**为什么要研究大模型的评测？**

百家争鸣，百花齐放。

1. 首先，研究评测对于我们全面了解大型语言模型的优势和限制至关重要。尽管许多研究表明大型语言模型在多个通用任务上已经达到或超越了人类水平，但仍然存在质疑，即这些模型的能力是否只是对训练数据的记忆而非真正的理解。例如，即使只提供LeetCode题目编号而不提供具体信息，大型语言模型也能够正确输出答案，这暗示着训练数据可能存在污染现象。
2. 其次，研究评测有助于指导和改进人类与大型语言模型之间的协同交互。考虑到大型语言模型的最终服务对象是人类，为了更好地设计人机交互的新范式，我们有必要全面评估模型的各项能力。
3. 最后，研究评测可以帮助我们更好地规划大型语言模型未来的发展，并预防未知和潜在的风险。随着大型语言模型的不断演进，其能力也在不断增强。通过合理科学的评测机制，我们能够从进化的角度评估模型的能力，并提前预测潜在的风险，这是至关重要的研究内容。
4. 对于大多数人来说，大型语言模型可能似乎与他们无关，因为训练这样的模型成本较高。然而，就像飞机的制造一样，尽管成本高昂，但一旦制造完成，大家使用的机会就会非常频繁。因此，了解不同语言模型之间的性能、舒适性和安全性，能够帮助人们更好地选择适合的模型，这对于研究人员和产品开发者而言同样具有重要意义。

**OpenCompass介绍**

上海人工智能实验室科学家团队正式发布了大模型开源开放评测体系 “司南” (OpenCompass2.0)，用于为大语言模型、多模态模型等提供一站式评测服务。其主要特点如下：

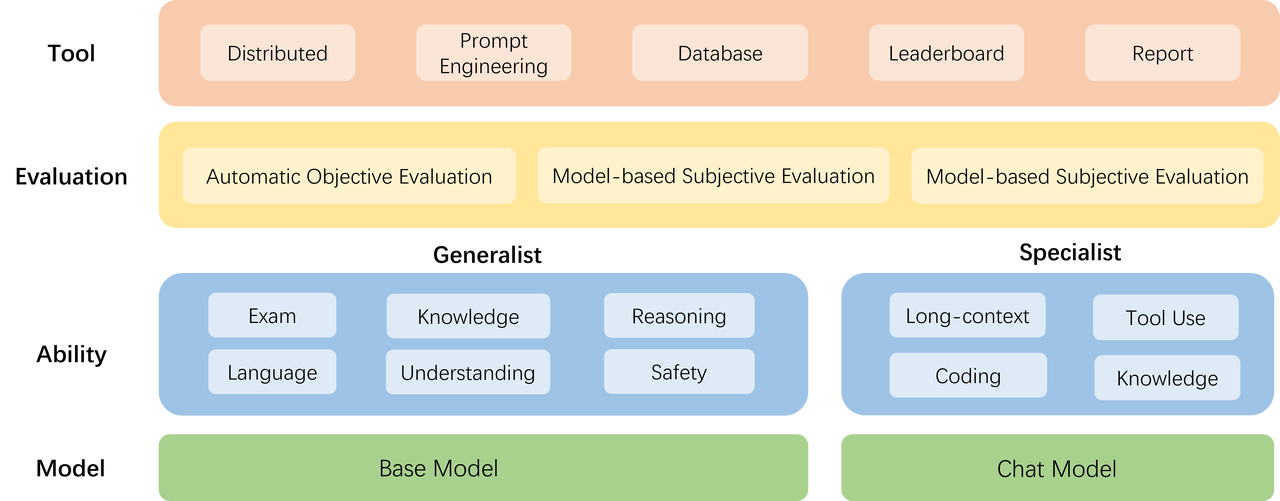
* 开源可复现：提供公平、公开、可复现的大模型评测方案
* 全面的能力维度：五大维度设计，提供 70+ 个数据集约 40 万题的的模型评测方案，全面评估模型能力
* 丰富的模型支持：已支持 20+ HuggingFace 及 API 模型
* 分布式高效评测：一行命令实现任务分割和分布式评测，数小时即可完成千亿模型全量评测
* 多样化评测范式：支持零样本、小样本及思维链评测，结合标准型或对话型提示词模板，轻松激发各种模型最大性能
* 灵活化拓展：想增加新模型或数据集？想要自定义更高级的任务分割策略，甚至接入新的集群管理系统？OpenCompass 的一切均可轻松扩展！

**评测对象**

本算法库的主要评测对象为语言大模型与多模态大模型。我们以语言大模型为例介绍评测的具体模型类型。

* 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。

**工具架构**

[](https://private-user-images.githubusercontent.com/148421775/321935556-705924f8-01b0-48f2-bca7-c660445b013f.png?jwt=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJpc3MiOiJnaXRodWIuY29tIiwiYXVkIjoicmF3LmdpdGh1YnVzZXJjb250ZW50LmNvbSIsImtleSI6ImtleTUiLCJleHAiOjE3MTc0OTQyOTAsIm5iZiI6MTcxNzQ5Mzk5MCwicGF0aCI6Ii8xNDg0MjE3NzUvMzIxOTM1NTU2LTcwNTkyNGY4LTAxYjAtNDhmMi1iY2E3LWM2NjA0NDViMDEzZi5wbmc_WC1BbXotQWxnb3JpdGhtPUFXUzQtSE1BQy1TSEEyNTYmWC1BbXotQ3JlZGVudGlhbD1BS0lBVkNPRFlMU0E1M1BRSzRaQSUyRjIwMjQwNjA0JTJGdXMtZWFzdC0xJTJGczMlMkZhd3M0X3JlcXVlc3QmWC1BbXotRGF0ZT0yMDI0MDYwNFQwOTM5NTBaJlgtQW16LUV4cGlyZXM9MzAwJlgtQW16LVNpZ25hdHVyZT00NjM2MjllZDkyMDM4ZTllYzRlODlmYjcyMWZkYTU5ZWJlNWEyYjhhYzMwNDYyMzY0MTNmNjg4OTI4NjFhNjZjJlgtQW16LVNpZ25lZEhlYWRlcnM9aG9zdCZhY3Rvcl9pZD0wJmtleV9pZD0wJnJlcG9faWQ9MCJ9.nvh3C1Rj24smOM1Eij75pIquvsKiIpMKAOVpmoBRJwY)

* 模型层：大模型评测所涉及的主要模型种类，OpenCompass 以基座模型和对话模型作为重点评测对象。
* 能力层：OpenCompass 从本方案从通用能力和特色能力两个方面来进行评测维度设计。在模型通用能力方面，从语言、知识、理解、推理、安全等多个能力维度进行评测。在特色能力方面，从长文本、代码、工具、知识增强等维度进行评测。
* 方法层：OpenCompass 采用客观评测与主观评测两种评测方式。客观评测能便捷地评估模型在具有确定答案（如选择，填空，封闭式问答等）的任务上的能力，主观评测能评估用户对模型回复的真实满意度，OpenCompass 采用基于模型辅助的主观评测和基于人类反馈的主观评测两种方式。
* 工具层：OpenCompass 提供丰富的功能支持自动化地开展大语言模型的高效评测。包括分布式评测技术，提示词工程，对接评测数据库，评测榜单发布，评测报告生成等诸多功能。

**设计思路**

为准确、全面、系统化地评估大语言模型的能力，OpenCompass 从通用人工智能的角度出发，结合学术界的前沿进展和工业界的最佳实践，提出一套面向实际应用的模型能力评价体系。OpenCompass 能力维度体系涵盖通用能力和特色能力两大部分。

**评测方法**

OpenCompass 采取客观评测与主观评测相结合的方法。针对具有确定性答案的能力维度和场景，通过构造丰富完善的评测集，对模型能力进行综合评价。针对体现模型能力的开放式或半开放式的问题、模型安全问题等，采用主客观相结合的评测方式。

**客观评测**

针对具有标准答案的客观问题，我们可以通过使用定量指标比较模型的输出与标准答案的差异，并根据结果衡量模型的性能。同时，由于大语言模型输出自由度较高，在评测阶段，我们需要对其输入和输出作一定的规范和设计，尽可能减少噪声输出在评测阶段的影响，才能对模型的能力有更加完整和客观的评价。 为了更好地激发出模型在题目测试领域的能力，并引导模型按照一定的模板输出答案，OpenCompass 采用提示词工程 （prompt engineering）和语境学习（in-context learning）进行客观评测。 在客观评测的具体实践中，我们通常采用下列两种方式进行模型输出结果的评测：

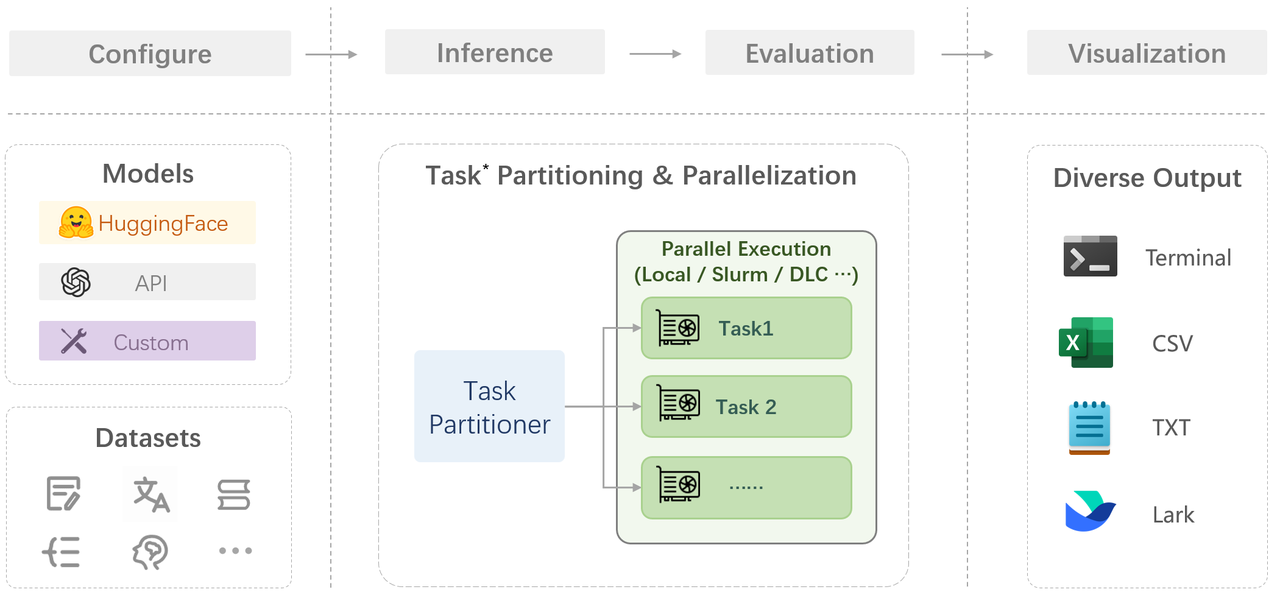
* 判别式评测：该评测方式基于将问题与候选答案组合在一起，计算模型在所有组合上的困惑度（perplexity），并选择困惑度最小的答案作为模型的最终输出。例如，若模型在 问题? 答案1 上的困惑度为 0.1，在 问题? 答案2 上的困惑度为 0.2，最终我们会选择 答案1 作为模型的输出。
* 生成式评测：该评测方式主要用于生成类任务，如语言翻译、程序生成、逻辑分析题等。具体实践时，使用问题作为模型的原始输入，并留白答案区域待模型进行后续补全。我们通常还需要对其输出进行后处理，以保证输出满足数据集的要求。

**主观评测**

语言表达生动精彩，变化丰富，大量的场景和能力无法凭借客观指标进行评测。针对如模型安全和模型语言能力的评测，以人的主观感受为主的评测更能体现模型的真实能力，并更符合大模型的实际使用场景。 OpenCompass 采取的主观评测方案是指借助受试者的主观判断对具有对话能力的大语言模型进行能力评测。在具体实践中，我们提前基于模型的能力维度构建主观测试问题集合，并将不同模型对于同一问题的不同回复展现给受试者，收集受试者基于主观感受的评分。由于主观测试成本高昂，本方案同时也采用使用性能优异的大语言模拟人类进行主观打分。在实际评测中，本文将采用真实人类专家的主观评测与基于模型打分的主观评测相结合的方式开展模型能力评估。 在具体开展主观评测时，OpenComapss 采用单模型回复满意度统计和多模型满意度比较两种方式开展具体的评测工作。

**快速开始**

**快速开始**

[](https://private-user-images.githubusercontent.com/148421775/321935388-276fa169-4937-4ef0-9620-1f7863054863.png?jwt=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJpc3MiOiJnaXRodWIuY29tIiwiYXVkIjoicmF3LmdpdGh1YnVzZXJjb250ZW50LmNvbSIsImtleSI6ImtleTUiLCJleHAiOjE3MTc0OTQyOTAsIm5iZiI6MTcxNzQ5Mzk5MCwicGF0aCI6Ii8xNDg0MjE3NzUvMzIxOTM1Mzg4LTI3NmZhMTY5LTQ5MzctNGVmMC05NjIwLTFmNzg2MzA1NDg2My5wbmc_WC1BbXotQWxnb3JpdGhtPUFXUzQtSE1BQy1TSEEyNTYmWC1BbXotQ3JlZGVudGlhbD1BS0lBVkNPRFlMU0E1M1BRSzRaQSUyRjIwMjQwNjA0JTJGdXMtZWFzdC0xJTJGczMlMkZhd3M0X3JlcXVlc3QmWC1BbXotRGF0ZT0yMDI0MDYwNFQwOTM5NTBaJlgtQW16LUV4cGlyZXM9MzAwJlgtQW16LVNpZ25hdHVyZT0xY2Y0NTNmYmU4M2FhNWQ5ZDYyNGE0NzBlOWJhNzczZmIwMzBiMmFmOTVlNjI4Y2E5ZDdjMjRkZDQxNTlmNzY0JlgtQW16LVNpZ25lZEhlYWRlcnM9aG9zdCZhY3Rvcl9pZD0wJmtleV9pZD0wJnJlcG9faWQ9MCJ9.4xIXgmiZZRIxrjYFuyDjo8Ju3Nw77_MWAKTsyqBWLoo)

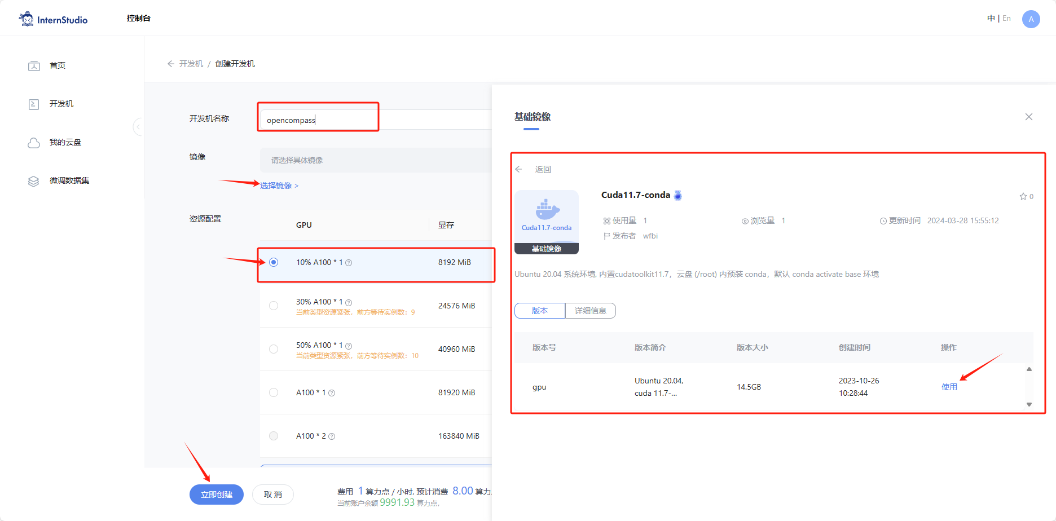
**概览**

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。如果需要了解该问题及解决方案，可以参考 FAQ: 效率。
* 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。 接下来，我们将展示 OpenCompass 的基础用法，展示书生浦语在 C-Eval 基准任务上的评估。它们的配置文件可以在 configs/eval\_demo.py 中找到。

**环境配置**

**创建开发机和 conda 环境**

在创建开发机界面选择镜像为 Cuda11.7-conda，并选择 GPU 为10% A100。 [](https://private-user-images.githubusercontent.com/102272920/323903366-1302d976-64db-4fc2-87f6-570e4a71147a.png?jwt=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJpc3MiOiJnaXRodWIuY29tIiwiYXVkIjoicmF3LmdpdGh1YnVzZXJjb250ZW50LmNvbSIsImtleSI6ImtleTUiLCJleHAiOjE3MTc0OTQyOTAsIm5iZiI6MTcxNzQ5Mzk5MCwicGF0aCI6Ii8xMDIyNzI5MjAvMzIzOTAzMzY2LTEzMDJkOTc2LTY0ZGItNGZjMi04N2Y2LTU3MGU0YTcxMTQ3YS5wbmc_WC1BbXotQWxnb3JpdGhtPUFXUzQtSE1BQy1TSEEyNTYmWC1BbXotQ3JlZGVudGlhbD1BS0lBVkNPRFlMU0E1M1BRSzRaQSUyRjIwMjQwNjA0JTJGdXMtZWFzdC0xJTJGczMlMkZhd3M0X3JlcXVlc3QmWC1BbXotRGF0ZT0yMDI0MDYwNFQwOTM5NTBaJlgtQW16LUV4cGlyZXM9MzAwJlgtQW16LVNpZ25hdHVyZT1lNzcyZjIyYTEzZGI5MDJjZGUxNGMzMjgxZmYxMTkzOTFkMTYxMmI5NjRkNjRkZGQ5MGY4M2M3NGJhYzMyMTljJlgtQW16LVNpZ25lZEhlYWRlcnM9aG9zdCZhY3Rvcl9pZD0wJmtleV9pZD0wJnJlcG9faWQ9MCJ9.whRaT04J7JnS0PFBt3VLCdvxj0hwECC97mgNM-hBLv4)

**安装**

**面向GPU的环境安装**

studio-conda -o internlm-base -t opencompass

source activate opencompass

git clone -b 0.2.4 https://github.com/open-compass/opencompass

cd opencompass

pip install -e .

**如果pip install -e .安装未成功,请运行:**

pip install -r requirements.txt

有部分第三方功能,如代码能力基准测试 HumanEval 以及 Llama 格式的模型评测,可能需要额外步骤才能正常运行，如需评测，详细步骤请参考安装指南。

**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 C-Eval 相关的配置

python tools/list\_configs.py internlm ceval

将会看到

+----------------------------------------+----------------------------------------------------------------------+

| Model | Config Path |

|----------------------------------------+----------------------------------------------------------------------|

| hf\_internlm2\_1\_8b | configs/models/hf\_internlm/hf\_internlm2\_1\_8b.py |

| hf\_internlm2\_20b | configs/models/hf\_internlm/hf\_internlm2\_20b.py |

| hf\_internlm2\_7b | configs/models/hf\_internlm/hf\_internlm2\_7b.py |

| hf\_internlm2\_base\_20b | configs/models/hf\_internlm/hf\_internlm2\_base\_20b.py |

| hf\_internlm2\_base\_7b | configs/models/hf\_internlm/hf\_internlm2\_base\_7b.py |

| hf\_internlm2\_chat\_1\_8b | configs/models/hf\_internlm/hf\_internlm2\_chat\_1\_8b.py |

| hf\_internlm2\_chat\_1\_8b\_sft | configs/models/hf\_internlm/hf\_internlm2\_chat\_1\_8b\_sft.py |

| hf\_internlm2\_chat\_20b | configs/models/hf\_internlm/hf\_internlm2\_chat\_20b.py |

| hf\_internlm2\_chat\_20b\_sft | configs/models/hf\_internlm/hf\_internlm2\_chat\_20b\_sft.py |

| hf\_internlm2\_chat\_20b\_with\_system | configs/models/hf\_internlm/hf\_internlm2\_chat\_20b\_with\_system.py |

| hf\_internlm2\_chat\_7b | configs/models/hf\_internlm/hf\_internlm2\_chat\_7b.py |

| hf\_internlm2\_chat\_7b\_sft | configs/models/hf\_internlm/hf\_internlm2\_chat\_7b\_sft.py |

| hf\_internlm2\_chat\_7b\_with\_system | configs/models/hf\_internlm/hf\_internlm2\_chat\_7b\_with\_system.py |

| hf\_internlm2\_chat\_math\_20b | configs/models/hf\_internlm/hf\_internlm2\_chat\_math\_20b.py |

| hf\_internlm2\_chat\_math\_20b\_with\_system | configs/models/hf\_internlm/hf\_internlm2\_chat\_math\_20b\_with\_system.py |

| hf\_internlm2\_chat\_math\_7b | configs/models/hf\_internlm/hf\_internlm2\_chat\_math\_7b.py |

| hf\_internlm2\_chat\_math\_7b\_with\_system | configs/models/hf\_internlm/hf\_internlm2\_chat\_math\_7b\_with\_system.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\_contamination\_ppl\_810ec6 | configs/datasets/contamination/ceval\_contamination\_ppl\_810ec6.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\_1cd8bf | configs/datasets/ceval/ceval\_ppl\_1cd8bf.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 |

| configuration\_internlm | configs/datasets/cdme/internlm2-chat-7b/configuration\_internlm.py |

| modeling\_internlm2 | configs/datasets/cdme/internlm2-chat-7b/modeling\_internlm2.py |

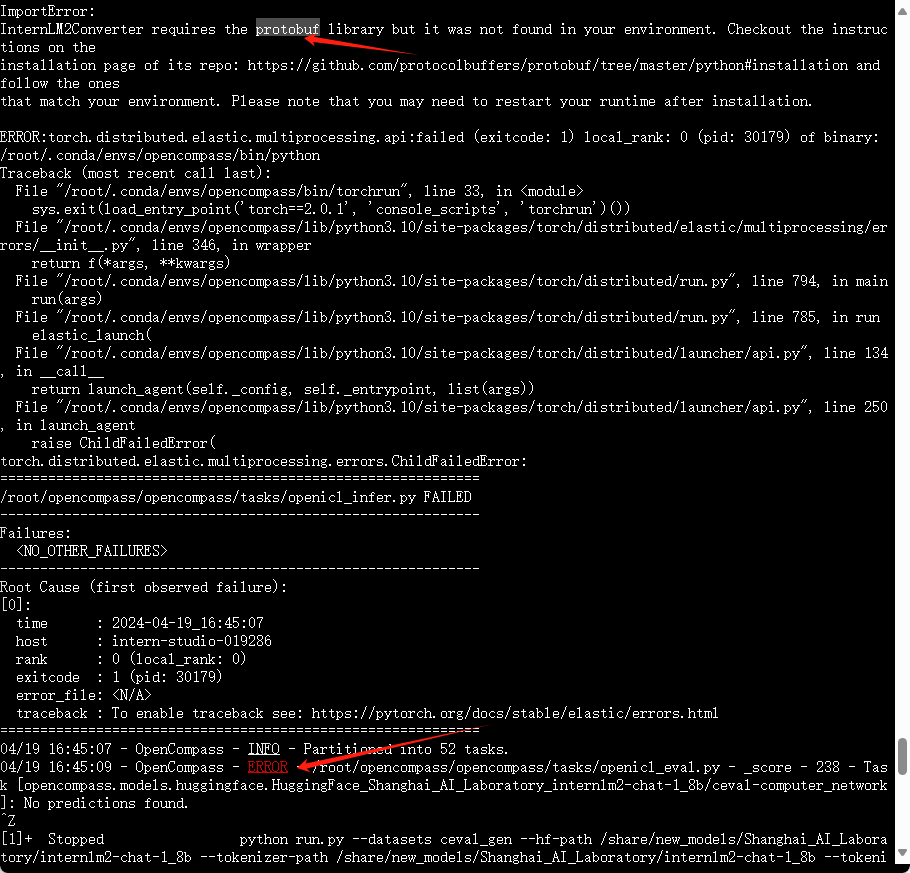
| tokenization\_internlm | configs/datasets/cdme/internlm2-chat-7b/tokenization\_internlm.py |

+--------------------------------+-------------------------------------------------------------------+

**启动评测 (10% A100 8GB 资源)**

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 InternLM2-Chat-1.8B 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/new\_models/Shanghai\_AI\_Laboratory/internlm2-chat-1\_8b --tokenizer-path /share/new\_models/Shanghai\_AI\_Laboratory/internlm2-chat-1\_8b --tokenizer-kwargs padding\_side='left' truncation='left' trust\_remote\_code=True --model-kwargs trust\_remote\_code=True device\_map='auto' --max-seq-len 1024 --max-out-len 16 --batch-size 2 --num-gpus 1 --debug

**遇到错误：**  
[](https://private-user-images.githubusercontent.com/102272920/323895192-6f0295e9-e1f6-479d-b9a0-18ef5a97c77d.png?jwt=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJpc3MiOiJnaXRodWIuY29tIiwiYXVkIjoicmF3LmdpdGh1YnVzZXJjb250ZW50LmNvbSIsImtleSI6ImtleTUiLCJleHAiOjE3MTc0OTQyOTAsIm5iZiI6MTcxNzQ5Mzk5MCwicGF0aCI6Ii8xMDIyNzI5MjAvMzIzODk1MTkyLTZmMDI5NWU5LWUxZjYtNDc5ZC1iOWEwLTE4ZWY1YTk3Yzc3ZC5wbmc_WC1BbXotQWxnb3JpdGhtPUFXUzQtSE1BQy1TSEEyNTYmWC1BbXotQ3JlZGVudGlhbD1BS0lBVkNPRFlMU0E1M1BRSzRaQSUyRjIwMjQwNjA0JTJGdXMtZWFzdC0xJTJGczMlMkZhd3M0X3JlcXVlc3QmWC1BbXotRGF0ZT0yMDI0MDYwNFQwOTM5NTBaJlgtQW16LUV4cGlyZXM9MzAwJlgtQW16LVNpZ25hdHVyZT01NTQwZDNiZmNmZjJjYjViNDdmMjQwMDJhYTM1NWU3ZmJjZTM4NTRjZThiMDAyNjcyZDE5Y2Y5MzJmYjViNzNhJlgtQW16LVNpZ25lZEhlYWRlcnM9aG9zdCZhY3Rvcl9pZD0wJmtleV9pZD0wJnJlcG9faWQ9MCJ9.LpbDcYRB21DCh9euK28zu6jps1DzEHceSsC3CXlx9XM)  
**解决方案：**

pip install protobuf

命令解析

python run.py

--datasets ceval\_gen \

--hf-path /share/new\_models/Shanghai\_AI\_Laboratory/internlm2-chat-1\_8b \ # HuggingFace 模型路径

--tokenizer-path /share/new\_models/Shanghai\_AI\_Laboratory/internlm2-chat-1\_8b \ # 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 1024 \ # 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

**遇到错误mkl-service + Intel(R) MKL MKL\_THREADING\_LAYER=INTEL is incompatible with libgomp.so.1 ... 解决方案：**

export MKL\_SERVICE\_FORCE\_INTEL=1

#或

export MKL\_THREADING\_LAYER=GNU

如果一切正常，您应该看到屏幕上显示 “Starting inference process”：

[2024-03-18 12:39:54,972] [opencompass.openicl.icl\_inferencer.icl\_gen\_inferencer] [INFO] Starting inference process...

评测完成后，将会看到：

dataset version metric mode opencompass.models.huggingface.HuggingFace\_Shanghai\_AI\_Laboratory\_internlm2-chat-1\_8b

---------------------------------------------- --------- ------------- ------ ---------------------------------------------------------------------------------------

ceval-computer\_network db9ce2 accuracy gen 47.37

ceval-operating\_system 1c2571 accuracy gen 47.37

ceval-computer\_architecture a74dad accuracy gen 23.81

ceval-college\_programming 4ca32a accuracy gen 13.51

ceval-college\_physics 963fa8 accuracy gen 42.11

ceval-college\_chemistry e78857 accuracy gen 33.33

ceval-advanced\_mathematics ce03e2 accuracy gen 10.53

ceval-probability\_and\_statistics 65e812 accuracy gen 38.89

ceval-discrete\_mathematics e894ae accuracy gen 25

ceval-electrical\_engineer ae42b9 accuracy gen 27.03

ceval-metrology\_engineer ee34ea accuracy gen 54.17

ceval-high\_school\_mathematics 1dc5bf accuracy gen 16.67

ceval-high\_school\_physics adf25f accuracy gen 42.11

ceval-high\_school\_chemistry 2ed27f accuracy gen 47.37

ceval-high\_school\_biology 8e2b9a accuracy gen 26.32

ceval-middle\_school\_mathematics bee8d5 accuracy gen 36.84

ceval-middle\_school\_biology 86817c accuracy gen 80.95

ceval-middle\_school\_physics 8accf6 accuracy gen 47.37

ceval-middle\_school\_chemistry 167a15 accuracy gen 80

ceval-veterinary\_medicine b4e08d accuracy gen 43.48

ceval-college\_economics f3f4e6 accuracy gen 32.73

ceval-business\_administration c1614e accuracy gen 36.36

ceval-marxism cf874c accuracy gen 68.42

ceval-mao\_zedong\_thought 51c7a4 accuracy gen 70.83

ceval-education\_science 591fee accuracy gen 55.17

ceval-teacher\_qualification 4e4ced accuracy gen 59.09

ceval-high\_school\_politics 5c0de2 accuracy gen 57.89

ceval-high\_school\_geography 865461 accuracy gen 47.37

ceval-middle\_school\_politics 5be3e7 accuracy gen 71.43

ceval-middle\_school\_geography 8a63be accuracy gen 75

ceval-modern\_chinese\_history fc01af accuracy gen 52.17

ceval-ideological\_and\_moral\_cultivation a2aa4a accuracy gen 73.68

ceval-logic f5b022 accuracy gen 27.27

ceval-law a110a1 accuracy gen 29.17

ceval-chinese\_language\_and\_literature 0f8b68 accuracy gen 47.83

ceval-art\_studies 2a1300 accuracy gen 42.42

ceval-professional\_tour\_guide 4e673e accuracy gen 51.72

ceval-legal\_professional ce8787 accuracy gen 34.78

ceval-high\_school\_chinese 315705 accuracy gen 42.11

ceval-high\_school\_history 7eb30a accuracy gen 65

ceval-middle\_school\_history 48ab4a accuracy gen 86.36

ceval-civil\_servant 87d061 accuracy gen 42.55

ceval-sports\_science 70f27b accuracy gen 52.63

ceval-plant\_protection 8941f9 accuracy gen 40.91

ceval-basic\_medicine c409d6 accuracy gen 68.42

ceval-clinical\_medicine 49e82d accuracy gen 31.82

ceval-urban\_and\_rural\_planner 95b885 accuracy gen 47.83

ceval-accountant 002837 accuracy gen 36.73

ceval-fire\_engineer bc23f5 accuracy gen 38.71

ceval-environmental\_impact\_assessment\_engineer c64e2d accuracy gen 51.61

ceval-tax\_accountant 3a5e3c accuracy gen 36.73

ceval-physician 6e277d accuracy gen 42.86

ceval-stem - naive\_average gen 39.21

ceval-social-science - naive\_average gen 57.43

ceval-humanities - naive\_average gen 50.23

ceval-other - naive\_average gen 44.62

ceval-hard - naive\_average gen 32

ceval - naive\_average gen 46.19