PR #21518 完整报告

sgl-project/sglang

[AMD] Fix Handle missing rope_theta in get_rope_config for Grok-1

合并时间：2026-04-01 01:58

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21518

执行摘要

本 PR 修复了 Grok-1 AMD INT4 模型加载时因缺失 rope_theta 属性导致的崩溃问题。通过将 grok.py 中的 get_rope_config 调用替换为本地属性提取逻辑，恢复了之前的行为并支持 Transformers v5。变更仅影响 AMD 特定的 Grok-1 INT4 模型，解决了 CI 测试失败。

功能与动机

AMD 夜间 CI 测试 nightly-8-gpu-grok1-int4 和 nightly-8-gpu-mi35x-grok1-int4 在加载 Grok-1 INT4 模型时崩溃，错误为 AttributeError: 'Grok1Config' object has no attribute 'rope_theta'。这是由于 PR #21135 将 grok.py 从安全的 getattr(config, "rope_theta", 10000) 迁移到共享的 get_rope_config(config) helper 函数，而 Grok-1 INT4 模型使用自定义 Grok1Config 类，未定义 rope_theta 属性。

实现拆解

只修改了 python/sglang/srt/models/grok.py 文件：

- 移除对 get_rope_config 的导入。
- 在类的 __init__ 方法中，将 rope_theta, _ = get_rope_config(config) 替换为本地逻辑：

```
rope_theta = getattr(config, "rope_theta", None)
if rope_theta is None:
    rope_params = getattr(config, "rope_parameters", None)
    rope_theta = rope_params["rope_theta"] if rope_params else 10000
```

这首先尝试 v4 路径的 rope_theta，然后 v5 路径的 rope_parameters["rope_theta"]，最后默认 10000。

评论区精华

Review 讨论较少，reviewer yctseng0211 直接批准。在 Issue 评论中，作者 michaelzhang-ai 强调了变更范围有限："no NVIDIA / common code impact"，并确认了 Grok-1 FP8 和 Grok-2 不受影响。yctseng0211 更新分支以包含 PR 21547 解决 stage-a failure。

风险与影响

- 风险：逻辑简单，回归风险低；但依赖于 rope_parameters 字典的存在，如果其他模型配置类似缺失，可能未覆盖。未添加单元测试，长期维护可能脆弱。
- 影响：仅修复 Grok-1 AMD INT4 模型加载，使相关 CI 测试通过。无其他模型或代码路径受影响，因为 grok.py 是 Grok-1 专用。

关联脉络

本 PR 直接关联到 #21135，后者引入了导致问题的变更。从近期历史 PR 看，涉及 AMD 的 PR 如 #21657 优化了 MoEGate 性能，但本 PR 专注于 bugfix，显示了在硬件特定模型配置处理上的持续改进。