PR #23883 完整报告

sgl-project/sglang

Enable DeepGemm warmup in DeepSeek-V4 cookbook

合并时间：2026-04-28 09:41

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23883

执行摘要

- 一句话：DeepSeek-V4 cookbook 启用 DeepGemm warmup
- 推荐动作：建议快速合并，此 PR 是文档 / 配置跟进，无技术风险。可精读第二个修复提交，了解如何修复遗漏引用错误。

功能与动机

依赖 PR#23756 和龙的新镜像，移除 SGLANG_JIT_DEEPGEMM_PRECOMPILE=0 以启用 DeepGemm warmup，从而提升 DeepSeek-V4 模型的启动和运行效率。

实现拆解

1. 移除 COMMON_ENV 常量：在 generateCommand() 函数（第 260 行附近）中删除 const COMMON_ENV = ["SGLANG_JIT_DEEPGEMM_PRECOMPILE=0"];，因为新镜像已默认支持 DeepGemm warmup，无需再禁用预编译。
2. 移除 buildRole() 中的重复声明：在 buildRole() 函数（第 520 行附近）中删除另一处相同的 const COMMON_ENV 声明，该声明是用于 PD 分离部署场景的。
3. 更新环境变量组装逻辑：在 generateCommand() 中将 const envAll = [...HW_ENV, ...recipeEnv, ...COMMON_ENV] 改为 const envAll = [...HW_ENV, ...recipeEnv]；在 buildRole() 中将 const envAll = [...HW_ENV, ...roleEnv, ...MNNVL_ENV, ...COMMON_ENV] 改为 const envAll = [...HW_ENV, ...roleEnv, ...MNNVL_ENV]。
4. 修复后续提交中的 ReferenceError：第二个提交修复了第一个提交中遗漏删除 ...COMMON_ENV 展开导致的引用错误。

关键文件：

- docs_new/src/snippets/autoregressive/deepseek-v4-deployment.jsx（模块 部署脚本；类别 source；类型 configuration）：唯一变更文件，移除了 DeepGemm warmup 禁用环境变量，启用默认 warmup。

关键符号：generateCommand, buildRole

评论区精华

无 review 评论或讨论。

- 暂无高价值评论线程

风险与影响

- 风险：该变更为纯文档 / 配置改动，风险极低。主要风险在于移除了 SGLANG_JIT_DEEPGEMM_PRECOMPILE=0 后，若环境未正确依赖新镜像（PR#23756），可能导致 DeepGemm warmup 失败或性能下降。但该风险已在 PR 描述中声明依赖关系。
- 影响：
  - 用户：使用 DeepSeek-V4 部署 cookbook 的用户将默认启用 DeepGemm warmup，预期获得更好的启动和运行性能。
  - 系统：无直接影响，因为仅为文档代码片段。
  - 团队：需要确保相关镜像和 PR#23756 已合并发布，否则用户可能遇到问题。
  - 风险标记：依赖未合并 PR

关联脉络

- PR #23756 Dependency for DeepGemm warmup support: 此 PR 明确声明依赖 PR#23756，该 PR 实现了 DeepGemm warmup 功能。