PR #47048 完整报告

vllm-project/vllm

[CI] Move distributed small LM eval to B200

合并时间：2026-07-01 01:34

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/47048

执行摘要

- 一句话：将分布式小模型 LM eval 从 2xL4 迁移到 2xB200 并增加 DiffusionGemma chat completions 支持
- 推荐动作：该 PR 属于常规 CI 维护，代码简单清晰，合并迅速。建议关注 CI 硬件迁移的策略，以及 chat completions 端点适配的模式，可作为后续类似评测扩展的参考。

功能与动机

现有分布式小模型 LM 评测运行在 2xL4 上，随着硬件升级，需要迁移到性能更强的 2xB200 以加速评测并获得更稳定的结果。DiffusionGemma 模型使用 chat completions 端点，而现有评测工具仅支持 completions 端点，因此需要扩展 gsm8k_eval.py 以支持 chat 调用。

实现拆解

1. CI 配置调整（.buildkite/test_areas/lm_eval.yaml）：将原有的 LM Eval Small Models (2xB200) 步骤重命名为 1xB200 并调整 key；将 LM Eval Small Models (2xL4) 步骤替换为 LM Eval Small Models Distributed (2xB200)，设备改为 b200-k8s，超时从 10 分钟调整为 120 分钟，保留 autorun_on_main。
2. 评测工具扩展（tests/evals/gsm8k/gsm8k_eval.py）：新增 call_vllm_chat_api 异步函数，与 call_vllm_api 类似但使用 /v1/chat/completions 端点，传递 model 和 messages 结构。修改 evaluate_gsm8k 函数，增加 model 和 use_chat_completions 参数，在 get_answer 内部根据标志选择调用 chat 或 completion API。
3. 测试用例更新（tests/evals/gsm8k/test_gsm8k_correctness.py）：run_gsm8k_eval 函数从配置中读取 model_name 和 use_chat_completions 并传递给 evaluate_gsm8k。
4. 模型配置添加（tests/evals/gsm8k/configs/DiffusionGemma-26B-A4B-it-FP8-dynamic.yaml）：增加 use_chat_completions: true 配置项，使该模型评测时使用 chat 端点。

关键文件：

- tests/evals/gsm8k/gsm8k_eval.py（模块 评测工具；类别 test；类型 test-coverage；符号 call_vllm_chat_api）：核心变更：新增 call_vllm_chat_api 函数，修改 evaluate_gsm8k 支持 chat completions，实现评测工具的扩展。
- .buildkite/test_areas/lm_eval.yaml（模块 CI 配置；类别 config；类型 configuration）：CI 配置文件更改：重命名和移动了两个 LM eval 步骤，实现了从 2xL4 到 2xB200 的迁移。

- tests/evals/gsm8k/test_gsm8k_correctness.py（模块 测试用例；类别 test；类型 test-coverage）：测试用例更新：传递新参数 model 和 use_chat_completions 给评测函数。
- tests/evals/gsm8k/configs/DiffusionGemma-26B-A4B-it-FP8-dynamic.yaml（模块 模型配置；类别 test；类型 test-coverage）：模型配置添加 use_chat_completions: true 以启用 chat 端点评测。

关键符号：call_vllm_chat_api, evaluate_gsm8k

关键源码片段

tests/evals/gsm8k/gsm8k_eval.py

核心变更：新增 call_vllm_chat_api 函数，修改 evaluate_gsm8k 支持 chat completions，实现评测工具的扩展。

```
async def call_vllm_chat_api(
    session: aiohttp.ClientSession,
    model: str,
    prompt: str,
    temperature: float,
    max_tokens: int,
    stop: list[str] | None = None,
    url: str | None = None,
    seed: int | None = None,
) -> tuple[str, int]:
    """Call vLLM's OpenAI-compatible chat completions endpoint for DiffusionGemma."""
    data = {
        "model": model,
        "messages": [{"role": "user", "content": prompt}],
        "temperature": temperature,
        "max_tokens": max_tokens,
        "stop": stop,
    }
    if seed is not None:
        data["seed"] = seed

    try:
        # POST 到 /v1/chat/completions 而非 /v1/completions
        async with session.post(f"{url}/v1/chat/completions",
                                json=data) as response:
            response.raise_for_status()
            result = await response.json()
            # 从 message.content 而非 choices[0].text 提取
            text = result["choices"][0]["message"]["content"] or ""
            completion_tokens = result.get("usage", {}).get("completion_tokens", 0)
            return text, completion_tokens
    except Exception as e:
        print(f"Error calling vLLM chat API ({type(e).__name__}): {e}")
```

```
    return "", 0
```

评论区精华

PR 未引发实质性讨论。审核人 mgoin 直接批准，表示“Nice! LGTM”。Claude bot 的自动评论未开启深入审查。

- 整体审查 (other): 已合并

风险与影响

- 风险：主要风险来自 CI 硬件迁移后 B200 的资源可用性和稳定性，但该步骤属于可选（optional），失败不会阻塞主线。新增的 chat completions 调用只针对 DiffusionGemma 模型，且代码结构与其他调用一致，出错概率低。未引入新的依赖或安全风险。
- 影响：影响范围仅限于 CI 中的 LM 评测步骤：分布式小模型评测将不再使用 2xL4，改为 2xB200；单设备 B200 评测步骤名称变更但逻辑不变。对用户无直接影响，对 CI 团队意味着需要管理 B200 资源。新增 chat completions 支持使得未来其他 chat 类模型可以复用该路径。
- 风险标记：CI 硬件迁移，新增 chat completions 调用路径

关联脉络

- 暂无明显关联 PR