PR #23951 完整报告

sgl-project/sglang

fix(openai): map reasoning.enabled to thinking AND enable_thinking

合并时间：2026-05-08 05:01

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23951

执行摘要

- 一句话：统一 reasoning.enabled 映射两种 chat_template key
- 推荐动作：建议合并。该 PR 是来自 #22254 的 9 行独立 bugfix，无外部依赖，解决模型间 thinking key 不一致问题。可后续跟进修复 review 指出的 enabled 与 effort='none' 同时设置时的优先级问题。

功能与动机

PR body 指出，此前 reasoning.enabled=true 仅在 chat_template_kwargs 中设置 thinking=True，该 key 仅被 deepseek-v3/kimi_k2 识别，而 qwen3/glm45/nemotron_3/interns1/mimo 检查的是 enable_thinking，导致功能对后者静默失效。

实现拆解

1. 在 python/sglang/srt/entrypoints/openai/protocol.py 的 normalize_reasoning_inputs 方法中，当 reasoning.enabled 为 true 时，除原有 ctk.setdefault("thinking", True) 之外，增加一行 ctk.setdefault("enable_thinking", True)，并添加注释说明不同模型检查不同键。
2. 在 test/registered/unit/entrypoints/openai/test_protocol.py 的 test_chat_completion_reasoning_effort 测试中，将预期结果从 {"thinking": True} 更新为 {"thinking": True, "enable_thinking": True}，以覆盖双 key 均被设置的情况。

关键文件：

- python/sglang/srt/entrypoints/openai/protocol.py（模块 协议解析；类别 source；类型 core-logic；符号 normalize_reasoning_inputs）：核心逻辑变更：在 normalize_reasoning_inputs 中为 enable_thinking 添加 setdefault 映射。
- test/registered/unit/entrypoints/openai/test_protocol.py（模块 协议测试；类别 test；类型 test-coverage；符号 test_chat_completion_reasoning_effort）：单元测试同步更新，验证双 key 均被设置。

关键符号：normalize_reasoning_inputs, test_chat_completion_reasoning_effort

关键源码片段

python/sglang/srt/entrypoints/openai/protocol.py

核心逻辑变更：在 normalize_reasoning_inputs 中为 enable_thinking 添加 setdefault 映射。

```
# 位于 python/sglang/srt/entrypoints/openai/protocol.py，normalize_reasoning_inputs 方法中
if enabled:
    ctk = values.get("chat_template_kwargs")
    if not isinstance(ctk, dict):
        ctk = {}
    # 不同模型检查不同的 chat template key:
    # - "thinking" 用于 deepseek-v3, kimi_k2
    # - "enable_thinking" 用于 qwen3, glm45, nemotron_3, interns1, mimo
    ctk.setdefault("thinking", True)
    ctk.setdefault("enable_thinking", True) # 新增行，覆盖更多模型
    values["chat_template_kwargs"] = ctk
```

test/registered/unit/entrypoints/openai/test_protocol.py

单元测试同步更新，验证双 key 均被设置。

```
# 位于 test/registered/unit/entrypoints/openai/test_protocol.py
# test_chat_completion_reasoning_effort 方法中
request = ChatCompletionRequest(
    model="test-model",
    messages=messages,
    reasoning={
        "enabled": True,
        "reasoning_effort": "high",
    },
)
self.assertEqual(request.reasoning_effort, "high")
# 验证两个 thinking key 均被设置
self.assertEqual(
    request.chat_template_kwargs,
    {"thinking": True, "enable_thinking": True},
)
```

评论区精华

Review 中 gemini-code-assist[bot] 指出一个高严重度逻辑漏洞：当 reasoning.enabled=true 且 reasoning_effort='none' 时，setdefault 会使 thinking 和 enable_thinking 保持 True，而 'none' 预期应禁用思考。但该 PR 未修复此问题，仅做映射补充。该问题在 reasoning_effort='none' 分支中已使用 setdefault 将两 key 设为 False，但由于 enabled 分支先执行，若两分支先后触发，『none』分支的 setdefault 不会覆盖已存在的 True 值。不过实际场景中同一请求通常不会同时设置 enabled=true 且 effort='none'，故问题影响有限。

- reasoning.enabled 与 reasoning_effort='none' 同时设置时 setdefault 逻辑缺陷 (correctness): 未解决。该 PR 专注于映射修复，未调整优先级逻辑。实际场景中同时设置 enabled=true 和 effort='none' 的可能性较低，但仍属潜在 bug。

风险与影响

- 风险：低风险。PR 仅新增一行 setdefault 并更新测试断言，逻辑简单。但遗留的『enabled 与 effort='none' 同时设置时的优先级』问题在极端场景下可能导致用户期望 disable 但实际 enable。此外，当用户显式传入 chat_template_kwargs 时，setdefault 保留用户指定值，符合预期。
- 影响：对用户：修复使用 reasoning.enabled=true 时 Qwen3/GLM45 等模型无思考输出的问题。对系统：变更范围极小（2 文件，9 行），无性能影响。对团队：统一不同模型家族的 thinking 映射方式，降低未来新增模型的认知成本。
- 风险标记：缺少边界值覆盖

关联脉络

- PR #22254 未提供具体标题：本 PR 从中拆分出的 9 行独立 bugfix。