PR #29937 完整报告

sgl-project/sglang

[NPU] [DOC] add missing DEEP_NORMAL_MODE_USE_INT8_QUANT for w8a8+deepep scenarios

合并时间：2026-07-03 10:18

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/29937

PR 分析报告：更新 Ascend NPU 文档

执行摘要

该 PR 主要对 Ascend NPU 文档进行维护：删除旧 best practice 文档（约 7000 行），在环境变量表中补充 DEEP_NORMAL_MODE_USE_INT8_QUANT 并标记为 deprecated，在优化文档中添加 NPU Graph 与 torch.compile 不兼容提示，同时在多个示例脚本中增加该环境变量设置。PR 主要影响 Ascend NPU 上使用 W8A8 量化 + DeepEP 的 DeepSeek/GLM5.2/Qwen3 模型部署。

功能与动机

根据 PR body，动机是 delete old best practice；add missing DEEP_NORMAL_MODE_USE_INT8_QUANT for w8a8+deepep scenarios。旧 best practice 文档内容陈旧且冗余，需要清除；同时 W8A8 量化场景需要明确的环境变量指导。

实现拆解

1. 删除旧 best practice 文档：删除 ascend_npu_best_practice.mdx（约 7000 行）。
2. 补充环境变量说明：在 ascend_npu_environment_variables.mdx 的环境变量表中新增 DEEP_NORMAL_MODE_USE_INT8_QUANT 条目，说明其在 W8A8 量化的 MoE 模型中启用 INT8 量化以减少通信量，并标记为 Deprecated（未来版本将移除，行为将自动推断）。
3. 更新优化文档：在 ascend_npu_optimization.mdx 中添加 Note，说明 --enable-torch-compile 与 NPU Graph 不兼容，启用 torch.compile 时必须通过 --disable-cuda-graph 禁用 NPU Graph。
4. 在多个配置示例中添加变量：在 GLM5.2 示例、Qwen3 模型教程以及旧 best practice 文件的多个脚本块中添加 export DEEP_NORMAL_MODE_USE_INT8_QUANT=1。
5. 修复格式问题：在 ascend_npu_support_features.mdx 中将 init-expert-location 的 HTML 转义字符 < 和 > 替换为实际字符。
6. 采纳 review 建议：在后续 commit 中移除了 low_latency（decode）模式下的冗余变量导出。

无。

评论区精华

- gemini-code-assist[bot] 指出 low_latency（decode）节点中设置 DEEP_NORMAL_MODE_USE_INT8_QUANT 是冗余的，仅 normal mode（prefill）有效。提交者采纳并移除了相关导出。

风险与影响

- 风险：文档删除可能导致用户找不到历史参考，但新文档更准确；deprecated 变量可能造成短时困惑，但已有明确说明。
- 影响：用户需更新部署脚本，在 W8A8+DeepEP 场景中添加该变量；deepseek 系列模型部署文档更加统一。

关联脉络

- 关联 PR #29828（GLM5.2 on ascend doc），该 PR 建立了 GLM5.2 文档，本 PR 在 GLM5.2 示例中添加了环境变量，属于后续完善。
- 该 PR 是 NPU 文档体系重构的一部分，后续可能继续清理和集中化环境变量说明。