OpenConmpass（司南）

一、为什么要做评测这件事情?

1、面向未来拓展能力维度

> 评测体系需增加新能力维度，比如数学，复杂推理，洛基推理，代码和智能体等，以全面评估模型性能

2、扎根通用能力，聚焦垂直行业

> 在医疗金融法律等专业领域，评测需要介个行业指示和规范贸易评估模型的行业实用性

3、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

4、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

二、大预言模型评测中的挑战

1、全面性

> 大模型应用场景千变万化
> 模型能力演进迅速
> 如何世纪和构造可拓展的能力维度体系

2、评测成本

> 评测数十万道题需要大量算力资源
> 基于人工打分的主管评测成本高昂

3、数据污染

> 海量语料不可避免带来评测集污染
> 需要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

4、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

如何启动openconmpass?

①、先clone相关代码

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

②、安装环境

```
pip install -r requirements.txt
```

③、准备数据集

```
cp yourdata.zip /root/opencompass/
unzip yourdata.zip
python tools/list_configs.py internlm ceval  # 可以查看相关配置
```

④、启动评测

```
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
```