PR #23690 完整报告

sgl-project/sglang

Small udpate gb300 recipe for deepseek v4

合并时间：2026-04-25 12:35

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23690

执行摘要

本 PR 仅更新了 DeepSeek-V4 部署文档中的 " 已验证配方 " 集合，将 gb300|small 硬件的 low-latency、balanced、max-throughput 三个配方标记为已验证，用户可以直接复制相关命令运行。变更极小（仅增加 3 行常量数据），无任何逻辑风险，建议快速合并。

功能与动机

随着 DeepSeek-V4 模型在 GB300 硬件上的验证完成，需要更新 VERIFIED_RECIPES 集合中的状态，使已验证的配方命令不再被注释掉。这样用户在查看文档时可以直接复制使用，无需手动解除注释。

实现拆解

变更文件：docs_new/src/snippets/autoregressive/deepseek-v4-deployment.jsx

1. 定位已验证集合: 在文件 VERIFIED_RECIPES 的 Set 字面量中，找到 h200|small|max-throughput 条目之后的位置。
2. 添加三项配方: 在该位置插入三个新的字符串元素:
   - "gb300|small|low-latency"
   - "gb300|small|balanced"
   - "gb300|small|max-throughput"
3. 效果: 这三条命令对应的部署脚本将从注释状态变为可直接运行的命令。未在集合中的其他配方仍保持注释状态并附带“正在验证”提示。

```
const VERIFIED_RECIPES = new Set([
    // ... 其他已验证配方 ...
    "h200|small|max-throughput",
    // 以下为本次新增：gb300|small 的低延迟、平衡、最大吞吐配方
    "gb300|small|low-latency",
    "gb300|small|balanced",
    "gb300|small|max-throughput",
    "h200|small|cp",
    // ... 其余已验证配方 ...
]);
```

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

核心变更文件：更新 VERIFIED_RECIPES 集合，增加了三个 gb300|small 的已验证配方。

```
const VERIFIED_RECIPES = new Set([
    // ... 其他已验证配方 ...
    "h200|small|max-throughput",
    // 以下为本次新增：gb300|small 的低延迟、平衡、最大吞吐配方
    "gb300|small|low-latency",
    "gb300|small|balanced",
    "gb300|small|max-throughput",
    "h200|small|cp",
    // ... 其余已验证配方 ...
]);
```

评论区精华

无 review 评论，仅包含 Mintlify 预览部署机器人自动评论和 gemini-code-assist 的配额提醒，无技术讨论。

风险与影响

- 风险: 极低。仅涉及常量数据定义，无任何运行时逻辑变更。若配方实际未完全验证，用户可能拿到未充分测试的命令，但此风险由人工验证流程控制，非代码问题。
- 影响: 正面，为 GB300 用户提供已验证的部署配方，降低使用门槛。仅影响文档展示。

关联脉络

- 历史关联: 与 PR #23684（DeepSeek-V4 环境变量文档）同属 DeepSeek-V4 文档系列更新。
- 演进方向: 随着更多硬件（如 GB300）和配方通过验证，后续会继续扩展 VERIFIED_RECIPES 集合，覆盖更多部署场景。