PR #22188 完整报告

sgl-project/sglang

[AMD] Fix test_kimi_k25_mxfp4.py : stage-c-test-large-8-gpu-amd-mi35x (linux-mi35x-gpu-8, 1)

合并时间：2026-04-08 04:48

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22188

执行摘要

- 一句话：通过固定模型版本修复 AMD CI 中 Kimi-K2.5-MXFP4 测试的权重加载失败问题。
- 推荐动作：该 PR 不值得精读，除非您负责 AMD CI 维护或 Kimi 模型集成。它只是一个简单的临时修复，设计决策单一（固定版本）。关注点在于团队如何处理上游依赖变更和 CI 稳定性，以及后续 PR 将如何解决根本问题。

功能与动机

PR body 明确指出，HuggingFace 模型更新（commit 94d8c1bd）量化了额外层，并将排除列表格式从正则表达式改为显式的每层名称，这暴露了 KimiK25ForConditionalGeneration 中的前缀不匹配问题，导致权重加载时出现 AssertionError。为了快速修复 CI 失败，选择将模型版本固定到最后一个已知良好的修订版本（b071bc6f）。

实现拆解

实现非常简单，仅修改了一个测试文件：在 test/registered/amd/test_kimi_k25_mxfp4.py 中添加了一个常量 KIMI_K25_MXFP4_REVISION，并将其作为 --revision 参数传递给测试服务器的启动参数。这确保了 CI 测试使用固定的模型版本，避免了上游变更的影响。

关键文件：

- test/registered/amd/test_kimi_k25_mxfp4.py（模块 测试 /AMD）：唯一被修改的文件，添加了模型版本常量并更新服务器启动参数，直接修复了 CI 失败。

关键符号：未识别

评论区精华

讨论主要集中在 HaiShaw 的评论上，他最初要求 "justify common code change from one specific quark model made, need strong justification to accept"，表明对为特定模型修改通用代码的担忧。但随后 yctseng0211 在 Issue 评论中澄清，已回滚所有对 kimi_k25.py 的更改，现在唯一的变化是将模型版本固定到已知良好版本，这 "zero risk to existing models"。HaiShaw 随后批准了 PR。这表明团队接受了这是一个低风险的临时修复，根本问题将在后续 PR 中解决。

- 通用代码修改的合理性 (design): 作者澄清已回滚模型文件更改，仅固定测试版本，风险为零，因此获得批准。

风险与影响

- 风险：风险极低：1) 仅影响 AMD CI 中的特定测试用例，不涉及生产代码或核心逻辑。2) 通过固定模型版本，避免了上游模型变更带来的不确定性，确保了测试的稳定性。3) 作者在 Issue 评论中明确表示 "zero risk to existing models"，且已回滚了最初对模型文件的修改。唯一潜在风险是如果固定的模型版本在未来因其他原因（如安全漏洞）变得不可用，但这是一个可管理的 CI 依赖问题。
- 影响：影响范围有限：1) 对用户无直接影响，这是一个内部测试修复。2) 对系统的影响仅限于恢复 AMD CI 中特定测试的通过状态，确保 CI 流水线的健康。3) 对团队的影响是解决了阻塞 CI 的失败，允许其他开发工作继续进行，同时为根本修复争取了时间。影响程度为低，因为不改变任何功能或性能。
- 风险标记：CI 依赖固定

关联脉络

- PR #21669 [AMD] Add Qwen3.5-397B FP8 nightly perf benchmarks for MI30x and MI35x: 同属 AMD 相关测试，涉及 AMD CI 和性能基准测试，共享类似的基础设施关注点。
- PR #22024 [NPU] enable mla prepare fused kernel only when being mla attn: 类似模型特定修复，但针对 NPU 后端；本 PR 是 AMD 测试的临时修复，都涉及硬件特定问题。