

PR #54852 完整报告

vllm-project/vllm

[CI][ROCm] Add DSpark evals

合并时间: 2026-09-03 02:14

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/54852>

执行摘要

- 一句话: 为 Kimi K3 和 DeepSeek V4 添加 MI355 ROCm DSpark 投机解码评估覆盖。
- 推荐动作: 这是一个常规的 CI 测试配置添加 PR, 风险较低, 主要价值在于提升测试覆盖。PR 作者已提供在目标硬件 (MI355) 上通过评估的证据, 且配置结构清晰。建议快速审查重点: 1) 确认 `lm_eval.yaml` 中 AMD 镜像任务的依赖和超时设置合理; 2) 检查新 YAML 配置文件中的关键参数 (如阈值、环境变量) 与现有 NVIDIA 配置或已知工作配置是否一致。适合作为低风险、高价值的维护性变更合入。

功能与动机

此 PR 堆叠在 #54817 之上, 旨在补充而非重复 draft #50519 中的工作。其核心动机是“使用当前的 `lm_eval` 镜像, 为 DeepSeek 在必需的 V2 运行器上, 并配置明确的 AITER MoE 路径, 为 Kimi K3 和 DeepSeek V4 添加 MI355 ROCm/AITER DSpark 评估覆盖”, 从而在 AMD MI355 硬件上建立对 DSpark 投机解码功能的持续评估。

实现拆解

1. 添加评估配置文件: 新增三个 YAML 格式的评估配置文件, 分别为 `DeepSeek-V4-Flash-DSpark-AITER-TEP4.yaml`、`Kimi-K3-pruned75-DSpark-AITER-TP4.yaml` 和 `Kimi-K3-pruned75-DSpark-TP4.yaml`。这些文件定义了模型名称、准确性阈值、投机解码参数 (如 `method: dspark`、`num_speculative_tokens`) 以及特定于平台的启动参数 (如 `--moe-backend aiter`) 和环境变量 (如 `VLLM_ROCM_USE_AITER: "1"`), 为测试提供了具体的运行时配置。
2. 配置 CI 流水线: 修改 `.buildkite/test_areas/lm_eval.yaml`, 为现有的 `:nvidia: (B200)` LM Eval Spec Decode 任务添加了一个 `mirror.amd` 区块。此区块定义了一个在 `mi355_4` 设备上运行的 AMD 镜像任务, 超时时间 120 分钟, 依赖于 `image-build-amd` 镜像, 并指定了详细的 `source_file_dependencies` 和使用 `models-spec-decode-rocm.txt` 配置列表的执行命令。这是实现跨平台 CI 覆盖的核心步骤。
3. 更新配置列表与文档: 创建新文件 `tests/evals/gsm8k/configs/models-spec-decode-rocm.txt`, 列出仅用于 ROCm 的评估配置。同时更新 `tests/evals/gsm8k/configs/models-spec-decode.txt`, 将其内容从仅包含 DeepSeek 的配置 (并已重命名) 调整为同时包含重命名后的 DeepSeek 配置和新增的 Kimi 配置, 确保 NVIDIA 平台的评估配置也同步更新。
4. 重命名现有配置: 将文件 `tests/evals/gsm8k/configs/DeepSeek-V4-Flash-DSpark-confidence-TP4.yaml` 重命名为 `DeepSeek-V4-Flash-DSpark-confidence-TEP4.yaml`, 以反映其

用于 TEP4（可能指特定配置或硬件）的用途，并保持命名一致性。此变更在 models-spec-decode.txt 中同步更新。

5. 测试验证：PR 描述中提到已在 4x MI355 上成功运行了 1,319 个问题的 GSM8K 评估，验证了 Kimi 和 DeepSeek 模型在新增配置下的准确性和投机解码接受率，表明配置有效。

关键文件：

- .buildkite/test_areas/lm_eval.yaml（模块 CI 流水线；类别 config；类型 configuration）：定义了 CI 评估流水线，本次修改为其添加了 AMD MI355 镜像任务，是实现跨平台测试覆盖的关键配置文件。
- tests/evals/gsm8k/configs/DeepSeek-V4-Flash-DSpark-AITER-TEP4.yaml（模块 评估配置；类别 test；类型 test-coverage）：新增的评估配置文件，为 DeepSeek V4 模型在 ROCm 平台上使用 AITER MoE 后端和 DSpark 投机解码定义了完整的测试参数，是验证特定硬件 - 软件组合的关键。

关键符号：未识别

关键源码片段

tests/evals/gsm8k/configs/DeepSeek-V4-Flash-DSpark-AITER-TEP4.yaml

新增的评估配置文件，为 DeepSeek V4 模型在 ROCm 平台上使用 AITER MoE 后端和 DSpark 投机解码定义了完整的测试参数，是验证特定硬件 - 软件组合的关键。

```
# DeepSeek V4 在 ROCm 平台上使用 AITER 后端的 DSpark 评估配置
model_name: "deepseek-ai/DeepSeek-V4-Flash-DSpark"
accuracy_threshold: 0.92 # 期望的 GSM8K 准确率下限
min_acceptance_length: 1.2 # DSpark 投机解码接受长度的宽松下限，用于冒烟测试
num_questions: 1319 # GSM8K 评估集大小
server_args: >-
  --moe-backend aiter # 关键：启用 ROCm AITER MoE 后端
  --attention_config.indexer_kv_dtype=fp8 # KV 缓存数据类型
  --speculative-config '{"method":"dspark", "num_speculative_tokens":7, "draft_sample_method":
    "probabilistic", "enable_adaptive_verification":false}' # DSpark 投机解码配置
env:
  VLLM_USE_V2_MODEL_RUNNER: "1" # 启用 V2 模型运行器
  VLLM_ROCM_USE_AITER: "1" # 环境变量，启用 ROCm AITER 支持
  VLLM_ROCM_USE_AITER_MOE: "1" # 环境变量，启用 ROCm AITER MoE 支持
```

评论区精华

Review 过程中未出现深度技术讨论。审阅者 shen-shanshan 直接给予了 APPROVED 状态。claude[bot] 的评论仅为说明性的，提示该仓库配置为手动代码审查。PR 提交者 AndreasKaratzas 在 issue 中发布 CI 运行命令并监控结果，显示了清晰的验证流程。整个评审焦点在于确认配置的正确性和 CI 任务的完整性。

- 暂无高价值评论线程

风险与影响

- 风险:

1. CI 配置依赖风险: PR 核心是添加 CI 评估配置, 主要风险在于配置参数 (如模型阈值、超时时间、设备依赖) 设置不当可能导致 CI 任务不稳定或失败。例如, `startup_max_wait_seconds: 1800` 和 `timeout_in_minutes: 120` 的设定需与实际硬件和模型加载时间匹配。
2. 平台特定代码路径风险: 新增的评估配置 (特别是 AITER 相关配置) 高度依赖 ROCm/AITER 内核和 MoE 后端的稳定性。如果底层 AITER 实现或 DSpark 投机解码在 ROCm 上存在未发现的 bug, 评估可能会失败。
3. 无源码逻辑变更风险: 此 PR 未修改 vLLM 核心源代码 (如 `vllm/` 目录下的 Python 文件), 因此对推理逻辑、性能或安全的直接影响极低。风险主要局限于 CI 基础设施和测试覆盖范围。

- 影响:

1. CI 与开发者影响: 直接影响 CI 流水线。为 ROCm 平台增加了重要的投机解码评估覆盖, 有助于提前捕获与 AMD 硬件相关的回归问题。变更了 NVIDIA B200 平台的 Spec Decode 评估配置, 可能影响其测试执行和触发条件。
2. 系统影响: 对 vLLM 运行时系统无直接影响。评估配置的准确性间接关系到用户对模型在特定硬件上性能 (如 DSpark 投机解码效率) 的预期。
3. 团队影响: 为维护 ROCm 支持的团队提供了更自动化的质量门禁。标准化的评估配置也有助于团队在不同模型和硬件组合间保持一致的测试标准。 - 风险标记: CI 配置依赖, 平台特定路径

关联脉络

- 暂无明显关联 PR