PR #27327 完整报告

sgl-project/sglang

Fix MMMU VLM eval max_tokens for CoT prompt

合并时间：2026-06-05 10:28

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27327

执行摘要

- 一句话：修复 MMMU VLM 评估 max_tokens 过短导致全部失败
- 推荐动作：建议合入。这是一个有明确回归根因的测试修复，变更量极小，且已在 H200 上验证所有模型通过阈值。合并后应观察夜间测试是否稳定通过。

功能与动机

PR #21841 改用了 CoT 提示（"Think step by step before answering"），但 max_tokens=30 太短，模型在输出 "Answer: X" 前无法完成推理，导致所有 VLM 模型在夜间 MMMU 准确率测试中全部失败。

实现拆解

仅修改一行：在 test/registered/eval/test_vlms_mmmu_eval.py 的 SimpleNamespace 配置中，将 max_tokens 参数从 30 改为 1024。

关键文件：

- test/registered/eval/test_vlms_mmmu_eval.py（模块 测试；类别 test；类型 test-coverage）：唯一变更文件，将 MMMU 评估的 max_tokens 从 30 改为 1024，修复全部 VLM 模型因 CoT 提示 token 不足而失败的回归问题。

关键符号：未识别

关键源码片段

test/registered/eval/test_vlms_mmmu_eval.py

唯一变更文件，将 MMMU 评估的 max_tokens 从 30 改为 1024，修复全部 VLM 模型因 CoT 提示 token 不足而失败的回归问题。

```
# test/registered/eval/test_vlms_mmmu_eval.py 第 94 行

args = SimpleNamespace(
    base_url=self.base_url,
    model=model_path,
    eval_name="mmmu",
    num_examples=100,
    num_threads=64,
```

```
    max_tokens=1024, # 修复：PR#21841 改为 CoT 提示后，原 30 个 token 不足以完成推理，
                     # 导致所有 VLM 模型夜间测试失败。
                     # 增加到 1024 后，所有模型均通过准确率阈值。
)
```

评论区精华

无讨论。PR 没有 review 评论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。仅修改测试配置参数，不影响任何生产逻辑。如果模型真实需要更多 token，1024 已足够；如果测试耗时过长，可后续再调优，但当前优先保证准确率测试通过。
- 影响：直接影响夜间 VLM MMMU 准确率测试，使所有模型从测试失败变为通过（如 Qwen2.5-VL-7B 从 0.28 升至 0.56）。不影响用户，不影响生产推理。
- 风险标记：低风险 仅测试参数调整

关联脉络

- PR #21841 [placeholder_title]: 该 PR 将 MMMU 评估提示改为 CoT 格式，但未相应增加 max_tokens，导致本 PR 修复的回归问题。