# PR #24435 完整报告

sgl-project/sglang

Update Qwen3-Coder docs_new NVIDIA guidance

合并时间：2026-06-02 04:38

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24435

## 执行摘要

- 一句话：迁移 Qwen3-Coder 文档至 docs_new 并更新 NVIDIA 部署指引
- 推荐动作：建议部署 Qwen3-Coder 的用户和文档维护者精读此 PR。值得关注的设计决策包括：如何在命令生成器中模块化硬件配置（modelConfigs 中直接定义 ep），以及移除显式 MoE 后端标志以利用默认值，这降低了用户配置负担。

## 功能与动机

随着 flashinfer_trtllm 成为默认且首选的 MoE 运行后端，旧的 Qwen3-Coder 部署文档需要更新以反映已验证的 NVIDIA 配置，并将内容从已归档的 sgl-cookbook 迁移到主仓库的 docs_new 目录，确保用户能获取最新且正确的部署命令和性能数据。

## 实现拆解

1. 更新部署命令生成器（docs_new/src/snippets/autoregressive/qwen3-coder-deployment.jsx）：在 modelConfigs 中为 B200 添加 ep: 8，为 GB200 添加 tp: 4, ep: 4；重构 EP 设置逻辑，优先使用 hwConfig.ep，否则根据量化类型回退；将 NVFP4 的 DP attention 与 EP 设置分离，并移除显式的 --moe-runner-backend 标志。
2. 更新文档正文（docs_new/cookbook/autoregressive/Qwen/Qwen3-Coder.mdx）：替换 NVIDIA 部分关于 MoE 后端的描述，明确 GB200 使用 --tp 4 --ep 4，B200 使用 --tp 8 --ep 8；新增 NVIDIA FP8 和 NVFP4 在低（1 并发）、中（16 并发）、高（64 并发）三种场景下的详细基准测试结果表格（吞吐量、延迟）；更新 GSM8K 准确率和输出吞吐量数据。
3. 根据 Review 反馈调整：在对话中，根据 reviewer 的建议修正了 B200 的 EP 配置，将生成器中的 B200 EP 从默认改为 8，使生成命令与文档中 benchmark 命令一致。

关键文件：

- docs_new/src/snippets/autoregressive/qwen3-coder-deployment.jsx（模块 部署命令生成器；类别 source；类型 core-logic）：包含命令生成逻辑的核心变更：新增 B200/GB200 的 ep 配置，重构 EP 设置优先级，移除 MoE 后端显式标志。直接影响用户在页面上看到的最终部署命令。
- docs_new/cookbook/autoregressive/Qwen/Qwen3-Coder.mdx（模块 用户文档；类别 other；类型 documentation）：用户可见的文档主体，包含更新后的部署指引、详细的基准测试结果和 GSM8K 准确率。是整个 PR 的对外交付内容。

关键符号：未识别

## 关键源码片段

### docs_new/src/snippets/autoregressive/qwen3-coder-deployment.jsx

包含命令生成逻辑的核心变更：新增 B200/GB200 的 *ep* 配置，重构 EP 设置优先级，移除 MoE 后端显式标志。直接影响用户在页面上看到的最终部署命令。

```
/*
 * 命令生成器中关键硬件配置与 EP 逻辑片段
 * modelConfigs 为每种硬件定义了 tp 和可选的 ep
 * 若硬件未指定 ep，则根据 NVFP4 量化回退为 1，否则由后续 fallback 处理
 */
const modelConfigs = {
  '480b': {
    baseName: '480B-A35B',
    mi300x: { tp: 8 },
    mi325x: { tp: 8 },
    mi355x: { tp: 8 },
    b200: { tp: 8, ep: 8 }, // B200 明确指定 ep=8
    gb200: { tp: 4, ep: 4 } // GB200 使用 tp=4, ep=4
  },
  '30b': {
    baseName: '30B-A3B',
    mi300x: { tp: 1 },
    mi325x: { tp: 1 },
    mi355x: { tp: 1 }
  }
};

const generateCommand = (values) => {
  // ... 前期验证省略 ...
  // TP setting
  cmd += ` \\
  --tp ${hwConfig.tp}`;

  // EP settings
  // 优先使用硬件配置中的 ep，若未定义且为 NVFP4 则设为 1，否则 null
  const ep = hwConfig.ep || (quantization === 'nvfp4' ? 1 : null);
  if (ep) {
    cmd += ` \\
  --ep ${ep}`;
  } else if (modelSize === '480b' && quantization === 'fp8') {
    // 480B FP8 需要 ep=2 满足 MoE 维度对齐
    cmd += ` \\
  --ep 2`;
  }

  // DP attention setting（独立于 EP，仅 NVFP4 启用）
```

```
  if (quantization === 'nvfp4') {
    cmd += ` \\
  --enable-dp-attention`;
  }

  // MOE runner backend（不再显式指定，默认使用 flashinfer_trtllm）
  if (isNvidia) {
    if (quantization === 'nvfp4') {
      cmd += ` \\
  --quantization modelopt_fp4`;
    }
  }
  // ... AMD 专属标志省略 ...
  return cmd;
};
```

## 评论区精华

- 代码健壮性建议：gemini-code-assist[bot] 指出三元运算符 quantization === 'nvfp4' ? 1 : null 在新增量化方法时可能脆弱，建议使用更健壮的配置映射。作者未直接修改，但该问题在当前上下文中影响不大。
- 命令与 GPU 明确性：维护者 zijiexia 要求补充完整的服务器启动命令并明确 GPU 型号。作者最初仅写 --tp 4 --ep 4，后更新为包含 sglang serve 的完整命令。
- B200 命令不一致：zijiexia 发现文档中 B200 的 benchmark 命令与生成器生成的不匹配（缺少 --ep 8）。作者承认并修复了生成器，同时接受 zijiexia 的建议改用 sglang serve 风格，确保一致。
- EP 逻辑健壮性 (design): 当前改动已使用 hwConfig.ep 优先，该问题影响有限，未进一步修改。
- 命令与 GPU 明确性 (question): 作者补充了 sglang serve 形式的完整命令，并区分 B200/GB200。
- B200 命令不一致 (correctness): 生成器配置和文档同步修正，接受 zijiexia 的建议改用 sglang serve 风格。

## 风险与影响

- 风险：低风险。主要风险来自 文档与命令生成器不一致：如果未来更改 modelConfigs 但忘记同步生成逻辑或文档，用户可能获得错误的部署参数。此外，移除 --moe-runner-backend 标志后，若用户使用的旧版 SGLang 中 flashinfer_trtllm 并非默认，可能导致回退到非最优后端。不过，当前变更基于已验证的最新版本，且通过 review 已消除明显不一致。
- 影响：
  - 用户：阅读 Qwen3-Coder 部署文档的用户将获得针对 B200/GB200 的准确命令和性能预期，减少试错成本。
  - 系统：无运行时影响，仅文档和前端代码变更。

- 团队：降低了文档维护成本（迁移到主仓库），但需确保后续对 MoE 后端的默认值变更时同步更新此处。
- 风险标记：文档与生成器一致性问题，默认后端变更可能影响旧版本用户

## 关联脉络

- PR #265 Codex/update qwen3 coder flashinfer trtllm: 关联的 sgl-cookbook issue，是本 PR 的前身，此次 PR 将其内容迁移到主仓库 docs_new。