

PR #43752 完整报告

vllm-project/vllm

[Quantization][CI] add humming lm-eval test

合并时间: 2026-06-23 02:23

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/43752>

执行摘要

- 一句话: 为 Humming 量化添加 lm-eval 测试配置与 CI 步骤
- 推荐动作: 此 PR 值得阅读以了解 vLLM 中如何为量化方法添加 lm-eval 测试。重点关注测试配置文件的组织方式和 CI 步骤的依赖设置。对于 Humming 量化开发者, 可参考此模式扩展更多模型的评估。

功能与动机

为 Humming 量化方案建立持续的正确性验证, 确保在 GSM8K 数学推理任务上的精度不退化。PR body 未提供详细说明, 但从 commit 和讨论可知作者希望将测试集成到 CI 中以防止回归。

实现拆解

实现拆解:

1. 添加模型评估配置文件: 在 tests/evals/gsm8k/configs/humming/ 目录下新增 4 个 YAML 文件, 分别定义 Qwen3-30B-A3B-MXFP4A16 和 gpt-oss-20b 在纯量化与激活 FP8 量化模式下的模型名称、精度阈值、fewshot 参数等。
2. 添加测试文件列表: 新增 config.txt 和 config-act-fp8.txt 两个列表文件, 供测试框架按批次执行。
3. 更新 CI 配置: 在 .buildkite/test_areas/lm_eval.yaml 中为 A100、H100、B200 三种 GPU 添加新的测试步骤, 设置 source_file_dependencies 以追踪相关源码变更, 步骤标记为 optional: true 和 temporary。
4. 修复源码 API: 在 vllm/model_executor/layers/quantization/humming.py 中将 layer.ensure_moe_quant_config_init() 改为 layer._ensure_moe_quant_config_init(), 对齐内部方法命名。
5. 更新依赖版本: 在 requirements/cuda.txt 中更新 humming-kernels 版本 (多次迭代后最终定为 0.1.6)。

关键文件:

- tests/evals/gsm8k/configs/humming/Qwen3-30B-A3B-MXFP4A16-humming-act-fp8.yaml (模块 测试配置; 类别 test; 类型 test-coverage): 新增的测试配置文件, 定义激活 FP8 量化模型的 GSM8K 评估参数, 包含环境变量设置。

- tests/evals/gsm8k/configs/humming/gpt-oss-20b-humming-act-fp8.yaml (模块 测试配置; 类别 test; 类型 test-coverage) : 新增的测试配置文件, 为 gpt-oss-20b 模型定义激活 FP8 量化模式的评估参数。
- vllm/model_executor/layers/quantization/humming.py (模块 量化模块; 类别 source; 类型 data-contract; 符号 HummingMoEMethod.process_weights_after_loading) : 核心源码变更, 修复了 MoE 量化配置初始化方法调用, 对齐内部 API。
- tests/evals/gsm8k/configs/humming/Qwen3-30B-A3B-MXFP4A16-humming.yaml (模块 测试配置; 类别 test; 类型 test-coverage) : 新增的测试配置文件, 定义纯量化 (MXFP4) 模型的 GSM8K 评估参数。
- tests/evals/gsm8k/configs/humming/gpt-oss-20b-humming.yaml (模块 测试配置; 类别 test; 类型 test-coverage) : 新增的测试配置文件, 为 gpt-oss-20b 模型定义纯量化模式的评估参数。
- .buildkite/test_areas/lm_eval.yaml (模块 CI 配置; 类别 config; 类型 configuration) : CI 配置更新, 新增 Humming 量化测试步骤, 定义源文件依赖和运行命令。
- tests/evals/gsm8k/configs/humming/config-act-fp8.txt (模块 测试列表; 类别 docs; 类型 documentation) : 新增的测试文件列表, 包含激活 FP8 量化模式的配置文件路径。
- tests/evals/gsm8k/configs/humming/config.txt (模块 测试列表; 类别 docs; 类型 documentation) : 新增的测试文件列表, 包含纯量化模式的配置文件路径。
- requirements/cuda.txt (模块 依赖文件; 类别 docs; 类型 documentation) : 更新 humming-kernels 依赖版本至 0.1.6, 对齐最新 humming 功能。

关键符号: HummingMoEMethod.process_weights_after_loading

关键源码片段

tests/evals/gsm8k/configs/humming/Qwen3-30B-A3B-MXFP4A16-humming-act-fp8.yaml

新增的测试配置文件, 定义激活 FP8 量化模型的 GSM8K 评估参数, 包含环境变量设置。

```
model_name: "nm-testing/Qwen3-30B-A3B-MXFP4A16" # 模型名称, 来自 nm-testing 仓库
accuracy_threshold: 0.86 # GSM8K 准确率最低阈值
num_questions: 1319 # 评估问题总数
num_fewshot: 5 # few-shot 示例数
server_args: >- # 服务器启动参数
    --enforce-eager                # 强制 eager 模式
    --max-model-len 8192           # 最大模型长度
    --tensor-parallel-size 1       # 张量并行度
    --quantization humming         # 使用 humming 量化
    --kernel-config.enable_flashinfer_autotune=False # 禁用 flashinfer 自动调优
env:
    VLLM_HUMMING_INPUT_QUANT_CONFIG: '{"dtype":"float8e4m3"}' #
    环境变量: 输入量化配置为 FP8
```

vllm/model_executor/layers/quantization/humming.py

核心源码变更, 修复了 MoE 量化配置初始化方法调用, 对齐内部 API。

```
# 在 process_weights_after_loading 方法中，MoE 量化配置初始化调用改为私有方法
# 变更行：layer.ensure_moe_quant_config_init() -> layer._ensure_moe_quant_config_init()
# 目的是对齐内部命名约定，避免外部直接调用
```

```
# use moe modular
experts: HummingIndexedExperts | HummingGroupedExperts
layer._ensure_moe_quant_config_init() # 调用私有方法初始化量化配置
assert self.moe_quant_config is not None
if get_humming_moe_gemm_type() == "indexed":
    experts = HummingIndexedExperts(layer, self.moe, self.moe_quant_config)
else:
    experts = HummingGroupedExperts(layer, self.moe, self.moe_quant_config)
self.experts = experts
```

评论区精华

在 Review 中，@AndreasKaratzas 指出 YAML 可能存在格式问题导致 bootstrap 失败，并询问两个 txt 文件重复的原因。@jinzhen-lin 回复已修复格式问题，并解释保留两个文件是因为未来可能仅有支持 w4a16 而不支持 fp8 激活的模型，届时 `config.txt` 和 `config-act-fp8.txt` 会分化。此外，@AndreasKaratzas 还询问了冗余的 build 文件，作者澄清是误提交并已删除。

- 配置重复与 bootstrap 问题 (question): @jinzhen-lin 修复了 YAML 格式，并解释保留两个文件以备未来仅支持 w4a16 的模型。

风险与影响

- 风险：主要风险在于 CI 步骤被标记为 optional 和 temporary，可能在实际生产 CI 中被忽略或过期，导致 Humming 量化得不到持续验证。另外，humming.py 中的方法名变更（`ensure_moe_quant_config_init` → `_ensure_moe_quant_config_init`）如果被其他外部插件依赖，可能造成兼容性问题；但当前是内部方法，风险可控。配置文件中的精度阈值（0.86 和 0.30）需要根据实际运行结果调整，若未来模型更新导致精度下降，测试可能误报失败。
- 影响：对用户无直接影响，变更集中在 CI 和测试配置。对系统而言，新增的 CI 步骤会增加少量构建时间（optional 所以不阻塞）。对团队而言，此 PR 为 Humming 量化建立了自动回归防护，有助于早期发现精度退化。影响程度中等，主要作用于测试基础设施。
- 风险标记：CI 步骤可选可能被忽视，方法名变更兼容性

关联脉络

- 暂无明显关联 PR