PR #21733 完整报告

sgl-project/sglang

[CI]Remove msgm-en and mmlu tests which cause timeout

合并时间：2026-03-31 16:10

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21733

执行摘要

本 PR 通过移除导致超时的 msgm-en 和 mmlu 测试，并将它们替换为 gsm8k 测试，优化了 CI 配置以提高稳定性。变更涉及 CI 工作流调整、依赖安装统一和多个测试文件更新，属于常规维护，风险较低但需关注测试覆盖变化。

功能与动机

动机：解决 CI 测试中的超时问题，确保测试流程更可靠。PR 标题明确指出“Remove msgm-en and mmlu tests which cause timeout”，从 Issue 评论中作者多次重跑测试验证通过（如使用 /rerun-ut 命令），表明这些测试在 CI 中频繁超时，影响开发效率。

实现拆解

主要变更点如下：

- CI 工作流文件：在 .github/workflows/pr-test.yml 中移除重复的 human-eval 安装步骤，简化流程。
- 依赖脚本：在 scripts/ci/cuda/ci_install_dependency.sh 中添加 human-eval 安装，统一管理依赖。
- 测试文件：在多个测试文件（如 test_data_parallelism.py、test_dp_attention.py、test_moe_eval_accuracy_large.py、test_moe_ep.py）中：
  - 将 MMLUMixin 和 MGSMEnMixin 替换为 GSM8KMixin。
  - 更新测试方法，例如将 test_mgsm_en 改为 test_gsm8k，并使用 run_eval_few_shot_gsm8k 函数。
  - 设置新的准确性阈值，如 gsm8k_accuracy_thres = 0.6。

评论区精华

由于没有正式的 review 讨论，评论区精华有限。Issue 评论中显示作者通过命令验证变更：

> Fridge003: /rerun-ut test_moe_eval_accuracy_large.py github-actions[bot]: 🔑 2-gpu-h100: View workflow run 这表示变更经过初步测试通过，但缺乏深度技术权衡或争议讨论。

风险与影响

- 技术风险：
  - 测试覆盖减少：移除 mmlu 和 msgm-en 测试可能遗漏模型在知识推理和数学问题上的性能问题。
  - 依赖安装变更：将 human-eval 安装移到脚本中，可能影响其他 CI 作业，需确保安装顺序无误。
  - 新测试准确性：gsm8k 测试的阈值（如 0.6）未经验证，可能引入评估偏差。
- 影响范围：
  - 对 CI：提高稳定性和效率，减少超时失败。
  - 对测试质量：测试重点从多领域评估转向数学问题，可能需补充其他评估以保持全面性。
  - 对团队：简化维护，但需监控测试结果并调整策略。

关联脉络

从历史 PR 分析，PR 21714 “Fix human-eval CI install on 5090 runners” 与本 PR 相关，因为它同样修复 human-eval 安装问题。这表明仓库近期在持续优化 CI 测试流程，以解决依赖和超时挑战。本 PR 进一步调整测试用例，反映了从特定数据集评估向更稳定测试的演进趋势。