PR #21898 完整报告

sgl-project/sglang

[CI] Remove crashing Kimi K2.5 EAGLE3/MTP variants, keep TP8 and TP8+DP8

合并时间：2026-04-02 11:27

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21898

执行摘要

- 一句话：移除导致夜间测试崩溃的 Kimi K2.5 MTP 变体，保留 TP8 和 TP8+DP8 配置。
- 推荐动作：该 PR 值得快速浏览，以了解 CI 测试配置的调整。关注点：1) 移除 MTP 变体的具体原因（OOM 和未知错误）；2) 新增 TP8+DP8 变体的配置；3) 测试覆盖范围的变化。对于负责 CI 或测试的工程师，建议检查是否有其他测试需要类似调整。

功能与动机

根据 PR body 描述，Kimi K2.5 的 MTP 变体在夜间测试中持续崩溃：TP8+MTP 变体因 OOM 被杀死（退出码 -9），TP8+DP8+MTP 变体因未知错误退出（退出码 1），在 H200 和 B200 GPU 上均出现此问题。移除这些不稳定变体是为了确保 CI 测试的可靠性，避免因测试崩溃导致的 CI 失败。

实现拆解

该 PR 仅修改了一个文件：test/registered/8-gpu-models/test_kimi_k25.py。主要改动包括：1) 移除与 EAGLE3 推测解码相关的配置（如 EAGLE3_DRAFT_MODEL_PATH 变量、eagle3_args 参数列表）；2) 更新测试类文档，从描述两个变体（基础版和 eagle3 版）改为仅描述基础版；3) 从测试用例的 model_settings 列表中移除两个 MTP 变体（TP8+MTP 和 TP8+DP8+MTP），并新增一个 TP8+DP8 变体（使用 base_args + dp_attn_args 参数，无推测解码）。

关键文件：

- test/registered/8-gpu-models/test_kimi_k25.py（模块 测试）：唯一修改的文件，移除了崩溃的 MTP 变体测试配置，并新增了 TP8+DP8 变体。

关键符号：TestKimiK25.test_kimi_k25

评论区精华

该 PR 没有 review 评论，仅有一条由作者提交的 PR body 描述。因此，没有关于设计权衡、争议或未解决疑虑的讨论。

- 暂无高价值评论线程

风险与影响

- 风险：技术风险较低：1) 回归风险：移除 MTP 变体可能减少对推测解码功能的测试覆盖，但保留了基础 TP8 和新增的 TP8+DP8 变体，核心功能测试仍在。2) 兼容性风险：无，因为这是测试配置的调整，不影响生产代码。3) 性能风险：无，不涉及性能优化。主要风险是测试覆盖范围可能不足，特别是对 EAGLE3 推测解码的测试缺失。
- 影响：对用户无直接影响，因为这是 CI 测试的内部调整。对系统影响：提高 CI 测试的稳定性，避免因测试崩溃导致的 CI 失败。对团队影响：简化测试配置，减少维护负担，但可能需后续补充推测解码的测试覆盖。影响范围限于测试基础设施，程度为低到中。
- 风险标记：测试覆盖减少

关联脉络

- PR #21767 [CI] add nvfp4 ci test for b200;: 同属 CI 测试配置调整，涉及多模态生成测试，但本 PR 专注于移除不稳定测试变体。
- PR #21890 Allow /rerun-test to checkout fork PR branch for trusted users: 同属 CI 基础设施改进，但本 PR 更侧重于测试稳定性维护。
- PR #21882 Add merge prohibition policy during CI maintenance mode: 同属 CI 维护相关，本 PR 是具体测试配置调整以提升稳定性。