PR #44383 完整报告

vllm-project/vllm

[Bugfix] Fix --enable-prompt-tokens-details omitting zero cached tokens

合并时间：2026-06-12 11:37

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/44383

执行摘要

- 一句话：修复 prompt_tokens_details 遗漏零缓存令牌
- 推荐动作：建议阅读本 PR，尤其是 Python 真值陷阱导致的行为偏差及其修复模式。对于维护 API 正确性的团队，这是值得关注的案例：一个空格的变化（and num_cached_tokens → and num_cached_tokens is not None）就能影响整个数据管道。disagg 路径的逐步补全也说明了全面的代码审查的重要性。

功能与动机

在启用 --enable-prompt-tokens-details 时，由于条件使用 and num_cached_tokens 的 truthiness 检查，cached_tokens=0 被当作 False 而省略 prompt_tokens_details 字段，导致客户端无法区分缓存缺失（None）和缓存命中（0）。关联 issue #44377 详细描述了该问题。

实现拆解

1. 识别问题：在 vllm/entrypoints/openai/chat_completion/serving.py 的流式生成器 chat_completion_stream_generator 和非流式生成器 chat_completion_full_generator 中，条件 and num_cached_tokens 和 and final_res.num_cached_tokens 将 int 0 视为 falsy。
2. 修改 chat 路径：将上述两处条件替换为 num_cached_tokens is not None 和 final_res.num_cached_tokens is not None。
3. 修改 completion 路径：在 vllm/entrypoints/openai/completion/serving.py 的 completion_stream_generator 和 request_output_to_completion_response 中做相同替换。
4. 修改 disagg 路径：在 vllm/entrypoints/serve/disagg/serving.py 的 serve_tokens_full_generator 和 serve_tokens_stream_generator 中做相同替换（该文件最初被遗漏，后补全）。
5. 添加回归测试：在 tests/entrypoints/serve/disagg/test_generate_stream.py 中新增 test_stream_prompt_tokens_details_zero_cached，验证 cached_tokens=0 被正确包含。
6. 更新现有测试：tests/entrypoints/openai/completion/test_completion.py 中的 test_single_completion 原断言 prompt_tokens_details=None，现更新为期望 cached_tokens=0，因为测试服务器默认启用 --enable-prompt-tokens-details 且禁用

prefix caching。

关键文件：

- tests/entrypoints/serve/disagg/test_generate_stream.py（模块 流式生成；类别 test；类型 test-coverage；符号 test_stream_prompt_tokens_details_zero_cached, mock_generate）：新增回归测试，验证零缓存令牌不被省略，覆盖流式与非流式场景。
- vllm/entrypoints/openai/chat_completion/serving.py（模块 聊天完成；类别 source；类型 core-logic；符号 chat_completion_stream_generator, chat_completion_full_generator）：核心聊天接口（流式与非流式），修复了 2 处 truthiness 检查，是主要修复文件。
- vllm/entrypoints/serve/disagg/serving.py（模块 分离生成；类别 source；类型 core-logic；符号 serve_tokens_full_generator, serve_tokens_stream_generator）：分离部署路径的 2 处修复，最初被遗漏后补全，体现了全面覆盖的重要性。
- vllm/entrypoints/openai/completion/serving.py（模块 完成；类别 source；类型 core-logic；符号 completion_stream_generator, request_output_to_completion_response）：完成接口的流式与非流式回复，统一修复以确保整体一致性。
- tests/entrypoints/openai/completion/test_completion.py（模块 完成测试；类别 test；类型 test-coverage）：更新现有测试断言以匹配新行为，确保 CI 通过。

关键符号：test_stream_prompt_tokens_details_zero_cached, mock_generate, chat_completion_stream_generator, chat_completion_full_generator, completion_stream_generator, request_output_to_completion_response, serve_tokens_full_generator, serve_tokens_stream_generator

关键源码片段

tests/entrypoints/serve/disagg/test_generate_stream.py

新增回归测试，验证零缓存令牌不被省略，覆盖流式与非流式场景。

```
@pytest.mark.asyncio
async def test_stream_prompt_tokens_details_zero_cached():
    """enable_prompt_tokens_details includes cached_tokens=0 in final usage.

    Regression test for https://github.com/vllm-project/vllm/issues/44377:
    zero cached tokens must not be treated as falsy and omitted.
    """
    engine = _mock_engine()

    async def mock_generate(*args, **kwargs):
        # 模拟生成器返回 num_cached_tokens=0 的场景
        yield _make_request_output(
            "req-1",
            token_ids=[10],
            finish_reason="stop",
            finished=True,
```

```
            num_cached_tokens=0, # 显式设为 0，触发目标条件
        )

    engine.generate = MagicMock(side_effect=mock_generate)
    serving = _build_serving_tokens(engine, enable_prompt_tokens_details=True)

    request = GenerateRequest(
        token_ids=[1, 2, 3],
        sampling_params=SamplingParams(max_tokens=10),
        model=MODEL_NAME,
        stream=True,
        stream_options=StreamOptions(include_usage=True),
    )

    response = await serving.serve_tokens(request)
    chunks = []
    async for chunk in response:
        chunks.append(chunk)

    parsed = _parse_sse_chunks(chunks)
    # Usage-only chunk (before [DONE])
    usage_chunk = parsed[-2]
    assert usage_chunk["choices"] == []
    # Zero cached tokens must be present, not omitted
    assert usage_chunk["usage"]["prompt_tokens_details"] is not None
    assert usage_chunk["usage"]["prompt_tokens_details"]["cached_tokens"] == 0
```

vllm/entrypoints/openai/chat_completion/serving.py

核心聊天接口（流式与非流式），修复了 2 处 truthiness 检查，是主要修复文件。

```
# 流式生成器 (line 884)
if self.enable_prompt_tokens_details and num_cached_tokens is not None:
    # 使用 is not None 而非隐式真值检查，确保 0 也能被包含
    final_usage.prompt_tokens_details = PromptTokenUsageInfo(
        cached_tokens=num_cached_tokens
    )

# 非流式生成器 (line 1308)
if (
    self.enable_prompt_tokens_details
    and final_res.num_cached_tokens is not None
):
    usage.prompt_tokens_details = PromptTokenUsageInfo(
        cached_tokens=final_res.num_cached_tokens
    )
```

vllm/entrypoints/serve/disagg/serving.py

分离部署路径的 2 处修复，最初被遗漏后补全，体现了全面覆盖的重要性。

```
# serve_tokens_full_generator (line 310)
if (
    self.enable_prompt_tokens_details
    and final_res.num_cached_tokens is not None
):
    # 注意：/coordinator 级别不提供此信息
    usage.prompt_tokens_details = PromptTokenUsageInfo(
        cached_tokens=final_res.num_cached_tokens
    )

# serve_tokens_stream_generator (line 427)
if self.enable_prompt_tokens_details and num_cached_tokens is not None:
    final_usage_info.prompt_tokens_details = PromptTokenUsageInfo(
        cached_tokens=num_cached_tokens
    )
```

评论区精华

> @chaunceyjiang 要求添加端到端测试（E2E test case）。 @sasindharan 回复已在 test_generate_stream.py 中添加回归测试，并指出 disagg 路径的 2 个位置也在测试中被覆盖。 @chaunceyjiang 最后批准（LGTM）。 讨论未出现重大分歧，作者通过单元测试弥补了 E2E 测试的不足。

- E2E 测试覆盖范围 (testing): reviewer 接受并批准 PR。

风险与影响

- 风险：风险较低。变更仅将 Python 隐式 truthiness 检查替换为显式 is not None 检查，对所有非 None 值的语义不变。主要风险在于遗漏其他类似条件，但已通过系统梳理 3 个文件共 6 处位置确保覆盖。disagg 路径在初始提交中被遗漏，后经补充说明已修复。潜在风险是在未来增加新服务路径时可能忽略同样模式，但代码审查时容易识别。无性能影响。
- 影响：对用户：先前当 num_cached_tokens=0 时，API 响应中 prompt_tokens_details 字段缺失或为 null，现在会正确返回 {"cached_tokens": 0}。这对依赖精确缓存计数的客户端（如计费系统、缓存策略）至关重要。对系统：仅条件判断变化，无额外开销。影响范围涉及 OpenAI 兼容的聊天 / 补全接口以及分离式部署（disagg）路径，但均属同一修复模式。
- 风险标记：Python truthiness 陷阱，初始遗漏分离路径，多入口条件需同步

关联脉络

- 暂无明显关联 PR