

PR #45913 完整报告

vllm-project/vllm

[Bugfix][test] Use Salesforce/wikitext for ppl tests

合并时间: 2026-06-17 18:37

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/45913>

执行摘要

- 一句话: 修复 PPL 测试数据集 ID 兼容性问题
- 推荐动作: 建议合并。这是针对上游库 API 变动的必要兼容性修复, 改动极小且已验证正确性。值得关注的是这类依赖升级引发的测试失败模式, 团队应建立定期检查测试依赖兼容性的机制。

功能与动机

`datasets >= 5.0` 和 `huggingface_hub >= 1.19` 不再接受遗留单段 ID `'wikitext'` (`HfUriError: 'Repository id must be namespace/name'`)。该 ID 已迁移至 `'Salesforce/wikitext'`。若不修复, 所有 `tests/models/language/generation_ppl_test/` 下的 PPL 测试都会在 `load_dataset()` 阶段失败, 无法计算任何 PPL。

实现拆解

对 `tests/models/language/generation_ppl_test/ppl_utils.py` 文件进行单行修改: 将 `load_dataset("wikitext", "wikitext-2-raw-v1", split="test")` 修改为 `load_dataset("Salesforce/wikitext", "wikitext-2-raw-v1", split="test")`。这是唯一需要修改的地方, 因为该文件中所有 PPL 测试共享这个 `load_dataset` 调用。提交者使用 Qwen 系列模型验证了 PPL 差异均小于 1% (最大 `+0.764%`), 确认变更不影响测试的正确性。

关键文件:

- `tests/models/language/generation_ppl_test/ppl_utils.py` (模块 测试工具; 类别 `test`; 类型 `test-coverage`): 唯一变更文件, 修改了 `load_dataset` 的数据集 ID 以兼容新版本 `datasets` 和 `huggingface_hub`。

关键符号: `wikitext_ppl_test`

关键源码片段

`tests/models/language/generation_ppl_test/ppl_utils.py`

唯一变更文件, 修改了 `load_dataset` 的数据集 ID 以兼容新版本 `datasets` 和 `huggingface_hub`。

```
# 文件 : tests/models/language/generation_ppl_test/ppl_utils.py
# 关键变更 : 更新数据集 ID 以兼容 datasets >= 5.0 的新要求
```

```

...
@torch.inference_mode
def wikitext_ppl_test(
    hf_runner,
    vllm_runner,
    model_info: GenerateModelInfo,
    max_length=MAX_LENGTH,
    vllm_extra_kwargs=None,
    atol=PPL_TOL,
):
    vllm_extra_kwargs = get_vllm_extra_kwargs(model_info, vllm_extra_kwargs)

    # 修复：使用完整的命名空间 ID 'Salesforce/wikitext'
    # 旧方式 'wikitext' 在 datasets >= 5.0 和 huggingface_hub >= 1.19 下会抛出 HfUriError
    dataset = load_dataset("Salesforce/wikitext", "wikitext-2-raw-v1", split="test")

    with vllm_runner(
        model_info.name,
        gpu_memory_utilization=0.7,
        max_model_len=max_length,
        max_num_seqs=1,
        **vllm_extra_kwargs,
    ) as vllm_model:
        ...

```

评论区精华

无。所有审核评论均为流程性内容（DCO 签署），无技术讨论。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更仅为更新数据集路径字符串，不涉及逻辑改动。PPL 验证结果表明数值差异极小（属于随机噪声范围），没有回归风险。潜在风险是 'Salesforce/wikitext' 在未来的库版本中再次变更，但这是外部依赖的正常演进，无法在本 PR 中预防。
- 影响：影响范围仅限于 tests/models/language/generation_ppl_test/ 下的 PPL 测试，确保持续集成中该测试套件不被上游库升级阻断。对用户和生产系统无影响。影响程度低。
- 风险标记：外部依赖变更

关联脉络

- PR #45180 fix(security): enforce audio decode duration limit in chat completions path: 同一兼容性系列（transformers 5.10），修复 fetch_images 相关依赖兼容性。
- PR #44559 [Bugfix][Gemma4] Pre-initialise streaming reasoning state when prompt ends inside an open <lchannel> (fixes #45834): 同一兼容性系列（transformers 5.10），修复 fetch_audio 相关依赖兼容性。