

PR #54745 完整报告

vllm-project/vllm

[CI] Disable CUDA graphs for GLM PCP evals

合并时间: 2026-09-02 04:24

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/54745>

执行摘要

- 一句话: 禁用 GLM PCP 评估配置中的 CUDA 图, 以修复 MRV2 稀疏 MLA PCP 的启动崩溃。
- 推荐动作:

功能与动机

修复 CI 通道阻塞。PR body 明确指出, 夜间构建 #86511 中, 两个 GLM PCP 配置在引擎初始化阶段失败, 报错信息为 'MRV2 sparse MLA PCP does not support CUDA graphs yet. Set -cc.cudagraph_mode=NONE.', 导致 B200 LM Eval PCP 通道在 GSM8K 评估开始前即告失败。该变更旨在恢复这些评估任务的可执行性。

实现拆解

本次变更为纯配置修改, 不涉及任何源码逻辑改动。

1. 变更入口与范围: 变更仅涉及两个特定的测试评估配置文件:
tests/evals/gsm8k/configs/GLM-5.2-NVFP4-TP1-PCP4-EP.yaml 和
tests/evals/gsm8k/configs/GLM-5.2-NVFP4-TP2-PCP2-EP.yaml。这两个文件定义了
在特定硬件配置下运行 GLM-5.2-NVFP4 模型 GSM8K 评估的服务器参数。
2. 核心配置改动: 在两个文件的 server_args 参数中, --compilation-config 的 JSON 字符串
值被修改。具体是将 cudagraph_mode 的值从不兼容的 `

关键文件:

- tests/evals/gsm8k/configs/GLM-5.2-NVFP4-TP1-PCP4-EP.yaml (模块 GLM-5.2-NVFP4-TP1-PCP4-EP; 类别 test; 类型 test-coverage): 测试配套; 包含测试覆盖调整; +1/-1
- tests/evals/gsm8k/configs/GLM-5.2-NVFP4-TP2-PCP2-EP.yaml (模块 GLM-5.2-NVFP4-TP2-PCP2-EP; 类别 test; 类型 test-coverage): 测试配套; 包含测试覆盖调整; +1/-1

关键符号: 未识别

评论区精华

- 暂无高价值评论线程

风险与影响

- 风险:
- 影响:
 - 风险标记: 暂无

关联脉络

- 暂无明显关联 PR