

PR #56934 完整报告

vllm-project/vllm

[XPU][CI]Skip test_chat_completion_with_tools in Intel GPU CI

合并时间: 2026-09-15 22:04

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/56934>

执行摘要

- 一句话: XPU CI 追加 deselect, 摘除 llama3.2 工具调用用例
- 推荐动作: 不建议精读源码 (本 PR 无源码), 但值得花 1 分钟了解“CI 豁免如何被记录”的反面案例: 本次跳过既没写失败原因、也没留恢复条件, 是典型的可维护性折损。建议关注三点: (1) 由 Intel 侧补一个 tracking issue, 写清 llama3.2 在 XPU 上的具体失败模式与预期恢复条件; (2) 考虑把多处 --deselect 抽成 job 级别的变量或 quarantine 列表, 便于批量回滚与审计; (3) 对 optional: true 的作业建立定期 review 机制, 避免 deselect 列表无限增长。

功能与动机

PR body 只有一句 “## Purpose Skip test_chat_completion_with_tools in Intel GPU CI”, 没有关联 Issue、没有贴失败日志或错误栈, 也没有说明是模型下载失败、xgrammar/tool parser 行为差异还是 XPU kernel 不支持。从既有配置可以推断: 同一作业此前已对 granite-3.0-8b 与 ministral-3b 做过同类 deselect, 本次只是把同一测试的 llama3.2 参数也加入排除列表, 即 Intel GPU 侧该用例持续不可用, 需要通过 CI 配置临时绕过以保持流水线可用。

实现拆解

1. 变更入口: 唯一改动文件是 .buildkite/intel_jobs/entrypoints_intel.yaml, 命中 job key entrypoints-integration-api-server-generate-intel (label :intel: (BMG) Entrypoints Integration (API Server Generate), timeout_in_minutes: 65, optional: true, VLLM_TEST_DEVICE: xpu) 。
2. 核心改动: 该 job 的 commands 是一段折叠标量, 先用 run-intel-test.sh 进入 tests/ 目录跑 pytest -v -s tool_use, 随后串行执行 tool_parsers、generate、anthropic、cohere 四组用例。本次在原有两条 --deselect 之后追加第三条: --deselect="tests/tool_use/test_chat_completions.py::test_chat_completion_with_tools[llama3.2]"。变更位于 pytest 选择器层面, 不会修改 tests/tool_use/test_chat_completions.py 本身, 因此主干 CI (NVIDIA 等) 仍会完整跑该参数。
3. 影响链路: deselect 只影响 Intel GPU 作业中 tool_use 目录的用例集合, 后续 entrypoints/tool_parsers、entrypoints/generate 等命令不受影响; source_file_dependencies (vllm/、tests/tool_use 等) 未变, 触发条件与作业粒度保持原样。

4. 测试与配置配套：没有源码、单元测试或文档改动，只有这一处 CI 配置；作业本身是 `optional: true`，失败本不阻塞合并门禁，因此该改动更偏向“减少噪音失败 / 缩短流水线耗时”，而非修复硬性阻断。
5. 提交与协作：4 个 commit 中包含两次 `Merge branch 'main' into xiangdong/skip_chat'`（其中一次由 maintainer jikunshang 完成），说明分支曾落后主干较多；期间机器人提示“This PR is 23 commits behind upstream main”，需要 rebase 后才能重新触发 CI。

关键文件：

- `.buildkite/intel_jobs/entrypoints_intel.yaml`（模块 CI 配置；类别 config；类型 configuration；符号 `entrypoints-integration-api-server-generate-intel, test_chat_completion_with_tools[llama3.2]`）：唯一改动文件，向 Intel GPU(BMG) Entrypoints Integration 作业的 `pytest` 命令追加一条 `--deselect`，摘除 llama3.2 参数的工具调用用例

关键符号：未识别

评论区精华

本 PR 没有任何 review 评论 (`review_comments_count = 0`)，唯一可见的交互是 CI 触发流程：jikunshang 第一次 `/ci run` 被机器人以“This PR is 23 commits behind upstream main”拒绝，第二次 `/ci run` 才成功触发 Buildkite #89075。审查侧由 zhenwei-intel 与 jikunshang 两人 APPROVED，均未留文字意见；claude[bot] 因 PR 来自 fork 而自动禁用审查，提示“A repository maintainer can comment @claude review to run a one-time review”。换言之，这次“跳过测试”的决定没有留下任何技术论证记录，是否属于合理的临时豁免无法从讨论中判断。

- CI 触发前的分支新鲜度要求 (question)：分支通过 merge main 追平主干后 CI 正常触发，本 PR 最终合入；也解释了提交历史中两次 `Merge branch 'main'` 的来源。
- 跳过测试缺少技术论证 (testing)：以静默批准收尾，`deselect` 的原因、影响范围与恢复条件均未被记录。

风险与影响

- 风险：
 1. 测试覆盖缺口：`test_chat_completion_with_tools[llama3.2]` 在 Intel GPU 上的工具调用链路从此不再被验证，此前已被摘除的还有 `granite-3.0-8b` 与 `minstral-3b` 参数，XPU 上 `test_chat_completions.py` 的有效覆盖面持续收窄，工具调用回归可能在 XPU 侧静默通过。
 1. 豁免缺少闭环：没有关联 Issue、没有失败日志、也没有 TODO/ 到期时间，`entrypoints_intel.yaml` 里的 `deselect` 列表正在成为“只增不减”的清单，后续维护者难以判断这些用例是否已可恢复。
 2. 误伤范围：`deselect` 采用带参数 id 的精确 node id，只会摘掉 llama3.2 这一个参数，其他模型参数仍会执行，但若失败原因是 `test_chat_completions.py` 的公共 fixture 或工具调用通用逻辑，则同目录其他用例仍会以失败形式暴露，问题只是从“失败”变成“可能更晚才被发现”。

3. 配置本身低风险：改动是 YAML 折叠标量中的字符串拼接，语法层面只需保证 `--deselect=` 与后续 `&&` 的相对位置正确；新增行放在 `--deselect` 与 `&& pytest ...` 之间，不会破坏 shell 连接符语义。 - 影响：对用户与线上服务无直接影响：该文件只作用于 Buildkite 上的 Intel GPU 作业，不进镜像、不改运行时行为。对 Intel/XPU 团队影响最直接——BMG 作业的 `tool_use` 阶段少跑一个用例，失败噪音下降，但 XPU 上工具调用能力的可见性同时下降，问题会被推到下游用户或后续 P1 修复。对主干团队影响有限：作业是 `optional: true`，其他厂商与 NVIDIA 流水线的 `test_chat_completion_with_tools` 覆盖保持不变。 - 风险标记：测试覆盖缺口，临时豁免无跟踪 Issue，跳过原因未记录，`deselect` 列表持续增长

关联脉络

- PR #56601 [CI] Extend Arm CPU kernel shard timeout: 同属异构硬件 CI 的“放宽约束以保流水线可用”类改动（Arm CPU 超时预算 30→60 分钟），与本 PR 的跳过策略动机一致，都是硬件侧稳定性维护。
- PR #56955 [CI] Fix NIXL push worker test stub after failure deferral: 同样是测试 /CI 与实现契约不同步后的适配修复，可作为对照：那边选择修正测试桩而不是跳过用例，本 PR 选择了直接 `deselect`。
- PR #56324 [CI] Shard V1 KV Connectors 1→4: 同为 Buildkite 作业配置层面的治理改动，反映仓库近期在 CI 时长、稳定性与分片策略上的持续调整。
- PR #56408 [Frontend] Use XGrammar schema constraints for DeepSeek V4.1: 改动 `tool_parsers/structural` tag 相关前端路径，属于 `tool_use` 类集成测试的上游依赖面；若 Intel GPU 上的工具调用失败与解析 / 约束路径相关，本 PR 的 `deselect` 可能掩盖该 PR 引入的 XPU 侧回归。