PR #22285 完整报告

sgl-project/sglang

Add CI tests for GLM-5

合并时间：2026-04-08 16:05

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22285

执行摘要

本 PR 为 GLM-5 模型添加 CI 测试，通过重命名现有测试文件并新增测试类，扩展了 8-GPU 测试套件。变更包括数据并行、张量并行和推测解码配置的准确性及速度基准测试，旨在增强模型测试覆盖，确保推理稳定性。

功能与动机

动机源于扩展 CI 测试以覆盖 GLM-5 模型的需求。PR 标题直接指出“Add CI tests for GLM-5”，虽无详细描述，但从同仓库历史 PR（如 #21710 涉及 GLM-5 性能基准）推断，此举旨在加强模型测试矩阵，预防回归。

实现拆解

实现集中在两个测试文件：

- test/registered/8-gpu-models/test_dsa_models_basic.py：重命名自 test_deepseek_v32_basic.py，添加 TestGLM5DP 和 TestGLM5TP 类，测试 8-GPU DP 和 TP 配置。
- 关键配置：模型路径 zai-org/GLM-5-FP8，超参数如 --tp 8、--dp 8。
- 测试方法：test_a_gsm8k（GSM8K 准确性）和 test_bs_1_speed（速度基准）。
- test/registered/8-gpu-models/test_dsa_models_mtp.py：重命名自 test_deepseek_v32_mtp.py，添加 TestGLM5DPMTP 和 TestGLM5TPMTP 类，支持 EAGLE 推测解码测试。
- 变更包括启用 SGLANG_ENABLE_SPEC_V2 环境变量，调整内存分数至 0.8。

评论区精华

无 review 评论，讨论为空。

风险与影响

风险：

- 测试阈值设置（如速度阈值 40 token/s）可能不合理，导致 CI 误报。
- 配置调整（如 --mem-frac 0.8）可能影响测试稳定性。
- CI 运行时间预估从 360 秒增至 720 秒，增加资源消耗。

影响：

- 正面：提升 GLM-5 模型测试覆盖，助益回归检测。
- 负面：新增测试可能延长 CI 流水线，但对用户无直接冲击。

关联脉络

从历史 PR 看：

- 21710 为 AMD 平台添加 GLM-5-FP8 夜间性能基准测试，与本 PR 共同扩展 GLM-5 测试生态。
- 22288 更新 H200/B200 测试模型，虽被撤销，但反映 CI 测试模型持续演进趋势。

本 PR 是 SGLang 测试基础设施常规扩展的一部分，强调对新兴模型如 GLM-5 的支持。