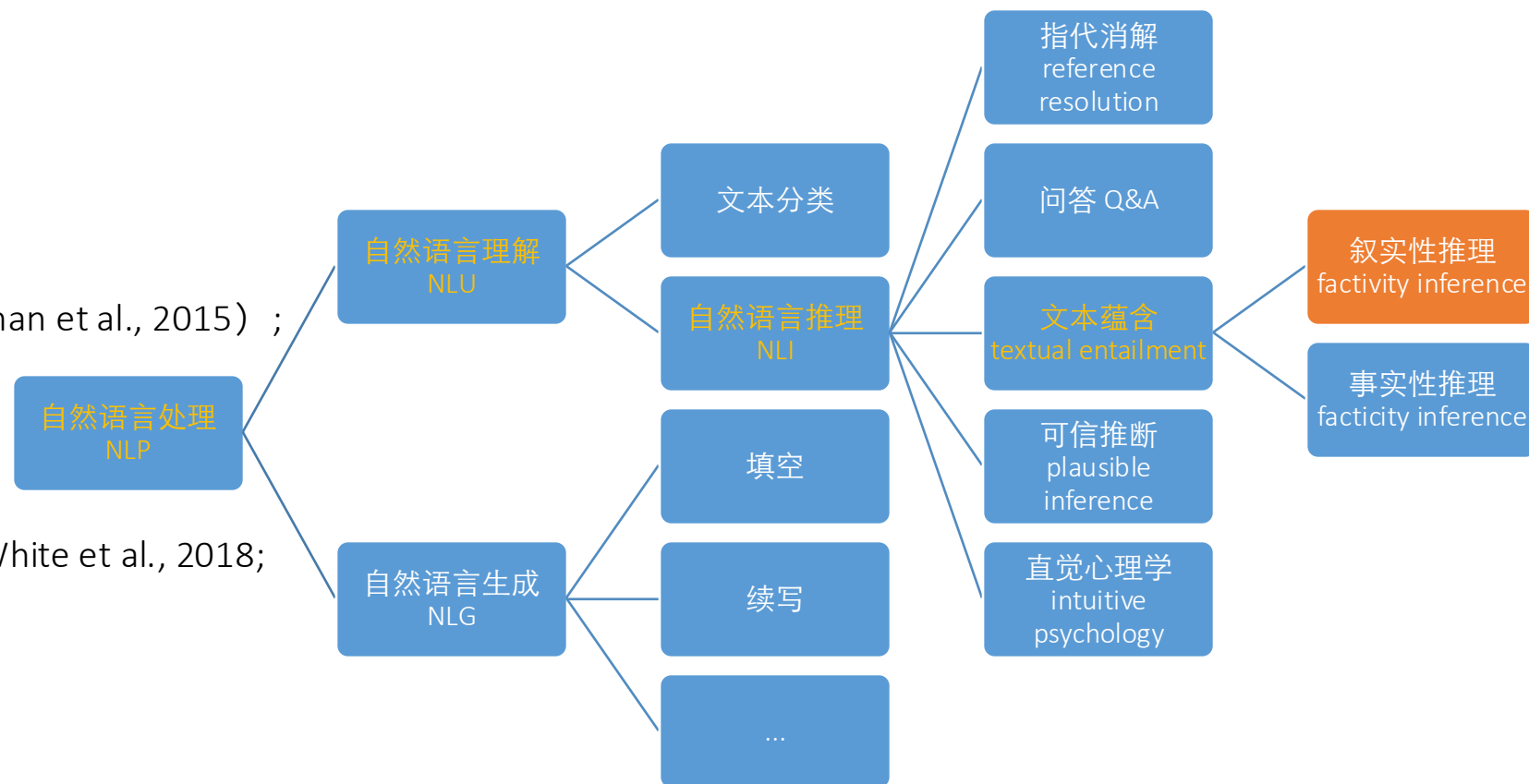
推断真值

叙实性和叙实性推理简介

- 什么是叙实性推理 (factivity inference) ?
- 在理论语言学研究 中, 叙实性是一个涉及词汇、句法、语义、语用等许多层面的复杂的研究课题。
- CL/NLP:

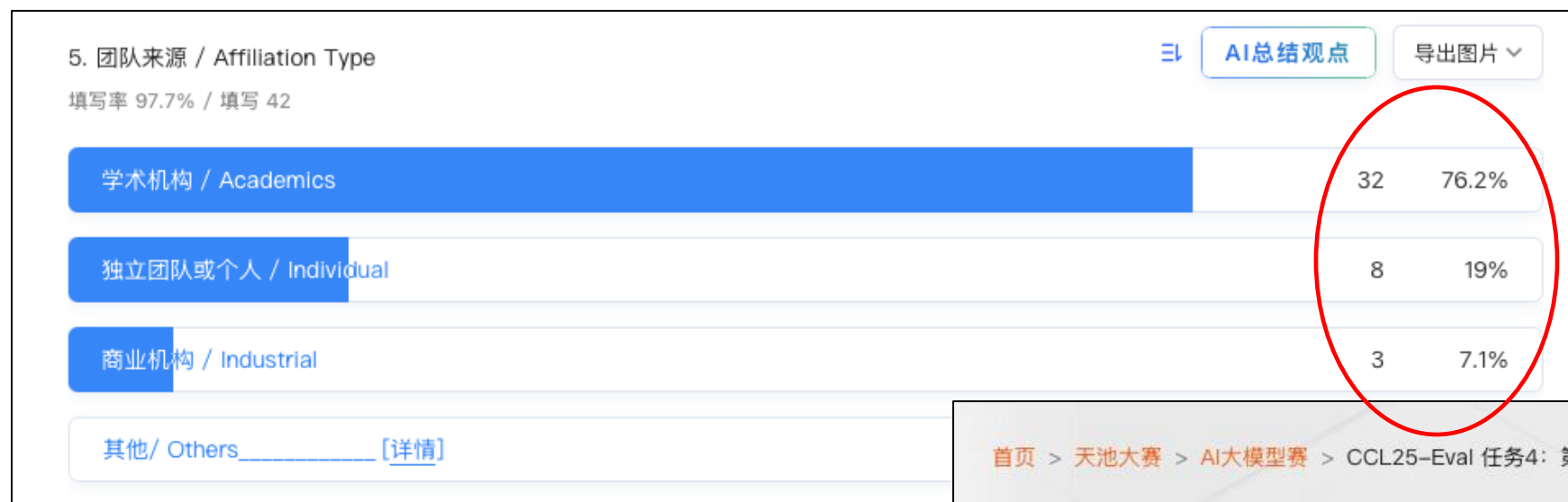
相关测试集:

GLUE (Wang et al., 2018) 的 MNLI (Bowman et al., 2015) ;
FactBank (Saurí & Pustejovsky, 2009) ;
UW (Lee et al, 2015) ;
MEANTIME (Minard et al., 2016) ;
UDS (White et al., 2016) ;
MegaVeridicality (White & Rawlins, 2018; White et al., 2018;
Rudinger et al., 2018) ;
CGC (Markowska et al., 2023) 。



FIE2025任务介绍——报名情况

- 截至4月27日，组织方共收到43份问卷报名和85个天池报名（含交叉），有30个队伍签署了参赛协议。



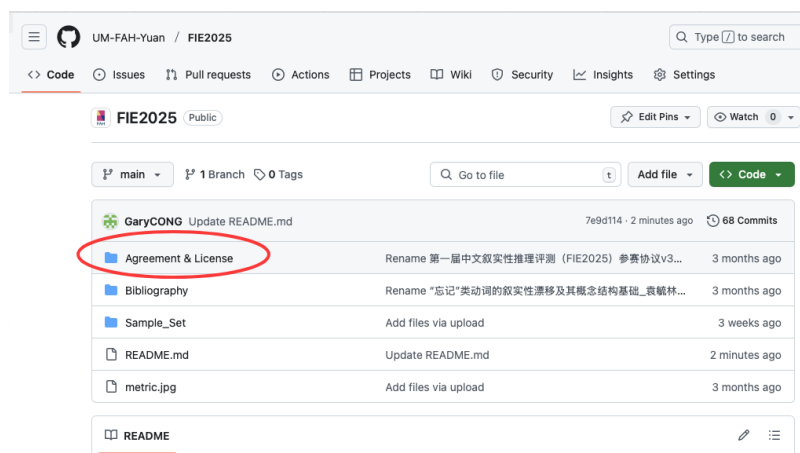
首页 > 天池大赛 > AI大模型赛 > CCL25-Eval 任务4: 第一届中文叙实性推理评测 (FIE2025)

CCL25-Eval 任务4: 第一届中文叙实性推理评测 (FIE2025)

赛事类型	AI大模型赛	奖金	¥0	团队	83	赛季2	2025-05-16	状态	进行中
举办方	TIANCHI 天池								

FIE2025任务介绍——报名提醒

- 参加本次评测既需要签署 《参赛协议》，也需要在 阿里云天池平台 上注册。
- 《参赛协议》请到Github任务网站<https://github.com/UM-FAH-Yuan/FIE2025/tree/main/Agreement%20%26%20License>中下载。
- 阿里云天池平台注册后请点击“报名参赛”。



- 由于在统计成绩时需要记录参赛队伍信息，如果在6月1日前未将签署好的参赛协议发送给组织方联系人，将无法参与最终排名。

FIE2025任务介绍——团队介绍



澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

FAH 人文學院
FACULDADE DE LETRAS
FACULTY OF
ARTS AND HUMANITIES

• 任务发起人:

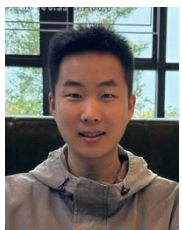


澳门大学教授
袁毓林



南京师范大学教授
李斌

• 任务联系人:



澳门大学博士生
丛冠良

• 团队成员:



吴俊潮
澳门大学博士生



寻天琦
澳门大学博士生



陈阳
澳门大学博士生



陈可盈
澳门大学硕士生



杨梓泓
澳门大学硕士生



汪月童
澳门大学本科生

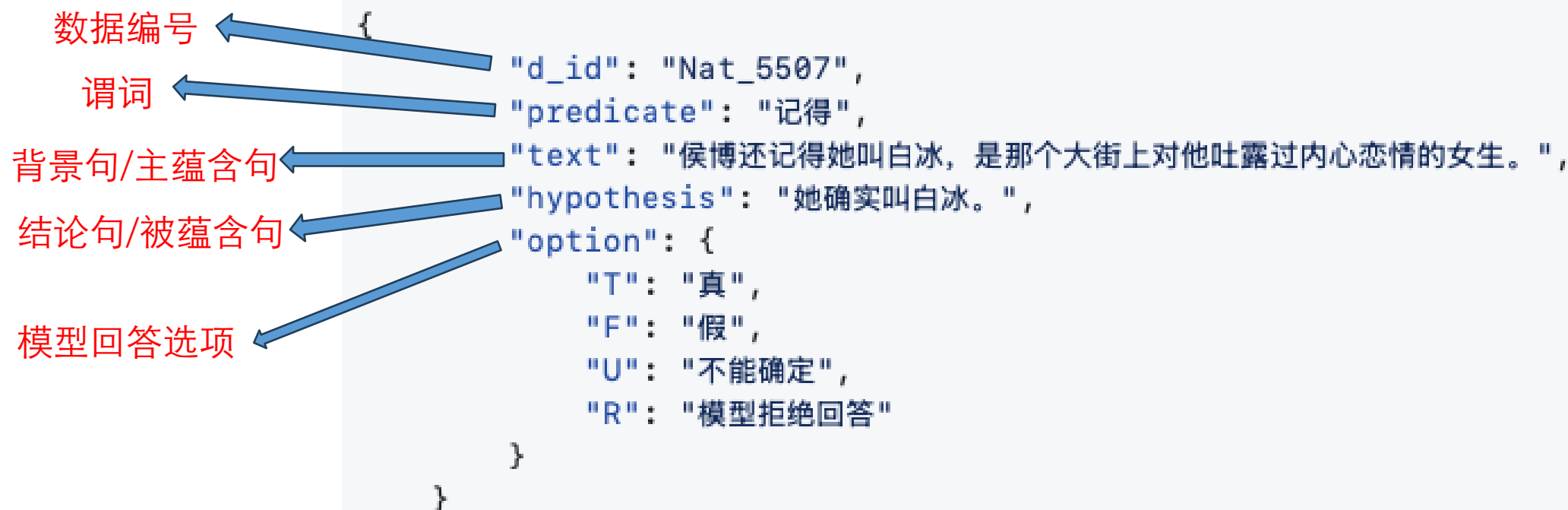
FIE2025任务介绍——数据集情况

数据规模：

- 样例集共包含**1000**条数据，分为**人造语料集**与**真实语料集**两个子集。其中，人造语料集300条，真实语料集700条。
样例集经过两次校订，目前最终版本已经上传至[Github-FIE2025](#)。
- 测试集共包含**2038**条数据，同样分为**人造语料集**与**真实语料集**两个子集。其中，人造语料集**562**（560+2）条，真实语料集**1476**（1440+36）条。
- 评测集中，有38条与“**假装**”“**装作**”有关的句子（人造2+真实36），这些句子只用于收集回答结果，无论模型如何作答，都**不计入最终成绩**；其余2000条正常计分。

FIE2025任务介绍——数据集情况

数据样例1：以真实语料集数据“Nat_5507”为例



数据样例2：以人造语料集数据“Art_0272”为例



- 若根据"text"判定"hypothesis"为真，意味着根据背景句推出结论句的真值为“真”，此时请输出“T”；
- 若根据"text"判定"hypothesis"为假，意味着根据背景句推出结论句的真值为“假”，此时请输出“F”；
- 若根据"text"不能判定"hypothesis"的真假，意味着根据背景句推出结论句的真值为“不能确定”，此时请输出“U”。
- 若模型拒绝回答，请输出“R”（出现此情况建议修改prompt）。

FIE2025任务介绍——评测过程

任务内容：

- 1. 自行选定若干大型语言模型（型号不限）。
- 2. 不微调赛道：利用测试集数据编制提示词，并以API访问的方式逐条发送给被试模型；（“不微调”指不修改权重等模型参数，只使用现有模型）
微调赛道：利用样例集数据对模型进行微调，并收集微调模型对测试集数据的回答。（“微调”指修改现有模型的参数，使用自制模型）
- 3. 要求模型仅以“text”字段值为依据来判断“hypothesis”字段值的真值情况（真/假/不能确定），记录模型的返回结果。
- 4. 最终将结果整理为JSON格式的数据文件。（可使用代码来统一提取结果，但不能人工修改）

FIE2025任务介绍——评测过程

最终提交的文件共3种：

- 1. **技术报告（上传至OpenReview）**：参与评测必须在OpenReview上提交技术报告，不提交技术报告的队伍成绩将不会被认可。
- 2. **测试结果文件（上传至天池页面）**：向天池平台提交1个JSON文件，其中将人造语料集和真实语料集的测试结果保存在一起。
- 3. **代码文件（保存备用）**：提交测试过程中使用的所有python代码文件，一个功能一个.py文件，并对其关键部分进行注释。
注意：所提交的代码需要能够完整复现测试及答案提取过程，否则无法认定成绩。

FIE2025任务介绍——评测过程

技术报告要求：

- 1.报告可由中文或英文撰写。
- 2.报告统一使用CCL 2025的论文模板。
- 3.报告正文不得超过4页（2左右，说明所使用的技术方法即可），参考文献页数不限。
- 4.报告应至少包含以下四个部分：模型介绍、评测结果、结果分析与讨论、参考文献。
- 5.技术报告通过OpenReview提交（<https://openreview.net>）（需提前注册网站账号）。



FIE2025任务介绍——评测过程

测试结果文件的数据格式要求：

- 1. 输出文件须为JSON格式，其中每条数据只需包含"d_id"和"answer"两个字段即可。
- 2. 由于所有题目都是单选题，一条数据的"answer"处只允许填写一个值。
- 3. 禁止对模型回答进行人工修正；
禁止人工诱导模型改变答案，如“你说的不对，正确答案应该是...”。
- 4. 测试结果文件的数据格式样例：

```
{  
    "d_id": "Nat_0001",  
    "answer": "T"  
}
```

FIE2025任务介绍——评测过程

- 5. 请在天池平台“提交结果”页面中点击“提交结果”按钮提交结果文件。
- 6. 天池平台“提交结果”页面将于**5月3日**开放，届时队伍可利用赛前的3天时间上传test文件以测试提交页面是否存在bug，如发现bug请及时联系组织方。
- 7. **务请看清赛道，勿在错误赛道提交文件！**



FIE2025任务介绍——评价标准

- 本次评测采用**总正确率**作为评价指标：

$$\text{total_acc} = \frac{\text{correct_art} + \text{correct_nat}}{\text{total_art} + \text{total_nat}}$$

- 其中，total_acc 为总正确率；
correct_art 为人造语料集中模型回答正确的数据量；
correct_nat 为真实语料集中模型回答正确的数据量；
total_art 为人造语料集中的数据总量；
total_nat 为真实语料集中的数据总量。
- 评测集中有38条与“**假装**”“**装作**”有关的句子（人造2+真实36），这些句子只用于收集回答结果，无论模型如何作答，都**不计入最终成绩**；其余2000条正常计分，也即 $\text{total_art} + \text{total_nat} = 2,000$ 。
- 微调赛道与不微调赛道的数据分开计算。

FIE2025任务介绍——奖项设置

任务奖项设置：

- 本届评测将设置一、二、三等奖，奖项按总正确率从高到低颁发（奖金待定）。
- 其中，一等奖0-1名，二等奖0-2名，三等奖0-3名。
- 微调赛道与不微调赛道分开排名。
- 荣誉证书将由中国中文信息学会颁发。

FIE2025任务介绍——时间安排

重要时间节点：

- 2025年5月3日：天池平台“提交结果”页面开放（用以测试提交程序）；
- 2025年5月6日-12日：组织方公布评测集；
- 2025年5月20日：参赛队伍在OpenReview中提交技术报告，用于审核；
- 2025年5月下旬：组织方对技术报告开展集中评审，提出修改意见；各参赛队伍修改技术报告；组织方将推荐排名靠前的队伍将技术报告扩写成论文投稿。
- 2025年6月1日：组织方公布中文叙实性推理评测结果，同时向工作坊提交任务综述。

完整时间安排请见Github任务网站<https://github.com/UM-FAH-Yuan/FIE2025/tree/main>

FIE2025任务介绍——预计问题

预计问题：

- 1. 模型拒绝回答——评测过程中超半数队伍反馈则排除此题（题目总数-1）。
- 2. 建议关闭深度推理。
- 3. 可以选择微调与不微调两个赛道其一参加，也可同时参加。
- 4. 数据清洗困难——可用代码统一提取答案，但不能人工修改。如果模型前后回答不一致，则需要修改prompt。注意要求模型只给出结果，不要解释。
- 5. 评测过程中可在天池平台实时查看排行榜——天池排行榜不代表最终结果（不签署协议或不提交技术报告的队伍的成绩不参与最终排名）
- 6. 评测过程中如遇问题，可随时在QQ群交流，或发邮件给联系人 guanliang.cong@connect.um.edu.mo。

Q & A

感谢垂注，敬请赐教！