OpenCompass大模型评测

官方文档：tutorial/opencompass/opencompass_tutorial.md

视频链接：OpenCompass 大模型评测 bilibili

OpenCompass介绍

模型评测的重要性：

- 普通用户：了解模型的特色能力和实际效果
- 开发者：监控模型能力变化，指导优化模型生产
- 管理机构：减少大模型带来的社会风险
- 产业界：找出最适合产业应用的模型，赋能真实场景

评测对象：

- 基座模型：一般是经过海量的文本数据以自监督学习的方式进行训练获得的模型（如OpenAI的GPT-3，Meta的LLaMA），往往具有强大的文字续写能力。
- 对话模型：一般是在的基座模型的基础上，经过指令微调或人类偏好对齐获得的模型（如OpenAI的ChatGPT、上海人工智能实验室的书生·浦语），能理解人类指令，具有较强的对话能力。

评测方式：

- 客观评测：问答题、都选题、判断题、分类题
- 主观评测：语言表达生动精彩、变化丰富、安全性

主流大模型评测框架

OpenMMLab bilibili

国内外评测体系的整体态势

	HELM	FlagEval	MMLU	Alpaca Eval	SuperCLUE	OpenLLM Leaderboard
机构	Stanford University	北京智源人工智能研究院 BEIJING ACADEMY OF ARTIFICIAL INTELLIGENCE	Berkeley UNIVERSITY OF CALIFORNIA	Stanford University	CLUE	Hugging Face
类型	客观评测	客观/主观评测	客观评测	主观评测	客观/主观评测	客观评测
量级	5W+ 英文题目	8W+ 中英双语	1W+ 英文题目	1K+ 英文题目	3K+ 中文题目	2W+ 英文题目

OpenCompass 能力框架

OpenMMLab bilibili

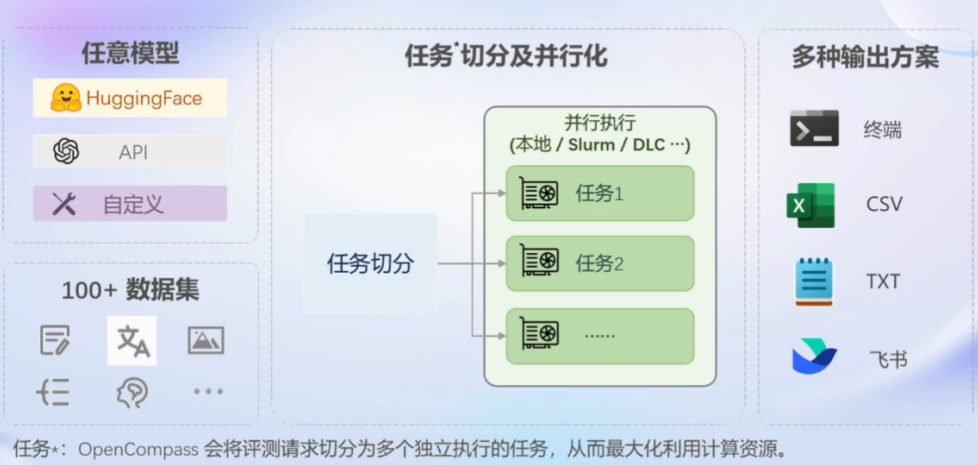
全球领先的大模型开源评测体系

6大维度，100+评测集，50万+评测题目

学科	语言	知识	理解	推理	安全
初中考试	字词释义	知识问答	阅读理解	因果推理	偏见 有害性
中国高考	成语习语	多语种知识问答	内容分析	常识推理	公平性 隐私性
大学考试	语义相似		内容总结	代码推理	真实性 合法性
语言能力考试	指代消解			数学推理	
职业资格考试	翻译				

OpenCompass 评测流水线设计

OpenMMLab bilibili



任务*：OpenCompass 会将评测请求切分为多个独立执行的任务，从而最大化利用计算资源。

大模型评测领域的挑战

在 OpenCompass 中评估一个模型通常包括以下几个阶段：**配置** -> **推理** -> **评估** -> **可视化**。

配置：这是整个工作流的起点。您需要配置整个评估过程，选择要评估的模型和数据集。此外，还可以选择评估策略、计算后端等，并定义显示结果的方式。

推理与评估：在这个阶段，OpenCompass 将会开始对模型和数据集进行并行推理和评估。**推理**阶段主要是让模型从数据集产生输出，而**评估**阶段则是衡量这些输出与标准答案的匹配程度。这两个过程会被拆分为多个同时运行的"任务"以提高效率，但请注意，如果计算资源有限，这种策略可能会使评测变得更慢。

可视化：评估完成后，OpenCompass 将结果整理成易读的表格，并将其保存为 CSV 和 TXT 文件。你也可以激活飞书状态上报功能，此后可以在飞书客户端中及时获得评测状态报告。

实战

基础作业

- 使用 OpenCompass 评测 InternLM2-Chat-7B 模型在 C-Eval 数据集上的性能

环境安装

其中,无法从`GitHub`上clone仓库，所以改用`gitee`

```
conda create --name opencompass --clone=/root/share/conda_envs/internlm-base
source activate opencompass
git clone https://github.com/open-compass/opencompass
cd opencompass
pip install -e .
```

数据准备

```
# 解压评测数据集到 data/ 处
cp /share/temp/datasets/OpenCompassData-core-20231110.zip /root/opencompass/
unzip OpenCompassData-core-20231110.zip

# 将会在opencompass下看到data文件夹
```

查看支持的数据集和模型

```
# 列出所有跟 internlm 及 ceval 相关的配置
python tools/list_configs.py internlm ceval
```

```
| Model                     | Config Path                                                    |
|---------------------------+----------------------------------------------------------------|
| hf_internlm2_20b          | configs/models/hf_internlm/hf_internlm2_20b.py                 |
| hf_internlm2_7b           | configs/models/hf_internlm/hf_internlm2_7b.py                  |
| hf_internlm2_chat_20b     | configs/models/hf_internlm/hf_internlm2_chat_20b.py            |
| hf_internlm2_chat_7b      | configs/models/hf_internlm/hf_internlm2_chat_7b.py             |
| hf_internlm_20b           | configs/models/hf_internlm/hf_internlm_20b.py                  |
| hf_internlm_7b            | configs/models/hf_internlm/hf_internlm_7b.py                   |
| hf_internlm_chat_20b      | configs/models/hf_internlm/hf_internlm_chat_20b.py             |
| hf_internlm_chat_7b       | configs/models/hf_internlm/hf_internlm_chat_7b.py              |
| hf_internlm_chat_7b_8k    | configs/models/hf_internlm/hf_internlm_chat_7b_8k.py           |
| hf_internlm_chat_7b_v1_1  | configs/models/hf_internlm/hf_internlm_chat_7b_v1_1.py         |
| internlm_7b               | configs/models/internlm/internlm_7b.py                         |
| ms_internlm_chat_7b_8k    | configs/models/ms_internlm/ms_internlm_chat_7b_8k.py           |
+---------------------------+----------------------------------------------------------------+
+---------------------------+----------------------------------------------------------------+
| Dataset                   | Config Path                                                    |
|---------------------------+----------------------------------------------------------------|
| ceval_clean_ppl           | configs/datasets/ceval/ceval_clean_ppl.py                      |
| ceval_gen                 | configs/datasets/ceval/ceval_gen.py                            |
| ceval_gen_2daf24          | configs/datasets/ceval/ceval_gen_2daf24.py                     |
| ceval_gen_5f30c7          | configs/datasets/ceval/ceval_gen_5f30c7.py                     |
| ceval_ppl                 | configs/datasets/ceval/ceval_ppl.py                            |
| ceval_ppl_578f8d          | configs/datasets/ceval/ceval_ppl_578f8d.py                     |
| ceval_ppl_93e5ce          | configs/datasets/ceval/ceval_ppl_93e5ce.py                     |
| ceval_zero_shot_gen_bd40ef | configs/datasets/ceval/ceval_zero_shot_gen_bd40ef.py           |
+---------------------------+----------------------------------------------------------------+
```

启动评测

确保按照上述步骤正确安装 OpenCompass 并准备好数据集后，可以通过以下命令评测 InternLM-Chat-7B 模型在 C-Eval 数据集上的性能。由于 OpenCompass 默认并行启动评估过程，我们可以在第一次运行时以 `--debug` 模式启动评估，并检查是否存在问题。在 `--debug` 模式下，任务将按顺序执行，并实时打印输出。

```
python run.py \
--datasets ceval_gen \
--hf-path /share/temp/model_repos/internlm-chat-7b/ \  # HuggingFace 模型路径
--tokenizer-path /share/temp/model_repos/internlm-chat-7b/ \  # HuggingFace
tokenizer 路径（如果与模型路径相同，可以省略）
--tokenizer-kwargs padding_side='left' truncation='left' trust_remote_code=True
\  # 构建 tokenizer 的参数
--model-kwargs device_map='auto' trust_remote_code=True \  # 构建模型的参数
--max-seq-len 2048 \  # 模型可以接受的最大序列长度
--max-out-len 16 \  # 生成的最大 token 数
--batch-size 2  \  # 批量大小
--num-gpus 1  # 运行模型所需的 GPU 数量
--debug
```

```
01/28 20:40:31 - OpenCompass - INFO - time elapsed: 19.18s
01/28 20:40:31 - OpenCompass - DEBUG - Get class `OpenICLEvalTask` from "task" registry in "opencompass"
01/28 20:40:31 - OpenCompass - DEBUG - An `OpenICLEvalTask` instance is built from registry, and its implementation can be fo
und in opencompass.tasks.openicl_eval
01/28 20:41:08 - OpenCompass - INFO - Task [opencompass.models.huggingface.HuggingFace_model_repos_internlm-chat-7b/ceval-fir
e_engineer]: {'accuracy': 22.58064516129032}
01/28 20:41:08 - OpenCompass - INFO - time elapsed: 23.06s
01/28 20:41:09 - OpenCompass - DEBUG - Get class `OpenICLEvalTask` from "task" registry in "opencompass"
01/28 20:41:09 - OpenCompass - DEBUG - An `OpenICLEvalTask` instance is built from registry, and its implementation can be fo
und in opencompass.tasks.openicl_eval
01/28 20:41:47 - OpenCompass - INFO - Task [opencompass.models.huggingface.HuggingFace_model_repos_internlm-chat-7b/ceval-env
ironmental_impact_assessment_engineer]: {'accuracy': 64.51612903225806}
01/28 20:41:47 - OpenCompass - INFO - time elapsed: 22.45s
01/28 20:41:48 - OpenCompass - DEBUG - Get class `OpenICLEvalTask` from "task" registry in "opencompass"
01/28 20:41:48 - OpenCompass - DEBUG - An `OpenICLEvalTask` instance is built from registry, and its implementation can be fo
und in opencompass.tasks.openicl_eval
01/28 20:42:25 - OpenCompass - INFO - Task [opencompass.models.huggingface.HuggingFace_model_repos_internlm-chat-7b/ceval-tax
_accountant]: {'accuracy': 34.69387755102041}
01/28 20:42:25 - OpenCompass - INFO - time elapsed: 22.49s
01/28 20:42:26 - OpenCompass - DEBUG - Get class `OpenICLEvalTask` from "task" registry in "opencompass"
01/28 20:42:26 - OpenCompass - DEBUG - An `OpenICLEvalTask` instance is built from registry, and its implementation can be fo
und in opencompass.tasks.openicl_eval
01/28 20:43:00 - OpenCompass - INFO - Task [opencompass.models.huggingface.HuggingFace_model_repos_internlm-chat-7b/ceval-phy
sician]: {'accuracy': 40.816326530612244}
```

除了通过命令行配置实验外，OpenCompass 还允许用户在配置文件中编写实验的完整配置，并通过 `run.py` 直接运行它。

评估结果

结果存储路径 `/root/opencompass/outputs/default/20240128_200437/summary/summary_20240128_200437.csv`

summary_20240128_2004

Delimiter: ,

	dataset	version	metric	mode	repos_internlm-chat-7b
30	ceval-middle_school_geography	-	-	-	-
31	ceval-modern_chinese_history	fc01af	accuracy	gen	73.91
32	l-ideological_and_moral_cultivation	-	-	-	-
33	ceval-logic	-	-	-	-
34	ceval-law	a110a1	accuracy	gen	25.00
35	al-chinese_language_and_literature	0f8b68	accuracy	gen	30.43
36	ceval-art_studies	2a1300	accuracy	gen	60.61
37	ceval-professional_tour_guide	4e673e	accuracy	gen	62.07
38	ceval-legal_professional	ce8787	accuracy	gen	39.13
39	ceval-high_school_chinese	-	-	-	-
40	ceval-high_school_history	-	-	-	-
41	ceval-middle_school_history	-	-	-	-
42	ceval-civil_servant	87d061	accuracy	gen	53.19
43	ceval-sports_science	-	-	-	-
44	ceval-plant_protection	-	-	-	-
45	ceval-basic_medicine	-	-	-	-
46	ceval-clinical_medicine	-	-	-	-
47	ceval-urban_and_rural_planner	95b885	accuracy	gen	45.65
48	ceval-accountant	002837	accuracy	gen	26.53
49	ceval-fire_engineer	bc23f5	accuracy	gen	22.58
50	ental_impact_assessment_engineer	c64e2d	accuracy	gen	64.52
51	ceval-tax_accountant	3a5e3c	accuracy	gen	34.69
52	ceval-physician	6e277d	accuracy	gen	40.82