

PR #29200 完整报告

sgl-project/sglang

[Cookbook] Nemotron3-Ultra: align MTP draft depth with NVIDIA reference (num_steps 5)

合并时间: 2026-06-25 03:12

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29200>

执行摘要

- 一句话: 修复 Nemotron3-Ultra MTP 参数与 NVIDIA 参考对齐
- 推荐动作: 该 PR 较为简单, 如果关注 Nemotron3-Ultra 模型部署, 可以精读此文档。主要为保持与其他框架的参数一致性。

功能与动机

PR body 中说明 'Fix the inconsistency in MTP compare to other frameworks (vLLM, TensorRT-LLM)'。Nemotron3-Ultra 模型的 MTP 层数 (draft depth) 在 NVIDIA 参考实现中为 5 步, 但先前文档中写为 3 步, 导致用户部署时参数错误。

实现拆解

在 `docs_new/src/snippets/autoregressive/nemotron3-ultra-deployment.jsx` 文件中, 将 MTP 配置的 `commandRule` 中 `--speculative-num-steps` 从 3 改为 5, 并将 `--speculative-num-draft-tokens` 从 4 改为 6 (因为 `num-steps` 5 且 `topk` 1, `draft tokens` = `num-steps` * `topk` = 5, 但文档中保持为 6 可能是为了预留一个额外 draft token, 此处按原逻辑对齐)。其余参数不变。

关键文件:

- `docs_new/src/snippets/autoregressive/nemotron3-ultra-deployment.jsx` (模块 文档示例; 类别 source; 类型 core-logic): 唯一变更的文件, 包含 Nemotron3-Ultra 部署参数模板。修改了 MTP 的 `commandRule`, 对齐 NVIDIA 参考的 `speculative-num-steps` 和 `speculative-num-draft-tokens`。

关键符号: 未识别

评论区精华

无 review 评论, PR 直接获得审核者 `zijiexia` 批准。

- 暂无高价值评论线程

风险与影响

- 风险: 该变更仅修改文档示例中的命令行参数, 不涉及任何运行时逻辑。风险极低: 错误参数可能导致 MTP 性能下降或推理失败, 但正确数值与官方参考一致, 无回归风险。

- 影响：影响范围仅限于 cookbook 文档页面的 Nemotron3-Ultra 部署示例。用户按照更新后的参数部署，将获得与 NVIDIA 参考一致的 MTP draft 深度。影响程度小，但改善了下游用户的部署准确性。
- 风险标记：极少变更

关联脉络

- 暂无明显关联 PR