PR #26969 完整报告

sgl-project/sglang

docs: add Nemotron 3 Ultra cookbook entry

合并时间：2026-06-04 15:14

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26969

执行摘要

本 PR 为 NVIDIA Nemotron 3 Ultra 550B 模型添加了完整的 Day-0 cookbook 文档，包括模型介绍、安装指引、交互式部署生成器和推理示例。核心改动是新增一个 500+ 行的 Markdown 页面和一个 400+ 行的 JSX 交互组件，后者通过验证矩阵确保仅输出经过测试的部署命令，有效降低用户配置错误风险。这是一次纯文档变更，不涉及运行时代码。

功能与动机

如 PR 描述所述："Adds SGLang Day-0 cookbook for NVIDIA Nemotron 3 Ultra (550B hybrid Transformer-Mamba MoE, 55B active, 1M context, text-only) following the existing Nemotron3-Super page layout." 随着 NVIDIA 发布新的混合 MoE 模型，需要在 SGLang 文档中提供即开即用的部署指南，帮助用户在多种 GPU 配置下快速启动推理服务。

实现拆解

1. 创建主文档页面：在 docs_new/cookbook/autoregressive/NVIDIA/Nemotron3-Ultra.mdx 中撰写完整的模型说明、架构特点、支持硬件 (H100/H200/B200/GB200/B300)、量化变体 (BF16/NVFP4)、安装命令、部署步骤、推理示例（包含 reasoning/tool-calling 输出）以及基准测试表格（数据暂标记为 TBD）。
2. 开发交互式部署生成器：新增 docs_new/src/snippets/autoregressive/nemotron3-ultra-deployment.jsx 文件，导出一个 React 组件。组件内部定义 VERIFIED_CONFIGS 数组作为单一事实来源，列出所有经验证的 {模型，硬件，TP} 组合。利用 findVerified、verifiedHardwareForModel、verifiedTpForModelHardware 等辅助函数动态过滤选项，确保用户选择仅限已验证组合。后续通过 generateCommand 函数拼装 sglang.launch_server 命令。
3. 注册页面导航：在 docs_new/docs.json 的 NVIDIA 分组下追加 "cookbook/autoregressive/NVIDIA/Nemotron3-Ultra" 条目，使页面出现在文档侧边栏。同时在 docs_new/cookbook/intro copy.mdx 的 NVIDIA 段落末尾添加条目并附带 NEW 标签徽章。
4. 添加 DP 注意力和 EP 开关：在生成器中加入 dpattention 和 expertparallel 选项，分别绑定 --dp --enable-dp-attention 和 --enable-ep 参数。DP 验证逻辑进一步根据模型类型 (BF16 仅 DP=2，NVFP4 DP=2/4/8) 和 TP 大小自动过滤可用选项。
5. 持续修正验证矩阵：在 review 过程中根据反馈多次调整 VERIFIED_CONFIGS，例如移除 BF16+GB200/GB300 TP=8（未验证），扩展 NVFP4 到 B300 TP=4/8，并更改默认值为 NVFP4+B200+TP=4+MTP。

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

交互式部署生成器的核心实现，包含 VERIFIED_CONFIGS 验证矩阵、动态选项过滤和命令生成逻辑。

```
export const Nemotron3UltraDeployment = () => {
  // 模型路径映射，BF16 和 NVFP4 两个量化变体
  const MODEL_PATHS = {
    bf16: 'nvidia/NVIDIA-Nemotron-3-Ultra-550B-A55B-BF16',
    nvfp4: 'nvidia/NVIDIA-Nemotron-3-Ultra-550B-A55B-NVFP4',
  };

  // 已验证的 {model, hardware, tp} 组合，未在列表中的组合会被 generateCommand 阻止
  // 必须与 MDX 文档中的"支持的 GPU"表格保持同步
  const VERIFIED_CONFIGS = [
    { model: 'bf16', hardware: 'h100', tp: '16', multinode: true },
    { model: 'bf16', hardware: 'h200', tp: '16', multinode: true },
    { model: 'bf16', hardware: 'b200', tp: '8' },
    { model: 'bf16', hardware: 'b300', tp: '8' },
    { model: 'nvfp4', hardware: 'b200', tp: '4' },
    { model: 'nvfp4', hardware: 'b200', tp: '8' },
    { model: 'nvfp4', hardware: 'b300', tp: '4' },
    { model: 'nvfp4', hardware: 'b300', tp: '8' },
    { model: 'nvfp4', hardware: 'gb200', tp: '4' },
    { model: 'nvfp4', hardware: 'gb300', tp: '4' },
  ];

  const findVerified = (model, hardware, tp) =>
    VERIFIED_CONFIGS.find((c) => c.model === model && c.hardware === hardware && c.tp ===
    tp);

  const verifiedHardwareForModel = (model) =>
    [...new Set(VERIFIED_CONFIGS.filter((c) => c.model === model).map((c) => c.hardware))];

  const verifiedTpForModelHardware = (model, hardware) =>
    [...new Set(VERIFIED_CONFIGS.filter((c) => c.model === model && c.hardware ===
    hardware).map((c) => c.tp))];

  const dpCandidatesForModel = (model) => (model === 'bf16' ? ['2'] : ['2', '4', '8']);

  const maxVerifiedTpForModelHardware = (model, hardware) => {
    const tps = verifiedTpForModelHardware(model, hardware).map(Number);
    return tps.length ? Math.max(...tps) : 0;
  };

  const verifiedDpForModelHardwareTp = (model, hardware, tp) => {
    const cap = Math.min(Number(tp) || 0, maxVerifiedTpForModelHardware(model, hardware));
    return dpCandidatesForModel(model).filter((d) => Number(d) <= cap);
  };
```

```
    // 后续 options 定义和 generateCommand 逻辑基于这些验证函数保证生成命令的有效性
};
```

评论区精华

- 阻止未验证命令：Fridge003 强调“需要阻止任何未验证的命令，例如所有 H200 命令对 NVFP4 检查点都应被阻止”。作者通过引入 VERIFIED_CONFIGS 表并动态禁用选项来解决。
- 添加 DP 注意力：Fridge003 建议增加 dp attention 配置，并指出 @guapisolo 正在处理。后续提交新增了 dpattention 选项，并根据模型类型和 TP 大小过滤。
- 专用 Docker 镜像：针对安装章节的 Docker 标签，Fridge003 建议使用 lmsysorg/sglang:dev-nemotron3-ultra 替代占位符。
- NEW 标签：Fridge003 要求在 frontmatter 添加 tag: NEW 以便侧边栏显示 NEW 徽章，已修复。

风险与影响

- 验证矩阵同步风险：VERIFIED_CONFIGS 与 MDX 文档中的“支持的 GPU”表格必须保持同步，否则用户可能参考文档手动构造未验证命令。建议后续增加自动化测试或单数据源。
- Docker 镜像占位符：安装章节当前使用开发镜像 dev-nemotron3-ultra，待正式版发布后需替换为稳定标签。
- 基准数据 TBD：速度与精度基准数据标记为 TBD，如果长时间未填充可能降低文档可信度。
- 无运行时代码影响：纯文档变更，不会引入运行时缺陷。

关联脉络

该 PR 是 SGLang 文档系列中针对 NVIDIA Nemotron 系列的新增入口，延续了 Nemotron3-Super 的页面结构。相关的底层支持 PR 包括：#26861（NVFP4 MoE 加载优化），为该模型的 NVFP4 部署提供性能基础。未来需要跟进基准测试数据填充和稳定 Docker 镜像发布。