PR #24279 完整报告

sgl-project/sglang

[CI] Temporarily disable marco/mcdse-2b-v1 in test_embedding_models

合并时间：2026-05-07 05:28

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24279

执行摘要

- 一句话：临时禁用闪崩的 mcdse 嵌入测试模型
- 推荐动作：建议快速合并以稳定 CI。PR 本身无争议，且附带了完善的根因分析和复现证据。值得关注的是其揭示的 sm_120 内核兼容性问题，后续修复 PR 应重点审查。

功能与动机

test_embedding_models.py 中 random.choice(MODELS) 选中 marco/mcdse-2b-v1 时，CI 随机失败（余弦相似度差异 ~0.30 vs HF 参考），阻塞 main 分支 PR 测试流水线。原作者在 PR body 中通过诊断 PR #24327 确认该模型是唯一罪魁祸首，并排除其他因素。

实现拆解

1. 修改模型配置字典：在 test/registered/prefill_only/test_embedding_models.py 中，将 marco/mcdse-2b-v1 条目注释掉，保留其他三个模型以维持测试覆盖率。
2. 添加详细注释：在注释代码旁说明禁用原因（HF 参考路径使用双向注意力而 SGLang 始终使用因果注意力，导致短提示词余弦差异 ~0.30），并附带失败 CI 运行链接供后续参考。
3. 回归验证：通过 CI 运行确认禁用后所有测试分区通过；同时通过诊断 PR #24327 精确复现原失败，验证定位正确。

关键文件：

- test/registered/prefill_only/test_embedding_models.py（模块 测试；类别 test；类型 test-coverage）：唯一变更文件：注释掉 marcro/mcdse-2b-v1 模型条目，并添加详细注释说明禁用原因和关联链接。

关键符号：未识别

关键源码片段

test/registered/prefill_only/test_embedding_models.py

唯一变更文件：注释掉 marcro/mcdse-2b-v1 模型条目，并添加详细注释说明禁用原因和关联链接。

```
# test/registered/prefill_only/test_embedding_models.py (modified)

MODEL_TO_CONFIG = {
```

```
    "Alibaba-NLP/gte-Qwen2-1.5B-instruct": (1, 1e-5),
    "intfloat/e5-mistral-7b-instruct": (1, 1e-5),
    # Temporarily disable: HF reference path in runners.py runs this Qwen2-VL
    # fine-tune with bidirectional attention (the non-sentence-transformers
    # branch in _get_sentence_transformer_embedding_model does not pass
    # is_causal=True), while SGLang's Qwen2-VL embedding is always causal —
    # producing ~0.30 cosine diffs vs HF on short prompts.
    # See https://github.com/sgl-project/sglang/actions/runs/25224929325/job/73966043206
    # "marco/mcdse-2b-v1": (1, 1e-5),
    "Qwen/Qwen3-Embedding-8B": (1, 1e-5),
    # Temporarily disable before this model is fixed
    # "jason9693/Qwen2.5-1.5B-apeach": (1, 1e-5),
}
```

评论区精华

无实质性讨论，只有 gemini-code-assist[bot] 的自动总结和 b8zhong 的批准。原作者在 PR body 中详尽分析了根因，但未引发进一步技术讨论。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。仅测试配置变更，不影响任何生产代码或模型推理路径。禁用 mcdse 后，CI 对 sm_120 + Qwen2-VL 内核路径的信号丢失，但这本就是不可靠测试，且原作者提出了明确的复现和修复计划。
- 影响：用户：无影响。系统：CI 回归测试不再因 mcdse 随机失败，提高流水线可靠性。团队：需跟踪 #24327 等后续修复，待底层 sm_120 内核问题解决后重新启用该模型。
- 风险标记：测试覆盖调整，临时禁用

关联脉络

- PR #24327 Diagnostic: force mcdse-2b-v1 in test_embedding_models: 诊断 PR，确认 mcdse 是唯一导致 CI 失败的模型；本 PR 的直接依据。