

AI 技术每日分析：行业产品、科学模型与推理基础设施竞速

中国高技术产业发展促进会新质生产力工作委员会

博雅云创 & 中科创新驱动

2026 年 9 月 11 日

摘要

今天 AI 技术方向的新变化，集中在“垂直行业产品、科学基础模型、推理基础设施和 Agent 商业化”四条线上。OpenAI 推出面向金融服务业的专门版本，把投行研究、股票研究、金融数据连接、权限和审计能力放进同一产品，说明通用模型正在进一步按强监管行业重构；IBM 与 NASA 开源月球基础模型和统一机器学习数据集，把 30 多个空间对齐数据层、9 种仪器和 4 项任务的数据整合为可复用科学 AI 底座；推理芯片公司 Positron 完成 8.75 亿美元融资，把“内存优先”架构推向下一代 Asimov 芯片和 Titan 系统；d-Matrix 选择接入 NVIDIA NVLink Fusion，专用推理 XPU 开始主动进入 GPU 主导的机架级生态；印度 PayU 则发布 Agent HQ，让中小商户直接“雇用”Agent 处理商品发现、客户沟通、拒付、退款和结算。AI 产业正在从“模型本身更强”转向“行业数据、推理成本、硬件互联和可直接购买的 Agent 工作单元”共同竞争。

Contents

一、OpenAI 推出金融服务专版：行业 AI 开始把数据、权限和审计一起产品化	2
--	---

二、IBM 与 NASA 开源月球基础模型：科学 AI 从单任务算法 转向可复用数据底座	2
三、Positron 融资 8.75 亿美元：推理时代开始围绕“内存容量与 带宽”重新设计芯片	3
四、d-Matrix 接入 NVLink Fusion：专用 XPU 不再只做 GPU 替代，而是进入异构机架	4
五、PayU 发布 Agent HQ：中小企业开始直接购买“可执行的数 字岗位”	5
趋势观察	6
参考文献	6

一、OpenAI 推出金融服务专版：行业 AI 开始把数据、权限和审计一起产品化

9 月 10 日，OpenAI 推出面向金融服务业的专门版本 ChatGPT for Financial Services，重点服务投资银行、股票研究等高信息密度、高合规要求的工作。Reuters 报道，该产品与 Morgan Stanley、Evercore 等金融机构共同打磨，并连接 LSEG、PitchBook、Daloopa 等数据源，同时支持金融机构既有数据订阅和内部资料。

与普通企业版相比，这次更值得关注的并不是“金融 Prompt 更懂行”，而是产品开始把受监管行业必须具备的治理能力放在同一层：角色访问控制、加密、审计日志导出，以及对可信外部数据和公司内部数据的统一检索。金融分析最怕的不是模型不会写摘要，而是数字出处不清、权限越界、数据版本混乱和结论无法追溯。

这说明行业 AI 的下一阶段不会只靠一个“金融模型”或“法律模型”解决问题。真正可落地的行业产品，需要把领域数据、 workflow 模板、身份权限、引用和审计共同设计。模型能力越来越通用以后，行业壁垒反而会更多落在数据连接、评价体系和合规控制上。

对于企业 AI 平台而言，这也意味着“垂直化”不等于重新训练一个大模型，而可能是围绕某个行业建立受治理的上下文与工具环境，让通用模型在正确的数据、权限和流程中工作。

二、IBM 与 NASA 开源月球基础模型：科学 AI 从单任务算法转向可复用数据底座

IBM 与 NASA 9 月 10 日宣布开源 NASA-IBM Lunar Foundation Model，并同步开放统一的机器学习就绪月球数据集。该数据集聚合了 30 多个空间对齐数据层，来自 9 种仪器和 4 项任务，其中包括 NASA Lunar

Reconnaissance Orbiter、GRAIL 以及日本 JAXA 的 SELENE/Kaguya 数据。

模型面向月球冰、火山地貌、陨石坑等科学任务。IBM 披露的技术结果显示，在部分月球冰潜力区域识别任务中，相较 SwinV2-B 基线可降低最高约 22% 的误差；在约 100 米上下文尺度的陨石坑任务中，用约一半训练数据取得接近 19% 的提升。上述结果来自 IBM 和 NASA 联合论文及发布材料，仍应结合具体任务和评测设置理解。

这条新闻的重要性在于“先统一科学数据，再做基础模型”。科研数据往往来自不同年代、不同仪器、不同分辨率和不同坐标系统，单纯把图片丢给模型并不能形成可靠科学结论。IBM 与 NASA 把多源观测先对齐成共同数据框架，再训练通用表示，使研究者可以在同一个底座上继续微调不同任务。

AI for Science 正在由“每个课题单独做一个模型”转向共享科学基础模型。对于材料、气象、地球科学和工程仿真同样如此：真正稀缺的资产通常不是网络结构，而是经过统一标定、可复用、可追溯的高质量科学数据体系。

三、Positron 融资 8.75 亿美元：推理时代开始围绕“内存容量与带宽”重新设计芯片

AI 推理硬件公司 Positron 9 月 10 日宣布完成 8.75 亿美元 Series C/C-1 融资，投后估值 50 亿美元。资金将用于完成下一代 Asimov 芯片流片、建设超过 2MW 的工程数据中心和仿真平台，并推动 Titan 推理系统量产。

Positron 选择了一条与主流 GPU 不同的“memory-first”路线。公司计划让 Asimov 采用 TSMC N3P 工艺，在 2026 年底流片、2027 年下半年生产，每颗芯片配置 288GB 至 2304GB 内存；Titan 将组合 4 至 8 颗

Asimov，面向超长上下文和超大参数模型。公司强调其下一代架构使用 LPDDR5X，尽量绕开当前紧张的 HBM 与 CoWoS 供应链。

这背后的产业逻辑很清晰。训练阶段更关注峰值计算，而大规模在线推理越来越受模型权重、KV Cache 和长期 Agent 状态占用的内存容量及带宽约束。Agent 持续运行、上下文不断变长之后，“能不能把更多状态放在高速内存中”会直接影响每 Token 成本和延迟。

Positron 披露，其第一代 Atlas 已在 Oracle Cloud Infrastructure 部署 50 多个机架。需要注意，关于 tokens per dollar、带宽利用率等性能优势主要来自公司自身口径，最终仍需要更多第三方 Benchmark 和真实负载验证。但 8.75 亿美元的新融资说明，资本已经开始把“推理专用架构”视为与训练 GPU 不同的独立市场。

四、d-Matrix 接入 NVLink Fusion：专用 XPU 不再只做 GPU 替代，而是进入异构机架

d-Matrix 9 月 10 日宣布，其下一代 Raptor XPU 将采用 NVIDIA NVLink Fusion，与 NVIDIA Vera CPU、ConnectX-9、BlueField-4 DPU 和 Spectrum-X Ethernet 等组件构建异构 AI 系统。Reuters 报道，Raptor 最终设计预计在 2026 年底完成，相关系统计划于 2027 年上市。

d-Matrix 长期聚焦低延迟生成式 AI 推理，并采用把计算更靠近内存的架构。此次选择 NVLink Fusion 很值得关注，因为它说明专用 AI 芯片公司的竞争方式正在变化：不一定要完全替代 GPU，而是通过高速互联进入 GPU 主导的数据中心机架，在最擅长的推理阶段承担特定负载。

Agent、语音助手和 Coding Agent 要求的不只是高吞吐，还要求低首 Token 延迟、稳定响应和成本可控。未来一个 AI 机架可能同时存在 GPU、专用 XPU、DPU 和不同类型内存，由运行时按工作负载自动分工。硬件竞争会越来越像“异构计算系统设计”，而不是单芯片跑分。

对开发者来说，真正决定新芯片能否进入市场的也不只是 TOPS，而是能否接入主流网络、调度、模型框架和推理服务栈。NVLink Fusion 等开放互联机制，正在成为新推理芯片缩短生态建设周期的重要入口。

五、PayU 发布 Agent HQ：中小企业开始直接购买“可执行的数字岗位”

印度支付平台 PayU 9 月 10 日在 Global Fintech Fest 2026 发布 Agent HQ，定位为面向中小商户的 AI Agent 商店。商户无需单独搭建模型基础设施或编写集成代码，可以在 PayU 现有后台中选择不同 Agent 并设置执行频率、边界和结果通知方式。

Agent HQ 一部分面向增长：帮助商品和服务被 AI 购物 Agent、Chat-GPT、Claude 以及 WhatsApp 等渠道发现，并支持建立面向网站和 App 的对话助手；另一部分面向运营，可处理交易跟进、拒付对账、退款升级和结算等任务。PayU 表示，Policybazaar、DailyObjects 等商户已经在使用相关能力。

这类产品的重要性不在于又出现一个“Agent Store”，而在于 Agent 第一次更像中小企业能够直接采购的工作单元。过去企业采用 AI 要先选模型、建知识库、做 API 集成，再设计流程；平台型公司现在可以把支付数据、业务流程和 Agent 打包成开箱即用的岗位能力。

对中小企业市场而言，Agent 商业化的关键可能不是最强推理，而是部署摩擦足够低、权限边界明确、每个任务能产生可核算价值。随着支付、CRM、电商和办公平台都开始内置 Agent 商店，“软件订阅”有可能进一步演变成“数字劳动力订阅”。

趋势观察

今天五条动态形成了一条完整的 AI 产业链：OpenAI 把通用模型封装成受监管行业产品，IBM 与 NASA 把科学数据加工成开放基础模型，Positron 和 d-Matrix 围绕推理经济性重构硬件，而 PayU 把 Agent 包装成商户可直接购买的工作能力。AI 竞争正在从模型参数和 Benchmark，扩展到“行业数据—推理基础设施—异构互联—Agent 交付”的系统竞争。下一阶段真正决定规模化的指标，会越来越接近单位任务成本、数据可信度、部署摩擦和业务结果。

参考文献

- Reuters | 《OpenAI launches ChatGPT for financial services industry》
| 2026-09-10 | 核验金融服务专版、合作机构、数据源与安全治理能力。来源链接
- OpenAI | 《AI for Financial Services》 | 2026-09 访问 | 补充金融行业的数据连接、安全、审计与工作流背景。来源链接
- IBM / PR Newswire | 《IBM and NASA Release Open-Source AI Model to Support Lunar Exploration》 | 2026-09-10 | 核验月球基础模型、统一数据集与技术评测。来源链接
- NASA / IBM 研究资料 | NASA-IBM Lunar Foundation Model 相关资料 | 2026-09-10 | 补充月球遥感、多模态与科学任务背景。来源链接
- Positron AI / PR Newswire | 《Positron AI Raises \$875 Million at a \$5 Billion Valuation...》 | 2026-09-10 | 核验融资、Asimov、Titan 及量产路线。来源链接
- Wall Street Journal | 《Positron Valued at \$5 Billion in New Funding as Demand for AI Chips Surges》 | 2026-09-10 | 交叉核验融资和推理

芯片市场背景。来源链接

- Reuters | 《Chip startup d-Matrix to use Nvidia chip-linking tech in AI servers》 | 2026-09-10 | 核验 NVLink Fusion、Raptor 及 2027 年系统路线。来源链接
- d-Matrix | NVLink Fusion 合作与推理架构资料 | 2026-09-10 | 补充 Raptor 及内存计算路线。来源链接
- The Economic Times | 《GFF 2026: PayU's new AI agent store aims to change how SMBs run businesses》 | 2026-09-10 | 核验 Agent HQ、无代码部署和运营 Agent 范围。来源链接
- PayU 相关发布资料 | Agent HQ 产品信息 | 2026-09-10 | 补充商户控制、交易运营和 Agentic Commerce 场景。
— 来源链接

联系我们，请扫描二维码



新质生产力工作委员会
官方公众号



工业智能算网
gyznsw.cn

新质生产力工作委员会：

中国高技术产业发展促进会新质生产力工作委员会，专注于推动工业人工智能、智能制造、数字化转型等前沿技术发展，为企业提供政策解读、技术咨询和产业对接服务。

工业智能算网：

专注于工业人工智能、新质生产力、工业软件 CAE、智能制造等前沿技术。提供每日动态分析、技术趋势解读、解决方案分享，推动工业智能化转型。

网站地址：<https://gyznsw.cn>