

PR #36246 完整报告

sgl-project/sglang

[AMD] Add Kimi-K2.7-Code-MXFP4 to cookbook

合并时间: 2026-08-25 17:37

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/36246>

执行摘要

- 一句话: cookbook 新增 Kimi-K2.7-Code MXFP4 部署指南与命令生成器
- 推荐动作: 值得精读 jsx 中 `getDynamicItems` 与 `condition` 的联动写法, 以及“文档命令与实测运行对齐”的迭代过程; 有 MI350X/MI355X 部署需求者可直接采用页面命令。整体为常规文档 PR, 不必深入代码审查。

功能与动机

PR body 指出, SGLang 已在 AMD MI350X/MI355X 上验证 Kimi-K2.7-Code 的 MXFP4 检查点, 但 cookbook 页面只展示了 INT4 检查点, 用户没有现成入口。本次变更把 `amd/Kimi-K2.7-Code-MXFP4` 的部署命令、配置要点和实测基准补进文档, 让用户可以直接复制使用。

实现拆解

1. 在 `docs/src/snippets/autoregressive/kimi-k27-code-deployment.jsx` 的 `options` 中新增 `quantization` 选项: 通过 `getDynamicItems` 依据当前硬件动态返回 INT4/MXFP4, MI350X/MI355X 默认 MXFP4, 其余硬件置灰并附 `disabledReason`。
2. 为 `reasoning`、`toolcall`、`dpattention` 三个选项增加 `condition` 条件, MXFP4 模式下隐藏这些控件, 因为解析器与 DP attention 在 MXFP4 检查点上自动检测, 避免用户配置冲突。
3. 在 `generateCommand` 中新增 `isMXFP4` 分支: 直接返回验证过的完整命令, 包含 11 个 AITER/ROCm 环境变量、`--tp 4`、`--attention-backend aiter`、`--mem-fraction-static 0.90`、`--kv-cache-dtype fp8_e4m3`、`--disable-radix-cache`、`--enable-aiter-allreduce-fusion` 等; 非 MXFP4 路径保持不变。
4. 在 `Kimi-K2.7-Code.mdx` 中把 MXFP4 检查点加入 Available Models, 并在 Configuration Tips 中补充支持硬件、TP=4、镜像版本、自动检测量化等说明; 新增 MXFP4 部署命令示例。
5. 新增 5.2 MXFP4 Benchmark 小节 (GSM8K 精度 + `bench_serving` 延迟), 并重排章节使官方 Model Card 数据在前。无测试或配置文件改动。

关键文件:

- `docs/src/snippets/autoregressive/kimi-k27-code-deployment.jsx` (模块 命令生成器; 类别 `source`; 类型 `core-logic`; 符号 `KimiK27CodeDeployment`, `generateCommand`, `getDynamicItems`): 交互式命令生成器的核心改动: 新增 Quantization 维度与 MXFP4 输

出分支，是本次文档功能的主要实现载体，直接影响用户复制到的命令。

- docs/cookbook/autoregressive/Moonshotai/Kimi-K2.7-Code.mdx（模块 部署文档；类别 other；类型 documentation）：cookbook 页面的用户可见内容：模型清单、配置提示、MXFP4 部署命令与基准测试，决定文档信息的完整性与可执行性。

关键符号：KimiK27CodeDeployment, generateCommand, getDynamicItems

关键源码片段

docs/src/snippets/autoregressive/kimi-k27-code-deployment.jsx

交互式命令生成器的核心改动：新增 Quantization 维度与 MXFP4 输出分支，是本次文档功能的主要实现载体，直接影响用户复制到的命令。

```
export const KimiK27CodeDeployment = () => {
  // 新增 Quantization 维度：根据所选硬件动态决定 INT4 / MXFP4 是否可用
  const options = {
    quantization: {
      name: 'quantization',
      title: 'Quantization',
      getDynamicItems: (values) => {
        // 只有 MI350X / MI355X 允许 MXFP4，其他硬件默认 INT4 并置灰 MXFP4
        const isMXFP4 = ['mi350x', 'mi355x'].includes(values.hardware);
        return [
          { id: 'int4', label: 'INT4', subtitle: 'Base checkpoint', default: !isMXFP4 },
          {
            id: 'mxfp4',
            label: 'MXFP4',
            subtitle: 'AMD FP4',
            default: isMXFP4,
            disabled: !isMXFP4,
            disabledReason: !isMXFP4 ? 'MXFP4 only on AMD MI350X/MI355X' : '',
          },
        ];
      },
    },
  },
  // MXFP4 模式下解析器与 DP attention 均自动检测，故在 UI 中隐藏对应选项
  reasoning: { condition: (values) => values.quantization !== 'mxfp4' },
  toolcall: { condition: (values) => values.quantization !== 'mxfp4' },
  dpattention: { condition: (values) => values.quantization !== 'mxfp4' },
};

const generateCommand = () => {
  const { hardware, quantization } = values;
  // MXFP4 走独立分支：直接返回与 launch_cookbook.sh 对齐的完整验证命令
  if (quantization === 'mxfp4') {
    const mxfp4Env = [
      'SGLANG_USE_AITER=1',
      'HIP_FORCE_DEV_KERNARG=1',
      'SGLANG_EXPERT_PARALLEL_SIZE=1',
    ],
```

```

'SGLANG_USE_DYNAMIC_MXFP4_LINEAR=0',
'TORCH_BLAS_PREFER_HIPBLASLT=1',
'TENSILE_STREAMK_DYNAMIC_GRID=6',
'AITER_QUICK_REDUCE_QUANTIZATION=INT4',
'AITER_USE_FLYDSL_MOE_SORTING=1',
'AITER_AR_1STAGE_MAX_KB=512',
'AITER_MXFP4_INTERMEDIATE=1',
'ROCM_QUICK_REDUCE_QUANTIZATION=INT4',
].join(' \\n');
return (
  mxfp4Env + ' \\n' +
  'sglang serve \\n' +
  ' --model-path amd/Kimi-K2.7-Code-MXFP4 \\n' +
  ' --tp 4 \\n' +
  ' --trust-remote-code \\n' +
  ' --attention-backend aiter \\n' +
  ' --mem-fraction-static 0.90 \\n' +
  ' --kv-cache-dtype fp8_e4m3 \\n' +
  ' --disable-radix-cache \\n' +
  ' --enable-aiter-allreduce-fusion \\n' +
  ' --host 0.0.0.0 \\n' +
  ' --port 30000'
);
}
// 非 MXFP4 分支沿用原逻辑，构建带 --reasoning-parser 等的命令
};
};

```

评论区精华

两位维护者 sogalin 与 1am9trash 均直接批准 (LGTM)，没有实质 review 意见。提交历史反映出作者在合并前反复校准文档：将 checkpoint 仓库 id 从 moonshotai 修正为 amd，补充验证镜像版本，并将生成命令升级为与 launch_cookbook.sh 完全一致的全量版本。

- 暂无高价值评论线程

风险与影响

- 风险：
 - 命令参数硬编码：MXFP4 分支把 11 个环境变量与启动参数写死，若 SGLang 或 ROCm 侧参数改名或默认值变化，文档会过期。
 - 参数不一致风险：PR body 中的验证命令使用 --mem-fraction-static 0.765，而最终合并代码与 mdx 为 0.90，两处口径不一致，读者可能困惑。
 - FP8 KV cache 精度权衡：mdx 已注明精度取舍，但默认开启可能影响精度敏感用户。
 - UI 依赖：MXFP4 会隐藏 parser/dpattention 选项，若用户通过其他途径绕过 UI 设置，生成的命令可能与预期不符。

- 影响：影响 AMD MI350X/MI355X 部署用户，提供开箱即用的 MXFP4 命令；对运行时零影响；与 Kimi-K2.6 页面 NVFP4 交互模式保持一致，降低后续维护复杂度。
- 风险标记：文档变更无测试覆盖，mem-fraction-static 正文与代码不一致 (0.765 vs 0.90)，生成命令携大量硬编码环境变量

关联脉络

- 暂无明显关联 PR