

PR #41661 完整报告

vllm-project/vllm

[Benchmark] Add --trust-remote-code flag to multi-turn benchmark

合并时间: 2026-05-05 16:00

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/41661>

执行摘要

- 一句话: 新增多轮基准测试的 --trust-remote-code 标志
- 推荐动作: 此 PR 改动简单直接, 可作为基础设施类变更的示例。如需了解 vLLM 基准测试工具链, 可精读该文件了解更多标志。

功能与动机

多轮基准测试脚本无法加载 HuggingFace Hub 上带有自定义 tokenizer 代码的模型 (如 Qwen、DeepSeek 等), 因为 `AutoTokenizer.from_pretrained` 默认 `trust_remote_code=False`。添加该标志可与 vLLM 其他工具的行为保持一致。

实现拆解

1. 在 `benchmarks/multi_turn/benchmark_serving_multi_turn.py` 的命令行参数解析器中新增 --trust-remote-code 标志 (store_true 类型)。
2. 修改 `AutoTokenizer.from_pretrained` 调用, 传入 `trust_remote_code=args.trust_remote_code`。
3. 无其他文件变更, 改动简洁。

关键文件:

- `benchmarks/multi_turn/benchmark_serving_multi_turn.py` (模块 基准测试; 类别 source; 类型 core-logic): 唯一变更文件, 新增 --trust-remote-code 参数并传递给 tokenizer 加载。

关键符号: 未识别

关键源码片段

`benchmarks/multi_turn/benchmark_serving_multi_turn.py`

唯一变更文件, 新增 --trust-remote-code 参数并传递给 tokenizer 加载。

```
# 在参数解析器中新增 --trust-remote-code 标志
parser.add_argument(
    "--trust-remote-code",
    action="store_true",
    help="Trust remote code when loading the tokenizer.",
```

)

```
# 在加载 tokenizer 时传递该参数
tokenizer = AutoTokenizer.from_pretrained(
    args.model, trust_remote_code=args.trust_remote_code
)
```

评论区精华

无实质性代码讨论；仅有自动 bot 评论和三位维护者的 Approval，其中 ivanium、youkaichao、ywang96 均批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。新增参数默认 False，完全向后兼容。trust_remote_code 是 HuggingFace 标准参数，传递后仅在用户显式指定时启用远程代码执行。
- 影响：影响范围仅限于 multi-turn 基准测试脚本的用户。受益者：需要使用自定义 tokenizer 模型的用户可正常运行基准测试。无性能、安全或兼容性影响。
- 风险标记：低风险

关联脉络

- 暂无明显关联 PR