PR #21740 完整报告

sgl-project/sglang

[CI] [Tracing] Add ci for tracing and fix bugs

合并时间：2026-04-03 01:50

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21740

执行摘要

本 PR 通过添加轻量级 OTLP 收集器和多种跟踪集成测试至 CI，显著提升 SGLang 跟踪功能的测试覆盖和稳定性，同时修复了 tokenizer manager 中的一个 bug，避免嵌入输出中的 token 计数错误。变更主要影响基础设施和测试流程，对用户透明但有益于开发者调试效率。

功能与动机

为什么做：根据 PR body 描述，目标是“实现简单的 OpenTelemetry Collector 并添加跟踪集成测试到 CI”，以取代依赖 Docker 的部署方式，简化测试环境并提高可靠性。此外，修复了一个追踪 bug，确保在处理 BatchEmbeddingOutput 时正确生成 token 使用属性。

实现拆解

按模块拆解改动：

1. bug 修复模块：修改 python/sglang/srt/managers/tokenizer_manager.py 中的 convert_to_span_attrs 函数，添加类型检查：python if not isinstance(recv_obj, BatchEmbeddingOutput): span_attrs[SpanAttributes.GEN_AI_USAGE_COMPLETION_TOKENS] = recv_obj.completion_tokens[i]
2. CI 依赖模块：更新 scripts/ci/cuda/ci_install_dependency.sh，在 EXTRAS 中添加“tracing”包，确保测试环境安装必要依赖。
3. 测试框架模块：移除旧手动测试 test/manual/test_tracing.py，新增自动化测试文件：
   - test/registered/observability/test_tracing.py：实现 LightweightOtlpCollector 类，支持 gRPC 接收和内存存储，覆盖 trace levels 0-3、批处理请求、并行采样等场景。
   - test/registered/observability/test_tracing_disaggregation.py：扩展至 PD 分离模式，验证分布式环境跟踪。
4. 文件组织模块：重命名 metrics 测试文件以统一 observability 目录结构。

评论区精华

review 讨论中最有价值的交锋：

- 安全绑定地址：gemini-code-assist[bot] 指出“绑定到 0.0.0.0 暴露测试收集器到本地网络”，建议改为 127.0.0.1 以提升安全性。

- 性能优化：同一 bot 建议“使用 ListFields() 仅提取设置字段”，减少数据冗余；并提议“减少 check_interval 从 2 秒到更短值”以加速 CI。
- 开发团队响应：sufeng-buaa 回应间隔优化，调整逻辑为“三次连续空检查”，平衡了测试速度和稳定性。

风险与影响

具体风险和影响分析：

- 风险：新测试可能轻微增加 CI 时间（本地约 130 秒），但优化后影响可控；绑定地址更改需确保所有连接端点一致，否则可能导致测试失败；protobuf 提取优化需验证数据完整性，避免断言错误。
- 影响：正面影响为主——提升跟踪测试覆盖，增强 CI 可靠性，修复 bug 防止数据错误；对开发者，简化测试部署（无需 Docker），但引入新测试框架需额外维护。

关联脉络

与历史 PR 和关联 Issue 的关系：

- 从近期历史 PR 看，本 PR 延续了 CI 改进趋势（如 PR 21950 修复 GPU 依赖），并关联测试基础设施优化（如 PR 21905 安全扫描）。
- 虽无直接关联 Issue，但体现了团队对 observability 模块（跟踪、指标）的持续投入，可能为未来调试和监控功能演进奠定基础。