

PR #28087 完整报告

sgl-project/sglang

[Doc] Fix some inconsistencies in the Nemotron Cookbook

合并时间: 2026-06-13 05:51

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/28087>

执行摘要

- 一句话: 修复 Nemotron Cookbook 文档不一致
- 推荐动作: 纯文档修订, 无明显风险, 可供用户直接参考。

功能与动机

根据 PR body, 作者指出文档中存在多处不一致: 命令需要简化、已验证的 `extra_buffer` 未在所有地方添加、Spec V2 标志已被设为默认故移除。

实现拆解

1. 在 JSX 生成器中新增 EP 字段并调整字段顺序: 将 EP 与 DP attention 互换位置, 使交互式配置面板更符合预期逻辑。
2. 简化 MTP 命令规则: 移除 Blackwell 硬件的条件分支和 `--mamba-scheduler-strategy extra_buffer` 以及 `--attention-backend trtllm_mha` 的硬编码, 改为在生成命令的循环中统一添加。
3. 统一添加 `extra_buffer` 和 `trtllm_mha`: 在命令构建函数中, 始终添加 `--mamba-scheduler-strategy extra_buffer`, 并在硬件为 Blackwell 时追加 `--attention-backend trtllm_mha`。
4. 移除 Spec V2 环境变量: 删除已废弃的 `SGLANG_ENABLE_SPEC_V2=1` 前缀。
5. 更新 MDX 文档: 补充 EP 和 MTP 的说明, 统一端口为 30000, 修正模型名称等。

关键文件:

- `docs_new/src/snippets/autoregressive/nemotron3-ultra-deployment.jsx` (模块 代码片段; 类别 source; 类型 core-logic): 核心交互式配置生成器, 包含字段顺序调整、命令构建逻辑重构、移除过时环境变量等主要变更。
- `docs_new/cookbook/autoregressive/NVIDIA/Nemotron3-Ultra.mdx` (模块 文档; 类别 other; 类型 core-logic): MDX 文档更新, 补充 EP 和 MTP 说明, 统一端口和模型名称。

关键符号: 未识别

关键源码片段

`docs_new/src/snippets/autoregressive/nemotron3-ultra-deployment.jsx`

核心交互式配置生成器，包含字段顺序调整、命令构建逻辑重构、移除过时环境变量等主要变更。

```
// 以下是变更后的命令构建函数 (generateCommand) 的完整实现片段
const generateCommand = (values) => {
  const { tp, kvcache, model, hardware } = values;
  const cfg = findVerified(model, hardware, tp);

  // 如果所选组合不在已验证矩阵中，则返回阻止消息
  if (!cfg) {
    return [
      '# This combination is not in the verified support matrix.',
      '# Please consult the verified matrix below for supported configurations.'
    ];
  }

  const modelPath = MODEL_PATHS[model] || MODEL_PATHS['bf16'];
  let cmd = `python3 -m sglang.launch_server \\n`;
  cmd += ` --model-path ${modelPath} \\n`;
  cmd += ` --trust-remote-code \\n`;
  cmd += ` --tp ${tp} \\n`;

  // 渲染所有选项的动态命令规则
  for (const [key, option] of Object.entries(options)) {
    if (option.commandRule) {
      const rule = option.commandRule(values[key], values);
      if (rule) {
        cmd += ` ${rule} \\n`;
      }
    }
  }

  // 始终添加 extra_buffer，因为此混合 Transformer-Mamba 模型需要
  cmd += ` --mamba-scheduler-strategy extra_buffer \\n`;
  // Blackwell 硬件需要 trtllm_mha 避免 flashinfer 默认破坏重叠调度器
  if (['b200', 'gb200', 'b300', 'gb300'].includes(hardware)) {
    cmd += ` --attention-backend trtllm_mha \\n`;
  }

  if (kvcache && kvcache !== 'none') {
    cmd += ` --kv-cache-dtype ${kvcache} \\n`;
  }

  if (cfg.multinode) {
    cmd += ` --node-rank <0|1> \\n`;
  }
  // ... 后续命令拼接省略
  return [cmd];
};
```

评论区精华

无 review 评论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅涉及文档和交互式配置生成器，不包含任何运行时逻辑。生成的命令可能因文档版本与运行时版本不一致而略微偏差，但属于文档维护的正常范畴。
- 影响：影响范围限于使用 Nemotron3-Ultra cookbook 文档的用户和开发者。交互式配置生成器的展示和命令行输出将更准确，端口和模型名称的统一降低用户困惑。
- 风险标记：低风险，仅文档变更

关联脉络

- PR #24955 Support Nemotron DP attention and MTP: 同属 Nemotron 模型功能线，该 PR 添加 DP attention 和 MTP 支持，当前 PR 更新相关文档。