PR #46733 完整报告

vllm-project/vllm

[Bugfix][Rust Frontend] Reject min_tokens above max_tokens

合并时间：2026-06-26 11:41

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/46733

执行摘要

该 PR 针对 Rust 前端多个入口（HTTP completions、generate、gRPC）未一致拒绝 min_tokens > max_tokens 无效请求的漏洞，在共享的文本降层函数中添加验证，并正确映射错误状态码。同时提取 is_request_validation_error 方法简化错误分类。改动涉及 6 个文件，新增 153 行，删除 28 行，包含充分的测试覆盖。

功能与动机

Python 端 SamplingParams 已认为 min_tokens > max_tokens 为非法组合，但 Rust 前端仅在 chat completions 路由有检查，其他路径如 /v1/completions 和 /generate 可能传递错误配置到引擎，导致 max_tokens=4 但 min_tokens=5 的不可能要求。PR body 明确指出："adds the validation to the shared Rust text lowering layer so all paths using it reject the invalid combination consistently."

实现拆解

1. 新增错误变体：在 rust/src/text/src/error.rs 中添加 Error::MinTokensExceedsMaxTokens，并实现 is_request_validation_error() 方法统一判断请求验证错误。
2. 共享降层检查：在 rust/src/text/src/lower.rs 的 lower_sampling_params 函数中，在解析 max_tokens 和 min_tokens 后插入 if min_tokens > max_tokens 检查。该函数被所有文本请求路径共享，一次添加覆盖全部入口。
3. 简化服务器错误路由：在 rust/src/server/src/error.rs 中删除自由函数 is_request_validation_error，改为直接调用 error.is_request_validation_error()，并简化 chat_submit_error 的模式匹配。
4. 修复 gRPC 错误映射：在 rust/src/server/src/grpc/mod.rs 中新增 text_error_to_status 函数，将验证错误映射为 tonic::Code::InvalidArgument，替换原来所有文本错误都硬编码为 Internal 的逻辑。
5. 聊天错误委托：在 rust/src/chat/src/error.rs 中为 vllm_chat::Error 添加 is_request_validation_error() 方法，代理到内部文本错误。
6. 测试配套：单元测试 lower_sampling_params_rejects_min_tokens_above_resolved_max_tokens 验证降层检查；集成测试 unary_generate_min_tokens_above_max_tokens_returns_invalid_argument 和 streaming_generate_min_tokens_above_max_tokens_returns_invalid_argument 验证 gRPC unary/streaming 路径均返回 InvalidArgument。

rust/src/text/src/lower.rs

共享文本降层的核心函数 lower_sampling_params，添加了 min_tokens > max_tokens 的检查，所有路径都经过此函数。

```
// lower_sampling_params 核心部分：解析 max_tokens 后检查 min_tokens
// 前面的参数 fallback 省略
let max_tokens = resolve_max_tokens(
    max_tokens,
    default_max_tokens,
    sampling_limits.max_model_len,
    prompt_len,
)?;
let min_tokens = min_tokens.unwrap_or(0);
// [ 关键添加 ] 拒绝 min_tokens 超过 max_tokens 的请求
if min_tokens > max_tokens {
    return Err(Error::MinTokensExceedsMaxTokens {
        min_tokens,
        max_tokens,
    });
}
// 后续参数处理（thinking_token_budget, frequency_penalty 等）
```

rust/src/text/src/error.rs

新增 Error::MinTokensExceedsMaxTokens 变体，实现 is_request_validation_error 方法，为所有错误分类提供统一接口。

```
#[error(
    "`min_tokens` must be less than or equal to `max_tokens`, \
     got min_tokens={min_tokens}, max_tokens={max_tokens}"
)]
MinTokensExceedsMaxTokens { min_tokens: u32, max_tokens: u32 },

impl Error {
    /// Whether this error represents invalid user request parameters.
    pub fn is_request_validation_error(&self) -> bool {
        match self {
            Self::PromptTooLong { .. }
            | Self::EmptyPromptTokenIds { .. }
            | Self::Logprobs(_)
            | Self::TokenIds(_)
            | Self::MinTokensExceedsMaxTokens { .. }
            | Self::InvalidThinkingTokenBudget
            // 空 prompt 可能通过 Llm wrapper 传递
            | Self::Llm(LlmError::EmptyPromptTokenIds { .. }) => true,
            _ => false,
        }
    }
}
```

评论区精华

> Codex bot: "gRPC 处理器当前用 Status::internal 包装验证错误，导致客户端验证失败误报为内部错误，应映射为 InvalidArgument。" — 作者随后添加 text_error_to_status 函数修复此问题。
>
> BugenZhao: " 这个函数 (is_text_request_validation_error) 与 HTTP 路由中的重复，最好将它降为 vllm_text::Error 的方法。" — 在第三个 commit 中由 BugenZhao 自己实现提取。

风险与影响

- 风险：错误映射从 Internal 改为 InvalidArgument 可能影响依赖 500 状态码的客户端，但这是正确行为；测试覆盖充分，回归风险低。
- 影响：用户得到更准确的 400 错误；系统验证逻辑统一；团队新增验证错误只需在 is_request_validation_error 中添加匹配臂，维护成本低。

关联脉络

该 PR 是 Rust 前端持续改进的一部分，与近期 #46696（TLS 实现切换）、#46719（测试夹具提取）、#46602（解析器统一）等 PR 共同推动 Rust 模块的健壮性和代码质量。