PR #25506 完整报告

sgl-project/sglang

[Doc] Fix several places for dpsk v4 cookbook

合并时间：2026-05-17 12:54

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25506

执行摘要

此 PR 修复 DeepSeek V4 cookbook 中的多处文档错误（缺失反引号、格式不一致），并在部署交互式命令生成器中新增 MegaMoE 功能的硬件兼容性逻辑，确保用户无法选择无效配置。变更涉及两个文件：JSX 组件和 MDX 文档，均为文档与 UI 层面的修正和增强。

功能与动机

修复 DeepSeek V4 cookbook 中的文档错误，包括 flashinfer_mxfp4 标识缺失反引号导致的渲染问题，以及 Notes: 标题未加粗的格式不一致。同时，补充 MegaMoE 功能的说明，涵盖其 W4A8 和 W4A4 变体、硬件与部署方案的限制，以及使用注意事项。

实现拆解

1. JSX 部署生成器 (deepseek-v4-deployment.jsx)：
   - 新增 MEGAMOE_UNSUPPORTED_RECIPES 和 MEGAMOE_UNSUPPORTED_HARDWARE 常量，定义不支持的部署方案（low-latency、cp）和硬件（h100、h200、h200-fp4）。
   - 新增 isMegamoeUnsupported(vals) 函数，用于判断当前配置是否支持 MegaMoE。
   - 在 resolveItems 函数中增加对 megamoe 选项的处理：当配置不支持时，将非 disabled 的选项禁用并显示原因（“MegaMoE is only supported on Blackwell” 或 “MegaMoE is not supported on this recipe”）。
   - 在 handleRadioChange 中添加回退逻辑：当切换硬件或部署方案导致 MegaMoE 不支持时，自动将 MegaMoE 选项重置为 disabled。
2. MDX 文档 (DeepSeek-V4.mdx)：
   - 修复 flashinfer_mxfp4 标识缺失的反引号。
   - 将 Notes: 改为加粗格式 **Notes:**。
   - 新增 MegaMoE 功能说明段落，包括 W4A8 和 W4A4 变体、硬件限制（仅 Blackwell）以及注意事项（如 --moe-runner-backend 不可手动设置）。
   - 修正 H200 原始 FP4 checkpoint 的说明，明确提供 Marlin 和 Flashinfer 两种 w4a16 MoE 内核选项。

docs_new/src/snippets/autoregressive/deepseek-v4-deployment.jsx

核心UI逻辑变更：新增MegaMoE硬件与部署方案兼容性判断函数isMegamoeUnsupported、相关常量及禁用逻辑，确保交互式命令生成器正确显示可用选项。

```
// MegaMoE is only supported on Blackwell with DeepEP-based recipes
// (balanced / max-throughput / pd-disagg). It's disabled on Hopper
// (H100 / H200 / H200-FP4) and on low-latency / cp recipes.
const MEGAMOE_UNSUPPORTED_RECIPES = new Set(["low-latency", "cp"]);
const MEGAMOE_UNSUPPORTED_HARDWARE = new Set(["h100", "h200", "h200-fp4"]);
const isMegamoeUnsupported = (vals) =>
  MEGAMOE_UNSUPPORTED_HARDWARE.has(vals.hardware) ||
  MEGAMOE_UNSUPPORTED_RECIPES.has(vals.recipe);

const resolveItems = (option, vals) => {
  // ... existing Marlin logic ...
  if (option.name === "megamoe" && vals && isMegamoeUnsupported(vals)) {
    const reason = MEGAMOE_UNSUPPORTED_HARDWARE.has(vals.hardware)
      ? "MegaMoE is only supported on Blackwell"
      : "MegaMoE is not supported on this recipe";
    return option.items.map((it) =>
      it.id === "disabled" ? it : { ...it, disabled: true, disabledReason: reason }
    );
  }
  return option.items;
};

const handleRadioChange = (optionName, value) => {
  setValues((prev) => {
    const next = { ...prev, [optionName]: value };
    // ... existing Marlin fallback ...
    // Switching to a hardware/recipe combo that doesn't support MegaMoE
    // while w4a8 / w4a4 is selected: fall back to disabled.
    if (
      (optionName === "hardware" || optionName === "recipe") &&
      next.megamoe !== "disabled" &&
      isMegamoeUnsupported(next)
    ) {
      next.megamoe = "disabled";
    }
    return next;
  });
};
```

评论区精华

> gemini-code-assist[bot]: flashinfer_mxfp4 标识缺少闭合反引号，导致) 和 . 被包含在代码块内。 结论：已修复。
>
> gemini-code-assist[bot]: Notes: 标题应加粗以与其他子标题格式一致。 结论：已修复。

风险与影响

该 PR 仅涉及文档和 UI 配置变更，无执行逻辑或底层代码修改，回归风险极低。用户将获得正确的文档渲染和更智能的 MegaMoE 配置体验。系统功能和性能不受影响。

关联脉络

与 PR#25406 (Mega MoE 解耦）和 PR#25412 (DSV4 cookbook 清理）密切相关，此 PR 在其基础上进一步修复文档错误并完善 MegaMoE UI 兼容性逻辑，形成完整的 MegaMoE 文档与用户体验链路。