PR #41575 完整报告

vllm-project/vllm

[ROCm][CI] Use vLLM generation defaults for DeepSeek prefetch-offload eval

合并时间：2026-05-06 09:08

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/41575

执行摘要

- 一句话：修复 DeepSeek V2-Lite 预取卸载评估脚本默认配置
- 推荐动作：该 PR 简单且经过充分审查，可以合并。不需要进一步审查。

功能与动机

确保定时运行的 DeepSeek V2-Lite 预取卸载准确性评估（H100-MI300）具有确定性，避免因模型 HF 生成默认值差异导致的非确定性结果。该评估用于 ROCm CI 中的回归检测。

实现拆解

在 .buildkite/scripts/scheduled_integration_test/deepseek_v2_lite_prefetch_offload.sh 脚本的 vllm serve 命令中添加 --generation-config vllm 标志。该标志强制 vLLM 使用内置生成配置，而不是依赖模型自身的 Hugging Face 默认值。

关键文件：

- .buildkite/scripts/scheduled_integration_test/deepseek_v2_lite_prefetch_offload.sh（模块 CI 脚本；类别 test；类型 test-coverage）：添加 --generation-config vllm 标志，以使用 vLLM 默认值进行确定性评估。

关键符号：未识别

关键源码片段

.buildkite/scripts/scheduled_integration_test/deepseek_v2_lite_prefetch_offload.sh

添加 --generation-config vllm 标志，以使用 vLLM 默认值进行确定性评估。

```
# ── Build optional vllm serve flags
# ──────────────────────────────────────────────────────

EXTRA_ARGS=()
if [[ -n "${ATTENTION_BACKEND:-}" ]]; then
  echo "Using attention backend: ${ATTENTION_BACKEND}"
  EXTRA_ARGS+=(--attention-backend "${ATTENTION_BACKEND}")
fi
```

```
# ... cleanup function ...

vllm serve "$MODEL" \
  --max-model-len 2048 \
  --offload-group-size 8 \
  --offload-num-in-group 2 \
  --offload-prefetch-step 1 \
  --offload-params w13_weight w2_weight \
  --generation-config vllm \  # <--- 新增：使用 vLLM 生成默认值而非模型的 HF 默认值
  --port "$PORT" \
  ${EXTRA_ARGS+"${EXTRA_ARGS[@]}"} &
SERVER_PID=$!
wait_for_server "$PORT"

# ... 运行 GSM8K 评估并检查准确性阈值 ...
```

评论区精华

该 PR 没有来自人工审查者的实质性讨论。机器人审查者（Claude、Gemini）未提供与变更相关的反馈。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅影响 ROCm 上 DeepSeek V2-Lite 预取卸载的定时评估脚本，不涉及生产代码或核心推理路径。
- 影响：影响范围仅限于 ROCm CI 中的定时评估脚本，使其在 GSM8K 评估中行为确定。不会影响用户或系统行为。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR