PR #35445 完整报告

sgl-project/sglang

[AMD] cookbook: serve Qwen3.5 MXFP4 on MI355X with an fp8_e4m3 KV cache

合并时间：2026-08-19 18:49

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/35445

执行摘要

本 PR 为 Qwen3.5 部署 cookbook 的 AMD MI355X MXFP4 配方新增 --kv-cache-dtype fp8_e4m3 参数，使文档生成的命令与公开基准测试所用的 KV 缓存配置完全一致，避免用户照抄命令时得到不同的性能配置。改动仅 1 行，属于纯文档片段修正，无运行时影响。

功能与动机

Qwen3.5 cookbook 中 MI355X 的 MXFP4 配方虽然标称基于 FP8 KV 缓存进行基准测试，但生成的部署命令却遗漏了 --kv-cache-dtype fp8_e4m3 参数。这导致用户直接复制命令时会使用默认的 FP16 KV 缓存，与文档声明的 FP8 配置不符，进而无法复现公开的 MI355X 性能数据。B200 NVFP4 配方早已包含该参数，本 PR 让 AMD FP4 配方与之对齐，确保文档与实际测量配置一致。

实现拆解

1. 定位生成逻辑： 在 docs/src/snippets/autoregressive/qwen35-deployment.jsx 的 Qwen35Deployment 组件中，找到 FP4 配方的后端参数生成分支。
2. 修改 MI355X 分支： 在 mi355x 分支中，紧接 --disable-radix-cache 之后新增一行 cmd += ' \ --kv-cache-dtype fp8_e4m3';，仅当硬件为 mi355x 且量化方式为 fp4 时添加该参数。
3. 保持其他分支不变： B200/B300 的 NVFP4 分支原本就有该参数，本次未修改；其他硬件和量化组合也不受影响。
4. 测试与 CI： 纯文档字符串生成逻辑改动，无新增测试文件；CI 已通过（PR Test 通过，Extra 失败但未给出具体原因）。

docs/src/snippets/autoregressive/qwen35-deployment.jsx

这是本 PR 唯一修改的文件，新增 --kv-cache-dtype fp8_e4m3 到 MI355X MXFP4 分支，使文档命令与基准配置对齐。

关键源码片段

docs/src/snippets/autoregressive/qwen35-deployment.jsx

这是本 PR 唯一修改的文件，新增 --kv-cache-dtype fp8_e4m3 到 MI355X MXFP4 分支，使文档命令与基准配置对齐。

```
// FP4-specific backend settings
```

```
if (quantization === 'fp4') {
  if (hardware === 'mi355x') {
    // AMD MXFP4 on MI355X: backend / --page-size 16 以及 INT8 量化的 ROCm quick all-reduce
    环境变量
    // 已由上面的 AMD backend 块输出（该配方使用 quick all-reduce 而非 AITER allreduce
    fusion）。
    // 在此处添加 FP4 专用参数:
    cmd += ' \
  --disable-radix-cache';
    cmd += ' \
  --kv-cache-dtype fp8_e4m3'; // 新增：与基准测试配置保持一致
    // 在开启 MTP 且 tp=2 时限制并发数，避免 OOM。
    if (speculative === 'enabled') {
      cmd += ' \
  --max-running-requests 128';
    }
  } else {
    // NVIDIA NVFP4 on Blackwell (B200 / B300)，此分支原本已包含 kv-cache-dtype fp8_e4m3
    ...
  }
}
```

评论区精华

无 review 评论，两位 reviewer 直接批准通过。

风险与影响

- 风险：几乎为零。改动仅影响文档片段生成的命令字符串，不涉及运行时代码。唯一行为变化是用户复制的命令会多出 --kv-cache-dtype fp8_e4m3，这可能改变 KV 缓存精度（从默认 FP16 变为 FP8），但这正是文档所声明并给出警告的基准配置。
- 影响：对用户，复制 MI355X MXFP4 命令的用户将获得与基准数据一致的配置；对系统，无运行时影响；对团队，文档与实测对齐，避免混淆。

关联脉络

本 PR 与近期 Qwen 系列模型的支持工作（如 Qwen3.8-27B 模型支持）属于同一功能线，持续完善 Qwen 模型在 SGLang 中的部署文档和配置。与 B200 NVFP4 配方相比，本 PR 只是补齐了 AMD 分支的对称性，整体方向是让 cookbook 覆盖更多硬件和量化组合，并确保文档命令与实际测量配置一致。