PR #34573 完整报告

sgl-project/sglang

docs(cookbook): add BF16 recipes to Nemotron 3.5 Lightning

合并时间：2026-08-12 23:15

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/34573

执行摘要

- 一句话：Nemotron 3.5 Lightning cookbook 新增 BF16 配方
- 推荐动作：建议按需阅读：如果你在维护 cookbook snippet 或需要为其他模型扩展量化档位，本 PR 是一个很好的模板——展示了 quantizations + modelNames（variant|quant 键）与 match 三维矩阵（硬件 x 量化 x 策略）的组织方式，以及 MTP 复用 EAGLE 路径、草稿模型独立于目标量化档位的约定。对一般读者价值有限，不值得深入精读。

功能与动机

PR body 说明：为 Nemotron 3.5 Lightning cookbook 增加 BF16 量化选项，镜像 NVFP4 单元，模型路径解析到 BF16 checkpoint。此前 cookbook 只提供 NVFP4 量化配方，使用 BF16 checkpoint 的用户需要手工拼接启动参数；本次改动把 BF16 支持纳入官方 cookbook 的渲染数据，降低使用门槛。

实现拆解

1. 变更入口：唯一改动文件 docs/src/snippets/configs/nvidia/nemotron-3.5-lightning.jsx，它是 Mintlify cookbook 页面在 hydration 时重新求值的 config 字面量，决定页面展示的硬件、量化、策略选项以及“复制命令”按钮生成的命令。
2. 量化维度扩展：quantizations 数组从仅 nvfp4 变为 nvfp4 + bf16；modelNames 新增 "default|bf16": "nvidia/NVIDIA-Nemotron-3.5-Lightning-30B-A3B-BF16"。variant|quant 键是渲染层解析真实 HF repo id 的约定，最终注入到 {{MODEL_NAME}}。
3. 配方矩阵补齐：新增 12 个 { match, env, flags } cell，覆盖 3 平台 x 4 策略。平台差异：B200（SM100）显式指定 --mamba-backend flashinfer、开启 --enable-mamba-cache-stochastic-rounding --mamba-cache-philox-rounds 5、--mem-fraction-static 0.85、--cuda-graph-max-bs-decode 16，DFlash 用 block size 6；H100（SM90）不指定 mamba 后端（默认 FA3 target attention），显存参数同为 0.85/16；DGX Spark（SM121）沿用 NVFP4 的 0.78/4 保守参数。策略差异：MTP 统一走 EAGLE 路径且 --speculative-draft-model-path 指回目标 {{MODEL_NAME}}；DFlash/DSpark 继续引用独立 NVFP4 草稿 checkpoint。
4. 配套变更：无测试、schema、部署配套改动；第二次提交 "match NVFP4 style for BF16 cells" 只是对齐代码风格。这类 snippet 属于数据契约，checkpoint 路径或推荐参数变化时需同步更新此文件。

关键文件：

- docs/src/snippets/configs/nvidia/nemotron-3.5-lightning.jsx（模块 文档配置；类别 docs；类型 configuration）：唯一变更文件，新增 BF16 量化选项、模型路径映射及 12 组面向 B200/H100/DGX Spark 的启动配方，是 cookbook 生成部署命令的数据源。

关键符号：未识别

评论区精华

本 PR 没有任何 review 评论或讨论线程（review_comments_count = 0），两位 reviewer b8zhong 与 zijiexia 均直接批准且未留下文字说明。没有可提炼的设计交锋或未解决疑虑；变更属于纯配置 / 文档扩展，风险面小。

- 暂无高价值评论线程

风险与影响

- 风险：
  1. 回归风险极低：改动完全位于 docs snippet，不进入运行时 import 路径，最多影响文档站渲染。
  2. 数据一致性风险："defaultlbf16" 映射到外部 HF repo nvidia/NVIDIA-Nemotron-3.5-Lightning-30B-A3B-BF16，若 checkpoint 不存在、改名或需要登录 token，cookbook 会生成失效命令；此文件没有自动化校验。
  3. 配置漂移风险：BF16 与 NVFP4 两套配方并存，且包含平台特有参数（B200 的 flashinfer 后端、随机舍入、DFlash block size 6 vs 4、显存比例 0.85 vs 0.78），模型版本升级时容易漏改某一档位。
  4. 参数声明风险：--mem-fraction-static 0.85、--cuda-graph-max-bs-decode 16 等为建议值，未经 CI 验证，不同显存 / 并发环境可能需要调整。 - 影响：影响范围仅限文档站 cookbook。用户侧，新增 BF16 选项卡后，B200/H100/DGX Spark 用户可以一键复制包含部署、压测、评测的完整命令，减少手工拼接错误；团队侧，配方数量翻倍（NVFP4 12 组 + BF16 12 组），维护成本小幅上升，但无需改动代码或测试，属于低风险、低成本的文档增强。 - 风险标记：文档配方无自动化校验，BF16 checkpoint 路径依赖外部 HF 仓库，与 NVFP4 配方并存存在参数漂移风险

关联脉络

- PR #34524 Fix DFlash sliding attention causality defaults: 本 PR 的 DFlash 配方依赖 --speculative-algorithm DFLASH 与 block size 参数，34524 刚修复了 DFlash 滑动注意力因果默认值回归，二者属于同一功能面。
- PR #32227 [XPU] Fix NemotronH (hybrid mamba2) launch on --device xpu: 同属 Nemotron-H 杂交 Mamba2 模型支持线，本 PR 的 BF16 配方是该模型 cookbook 的进一步补充。
- PR #34379 [AMD] GLM 5.2 MXFP4 SGLANG COOKBOOK: 同为 cookbook snippet 量化配方先例，展示了相同的 quantizations + modelNames 映射模式，说明该 snippet 结构是

cookbook 的通用约定。