PR #42116 完整报告

vllm-project/vllm

[Frontend] add support for thinking_token_budget in completions

合并时间：2026-05-14 04:01

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42116

执行摘要

- 一句话：为 Completions API 添加 thinking_token_budget 参数
- 推荐动作：该 PR 变更简单清晰，值得快速合入。建议后续为 thinking_token_budget 补充范围校验和测试。有权限的维护者已批准合入，无需过多关注。

功能与动机

根据 PR body 描述，目的是在 completions API 中支持传入 thinking_token_budget 参数（该参数在 chat completions 中已实现）。这是对已有功能的补齐，使得 completion 请求也能像 chat 请求一样控制推理模型的思考 token 预算，对使用 reasoning model（如 Cohere 等）的用户是有意义的。

实现拆解

1. 在 CompletionRequest 模型中新增字段：在 vllm/entrypoints/openai/completion/protocol.py 的 CompletionRequest 类中，于 repetition_detection 字段之后、# --8<-- [end:completion-extra-params] 标记之前，新增一个类型为 int | None、默认值为 None 的 thinking_token_budget 字段，并附上说明："推理模型思考操作的最大 token 数，-1 表示无限制"。
2. 在 to_sampling_params 方法中传递该字段：在同一个文件的 to_sampling_params 方法中，向 SamplingParams.from_optional 调用增加关键字参数 thinking_token_budget=self.thinking_token_budget。这样，当 completion 请求转换为采样参数时，该值会被带入后续推理流程。
3. 无测试、配置或文档配套改动：本次变更未添加单元测试、集成测试或文档更新。

关键文件：

- vllm/entrypoints/openai/completion/protocol.py（模块 前端 API；类别 source；类型 core-logic）：唯一变更文件，新增 thinking_token_budget 字段定义并传递到 SamplingParams

关键符号：未识别

关键源码片段

vllm/entrypoints/openai/completion/protocol.py

唯一变更文件，新增 thinking_token_budget 字段定义并传递到 SamplingParams

```
# vllm/entrypoints/openai/completion/protocol.py ( 片段 )

class CompletionRequest(OpenAIBaseModel):
    # ... 其他字段省略 ...

    repetition_detection: RepetitionDetectionParams | None = Field(
        default=None,
        description="...",
    )

    # 新增字段：允许用户在 completion 请求中设置 thinking token budget
    thinking_token_budget: int | None = Field(
        default=None,
        description=(
            "Maximum number of tokens allowed for thinking operations "
            "(reasoning models). -1 = unlimited."
        ),
    )

    # --8<-- [end:completion-extra-params]

    def to_sampling_params(self, ...) -> SamplingParams:
        # ... 其他参数准备逻辑 ...
        return SamplingParams.from_optional(
            n=self.n,
            # ... 其他已有参数 ...
            extra_args=extra_args or None,
            skip_clone=True,
            repetition_detection=self.repetition_detection,
            thinking_token_budget=self.thinking_token_budget, # 新增：透传 thinking token 预算
        )
```

评论区精华

唯一有价值的讨论来自自动代码审查机器人 gemini-code-assist[bot]，它建议对新字段添加 ge=-1 和 le=_INT64_MAX 的取值约束，以与其他整数参数（如 seed、truncate_prompt_tokens）保持一致。PR 作者 walterbm 询问维护者 aarnphm 是否需要此验证，后者回复“I don't think this matter. the gemini-bot is very noisy as of late”，因此未采纳该建议。整个 review 过程简单，aarnphm 直接批准。

- thinking_token_budget 字段是否需要添加范围校验 (correctness): 未添加范围校验，保持与 chat completions 一致的做法。

风险与影响

- 风险:

1. 缺少年度整型范围校验：新字段未添加 ge=-1 或 le 限制，如果传入极端值可能会导致下游 SamplingParams 处理异常，但考虑到 chat completions 中同样没有此类硬校验，影响可控。
2. 未覆盖测试：没有对应的单元测试或集成测试，今后如果 to_sampling_params 的内部逻辑或签名发生变化，该字段容易被遗漏。
3. 兼容性：新增可选字段，对现有请求无影响，向后兼容良好。 - 影响：影响范围：仅影响使用 completion API 且需要控制推理模型思考 token 预算的用户。影响程度：低。这是一个小的功能补齐，不涉及现有行为变更；但为后续支持 Cohere 等 reasoning model 在 completion 端的使用扫清了障碍。团队影响：降低维护者回答 thinking_token_budget 在 completion 中为何不支持的问题的频率。

- 风险标记：缺少测试覆盖，缺少输入校验

关联脉络

- 暂无明显关联 PR