PR #22982 完整报告

sgl-project/sglang

[Docs] fix profiling endpoint

合并时间：2026-04-17 00:51

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22982

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，无需深入技术分析。对于需要了解性能剖析端点的开发者，可快速浏览以确认正确的 API 使用方式。

功能与动机

根据 PR body 中的描述，原文档指导用户使用 /end_profile 来停止剖析，但实际运行的服务器暴露的是 /start_profile 和 /stop_profile 端点。这导致了文档与实际 API 行为不一致，可能误导用户。

实现拆解

1. 修正端点名称：将文档中所有提及 /end_profile 的地方统一替换为 /stop_profile，包括端点描述、参数说明和示例命令。
2. 澄清输出目录行为：在“剖析服务器”章节开头，新增一段说明，解释 output_dir 参数在 bench_serving --profile 客户端调用和直接调用 /start_profile 时的不同行为，并建议同时设置 SGLANG_TORCH_PROFILER_DIR 环境变量以避免混淆。
3. 同步参数说明：在 /start_profile 的参数列表中，将 num_steps 参数的描述从“手动停止使用 /end_profile”更新为“手动停止使用 /stop_profile”。
4. 更新示例命令：将所有示例中的 curl -X POST http://127.0.0.1:30000/end_profile 命令更新为使用 /stop_profile 端点。

关键文件：

- docs/developer_guide/benchmark_and_profiling.md（模块 开发者指南；类别 docs；类型 documentation）：这是唯一被修改的文件，包含了性能剖析的完整文档，端点和参数描述的修正直接影响用户操作。

关键符号：未识别

评论区精华

该 PR 仅有一次由 b8zhong 的批准，没有实质性的 review 评论或讨论。这表明变更直接且无争议，属于简单的文档同步修复。

- 暂无高价值评论线程

风险与影响

- 风险：技术风险极低：
  - 此 PR 仅修改文档，不涉及任何源代码、配置或测试，因此不存在回归、性能、安全或兼容性风险。
  - 唯一潜在风险是文档更新可能仍存在其他未同步的端点引用，但基于当前变更范围，风险可控。
- 影响：影响范围有限：
- 用户影响：正面影响。修复了文档错误，使用户能正确使用 /stop_profile 端点停止剖析，避免操作失败或困惑。
- 系统影响：无。文档变更不影响系统运行。
- 团队影响：维护了文档的准确性，减少了用户支持成本。
- 风险标记：文档不一致

关联脉络

- PR #22523 [Doc] correct the HTTP endpoint for stopping profiling in benchmark_and_profiling.md: 两者都修改了同一个文件 docs/developer_guide/benchmark_and_profiling.md，且都涉及性能剖析端点的文档修正。PR 22523 可能已部分修正了端点，但本 PR 进一步更新了更多细节（如参数描述和示例命令）。