PR #22399 完整报告

sgl-project/sglang

[CI] Add GLM-5.1 nightly tests and update Qwen3.5 model

合并时间：2026-04-09 08:04

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22399

执行摘要

本 PR 为 sglang 项目 CI 系统新增 GLM-5.1 FP8 模型的夜间测试，覆盖 H200/B200 GPU 集群的三种并行配置变体，同时更新 Qwen3.5 测试至 FP8 版本并集成数据并行注意力机制。这些变更扩展了大模型测试覆盖，确保 CI 与最新模型版本和并行策略同步，对开发团队验证模型兼容性和性能有积极影响，但需注意外部模型依赖和测试资源消耗的风险。

功能与动机

PR 的主要目标是扩展夜间测试套件，以覆盖新发布的 GLM-5.1 FP8 模型和更新后的 Qwen3.5 FP8 模型。根据 PR body，具体动机包括：

- 添加 GLM-5.1 FP8 测试：为 H200/B200 GPU 集群（nightly-8-gpu-common 套件）新增测试，包含 TP8、TP8+DP8 和 TP8+DP8+MTP 三种变体，以验证不同并行配置下的模型表现。
- 更新模型引用：将 GB300 测试中的 GLM-5 模型升级至 GLM-5.1，确保测试使用最新版本。
- 升级Qwen3.5测试：将模型路径更新为FP8精度版本（Qwen/Qwen3.5-397B-A17B-FP8），并集成来自 PR #22288 的 DP-attention 变体，增强测试覆盖。

这些变更是为了保持 CI 测试与模型演进的同步，防止因模型过时而产生测试缺口。

实现拆解

实现涉及三个测试文件的变更，按模块拆解如下：

文件路径	变更类型	关键改动	所属模块
test/registered/8-gpu-models/test_glm_51_fp8.py	新增	定义 GLM-5.1 FP8 测试类，包含三种并行变体：TP8、TP8+DP8、TP8+DP8+MTP，使用 gsm8k 数据集（baseline_accuracy=0.92）进行准确性和性能测试。	CI测试

文件路径	变更类型	关键改动	所属模块
test/registered/8-gpu-models/test_qwen35.py	修改	更新模型路径至 FP8 版本，添加 DP-attention 变体（--dp=8 --enable-dp-attention），扩展测试覆盖至 TP8+DP8 和 TP8+DP8+MTP 配置。	CI测试
test/registered/gb300/test_glm5_fp8.py	修改	将模型路径从 zai-org/GLM-5-FP8 更新为 zai-org/GLM-5.1-FP8，同步更新测试名称和文档字符串。	CI测试

关键代码逻辑示例（来自 test_glm_51_fp8.py）：

```
variants = [
    ModelLaunchSettings(
        GLM_51_FP8_MODEL_PATH,
        tp_size=8,
        extra_args=COMMON_ARGS,
        variant="TP8",
    ),
    ModelLaunchSettings(
        GLM_51_FP8_MODEL_PATH,
        tp_size=8,
        extra_args=COMMON_ARGS + dp_args,
        variant="TP8+DP8",
    ),
    ModelLaunchSettings(
        GLM_51_FP8_MODEL_PATH,
        tp_size=8,
        extra_args=COMMON_ARGS + dp_args + MTP_ARGS,
        variant="TP8+DP8+MTP",
        env={"SGLANG_ENABLE_SPEC_V2": "1"},
    ),
]
```

评论区精华

由于 review_comments_count 为 0，没有记录 review 讨论。但从提交历史中可推断一个潜在讨论点：第二个提交“Revert GLM-5.1 naming in test_glm5_nvfp4.py”表明，在初始实现后可能发现 GLM-5.1 NVFP4 模型不存在，因此回退了相关命名更改以保持一致性。这提示了模型命名验证的重要性，但具体讨论细节未公开。

风险与影响

风险分析：

1. 外部模型依赖：测试依赖 Hugging Face 模型仓库（如 zai-org/GLM-5.1-FP8），若模型被删除或更新，可能导致 CI 失败。
2. 新增并行变体复杂度：DP-attention 变体（--enable-dp-attention）可能引入未充分测试的代码路径，增加调试难度。
3. 测试时间增长：夜间测试估计时间较长（1800 秒），可能影响 CI 资源利用和反馈速度。
4. 模型版本差异：GLM-5 到 GLM-5.1 的更新可能带来行为变化，需确保准确度阈值（0.92）仍适用。

影响分析：

- 对 CI 系统：扩展测试覆盖，提升对最新模型和并行策略的验证能力，但可能增加夜间 CI 负载。
- 对开发团队：提供更全面的模型兼容性反馈，有助于及早发现回归问题。
- 对用户：无直接影响，属于内部测试改进。

关联脉络

从近期历史 PR 看，本 PR 与多个 CI 测试优化 PR 相关：

- PR #22288：Qwen3.5 的 DP-attention 变体来源，表明这是功能集成的一部分。
- PR #22346 和 #22237：同属 CI 测试调整，关注资源分配和准确度阈值，反映项目对测试稳定性和准确性的持续优化。
- 更广泛的趋势：近期 PR（如 #22400、#22395）显示项目正加强 CI 效率和覆盖，本 PR 延续了这一方向，专注于大模型测试的扩展和更新。

整体上，本 PR 是 sglang 项目 CI 测试演进的一部分，旨在确保测试套件与快速迭代的模型生态保持同步。