

PR #41335 完整报告

vllm-project/vllm

[ROCm][CI] Align spec decode logprob test prefill settings

合并时间: 2026-05-04 12:33

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/41335>

执行摘要

- 一句话: 对齐 ROCm 投机解码测试预填设置
- 推荐动作: 建议合并。这是一个低风险、高收益的测试稳定性修复, 解决了 ROCm CI 中的间歇性失败。值得关注的是其设计模式——使用共享配置字典确保两个实验组配置一致, 可用于其他比较性测试。

功能与动机

ROCm 上的浮点归约对批处理几何和调度敏感, 参考 LLM (默认调度) 与投机解码 LLM (分块预填) 使用不同的归约顺序, 会产生数值偏差 (观察到的差异约 0.235), 超出测试容差, 导致测试误报失败。PR body 明确说明此变更意在隔离投机解码正确性。

实现拆解

1. 提取共享配置: 在 `test_spec_decode_logprobs` 函数中, 将参考 LLM 和投机解码 LLM 共用的参数 (`max_logprobs`、`seed`、`logprobs_mode` 等) 提取为字典 `llm_kwargs`, 并统一设置 `enable_chunked_prefill=True` 和 `max_num_batched_tokens=32`。
2. 统一两 LLM 实例: 参考 LLM 和投机解码 LLM 均通过 `**llm_kwargs` 接收参数, 确保两者使用完全相同的调度和批处理配置。
3. 更新注释: 调整 `ROCM_DETERMINISM_KWARGS` 上方的注释, 使其描述更精确, 强调调度差异是数值偏差的根源。
4. 删除冗余参数: 投机解码 LLM 构造时移除重复的单独参数 (如 `enable_chunked_prefill=True` 等), 由 `llm_kwargs` 统一提供。

关键文件:

- `tests/v1/sample/test_logprobs.py` (模块 `logprobs` 测试; 类别 `test`; 类型 `test-coverage`; 符号 `test_spec_decode_logprobs`): 唯一的变更文件, 通过提取共享配置 `llm_kwargs` 并统一预填设置修复 ROCm 数值偏差问题。

关键符号: `test_spec_decode_logprobs`

关键源码片段

[tests/v1/sample/test_logprobs.py](#)

唯一的变更文件, 通过提取共享配置 `llm_kwargs` 并统一预填设置修复 ROCm 数值偏差问题。

```
# 在 test_spec_decode_logprobs 函数中，原来 ref_llm 使用默认调度，
# spec_llm 强制 enable_chunked_prefill=True & max_num_batched_tokens=32。
# 现在统一为 llm_kwargs，两者使用相同配置，消除数值噪声：
```

```
max_model_len = 256
llm_kwargs = dict(
    max_logprobs=5,
    max_model_len=max_model_len,
    seed=42,
    logprobs_mode=logprobs_mode,
    gpu_memory_utilization=0.4,
    # 强制相同的分块预填设置，使比较仅反映投机解码差异
    enable_chunked_prefill=True,
    max_num_batched_tokens=32,
    enable_prefix_caching=False,
    **ROCM_DETERMINISM_KWARGS,
)

# 参考 LLM
ref_llm = LLM(
    model=model_name,
    **llm_kwargs,
)

# 投机解码 LLM
spec_llm = LLM(
    model_name,
    speculative_config=spec_config_with_len,
    **llm_kwargs,
)
```

评论区精华

Bortlesboat 指出：之前的设置隐含地测试了“调度切换下 logprobs 的稳定性”，但这不是测试的目的。通过强制两实例使用相同预填配置，测试现在只隔离投机解码正确性。共享

`llm_kwargs` 结构也杜绝了未来两配置漂移的可能。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更仅修改一个测试函数内的配置提取方式，不涉及生产代码。但需注意：强制预填设置可能降低测试对默认调度场景的覆盖，但测试本身的目标是验证投机解码正确性，而非调度稳定性。
- 影响：影响限于 ROCm 平台上的 `test_spec_decode_logprobs` 测试的稳定性。变更后该测试不再因数值噪声误报失败，能更准确地反映投机解码的正确性。对其他平台（如 NVIDIA）无影响，因为 `ROCM_DETERMINISM_KWARGS` 条件为空字典。
- 风险标记：测试覆盖变更

关联脉络

- PR #33493 （假设的引用 PR） ROCm 确定性修复：PR 注释中提及 TODO 引用该 PR 的后续修复，用于最终移除 VLLM_ROCM_USE_SKINNY_GEMM 的环境变量工作区。