PR #49749 完整报告

vllm-project/vllm

[CI] Stabilize memory-sensitive compile and structured output tests

合并时间：2026-07-25 04:45

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/49749

执行摘要

- 一句话：稳定内存敏感的 CI 测试
- 推荐动作：值得快速合并以恢复 CI 健康，但应跟进 njhill 建议的根本原因调查，考虑在 EngineCore 层面修复循环引用或改进 GC 兼容性，而非依赖 finalizer。

功能与动机

参考关联 Issue #49672，BAAI/bge-multilingual-gemma2 的编译测试在 VLLM_COMPILE + inductor 路径下出现 KV 缓存内存不足（Available KV cache memory: -0.45 GiB）。PR body 指出结构化输出测试因依赖 GC 回收引擎，参数化用例间残留约 30.7 GiB 内存导致后续用例 OOM。

实现拆解

1. 调整编译测试参数 (tests/compile/fullgraph/test_basic_correctness.py)：为 BAAI/bge-multilingual-gemma2 的 TestSetting 添加 --gpu-memory-utilization 0.98。
2. 注册结构化输出测试 finalizer (tests/entrypoints/llm/test_struct_output_generate.py)：在 test_structured_output 函数中新增 request: pytest.FixtureRequest 参数，并在创建 LLM 实例后调用 request.addfinalizer(llm.llm_engine.engine_core.shutdown) 确保每个用例结束时主动释放引擎资源。 两项变更均仅涉及测试文件，未影响生产代码。

关键文件：

- tests/entrypoints/llm/test_struct_output_generate.py（模块 结构化输出测试；类别 test；类型 test-coverage）：新增 request.addfinalizer 调用，主动释放 LLM 引擎资源，解决参数化测试用例间 GPU 内存残留问题。
- tests/compile/fullgraph/test_basic_correctness.py（模块 编译测试；类别 test；类型 test-coverage）：为 BAAI/bge-multilingual-gemma2 的 test_setting3 增加 --gpu-memory-utilization 0.98，缓解 PyTorch 2.13 升级后 KV 缓存预算缩减导致的 OOM。

关键符号：未识别

关键源码片段

tests/entrypoints/llm/test_struct_output_generate.py

新增 request.addfinalizer 调用，主动释放 LLM 引擎资源，解决参数化测试用例间 GPU 内存残留问题。

```
# tests/entrypoints/llm/test_struct_output_generate.py 关键变更

@pytest.mark.parametrize(...)
def test_structured_output(
    request: pytest.FixtureRequest, # 新增 FixtureRequest 参数
    backend: str,
    tokenizer_mode: str,
    model_name: str,
    speculative_config: dict[str, Any],
):
    # ... 原有代码
    llm = LLM(
        model=model_name,
        enforce_eager=True,
        max_model_len=1024,
        # ... 其他参数
    )
    # 注册 finalizer：每个用例结束后主动关闭 EngineCore，
    # 避免依赖 GC 回收导致的内存残留（曾导致后续用例 OOM）。
    request.addfinalizer(llm.llm_engine.engine_core.shutdown)
    # ... 后续测试逻辑
```

tests/compile/fullgraph/test_basic_correctness.py

为 BAAI/bge-multilingual-gemma2 的 test_setting3 增加 --gpu-memory-utilization 0.98，缓解 PyTorch 2.13 升级后 KV 缓存预算缩减导致的 OOM。

```
# tests/compile/fullgraph/test_basic_correctness.py 关键变更

TestSetting(
    model="BAAI/bge-multilingual-gemma2",
    model_args=[
        "--runner",
        "pooling",
        "--dtype",
        "bfloat16",
        "--max-model-len",
        "2048",
        # 新增：将 GPU 内存利用率从默认值提高至 0.98，
        # 以应对 PyTorch 2.13 编译路径下 KV 缓存预算减少约 1.66 GiB 的情况。
        "--gpu-memory-utilization",
        "0.98",
    ],
    pp_size=1,
    tp_size=1,
    attn_backend=ATTN_BACKEND,
    method="encode",
```

)

评论区精华

讨论集中于结构化输出测试的 finalizer 方案。njhill 建议优先根除导致 GC 失效的循环引用，而不是用 finalizer 掩盖问题；作者 ZJY0516 认同该观点，但指出不能长期让 CI 保持失败，因此作为临时修复先合并。最终 mgoin 批准合并以恢复 CI 健康，并声明 njhill 后续将处理根本原因。

- 使用 finalizer 还是修复 GC 循环引用 (design): 临时接受 finalizer，后续由 njhill 跟进根本原因修复。

风险与影响

- 风险：风险极低。变更仅影响测试配置和测试清理流程，不涉及生产逻辑。但 addfinalizer 的实现直接调用 EngineCore.shutdown，若该接口因异常抛出，可能导致测试终止时资源未完全释放（但 pytest 会捕获异常并继续执行后续 cleanup）。此外，0.98 的 GPU 内存利用率可能在其他 GPU 型号上仍不充分，需依赖后续运行时内存分析。
- 影响：直接影响：稳定两个频繁失败的内存敏感 CI 测试（distributed-compile 和 structured-output），提升 CI 可靠性。间接影响：提示团队关注 PyTorch 2.13 升级后的内存预算变化及垃圾回收问题。影响范围限于 CI 流程，用户无感知。
- 风险标记：临时方案，测试弱稳定性

关联脉络

- PR #49672 [CI Failure]: distributed-compile - test_compile_correctness[test_setting3]: 本 PR 直接修复该 Issue 描述的编译测试 OOM 问题。