vLLM 2026 年第 18 周周报 (04/27 - 05/03)

vllm-project/vllm

周期：2026-04-27 至 2026-05-03

来源 PR：174・重点 PR：24・自动生成

原文链接：http://prhub.com.cn/vllm-project/vllm/reports/2026-04-27-to-2026-05-03

执行摘要

本周 vLLM 项目迎来高速发展期，174 个 PR 中 24 个被标记为重点。核心主线围绕 DeepSeek V4 的首次完整集成——包含模型架构、量化后端、推测解码的全面落地；同时注意力后端抽象化取得关键进展，vLLM IR 编译框架功能增强，ROCm 平台性能大幅提升。多模态模型家族持续扩大，前端 API 进一步对齐 OpenAI 标准。尽管整体进展迅猛，但测试覆盖和核心路径变更的回归风险仍是团队需要重点关注的方向。

本周重点变化

DeepSeek V4 全面集成

DeepSeek V4 成为本周最受关注的主题，涉及至少 7 个重点 PR。#40860 作为基础模型引入，整合了 MLA 注意力、MegaMoE、MXFP4 量化及 MTP 推测解码，#41083 和 #40950 进一步补充 MXFP4 MoE 后端和数值 clamp，而 #41217（ROCm）和 #41061（多流 GEMM）则针对特定平台优化。这套组合拳标志着 vLLM 对新一代 MoE 架构的完整支持初具雏形。

注意力后端与编译系统重构

注意力系统迎来重要抽象：#32623 将 MLA prefill 后端设计为可插拔，并彻底移除了 cuDNN 依赖，使后端选择更加灵活。vLLM IR 框架通过 #36823 引入 maybe_inplace 重载机制，为 fused 算子和内存优化开辟新路径。这些改动虽然不直接产生用户可见的性能收益，但为未来的架构演进奠定了坚实基础。

多平台性能优化

ROCm 社区贡献活跃：#37646 通过 AITER 融合 Allreduce 与 RMSNorm，在 MI300X 上获得最高 13% 吞吐提升；#41217 针对 DeepSeek V3.2 的 MLA 注意力做了深度优化，引入专用 Triton kernel 和 FP8 稀疏支持。CPU 后端也不遑多让，#39445 首次引入 FP8 KV 缓存量化以减少内存带宽，对 AMX/AVX-512 平台意义重大。

模型支持扩展

本周新增了 Moondream3、MiMo-V2.5、Laguna XS.2、Cohere MoE、EagleMistral 等多个模型，覆盖多模态、推测解码、工具调用等方向。Qwen2.5-VL 的 ViT CUDA 图支持 (#40830) 是该系列模型推理加速的重要一步。

前端与工具调用

prompt_embeds (#40720) 允许在 Chat API 中混入预计算嵌入，丰富了多模态输入手段。Responses API 实现流式工具调用 (#40700, #41110) 并支持 required 和 named 的 tool_choice，向功能完备迈出一大步。system_fingerprint (#40537) 的加入提升了 API 兼容性。

基础设施与代码健康

KV offloading 模块持续重构 (#40538, #39186, #41228)，提升可维护性和功能完备性；大量模型权重加载迁移至 AutoWeightsLoader，代码更统一；CI 增加 eval 测试、减少 flaky 测试、优化 Docker 镜像体积。

模块与主题趋势

- 模型支持：DeepSeek 系列无疑是本周最大热点，从 V3.2 优化到 V4 完整加入，变动涉及底层 kernel、量化、注意力等多个层面。此外，多模态模型如 Moondream3、MiMo-V2.5、Laguna XS.2 的加入，表明 vLLM 正在向更广泛的社区需求开放。
- 性能与量化：FP8/MXFP4 成为性能优化的主流选择，从 ViT 编码器到 MoE 层、从 GPU 到 CPU，量化相关 PR 密度很高（NVFP4、Humming MXFP4 等）。Triton kernel 替代成为普遍模式，带来 5-58 倍加速。
- 基础架构：注意力后端抽象和 vLLM IR 的改进是本周架构演进的两大核心。KV offload 模块也经多次重构，逐渐成熟。
- 多平台：ROCm 本周贡献了多个高 importance 的 PR，覆盖融合运算、DeepSeek 优化、量化修复等。XPU、CPU 也有少量但关键的改进。
- 测试与 CI：CI 改进数量可观，但测试覆盖仍是最大风险点，尤其在新模型和核心路径的 PR 中常被提及。

风险观察

1. 测试覆盖严重不足：本周所有重点 PR 中，“缺少测试覆盖”被提及 34 次，为最高频风险。DeepSeek V4、MiMo-V2.5 等新模型引入大量新代码，但测试用例跟进不足，可能隐藏运行时错误。
2. 核心路径变更风险：32 个 PR 被标记为“核心路径变更”，涉及注意力、调度器、KV 缓存管理等关键模块。这类变更影响面大，一旦出现问题可能影响整个推理链路。
3. 依赖版本稳定性：部分 PR 依赖 tilelang、flashinfer 等库的特定版本，但未在 PR 中明确锁定，可能导致不同环境构建不兼容。
4. 跨平台兼容性：一些针对 SM90+ 的优化（如 NVFP4、FlashInfer FP8）未充分验证软件栈，对其他 GPU 或 ROCm 平台的有效性存疑。
5. 全局配置副作用：如 #41129 (Laguna XS.2) 中直接修改全局 eplb_config，可能影响其他依赖该配置的模型，是潜在的隐蔽风险。

重点 PR 速览

PR #	重要性	摘要
#40860	9.5	DeepSeek V4 初始集成：引入基础模型、MLA、MegaMoE、MXFP4 量化及 MTP 推测解码
#32623	9.4	MLA prefill 后端抽象化，移除 cuDNN，实现可插拔选择机制
#36823	9.4	vLLM IR 添加 may_inplace 重载，支持 fused_add_rms_norm 等内存优化
#39186	9.4	逐作业 KV 卸载完成通知，加速前缀缓存重用，提升 CPU offloading 效率
#41083	9.2	Humming MXFP4 MoE 后端，DeepSeek V4 推理性能提升超 40%
#40720	9.2	Chat API 支持 prompt_embeds 内容部分，扩展多模态输入
#38065	9.2	ViT 注意力 FP8 量化加速，长序列场景提速 1.2x
#40967	9.2	MiMo-V2.5 系列模型支持，覆盖 Omni 和 MTP 推测解码
#39445	9.2	CPU 后端 FP8 KV 缓存量化，降低内存带宽需求
#34668	9.2	思考预算移出 LogitsProcessor，兼容投机解码

后续建议

1. 补齐测试覆盖：优先为 DeepSeek V4 和相关量化后端添加端到端和单元测试，确保多 GPU、EP 等场景下功能正确。
2. 回归验证核心模块：注意力后端抽象和 IR 编译管道改动影响深远，建议在 CI 中增加更全面的 benchmark 和正确性测试。
3. 统一量化后端注册：Humming、NVFP4、MXFP4 等后端逐步集成，建议标准化注册流程和配置接口，降低使用门槛。
4. 关注全局副作用：新模型应避免对全局配置的副作用，推动模型负载隔离。
5. 持续迭代跨平台支持：ROCm 性能优化需要在高并发场景下进一步验证；CPU FP8 attention 应加入更多模型测试。
6. 依赖版本管理：对高频依赖如 tilelang、flashinfer 应明确版本下限，避免构建破碎。