

AI 前沿发展日报 | 2026 - 07 - 19 (Asia)

日期：2026 - 07 - 19；覆盖窗口：2026 - 07 - 18 00:00 至 2026 - 07 - 19 00:00 (Asia)，仅纳入截至发稿时经官方页面、一级媒体、独立评测或原始论文重新核验的事实。

今日要点

周末新增信号不多，但指向同一件事：AI 的价值判断正从发布声量转向真实流量、独立评测与资本市场反馈。OpenRouter 的周度 token 使用量前五名已全部由中国开放权重模型占据；Thinking Machines 的 Inklings 则在 ARC Prize 独立评测中击败所有模型，成为唯一进入前二的开放权重模型。二级市场随之重新计价高投入路线，Meta 单日下跌 5.2%，NVIDIA 下跌 2.2%。政策侧，美国两党议员首次尝试把致命自主武器的人工批准写入联邦法律，而司法系统已经用处罚和核验要求把生成式 AI 的责任落到具体使用者。中长期看，模型竞争正在同时接受三张成绩单：使用量、可验证能力和风险责任。

今日结论

1. 开放权重模型的竞争已从“能否追平”进入“能否拿走工作负载”。企业采购应把真实任务流量、单位成功成本和可控部署权放到同一张评估表上。
2. 独立评测开始改变新模型的可信度分配。Inklings 的 ARC-AGI 成绩说明美国开放模型也在追赶，但尚未覆盖 ARC-AGI-3，不能把单一榜单外推为全面领先。
3. AI 的责任边界正在由原则转成操作要求：武器目标需非 AI 来源复核、诉讼材料需人工核验。高风险行业应预先设计“谁批准、用什么证据复核、如何留痕”。

AI 产品与应用

开放模型已经取得可见的开发者工作负载。Axios 7 月 18 日核对 OpenRouter 周报后指出，按 token 使用量计，腾讯、小米、DeepSeek、MiniMax 与智谱的模型占据前五，且五者均为开放权重模型；OpenRouter 自有研究显示，中国开放模型的周度份额曾由 2024 年末约 1.2% 升至接近 30%。这不是总体 AI 市场份额，也不包含大量厂商直连和订阅流量，但它至少说明，在可自由路由的开发者场景中，低价、可定制和可自托管已经转化为真实调用。对企业而言，默认把最贵模型用于所有任务的做法正在失效，更合理的架构是按任务难度分层路由，并持续监控成功率、延迟与返工成本。来源：Axios 7 月 18 日报道 (<https://www.axios.com/2026/07/18/china-ai-open-router>)、OpenRouter State of AI (<https://openrouter.ai>)

模型与技术进展

Inkling 获得首轮独立推理评测验证。ARC Prize 于 7 月 17 日更新结果：Thinking Machines 的 9750 亿参数 MoE 模型 Inkling 在 ARC-AGI-1 与 ARC-AGI-2 中分别取得 55% 和 36.5%，平均每题成本分别为 0.30 美元和 0.64 美元，是该机构当前测过的开放重模型最高分；ARC-AGI-3 尚未评测。与前一天仅讨论模型卡不同，今天的新增变量是第三方给出了任务级结果和成本。Inkling 采用 Apache 2.0、410 亿激活参数并支持最多 100 万 token，上游可定制性明确；但其体量仍要求昂贵部署，企业应先验证自有任务收益，而不是把“开放”误解为“便宜”。来源：ARC Prize 验证结果 (<https://arcprize.org/results/thinky-inkling>)、Thinking Machines 官方发布 (<https://thinkingmachines.ai/news/introducing-inkling/>)、模型卡 (<https://thinkingmachines.ai/model-card/inkling/>)。

投融资与商业动态

资本市场开始为“更便宜的同级能力”重估高投入 AI 交易。7 月 17 日美股收盘，S&P 500 下跌 1.0%，纳斯达克下跌 1.4%；Meta 跌 5.2%，NVIDIA 跌 2.2%。归因于投资者重新审视拥挤的半导体交易和巨额 AI 资本开支，Kimi K3 的低价开放路线进一步放大担忧；AP 也指出，市场在质疑芯片与内存需求能否由足够的利润和生产率回报支撑。不能把单日跌幅解释为 AI 投资周期反转，但估值逻辑正在变化：模型能力若更快商品化，基础设施支出就必须用利用率、外部收入和应用毛利来证明。来源：Reuters (<https://in.marketscreener.com/news/wall-st-slides-after-ai-7f51dada89ff25>)、AP (<https://apnews.com/article/67f5a>)。

政策、伦理与安全

美国两党议员把致命自主武器的人工控制推进到法案文本。众议员 Don Beyer、Tom Barrett 与 Sara Jacobs 于 7 月 17 日提出《Human Authority over AI Act》，要求美军任何有意致命的自主或 AI 武器使用都须有人类监督、批准或处于决策回路中；法案生效后前五年，指挥官还用非 AI 来源核验 AI 生成目标，导弹防御系统除外。它比五角大楼现有“适当程度的人类判断”政策更具体，也更强，因为目标是写入联邦法律。对国防承包商和其他高风险 AI 供应商，真正可迁移的要求是双重证据、明确批准者与可审计记录。法案目前只是提案，最可能在年度国防授权法案中继续博弈。来源：Axios (<https://www.axios.com/2026/07/17/bipartisan-ai-weapons>)。

美国法院正在以个案方式建立 AI 使用责任。Axios 7 月 18 日汇总称，Northwest 对 112 名法官的调查中，超过 60% 使用过至少一种 AI 工具，22.4% 每周或每天使用；与此同时，密歇根一名联邦法官本周认定政府在移民诉讼文件中很可能引用了 AI 生成的虚假案例，并要求今后不得再提交不存在的判例。商业含义不是法律行业会否采用 AI，而是采用已经发生、核验制度却落后。律所、合规和公共部门应把引用溯源、机密信息隔离和最终签署人责任设为上线门槛。来源：Axios 7 月 18 日报道 (<https://www.axios.com/2026/07/18/federal-judge-ai-fraud>)。

2026/07/18/ai-lawyers-judges-balancing-act)、Northwestern.edu/stories/2026/03/northwestern-number-of-federal-judges-are-already-using-ai-to-: // f i n g f x . t h o m s o n r e u t e r s . c o m / g f x / l e g a l d o c s / m o v a z w r % 2 0 - % 2 0 W D % 2 0 M i c h % 2 0 2 0 2 6 0 7 1 6 . p d f) 。

X 平台高信号

1. 信号标题：OpenRouter 周度 token 使用量前五名全部来自中国开放模型

类型：硬采用数据

来源或链接：<https://www.axios.com/2026/07/18/china-ai-cpic-openai>

核心观点：7 月 18 日新增事实是，腾讯、小米、DeepSeek、MiniMax 与智谱已占据 Router 周度 token 使用量前五名；这比继续讨论某一款新模型的榜单名次更能反映开发者实际选择。

为什么重要：在可自由切换供应商的平台上，开放权重和低成本已经从产品卖点转化为调用量。

影响：企业模型路由应增加开放模型候选，但必须把 OpenRouter 样本与全市场份额区分开，不能忽略厂商直连和订阅流量。

验证状态：排名由 Axios 核对，长期趋势由 OpenRouter 官方数据支持；不代表整体企业 AI 市场份额。

2. 信号标题：Inklings 成为 ARC Prize 当前得分最高的开放权重模型

类型：独立能力与成本验证

来源或链接：<https://arcprize.org/results/thinky-inklings>

核心观点：ARC Prize 7 月 17 日发布后新增验证：Inklings 在 ARC-AGI-1-2 得分 79.5% 和 36.5%，平均每题成本为 0.30 美元和 0.64 美元。

为什么重要：独立任务级评测把厂商模型卡中的能力声明转化为可比较的成绩与成本。

影响：企业可把 Inklings 纳入可定制模型短名单，但它尚未测试 ARC-AGI-3，且推理硬件总成本仍需单独测量。

验证状态：ARC Prize 已验证两项成绩；ARC-AGI-3 未评测，其他企业任务表现仍待独立复测。

3. 信号标题：Grok Build 开放源代码并允许完全本地推理

类型：开发工具访问变化

来源或链接：<https://x.ai/news/grok-build-open-source>

核心观点：SpaceX AI 公开了 Grok Build 的 agent loop、工具调用、终端系统代码；用户可自行编译，并在 config.toml 中接入本地推理，而不必绑定 Grok 模型。

为什么重要：开放的不只是模型权重，而是决定上下文装配、工具分派和执行边界的 code

ng agent harness。

影响：工程团队可审计并改造执行逻辑、接入私有模型，但需要自行承担沙箱、权限、升级兼容和供应链安全。

验证状态：源代码与本地优先模式已由 SpaceX AI 官方页面和 GitHub 仓库核实；生产稳定性需使用方验证。

4. 信号标题：美国两党法案要求 AI 武器目标用非 AI 来源复核

类型：军事 AI 规则变化提案

来源或链接：<https://www.axios.com/2026/07/17/bipartisan-right-ai-weapons>

核心观点：新法案不要求致命武力有人批准，还规定生效后前五年必须以非 AI 来源核验 AI 生成目标，导弹防御例外。

为什么重要：它把“人工在环”拆成了可检查的批准与证据要求，缩小了口号式承诺的空间。

影响：若进入年度国防授权法案，国防 AI 供应商将需要双源核验、决策日志和明确的指挥责任链。

验证状态：法案已提出并由 Axios 核实；尚未通过众议院、参议院或签署成法。

5. 信号标题：超过 60% 受访法官已用 AI，法院同步收紧虚假引用责任

类型：专业行业采用与责任变化

来源或链接：<https://www.axios.com/2026/07/18/ai-lawyers>

核心观点：Northwestern 对 112 名法官的调查显示 60% 用过至少一种 AI 工具。歇根联邦法院又要求政府确保未来文件不再包含不存在的判例，说明采用和问责正在同时加速。

为什么重要：司法系统不是在等待统一政策后才使用 AI，个案处罚与审判者实践正先行塑造边界。

影响：法律与合规团队需要强制引用回链、人工复核、机密数据隔离和最终签署人确认，不能仅靠员工培训。

验证状态：调查比例、法院文件与法官要求已由原始机构或 Axios 核实；调查样本有限，不代表全美法官总体比例。

前沿研究速递

1. LongStraw：在固定显存预算下把强化学习上下文推过 200 万 token

做了什么：研究者提出分段重放与分布式梯度组合方法，使长上下文强化学习后训练不必同时保留整条轨迹；在 8 张 H20 上完成 Qwen3.6-27B 的 210 万位置分组计分与反

压力测试，在 32 张 H20 上跑通 GLM-5.2 全 78 层的 210 万 token 推理。arXiv:2607.14952 (<https://arxiv.org/abs/2607.14952>)

新在哪里：以往百万 `token` 更多停留在推理阶段，这项工作把后训练执行能力推进到同一数量级；扩大分组规模只增加 `0.21GB` 峰值已分配显存，单独压力测试达到 `446` 万位置。

潜在应用方向：长周期 `coding agent`、深度研究、超长工具轨迹、持续记忆任务和复杂工作流的强化学习。

一句话判断：这是长上下文 `agent` 后训练的重要工程进展，但论文明确承认部分梯度路径尚不完整，目前证明的是执行容量，不是完整训练正确性。

2. Partition, Prompt, Aggregate：用概率自洽性检查模

做了什么：研究者把总体递归切分为子群，分别询问模型，再检验按先验权重汇总后的估计是否与直接总体估计满足全概率公式。来源：`arXiv:2607.15277` (<https://arxiv.org/abs/2607.15277>)。

新在哪里：多种前沿模型普遍违反基本一致性；在 `persona` 提示实验中，从更细子群重建的总体估计反而往往更接近人类参考数据，论文称之为“宏观谬误”。该指标不依赖外部正确答案，可直接检查模型自身概率判断是否互相矛盾。

潜在应用方向：市场规模估计、用户分群预测、风险评分、民调模拟、定价研究和需要聚合判断的企业分析。

一句话判断：让模型回答一个总体比例并不可靠；先分群、再汇总、最后做一致性检查，可能成为 `AI` 辅助决策的低成本质量门槛。