

PR #27726 完整报告

sgl-project/sglang

[Docs] Update MegaMoE handling and rerun benchmarks

合并时间: 2026-06-10 10:00

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/27726>

执行摘要

本 PR 更新 DeepSeek-V4 cookbook 页面的 Playground 组件和 Benchmark 数据，将 MegaMoE 从独立 axis 合并到 moe axis 内，高吞吐策略从 W4A4 改为 W4A8 以避免精度下降，并刷新所有基准数字。同步更新 cookbook-add-model 技能模板和文档。

功能与动机

根据 PR body，变更动机为：

1) 更新 Playground 中的 MegaMoE 处理； 2) 高吞吐策略从 W4A4 改用 W4A8 因精度下降问题； 3) 更新基准结果。此前 MegaMoE 是一个独立 playground axis，现在将其整合到 moe axis 作为后端选项，并增加量化子选择，简化 UI。

实现拆解

1. Playground 引擎 (`_playground.jsx`): 移除 `megamoe` axis 常量定义和 `AXIS_HANDLERS` 条目；在 `moe` axis handler 中增加 `mmQuant` 状态 (`'w4a8'/'w4a4'`)，新增 `deriveFromBase` 从 base cell 的环境变量 (`SGLANG_OPT_DEEPGEMM_MEGA_MOE_USE_FP4_ACTS`) 派生默认量化；`revertHidden` 增加 MegaMoE 可用性门控 (`requiresHw/excludesStrategy`)；`apply` 方法注入对应环境变量。
2. 模型配置 (`deepseek-v4.jsx`): 将 `moe` 轴 backend option 中的 `megamoe` 条目的禁用逻辑替换为 `requiresHw` 门控；新增 `megamoeQuant` 子配置块 (含 `stripEnv` 和 W4A8/W4A4 的环境变量)；调整高吞吐策略默认使用 W4A8；增加 `--warmup-requests 64` 到 speed 命令；`numPromptsByConc` 中并发 1 时从 8 改为 32。
3. 基准数据 (`deepseek-v4-benchmarks.jsx`): 基于 `sglang v0.5.12.post1` 重新测量 B200/B300 各策略的 TTFT、TPOT 和 `tokens_per_sec_per_gpu`，部分数值显著提升 (如 B200 low-latency flash TTFT 从 428ms 降至 87ms)，并移除部分 `pro high-throughput` 条目。
4. 技能模板与文档: 同步更新 `.claude/skills/cookbook-add-model/` 下的 `config.jsx.tmpl` (增加 MegaMoE/megamoeQuant 注释模板)、`engine-axis.md` (轴数量从 8 减为 7, MegaMoE 归入 moe) 和 `authoring-reference.md` (moe 轴说明更新)。

`docs_new/src/snippets/_playground.jsx`

核心 Playground 引擎变更：将 MegaMoE 从独立 axis 迁移到 moe axis 内，作为后端选项并添加量化子选择，同时更新注释和条件逻辑。

```

// moe axis handler 关键方法（合并 MegaMoE 后）
moe: {
  initState: () => ({ backend: null, ep: null, mmQuant: null }),

  // 从 base cell 推导初始状态：检查环境变量判断是否为 W4A4
  deriveFromBase: (cell, fc, h) => {
    const flags = (cell && cell.flags) || [];
    const baseEnv = (cell && cell.env) || [];
    const a2a = h.findFlagArg(flags, "--moe-a2a-backend");
    const runner = h.findFlagArg(flags, "--moe-runner-backend");
    // 检测 FP4 activations 环境变量以区分 W4A8/W4A4
    const fp4Acts = baseEnv.some(
      (e) => e.startsWith("SGLANG_OPT_DEEPGEMM_MEGA_MOE_USE_FP4_ACTS"));
    return {
      backend: a2a || runner || null,
      ep: h.parseIntFlag(flags, "--ep"),
      mmQuant: fp4Acts ? "w4a4" : "w4a8",
    };
  },

  // 回退隐藏：当用户选择被隐藏时置 null；新增 MegaMoE 可用性门控
  revertHidden: (value, fc, base, h) => {
    let changed = false;
    const next = { ...value };
    if (next.backend !== null && fc.backend?.options
      && h.isHidden(fc.backend.options, next.backend, base)) {
      next.backend = null; changed = true;
    }
    // 检查 MegaMoE 选项是否因硬件 / 策略被隐藏
    const mmOpt = (fc.backend?.options || []).find((o) => o.id === "megamoe");
    const mmAvail = !!mmOpt
      && (!mmOpt.requiresHw || mmOpt.requiresHw.includes(base.hw))
      && (!mmOpt.excludesStrategy || !mmOpt.excludesStrategy.includes(base.strategy));
    if (next.backend === "megamoe" && !mmAvail) {
      next.backend = null; changed = true;
    }
    if (next.ep !== null && fc.ep?.values
      && h.isHidden(fc.ep.values, next.ep, base)) {
      next.ep = null; changed = true;
    }
    return changed ? next : value;
  },

  // apply 方法也相应更新（注入 MegaMoE 量化环境变量）
}

```

[docs_new/src/snippets/configs/deepseek-ai/deepseek-v4.jsx](#)

DeepSeek-V4 配置更新：调整 MegaMoE backend 选项（移除原有 megamoe axis 配置，改为 moe 轴内的子选项）、增加 megamoeQuant 量子化配置、更新高吞吐策略为 W4A8、调整 numPromptsByConc 和添加 warmup-requests。

```
// moe 轴 backend 选项中的 MegaMoE 条目（新增 requiresHw 硬件门控）
options: [
  { id: null, label: "Inherited" },
  { id: "deepep", label: "DeepEP",
    flags: ["--moe-a2a-backend deepep"] },
  // Blackwell-only; 无策略门控，playground 允许任何策略实验
  { id: "megamoe", label: "MegaMoE",
    flags: ["--moe-a2a-backend megamoe"],
    requiresHw: ["b200", "b300", "gb200", "gb300"] },
  { id: "flashinfer_mxfp4", label: "FlashInfer (MXFP4)",
    flags: ["--moe-runner-backend flashinfer_mxfp4"] },
  { id: "marlin", label: "Marlin (W4A16)",
    flags: ["--moe-runner-backend marlin"] },
],

// 新增：MegaMoE 量子化选择（仅在 backend === "megamoe" 时显示）
megamoeQuant: {
  // 选择任一量化都会 strip DeepEP 环境变量
  stripEnv: ["SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK"],
  options: [
    { id: "w4a8", label: "W4A8",
      // 仅设置 NUM_MAX
      env: ["SGLANG_OPT_DEEPEGEMM_MEGA_MOE_NUM_MAX_TOKENS_PER_RANK=8320"] },
    { id: "w4a4", label: "W4A4",
      // 额外设置 FP4 激活和 MXF4 内核
      env: [
        "SGLANG_OPT_DEEPEGEMM_MEGA_MOE_NUM_MAX_TOKENS_PER_RANK=8320",
        "SGLANG_OPT_DEEPEGEMM_MEGA_MOE_USE_FP4_ACTS=1",
        "SGLANG_OPT_DEEPEGEMM_MEGA_MOE_USE_MXF4_KIND=1",
      ] },
  ],
},
```

评论区精华

无 review 评论，仅有 wisclmy0611 的 approve。PR 为自包含文档更新，未引发技术讨论。

风险与影响

- 风险：很低。主要关注文档配置与运行时实际支持的选项可能存在偏差（如 requiresHw 与实际硬件限制）；基准数字基于特定 sglang 版本，用户运行结果可能不同。
- 影响：仅影响 DeepSeek-V4 cookbook 页面的 Playground 组件和 Benchmark 卡片，用户通过文档获取的配置示例将反映 MegaMoE 的新交互方式；cookbook 作者需注意 MegaMoE 不再作为独立 axis。

关联脉络

本 PR 与近期的 DeepSeek-V4 cookbook 相关工作（如 27599 spec 更新）相关，属于持续完善文档和用户体验的一部分。无关联 Issue。