PR #45468 完整报告

vllm-project/vllm

[Bugfix] Reject structured outputs for diffusion decoders with a clear error

合并时间：2026-06-14 03:04

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/45468

执行摘要

- 一句话：拒绝扩散模型的结构化输出请求
- 推荐动作：值得精读，尤其是理解扩散模型与结构化输出不兼容的根因。设计决策（在 _validate_structured_outputs 中提前拦截）简洁合理，可作为类似跨模块兼容性守卫的参考。

功能与动机

用户在 issue #45436 中报告，对 DiffusionGemma 模型使用 response_format 或 structured_outputs 参数时，服务器返回 HTTP 500 内部错误，日志显示 FSM 拒绝 token 序列。根因是 Diffusion LLM 的并行去噪过程与结构化输出的自回归 FSM 不兼容。PR 作者在 issue 评论中指出，即使生成合法 JSON 也会因非停止 token 被拒绝。

实现拆解

1. 修改 _validate_structured_outputs 方法：在 vllm/sampling_params.py 中，为该方法的签名新增 model_config 参数，方法体开头增加 if model_config.is_diffusion: 检查，当检测到扩散模型且提供了结构化输出约束时，直接抛 ValueError。
2. 更新 verify 方法调用：在 SamplingParams.verify() 中，将 _validate_structured_outputs 的调用签名由 (structured_outputs_config, tokenizer) 改为 (model_config, structured_outputs_config, tokenizer)，确保 model_config 传入。
3. 新增测试文件：在 tests/v1/structured_output/test_validation.py 中新增两个测试用例：test_structured_outputs_rejected_for_diffusion_models 验证结构化输出请求应被拒绝并包含特定错误消息；test_plain_request_allowed_for_diffusion_models 验证无结构化输出的普通请求不受影响。测试使用 _StubModelConfig 模拟扩散模型配置。

关键文件：

- vllm/sampling_params.py（模块 采样参数；类别 source；类型 core-logic；符号 SamplingParams._validate_structured_outputs, SamplingParams.verify）：核心修改文件：在 _validate_structured_outputs 方法中新增扩散模型守卫，并修改 verify 方法调用签名，以传递 model_config。
- tests/v1/structured_output/test_validation.py（模块 验证层；类别 test；类型 test-coverage；符号 _StubModelConfig, init, test_structured_outputs_rejected_for_diffusion_models, test_plain_request_allowed_for_diffusion_models）：新增测试文件，包

含两个测试用例，覆盖扩散模型结构化输出被拒绝和普通请求不受影响的场景。

关键符号：SamplingParams._validate_structured_outputs, SamplingParams.verify

关键源码片段

vllm/sampling_params.py

核心修改文件：在 _validate_structured_outputs 方法中新增扩散模型守卫，并修改 verify 方法调用签名，以传递 model_config。

```
# vllm/sampling_params.py (head)

def _validate_structured_outputs(
    self,
    model_config: ModelConfig,
    structured_outputs_config: StructuredOutputsConfig | None,
    tokenizer: TokenizerLike | None,
) -> None:
    if structured_outputs_config is None or self.structured_outputs is None:
        return

    # 扩散语言模型并行去噪整块画布，而非从左到右逐 token 采样，
    # 这与语法 FSM 的要求不兼容。没有此检查，请求会在生成中途
    # 因 FSM 拒绝而失败（HTTP 500）。参见 issue #45436。
    if model_config.is_diffusion:
        raise ValueError(
            "Structured outputs are not yet supported for diffusion "
            "language models. Remove the structured output constraint "
            "(e.g. `response_format`, `structured_outputs`) from the "
            "request."
        )

    # 原有 tokenizer 检查保持不变
    if tokenizer is None:
        raise ValueError(
            "Structured outputs requires a tokenizer so it can't be used with 'skip_tokenizer_init'"
        )
    # 后续 backend 兼容性检查 ...
```

tests/v1/structured_output/test_validation.py

新增测试文件，包含两个测试用例，覆盖扩散模型结构化输出被拒绝和普通请求不受影响的场景。

```
# tests/v1/structured_output/test_validation.py (new file)

import pytest

from vllm.config import StructuredOutputsConfig
from vllm.sampling_params import SamplingParams, StructuredOutputsParams
```

```
pytestmark = pytest.mark.cpu_test

JSON_SCHEMA = {
    "type": "object",
    "properties": {
        "invoice_id": {"type": "string"},
        "customer": {"type": "string"},
    },
    "required": ["invoice_id", "customer"],
    "additionalProperties": False,
}

# 用于测试的 stub ModelConfig，仅暴露 is_diffusion 属性
class _StubModelConfig:
    def __init__(self, is_diffusion: bool):
        self.is_diffusion = is_diffusion


def test_structured_outputs_rejected_for_diffusion_models():
    """扩散模型结构化输出请求应被拒绝，并返回包含
    "not yet supported for diffusion" 的 ValueError。"""
    params = SamplingParams(
        structured_outputs=StructuredOutputsParams(json=JSON_SCHEMA)
    )
    with pytest.raises(ValueError, match="not yet supported for diffusion"):
        params._validate_structured_outputs(
            _StubModelConfig(is_diffusion=True),
            StructuredOutputsConfig(),
            tokenizer=None,
        )


def test_plain_request_allowed_for_diffusion_models():
    """不带结构化输出的普通请求不应被拦截。"""
    params = SamplingParams()
    params._validate_structured_outputs(
        _StubModelConfig(is_diffusion=True),
        StructuredOutputsConfig(),
        tokenizer=None,
    )
```

评论区精华

讨论较少，核心思路在 issue 中已明确。reviewer mgoin 在 Slack 上提及了相关话题。独立验证者 stonehall-simon 在 RTX PRO 6000 上使用原始镜像确认了改动生效：结构化请求返回 HTTP 400，普通请求保持 HTTP 200。

- 独立验证结果 (testing): 改动在目标硬件上验证通过。

风险与影响

- 风险：风险极低。变更仅在校验阶段增加提前返回，对非扩散模型无任何逻辑影响。测试覆盖了正向和负向场景。唯一需要注意的是，若未来扩散模型开始支持结构化输出，需要在此处移除该守卫。
- 影响：影响范围限于请求校验阶段，对系统性能无影响。用户侧影响：使用扩散模型的结构化输出请求将从不明原因的 HTTP 500 变为清晰的 HTTP 400 错误，用户体验提升。不影响其他模型或非结构化输出场景。
- 风险标记：兼容性守卫，测试覆盖完善

关联脉络

- PR #45163 [Model] Add DiffusionGemma model support: 引入了扩散模型，是导致该 bug 的直接原因。