PR #21794 完整报告

sgl-project/sglang

Switch MooncakeSpec to EAGLE3 + Llama-3.1

合并时间：2026-04-01 08:12

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21794

执行摘要

本 PR 将 MooncakeSpec 测试中的模型从 Llama-2 EAGLE 切换为 Llama-3.1 EAGLE3，并调整了精度参数、阈值和并行度，以解决因 chat template wrapping 导致的评估分数下降问题，确保 CI 测试通过。

功能与动机

主要动机是解决 Llama-2 模型在 Chat API 下 GSM8K 分数下降至约 0.11 的问题，这阻碍了评估统一流程（关联 Issue #21667）。通过切换到 Llama-3.1 EAGLE3，可以恢复正常评估分数，避免 CI 阻塞。

实现拆解

所有变更集中在 test/registered/disaggregation/test_disaggregation_basic.py 文件的 TestDisaggregationMooncakeSpec 类中：

- 模型常量更新：
- 参数添加：在 spec_args 中添加 --dtype=float16，以支持 EAGLE3 + Llama-3.1。
- 阈值调整：将 test_gsm8k 中的断言从 self.assertGreater(metrics["accuracy"], 0.20) 改为 self.assertGreater(metrics["accuracy"], 0.74)。
- 并行度增加：将 parallel 参数从 2 改为 128。

评论区精华

本 PR 未收到任何 review 评论，因此无讨论内容可提炼。

风险与影响

- 风险：阈值调整可能因环境变化导致测试不稳定；新模型和参数需确保兼容性；并行度增加可能提高资源消耗。
- 影响：仅影响 CI 测试流程，对用户无直接感知，有助于团队维护测试一致性。

关联脉络

本 PR 是测试维护的一部分，与其他评估和 disaggregation 相关 PR 形成协同：

- PR #21785 添加了 CompletionSampler，支持非聊天模型评估，与本 PR 的 eval unification 目标一致。

- PR #21760 清理了 disaggregation 模块的冗余代码，反映了对该模块的持续优化。