PR #33481 完整报告

sgl-project/sglang

Add NVIDIA Nemotron 3.5 Lightning cookbook

合并时间：2026-08-11 21:00

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/33481

执行摘要

该 PR 为 SGLang 文档站新增 NVIDIA Nemotron 3.5 Lightning cookbook 页面，通过配置驱动（config + benchmarks + Deployment/Playground 组件）为 B200、H100、DGX Spark 三个单 GPU 平台生成 NVFP4 量化部署命令，覆盖基础服务与 MTP、DFlash、DSpark 三种投机解码策略，并提供 TP、EP、MoE backend、parser 等 Playground 探索轴。改动集中在 5 个文档相关文件（+537/-1），不涉及运行时代码；11 个 commit 经历了“加 B200 → 收敛到 H100/DGX Spark → 按 review 意见精简 flag → 重加 B200 → 修正 Docker tag 拼写”的演进，最终获得 b8zhong 与 JustinTong0323 两个 APPROVED 后合入。

功能与动机

Nemotron 3.5 Lightning 是 NVIDIA 的 30B-A3B 混合推理 LLM，模型支持由 PR#33554 落地。PR body 的定位是“Adds a cookbook page for NVIDIA Nemotron 3.5 Lightning with single-GPU NVFP4 recipes for H100 and DGX Spark, each covering base serving plus MTP, DFlash, and DSpark speculative decoding, along with playground axes”。即：在模型支持就绪后，把经过验证的单卡启动命令、投机解码组合、推理 / 工具解析器用法沉淀为面向用户的文档，并把 NVIDIA 模型族在 cookbook 首页的入口切换到新页面（intro.mdx 的 NVIDIA 卡片 href 变更印证了这一点）。

实现拆解

1. 新增 cookbook 页面：docs/cookbook/autoregressive/NVIDIA/Nemotron3.5-Lightning.mdx（+162 行）是页面主体，包含安装 Accordion（Python 路径用 uv pip install 从 refs/pull/33554/head 安装以拿到 Nemotron 3.5 模型支持，Docker 路径拉取 lmsysorg/sglang:dev-nemotron3-5-lightning 多架构镜像）、checkpoint 对照表（NVFP4 主模型 / BF16 参考 / DFlash 与 DSpark 两个 W4A16 draft 模型）、OpenAI 接口示例（--reasoning-parser nemotron_3 把思维链写入 message.reasoning_content，--tool-call-parser qwen3_coder 返回结构化 message.tool_calls）。
2. 新增配置驱动数据：docs/src/snippets/configs/nvidia/nemotron-3.5-lightning.jsx（+355 行）导出单一 config 字面量，声明平台、策略、模型名、镜像，核心是 cells 数组——每个元素用 match 键命中一条已验证配方。B200 显式加 --mamba-backend flashinfer 与 mamba 缓存随机舍入；H100 依赖 FA3 默认 target attention；DGX Spark 用 --mem-fraction-static 0.78 与 --cuda-graph-max-bs-decode 4 的保守组合。
3. 新增基准数据脚手架：...-benchmarks.jsx（+18 行）导出 12 条与 cells 一一对应的 match 记录，数值全部 pending，卡片在数据回填前渲染 pending 占位。

4. 注册与导航：docs/docs.json 在 NVIDIA group 下追加新页面条目；docs/cookbook/autoregressive/intro.mdx 把 NVIDIA 卡片 href 从 Nemotron3-Ultra 切换到新页面。
5. Review 驱动的演进与配套：commit cc788103 落实“删除等于默认值 flag”的意见，把每个 cell 从 26-34 个 flag 降到 6-13 个量级；commit 9e92e39e 把安装命令指向 PR#33554；最后一笔修正 Docker tag 拼写（lighting → lightning）。纯文档改动，无测试配套；CI 主线通过、extra 线失败（具体原因未在材料中说明）。

docs/src/snippets/configs/nvidia/nemotron-3.5-lightning-benchmarks.jsx

为 Deployment 卡片提供与 cells 一一对应的基准数据骨架，当前全部 pending，决定页面性能数据的展示形态。

```
// 每个 match 记录与 config.cells 一一对应；数值尚未回填，卡片先渲染 "pending"，
// 待实测跑完后再把速度 / 准确率数据填进来。
export const benchmarks = [
  { match: { hw: "b200", variant: "default", quant: "nvfp4", strategy: "balanced", nodes: "single" }
  },
  { match: { hw: "b200", variant: "default", quant: "nvfp4", strategy: "mtp", nodes: "single" } },
  { match: { hw: "b200", variant: "default", quant: "nvfp4", strategy: "dflash", nodes: "single" } },
  { match: { hw: "b200", variant: "default", quant: "nvfp4", strategy: "dspark", nodes: "single" } },
  { match: { hw: "h100", variant: "default", quant: "nvfp4", strategy: "balanced", nodes: "single" }
  },
  { match: { hw: "h100", variant: "default", quant: "nvfp4", strategy: "mtp", nodes: "single" } },
  { match: { hw: "h100", variant: "default", quant: "nvfp4", strategy: "dflash", nodes: "single" } },
  { match: { hw: "h100", variant: "default", quant: "nvfp4", strategy: "dspark", nodes: "single" } },
  { match: { hw: "dgx-spark", variant: "default", quant: "nvfp4", strategy: "balanced", nodes:
  "single" } },
  { match: { hw: "dgx-spark", variant: "default", quant: "nvfp4", strategy: "mtp", nodes: "single" } }
  ,
  { match: { hw: "dgx-spark", variant: "default", quant: "nvfp4", strategy: "dflash", nodes: "single"
  } },
  { match: { hw: "dgx-spark", variant: "default", quant: "nvfp4", strategy: "dspark", nodes:
  "single" } },
];
```

评论区精华

> b8zhong（install 段落）：“Change this to pip install of the PR of https://github.com/sgl-project/sglang/pull/33554”
>
> b8zhong（config 多处）：“Remove all flags that are set by default” / “Remove all redundant flags” / “Delete” —— commit cc788103 落实，说明每个 cell 曾带 26-34 个 flag，而其他 cookbook 页面只有 6-13 个。
>
> b8zhong：“Why shorten the context length” —— 未在材料中捕获明确答复，属于未决疑问。

另外，合并后 b8zhong 留言“Please update all commands to release repos”，作者回复已处理；但合并时文件仍引用 refs/pull/33554/head，二者存在出入，建议以仓库当前 main 状态

为准。

风险与影响

- 安装源非发布形态：Python 安装命令固定到 refs/pull/33554/head，一旦该 PR 的 ref 变更或被回收，文档命令无法复现当时的运行环境；b8zhong 合并后也要求统一改为 release 仓库。影响面：跟随文档安装的用户。
- 配方与运行时默认值耦合：cells 里的 flags 依赖 SGLang 对 NemotronH 的默认解析（attention backend、mamba backend、MoE runner、mem-fraction 等）。这些默认值在后续版本调整时，文档配方会静默失效或偏离最优配置，属于持续维护成本。
- 性能数据可信度：benchmarks 全部是 pending，H100 曾在 commit 中标记 unverified pending eval data；在数据回填之前，页面不构成任何吞吐 / 延迟选型依据。
- 镜像 tag 语义：历史上发布过拼写错误的 dev-nemotron3-5-lighting tag，正确 tag 已发布且内容一致，但错误 tag 仍被保留刷新，存在用户误用旧 tag 的混淆风险。

影响范围上，用户侧新增了三个单卡平台 × 四种策略的 NVFP4 部署路径；文档站 NVIDIA 模型族首页入口从 Nemotron3-Ultra 切换到新页面；团队侧沉淀了“config + benchmarks + Deployment/Playground”的 cookbook 模板和“只写非默认 flag”的约定。

关联脉络

- PR#33554（Nemotron 3.5 Lightning 模型支持）：本 PR 的安装命令与 dev-nemotron3-5-lightning 镜像都依赖它，是功能链路的前置。
- PR#34363（Ling-3.0-flash cookbook）：同一套 config + benchmarks + Deployment/Playground 的 cookbook 模板，说明该模式正成为新模型文档的标准做法。
- PR#33865（DSpark + DeepSeek V4 兼容修复）：DSpark 是本次 cookbook 覆盖的三种投机解码之一，可关联观察 DSpark 在不同模型族上的演进与修复。