

PR #27235 完整报告

sgl-project/sglang

refactor: replace oversized 1.3MB tiny_tokenizer.json fixture with a genuinely tiny byte-level BPE fixture

合并时间: 2026-06-09 08:40

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/27235>

执行摘要

将 `experimental/sgl-router/tests/fixtures/tiny_tokenizer.json` 从 1.3 MB 的完整 GPT-2 BPE tokenizer 替换为几 KB 的真正小型 byte-level BPE tokenizer，同时更新了一个相关测试断言。所有 11 个引用该 fixture 的测试仍保持绿色。仓库体积减少约 1.3 MB，不涉及任何生产代码变更。

功能与动机

Issue #27219 指出 `tiny_tokenizer.json` 实际包含完整的 GPT-2 BPE tokenizer (~50k 词汇表 + 完整 merges 表)，大小达到 1.3 MB，与文件名中的 "tiny" 严重不符。该文件作为 fixture 被 11 个测试用例引用用于构建 `TokenizerRegistry`，因此不能直接删除。需要替换为一个真正小巧的 tokenizer，同时保持所有测试通过。

实现拆解

1. 替换测试 fixture 文件 `experimental/sgl-router/tests/fixtures/tiny_tokenizer.json`: 重新生成一个真正小的 byte-level BPE tokenizer，仅包含 256 个 byte token (0-255) 加上一个 `<lendoftextl>` token，merges 表为空。新文件从 1.3 MB 缩减至几 KB，同时由于是 byte-level，对 ASCII/bytes 无损编码，确保了 `encode/decode` round-trip 断言继续有效。
2. 更新 Rust 测试断言 `experimental/sgl-router/src/tokenizer/mod.rs` 中的 `decode_complete_returns_string_on_partial_utf8` 测试: 原先使用了 GPT-2 特有的 token ID `[47249, 222]` 来表示 emoji 🍌。在新的 byte-level tokenizer 中，emojis 编码为原始 UTF-8 字节序列 `[240, 159, 152, 128]`。因此更新了断言值及其注释，同时注意到传递给 `decode_complete` 的切片也从 `&full[..1]` (之前第一个 token 即覆盖了部分 UTF-8 前缀) 改为 `&full[..3]` (现在需要 3 个 token 才能覆盖同样的 3 个字节前缀)，以继续触发 `DecodeResult::Partial` 分支。

`experimental/sgl-router/tests/fixtures/tiny_tokenizer.json`

被替换的核心测试 fixture，从 1.3MB 缩减至几 KB，同时保持了对 ASCII/bytes 无损的 tokenization 能力。

```
// 新的 tiny_tokenizer.json 核心结构: 极小 byte-level BPE
{
  "version": "1.0",
  "added_tokens": [{ "id": 256, "special": true, "content": "<lendoftextl>" }],
  "pre_tokenizer": { "type": "ByteLevel" },
```

```

"post_processor": {"type": "ByteLevel"},
"decoder": {"type": "ByteLevel"},
"model": {
  "vocab": {
    "\u0100": 0, "\u0101": 1, ..., "\u00ff": 255, // 256 个 byte tokens
    "<lendoftextl>": 256
  },
  "merges": [] // 空 merges 表, 仅依赖 byte-level 编码
}
}

```

experimental/sgl-router/src/tokenizer/mod.rs

唯一的源码改动文件，更新了 test `decode_complete_returns_string_on_partial_utf8` 中的断言，以适配新 tokenizer 的字节级特性。

```

// 片段来自 tiny_tokenizer.json 替换后，用于测试 partial UTF-8 decode 分支的代码调整
#[test]
fn decode_complete_returns_string_on_partial_utf8() {
  let r = TokenizerRegistry::load_from_config(&cfg()).unwrap();
  let t = r.get("tiny").unwrap();

  // 新 tokenizer 是 no-merge byte-level BPE: emoji 🌀 被编码为 4 个原始字节 token [240, 159, 152, 128]
  let full = adapter::encode(&t, "🌀").unwrap();
  assert_eq!(
    full,
    vec![240, 159, 152, 128],
    "fixture tokenisation drift: '🌀' no longer encodes to [240, 159, 152, 128]"
  );

  // 只解码前 3 个字节（一个 4 字节 UTF-8 码点的前缀），触发 partial UTF-8 分支
  let s = adapter::decode_complete(&t, &full[..3], false).unwrap();
  assert_eq!(s, "\u{FFFD}"); // 预期替换字符
}

```

评论区精华

PR 本身没有产生 review 讨论（0 review comments），只有一个 reviewer 直接批准。合并者在合并后对作者表示感谢。

风险与影响

- 风险：极低。仅修改测试 fixture 和单一测试断言，不涉及推理路径或生产代码。风险在于可能破坏 11 个引用该 fixture 的测试，但全部通过，且新 fixture 对 ASCII/bytes 无损，满足所有 round-trip 和 count > 0 断言。
- 影响：仓库体积减少约 1.3 MB，对团队透明，对用户无影响。所有测试继续运行正常。

关联脉络

该PR直接对应issue#27219 (Oversizedtestfixture)，是测试基础设施清理工作的一部分。
近期历史 PR 中未见其他直接关联的 fixture 改动。