PR #25198 完整报告

sgl-project/sglang

[Docs] Update Nemotron3-Nano-Omni cookbook to reflect new model paths

合并时间：2026-06-04 05:55

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25198

执行摘要

更新 Nemotron3-Nano-Omni 部署文档和交互式代码片段中的模型路径，统一使用新的推理 BF16/FP8/NVFP4 HuggingFace ID，并调整默认选项。

功能与动机

原始文档引用了旧的模型 ID（如 nvidia/Nemotron-3-Nano-Omni-30B-A3B-Reasoning），这些路径已被更新为新格式（如 nvidia/Nemotron-3-Nano-Omni-30B-A3B-Reasoning-BF16）。作为 #25023 和 #25024 的配套更新，确保用户获得正确的模型路径。

实现拆解

1. 更新代码片段中的模型路径映射：在 docs_new/src/snippets/autoregressive/nemotron3-nano-omni-deployment.jsx 中，移除 reasoning 选项，将 MODEL_PATHS 中的路径改为带有 Reasoning- 前缀的新 ID，同时将默认模型从 reasoning 改为 bf16。
2. 同步 cookbook 文档：在 docs_new/cookbook/autoregressive/NVIDIA/Nemotron3-Nano-Omni.mdx 中，更新了模型变体列表、示例命令、API 使用示例中的模型名称和链接。

docs_new/src/snippets/autoregressive/nemotron3-nano-omni-deployment.jsx

核心部署代码片段，定义了模型路径映射和默认选项

```
// 定义模型路径映射，所有变体统一使用新的推理路径
const MODEL_PATHS = {
  bf16: 'nvidia/Nemotron-3-Nano-Omni-30B-A3B-Reasoning-BF16',
  fp8: 'nvidia/Nemotron-3-Nano-Omni-30B-A3B-Reasoning-FP8',
  nvfp4: 'nvidia/Nemotron-3-Nano-Omni-30B-A3B-Reasoning-NVFP4',
};

const options = {
  // ... 其他选项
  model: {
    name: 'model',
    title: 'Model',
    items: [
      { id: 'bf16', label: 'BF16', default: true }, // BF16 改为默认选中
      { id: 'fp8', label: 'FP8', default: false },
      { id: 'nvfp4', label: 'NVFP4', default: false },
```

```
      ],
    },
  };

  const generateCommand = (values) => {
    // ... 验证逻辑
    const modelPath = MODEL_PATHS[model] || MODEL_PATHS.bf16; // 默认回退改为 bf16
    // ... 构建命令
  };
```

评论区精华

无 review 评论。

风险与影响

- 风险：无技术风险，仅涉及文档字符串更新，不影响运行时代码。
- 影响：用户通过 cookbook 快速部署时将获得正确的模型 ID，避免路径错误。

关联脉络

本 PR 是 #25023 和 #25024 的配套文档更新，确保新模型路径在用户文档中同步。