

PR #52961 完整报告

vllm-project/vllm

Add Laguna-XS-2.1-INT4 to nightly CI

合并时间: 2026-09-01 02:22

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/52961>

执行摘要

- 一句话: 新增 Laguna INT4 夜间评估, NVFP4 迁移 chat 协议
- 推荐动作: 不值得作为源码精读对象, 但值得快速浏览: 它是理解 vLLM nightly eval 配置结构的最小样例。值得关注的设计决策是 " 评估协议 (chat completions vs raw completions) 与精度门限联动调整 ", 以及 " 跨硬件运行矩阵拆分 (models-small vs models-blackwell) " 的做法。对维护 nightly CI 的工程师而言, 这是一个低风险、易回滚的配置型变更。

功能与动机

PR body 非常简略, 仅说明 "I also took the liberty of migrating the NVFP4 variant to use chat completions too, since it improves scores"。结合变更内容推断, 动机有两层: 一是为 poolside/Laguna-XS-2.1-INT4 建立 nightly 精度回归监控, 避免 INT4 量化小模型在后续版本中静默退化; 二是 NVFP4 变体切换到 chat completions 协议后评估分数提升, 需要同步上调 accuracy_threshold 以保持门限有效性。

实现拆解

整个变更都是 nightly eval 配置调整, 没有源码主路径改动, 按 4 步拆解:

1. 新增 INT4 夜间评估配置: 新增 tests/evals/gsm8k/configs/Laguna-XS-2.1-INT4.yaml, 声明模型 poolside/Laguna-XS-2.1-INT4、accuracy_threshold: 0.90、num_questions: 1319、num_fewshot: 5、max_tokens: 1024、use_chat_completions: true、startup_max_wait_seconds: 1200, 服务参数为 --enforce-eager、--max-model-len 4096、--trust-remote-code。该配置为模型建立了 nightly GSM8K 精度基线。
2. NVFP4 变体协议迁移与门限联动: 修改 tests/evals/gsm8k/configs/Laguna-XS.2-NVFP4.yaml, 新增 max_tokens: 1024 与 use_chat_completions: true, 并将 accuracy_threshold 从 0.86 上调至 0.90。PR body 说明 chat completions 可提升评估分数, 门限同步上调是为了让门限继续反映真实精度水平。
3. 运行矩阵调整: 从 tests/evals/gsm8k/configs/models-small.txt 移除 INT4 条目 (同时补上该文件缺失的末尾换行符), 并在 models-blackwell.txt 中加入 Laguna-XS-2.1-INT4.yaml。由此 INT4 评估被限定在 Blackwell GPU runner 上运行, 与 NVFP4 同处 blackwell 清单, 避免在非 Blackwell 硬件上出现不支持或结果失真的情况。

4. CI 验证与收尾：全程两次 Buildkite CI (#84627、#86394)，期间修复 pre-commit 失败并合并 main，最终由 mgoin approve 合并。无测试源码、schema 或部署配套变更。

关键文件：

- tests/evals/gsm8k/configs/Laguna-XS-2.1-INT4.yaml (模块 评估配置；类别 test；类型 test-coverage)：本 PR 的核心新增，为 poolside/Laguna-XS-2.1-INT4 建立 nightly GSM8K 精度监控基线，包含 0.90 门限与 chat completions 协议。
- tests/evals/gsm8k/configs/Laguna-XS.2-NVFP4.yaml (模块 评估配置；类别 test；类型 test-coverage)：将 NVFP4 变体从 raw completions 迁移到 chat completions，门限 0.86 上调至 0.90，体现评估协议对分数的直接影响。
- tests/evals/gsm8k/configs/models-blackwell.txt (模块 模型清单；类别 docs；类型 documentation)：将 INT4 评估注册到 Blackwell 运行矩阵，决定该模型只在 Blackwell runner 上执行。
- tests/evals/gsm8k/configs/models-small.txt (模块 模型清单；类别 docs；类型 documentation)：将 INT4 从通用 small 矩阵摘除，确保只在 Blackwell 环境评估；同时修复文件缺失的末尾换行符。

关键符号：未识别

关键源码片段

tests/evals/gsm8k/configs/Laguna-XS-2.1-INT4.yaml

本 PR 的核心新增，为 poolside/Laguna-XS-2.1-INT4 建立 nightly GSM8K 精度监控基线，包含 0.90 门限与 chat completions 协议。

```
# poolside/Laguna-XS-2.1-INT4 夜间 GSM8K 评估配置 (新增)
# 目标：持续监控 INT4 量化小模型在 GSM8K 上的精度回归
model_name: "poolside/Laguna-XS-2.1-INT4"
accuracy_threshold: 0.90 # 精度门限，低于 0.90 的夜间运行会被标记为失败
num_questions: 1319 # 与 NVFP4 变体保持一致的问题数
num_fewshot: 5 # 5-shot 评估
max_tokens: 1024 # 限制生成长度，控制单次评估耗时
use_chat_completions: true # 使用 chat completions 协议评估，分数更高
startup_max_wait_seconds: 1200 # 拉起服务端的超时上限
server_args: >-
  --enforce-eager
  --max-model-len 4096
  --trust-remote-code # 模型仓库依赖远程代码执行
```

tests/evals/gsm8k/configs/Laguna-XS.2-NVFP4.yaml

将 NVFP4 变体从 raw completions 迁移到 chat completions，门限 0.86 上调至 0.90，体现评估协议对分数的直接影响。

```
# poolside/Laguna-XS.2-NVFP4 夜间 GSM8K 评估配置 (修改)
# 本 PR 将评估协议从 raw completions 迁移到 chat completions,
# 由于 chat 协议下分数提升，精度门限从 0.86 上调到 0.90
model_name: "poolside/Laguna-XS.2-NVFP4"
```

```
accuracy_threshold: 0.90 # 由 0.86 上调而来
num_questions: 1319
num_fewshot: 5
max_tokens: 1024 # 新增: 限制输出长度, 控制评估耗时
use_chat_completions: true # 新增: 迁移到 chat completions 协议
startup_max_wait_seconds: 1200
server_args: >-
  --enforce-eager
  --max-model-len 4096
  --trust-remote-code
```

评论区精华

该 PR 的实质 review 讨论很少:

- mergify[bot] 在评论中指出 pre-commit 检查失败, 要求运行 `uv pip install pre-commit>=4.5.1` 与 `pre-commit run --all-files` 后提交修复; 从后续提交历史看已修复并通过。
- claude[bot] 说明 fork PR 默认禁用自动 review, 维护者可评论 `@claude review` 触发。
- mgoin 两次触发 `/ci run` (Buildkite #84627、#86394) 并直接 approve, 无文字评价, 说明维护者对该变更的接受度较高。
- pre-commit 检查失败与 CI 重跑 (style): 后续提交 4e01406 通过 CI, mgoin approve 并合并, 问题已解决。
- NVFP4 迁移 chat completions 与阈值上调 0.90 (design): 变更直接落地在 Laguna-XS.2-NVFP4.yaml, 无反对意见, mgoin approve。

风险与影响

- 风险: 风险集中在 nightly CI 稳定性与供应链两方面:
 1. 门限提升可能引入 flaky: Laguna-XS.2-NVFP4.yaml 的 `accuracy_threshold` 从 0.86 提到 0.90, 依赖 chat completions 协议稳定提升分数这一假设。若模型仓库更新、批次波动或服务端行为变化, 0.90 门限可能让夜间评估出现偶发失败。
 2. --trust-remote-code 供应链风险: 两个 YAML 都带 `--trust-remote-code`, nightly 环境会执行模型仓库里的远程代码; 一旦 poolside 的模型仓库被篡改, CI 执行链会直接受影响。
 3. Blackwell 硬件依赖: INT4 配置只注册在 models-blackwell.txt, 若 Blackwell runner 可用性不足或队列积压, INT4 的精度监控覆盖会静默减少甚至中断。
 4. 评估基线不可比: NVFP4 的协议切换使旧门限 (0.86) 下的历史结果与新门限 (0.90) 不可直接对比, 后续回溯精度趋势时需要记录协议切换时点。- 影响: 影响范围限于内部 nightly eval CI, 不触及用户请求路径与产品代码:
- 对用户: 无直接影响。
- 对系统: nightly eval 新增 1 个模型运行 (INT4), NVFP4 的评估协议、`max_tokens` 与门限均改变, 评估结果基线需要重新解读。
- 对团队: poolside 系列模型的精度回归更早暴露; 同时该 PR 提供了一份最小可复用的 nightly eval 配置模板, 后续新增模型的监控配置可参照此写法。

- 风险标记：夜间门限提升可能引入 flaky, trust-remote-code 供应链风险，Blackwell 硬件可用性依赖，评估协议切换使历史基线不可比

关联脉络

- 暂无明显关联 PR