PR #28116 完整报告

sgl-project/sglang

chore: bump tokenspeed_mla 0.1.1 -> 0.1.6

合并时间：2026-06-13 10:57

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/28116

执行摘要

- 一句话：tokenspeed_mla 版本 0.1.1 → 0.1.6
- 推荐动作：值得合并，性能提升显著且风险低。推荐关注其他使用 tokenspeed_mla 的模型（如 DeepSeek V4）的回归测试结果。

功能与动机

PR body 指出 SGLang 当前锁定 tokenspeed_mla==0.1.1，而 0.1.5 起内核引入了 FP8 P pre-scale 和重写的 softmax 规约，使 B200 上 MLA extend/prefill 注意力时间有约 12% 的优化空间。直接升级至最新 0.1.6 可以无代码改动地获取这些性能提升。

实现拆解

1. 依赖版本修改：在 python/pyproject.toml 中将 tokenspeed_mla 的 pin 从 0.1.1 改为 0.1.6，其余保持不变。
2. 验证测试：在 DeepSeek-V3-0324-FP4 模型上以 --attention-backend tokenspeed_mla 运行 GSM8K *e2e* 测试，得分 0.9550 超过 0.93 阈值；同时 test_tokenspeed_mla.py 单元测试在 B200 和 H100 上通过。
3. 基准评估：通过 torch.profiler 获取纯 GPU 内核时间，对比自注意力、前缀 chunk 和完整前缀场景，0.1.6 相对 0.1.1 有 1.13～1.15x 加速。

关键文件：

- python/pyproject.toml（模块 依赖管理；类别 config；类型 configuration）：唯一变更文件，将 tokenspeed_mla 版本从 0.1.1 提升至 0.1.6

关键符号：未识别

关键源码片段

python/pyproject.toml

唯一变更文件，将 tokenspeed_mla 版本从 0.1.1 提升至 0.1.6

```
# python/pyproject.toml
# 依赖列表中 tokenspeed_mla 版本从 0.1.1 更新为 0.1.6
# 新版内核包含 FP8 P pre-scale 和优化的 softmax 规约，可提升 MLA 注意力性能
dependencies = [
```

```
    ...
    "tokenspeed_mla==0.1.6", # 原为 0.1.1
    ...
]
```

评论区精华

无 review 评论，仅作者请求 rerun 测试并通过。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。此为纯依赖升级，后端 API 未变，且已通过准确率测试和单元测试验证。但若新版本引入其他未覆盖场景的数值差异，可能影响特定模型或配置。
- 影响：直接影响 Blackwell 系列 GPU 上使用 tokenspeed_mla 注意力后端的 MLA 模型（如 DeepSeek V3/V4），预计提升约 12% 的 extend/prefill 注意力性能。对其他硬件或后端无影响。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR