

PR #52098 完整报告

vllm-project/vllm

[Frontend] Log output token IDs at DEBUG level

合并时间: 2026-08-13 21:11

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/52098>

执行摘要

- 一句话: 输出 token IDs 日志移至 DEBUG 级别, INFO 仅保留文本
- 推荐动作: 值得快速阅读: 其设计决策 (复用既有 `log_inputs` 的级别拆分模式而非新增 CLI 标志) 体现了最小化配置面的思路, 对同类日志需求有参考价值; 测试更新到位, 新增了 DEBUG 开关的边界测试, 是一个低风险的小特性。

功能与动机

Issue #51912 提出在生产环境中希望保留可读生成文本和 `finish_reason`, 同时省略 `output_token_ids`, 因为长 token ID 列表增加日志量且可被解码回文本带来隐私顾虑。维护者 DarkLight1337 在 review 中明确建议: 不如像 `log_inputs` 一样, 让 output token IDs 只在 DEBUG 级别记录, 而不是新增布尔开关。

实现拆解

1. 核心逻辑改造: 在 `vllm/entrypoints/serve/utils/request_logger.py` 的 `RequestLogger.log_outputs` 中, 将 `output_token_ids` 的截断与记录从原 INFO 日志调用中拆出, 放入新增的 `logger.debug` 分支; 并先用 `logger.isEnabledFor(logging.DEBUG)` 判断, 避免在未启用 DEBUG 时执行不必要的 list 转换。文本截断逻辑保持不变, 仍在 INFO 路径执行。
2. 日志内容拆分: INFO 日志消息改为 "Generated response %s%s: output: %r, finish_reason: %s", 不再包含 `output_token_ids`; DEBUG 日志消息为 "Generated response %s%s details: output_token_ids: %s", 流式 delta/complete 的 `stream_info` 标记在两个级别中保持一致。
3. CLI 帮助同步: 在 `vllm/entrypoints/openai/cli_args.py` 中更新 `enable_log_outputs` 的 docstring, 说明输出文本和 finish reason 记录在 INFO, 输出 token IDs 记录在 DEBUG, 删除原先“信息最多只在 INFO 级别”的旧描述。
4. 测试配套: 在 `tests/entrypoints/serve/utils/test_request_logger.py` 中调整原有断言 (INFO 参数不再包含 token IDs), 并新增 `test_request_logger_log_output_token_ids_require_debug`, 验证当 logger 未启用 DEBUG 时 debug 不会被调用、截断行为依然正确。

关键文件:

- `vllm/entrypoints/serve/utils/request_logger.py` (模块 请求日志; 类别 source; 类型 core-logic; 符号 `log_outputs`): 核心实现文件, `log_outputs` 方法完成了 INFO/DEBUG

的日志级别拆分，是本次变更的主体。

- tests/entrypoints/serve/utils/test_request_logger.py (模块 日志测试; 类别 test; 类型 test-coverage; 符号 test_request_logger_log_output_token_ids_require_debug) : 测试配套, 更新原有断言并新增 DEBUG 级别边界测试, 验证 INFO/DEBUG 分离行为。
- vllm/entrypoints/openai/cli_args.py (模块 CLI 参数; 类别 source; 类型 documentation) : 同步更新 enable_log_outputs 的帮助文本, 反映 INFO/DEBUG 分级, 避免文档与实现不一致。

关键符号: log_outputs

关键源码片段

vllm/entrypoints/serve/utils/request_logger.py

核心实现文件, `log_outputs` 方法完成了 INFO/DEBUG 的日志级别拆分, 是本次变更的主体。

```
def log_outputs(
    self,
    request_id: str,
    outputs: str,
    output_token_ids: Sequence[int] | None,
    finish_reason: str | None = None,
    is_streaming: bool = False,
    delta: bool = False,
) -> None:
    # 先对输出文本按 max_log_len 截断, 这部分仍属于 INFO 级别的人类可读内容
    max_log_len = self.max_log_len
    if max_log_len is not None and outputs is not None:
        outputs = outputs[:max_log_len]

    stream_info = ""
    if is_streaming:
        # 区分 streaming delta 与 streaming complete, 方便定位日志来源
        stream_info = " (streaming delta)" if delta else " (streaming complete)"

    # 只有当日志级别允许 DEBUG 时才拼装并发送 token IDs,
    # 避免在 INFO 场景下产生不必要的 list 转换开销
    if logger.isEnabledFor(logging.DEBUG):
        if max_log_len is not None and output_token_ids is not None:
            # token IDs 是序列, 先转 list 再截断, 保持与文本截断一致
            output_token_ids = list(output_token_ids)[:max_log_len]
        logger.debug(
            "Generated response %s%s details: output_token_ids: %s",
            request_id,
            stream_info,
            output_token_ids,
        )

    # INFO 级别只保留文本与 finish_reason, 满足排障同时控制日志量
```

```
logger.info(
    "Generated response %s%s: output: %r, finish_reason: %s",
    request_id,
    stream_info,
    outputs,
    finish_reason,
)
```

评论区精华

核心讨论围绕设计方向：DarkLight1337 提出“I think it would be better to handle this like `log_inputs`, i.e., the output tokens should only be logged at DEBUG level”，否定了作者初始新增 CLI 标志的方案。作者采纳并回复：“addressed in 58eeb18. I removed the new CLI flag and changed output logging to follow the existing request-input split”。该决策避免了新增配置面，保持了日志行为的统一语义。

- 输出 token IDs 应仅记录在 DEBUG 级别 (design): 作者采纳建议，移除新 CLI 标志，改为将 `output_token_ids` 输出到 DEBUG 日志，INFO 仅保留文本与 `finish_reason`。

风险与影响

- 风险：1) 日志行为变更：依赖 INFO 日志中 `output_token_ids` 的运维脚本或监控系统在升级后会缺失该字段，需要改为读取 DEBUG 日志或调整日志配置；2) DEBUG 级别下 `output_token_ids` 可能为 None，此时日志会打印 None，解析方需容错；3) 测试使用 mock logger，若真实 logger 的 `isEnabledFor` 行为与预期不一致（如自定义 filter），可能出现 DEBUG 日志未按预期输出；4) 影响面集中在 `--enable-log-outputs` 启用的服务端，默认配置 (INFO) 无变化，整体风险较低。
- 影响：影响所有启用 `--enable-log-outputs` 的 vLLM 服务的请求日志输出：INFO 日志不再包含 token IDs，DEBUG 日志新增 token IDs。对默认配置 (INFO) 无影响，但对依赖 INFO 日志中 token IDs 的运维工具属于行为变更。改动范围小，集中在入口层请求日志模块，不涉及模型推理路径，对性能和功能性无影响。
- 风险标记：日志行为变更，DEBUG 级别依赖，默认配置无影响

关联脉络

- 暂无明显关联 PR