PR #25378 完整报告

sgl-project/sglang

[Doc] Update MegaMoE usage

合并时间：2026-05-15 17:17

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25378

执行摘要

- 一句话：更新 DeepSeek V4 部署文档中的 MegaMoE 配置
- 推荐动作：这是一个小规模的文档更新，不需要深度 review。合并后有助于用户了解 MegaMoE 的可选配置。

功能与动机

补充 MegaMoE 的 w4a4 内核自定义配置选项，帮助用户在 B200/B300 Pro 等平台通过额外环境变量提升性能。

实现拆解

1. 在 docs_new/src/snippets/autoregressive/deepseek-v4-deployment.jsx 的 Enabling MegaMoE 代码块中，在 SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK=0 之后添加两行注释和两个可选的 export 环境变量（SGLANG_OPT_DEEPGEMM_MEGA_MOE_USE_FP4_ACTS=1, SGLANG_OPT_DEEPGEMM_MEGA_MOE_USE_MXF4_KIND=1）。
2. 在代码块下方的说明段落中增加一行解释，说明这两个变量用于自定义 w4a4 MegaMoE 内核，并提及性能提升和精度影响。

关键文件：

- docs_new/src/snippets/autoregressive/deepseek-v4-deployment.jsx（模块 文档片段；类别 docs；类型 documentation）：唯一变更文件，更新了 MegaMoE 的 w4a4 内核环境变量示例和描述。

关键符号：未识别

关键源码片段

docs_new/src/snippets/autoregressive/deepseek-v4-deployment.jsx

唯一变更文件，更新了 MegaMoE 的 w4a4 内核环境变量示例和描述。

```
// 在 Enabling MegaMoE 代码块的末尾
// 增加以下两行可选环境变量
# Optional env vars for custom w4a4 MegaMoE kernel
SGLANG_OPT_DEEPGEMM_MEGA_MOE_USE_FP4_ACTS=1
```

SGLANG_OPT_DEEPGEMM_MEGA_MOE_USE_MXF4_KIND=1

// 同时在说明段落中增加一行解释

<code>SGLANG_OPT_DEEPGEMM_MEGA_MOE_USE_FP4_ACTS=1 SGLANG_OPT_DEEPGEMM_MEGA_MOE_USE_MXF4_KIND=1</code> for customized w4a4 MegaMoE kernel. With this kernel, the performance can be increased with negligible accuracy drop (~89.5 GPQA for Pro model)

评论区精华

Reviewer gemini-code-assist[bot] 指出两个问题：1）新增的环境变量使用了 export 关键字，与已有变量不一致，建议去掉 export；2）说明段落中两个变量被拼接在同一个 <code> 块内，缺少分隔符和换行，建议修复格式。作者已采纳建议，去除了 export 并调整了格式。

- export 关键字一致性 (style): 作者已采纳并修复，去除了 export。
- 说明段落格式问题 (style): 作者已采纳并修复，格式已调整。

风险与影响

- 风险：无技术风险。该 PR 仅更改文档中的示例代码和描述文本，不涉及任何运行时逻辑。
- 影响：影响范围仅限于 DeepSeek V4 部署文档的交互式代码片段，用户可参考更新的示例启用 w4a4 MegaMoE 内核。该变更属于增量增强，不影响现有功能。
- 风险标记：暂无

关联脉络

- PR #25369 Add hicache feature in dsv4 cookbook: 同为 DeepSeek V4 部署文档的更新，更改了同一文件。