PR #34379 完整报告

sgl-project/sglang

[AMD] GLM 5.2 MXFP4 SGLANG COOKBOOK

合并时间：2026-08-12 18:39

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/34379

执行摘要

- 一句话：GLM-5.2 cookbook 新增 AMD MI355X MXFP4 部署配方
- 推荐动作：值得浏览（不必精读）：这是标准的 cookbook 配置扩展 PR，展示了 SGLang 文档配置系统（Mintlify snippets + JSX config）如何组织多硬件 / 多量化组合的部署矩阵，以及如何在未验证状态下通过 disable/enable 门控与 verified: false 标记管理文档诚实性。值得关注的设计决策： 1) 用 "hwlquant" 复合键在 dockerImages 中做细粒度镜像覆盖； 2) 用 enable/enableReason 门控只对特定 hw+quant 组合开放 MTP 选项； 3) 对 gfx950 专属能力（MXFP4、MTP 3-1-4）的显式范围界定。后续可跟进其 TODO 中的准确率 / 速度数据是否在后续 PR 中补齐并翻转 verified 标记。

功能与动机

AMD 为 MI355X (gfx950) 发布了 Quark 量化的 MXFP4 版 GLM-5.2（amd/GLM-5.2-MXFP4），而现有 cookbook 只覆盖 FP8/BF16（全 AMD SKU）和 NVFP4（NVIDIA Blackwell）。PR body 说明目标是让 MI355X 用户获得可直接运行的文档化配方，并沿用 amd/GLM-5.1-MXFP4 的先例。Issue 评论中 seungrokj 提到需要合入以将 SGL 性能数据推送到 SA（"We need this merged to push SGL perf data to SA"），说明该配方也是 AMD 侧性能数据采集的前置条件。

实现拆解

实现分为两步，全部为文档 / 配置改动，不触碰运行时代码：

1. 注册 MXFP4 量化选项与模型映射（docs/src/snippets/configs/zai-org/glm-5.2.jsx）：在 quantizations 列表新增 { id: "mxfp4", label: "MXFP4" }，在 modelNames 中新增 "defaultlmxfp4": "amd/GLM-5.2-MXFP4" 映射，并在 dockerImages 中为 MI355X + MXFP4 组合固定独立的镜像 lmsysorg/sglang-rocm:v0.5.16-rocm720-mi35x-20260728（比现有 MI355X FP8/BF16 镜像更新）。
2. 新增 MI355X MXFP4 部署 cell 与 MTP 推测解码选项（同文件）：新增三个 hw: "mi355x", quant: "mxfp4" 的 cell（low-latency / balanced / high-throughput），统一使用 tp=4（4-bit MoE 权重可放入 4-GPU 切片）、--trust-remote-code（Quark 自定义量化配置需要）和 --kv-cache-dtype fp8_e4m3，参数继承自已验证的 amd/GLM-5.1-MXFP4 MI355X 配方；三个 cell 均标记 verified: false。同时在 playgroundFeatures.speculative.options 中新增 mtp-314（EAGLE / MTP 3-1-4）选项，通过 enable: { hw: ["mi355x"], quant: ["mxfp4"] } 条件门控，仅对 MI355X MXFP4 开

放，并给出 GSM8K em_strict 0.971 的验证数据与 SGLANG_SIMULATE_ACC_LEN=2.99 的基准建议。需要注意：MXFP4 是 gfx950 专属，未对 MI300X/MI325X (gfx942) 开放，且 DSA-CP、MTP 516/112 等在 AMD 上仍被 disable 门控限制。

3. 更新 Markdown 文档（docs/cookbook/autoregressive/GLM/GLM-5.2.mdx）：模型介绍段落补充 AMD MXFP4 build 及其未验证状态；模型变体表新增 GLM-5.2-MXFP4 行与 HF 资源链接 https://huggingface.co/amd/GLM-5.2-MXFP4；Configuration Tips 新增 "MI355X MXFP4 (gfx950-only)" 提示，说明 tp=4、--trust-remote-code、--kv-cache-dtype fp8_e4m3 等要求。
4. 测试与验证配套：无测试文件变更（纯文档改动，PR checklist 标注 N/A）；准确率（GSM8K/AIME25 via sgl-eval）与速度（benchmark_serving）结果均留 TODO，配方以 unverified 状态合入。

关键文件：

- docs/src/snippets/configs/zai-org/glm-5.2.jsx（模块 文档配置；类别 source；类型 core-logic；符号 config.quantizations, config.modelNames, config.dockerImages, config.playgroundFeatures.speculative.options）：cookbook 的核心配置源文件：注册 mxfp4 量化、映射 amd/GLM-5.2-MXFP4 模型名、固定 MI355X MXFP4 专用镜像、新增 3 个 MI355X MXFP4 部署 cell 和 1 个 MTP 3-1-4 门控选项，是本次变更的主体（+1258/-1151）。
- docs/cookbook/autoregressive/GLM/GLM-5.2.mdx（模块 文档页面；类别 other；类型 core-logic）：面向读者的 Markdown 文档：模型介绍、变体表和资源链接补充 AMD MXFP4 build，Configuration Tips 新增 MI355X MXFP4 (gfx950-only) 配置说明，明确 unverified 状态（+8/-2）。

关键符号：未识别

评论区精华

该 PR 的 review 评论极少：1am9trash 直接 APPROVED（"LGTM"），HaiShaw 也 APPROVED（无正文），无 review 评论线程。Issue 评论中 seungrokj 请求 sogalin 尽快 review 以推进合入（用于推送 SGL 性能数据到 SA），functionstackx 引用了 HaiShaw。主要讨论点实质上是流程性的：待办事项（Accuracy/Speed Tests TODO）与 unverified 状态已在 PR body 和文档中显式声明，没有引发关于配置正确性的技术争议。

- 暂无高价值评论线程

风险与影响

- 风险：主要风险集中在文档 / 配置层面而非运行时： 1) 新配方全部标记 verified: false，未经过 GLM-5.2 实际基准验证，参数（tp=4、--kv-cache-dtype fp8_e4m3 等）是从 amd/GLM-5.1-MXFP4 推断迁移而来，同架构家族推断合理但存在偏差可能； 2) 镜像固定策略引入依赖风险：dockerImages["mi355x|mxfp4"] 指向 v0.5.16-rocm720-mi35x-20260728（比现有 MI355X 镜像更新），若该镜像在 gfx950 上存在类似此前 FP8 误编译（见文档中关于 gfx950 block-FP8 的 Note）的隐藏问题，文

档用户会直接受影响； 3) mtp-314 选项依赖 gfx950 spec-decode draft kernel 在 num-steps=3 内的构建包络（≤3），若上游内核变更可能失效； 4) 配置键 "mi355xlmxfp4" 的复合键写法依赖 snippets 引擎的解析约定，改动无测试覆盖； 5) 文档中引用的模型链接与镜像 tag 均为外部资源，存在失效风险。整体不影响既有运行时路径，回归风险很低。

- 影响：影响范围：面向文档用户（部署 GLM-5.2 到 AMD MI355X 的用户）和 AMD 工程团队（性能数据采集流程）。对现有 FP8/BF16/NVFP4 配方无破坏性影响，新增选择器与配置键均为增量式。影响程度中等偏下：属于 cookbook 能力扩展，为 MI355X 用户提供可直接复制的 MXFP4 部署路径，同时新增了一条 MTP 3-1-4 推测解码选项（AMD 侧首批被 enable 的 spec-decode 选项之一）。对团队而言，提供了 AMD gfx950 平台 GLM-5.2 MXFP4 的标准化部署基线，但 unverified 状态意味着准确性 / 性能数据尚未闭环。
- 风险标记：配方未经验证，依赖外部镜像 tag，无测试覆盖，配置键无 schema 校验

关联脉络

- PR #34476 [AMD][DI][CI] Add GLM-5.2 MXFP4 wide-EP16 2P1D nightly recipes: 同为 AMD MI355X + GLM-5.2 MXFP4 主题，一个补 cookbook 配方，一个补 CI 夜间测试配方，形成配套的部署文档与验证基础设施闭环。
- PR #33997 Bump FlashInfer to 0.6.17 and remove Kimi K3 workarounds: 与 MXFP4 量化（mxfp4.py）相关，是本 PR 引用的量化技术栈（Quark 量化 + MXFP4）的邻近改动，影响 MI355X 上的量化推理路径。
- PR #33945 feat: support deterministic FA4 for GLM-4.7-Flash: 同为 quant + 文档相关的 cookbook 演进（server_args 文档、注册测试），体现仓库在 Blackwell/AMD 量化部署文档上的持续投入脉络。