PR #44234 完整报告

vllm-project/vllm

[Test][BugFix] Fix double-BOS in PD+specdec acceptance test

合并时间：2026-06-02 05:31

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/44234

执行摘要

- 一句话：修复 PD+SD 测试中重复 BOS 问题
- 推荐动作：建议合并。修复虽小但提升了测试质量，防止未来因 token 不一致导致的误判。

功能与动机

作者在调查 MRV2 测试失败时发现，PD+specdec acceptance 测试使用的 prompt 已包含 BOS，但 /completions API 默认会再添加一个 BOS，导致双重 BOS，使 acceptance rate 比预期低约 5%。基线是在 add_special_tokens=False 条件下生成的，但测试未对齐该设置。虽然之前测试大多能通过（因为阈值较宽松），但该不一致可能掩盖回归或导致后续偶发失败。

实现拆解

1. 在 tests/v1/kv_connector/nixl_integration/test_spec_decode_acceptance.py 中的 test_spec_decode_acceptance_length 函数内，对 /completions API 的调用添加 extra_body 参数，设置 add_special_tokens 为 False。
2. 这样，API 不会在已包含 BOS 的 prompt 前再添加额外 BOS，使得测试输入与基线生成时的 tokenization 一致。
3. 变更仅涉及 4 行添加（注释 + 参数），无其他文件修改，影响范围小且明确。
4. 该修复使测试的 acceptance length 更接近真实值，避免因 tokenization 差异导致的误报或漏报。

关键文件：

- tests/v1/kv_connector/nixl_integration/test_spec_decode_acceptance.py（模块 测试脚本；类别 test；类型 test-coverage；符号 test_spec_decode_acceptance_length）：唯一修改的文件，修复了测试中因重复 BOS 导致 acceptance length 偏低的问题。

关键符号：test_spec_decode_acceptance_length

关键源码片段

tests/v1/kv_connector/nixl_integration/test_spec_decode_acceptance.py

唯一修改的文件，修复了测试中因重复 BOS 导致 acceptance length 偏低的问题。

```
def test_spec_decode_acceptance_length():
    """..."""
```

```
# ... 先前代码不变 ...
for i, prompt in enumerate(prompts):
    resp = client.completions.create(
        model=MODEL_NAME,
        prompt=prompt,
        max_tokens=DEFAULT_OUTPUT_LEN,
        temperature=0.0,
        top_p=1.0,
        # 这里修复：prompt 已经 chat-templated（包含 BOS），
        # 设置 add_special_tokens=False 防止 API 再添加一个 BOS，
        # 避免与基线（add_special_tokens=False）产生 tokenization 差异。
        extra_body={"add_special_tokens": False},
    )
    # ... 后续不变 ...
```

评论区精华

仅有 1 位 reviewer 批准（yewentao256），无讨论线程。LGTM。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。仅修改测试请求参数，不影响任何生产代码或推理逻辑。若某些模型或 API 版本不支持 add_special_tokens 参数，该测试可能失败，但该参数是 OpenAI 兼容 API 的标准字段，且 vLLM 支持此参数。
- 影响：影响范围仅限该测试文件。修复后，测试的 acceptance length 比较会更准确，提升 CI 的可靠性，尤其是在回归检测中减少假阴性。对用户无影响。
- 风险标记：仅测试变更，低风险

关联脉络

- PR #41294 [ROCm][CI] Fix and stabilize EAGLE3 acceptance tests: 同为 spec decode acceptance 测试的修复，稳定 CI。
- PR #44078 [MRV2] Remove Eagle's dedicated CUDA graph pool: 作者提及该修复是在调查 MRV2 测试失败时发现的，MRV2 相关改动可能因测试不稳定而受影响。