PR #24079 完整报告

sgl-project/sglang

[Bench] fix bench_hf.py KeyError + reduce print spam + add --limit

合并时间：2026-04-30 02:43

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24079

执行摘要

- 一句话：修复 bench_hf.py KeyError 并添加 --limit 参数
- 推荐动作：建议合并，变更小且明确修复 bug 并提升开发效率。对于关注 MMMU 评测流程的开发者可快速浏览变更。

功能与动机

PR body 明确指出：Fix the KeyError: 'original_response' bug。同时通过添加 --limit 参数便于调试和冒烟测试。

实现拆解

1. 修复 KeyError：在 InternVL 分支（第 102 行）和非 InternVL 分支（第 150 行）的 process_result 调用前，添加 sample["original_response"] = response 赋值，确保 process_result 能正常访问该键。
2. 减少日志噪音：删除两处 print(f"response: {response}")，避免每次推理都打印冗长的 response 内容。
3. 添加 --limit 参数：在 samples = prepare_samples(eval_args) 之后，通过 if getattr(args, "limit", None): 截取前 N 个样本，并打印提示信息。同时在 argparse 中添加 --limit 参数。

关键文件：

- benchmark/mmmu/bench_hf.py（模块 基准测试；类别 source；类型 core-logic；符号 eval_mmmu）：唯一变更文件：修复 KeyError、减少打印噪音、添加 --limit 参数。

关键符号：eval_mmmu

关键源码片段

benchmark/mmmu/bench_hf.py

唯一变更文件：修复 KeyError、减少打印噪音、添加 --limit 参数。

```
    # 在调用 process_result 前设置 sample["original_response"] 以避免 KeyError
    # 同时移除冗余的 print 语句以减少日志噪音

    if "InternVL" in args.model_path:
```

```
            # ... InternVL 处理 ...
            sample["original_response"] = response # 新增：确保键存在
            process_result(response, sample, answer_dict, out_samples)
            continue

        # ... 其他模型处理 ...
        sample["original_response"] = response # 新增：确保键存在
        process_result(response, sample, answer_dict, out_samples)
```

评论区精华

仅有一条来自 gemini-code-assist[bot] 的自动提示（每日配额限制），无人工讨论。后续 reviewer yhyang201 直接批准，无未解决问题。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅 1 个文件变更，+12/-2，逻辑简单。--limit 参数默认 None，不影响原有行为；修复 KeyError 使流程正确；移除冗余 print 无副作用。
- 影响：影响范围仅限于 benchmark/mmmu/bench_hf.py 脚本使用者。提高脚本稳定性和调试便利性。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR