

PR #37109 完整报告

sgl-project/sglang

[Docs] Add NVFP4 section to GLM-5.3-Flash cookbook

合并时间: 2026-09-01 14:14

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/37109>

执行摘要

本 PR 为 GLM-5.3-Flash 部署食谱添加了 NVFP4 量化部分，通过更新配置文件和基准数据，扩展了模型在 Blackwell GPU 上的部署选项。变更纯文档性质，风险低，但需确保配置与验证硬件匹配。

功能与动机

动机是支持 RadixArk/GLM-5.3-Flash-NVFP4 检查点，该检查点使用 NVIDIA Model Optimizer 0.46.0 进行 NVFP4 W4A4 量化 (abs-max scaling, group size 16, MoE/MLP-only)。PR body 中说明: "Adds an NVFP4 quantization section to the GLM-5.3-Flash cookbook page covering the RadixArk/GLM-5.3-Flash-NVFP4 checkpoint..."，为用户提供了量化模型的部署指南。

实现拆解

1. 扩展配置维度: 在 docs/src/snippets/configs/zai-org/glm-5.3-flash.jsx 的 matchDims 中添加 quant 维度，定义 FP8 和 NVFP4 选项，并通过 disabled 函数限制 NVFP4 仅适用于 Blackwell 硬件 (gb300、gb200、b200、b300)。
2. 更新模型与精度映射: 在同一文件中添加模型名称 nvfp4: "RadixArk/GLM-5.3-Flash-NVFP4" 和精度标签 aime2026_pct。
3. 添加部署食谱: 在 cells 数组中为所有 Blackwell 硬件添加 NVFP4 配置单元，包括低延迟和高吞吐模式。GB300 食谱标记为已验证，其他硬件标记为未验证，并设置相应的服务器标志 (如 --quantization modelopt_fp4、--moe-runner-backend flashinfer_cutlass)。
4. 更新基准数据: 在 docs/src/snippets/configs/zai-org/glm-5.3-flash-benchmarks.jsx 中添加 NVFP4 基准条目，包括速度 (tokens_per_sec_per_gpu) 和准确性 (GSM8K、AIME 2026) 数据，基于 4x GB300 测量。
5. 清理过时配置: 在 docs/cookbook/autoregressive/GLM/GLM-5.3-Flash.mdx 中移除弃用的 --disable-shared-experts-fusion 标志。

[docs/src/snippets/configs/zai-org/glm-5.3-flash-benchmarks.jsx](#)

基准测试数据文件，添加了 NVFP4 的性能测量和评估结果，为用户提供量化模型的实际表现参考。

```
// 添加 NVFP4 低延迟基准条目
{
  match: { hw: "gb300", strategy: "low-latency", quant: "nvfp4", kvDsaPair: "bf16-tilelang" },
```

```
sglang_version: "033446bb05",
latencyPercentile: "Mean",
speed: [
  {
    workload: {
      dataset: "random",
      isl: 1024,
      osl: 256,
      max_concurrency: 16,
      num_prompts: 80,
    },
    ttft_ms: 242.66,
    tpot_ms: 16.95,
    tokens_per_sec_per_gpu: 1022.61,
  },
],
accuracy: { gsm8k_pct: 97.14, aime2026_pct: 92.45 },
notes:
  "RadixArk/GLM-5.3-Flash-NVFP4 — NVFP4 W4A4 post-training quantization ... Measured on
  4x GB300 with the lmsysorg/sglang:glm-5.3-flash image (PR #36507 head 033446bb05),
  adaptive MTP 5/1/6, BF16 KV + TileLang DSA, and the flashinfer_cutlass MoE runner.",
},
```

评论区精华

无 review 讨论，仅有一个批准评论（reviewer: zijiaxia，状态：APPROVED，内容为空）。因此没有技术交锋或设计争议。

风险与影响

- 风险：配置准确性风险较低，NVFP4 选项已限制为 Blackwell 硬件；基准数据时效性风险，已记录版本和测量条件。
- 影响：用户扩展了部署选择，系统无代码变更，团队需维护文档与实际验证同步。

关联脉络

本 PR 与历史 PR #36507（GLM-5.3-Flash 支持）相关，后者为当前文档提供了模型基础。PR #35120 涉及 NVFP4 内核，展示了量化技术的演进。当前 PR 独立完成了文档扩展，无其他近期 PR 直接关联。