

# 书生二期第七课笔记

复习链接：

视频链接：[https://www.bilibili.com/video/BV1Pm41127jU/?spm\\_id\\_from=333.788&vd\\_source=cb7df683bee4f62cb6d1e0e36524b4ff](https://www.bilibili.com/video/BV1Pm41127jU/?spm_id_from=333.788&vd_source=cb7df683bee4f62cb6d1e0e36524b4ff)

课程文档：<https://github.com/InternLM/Tutorial/blob/camp2/opencompass/readme.md>

课程作业：<https://github.com/InternLM/Tutorial/blob/camp2/opencompass/homework.md>

构建一个领域的测评数据机是比较重要的事儿。





## 面向未来 拓展能力维度

评测体系需增加新能力维度，如数学、复杂推理、逻辑推理、代码和智能体等，以全面评估模型性能。



## 扎根通用能力 聚焦垂直行业

在医疗、金融、法律等专业领域，评测需结合行业知识和规范，以评估模型的行业适用性。



## 高质量 中文基准

针对中文场景，需要开发能准确评估其能力的中文评测基准，促进中文社区的大模型发展。



## 性能评测 反哺能力迭代

通过深入分析评测性能，探索模型能力形成机制，发现模型不足，研究针对性提升策略。



## 全面性



- 大模型应用场景千变万化
- 模型能力演进迅速
- 如何设计和构造可扩展的能力维度体系

## 评测成本



- 评测数十万道题需要大量算力资源
- 基于人工打分的主观评测成本高昂

## 数据污染



- 海量语料不可避免带来评测集污染
- 亟需可靠的数据污染检测技术
- 如何设计可动态更新的高质量评测基准

## 鲁棒性



- 大模型对提示词十分敏感
- 多次采样情况下模型性能不稳定

# OpenCompass 2.0 司南大模型评测体系开源历程

OpenMMLab bilibili



## OpenCompass 助力大模型产业发展和学术研究

OpenMMLab bilibili

广泛应用于头部大模型企业和科研机构



获得 Meta 官方推荐  
唯一国产大模型评测体系


These types of projects provide a quantitative way of looking at the models performance in simulated real world examples. Some of these projects include the [LM Evaluation Harness](#) (used to create the [HF leaderboard](#)), [HELM](#), [BIG-bench](#) and [OpenCompass](#).

社区支持最完善的评测体系之一  
100+ 评测集 50万+ 题目



我们如何评测大模型？

OpenMMLab




### 基座模型

- 海量数据无监督训练(Base)

中国的首都是哪里？

俄罗斯的首都是哪里？

美国的首都是哪里？




### 对话模型

- 指令数据有监督微调(SFT)
- 人类偏好对齐(RLHF)


中国的首都是哪里？

北京



### 公开权重的开源模型

- 使用GPU/推理加速卡进行本地推理



### API 模型

- 发送网络请求获取回复

04:36 / 45:36

发个友善的弹幕见证弹幕礼仪

发送

1080P 高清 选集 1.5x

客观评测与主观评测

OpenMMLab

### 客观问答题

中国的首都是哪里？

北京  
北京。  
中国的首都是北京。  
中国的首都是北京（Beijing）。  
答案是北京。

### 客观选择题

中国的首都是哪里？  
A. 北京 B. 上海 C. 香港 D. 东京

A  
A.  
北京.  
A. 北京

### 开放式主观问答

问题：写一首七言律诗，表达对龙年春节的期待

模型A	模型B
龙舞九天迎春到， 梅花映雪报年丰。 家家户户贴对联， 喜气洋洋乐融融。 团圆饺子包幸福， 亲情友情暖如春。 国家昌盛民安泰， 龙年春节喜无穷。	龙年春节将到时， 万家灯火照街衢。 瑞雪纷飞银装素， 爆竹声声乐无边。 亲人团聚笑声起， 美食佳肴满桌铺。 祈愿来年顺顺利， 龙腾虎跃福满门。

比较：A 比 B 好/差  
打分：A 4分/ B 3分

05:41 / 45:36

发个友善的弹幕见证弹幕礼仪

发送

1080P 高清 选集 1.5x



## 提示词工程(举例)

### 明确性

不好的示例：请写一篇关于人工智能的全面介绍。  
(目标过于宽泛，缺乏具体要求和细节)

### 概念无歧义

不好的示例：我想了解苹果。  
(问题内容存在歧义)

### 逐步引导

不好的示例：告诉我怎么做蛋糕。(缺少指引)

### 具体描述

不好的示例：告诉我一个笑话。(缺少具体细节)

### 迭代反馈

不好的示例：这个回答不够详细。(缺少清晰的反馈)

好的示例：请为我写一篇关于人工智能的科普文章，要求涵盖其发展历程、应用领域以及对社会的影响。(目标明确，完整覆盖了主题内容)

好的示例：请描述苹果公司的创始人史蒂夫·乔布斯的职业生涯。  
(给出了准确的概念)

好的示例：首先，我需要准备哪些材料来制作巧克力蛋糕？接下来，烘焙过程中有哪些关键步骤？(给出了思路引导)

好的示例：请创作一个关于时间旅行的幽默故事。  
(给出了具体的问题描述和范围)

好的示例：你能更详细地解释人工智能能在医疗诊断中的应用吗？(明确清晰的修改建议)

## 提示词工程(举例)

### 小样本学习

### 思维链技术

#### 文本分类

示例1: "苹果发布新款iPhone" - 主题：科技  
示例2: "美国大选结果揭晓" - 主题：政治  
示例3: "全球气候变化影响报告" - 主题：环境  
问题："科学家发现新的抗癌药物" - 主题：

#### 情感分析

示例1: "我今天非常开心！" - 情感：正面  
示例2: "这个产品真是糟糕透了。" - 情感：负面  
示例3: "天气预报说今天会下雨。" - 情感：中性  
问题："我对这个结果感到失望。" - 情感：

#### 机器翻译

示例1: "你好" - "Hello"  
示例2: "我需要帮助" - "I need help"  
示例3: "这个餐厅的食物很美味" - "The food in this restaurant is delicious"  
问题：我今天遇到了一个老朋友。

#### 直接提问

问：自助餐厅有23个苹果。如果他们用了20个来做午餐，并买了6个，他们现在有多少个苹果？

自助餐厅现在有49个苹果。  
回答错误

#### 思维链

问：自助餐厅有23个苹果。如果他们用了20个来做午餐，并买了6个，他们现在有多少个苹果？**让我们一步一步思考这个问题**

好的，我们来一步一步解决这个问题。1. 自助餐厅原本有23个苹果。2. 他们用了20个苹果来做午餐，所以我们需要从23个中减去20个苹果： $23 - 20 = 3$ 个苹果。3. 然后，他们又买了6个苹果，所以我们需要在剩下的苹果数量上加上这6个： $3 + 6 = 9$ 个苹果。所以，自助餐厅现在有9个苹果。  
回答正确

## 模型输入



30%位置 → 小明在上海人工智能实验室实习



问题：小明在上哪里实习？

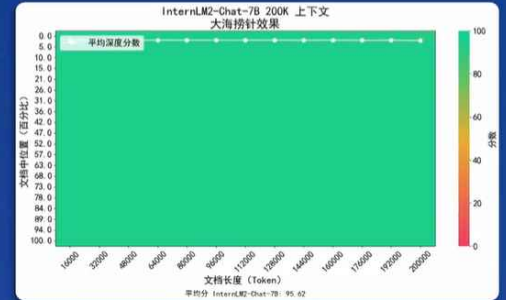
文档总长度200K

## 模型输出

回答：上海人工智能实验室

## 能力分析与性能示意

- 指令跟随能力
- 长文本建模能力
- 信息抽取能力



# 汇集社区力量：工具 - 基准 - 榜单 三位一体

提供高时效性  
高质量评测集



发布权威榜单  
洞悉行业趋势

支撑高效评测  
支持能力分析



致力于探索最先进的语言与视觉模型，为工业界和研究社区提供全面、客观、中立的评测参考

有言头部企业商业模式(热门开源模型)尚未对公众开放的内源模型

有需求即合商业模式(热门开源模型)| 尚未对公众开放的内源模型

[illegible]

教育头条公众号金程教育 | 热门课程 | 尚未对公众开放的内训精英

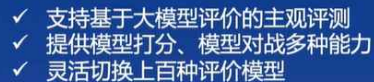
● 最新分類法

OpenMMLab OpenCompass 司南

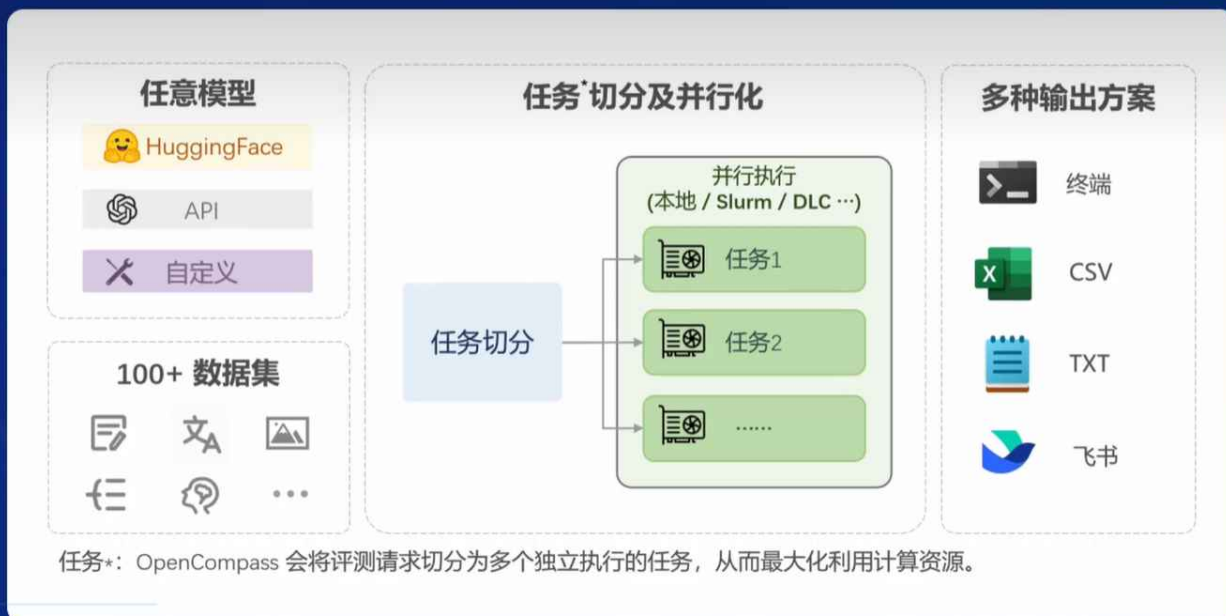
- ✓ 提供多种数据污染检测方法
- ✓ 支持包括GSM-8K, MMLU等主流数据集上的污染检测



- ✓ 支持1M长度大海捞针测试
- ✓ 支持多个主流长文本评测基准







## CompassKit : 大模型评测全栈工具链



### VLMEvalKit 多模态评测工具

一站式多模态评测工具, 支持主流多模态模型和数据集, 助力社区比较不同多模态模型在各种任务上的性能。



### Code-Evaluator 代码评测工具

提供基于 docker 的统一编程语言评测环境, 确保代码能力评测的稳定性和可复现性。

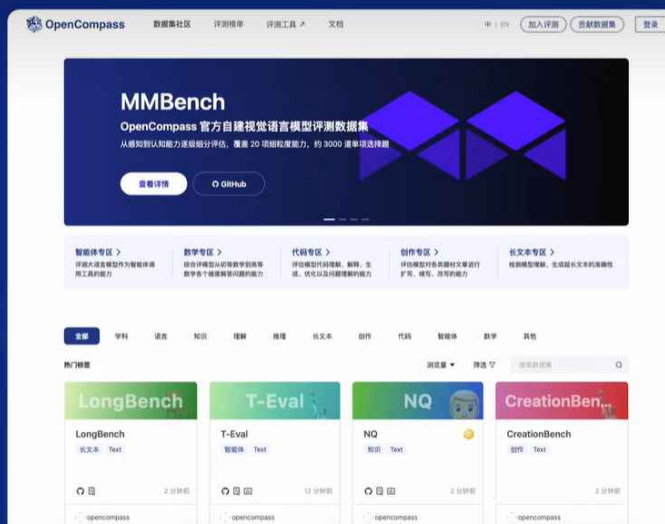


### MixtralKit MoE 模型入门工具

为 MoE 模型初学者提供学习资料、模型架构解析、推理与评测教程等入门工具。



## 开源开放，共建共享的大模型评测基准社区



## 能力维度全面升级

## 基础能力

考察大模型在如语言、知识、理解、数学、代码、推理等维度上的基本功



语言



知识



理解



数学



代码



推理

## 综合能力

考察大模型综合运用各类知识、理解与分析、多步推理、代码工具等来完成复杂任务的能力水平



考试



对话



创作



智能体



评价



长文本

## MathBench



多层次数学能力评测基准

## CriticBench



多维度的LLM反思能力评估基准

## T-Eval



大模型细粒度工具能力评测基准

## CreationBench



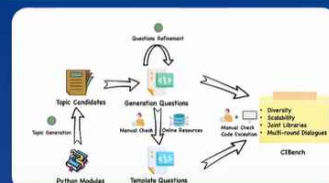
多场景中文创作能力评测基准

## F-Eval



大模型基础能力评测

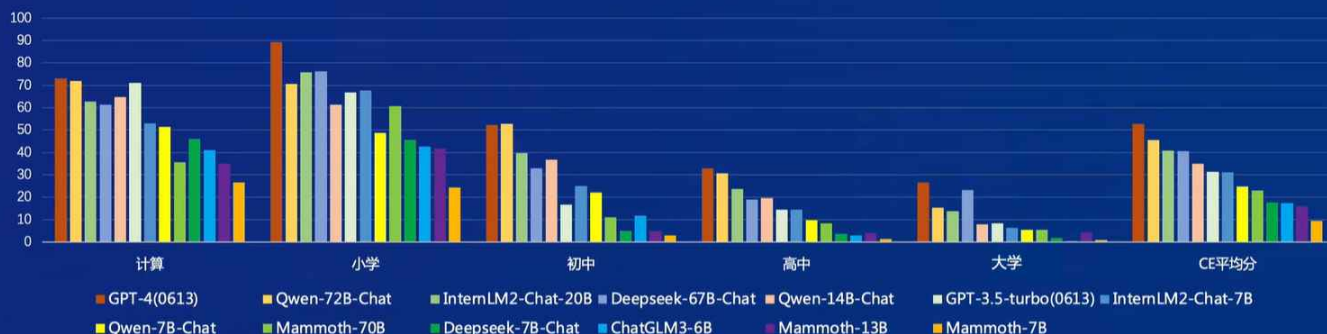
## CIBench



代码解释器能力评测基准

## MathBench:多层次数学能力评测基准

### MathBench性能



### 三级梯度难度:

- 初级：注重基础运算(计算与小学)
- 中级：基础应用和理解(初中)
- 高级：复杂运算及推理(高中和大学)

### 多语言题目来源：

- 中文：中考、高考和研究生考试等
- 英文：SAT、AMC、GRE考试等

### 循环评估Circular Evaluation

问：自助餐厅已有有23个苹果。新买了6个，他们现在有多少个苹果？  
 A: 1个, B: 17个, C: 29个, D: 138个  
 问：自助餐厅已有有23个苹果。新买了6个，他们现在有多少个苹果？  
 A: 7个, B: 29个, C: 138个, D: 1个  
 问：自助餐厅已有有23个苹果。新买了6个，他们现在有多少个苹果？  
 A: 29个, B: 138个, C: 1个, D: 7个  
 问：自助餐厅已有有23个苹果。新买了6个，他们现在有多少个苹果？  
 A: 138个, B: 1个, C: 17个, D: 29个

- 任务和软件的多样性
- 评测基准具有高度可拓展性
- 多难度等级设计（10轮以上交互）

能力	Python库
数据读取与操作	Pandas
数据可视化	Matplotlib Seaborn
机器学习	Scikit-learn LightGBM
深度学习	PyTorch TensorFlow
自然语言处理	NLTK
计算机视觉	OpenCV

13:09 / 45:36



发个友善的弹幕见证弹幕礼仪

发送

1080P 高清 选集 1.5x

Load the Penguins dataset provided by Seaborn. Drop the rows containing missing data from the dataset.

```

(1): import seaborn as sns

# Load the Penguins dataset
penguins = sns.load_dataset('penguins')

# Drop rows with missing data
penguins_cleaned = penguins.dropna()
    
```

Plot a univariate distribution of observations for the 'flipper\_length\_mm' column. Add color to the plot based on the 'species' column to differentiate between the species.

```

(2): import seaborn as sns
import matplotlib.pyplot as plt

# Create the plot with colors based on the 'species' column
sns.histplot(data=penguins_cleaned, x='flipper_length_mm', hue='species', kde=True)

# Show the plot
plt.show()
    
```

Fill the area under the lines of the plot with color for better visualization. Calculate the mean of the 'flipper\_length\_mm' column and add a vertical line to the plot at this mean value.

```

(3): import seaborn as sns
import matplotlib.pyplot as plt

# Create the plot with colors based on the 'species' column
# Fill the area under KDE plot
sns.histplot(data=penguins_cleaned, x='flipper_length_mm', hue='species', kde=True, element='poly')

# Calculate the mean of the 'flipper_length_mm' column
mean_flipper_length = penguins_cleaned['flipper_length_mm'].mean()

# Add a vertical line to the plot at the mean value
plt.axvline(x=mean_flipper_length)
    
```

规划

检索

指令  
遵循

推理

理解

评价

**(a) Plan:**

The plan to solve (Question) is:  
First, call (Tool Name) to get ...;  
Second, call (Tool Name) to get ...;  
...  
Finally, return ... as the answer.

**(b) Reason:**

User's question is (Question).  
I have known (Previous Rounds' Return).  
My thought and goal at current step is (Current Thought).

**(c) Retrieve:**

My goal is (Current Thought) at this step.  
From the (Tool List), the useful tool is (Tool Name).

**(d) Understand:**

My goal is (Current Thought) at this step.  
I want to use (Tool Name and Document).  
The proper parameters to call the tool is (Parameter List).

**(e) Instruct:**

The tool's Document is (Document File):  
I want to call it with (Parameter List).  
To call the tool successfully, the right command is (Call Command).

**(f) Review:**

I want to know (thought) at this step.  
The tool returns (tool response).  
The return [succeeds / fails] to provide enough information.

Human Query: (Query Description)

Tool Set: (Tool Document List)

Before Tool Calls:  
Design a plan to call tools sequentially.

Multi-Round Tool Calls: Each Round

Before calling:  
Reason about what to do next.  
Retrieve useful tool for this query.  
Understand proper parameters.

At calling:  
Follow document and instruction of parameters to form right commands.

After calling:  
Review the return of tools.

After Tool Calls:  
Summarize and output final answer.

- 全面且细粒度的评测方式
- 高质量的评测数据
- 剖析模型的工具调用能力



# 群策群力：携手行业领先共建繁荣生态

OpenMM Lab Bilibili

## OpenFinData全场景金融评测基准



东方财富

### OpenFinData 金融全场景评测数据集介绍

## LawBench大模型司法能力基准



### LawBench

评估大语言模型 (LLMs)  
在高度专业化法律领域的综合评估基准

## MedBench中文医疗大模型评测基准

### MedBench

MedBench致力于打造一个科学、公平且严谨的中文医疗大模型评测体系及开放平台。我们基于医学权威标准，不断更新维护高质量的医学数据集，全方位多维度量化模型在各个医学维度的能力。

## SecBench网络安全评测基准 Tencent 腾讯

### SecBench 网络安全大模型评测

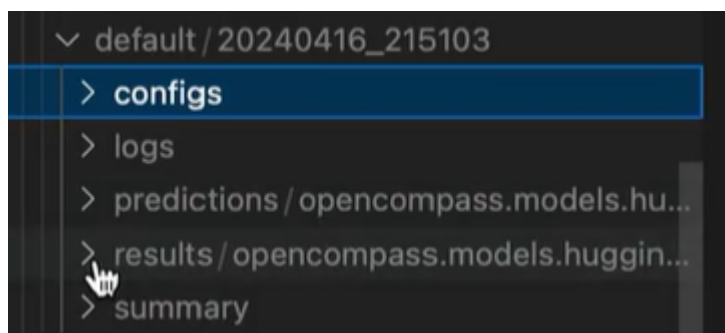
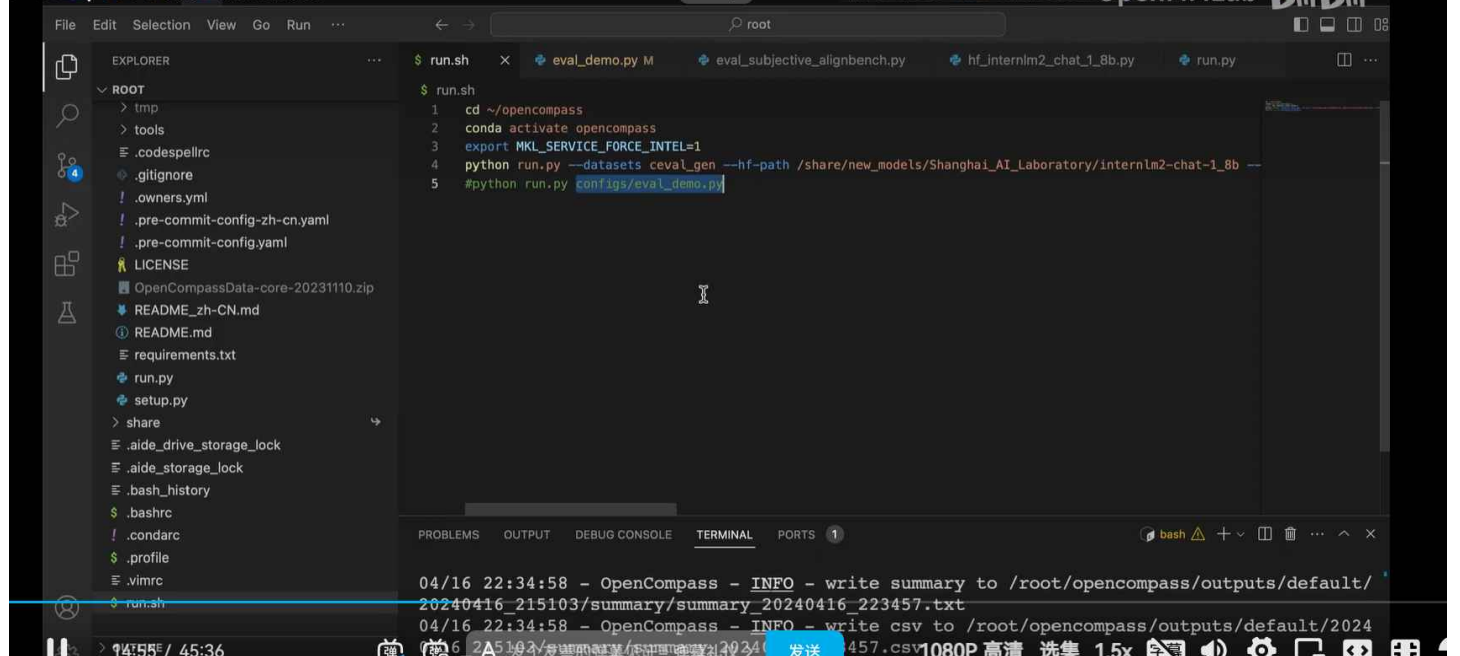
腾讯安全实验室联合业界顶尖网络安全专家，构建首个面向网络安全大模型的综合评估基准。该基准旨在全面评估大模型在网络安全领域的综合能力，包括漏洞挖掘、恶意代码分析、威胁情报处理等。SecBench 包含丰富的测试用例和自动化评估工具，为业界提供公平、公正、客观的评测平台。

立即下载

腾讯安全实验室  
Tencent Security Lab

腾讯安全科技实验室  
Tencent Security Technology Lab

## OpenCompass 大模型评测实战



跑的数据的总结

```
File Edit Selection View Go Run ... root
EXPLORER
ROOT
  > agieval
  > anli
  > anthropics_evals
  > apps
  > ARC_c
  > ARC_e
  > bbh
  > ceval
    > .ipynb_checkpoints
    > ceval_clean_ppl.py
    > ceval_gen_2daf24.py
    > ceval_gen_5f30c7_new.py U
    > ceval_gen_5f30c7.py M
    > ceval_gen.py
    > ceval_internal_ppl_1cd8bf.py
    > ceval_ppl_1cd8bf.py
    > ceval_ppl_93e5ce.py
    > ceval_ppl_578f8d.py
    > ceval_ppl.py
    > ceval_zero_shot_gen_bd40ef.py
  > ChemBench
  > CIBench
  > civilcomments

run.sh ceval_gen_5f30c7.py M ceval.py ceval2.py U ceval_gen_5f30c7_new.py U x
opencompass > configs > datasets > ceval > ceval_gen_5f30c7_new.py
3 from opencompass.openicl.icl_inferencer import GenInferencer
4 from opencompass.openicl.icl_evaluator import AccEvaluator
5 from opencompass.datasets import CevalDataset
6 from opencompass.utils.text_postprocessors import first_capital_postprocess
7
8 ceval_subject_mapping = {
9     'computer_network': ['Computer Network', '计算机网络', 'STEM'],
10    '': []
11 }
12 ceval_all_sets = list(ceval_subject_mapping.keys())
13
14 ceval_datasets = []
15 for _split in ["val"]:
16     for _name in ceval_all_sets:
17         _ch_name = ceval_subject_mapping[_name][1]
18         ceval_infer_cfg = dict(
19             ice_template=dict(
20                 type=PromptTemplate,
21                 template=dict(
22                     begin="</E>",
23                     round={
24                         dict(
25                             role="HUMAN",
26                             prompt=
27                                 f"以下是中国关于{_ch_name}考试的单项选择题，请选出其中的正确答案。\\n{{question}}\\nA. {{A}}
28                                 ),
29                                 dict(role="BOT", prompt="(answer)"),
30                             ),
31                             ice_token="</E>"
```

