

PR #33251 完整报告

sgl-project/sglang

docs: drop unreachable inkling LoRA benchmark entry

合并时间: 2026-08-02 14:58

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/33251>

执行摘要

本 PR 仅删除 `docs_new/src/snippets/configs/thinkingmachines/inkling-benchmarks.jsx` 中一条不可达的 `h200 / lora / bf16 / balanced / single` LoRA 基准条目。该条目没有对应单元格，也没有携带任何数据，删除后不影响任何渲染输出，纯粹消除文档对不存在单元格的错误声明。

功能与动机

PR body 说明: `inkling.jsx` 未声明 `h200 / lora / bf16 / balanced / single` 这个单元格，而 `_deployment.jsx` 的 `findBenchmark` 与 `findCell` 使用相同的 `match` 元组定位，因此任何 picker 选择都无法解析该条目。作者还提到，本次清除是在为后续新增静态检查（校验每个 `benchmarks` 的 `match` 元组都指向真实单元格）做准备，静态检查本身不在本 PR 内。

实现拆解

- 核对单元格声明：对照 `inkling.jsx` 中实际存在的 LoRA 单元格，确认 `h200 / bf16 / balanced / single` 缺少对应单元格。
- 确认不可达性：`_deployment.jsx` 的 `findBenchmark` 与 `findCell` 使用同一组 `match` 元组，单元格不存在意味着该条目永远无法被 picker 选中。
- 删除条目：从 `inkling-benchmarks.jsx` 的 `benchmarks` 数组移除该行，净删 1 行，不涉及其他字段或数据。
- 回归验证：运行 `node docs_new/scripts/check_cookbook_configs.mjs`，前后均只有 `popular-models.jsx: no export const config` 这一既有失败，无新增问题。

`docs_new/src/snippets/configs/thinkingmachines/inkling-benchmarks.jsx`

唯一变更文件，删除一个不可达的 `h200 / lora / bf16 / balanced / single` 基准条目，消除与 `inkling.jsx` 单元格定义的不一致。

```
// LoRA 基准条目: match 元组必须对应 inkling.jsx 中真实存在的单元格,
// 否则 _deployment.jsx 的 findBenchmark 永远无法选中该条目。
// 本次变更删除了 h200 / lora / bf16 / balanced / single 这条不可达条目。
export const benchmarks = [
  // nvfp4 量化 + single 节点: b200 / b300 / gb200 / gb300 / h200 共 5 条
  { match: { hw: "b200", variant: "lora", quant: "nvfp4", strategy: "balanced", nodes: "single" } },
  { match: { hw: "b300", variant: "lora", quant: "nvfp4", strategy: "balanced", nodes: "single" } },
  { match: { hw: "gb200", variant: "lora", quant: "nvfp4", strategy: "balanced", nodes: "single" } },
  { match: { hw: "gb300", variant: "lora", quant: "nvfp4", strategy: "balanced", nodes: "single" } },
```

```
{ match: { hw: "h200", variant: "lora", quant: "nvfp4", strategy: "balanced", nodes: "single" } },  
// bf16 量化 + multi-2 节点: gb300  
{ match: { hw: "gb300", variant: "lora", quant: "bf16", strategy: "balanced", nodes: "multi-2" } },  
];
```

评论区精华

该 PR 没有任何 review 评论，[wisclmy0611](#) 直接批准。唯一 issue 评论是 Gemini Code Assist 的自动通知，说明审查服务已停止，无实质技术讨论。PR body 中作者已完整给出不可达条目的判断依据和验证结果。

风险与影响

风险极低。删除的条目无数据，且查找逻辑无法命中，因此页面渲染与 picker 行为完全不受影响。唯一遗留场景是后续若在 [inkling.jsx](#) 中新增 [h200 / bf16](#) 单元格，需要同步恢复该条目；届时规划中的静态检查可以兜底。影响范围仅限单文件、单行。

关联脉络

PR body 提到该清理源自一个计划中的静态检查，本 PR 为其扫清障碍。与仓库近期 [#33116](#)（Inkling 短卷积状态重构）同属 Inkling 支持线，但本 PR 仅涉及文档配置，无代码依赖。未来应关注该静态检查落地，以杜绝类似不可达条目再次出现。