PR #22346 完整报告

sgl-project/sglang

[CI] Set RUNAI_STREAMER_MEMORY_LIMIT=0 for stage-b-test-1-gpu-small

合并时间：2026-04-08 17:23

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22346

执行摘要

本 PR 在 stage-b-test-1-gpu-small CI 作业中设置 RUNAI_STREAMER_MEMORY_LIMIT=0 环境变量，以解决 runai 模型加载测试因内存缓冲限制导致的性能瓶颈和超时问题。通过启用无限缓冲，将测试吞吐量从 42 秒 / 迭代提升至 3 秒 / 迭代，确保测试在 30 分钟超时内完成，提升 CI 可靠性。变更仅涉及 CI 工作流文件，风险低，不影响生产代码。

功能与动机

- 问题背景：test_runai_model_loader 测试从公共 GCS 桶流式加载 CodeGemma 2B 模型（5.6 GiB，165 张量）时，流式加载器的默认内存缓冲上限导致其在预取前阻塞消费，使吞吐量从 ~3 秒 / 迭代降至 ~42 秒 / 迭代，总耗时 18 分钟（5.3 MiB/s），引发 CI 作业超时（失败日志）。
- 解决方案：设置 RUNAI_STREAMER_MEMORY_LIMIT=0 以启用无限内存缓冲，避免阻塞，匹配已有工作流（如 diffusion-ci-gt-gen 和 pr-test-multimodal-gen）的做法，并遵循 PR #17636 的先例。

实现拆解

仅修改一个文件，具体变更如下：

文件路径	变更内容	作用
.github/workflows/pr-test.yml	在 stage-b-test-1-gpu-small 作业的 env 部分添加 RUNAI_STREAMER_MEMORY_LIMIT: 0	为特定 CI 作业设置环境变量，使流式加载器无内存缓冲限制

代码片段：

```
env:
  CONTINUE_ON_ERROR_FLAG: ${{ needs.check-changes.outputs.continue_on_error == 'true' && '--continue-on-error' || '' }}
  RUNAI_STREAMER_MEMORY_LIMIT: 0
```

评论区精华

Review 中无实质性技术讨论，仅 hnyls2002 批准 PR。Issue 评论显示验证过程：

- alexnails 执行 /rerun-stage stage-b-test-1-gpu-small 触发重跑。
- hnyls2002 执行 /rerun-test test_runai_model_loader.py --branch main (2 tries) 并确认成功（重跑链接）。

风险与影响

- 技术风险：
  - 无限内存缓冲可能增加 CI 节点的内存使用，但已有类似工作流采用相同设置，表明风险可控。
  - 变更仅限 CI 环境，不涉及生产代码，回归风险小。
  - 缺少内存使用量监控，但测试模型大小固定（5.6 GiB），影响有限。
- 影响分析：
  - 用户：无感知，不影响系统功能。
- 系统：提升 CI 测试稳定性，避免超时失败。
- 团队：减少 CI 重试成本，加快测试反馈循环。

关联脉络

- 历史 PR 关联：PR body 提及遵循 PR #17636，表明此环境变量设置已有先例，是 CI 优化的一致模式。
- 跨 PR 趋势：近期多个 PR（如 #22292、#22301、#22298）聚焦 CI 优化，包括测试超时修复、资源管理和工作流调整，本 PR 延续了这一基础设施改进方向，针对特定测试场景的性能调优。