PR #21931 完整报告

sgl-project/sglang

[CI] Migrate mgsm_en eval to gsm8k to remove openaipublic dependency

合并时间：2026-04-08 07:29

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21931

执行摘要

- 一句话：迁移 CI 测试数据集从 MGSM-EN 到 GSM8K，移除外部依赖并调整阈值。
- 推荐动作：建议 CI 维护者、测试工程师和关注模型准确性的开发者精读，重点关注阈值校准策略和依赖管理决策；对一般工程师，了解变更背景即可，无需深入代码细节，但可参考如何优化 CI 稳定性。

功能与动机

PR body 中指出，多个 CI 测试运行时从 https://openaipublic.blob.core.windows.net/ 下载数据集，这是一个由 OpenAI 托管的外部 Azure blob，导致 CI 依赖脆弱——如果端点慢、限速或不可用，测试会因无关原因失败。MGSM-EN 数据集下载 12 个 TSV 文件且无缓存，而 GSM8K 使用 download_and_cache_file 缓存到 /tmp/test.jsonl，首次获取后离线运行，旨在消除 CI 不稳定因素。

实现拆解

实现分为两部分：一是将 8 个测试文件中的 eval_name="mgsm_en" 统一替换为 eval_name="gsm8k"，涉及分布式、量化、AMD 准确性和调度器测试；二是基于 GSM8K（5-shot/CoT）基准重新校准阈值，将所有阈值设置为基准值的 95% 以考虑硬件差异，并修正了如 Mistral-7B、Llama-3.1-70B 等模型的错误阈值；同时调整测试参数，如移除硬编码 num_examples 以使用默认值。

关键文件：

- test/registered/amd/accuracy/mi30x/test_gsm8k_eval_amd.py（模块 testing）：阈值调整最全面，涉及 AMD GPU 多模型测试，涵盖 FP8 和 FP16 变体，是迁移的核心文件之一。
- test/registered/eval/test_text_models_gsm8k_eval.py（模块 testing）：通用文本模型 GSM8K 评估文件，阈值更新影响广泛，包括多个主流模型，是 CI 测试套件的关键部分。
- test/registered/quant/test_quantization.py（模块 testing）：量化模型测试文件，阈值重新校准以匹配 GSM8K 格式，影响 AWQ 和 GPTQ 等量化方法的准确性验证。
- test/registered/piecewise_cuda_graph/test_piecewise_cuda_graph_support_1_gpu.py（模块 testing）：CUDA 图测试文件，调整参数和阈值以适配 GSM8K，用于验证 piecewise CUDA graph 不影响模型准确性。

关键符号：test_gsm8k, test_gsm8k_all_models, test_1_gsm8k_has_prefill_delayer, test_2_gsm8k_no_prefill_delayer

评论区精华

review 中仅有一个 bot 评论，由 gemini-code-assist[bot] 发表，确认变更内容无误，表示 ' 没有反馈可提供 '，表明变更被简单接受，没有实质性讨论、争议或未解决疑虑。

- 变更确认 (other): 变更被接受，无需修改。

风险与影响

- 风险：风险包括：阈值调整基于 GSM8K 基准，但校准可能不准确，导致测试假阳性（过于宽松，遗漏回归）或假阴性（过于严格，误报失败）；GSM8K 数据集虽然缓存，但 GitHub 可用性也可能影响 CI 稳定性；迁移后测试覆盖范围是否与 MGSM-EN 一致存在不确定性，可能遗漏某些模型行为或边缘案例。具体文件如 test/registered/amd/accuracy/mi30x/test_gsm8k_eval_amd.py 中的阈值变化需谨慎验证。
- 影响：对终端用户无直接影响，主要影响 CI 流程和团队开发体验；系统层面，移除外部依赖提高 CI 稳定性、可重复性和离线能力，减少因网络问题导致的失败；团队层面，开发者将受益于更可靠的测试套件，加速代码合并流程，但需适应新阈值可能带来的测试结果变化。
- 风险标记：阈值校准风险，外部依赖移除，测试覆盖一致性

关联脉络

- PR #22288 [CI] Update nightly test models for H200/B200: 同为 CI 测试更新，涉及模型和依赖调整，反映 CI 基础设施的持续优化趋势。
- PR #22188 [AMD] Fix test_kimi_k25_mxfp4.py : stage-c-test-large-8-gpu-amd-mi35x (linux-mi35x-gpu-8, 1): 修复 AMD 测试中的权重加载问题，与本 PR 的 AMD 准确性测试相关，共同提升 AMD 平台 CI 稳定性。
- PR #22282 [tiny] migrate /get_server_info; print accept length in accuracy tests: 迁移废弃端点并在精度测试中打印信息，类似本 PR 的测试基础设施更新，显示团队对 CI 细节的改进关注。