

PR #30101 完整报告

sgl-project/sglang

Adjust KL_THRESHOLD for log probability calculations

合并时间: 2026-07-04 14:53

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/30101>

执行摘要

将 LoRA logprob 测试中的 KL 散度阈值从 $5e-3$ 调整至 $6e-3$ ，以匹配更新后的参考 KL 值 ($5.008e-3$)，防止 CI 测试误报失败。

功能与动机

LoRA logprob 精度测试使用 KL 散度衡量 SGLang 输出与参考数据之间的差异。近期模型或环境变化导致参考 KL 值从 $4.766e-3$ 变为 $5.008e-3$ ，原有阈值 $5e-3$ 已不足以容纳新值，需上调至 $6e-3$ 。

实现拆解

- 修改文件 `test/registered/lora/test_lora_qwen3_30b_a3b_instruct_2507_logprob_diff.py` 第 49 行常量 `KL_THRESHOLD`。
- 从 $5e-3$ 改为 $6e-3$ ，并添加注释说明旧值与新 KL 值。

无。

评论区精华

无 reviewer 讨论。

风险与影响

- 风险较低：阈值上调幅度小 ($5e-3 \rightarrow 6e-3$)，实际 KL 值接近 $5e-3$ ，仍能捕获异常大的偏差。
- 影响范围仅限于此测试文件，无用户可见的功能变更。

关联脉络

无直接关联的 PR。建议确认 KL 值漂移是否由近期模型或 kernel 变更引起。