

AI前沿发展日报 | 2026-09-15 (Asia/Shanghai)

日期：2026-09-15；覆盖窗口：2026-09-14 06:49 至 2026-09-15 06:50 (Asia/Shanghai)；以已公开且可核验的信息为准，窗口外内容明确标为近期补充；仅公布海外日期的公告不推定具体北京时间。

今日要点

- 产业新增事实开始落在交付条件上。小马智行与广汽商用车发布新款无人重卡，富士通给出 MONAKA 处理器销售及服务器后续交付安排；两者都是商业化进展，尚不能把规划中的规模和降本效果当成兑现结果。小马智行公告 (<https://ir.pony.ai/zh-hant/news-releases/news-release-details/pony-ai-inc-unveils-new-gen-4-robotruck-collaboration-gac>)；富士通新闻稿转载 (<https://www.hpcwire.com/off-the-wire/fujitsu-sets-november-launch-for-fujitsu-monaka-cpu-and-server/>)
- 基础设施并购瞄准冷却、配电与水的组合。AirJoule 收购 BitSink，首期对价与未来收入挂钩的追加对价分开设置；中期看点是设备交付和业务整合，而非交易标题中的最大金额。公司新闻稿转载 (<https://www.hpcwire.com/off-the-wire/airjoule-acquires-bitsink-to-expand-into-ai-and-hpc-data-center-infrastructure/>)
- 微软把模型行为要求交给公众审阅。新增事实是 MAI 行为准则草案和六周征询期，不再重复此前“行业应否放缓”的争论。Microsoft AI 原文 (<https://microsoft.ai/news/mai-code-of-conduct/>)
- 可核实的当日 X 原帖信号偏少。本期该栏以五项一手产品条款和采用数据补充，逐条注明验证边界；涉及此前发布的信息，不重标为今日首发。

今日结论

1. 把实体 AI 的采购验收放到完整运营账上。硬件物料降价、芯片吞吐提升都只是中间变量；运输业务要看接管、维护与车辆利用率，推理业务要看实际模型、机房条件和可交付时间。
2. 优先评价已经完成的业务，而不是代理活动总量。工作单元、会话数和消息数能说明使用规模，但不能直接证明任务成功、收入增长或人力节省；下一步应把它们与工单关闭、成交和返工记录关联。
3. 将新增控制能力转成可执行的配置与验收项。检索请求可关闭哪些收费附加项、模型能否响应中止、旧模型是否继续供应，均应落实到实际调用与行为检查，避免仅凭公告标题调整系统。

AI 产品与应用

小马智行与广汽商用车发布新重卡，量产仍是年内计划

窗口内新增 | 公告发布于北京时间 2026-09-14 20:00。新车型基于广汽商用车 T9 纯电重卡平台，接入小马智行第四代 L4 自动驾驶系统，计划于 2026 年内开始量产，面向干线货运、专线物流和港口运输。公司投资者公告 (<https://ir.pony.ai/zh-hant/news-releases/news-release-details/pony-ai-inc-unveils-new-gen-4-robotruck-collaboration-gac>)

为什么重要：公司披露自动驾驶套件物料成本较上一代下降 70%，并采用与第七代 Robotaxi 相同的域控制器。这里的实质进展是车型与制造平台结合，以及部件复用；不能把套件降价外推为整车成本同比例下降。

商业影响：物流企业可以据此启动特定线路的运营测算。公司预计每吨公里运输成本降低 30%，但这是前瞻估计，未提供本期可独立验证的规模运营结果；人员配置、接管频率和维护支出仍会影响实际收益。

证据边界：已核验公司公告的实时搜索正文；原网页直接读取受限。产品计划和成本数据均为公司披露，不等于已取得所有地区的无人运营许可。

模型与技术进展

富士通 MONAKA：为受供电和散热限制的机房提供 CPU 推理选项

窗口内新增 | 公司新闻稿日期 2026-09-14。富士通宣布 MONAKA CPU 从 2026 年 11 月开始销售，并规划搭载该处理器的服务器。新闻稿强调单靠 CPU 构建推理平台、适配风冷机房；性能主张包括最高 3.8GHz、8800MT/s 内存传输速度，以及相较其他 CPU 的两倍 AI 推理吞吐。富士通新闻稿，HPCwire 完整转载 (<https://www.hpcwire.com/off-the-wire/fujitsu-sets-november-launch-for-fujitsu-monaka-cpu-and-server/>)

为什么重要：这条路线争取的是不能轻易扩电或改造液冷的本地部署场景。它扩大了推理硬件的候选范围，但上述比较未在所读新闻稿中充分列出模型、精度与对照平台，不能推导为普遍超越 GPU。

商业影响：金融、制造等需要本地处理数据的团队，可要求供应商以自身模型做容量和延迟验证。公告同时列出 2026 年 11 月销售安排与 2027 年 4 月起日本、欧洲服务器陆续发货安排；销售启动与实际交付必须分开核对，本期不将其写成现货可用。

验证状态：依据署名富士通的公司新闻稿转载，属于一手声明的分发版本；未进行硬件实测。

投融资与商业动态

AirJoule 收购 BitSink：2,700 万美元首期对价，追加付款取决于收入

窗口内新增 | 公司公告日期 2026-09-14。AirJoule 宣布收购 BitSink 全部股权，首期支付 1,800 万美元现金及价值 900 万美元的普通股；交割后连续三个日历年的收入里程碑，还可能触发最高 4,000 万美元的股票对价。AirJoule 新闻稿，HPCwire 转载 (<https://www.hpcwire.com/off-the-wire/airjoule-acquires-bitsink-to-expand-into-ai-and-hpc-data-center-infrastructure/>)

为什么重要：BitSink 提供冷却、配电和机架设备；AirJoule 希望把这些能力与利用低品位余热从空气中制水的技术结合。收购对象是可制造和交付的基础设施业务，商业落点较“新增算力愿景”更具体。

商业影响：对数据中心运营商，价值取决于组合系统能否降低现场工程复杂度，并稳定提供冷却和水。对交易观察者，首期 2,700 万美元与最高合计 6,700 万美元应严格区分：后者包含尚未兑现的业绩条件，不是已经全部支付的收购款。

验证状态：金额与交易结构已由收购方新闻稿核验；整合后的性能、客户增长与协同收益仍属于未来兑现事项，本期不采用新闻稿中的行业市场规模预测。

政策、伦理与安全

MAI 行为准则进入六周公众征询，重点是可中止和不自行扩权

窗口内新增 | 官方公告日期 2026-09-14。Microsoft AI 发布 MAI 模型行为准则首份草案，开放六周反馈，并计划在 2026 年内发布修订版。其要求包括不抵抗人类中断、纠正或关闭，不自行扩大工作范围、不接受无人赋予的目标，以及不向审计人员隐藏推理。官方征询公告 (<https://microsoft.ai/news/mai-code-of-conduct/>)

为什么重要：草案把“人类保持控制”细化为可讨论的模型行为，而不是停留在行业原则。新增的是公开文本、反馈入口和修订流程；这不是政府法规，也不是对当前所有产品行为的实测保证。

商业影响：使用 MAI 的企业可把中止后的动作、权限范围和审计可见性纳入供应商评估，并针对真实业务冲突提交意见。适用对象是微软自研 MAI 模型，不能据此推定 Azure 托管的所有第三方模型采用相同规则。

X 平台高信号

本栏原帖核验不足，以下为一手来源补充，不能视作五条已验证的 X 发帖。只保留可核验的访问、计费、采用数据与统计口径变化；“为什么重要”和“影响”为本报分析。

1. Agent Search 可按请求关闭收费附加项

信号标题：查询级附加项控制进入正式 API

类型：计费控制 / 正式可用

来源或链接：Google Cloud 发布说明 (<https://docs.cloud.google.com/generative-ai-app-builder/docs/release-notes>)

核心观点：2026-09-14 发布说明确认，searchAddonSpec 在 v1 API 正式可用，可为单个程序化搜索请求关闭附加项。

为什么重要：团队可以按任务需要决定是否使用附加能力，无须对全部查询采用同一配置。

影响：高频简单查询可单独设置；节省幅度取决于原有配置与流量，不等于搜索基础价格普遍下降。

验证状态：已直接核验官方发布说明；未核验对应 X 原帖，也未实测账单变化。

2. DeepSeek V4 Pro 继续供应，不能沿用自动切换传闻

信号标题：原定切换日期已过，当前官方条款仍保留 Pro

类型：模型访问 / 供应安排纠偏

来源或链接：DeepSeek 当前模型与价格页 (https://api-docs.deepseek.com/quick_start/pricing/)

核心观点：截至 2026-09-15，官方明确在 2026-09-14 之后继续提供 V4 Pro API，计费方式保持不变；模型表仍将其列为 DeepSeek-V4-Pro-0813。此前流传的 Pro 全量转到 V4.1 Flash 安排不应再用于采购判断。

为什么重要：旧模型名继续可用，与后台实际换成另一个模型，是两种不同的兼容性状态。

影响：依赖 Pro 的任务无需仅因原定日期而迁移；若主动改用 Flash，应重新验收输出与费用。

验证状态：当前一手条款已直接核验；官网未列本次修订的精确时刻，因此不宣称撤回决定发生于今日；X 原帖未直接核验。

3. Salesforce 披露 70 亿工作单元，规模口径不能等同已完成订单

信号标题：Agentforce 与 Slack 的使用规模有了量化披露

类型：采用数据 / 公司自报

来源或链接：Salesforce 2026-09-14 公告 (<https://www.salesforce.com/ap/news/press-releases/2026/09/14/sg-salesforce-expands-agentforce-with-a-new-portfolio-of-ai-agents-built-for-high-value-work/>)

核心观点：公司披露累计交付 70 亿 Agentic Work Units (AWU)，其中第二季度为 32 亿；同日发布按岗位组织的代理组合，但跨数周推进销售工作的 Hunter 仍在试点，正式可用计划为 2026 年 11 月。

为什么重要：规模数据证明公司开始提供活动量口径，但不能直接换算成独立客户数或完整业务任务数。

影响：企业选型应分别核对产品可用阶段、AWU 计量定义和自己的业务完成率，避免把总量当成收入贡献。

验证状态：已直接核验公司公告；数据未经独立审计，本期未核验对应 X 原帖。未将 Hunter 描述成已经全面商用。

4. GitHub 新增独立 Agents 窗口的使用统计

信号标题：代理窗口活跃用户与会话数可单独追踪

类型：采用统计 / 数据访问

来源或链接：GitHub 官方更新 (<https://github.blog/changelog/2026-09-11-add-vs-code-agents-to-copilot-usage-metrics/>)

核心观点：近期补充，发布于 2026-09-11。Copilot 指标增加 VS Code Agents 独立窗口的日活用户、会话与用户消息统计，覆盖企业和组织层面的 1 日、28 日报告。

为什么重要：这些统计与编辑器内 Agent Mode 分开，口径不一致会让内部采用率比较失真。

影响：管理员可分别跟踪不同入口；字段缺失不能直接解释为零使用，也不能用消息数量衡量开发产出。

验证状态：已直接核验一手更新；需要启用使用指标政策并具备相应查看权限。X 原帖未直接核验，非今日首发。

5. Kiro 学生访问扩围，额度不可结转或额外购买

信号标题：新增大学覆盖与一年期免费访问有明确条件

类型：访问资格 / 免费额度

来源或链接：Kiro 扩围公告 (<https://kiro.dev/blog/students-2026/>)；学生计划正式条款 (<https://kiro.dev/students/terms/>)

核心观点：近期补充，公告与条款更新于 2026-09-08。计划新增覆盖 16 个国家的 121 所大学；合资格学生每月获得 1,000 credits，持续 12 个月，需年满 18 岁并验证在读身份。

为什么重要：这降低了学习和试做项目的门槛，但不是无限量的生产服务。

影响：课程安排须考虑额度不结转、期间不能另购额外 credits，以及到期回落至免费层；学校参与不代表学校背书。

验证状态：已直接核验一手公告及正式条款；未验证具体个人资格或对应 X 原帖，不将资格覆盖数写成活跃学生数。

前沿研究速递

RD-Forget：保留历史，但控制过时信息何时参与回答

近期补充 | 初稿提交于北京时间 2026-09-09 22:48，非当日新论文。arXiv 原文 (<https://arxiv.org/abs/2609.10263>)

做了什么：保留原始观察记录，为每个问题构建单独的内存视图；回答当前状态时抑制被替代的旧事实，查询历史时仍允许取回。

新在哪里：不靠直接删除历史解决事实更新，而是把“存着什么”与“这次使用什么”分开；无需重新训练模型，并保留多跳推理需要的关联。

潜在应用方向：客户资料更新、合同版本查询、长期个人助手，以及同时回答现状与历史的问题库。

一句话判断：它提出了实用的内存管理方法，但回答时抑制旧信息不等于从存储中删除数据；效果仍需在真实业务上复现。

织物识别的选择性模型调用：把昂贵推理留给难例

近期补充 | 初稿提交于北京时间 2026-09-09 19:40，非当日新论文。arXiv 原文 (<https://arxiv.org/abs/2609.10065>)

做了什么：让卷积网络处理多数织物结构图片，仅在难例中调用视觉语言模型，并把后者的选择限制在符合织物分类体系的三个候选项内。

新在哪里：作者在自建 14 类基准上报告，将视觉语言模型触发比例从 60% 压至不足 10%，API 成本约降 90%，未测得准确率损失；这是该实验设置中的结果。

潜在应用方向：纺织供应链资料结构化、面料入库识别和需要控制调用费用的工业视觉系统。

一句话判断：有价值的是把调用条件与专业分类约束一起设计；迁移到其他工厂前，仍须验证不同光照、织物和采集设备下的错误率。