

PR #34461 完整报告

sgl-project/sglang

[ROCm] Extend the gfx950 extend-attention tile to head_dim <= 128: -43% kernel, -14% TTFT, bit-identical

合并时间: 2026-08-24 15:29

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/34461>

执行摘要

- 一句话:
- 推荐动作:

功能与动机

实现拆解

1. 测试与验证 - [test/registered/attention/test_triton_attention_kernels.py](#) (test-coverage) : 测试配套; 包含 测试覆盖调整、控制流调整、配置键调整; +12/-10
2. 配置与运行环境 - [python/sglang/kernels/ops/attention/extend_attention.py](#) (infrastructure) : 部署 / 基础设施; 包含 控制流调整; +5/-6

关键文件:

- [test/registered/attention/test_triton_attention_kernels.py](#) (模块 `triton/attention/kernels`; 类别 `test`; 类型 `test-coverage`) : 测试配套; 包含 测试覆盖调整、控制流调整、配置键调整; +12/-10
- [python/sglang/kernels/ops/attention/extend_attention.py](#) (模块 `sglang/kernels`; 类别 `infra`; 类型 `infrastructure`) : 部署 / 基础设施; 包含 控制流调整; +5/-6

关键符号: 未识别

评论区精华

- 暂无高价值评论线程

风险与影响

- 风险:
- 影响:
 - 风险标记: 暂无

关联脉络

- 暂无明显关联 PR