七 OpenCompass笔记等

https://www.bilibili.com/video/BV1Pm41127jU/?spm_id_from=333.999.0.0&vd_source=628bd23c68972a5700988ed6b195c8e1

第七章介绍了大模型评测的重要性、司南评测体系2.0的实现方法和特点、以及OpenCompass的执行流程和自建数据集的方法。

一、大模型评测的重要性和挑战

评测能够全面了解大型语言模型的优势和局限。

聚焦垂直领域，进行针对性的模型评测。

欧拉评测体系在头部研究机构中广泛使用，是国产评测体系。

二、Open Pass评测工具升级

结合社区力量，定期更新榜单和模型性能。

支持自定义模型和数据集，并进行任务分割和并行化处理。

自研数据集如max bench和critical bench，专注于评测梯度难度和各种知识能力。

三、使用OpenAI API进行自然语言处理任务评测

指定评测数据集、模型路径和必要的参数。
通过命令行和Python方式启动评测，查看结果。
四、Open Compass执行流程和自建数据集方法

需要关注PARTITIONER、RUNNER、SUMMARIZER和TASKS文件。
实现新数据集时修改配置文件CONFIG、DATASET和Python文件，并返回一个包含字典和REVT的列表。
Python文件中需导入新实现的类，完成数据集的实现。
五、实现新数据集

修改类名和import语句，实现数据集的读取和格式化。
最后返回一个DATASET字典，并遍历所有子集，读取相应文件。
大模型评测概述

评测有助于指导改进人类与大型语言模型的交互，预测未来发展，预防未知风险。
了解不同模型间的性能、适用性和安全性对研究人员和产品开发者具有重要意义。

OpenCompass介绍

OpenCompass2.0提供一站式评测服务，特点包括开源可复现、全面能力维度、丰富模型支持、分布式高效评测、多样化评测范式、灵活拓展。

评测对象

主要评测语言大模型和多模态大模型，关注基座模型和对话模型。

工具架构

模型层关注基座模型和对话模型。能力层从通用能力和特色能力两方面进行评测维度设计。方法层采用客观评测与主观评测相结合的方式。

设计思路

OpenCompass旨在准确、全面、系统化评估大语言模型的能力。

评测方法

客观评测：通过定量指标比较模型输出与标准答案的差异。

主观评测：基于人类的主观感受进行评测，结合模型辅助和人类反馈。

评测实践

配置评测环境、数据准备、支持的数据集和模型、启动评测。

概览

OpenCompass包含配置、推理、评估、可视化等阶段，从模型和数据集选择到评估结果的可视化呈现。

大语言模型评测中的挑战

全面性

- 大模型应用场景千变万化
- 模型能力演进迅速
- 如和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

能力维度全面升级

OpenMMLab OpenCompass 司南 bilibili

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

开发机 WebIDE

CPU 12.21% 内存 0.78 / 72 GB 1.09% GPU-1: 30% Nvidia A100 0% 显存 0 / 24566 MiB

OpenMMLab Bilibili

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

OUTLINE

TIMELINE

$ run.sh | ceval_gen_5f30c7.py M | ceval.py | ceval2.py U | ceval_gen_5f30c7_new.py U

opencompass > configs > datasets > ceval > ceval_gen_5f30c7_new.py > _split

```
10
11  (ceval_subject_mapping.keys())
12
13
14  :
15  l_all_sets:
16  val_subject_mapping[_name][1]
17  fg = dict(
18  ate=dict(
19  PromptTemplate,
20  ate=dict(
21  egin="</E>",
22  ound=[
23      dict(
24          role="HUMAN",
25          prompt=
26          f"以下是中国关于{_c[_name}考试的单项选择题，请选出其中的正确答案。\n{{question}}\nA. {{A}}\nB. {{B}}\nC. {{C}}
27      ),
28      dict(role="BOT", prompt="{answer}"),
29  ),
30  oken="</E>",
31
32  =dict(type=FixKRetriever, fix_id_list=[0, 1, 2, 3, 4]),
33  r=dict(type=GenInferencer),
34
35
36  g = dict(
37  =dict(type=AccEvaluator),
38  processor=dict(type=first_capital_postprocess))
39
40  s.append(
```

Ln 24, Col 10 (719 selected) Spaces: 4 UTF-8 LF Python 3.11.5 ('base': conda) Layout: U.S.