vLLM 周报：2026 年第 22 周（05/25 - 05/31）

vllm-project/vllm

周期：2026-05-25 至 2026-05-31

来源 PR：199 • 重点 PR：24 • 自动生成

原文链接：http://prhub.com.cn/vllm-project/vllm/reports/2026-05-25-to-2026-05-31

执行摘要

本周（2026-05-25 至 2026-05-31）vLLM 仓库共合并 199 个 PR，其中重点 PR 24 个。整体变化主线围绕 MoE 量化 oracle 模块化重构、ROCm 平台性能跃升、Rust 前端能力增强以及 DeepSeek V4 模型深度迭代展开。团队在保持高频迭代的同时，通过大量 bugfix 和代码清理巩固了代码基础。值得关注的风险集中在核心路径变更的测试覆盖缺口和新平台 / 新模型的可维护性上。

本周重点变化

MoE 量化 oracle 模块化重构

由 bnellnm 主导的 MoE 量化后端重构取得重要进展。PR #42647 将 MoeWNA16Method 迁移至 MK oracle，统一后端选择；#42553 将 CompressedTensorsWNA16MarlinMoEMethod 接入 oracle 并新增 FlashInfer Monolithic 支持；#42789 引入 CPU 端 W4A8 oracle 架构。这些 PR 共同构建了可扩展的 MoE 后端选择框架，未来增加新后端将更加简洁。同时 #43108 移除了不再需要的 supports_expert_map，#43727 清理了 inplace 机制，体现了重构与清理并进的风格。

ROCm 平台多重重磅发布

AMD ROCm 团队本周成果丰硕：#41394 为 RDNA3 (gfx1100) 引入原生 HIP W4A16 GPTQ 内核，性能可达 Triton bf16 的 2.5-4.2 倍；#43679 启用 Tilelang 实现的 MHC 内核，覆盖 CUDA 和 ROCm；#42595 修复了 DSv4 上 AITER MxFP4 MoE 的三大 bug；#43898 消除了稀疏注意力中的 GPU 气泡。此外，CI 工作负载从 MI300 迁移到 MI325（#43824），并升级 ROCm 至 7.2.3（#43136）。这些改进显著提升了 AMD 平台尤其是消费级 GPU 的推理竞争力。

Rust 前端能力扩张

BugenZhao 领衔的 Rust 前端持续进化：#43469 引入 mock engine 支持压力测试；#43662 统一工具解析器流式 / 非流式行为；#43670 优化多模态提示扩展实现 7.3x 加速；#43854 新增 /version 端点；#43872 添加 hy_v3 工具解析器；#43850 将 Gemma4 参数扫描性能提升约 600 倍。同时测试体系同步完善（#43582 往返测试）。这些 PR 使 Rust 前端在功能完备性和性能上持续接近 Python 版本。

DeepSeek V4 深度迭代

DeepSeek V4 模型成为本周修改最频繁的模型，涉及近 20 个 PR。关键改进包括：compressor 重构并修复 ROCm 兼容性（#43710）、移除 MegaMoE 和 AMD 路径以简化架构（#43629, #43829, #43891）、融合 Q head padding 消除 F.pad（#43162）、CuTe DSL 稀疏压缩器（#43584）、Move MegaMoE 输入准备 kernel 到独立文件（#43632）。这些变更表明团队正集中力量优化 DSV4 的推理性能与代码结构。

模块与主题趋势

模块	趋势	代表 PR
MoE 重构	量化后端全面 oracle 化，统一选择机制	#42647, #42553, #42789, #43108, #43727
ROCm	原生 kernel 加速、Tilelang 集成、Bug 修复	#41394, #43679, #42595, #43898, #43120, #43781
Rust 前端	功能完善、性能优化、测试增强	#43469, #43662, #43670, #43854, #43872, #43850
DeepSeek V4	重构精简、融合 kernel、 ROCm 修复	#43710, #43629, #43162, #43584, #43905
KV offload	精细化策略、生命周期管理	#43205, #43797, #43870, #39983
新模型 / 解析器	Step-3.7, GLMGA, Cosmos3, MiniCPM5 等	#43859, #43575, #43356, #43175
CI/Infra	arm64 构建、Rust 标签自动标记、测试优化	#41303, #43866, #43824, #43815

风险观察

本周 PR 中高频出现“核心路径变更”和“缺少测试覆盖”风险标签（分别为 26 和 23 次）。主要风险集中在：

- MoE oracle 重构：多个 PR 修改了量化权重处理和后端选择逻辑，但测试仅覆盖主要后端，组合场景（如 TP>1 + spec decode）验证不足。
- 新平台 kernel：RDNA3 内核（#41394）无 CI 覆盖，Tilelang MHC（#43679）对 warp size 硬编码 32，虽经解释但仍存隐忧。
- 废弃流程：JAISLMHeadModel 废弃（#43784）可能影响用户，需确保迁移路径清晰。
- 分布式复杂性：多个 bugfix 针对 KV connector、spec decode 等交叉场景（#43719, #42585），时序依赖风险难以完全通过单机测试覆盖。

重点 PR 速览

1. #41394- [Kernel][ROCm] Native W4A16 kernel for AMD RDNA3：为 gfx1100 实现高性能量化内核，decode 和 prefill 均超越现有方案，是 ROCm 消费级 GPU 推理的重要突破。
2. #42647 & #42553- MoE oracle 重构：统一 WNA16 和 CompressedTensors 量化后端，新增 FlashInfer Monolithic 支持，为后续所有 MoE 量化方法提供可扩展框架。
3. #43205- [KV Offload] per-request 卸载策略：引入 OffloadPolicy 和生命周期钩子，使次级层能按需控制块卸载，是 KV offload 智能化的关键一步。
4. #43469- [Rust Frontend] mock engine：在无 GPU 环境下对前端进行压力测试，基准达 2.1M token/s，是 Rust 前端独立迭代的基础设施。
5. #38445- [PERF] MiniMax-M2 gate kernel：融合 FP32 路由 GEMM 核，低并发下吞吐提升 32%，展示了为特定模型定制 kernel 的极致优化。
6. #43859- [Model] Support Step-3.7-Flash：支持最新多模态 MoE 模型及 MTP 推测解码，per-group slot mapping 设计值得借鉴。
7. #42789- [MoE Refactor] CPU W4A8 oracle：将 CPU 端 W4A8 量化纳入 oracle 架构，推动 CPU 推理现代化。
8. #43575- [feat] GLMGA processor：新增 GLMGA 多模态模型支持，但存在除零和类型安全等问题，需要后续修复。

后续建议

- 加强 oracle 重构的测试覆盖：建议为核心 MoE 后端（Marlin, FlashInfer, Triton）添加跨平台、跨并行配置的集成测试。
- 持续关注 ROCm 稳定性：RDNA3 内核和 Tilelang 虽已合并，但缺少 CI 回归，建议推动 CI 引入 RDNA3 或其他 ROCm 设备。
- Rust 前端功能追赶：继续补齐 Python 前端已有功能（如完整工具调用、reasoning），并关注 mock engine 中已指出的潜在问题（直接索引、忙循环）。
- DeepSeek V4 收敛：本周重构动作频繁，下一步应稳定代码并完善端到端性能基准，确保改动不引入回归。
- 排查废弃模型的用户影响：对 JAIS 等废弃模型提供明确的迁移指南和版本支持说明，降低社区升级成本。