PR #27247 完整报告

sgl-project/sglang

[AMD] Guard aiter greedy_sample OOB token id (fixes VLM MMMU CI)

合并时间：2026-06-05 03:53

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27247

执行摘要

- 一句话：修复 AMD CI 上 aiter 贪婪采样越界 token 问题
- 推荐动作：此 PR 值得快速合并。变更简洁、目标明确、风险可控。建议合并后跟踪 aiter 上游修复进度，待修复后移除该环境和相关测试变通。同时，建议根据 hubertlu-tw 的建议增强 test_aiter_greedy_sample_amd.py 测试覆盖，防止类似回归再次发生。

功能与动机

在 AMD CI 上，test_vlm_mmmu_benchmark (MiniCPM-V-2.6, MMMU) 会间歇性因 lmms-eval 'IndexError: list index out of range' 失败。根本原因是 aiter greedy_sample 内核对于全 NaN 或全 -inf 的 logits 行返回了与 vocab_size 相等的越界 token id，导致下游解码为空字符串、OpenAI 兼容服务器返回 content=None、eval 收集结果列表为空，最终触发 IndexError。该问题影响约 8% 的 MMMU 输入，而 torch.argmax 处理相同输入时始终返回合法（尽管可能错误）的索引，因此评估得以继续。

实现拆解

1. 配置防护开关：在 sampler.py 中新增模块级常量 _disable_aiter_greedy_sample，通过 get_bool_env_var('SGLANG_DISABLE_AITER_GREEDY_SAMPLE') 读取环境变量，并添加详细注释说明 aiter 内核的 bug 及回退含义。
2. 修改贪婪采样逻辑：将 forward 方法中贪婪分支的条件从 if _use_aiter: 改为 if _use_aiter and not _disable_aiter_greedy_sample:，当开关启用时，即使 _use_aiter 为真，也回退到 torch.argmax 路径，从而避免触发 aiter 内核的越界行为。
3. 测试环境注入：在 test_vlm_models.py 中新增 is_in_amd_ci 导入，在 test_vlm_mmmu_benchmark 方法内部，若在 AMD CI 中则构造 custom_env 字典（{'SGLANG_DISABLE_AITER_GREEDY_SAMPLE': '1'}）并传给 _run_vlm_mmmu_test，使该环境变量仅在 AMD CI 的 MMMU 测试生效，不影响其他场景。

关键文件：

- python/sglang/srt/layers/sampler.py（模块 采样器；类别 source；类型 core-logic；符号 _disable_aiter_greedy_sample, Sampler.forward）：核心变更：添加环境变量开关并修改贪婪采样分支，是修复逻辑的所在地。
- test/registered/models/test_vlm_models.py（模块 VLM 测试；类别 test；类型 test-coverage；符号 TestVLMModels.test_vlm_mmmu_benchmark）：测试变更：在

AMD CI 的 MMMU 测试中设置防护环境变量，确保评测通过。

关键符号：_disable_aiter_greedy_sample, Sampler.forward, TestVLMModels.test_vlm_mmmu_benchmark

关键源码片段

python/sglang/srt/layers/sampler.py

核心变更：添加环境变量开关并修改贪婪采样分支，是修复逻辑的所在地。

```
# 在文件顶部（第 49-54 行，附近），新增防护开关常量
# The aiter greedy_sample kernel can return an out-of-range token id (== vocab_size,
# e.g. 151666 for MiniCPM-V) for all-NaN / all -inf logit rows on ROCm, which decodes
# to an empty string and breaks downstream consumers. Set this to 1 to fall back to
# torch.argmax (which always returns a valid index). Default off so behavior is
# unchanged elsewhere.
_disable_aiter_greedy_sample = get_bool_env_var("SGLANG_DISABLE_AITER_GREEDY_SAMPLE")
# ...
# 在 forward 方法的贪婪采样分支中（第 119-126 行）
if sampling_info.is_all_greedy:
    # 只有当 aiter 可用且未禁用时才使用 aiter 内核
    if _use_aiter and not _disable_aiter_greedy_sample:
        batch_next_token_ids = torch.empty(
            logits.shape[0], device=logits.device, dtype=torch.int32
        )
        _aiter_greedy_sample(batch_next_token_ids, logits)
    else:
        # 回退到 torch.argmax，始终返回 [0, vocab_size-1] 范围内的合法索引
        batch_next_token_ids = torch.argmax(logits, -1)
```

test/registered/models/test_vlm_models.py

测试变更：在 AMD CI 的 MMMU 测试中设置防护环境变量，确保评测通过。

```
# 更新导入：新增 is_in_amd_ci
from sglang.test.test_utils import is_in_amd_ci, is_in_ci
# ...
# 在 test_vlm_mmmu_benchmark 方法中（第 39-51 行）
for model in models_to_test:
    with tempfile.TemporaryDirectory(...) as temp_dir:
        # On AMD CI, the aiter greedy_sample kernel returns an out-of-range
        # token id (== vocab_size) for degenerate (all-NaN / all -inf) logit
        # rows, producing empty completions that crash the MMMU eval. Disable
        # it there so greedy sampling falls back to torch.argmax.
        custom_env = None
        if is_in_amd_ci():
            custom_env = {"SGLANG_DISABLE_AITER_GREEDY_SAMPLE": "1"}
        self._run_vlm_mmmu_test(model, temp_dir, custom_env=custom_env)
```

评论区精华

复审者 HaiShaw 在批准前要求作者在 aiter 仓库提交 issue 以跟进根本修复（'please open an aiter issue as follow up'）。此外，hubertlu-tw 指出已有 test/registered/ops/test_aiter_greedy_sample_amd.py 测试脚本，建议增强该测试覆盖范围，以便在 aiter 上游修复前即可发现此类回归。

- 在 aiter 仓库提交上游 issue 跟踪根本修复 (other): 作者应创建 aiter issue 并跟进上游修复。
- 增强现有 aiter greedy_sample 测试覆盖 (testing): 建议后续加强测试，但当前 PR 未做修改。

风险与影响

- 风险：该变更本身风险极低：
  - 默认行为不变（环境变量未设置时功能等同原代码）。
  - 仅在 AMD CI 的 VLM MMMU 测试中显式设置该变量，不影响其他硬件平台、其他采样策略或生产部署。
  - 若未来 aiter 内核修复了越界 bug，移除该开关也不会影响正确性，届时只需清理环境和测试代码即可。
  - 潜在风险：若 _preprocess_logits 未能过滤所有 NaN/inf（例如由于自定义 logit 处理器），回退到 torch.argmax 会输出一个虽然合法但语义错误的 token，可能掩盖模型本身的数值稳定性问题。
  - 影响：对用户：无直接影响；环境变量默认关闭，普通用户不需要感知。对 AMD CI：MMMU 评测将从间歇性失败变为稳定通过，提升 CI 可靠性。对团队：新增了一个运行时防护开关，但语义明确（仅覆盖 aiter 贪婪采样路径），且仅用于 AMD 平台，维护成本低。
  - 风险标记：环境变量防护，仅 AMD CI 测试变通，上游依赖未修复

关联脉络

- PR #26383 引入 aiter greedy_sample 内核用于 AMD 贪婪解码：当前 PR 修复了 #26383 引入的 aiter greedy_sample 内核中的越界 bug。