

# PR #27019 完整报告

sgl-project/sglang

feat(api): add require\_reasoning field for engine's generate api

合并时间: 2026-06-03 01:36

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/27019>

## 执行摘要

- 一句话: Engine generate API 新增 `require_reasoning` 参数
- 推荐动作: 此 PR 是简单的参数透传, 阅读价值有限, 但可作为 API 扩展模式的参考。建议后续补充单元测试和文档。

## 功能与动机

根据 PR body, `require_reasoning` 是控制推理 schema 和追踪推理 token 的重要字段, 但未在 engine 的 generate API 中暴露, 而该 API 被 dynamo 等其他生态系统使用, 因此需要补充。

## 实现拆解

1. 在 `python/sglang/srt/entrypoints/engine.py` 的 `generate` 方法参数列表中添加 `require_reasoning: bool = False` 参数。
2. 在 `generate` 方法内部构造 `GenerateReqInput` 对象时传入 `require_reasoning=require_reasoning`。
3. 类似地在 `async_generate` 方法参数列表中添加同一参数并传入 `GenerateReqInput` 构造调用。此 PR 不涉及测试、配置或部署配套变更。

关键文件:

- `python/sglang/srt/entrypoints/engine.py` (模块引擎; 类别 source; 类型 core-logic; 符号 `generate`, `async_generate`) : 唯一变更文件, 在 `generate` 和 `async_generate` 方法中添加 `require_reasoning` 参数并传递至 `GenerateReqInput`。

关键符号: `generate`, `async_generate`

## 关键源码片段

`python/sglang/srt/entrypoints/engine.py`

唯一变更文件, 在 `generate` 和 `async_generate` 方法中添加 `require_reasoning` 参数并传递至 `GenerateReqInput`。

```
# python/sglang/srt/entrypoints/engine.py (部分)
# 在 generate 方法签名中添加 require_reasoning 参数
class Engine:
```

```

def generate(
    self,
    # ... 其他参数 ...
    custom_logit_processor: Optional[Union[List[str], str]] = None,
    require_reasoning: bool = False, # [ 新增 ] 控制推理 schema 和追踪推理 token
    return_hidden_states: bool = False,
    # ...
) -> Union[Dict, Iterator[Dict]]:
    # ...
    obj = GenerateReqInput(
        # ...
        custom_logit_processor=custom_logit_processor,
        require_reasoning=require_reasoning, # [ 新增 ] 传递给 GenerateReqInput
        return_hidden_states=return_hidden_states,
        # ...
    )
    # ...

# 同步修改 async_generate 方法, 保持行为一致
async def async_generate(
    self,
    # ...
    custom_logit_processor: Optional[Union[List[str], str]] = None,
    require_reasoning: bool = False, # [ 新增 ] 与 generate 保持一致
    return_hidden_states: bool = False,
    # ...
) -> Union[Dict, Iterator[Dict]]:
    # ...
    obj = GenerateReqInput(
        # ...
        custom_logit_processor=custom_logit_processor,
        require_reasoning=require_reasoning, # [ 新增 ]
        return_hidden_states=return_hidden_states,
        # ...
    )
    # ...

```

## 评论区精华

Review 评论极少，只有 JustinTong0323 的批准，没有实质性讨论或争议。

- 暂无高价值评论线程

## 风险与影响

- 风险：风险极低。变更仅添加了一个带默认值 False 的可选参数，不会影响现有调用行为；但缺少测试覆盖和文档更新，未来若 require\_reasoning 逻辑变更可能存在回归风险。
- 影响：对用户：提供了控制推理行为的新能力，兼容旧调用（默认 False）。对系统：无运行时性能影响，仅为参数透传。对团队：涉及接口扩展，需注意与其他 API 的同步更新。

- 风险标记: 缺少测试覆盖, 缺少文档更新

## 关联脉络

- 暂无明显关联 PR