

PR #48703 完整报告

vllm-project/vllm

[XPU] [UT] [CI] add xpu config to run gpt-oss accuracy in ut and ci

合并时间: 2026-07-29 14:16

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/48703>

执行摘要

- 一句话: 为 XPU 添加 GPT-OSS 准确性测试配置和 CI 步骤
- 推荐动作: 该 PR 属于测试 /CI 基础设施增强, 值得 XPU 相关开发人员了解 CI 覆盖的扩展。核心逻辑开发人员可忽略。

功能与动机

在 XPU 设备上集成 GPT-OSS 准确性测试到单元测试和 CI 流程中, 确保模型在 Intel GPU 上的精度满足要求。此前缺少相应的配置文件和 CI 步骤。

实现拆解

1. 新增默认 attention 配置: 文件 tests/evals/gpt_oss/configs/gpt-oss-20b-xpu-baseline.yaml, 指定模型名称 openai/gpt-oss-20b、精度阈值 0.568 和 reasoning_effort: low, 不传递额外 server_args。
2. 新增 Triton Attention 配置: 文件 tests/evals/gpt_oss/configs/gpt-oss-20b-xpu-triton-attn.yaml, 除基线参数外添加 server_args: "--attention-backend TRITON_ATTN", 以覆盖不同的 attention 实现。
3. 创建设置列表: 文件 tests/evals/gpt_oss/configs/models-xpu.txt, 列出上述两个配置文件名, 供测试框架自动发现。
4. 配置 CI 步骤: 在 .buildkite/intel_jobs/test-intel.yaml 中新增 XPU GPQA Eval (GPT-OSS) 步骤, 安装 gpt-oss[eval]==0.0.5 并执行 pytest evals/gpt_oss/test_gpqa_correctness.py, 触发条件为 source_file_dependencies 中列出的相关文件变更。

关键文件:

- tests/evals/gpt_oss/configs/gpt-oss-20b-xpu-triton-attn.yaml (模块 测试配置; 类别 test; 类型 test-coverage): 新增的 XPU 配置, 用于 GPT-OSS 评估使用 Triton Attention 后端
- tests/evals/gpt_oss/configs/gpt-oss-20b-xpu-baseline.yaml (模块 测试配置; 类别 test; 类型 test-coverage): 新增的 XPU 基线配置, 用于默认 attention 后端的 GPT-OSS 评估
- .buildkite/intel_jobs/test-intel.yaml (模块 CI 流水线; 类别 config; 类型 configuration): CI 配置中新增 XPU GPQA Eval 步骤, 集成 GPT-OSS 测试到 XPU CI

- tests/evals/gpt_oss/configs/models-xpu.txt (模块 测试清单; 类别 docs; 类型 documentation) : 列出 XPU 评估配置文件名, 供测试框架读取

关键符号: 未识别

关键源码片段

tests/evals/gpt_oss/configs/gpt-oss-20b-xpu-triton-attn.yaml

新增的 XPU 配置, 用于 GPT-OSS 评估使用 Triton Attention 后端

```
# SPDX-License-Identifier: Apache-2.0
# SPDX-FileCopyrightText: Copyright contributors to the vLLM project
model_name: openai/gpt-oss-20b
metric_threshold: 0.568
reasoning_effort: low
server_args: "--attention-backend TRITON_ATTN"
```

评论区精华

PR 由 fork 仓库提交, Claude bot 评论表示自动化审查被禁用, 需 maintainer 触发。最终由 jikunshang 批准合并。无实质性技术讨论。

- 暂无高价值评论线程

风险与影响

- 风险: 变更范围有限, 仅涉及测试配置和 CI 步骤, 不修改任何核心代码或运行时逻辑。主要风险在于 CI 步骤依赖正确的基础镜像和环境 (如 gpt-oss[eval]==0.0.5 包可用), 但该风险由已有的 CI 基础设施管理。硬件要求 (mem: 24+) 可能在某些 agent 上不可用, 但已有标签保证。
- 影响: 对用户: 无直接感知, 对 XPU 设备上的 GPT-OSS 评估提供了自动化 CI 覆盖。对系统: 新增一个 CI 步骤, 增加了 CI 总执行时间, 但仅当 XPU 相关文件变更时触发。对团队: Intel GPU 开发者将获得更频繁的准确性回归检测。
- 风险标记: CI 配置变更, 新增测试配置

关联脉络

- 暂无明显关联 PR