PR #15562 完整报告

sgl-project/sglang

[Feature] Add Reasoning Tokens Usage

合并时间：2026-04-04 17:18

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/15562

执行摘要

- 一句话：添加推理令牌使用统计，修复当前字段始终为 0 的问题。
- 推荐动作：该 PR 值得精读，特别是设计决策：将逻辑放在输出处理器而非服务器进程以避免重新标记化复杂性，以及如何处理推测解码场景的统一令牌 ID 格式。

功能与动机

PR body 中指出：'SGLang currently returns token usage information, but the reasoning_tokens field is always 0, which makes it unusable as a statistical metric.' 这影响了分析和监控，用户无法获取正确的推理令牌计数。

实现拆解

实现分为四个主要部分：1) 在 schedule_batch.py 的 Req 类中添加 reasoning_tokens 字段和 update_reasoning_tokens 方法，用于基于 think_end_id 检测推理阶段结束并累加令牌。2) 在 scheduler_output_processor_mixin.py 中添加 _maybe_update_reasoning_tokens 方法，在预填充和解码阶段调用更新逻辑，并统一处理推测解码的令牌 ID 格式。3) 更新 OpenAI API 端点文件（如 serving_chat.py）和 usage_processor.py，将推理令牌纳入响应统计计算。4) 新增测试文件验证功能正确性。

关键文件：

- python/sglang/srt/managers/schedule_batch.py（模块 scheduling）：添加推理令牌计数核心逻辑，包括 update_reasoning_tokens 方法和相关状态字段
- python/sglang/srt/managers/scheduler_output_processor_mixin.py（模块 scheduling）：集成推理令牌更新到调度输出处理，添加 _maybe_update_reasoning_tokens 方法并处理推测解码
- python/sglang/srt/entrypoints/openai/usage_processor.py（模块 api）：更新使用统计计算以包含推理令牌，影响所有 API 响应的令牌使用字段
- test/registered/openai_server/features/test_reasoning_usage_tokens.py（模块 test）：新增测试验证推理令牌计算正确性，覆盖正常、推测解码和流式场景

关键符号：update_reasoning_tokens, _maybe_update_reasoning_tokens, calculate_token_usage

评论区精华

Review 中核心讨论点：1) JustinTong0323 询问推测解码场景的处理，作者随后添加支持，统一令牌 ID 格式。2) CatherineSue 建议重命名 _reasoning_over 为 _is_reasoning_over，作者采纳以提升可读性。3) 关于 gRPC 中 getattr 的使用，cklxx 快速修复为直接访问属性。未解决疑虑：用户是否可以在不启用推理解析器的情况下获取推理令牌？讨论中提及但未达成结论。

- 推测解码处理 (correctness): 作者更新代码，在 process_batch_result_decode 中统一 next_token_ids 格式并调用更新方法
- 变量命名优化 (style): 作者采纳建议，重命名变量
- gRPC 属性访问设计 (design): cklxx 修改代码，移除不必要的 getattr 用法

风险与影响

- 风险：技术风险包括：1) 回归风险：核心调度文件（schedule_batch.py 和 scheduler_output_processor_mixin.py）变更可能影响其他请求处理逻辑。2) 性能风险：添加额外计数逻辑，但作者称开销小，需监控对高并发场景的影响。3) 兼容性风险：API 响应添加新字段，保持向后兼容，但客户端可能需要适配。4) 测试覆盖：新增测试覆盖正常和推测解码场景，但边缘案例（如空推理令牌列表）需进一步验证。
- 影响：影响范围：1) 用户：现在可以正确获取推理令牌统计，用于成本监控、分析和计费，提升用户体验。2) 系统：无破坏性变更，API 响应添加字段，不影响现有功能。3) 团队：需要更新相关文档以反映新字段，并确保在后续开发中一致使用。
- 风险标记：核心调度变更，推测解码兼容性，缺少文档更新

关联脉络

- PR #15875 Fix: add reasoning tokens usage: 重复 PR，本 PR cherry-pick 了其中的测试代码
- PR #14404 未知（讨论中提及）：讨论中提到的另一个重复 PR，显示该功能有多人尝试实现
- PR #21080 [Speculative Decoding] Add FA4-based Spec Support: 涉及推测解码功能，与本 PR 的推测解码处理逻辑相关