PR #21785 完整报告

sgl-project/sglang

Add CompletionSampler for non-chat eval in run_eval

合并时间：2026-04-01 07:33

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21785

执行摘要

本 PR 新增了 CompletionSampler 类，使模型评估工具能通过 OpenAI 的 /v1/completions API 进行非聊天评估，解决了某些模型在 Chat API 下性能下降的问题，并统一了所有评估路径到 run_eval，便于 CI 回归测试。该变更属于有意义的功能改进，影响范围有限，但需注意 API 兼容性和错误处理风险。

功能与动机

根据 PR body 描述，动机源于某些 CI 测试模型（如 DeepSeek-V3 INT8 量化检查点、Llama-2-based EAGLE 模型）在 GSM8K 评估中通过 Completion API 表现良好，但在 Chat API 下因聊天模板包装得分接近零。为了将所有评估路径统一到 run_eval 以支持回归 CI，需要添加一个 Completion API 采样器。

实现拆解

实现方案包括两个关键文件修改：

- python/sglang/test/simple_eval_common.py：新增 CompletionSampler 类，使用 client.completions.create() 发送原始文本提示，代码示例：
- python/sglang/test/run_eval.py：新增 --api 参数（默认 'chat'，选项为 'chat' 或 'completion'），在 run_eval_once 函数中根据参数选择采样器。第二个提交提取了公共参数（如 model、max_tokens）以避免代码重复。

评论区精华

没有 review 讨论，因此无讨论要点。

风险与影响

- 技术风险：Completion API 可能不支持某些模型参数（如 reasoning_effort），CompletionSampler.__call__ 中对异常的捕获可能返回空字符串，影响评估准确性；测试计划未包含具体代码，需验证功能覆盖。
- 影响范围：用户（模型评估者）可更准确地评估非聊天模型；系统减少了评估脚本维护复杂度；团队通过统一路径提升了 CI 测试效率。影响程度中等，主要限于评估工具模块。

关联脉络

从近期历史 PR 分析中，本 PR 与 #21751（修复 CI 环测试超时）和 #21753（优化 CI 重测试检测）相关，因为它们都关注 CI 测试的稳定性和功能增强，反映了团队在提升评估和测试一致性方面的持续努力。