

PR #43262 完整报告

vllm-project/vllm

update GPU json file based on h200 recipes

合并时间: 2026-05-21 11:57

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/43262>

执行摘要

本 PR 重构了 H200 性能基准测试的 JSON 配置文件，引入 defaults 段以减少重复，并新增了多种随机输入输出长度的测试，同时移除了旧模型测试。变更仅涉及一个文件，但 review 中指出了两处潜在问题（语法错误和参数缺失），可能影响 CI 稳定性。

功能与动机

PR body 明确说明: "update vLLM Benchmark Suite json file for H200 based on vLLM recipes", 目的是使基准测试配置与 H200 设备上的官方 recipes (如 [Llama-3.1-8B-Instruct](#)、[Llama-3.3-70B-Instruct](#)、[Gemma-4-E4B-it](#)) 保持一致，从而在 H200 裸机上执行准确的性能测试。

实现拆解

1. 结构重构: 将原本的 JSON 数组改为对象，包含 defaults 和 tests 两个顶级键。defaults 定义了所有测试共享的参数（模型、TP、数据集、QPS 列表等），大幅减少每个测试项的重复内容。
2. 测试重组: 每个测试仅覆盖与其默认值的差异字段，便于维护和新增。例如 serving_llama8B_tp1_sharegpt 只需指定 dataset_name 和 dataset_path。
3. 新增随机输入长度测试: 为 Llama-8B 添加了四种组合（128/128、128/2048、2048/128、2048/2048），以评估不同输入输出长度下的性能表现。
4. 添加新模型测试:
 - serving_llama70B_tp4_random_128_128: 使用 Llama-3.3-70B-Instruct, TP=4 (需注意 review 指出的缺失问题)。
 - serving_gemma4-e4b_tp1_random_128_128: 使用 Google Gemma-4-E4B-it, 启用了 enable_auto_tool_choice、tool_call_parser 等参数。
5. 移除旧测试: 删除了 serving_mixtral8x7B_tp2_sharegpt 和 serving_llama70B_tp4_sharegpt_specdecode, 使测试集更聚焦于 H200 常见模型。

[.buildkite/performance-benchmarks/tests/serving-tests.json](#)

唯一变更文件，完整重构了 H200 性能基准测试配置，引入 defaults 结构并调整测试集。

```
/* 新结构使用 defaults 定义公共参数，tests 只保留增量 */
{
  "defaults": {
    "qps_list": ["inf"],
```

```

"max_concurrency_list": [12, 16, 24, 32, 64, 128, 200],
"server_parameters": {
  "model": "meta-llama/Llama-3.1-8B-Instruct",
  "tensor_parallel_size": 1,
  "dtype": "bfloat16"
},
"client_parameters": {
  "model": "meta-llama/Llama-3.1-8B-Instruct",
  "backend": "vllm",
  "ignore-eos": "",
  "temperature": 0,
  "num_prompts": 200
}
},
"tests": [
  {
    "test_name": "serving_llama8B_tp1_sharegpt",
    /* 只覆盖 sharegpt 相关参数 */
    "server_parameters": { "tensor_parallel_size": 1 },
    "client_parameters": {
      "dataset_name": "sharegpt",
      "dataset_path": "./ShareGPT_V3_unfiltered_cleaned_split.json"
    }
  },
  {
    "test_name": "serving_llama8B_tp1_random_128_128",
    "server_parameters": { "tensor_parallel_size": 1 },
    "client_parameters": {
      "dataset_name": "random",
      "random-input-len": 128,
      "random-output-len": 128
    }
  }
]
/* 更多测试 ... */
}

```

评论区精华

- gemini-code-assist[bot]: 指出第 31–35 行存在 JSON 语法错误，一个独立的 client_parameters 对象缺少 test_name，会导致解析失败。
- gemini-code-assist[bot]: 指出 serving_llama70B_tp4_random_128_128 测试未设置 tensor_parallel_size，默认使用 defaults 中的 1，70B 模型在单卡 H200 上很可能 OOM。
 - 人工审核者 bigPYJ1151 批准了 PR，未对上述问题提出修改要求。

风险与影响

- 技术风险

- JSON 语法错误将直接导致 CI 基准测试流水线解析失败，需要紧急修复。
- 70B 测试缺少 TP=4 设置会导致 OOM 或加载失败，该测试将失去意义。
- 影响范围：仅影响 H200 环境下的 CI 性能基准测试。对用户无影响，但可能影响内部性能回归检测的可靠性。
- 影响程度：中等。若这些问题未被修复，会导致测试流水线不稳定，性能数据缺失。

关联脉络

本 PR 属于持续配置维护的一部分，与近期其他基准测试调整（如 #43186 降低阈值）共同优化 CI 稳定性。目前无直接关联的 issue 或其他 PR。