PR #22523 完整报告

sgl-project/sglang

[Doc] correct the HTTP endpoint for stopping profiling in `benchmark_and_profiling.md`

合并时间：2026-04-17 00:54

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22523

执行摘要

- 一句话：修正性能剖析文档中停止剖析的 HTTP 端点名称。
- 推荐动作：该 PR 变更简单直接，仅修正文档错误，无需精读。但值得关注 review 中暴露的文档与实现不一致问题（如 start_step 参数），建议后续跟进全面文档审查。

功能与动机

根据 PR body 的描述，本次修改是为了修正 developer_guide/benchmark_and_profiling.md 文档中停止剖析的 HTTP 端点，将其从 /end_profile 更正为 /stop_profile。这是对文档错误的直接修复，以确保用户能够根据文档正确使用剖析功能。

实现拆解

1. 修改文档端点引用：在 docs/developer_guide/benchmark_and_profiling.md 文件中，将所有提及停止剖析的 HTTP 端点从 /end_profile 替换为 /stop_profile。 - 涉及文件：docs/developer_guide/benchmark_and_profiling.md - 关键变更：修改了端点名称的文本描述和示例代码中的 curl 命令。 - 原因：确保文档与实际后端实现的 API 一致，避免用户使用错误的端点。 - 影响：用户将根据正确的端点名称来停止剖析会话。
2. 发现文档与实现不一致：在 review 过程中，gemini-code-assist[bot] 指出文档中提到的 start_step 参数在后端 python/sglang/srt/utils/profile_utils.py 的 configure 方法中尚未支持（存在 assert start_step is None），这可能导致用户混淆或运行时错误。但本次 PR 未解决此问题，作者 cs-cat 建议后续通过新的 PR 全面审查文档。

关键文件：

- docs/developer_guide/benchmark_and_profiling.md（模块 开发者指南；类别 docs；类型 documentation）：这是本次 PR 唯一修改的文件，包含了性能剖析的详细指南，端点名称的更正直接影响用户操作。

关键符号：未识别

评论区精华

review 中的核心讨论围绕文档与后端实现的一致性展开：

- 端点名称更正：本次 PR 的主要目的是将 /end_profile 更正为 /stop_profile，这一变更得到了认可。

- 参数支持不一致：gemini-code-assist[bot] 指出文档中描述的 start_step 参数在后端尚未实现（assert start_step is None），这可能导致用户困惑。
- 后续行动：作者 cs-cat 回应建议需要全面审查整个文档，并通过新的 PR 来解决此类不一致问题，但本次 PR 仅聚焦于端点名称的修正。
  - 文档中 start_step 参数与后端实现不一致 (correctness): 作者 cs-cat 建议后续通过新 PR 全面审查文档来解决此类不一致，本次 PR 未修改此部分。

风险与影响

- 风险：技术风险较低：
- 回归风险：无，本次变更仅涉及文档文本修改，不涉及任何源代码、配置或测试逻辑。
- 兼容性风险：无，文档更正后与实际 API 一致，不会引入兼容性问题。
- 安全风险：无。
- 未解决风险：文档中仍存在 start_step 参数描述与后端实现不一致的问题，这可能导致用户尝试使用未支持的功能时遇到断言错误或混淆，但此风险并非本次 PR 引入，且已通过 review 讨论暴露。
- 影响：影响范围有限：
- 对用户的影响：正面影响，用户将获得正确的 API 端点信息，避免因文档错误而无法停止剖析会话。影响程度为低，仅涉及文档使用者。
- 对系统的影响：无，不改变任何系统行为或性能。
- 对团队的影响：提醒团队注意文档与代码实现的一致性，可能促使后续更全面的文档审查。
- 风险标记：文档与实现不一致

关联脉络

- PR #22975 [NPU] [DOC] Update npu best practice docs to match latest code: 同为文档更新 PR，旨在确保文档与代码实现同步，体现了团队对文档准确性的持续关注。
- PR #22923 docs: fix incorrect default max-payload-size in gateway config reference: 同为文档修正 PR，修正配置默认值错误，与本次 PR 类似，属于文档维护性质。