PR #28590 完整报告

sgl-project/sglang

[Docs] DeepSeek-V4 cookbook: drop --disable-flashinfer-autotune from GB300 Flash low-latency

合并时间：2026-06-18 15:05

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/28590

执行摘要

该 PR 对 DeepSeek-V4 cookbook 文档进行了两处修正：从 GB300 Flash 低延迟服务的示例启动命令中移除已不再需要的 --disable-flashinfer-autotune 参数，并为之前为空的 benchmark 单元格补入了 v0.5.13.post1 版本的实测性能数据。变更仅涉及文档模板，无任何运行时影响。

功能与动机

根据 PR 描述，移除 --disable-flashinfer-autotune 是因为在 v0.5.13.post1 上该参数不再必要——flashinfer 的 autotune 机制已默认优化或不再干扰低延迟场景。同时，补全 GB300 Flash 低延迟的 benchmark 数据为用户提供可参考的性能指标。

实现拆解

1. 填充 benchmark 数据 - 在 deepseek-v4-benchmarks.jsx 中，为 match: { hw: "gb300", variant: "flash", quant: "fp4", strategy: "low-latency", nodes: "single" } 配置项添加了 sglang_version: "0.5.13.post1" 和两个 workload 的性能数据（concurrency=1 和 concurrency=16），包括 median TTFT、median TPOT 和 tokens/sec/GPU。
2. 删除多余启动参数 - 在 deepseek-v4.jsx 的 GB300 Flash 低延迟配置 flags 数组中，删除了 --disable-flashinfer-autotune 这一行，使命令简洁且与实际推荐配置一致。

docs_new/src/snippets/configs/deepseek-ai/deepseek-v4-benchmarks.jsx

为之前为空的 GB300 Flash 低延迟 benchmark 单元格补充了 v0.5.13.post1 的实际性能数据，包括 TTFT、TPOT 和 tokens/sec/GPU。

```
// ========================================================================
// GB300 + FP4
// ========================================================================
{
  match: { hw: "gb300", variant: "flash", quant: "fp4", strategy: "low-latency", nodes: "single" },
  sglang_version: "0.5.13.post1", // 新增：标记此数据使用的 sglang 版本
  speed: [ // 新增：填充之前为空的 benchmark 数据
    { workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 1 },
      ttft_ms: 463, tpot_ms: 4.19, tokens_per_sec_per_gpu: 35 },
    { workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 16 },
      ttft_ms: 436, tpot_ms: 8.93, tokens_per_sec_per_gpu: 336 },
  ],
```

```
},
```

docs_new/src/snippets/configs/deepseek-ai/deepseek-v4.jsx

删除 GB300 Flash 低延迟服务启动命令中不再需要的 --disable-flashinfer-autotune 参数，确保文档中的命令准确可用。

```
// GB300 Flash 低延迟配置的启动 flags 数组（已删除 --disable-flashinfer-autotune）
flags: [
  "--trust-remote-code",
  "--model-path {{MODEL_NAME}}",
  "--tp 4",
  "--moe-runner-backend flashinfer_mxfp4",
  "--speculative-algorithm EAGLE",
  "--speculative-num-steps 3",
  "--speculative-eagle-topk 1",
  "--speculative-num-draft-tokens 4",
  "--chunked-prefill-size 4096",
  // 删除了 "--disable-flashinfer-autotune" 这一项
  "--swa-full-tokens-ratio 0.1",
  "--host {{HOST_IP}}",
  "--port {{PORT}}",
],
```

评论区精华

无 review 讨论。reviewer wisclmy0611 直接批准了 PR。

风险与影响

风险：无。仅文档修改，不影响任何运行时代码。 影响：仅影响 DeepSeek-V4 cookbook 文档的读者。用户将看到正确的启动命令和完整的 benchmark 数据。

关联脉络

该 PR 与近期 DeepSeek-V4 相关的文档完善（如 PR #28613 添加压缩状态 dtype 提示）和 benchmark 基础设施改进（如 PR #28592 重构 benchmark 启动逻辑）同属持续优化 DeepSeek-V4 用户体验的系列工作。