

【简摘】AI 情报速览 | 2026-10-12

概览：本期焦点是 Agent 的“自我改进”与“长周期运行”：Sakana AI 的 MASS 把多 Agent 团队的经验训练回模型；Decision 3.0 与 AesCode 两个开放权重模型家族上线；Qwen Code 和 OpenClaw 都在补持久化与重启恢复。

标签：【官方确认】【媒体报道·待确认】【社区项目·作者自测】。时间均为北京时间。

1. Sakana AI 发布 MASS 【官方确认】

- 同一个 Qwen3.6-27B 同时担任执行者、优化者和评审者，再用 LoRA 学习团队的优秀轨迹。两轮后，测试任务对底模的胜率为 69.9%。
- 研究类基准的单位 Token 得分为底模的 1.2—1.6 倍，软件工程类基准为 0.94—0.96 倍。代码未指定开源许可证，不发布权重。
- 来源：[arXiv](#)

2. Decision 3.0 六模型上线 【官方确认】

- 参数量 0.59B 至 26.09B，Apache 2.0。一次调用回答多选、是非、评分问题并直接返回概率。d3-edge 纯文本延迟 6.7 ms (MI325X)。
- 小模型并非全面领先：d3-edge 的新领域分低于上一代。权重于 10-10 至 10-11 上传。
- 来源：[HF Collection](#)

3. 微软 AesCode-8B/32B 【官方确认】

- 以生成图为版式参考，输出可编辑的 HTML 信息图。300 样本评测中 32B 得 85.89，GPT-5.5 带参考图为 81.28；优势主要来自程序化检查一侧。
- 权重 Apache 2.0，代码 MIT，训练数据未公开。
- 来源：[GitHub](#)

4. llama.cpp b11556 接入 Bonsai 2 旋转基 【官方确认】

- 10-11 21:31 发布，上游加入 Hadamard 旋转运行时。上游尚无 PTQ1_0 / PQ2_0 打包类型，官方 5.95 GB 文件仍需 Prism fork。Bonsai 2 于 9-17 发布。
- 来源：[PR #29600](#)

5. llama.cpp b11557—b11559 【官方确认】

- 内容为 RPC 分级调试日志、OpenCL 的 Q4_K / rms_norm 修复，以及 Gemma 4 GPU 解码与 DK512 / DK64 / DK128 优化。后两个构建于 10-12 凌晨发布。

- 来源: [b11559](#)

6. Qwen Code v0.25.1 正式版【官方确认】

- 10-11 18:20 发布, Desktop v0.25.1 与 SDK v0.1.19 同日跟进。重点是 Managed Agent 的持久记录、多 Agent 会话(预算、取消、团队), 以及实验性 K8s Runtime。
- 来源: [Release](#)

7. OpenClaw v2026.10.5-beta.1【官方确认】

- 10-11 21:08 发布。Gateway 重启后会话可以恢复, 读取移到 Worker 线程, 修复 SQLite 死锁; 新增 Worker 本地推理和 X 提及插件。
- 来源: [Release](#)

8. Unsloth Qwen-Image-2.1-Turbo 量化【官方确认】

- INT8-ConvRot 版 7.26 GB, LPIPS 0.037, 为官方推荐; FP8 版 0.125。采用 Qwen Research License, 非 Apache。仓库创建于 10-10。
- 来源: [HF](#)

9. OpenJet Megakernel 与 FineEnvs【社区项目·作者自测】

- OpenJet 在 RTX 3090 上运行 Qwen3.8-27B: 写代码 140 vs 73 tok/s, 提升主要来自更多草稿 Token; 许可证 AGPL-3.0-only。
- FineEnvs 把 405 个 Lean 命题做成 369 个 Agent 环境, 许可证 Apache-2.0。
- 来源: [OpenJet](#)

10. 英国加强 AI 相关网络防御【媒体报道·待确认】

- 《卫报》10-12 报道: AI 大臣 Narayan 将呼吁前沿 AI 公司加强安全, 并称不能完全依赖企业自律。属于政策方向, 并非已生效法律。
- 来源: [The Guardian](#)