

# PR #39541 完整报告

vllm-project/vllm

[CI] Add DGX Spark GPQA smoke test

合并时间: 2026-06-23 05:52

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/39541>

## 执行摘要

- 一句话: 为 DGX Spark 添加 GPQA 烟雾测试 CI 步骤
- 推荐动作: 本 PR 属于常规 CI 基础设施扩展, 无核心逻辑变更。值得关注的点:
  - DGX Spark (SM120) 作为新硬件平台的支持入口。
  - 镜像构建中 CUDA 架构列表的调整方法。
  - 路径错误的修正应尽快合并以避免 CI 失败。 审阅者 @AndreasKaratzas 提出的配置复用建议可能驱动后续重构。

## 功能与动机

本 PR 的目的是在 DGX Spark CI runner 上运行 GPQA 评估烟雾测试, 以验证该硬件平台的基本功能。PR body 明确说明: “Smoke test to validate the dgx-spark CI runner - Runs GPQA eval with gpt-oss-20b on SM120 (no TP, baseline config)”。

## 实现拆解

本 PR 的变更可分为三个部分:

1. 新增 DGX Spark 模型配置: 在 tests/evals/gpt\_oss/configs/ 下新增 gpt-oss-20b-sm120.yaml 文件, 定义使用 openai/gpt-oss-20b 模型, 设置准确率阈值为 0.568 和 reasoning\_effort 为 low。此外新增 models-spark.txt 文件, 引用该配置文件名。
2. 添加 Buildkite CI 步骤: 在 .buildkite/test\_areas/lm\_eval.yaml 中新增一个可选 CI 步骤, 使用 device: dgx-spark, num\_devices: 1 (无 TP), 超时 120 分钟, 依赖 csrc/vllm/model\_executor/layers/quantization 和 tests/evals/gpt\_oss/, 命令安装 gpt-oss[eval]==0.0.5 并运行测试。
3. 适配 arm64 Docker 镜像构建: 修改 .buildkite/image\_build/image\_build\_arm64.sh, 将 torch\_cuda\_arch\_list 从仅 9.0 (GH200) 改为 9.0 12.0, 其中 12.0 对应 SM121 (DGX Spark/GB10), 并更新注释说明。

关键文件:

- tests/evals/gpt\_oss/configs/gpt-oss-20b-sm120.yaml (模块 模型配置; 类别 test; 类型 test-coverage): 核心模型配置文件, 定义 DGX Spark 上运行的模型名称、准确率阈值和推理努力级别。

- `.buildkite/test_areas/lm_eval.yaml` (模块 CI 配置; 类别 config; 类型 configuration) : Buildkite 主配置, 新增 DGX Spark 的 CI 步骤。
- `.buildkite/image_build/image_build_arm64.sh` (模块 构建脚本; 类别 other; 类型 core-logic) : 修改 arm64 Docker 镜像构建参数, 添加 SM121 支持。
- `tests/evals/gpt_oss/configs/models-spark.txt` (模块 测试列表; 类别 docs; 类型 documentation) : 模型列表文件, 引用新配置, 供测试框架加载。

关键符号: 未识别

## 关键源码片段

### `tests/evals/gpt_oss/configs/gpt-oss-20b-sm120.yaml`

核心模型配置文件, 定义 DGX Spark 上运行的模型名称、准确率阈值和推理努力级别。

```
# SPDX-License-Identifier: Apache-2.0
# SPDX-FileCopyrightText: Copyright contributors to the vLLM project
# 针对 DGX Spark (SM120) 的 GPQA 评估配置
# 模型 : gpt-oss-20b, 不使用 TP, 单设备运行
model_name: "openai/gpt-oss-20b"
metric_threshold: 0.568
reasoning_effort: "low"
```

### `.buildkite/image_build/image_build_arm64.sh`

修改 arm64 Docker 镜像构建参数, 添加 SM121 支持。

```
# build for arm64 GPU targets: Grace/GH200 (sm_90) and DGX Spark/GB10
# (sm_121, family-covered by 12.0 under CUDA 13)
docker build --file docker/Dockerfile \
  --platform linux/arm64 \
  --build-arg max_jobs=16 \
  --build-arg nvcc_threads=4 \
  --build-arg torch_cuda_arch_list="9.0 12.0" \
  --build-arg USE_SCCACHE=1 \
  --build-arg buildkite_commit="$BUILDKITE_COMMIT" \
  --tag "$REGISTRY"/"$REPO":"$BUILDKITE_COMMIT"-arm64 \
  ...
```

## 评论区精华

1. CI 配置路径错误: @gemini-code-assist[bot] 指出 `--config-list-file=configs/models-spark.txt` 路径不正确, 实际文件位于 `tests/evals/gpt_oss/configs/models-spark.txt`, 假设测试从 `tests/` 目录执行, 应改为 `evals/gpt_oss/configs/models-spark.txt`。
2. 缺少 `key` 属性: @AndreasKaratzas 注意到新步骤没有 `key` 属性, 但观察到其他测试组也缺少该属性, 因此未强制要求。
3. 配置重复性技术债务: @AndreasKaratzas 评论指出, 新的配置文件与 baseline 配置几乎一样 (仅无 TP), 每新增一个模型评估会增加两个几乎重复的文件, 形成增量技术债务, 建议长期寻找更好的解决方案。

- CI 配置路径错误 (correctness): 路径需要修正, 但 PR 最终合并时是否已修正? 从提交历史看, 最后的提交是 [CI] Build arm64 CI image for DGX Spark (sm\_121), 未包含该路径修复, 因此问题可能仍未解决。
- 缺少 key 属性 (style): 无需修复, 符合现有风格。
- 配置重复性技术债务 (design): 未在本次 PR 中解决, 但提出长期关注点。

## 风险与影响

- 风险:
  1. CI 配置错误风险: Review 中发现的路径错误可能导致测试步骤因找不到配置文件而失败, 需要修正。 - 涉及文件: .buildkite/test\_areas/lm\_eval.yaml 第 235 行。
  2. arm64 镜像构建风险: 将 CUDA 架构列表从 9.0 改为 9.0 12.0 会增加镜像构建时间, 并可能引入与 SM121 相关的编译兼容性问题, 但这是适配新硬件的必要改动。 - 涉及文件: .buildkite/image\_build/image\_build\_arm64.sh。
  3. 测试本身风险较低: 该步骤为 optional (optional: true), 不会阻塞主线 CI。 - 影响: 影响范围: 仅限于 DGX Spark (SM120) 硬件平台, 影响面小。影响程度:
- 对用户: 无直接影响, 仅新增 CI 验证流程。
- 对系统: arm64 镜像构建时间可能因增加一个架构而延长, 但影响有限。
- 对团队: 新增 CI 步骤后, 需维护模型配置列表, 如 @AndreasKaratzas 指出的配置重复问题。
- 风险标记: 路径配置可能错误, 缺少测试覆盖, 配置重复性技术债务

## 关联脉络

- PR #43752 [Quantization][CI] add humming lm-eval test: 同样是添加 lm-eval 测试配置和 CI 步骤, 与本 PR 相似的模式, 均在 .buildkite/test\_areas/lm\_eval.yaml 中新增步骤。