

GPT-6 Astra安全争议发酵;阿里云寒武纪入PyTorch董事会;Mistral创欧洲融资纪录

## 头部企业动态

- OpenAI 9月3日发布 GPT-6 Astra 并公开系统卡:网络安全能力首次触及 Preparedness 框架 Critical 级,生物化学为 High 级;向 ChatGPT Plus/Pro/企业版、API 及 AWS 分批开放。 **A · 96% 已核实**
- PyTorch 基金会 9月8日在上海宣布:阿里云、寒武纪以白金会员加入并各获理事会与技术咨询委员会席位,蚂蚁集团为金牌会员;与华为同台做主题演讲。 **A · 97% 已核实**
- Mistral AI 9月8日宣布完成 30 亿欧元 D 轮,投后估值超 210 亿欧元,三星领投,EQT 旗下基金与 PSG 联合领投,称为欧洲科技公司最大股权融资。 **B+ · 91% 多源一致**
- OpenAI 9月6日发文《研究加速:OpenAI 内部视角》,称已实现「自动化研究实习生」目标,研究部门每人日消耗 3.1 个 agent 工作日,目标 2028 年 3 月做出自动化 AI 研究员。 **B+ · 89% 多源一致**

## 模型发布

- Anthropic 9月1日发布 Claude Fable 5.1 与 Mythos 5.1,同一底层模型、安全护栏不同;缓存读取价降至每百万 0.25 美元(降 75%),Terminal-Bench 4.0 由 42.0% 升至 55.8%。 **B+ · 77% 多源一致 · 链接探活失败**
- Google 9月2日发布 Gemini 3.8 Flash,基于 3.7 Flash 迭代,定价每百万 0.75/3.75 美元至年底;同步推出 3.8 Flash Cyber,仅向 Fairwind 计划内的防御方开放。 **A · 96% 已核实**
- DeepSeek 9月8日在官方交流群开放 V4.1 Flash 中间版内测:新模型结构、原生多模态,模型 ID 标注 9月10日到期,计费同 V4 Flash,每账号限 20 并发。 **B- · 80% 同源引述**

## 热点与趋势

- 前沿模型网络攻防能力集体触顶:OpenAI Astra 系统卡承认思维链可监控性较前代下降,并称若持续恶化将难以侦测失准行为;Google 同期将 Cyber 变体限定给受审防御机构。 **B+ · 90% 多源一致**
- The Information 9月3日报道 Thinking Machines Lab 洽谈以不低于 400 亿美元估值融资,英伟达拟参投;各家媒体转述英伟达出资额从 25 亿到 30 亿美元不一,交易尚未官宣。 **B- · 63% 同源引述 · 链接探活失败**
- 开源 AI 栈治理向中国厂商倾斜:PyTorch 基金会称中国已有 250 余家机构贡献 PyTorch、vLLM、Ray 等项目,阿里云 Qwen 团队与华为昇腾在会上分别讲多集群推理与多后端互操作。 **A · 96% 已核实**

## 主动剔除(未核实,不外发)

- 新智元「老黄 30 亿美元押注穆拉蒂」:公众号原文被验证墙拦截,且金额与 The Information 报道的约 25 亿美元不一致,仅保留估值谈判本身并降为 B-。
- Meta 超级应用 Muse(内部代号 Hatch)及最高每月 199.99 美元定价:仅 TestingCatalog 基于内测泄露的报道,Meta 未官宣,不采用。
- MiniCPM5-2B 发布:PaperWeekly 公众号原文抓取被拒,未找到独立一手源,不采用。
- 微软与 Mistral 数十亿美元欧盟数据中心协议、Anthropic 获 Google 400 亿美元投资及 3500 亿估值:仅见聚合页摘要,未找到原文,不采用。
- 新智元「Claude 黎曼猜想最大突破」「GPT-5.6 一天改写数学史」「Fable 5.1 改写 API 切断蒸馏」等标题:原文均被拦截,未核到官方或论文一手源,不采用。

## 信息源清单(可审计)

| 条目                       | 等级 | 可信度 | 探活 | 来源                                                                                                                        |
|--------------------------|----|-----|----|---------------------------------------------------------------------------------------------------------------------------|
| OpenAI 9月3日发布 GPT-6 Astr | A  | 96% | √活 | <a href="https://deploymentsafety.openai.com/gpt-6-astra...">deploymentsafety.openai.com/gpt-6-astra...</a>               |
| PyTorch 基金会 9月8日在上海宣布:阿里 | A  | 97% | √活 | <a href="https://pytorch.org/blog/alibaba-cloud-ant-group-cambr...">pytorch.org/blog/alibaba-cloud-ant-group-cambr...</a> |
| Mistral AI 9月8日宣布完成 30 亿 | B+ | 91% | √活 | <a href="https://techcrunch.com/2026/09/08/mistral-raises-e3b-a...">techcrunch.com/2026/09/08/mistral-raises-e3b-a...</a> |
| OpenAI 9月6日发文《研究加速:OpenA | B+ | 89% | √活 | <a href="https://unite.ai/openai-hits-goal-of-building-an-autom...">unite.ai/openai-hits-goal-of-building-an-autom...</a> |
| Anthropic 9月1日发布 Claude  | B+ | 77% | ✗死 | <a href="https://venturebeat.com/technology/anthropics-claude-f...">venturebeat.com/technology/anthropics-claude-f...</a> |
| Google 9月2日发布 Gemini 3.8 | A  | 96% | √活 | <a href="https://deepmind.google/models/model-cards/gemini-3-8-...">deepmind.google/models/model-cards/gemini-3-8-...</a> |
| DeepSeek 9月8日在官方交流群开放 V4 | B- | 80% | √活 | <a href="https://ithome.com/0/999/795.htm...">ithome.com/0/999/795.htm...</a>                                             |
| 前沿模型网络攻防能力集体触顶:OpenAI As | B+ | 90% | √活 | <a href="https://deploymentsafety.openai.com/gpt-6-astra...">deploymentsafety.openai.com/gpt-6-astra...</a>               |
| The Information 9月3日报道 T | B- | 63% | ✗死 | <a href="https://pymnts.com/news/artificial-intelligence/2026/t...">pymnts.com/news/artificial-intelligence/2026/t...</a> |
| 开源 AI 栈治理向中国厂商倾斜:PyTorch | A  | 96% | √活 | <a href="https://pytorch.org/blog/alibaba-cloud-ant-group-cambr...">pytorch.org/blog/alibaba-cloud-ant-group-cambr...</a> |

\*评分:A 官方一手抓原文 / B+ ≥2独立源 / B 单家权威或未复核 / B- 同源引述 / C 二手(不采)。本期 WebSearch=16 WebFetch=28 成本≈\$2.4144375。自动生成,源已分级核验。\*