

PR #29266 完整报告

sgl-project/sglang

Sync Gemma4 hardware table with Blackwell recipes

合并时间: 2026-06-25 22:57

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29266>

执行摘要

该 PR 同步了 Gemma4 cookbook 文档中的硬件需求表格与已有的 Blackwell 部署配方，为 E2B、E4B、12B、31B 和 26B-A4B 模型行添加了 B200/B300 支持，并明确 31B 在 B200/B300 上使用 TP=1。变更仅涉及一个 Markdown 文件中的 HTML 表格，无代码改动，风险极低。

功能与动机

Gemma4 的命令选择器已暴露 B200/B300 的部署命令（见 [gemma4-deployment.jsx](#)），但硬件需求表格仍过时，缺少这些硬件条目。PR body 指出: "...readers could see a generated Blackwell command while the table did not list that hardware." 此变更旨在消除文档与配方之间的不一致，让用户在查阅硬件需求时也能找到 B200/B300 的支持信息。

实现拆解

- 定位文件: docs_new/cookbook/autoregressive/Google/Gemma4.mdx 中的硬件需求表格。
- 添加 B200/B300 条目: 在 E2B, E4B, 12B, 31B, 26B-A4B 共 5 行的硬件列中插入 "1x B200 / 1x B300" (12B 为 "1x B200 / 1x B300", 其余类似)，并保留原有 H200 和 AMD 条目 (如 MI300X, MI325X, MI355X)。
- 更新 TP 列 (仅 31B): 从 2 (H200) / 1 (AMD) 改为 2 (H200) / 1 (B200/B300/AMD), 明确 B200/B300 也使用 TP=1。
- 验证: 使用 `git diff --check` 和静态文档检查确保格式正确，并引用已完成 B200 配方测试作为证据。

因变更仅涉及 HTML 表格文本，无代码逻辑，此处省略源码片段。

评论区精华

该 PR 无 review 评论，无技术讨论。

风险与影响

- 风险: 极低。仅修改文档文本，无运行时影响。唯一潜在风险是未来 Blackwell 硬件配置发生变化时表格可能再次过时，但当前已与配方一致。
- 影响: 用户可直接在硬件需求表格中看到 B200/B300 支持情况，减少配置时的困惑；团队文档与代码配方的同步得到改善。

关联脉络

与 #29252 (Tune Gemma4 26B-A4B B200 memory recipe) 属于同一维护周期：前者验证了 B200 配方的可行性，后者将验证结果同步到硬件需求表格。此外，与 #29302 (fix diffusion docs and cookbook drift) 同属作者持续维护 cookbook 文档一致性的系列工作。