

vllm-project/vllm 周报 · 2026 第 37 周 (09-07 至 09-13)

vllm-project/vllm

周期: 2026-09-07 至 2026-09-13

来源 PR: 331 · 重点 PR: 24 · 自动生成

原文链接: <http://prhub.com.cn/vllm-project/vllm/reports/2026-09-07-to-2026-09-13>

1. 执行摘要

本周 vllm-project/vllm 共合并 331 个 PR, 其中 24 个为重点 PR, 平均重要度 6.37。变化集中在三条主线: DeepSeek V4/V4.1 在 CUDA、ROCm、CPU 与前端四层同步铺开; 去 JIT 预热 (warmup) 迁移系列推进到第 2/3/4/5/6/9 个 PR; ROCm/AITER 集成与 KV Connector、调度器修复保持高强度。与此同时, 核心路径变更以 48 次高居风险类别首位, 并伴随 MRV2 预热回滚等操作, 说明主线代码仍处高波动期。本周管理重点应是降回归风险, 而非单纯庆祝功能落地。

2. 本周重点变化

2.1 DeepSeek V4/V4.1 全平台铺开

DeepSeek 系模型本周横跨四个层次落地: [Add DeepSeek-V4 CPU backend](#) (#55355) 把稀疏 MLA、Lightning Indexer、mHC 门控移植为 AVX512/AMX 内核并解耦 CPU 后端; [Enable DeepSeek V4 Vision](#) (#55107) 把视觉包装类上移到平台中立的 common 层, 并在 ROCm 打通多模态前缀注意力; V4.1-Flash 则一次性补齐模型定义 (#56228)、Rust/Python 双栈前端 (#56208)、Engram DP 分片与异步预取 (#56512)、输入元数据 Triton 融合 (#56562, decode TPOT 约降 15%)。这是一条从能跑走向多平台、低延迟、可扩展的完整链路, 建议作为下一阶段资源投放重点。

2.2 去 JIT 预热框架高速迭代, 同时出现回滚

warmup 系列本周合并了第 2/3/4/5/6/9 号 PR (#53564、#53566、#50176、#50178、#56323), 把序列 /DCP、FA4 MLA、CuTeDSL 注意力、MHC TileLang 与采样 /DFlash 内核统一迁入 [VllmTritonJitKernel](#) / [VllmCuTeDSLJitKernel](#) 声明式契约。多轮 review 暴露两个系统性坑: warmup 张量 dtype 与运行时不一致会让预热失效并重新触发 JIT; 预热覆盖不全则表现为请求时 10-20 秒停顿 (#56452)。配套的 Qwen 预热扩展 (#54797) 也同周落地。但 #56654 明确回滚了 MRV2 专属迁移、恢复直调内核, 说明该框架仍在收敛, 不宜视为完全稳定。

2.3 注意力后端架构扩张与抽象收紧并行

新增可复用的复合后端 [composite.py](#), 把图像掩码批次路由到 Triton、纯因果查询路由到 FlashInfer/FA4 (#56305); 稀疏 MLA 模型开启 PCP+DCP 联合 (#56157), 明确所有 rank 派生一致 collective 形状的不变式; HiSparse 落地 host-resident 稀疏 MLA 缓存层 (#53781), 并补齐 TP rank 共享主机池 (#56629) 与缓存指标 (#56061); XQA 解码扩展

到 SM90 (#50439) 。与此同时抽象层开始收紧默认值，例如注意力实现的 DCP 支持改为显式声明 (#55780) ，后端选择优先级成为 review 焦点。

2.4 水印功能从 0 到 1

原生 Gumbel-max 文本水印 (#54053) 与双钥投机解码水印 (#56122) 在三天内相继合并，覆盖生成、检测、Triton 融合采样内核与预热。设计上做了明确取舍：Philox 保留为唯一 PRF (HMAC-SHA256 因 GPU 性能过慢移除) ，密钥域分离、敏感字段 `exclude=True` 防护到位。该功能开启推理引擎内置可检测性的新方向，但低检测率成因、密钥安全边界与采样路径性能仍需继续观察。

3. 模块与主题趋势

本周标签分布高度集中在修复与性能两条线：

主题	数量	解读
bugfix	134	约四成 PR 是修复，主线仍在补稳定性欠账
performance	61	内核、预热、元数据融合是主要抓手
v1	59	v1 架构持续推进
rocm	54	ROCm 平台工作量突出，仅次于 v1
ci/build	51	CI 门禁、超时与平台跳过密集调整
test	46	测试补强与平台适配并行
kernel	38	含 DeepSelect 等第三方内核集成
kv-connector	34	KV 传输 / 卸载为第二大功能主线

热点文件前十中，`scheduler.py` (5 次)、`single_type_kv_cache_manager.py` (5 次)、`model_runner.py` (5 次)、`config/vllm.py` (5 次)、`_aiter_ops.py` (6 次) 全部属于核心路径或跨平台耦合点，说明改动重心正从新增模型转向调度、缓存与平台抽象的深水区。作者分布上，khluu 与 AndreasKaratzas 各 13 个 PR，taneem-ibrahim 11 个，LopezCastroRoberto、BugenZhao、yewentao256 各 7 个——CI/ 平台侧与内核 warmup、Rust 前端化、KV/ 调度修复分别由固定几人长期投入，属于典型的关键人依赖结构，需要关注单点风险。

4. 风险观察

- 核心路径变更 (48 次) 为最大风险源：注意力、调度器、量化与解码头同时被改动，且一周内多次回滚 (#56654 MRV2 预热、#56429 KimiK3 DCP、#56546 agent 文档) ，回归窗口叠加。
- 测试与 CI 覆盖被跳过稀释：缺少测试覆盖 13 条；XPU/ROCm 多个用例被临时跳过 (#56463、#56555、#56038、#56179) ，Torch ABI 与 MI300 超时翻倍，真实退化可能被掩盖。

- 3. 预热契约失配风险未根除：#50176、#56323 讨论均指出 dtype 与键推导不一致会导致预热无效；MRV2 回滚后该路径恢复运行时 JIT，首请求延迟存在回归可能。
- 4. 第三方内核与依赖契约耦合加剧：DeepSelect、AITER、Mooncake、TileLang、CUTLASS DSL 4.7 等引入版本与 ABI 依赖，ROCm 侧 gfx950 专用路径尤其明显。
- 5. 部分修复留尾：PP 场景下 Engram hash 解引用、HiSparse NIXL 失败失效范围、CUDA/common 侧 SWA clamp 缺陷等被明确留在后续 PR，需跟踪闭环。

5. 重点 PR 速览

PR	主题	为什么重要
#56305	复合注意力后端	引入可复用 composite 机制，后端优先级顺序成为核心讨论点
#56512	Engram DP 分片 + 异步预取	V4.1 性能与内存优化关键，但 PP 边界与 CI 覆盖被 reviewer 质疑
#56464	DeepSelect TopK	DSA 解码 top-k 数倍提速，建立第三方内核合规接入范式
#56157	PCP+DCP 稀疏 MLA	修复集体通信挂起，同时暴露 PIECEWISE 下 decode 性能退化
#54053 / #56122	Gumbel 水印 + 双钥投机水印	水印功能从生成到投机解码补齐闭环
#54985	Elastic EP 复用 CUDA graph	16 到 24 卡停机时间从 40 秒降至近零，扩容不再打断服务
#55355	DeepSeek-V4 CPU 后端	AVX512/AMX 内核加模块化解耦，CPU 推理能力补全
#53781 / #56629	HiSparse	host-resident 稀疏 MLA 层与 TP 共享主机池，KV 分层架构关键一步
#56654	MRV2 预热回滚	反向信号：预热框架尚未稳定的直接证据
#55522 / #56433	RDNA3 与 AITER 修复	ROCm 量化迁移与静默数值错误修复，属于隐性正确性风险

6. 后续建议

1. 把 DeepSeek V4/V4.1 作为跨团队专项跟踪：模型定义、CPU/ROCm 后端、Engram、前端四条线已同周落地，建议建立统一的冒烟与精度门禁，避免各平台各自为战。
2. 为 warmup 契约补充一致性校验：在注册期自动核对 dtype 与编译键推导，防止预热了但仍然编译的静默失效；同时明确 MRV2 路径的短期方案与长期目标。
3. 收紧 CI 跳过与超时策略：对本周新增的临时 skip 建立到期清理机制，并把关键平台（XPU、ROCm、B200）的覆盖缺口登记为显式技术债。
4. 高风险改动要求配套回滚预案：注意力后端选择、调度器抢占、量化内核选择等核心路径变更建议强制附带回滚开关或对照基准。
5. 统一第三方内核接入规范：DeepSelect 的 CMake FetchContent 加 stable ABI 模式可作为范式，但需要统一版本契约、失败兜底与精度回归规范。