

PR #28153 完整报告

sgl-project/sglang

fix(sampling): reject non-finite temperature in SamplingParams.verify

合并时间: 2026-06-14 14:41

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/28153>

执行摘要

- 一句话: 修复非有限温度导致服务拒绝访问漏洞
- 推荐动作: 值得精读, 因为它展示了一个容易被忽视的安全漏洞 (IEEE-754 特性导致单侧比较失效) 以及简单的修复方法。代码改动极小但影响安全, 适合作为输入验证的最佳实践示例。

功能与动机

PR body 指出, 一个带有 NaN 或 +/-Inf 的 temperature 参数会导致整个服务器崩溃, 这是一个请求级拒绝服务漏洞。由于 temperature 是通过 JSON 公开的参数, 任何客户端都可以触发该漏洞。

实现拆解

1. 修改验证逻辑 (python/sglang/srt/sampling/sampling_params.py) : 在 SamplingParams.verify() 方法中, 将原有的单侧比较 `if self.temperature < 0.0` 改为 `if not math.isfinite(self.temperature) or self.temperature < 0.0`, 并相应地更新错误消息。
2. 更新导入 (python/sglang/srt/sampling/sampling_params.py) : 新增 `import math`。
3. 添加回归测试 (test/registered/unit/sampling/test_sampling_params.py) : 在 TestSamplingParamsVerify 类中添加两个测试用例 `test_nan_temperature_raises` 和 `test_inf_temperature_raises`, 确保 NaN 和 Inf 温度被正确拒绝。

关键文件:

- python/sglang/srt/sampling/sampling_params.py (模块 采样参数; 类别 source; 类型 core-logic; 符号 verify) : 包含核心验证逻辑的修改, 通过增加 `isfinite` 检查来拒绝非有限温度值。
- test/registered/unit/sampling/test_sampling_params.py (模块 采样参数; 类别 test; 类型 test-coverage; 符号 `test_nan_temperature_raises`, `test_inf_temperature_raises`) : 添加了两个验证非有限温度值的回归测试用例, 确保修复被正确覆盖。

关键符号: `verify`

关键源码片段

[python/sglang/srt/sampling/sampling_params.py](#)

包含核心验证逻辑的修改，通过增加 `isfinite` 检查来拒绝非有限温度值。

```
# python/sglang/srt/sampling/sampling_params.py
import math # 新增导入

class SamplingParams:
    # ... __init__ 代码保持不变 ...

    def verify(self, vocab_size):
        # 修复：原条件 self.temperature < 0.0 无法捕获 NaN/Inf（因为 nan < 0 为 False），
        # 新增 math.isfinite() 检查，确保温度是有限非负值。
        if not math.isfinite(self.temperature) or self.temperature < 0.0:
            raise ValueError(
                f"temperature must be a non-negative finite number, got {self.temperature}."
            )
        # 以下其他字段的验证保持不变
        if not 0.0 < self.top_p <= 1.0:
            raise ValueError(f"top_p must be in (0, 1], got {self.top_p}.")
        # ...
```

test/registered/unit/sampling/test_sampling_params.py

添加了两个验证非有限温度值的回归测试用例，确保修复被正确覆盖。

```
# test/registered/unit/sampling/test_sampling_params.py
class TestSamplingParamsVerify(CustomTestCase):
    VOCAB_SIZE = 32000

    def _make(self, **kwargs):
        defaults = dict(temperature=1.0, top_p=1.0, top_k=10, min_p=0.0)
        defaults.update(kwargs)
        return SamplingParams(**defaults)

    def test_negative_temperature_raises(self):
        sp = self._make(temperature=-0.5)
        with self.assertRaises(ValueError):
            sp.verify(self.VOCAB_SIZE)

    # 新增：验证 NaN 温度被拒绝
    def test_nan_temperature_raises(self):
        """verify() must reject NaN temperature; the bare < 0.0 check alone lets it through."""
        sp = self._make(temperature=float("nan"))
        with self.assertRaises(ValueError):
            sp.verify(self.VOCAB_SIZE)

    # 新增：验证 Inf 温度被拒绝
    def test_inf_temperature_raises(self):
        """verify() must reject non-finite (inf) temperature."""
        sp = self._make(temperature=float("inf"))
        with self.assertRaises(ValueError):
            sp.verify(self.VOCAB_SIZE)
```

评论区精华

作者在 issue 评论中补充了与 vLLM 的对比分析，指出 vLLM 同样存在该验证漏洞，但由于其采样器内部处理了非有限概率值而未崩溃。这进一步强调了在验证层防御的重要性。无其他讨论。

- 暂无高价值评论线程

风险与影响

- 风险：变更极小且仅在验证逻辑中，无回归风险。错误消息更新不影响兼容性。添加的测试覆盖了两个典型非有限情况，确保验证生效。
- 影响：修复了请求级拒绝服务漏洞，防止恶意用户通过发送非有限温度值导致整个 SGLang 服务器崩溃。影响范围包括 `/generate`、`/v1/chat/completions`、`/v1/completions` 等所有接受 `temperature` 参数的端点。对正常用户无影响。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR