PR #27296 完整报告

sgl-project/sglang

Add --enable-symm-mem for Qwen3.5

合并时间：2026-06-05 06:23

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27296

执行摘要

- 一句话：为 Qwen3.5 添加 H100 FP8 下的 --enable-symm-mem 支持
- 推荐动作：建议合并，属于有益的文档改进。无需深度审查。

功能与动机

在 H100 FP8 部署中，启用 NCCL 对称内存可获得最佳性能。参考了 InferenceX 仓库的相关 PR。

实现拆解

1. 在部署配置片段中添加条件判断：在 docs_new/src/snippets/autoregressive/qwen35-deployment.jsx 文件中，在 // Chunked prefill tuning 区块之前插入新的条件块，当硬件为 h100、量化类型为 fp8 且 TP > 1 时，向命令行追加 --enable-symm-mem 参数。
2. 在文档中添加提示：在 docs_new/cookbook/autoregressive/Qwen/Qwen3.5.mdx 的配置技巧小节中新增一条针对 H100 FP8 的说明，建议添加 --enable-symm-mem。

关键文件：

- docs_new/src/snippets/autoregressive/qwen35-deployment.jsx（模块 部署片段；类别 source；类型 core-logic）：核心变更：在部署配置生成逻辑中新增 --enable-symm-mem 条件追加。
- docs_new/cookbook/autoregressive/Qwen/Qwen3.5.mdx（模块 文档；类别 other；类型 core-logic）：文档配套更新：在配置技巧中添加对应说明。

关键符号：未识别

评论区精华

无 review 讨论。PR 由作者发起，合并者直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：变更仅涉及文档和交互式配置片段，不影响运行时逻辑。不会引入回归或安全问题。

- 影响：影响范围仅限于使用 H100 FP8 运行 Qwen3.5 的用户，通过文档获取优化建议。对系统整体无影响。
- 风险标记：无风险

关联脉络

- PR #1544 InferenceX 的参考 PR: PR body 中直接引用为参考来源