PR #29380 完整报告

sgl-project/sglang

[Docs] Add NVFP4 quantization to GLM-5.2 cookbook

合并时间：2026-06-26 15:40

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/29380

执行摘要

本 PR 为 GLM-5.2 部署文档新增 NVFP4 量化选项，限定 Blackwell Ultra（B300/GB300），同时优化 _deployment.jsx 以支持 per-quant Docker 镜像映射。

功能与动机

NVIDIA Model Optimizer 提供了 NVFP4 精度的 GLM-5.2 checkpoint（nvidia/GLM-5.2-NVFP4），仅量化 MoE 专家线性层，精度接近 FP8 基线。需要将其集成到 GLM-5.2 部署文档中，让 Blackwell Ultra 用户能够一键部署。

实现拆解

1. 添加 NVFP4 量化选项：在 glm-5.2.jsx 的 quantizations 数组中增加 { id: "nvfp4", label: "NVFP4" }，并设置对应的模型名 nvidia/GLM-5.2-NVFP4。
2. 创建 deployment cells：为该量化新增四个已验证的部署单元（B300 low-latency、B300 balanced、GB300 low-latency、GB300 balanced），使用 TP4 和 --quantization modelopt_fp4，其中 low-latency 策略启用 EAGLE 推测解码。
3. 修改 Docker 镜像解析引擎：在 _deployment.jsx 中将镜像选择逻辑从仅按 hw 查找改为优先按 hwlquant 再按 hw，使 NVFP4 能自动使用专用开发镜像 lmsysorg/sglang:dev-glm52-nvfp4。
4. 更新主文档页面和基准桩：在 GLM-5.2.mdx 中添加 NVFP4 模型介绍段落、模型表行和资源行，并在部署面板上方增加 Note 提示镜像自动切换行为；在 glm-5.2-benchmarks.jsx 中添加无数据的 NVFP4 benchmark 占位条目。
5. 同步技能模板和参考文档：更新 .claude/skills/cookbook-add-model/templates/config.jsx.tmpl 和 authoring-reference.md 中的 dockerImages 说明，引导后续作者使用 hwlquant 复合键。

docs_new/src/snippets/configs/zai-org/glm-5.2.jsx

核心配置文件，添加 NVFP4 量化选项、模型名、per-quant Docker 镜像及四个已验证的 deployment cells

```
// 新增 NVFP4 量化选项、模型名、per-quant Docker 镜像及已验证 cell
export const config = {
  quantizations: [
    { id: "fp8", label: "FP8" },
    { id: "bf16", label: "BF16" },
```

```
    { id: "nvfp4", label: "NVFP4" }, // 新增 NVFP4 量化
  ],
  modelNames: {
    "defaultlfp8": "zai-org/GLM-5.2-FP8",
    "defaultlbf16": "zai-org/GLM-5.2",
    "defaultlnvfp4": "nvidia/GLM-5.2-NVFP4", // NVIDIA 官方 NVFP4 checkpoint
  },
  dockerImages: {
    h200: "lmsysorg/sglang:latest",
    b200: "lmsysorg/sglang:latest",
    gb300: "lmsysorg/sglang:latest",
    b300: "lmsysorg/sglang:latest",
    // NVFP4 需要包含 modelopt_fp4 支持的开发镜像（per-quant 覆盖）
    "b300lnvfp4": "lmsysorg/sglang:dev-glm52-nvfp4",
    "gb300lnvfp4": "lmsysorg/sglang:dev-glm52-nvfp4",
  },
  cells: [
    // 已验证的 NVFP4 B300 Low-Latency cell
    {
      match: { hw: "b300", variant: "default", quant: "nvfp4", strategy: "low-latency", nodes:
      "single" },
      verified: true,
      env: [],
      flags: [
        "--trust-remote-code",
        "--model-path {{MODEL_NAME}}",
        "--tp 4",
        "--quantization modelopt_fp4",
        "--speculative-algorithm EAGLE",
        "--speculative-num-steps 5",
        "--speculative-eagle-topk 1",
        "--speculative-num-draft-tokens 6",
        "--chunked-prefill-size 131072",
        "--mem-fraction-static 0.70",
        "--host {{HOST_IP}}",
        "--port {{PORT}}",
      ],
    },
    // 其余 NVFP4 cells 类似（B300 balanced、GB300 low-latency、GB300 balanced）
  ],
};
```

docs_new/src/snippets/_deployment.jsx

部署引擎核心，修改 Docker 镜像解析逻辑以支持 per-quant 键

```
// 在 renderCommand 函数中，docker 模式下的镜像选择逻辑变更
if (mode === "docker") {
  // 之前的逻辑：仅按 hw 取镜像
  // const image = (config.dockerImages && config.dockerImages[sel.hw]) || "lmsysorg/sglang:
```

```
    dev";

    // 修改后：优先按 hwlquant 键，再按 hw，最后 fallback 到 :dev
    const di = config.dockerImages || {};
    const image = di[`${sel.hw}|${sel.quant}`] || di[sel.hw] || "lmsysorg/sglang:dev";

    // … 后续端口解析和命令生成不变
  }
```

评论区精华

本 PR 无审核评论，由 Fridge003 直接批准。

风险与影响

风险：

- _deployment.jsx 的镜像选择逻辑变更为先匹配 hwlquant 再回退 hw，若其他模型使用了类似格式的键名可能导致误匹配，但当前仅 GLM-5.2 定义了该键，影响可控。
- NVFP4 cells 的 benchmark 数据仍为占位符，可能给用户带来性能预期偏差。

影响：

- 用户：GLM-5.2 用户可通过文档页面一键生成 NVFP4 部署命令，Docker 模式自动使用正确镜像。
- 系统：所有 cookbook 页面都能利用 per-quant 镜像重写能力。
- 团队：更新了技能文档，引导后续作者使用 hwlquant 复合键。

关联脉络

本 PR 是 GLM-5.2 cookbook 的后续功能扩展，在已有 FP8/BF16 基础上增加 NVFP4 量化选项，完善对 Blackwell Ultra 硬件的覆盖。未发现与其他活跃 PR 的直接依赖。