PR #38654 完整报告

vllm-project/vllm

[Bugfix] Fix `vllm bench serve` to count multimodal tokens in "total input tokens"

合并时间：2026-04-14 19:00

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38654

执行摘要

- 一句话：修复多模态模型基准测试中总输入令牌计数不包含图像令牌的问题。
- 推荐动作：该 PR 值得精读，因为它揭示了基准测试工具在多模态场景下的一个常见陷阱：客户端与服务器令牌计数的不一致。关注点包括：1. 如何通过服务器返回的 usage 字段校正客户端计算。2. review 中关于流式响应处理结构的讨论，虽然未在本 PR 解决，但值得注意。3. 修改的简洁性体现了 bugfix 的最佳实践：最小化变更，聚焦问题本身。

功能与动机

根据 PR 描述，当对多模态模型进行基准测试时，vllm bench serve 报告的 total_input_tokens 仅基于客户端纯文本提示长度，排除了图像 / 编码器令牌。虽然服务器已通过流式响应中的 usage.prompt_tokens 字段报告了正确的计数（文本 + 图像令牌），但该值未被捕获。这导致基准测试结果不准确，无法反映实际的预填充大小。PR 作者提供的测试案例显示，修复前总输入令牌为 512（仅文本），修复后为 1606（文本 + 图像令牌）。

实现拆解

实现方案分为两个关键修改点：1. 在 vllm/benchmarks/lib/endpoint_request_func.py 中，为 OpenAI completions 和 chat completions 的异步请求函数添加对 usage.prompt_tokens 的捕获逻辑，当该字段存在时将其赋值给 output.prompt_len。2. 在 vllm/benchmarks/serve.py 的 calculate_metrics 函数中，将输入令牌累计的来源从 input_requests[i].prompt_len 改为 outputs[i].prompt_len，从而使用服务器返回的实际令牌计数。

关键文件：

- vllm/benchmarks/lib/endpoint_request_func.py（模块 benchmarks）：核心修改文件，负责捕获服务器返回的 prompt_tokens 字段，是修复的关键逻辑所在。
- vllm/benchmarks/serve.py（模块 benchmarks）：修改指标计算逻辑，使用服务器返回的 prompt_len 替代客户端计算的 prompt_len，完成修复闭环。

关键符号：async_request_openai_completions, async_request_openai_chat_completions, calculate_metrics

评论区精华

review 中主要讨论了代码结构的潜在问题。gemini-code-assist[bot] 指出，当前实现使用 elif usage := data.get("usage"): 结构，假设 usage 和 choices 不会出现在同一个流式块中。虽然 vLLM 中常见，但某些 OpenAI 兼容提供商可能在最终块中同时包含两者，导致如果 choices 为真，elif 块会被跳过，prompt_len 等指标无法更新。建议独立检查 usage 以确保跨后端的鲁棒性。作者 mgehre-amd 回应称这是预先存在的问题，与本 PR 无关。最终 PR 被合并，未对建议的结构修改进行进一步调整。

- 流式响应中 usage 字段捕获的代码结构问题 (correctness): 作者回应这是预先存在的问题，与本 PR 无关，未进行修改。

风险与影响

- 风险：技术风险较低：1. 回归风险：修改仅影响基准测试的指标计算逻辑，不改变核心推理路径，但需确保 output.prompt_len 在所有情况下都被正确设置，否则可能导致指标计算错误。2. 兼容性风险：依赖服务器返回的 usage.prompt_tokens 字段，如果某些后端不提供该字段，prompt_len 可能保持为 None，导致计算时可能出错（但原逻辑也存在类似问题）。3. 代码结构风险：如 review 所指，elif 结构可能在某些提供商场景下错过 usage 数据，但这被标记为预先存在的问题。
- 影响：影响范围：1. 对用户：多模态模型基准测试用户将获得更准确的输入令牌计数，有助于性能评估和容量规划。2. 对系统：仅影响基准测试工具的输出指标，不改变推理引擎或 API 行为。3. 对团队：修复了一个长期存在的指标偏差问题，提升了基准测试的可靠性。影响程度：中等，因为它修正了关键性能指标的准确性，但仅限于基准测试场景。
- 风险标记：依赖外部字段，预先存在的结构问题

关联脉络

- PR #39473 fix: handle ImportError in load_audio: 同属多模态相关 bugfix，涉及音频处理，但本 PR 聚焦基准测试指标。
- PR #38061 [MM][Perf][CG] Support ViT full CUDA graph for Qwen3-VL video inference: 同属多模态性能优化，本 PR 的测试案例使用了 Qwen 多模态模型，关联性强。
- PR #39753 [Model] Use mm_features for Ernie-4.5 VL M-RoPE: 同属多模态模型相关改进，涉及图像特征处理，与本 PR 的多模态令牌计数主题相关。