亿欧智库

2024 中国信创+AI趋势洞察报告

亿欧智库 https://www.iyiou.com/research

亿欧智库

目录

CONTENTS

生成式AI爆发，信创产业迎来新拐点，信创+AI深度融合共促发展

- 当前，人工智能技术范式变革和全球数智化的加速转型为算力基础设施业务提供了新的成长动能，也为我国信创产业提供了新的发展方向。
- 从1993年中软推出第一代基于UNIX为底层的国产Linux操作系统“COSIX 1.0”，到2013年浪潮天梭K1小型机系统上市，再到飞腾、鲲鹏等芯片技术逐渐成熟，麒麟操作系统、达梦数据库党政和金融领域被广泛引用，我国自主安全的信息技术不断突破创新，逐步与国际先进水平接轨。
- 如今，生成式AI的爆发将信创产业推向关键的技术升级拐点：更高的算力需求、更高效的算法优化、更庞大的数据规模提供了可观的产业升级机会。另一方面，要解决AI领域备受关注的安全性问题，需要依托自主安全的产业链和生态环境。在未来，信创产业与AI产业有望深度融合，共促发展。

信创

生成式AI爆发
信创产业迎来关键拐点

面对目前AI大模型发展浪潮下算力需求的爆炸式增长以及AI能力的革命性跃升，为了在全球科技竞争中占据主动地位并确保核心技术自主安全，我国信创产业积极拥抱技术变革。

生成式AI的崛起，不仅对硬件基础设施提出了更高的算力要求，还对基础软件的资源管理能力提出挑战，更考验对大模型的设计能力和生态应用能力。为此，我国信创产业亟需在技术，质量以及规模上进行突破，以应对生成式AI全面井喷的巨量需求。

AI

AI的安全发展
离不开自主安全的产业链条与生态环境

AI产业在为社会生产力带来提升的同时也向人类提出了前所未有的安全挑战。目前已知的能够由AI技术实现的安全威胁包括：由AI增强的网络诈骗，用于网络攻击的破坏性代码生成，以及由目标偏差导致的破坏性行为。

我国信创产业通过推动关键技术的自主安全、构建本土化技术生态、强化数据安全与隐私保护，为AI产业提供了安全的发展环境与关键的政策和资金支持，同时加速了AI产业链的完善与升级，共同推动了AI技术创新和商业化落地。

数据来源：公开资料，亿欧智库

《2024政府工作报告》关注科技创新、新型基础设施建设，强调超前建设数字基础设施

- 2024年《政府工作报告》明确提出“适度超前建设数字基础设施，加快形成全国一体化算力体系，培育算力产业生态”，为数字经济的深化发展奠定了坚实基础。在政府工作报告中被首次提出的“人工智能+”行动，核心在于推动人工智能技术与各行各业的深度融合，创造新的产品、服务和商业模式，从而推动传统行业的转型升级和社会经济结构的变革。
- 在产业下游，“人工智能+”行动通过将AI技术与制造、医疗、教育、交通等传统行业深度融合。例如，AI在制造业中的应用已涵盖精细化生产、工艺优化、设备维护等多个环节，显著提升生产效率。而在产业上游，“适度超前建设”数字基础设施，特别是智能算力中心的布局，不仅将推动GPU、NPU等异构芯片的国产化，还带动了液冷技术等高效散热方案的创新，进一步降低了算力成本与能耗。这种上下联动的产业升级模式，正在加速全要素生产率的提升，为经济高质量发展注入新动能。

2024年政府工作总体目标

首次提出 “人工智能+” 行动

深入推进数字经济创新发展：深化大数据、人工智能等研发应用，开展“人工智能+”行动，打造具有国际竞争力的数字产业集群。适度超前建设数字基础设施，加快形成全国一体化算力体系，培育算力产业生态。

积极培育新兴产业和未来产业：加快智能网联新能源汽车、前沿新兴氢能、新材料、创新药等产业发展，积极打造生物制造、商业航天、低空经济等新增长引擎。制定未来产业发展规划，开辟量子技术、生命科学等新赛道。

积极扩大有效投资：重点支持科技创新、新型基础设施、节能减排降碳，加强民生等经济社会薄弱领域补短板，推动各类生产设备、服务设备更新和技术改造。2024年中央预算内投资拟安排7000亿元。

形成以人工智能为引擎的新质生产力

人工智能技术快速发展的背后，需要强大的算力作为支撑

适度超前建设数字基础设施 » **“人工智能+”新业态新模式** » **形成一体化算力产业生态**

适度，一方面意味着算力投资不能过度与实际需求脱节；

另一方面，因为技术迭代较快，算力投资应是持续性的长期投入。

数据来源：公开资料，亿欧智库

政策引导算力部署加速，各省市政府大力推动AI基础设施国产化

- 2023年10月工业和信息化部等六部门联合印发《算力基础设施高质量发展行动计划》，将算力定义为集信息计算力、网络运载力、数据存储力于一体的新型生产力。主要目标包括2025年计算力规模超过300EFLOPS，智能算力占比达到35%，重点应用场所光传送网（OTN）覆盖率达到80%，存储总量超过1800EB，先进存储容量占比达30%以上。
- 各省市政府发布引导智算中心发展相关政策，大力推动AI基础设施国产化进程。其中宁夏政府对国产化率90%以上的数据中心给予补贴奖励，青岛人工智能产业园要求国产人工智能加速卡占比≥70%。

亿欧智库：《算力基础设施高质量发展行动计划》量化指标

	序号	指标	2023年	2024年	2025年
计算力	1	算力规模（EFLOPS）	220	260	300
	2	智能计算中心（个）	30	40	50
	3	智能算力占比（%）	25	30	35
运载力	4	重点应用场所光传送网(OTN)覆盖率（%）	50	65	80
	5	SRv6等创新技术使用占比（%）	20	30	40
	6	国家枢纽节点数据中心集群间网络时延达标率(%)	65	75	80
存储力	7	存储总量（EB）	1200	1500	1800
	8	先进存储容量占比（%）	25	28	30

亿欧智库：我国引导智算中心发展相关政策

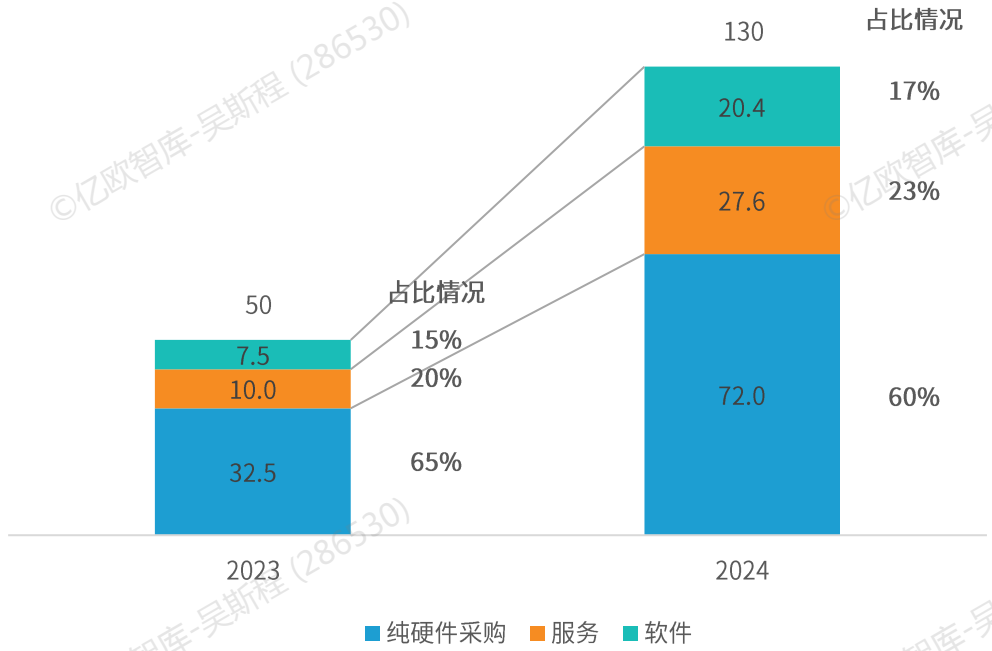
部门	政策/公告	内容
宁夏自治区人民政府办公厅	《自治区人民政府办公厅关于促进全国一体化算力网络国家枢纽节点宁夏枢纽建设若干政策的意见》	加大信创企业扶持力度，对基础软硬件实现国产化率90%以上的数据中心，给予企业最高不超过1000万元奖励。重点加强基础芯片、自主指令集的产学研及配套产业建设
宁夏自治区发展改革委	《全国一体化算力网络国家枢纽节点宁夏枢纽建设2023年工作要点》	推动龙芯中科、中电子、中兴、华为等自主安全产业园区落地，鼓励亿国产化CPU、GPU、操作系统等自主安全产品为底座的信创云平台自主研发，打造安全可信计算、网络和存储能力。
青岛人工智能产业园	《青岛“海之心”人工智能计算中心项目》	要求国产人工智能加速卡占比≥70%
成都市经信局	《成都市围绕超算智算加快算力产业发展的政策措施》	鼓励智算中心建设国产自主安全、安全可靠的人工智能算力基础设施和技术路线生态。
南京麒麟科创园	《南京智能计算中心算力推广办法（试行）》	以“算力券”和“算力折扣”两种形式为广大企业提供更优质的算力资源、更有力的服务保障。
北京市人民政府	《北京市加快建设具有全球影响力的人工智能创新策源地实施方案（2023-2025年）》	在人工智能产业聚集区新建或改建升级人工智能商业化算力中心，加强国产芯片部署应用，推动自主安全软硬件算力生态建设
成都市经信局	《成都市加快大模型创新应用推进人工智能产业高质量发展的若干措施》	引导国家超算成都中心、成都智算中心合理扩容，支持鲲鹏、昇腾、海光等自主安全芯片部署，提高自主研发算力设备比例。

数据来源：公开资料，亿欧智库

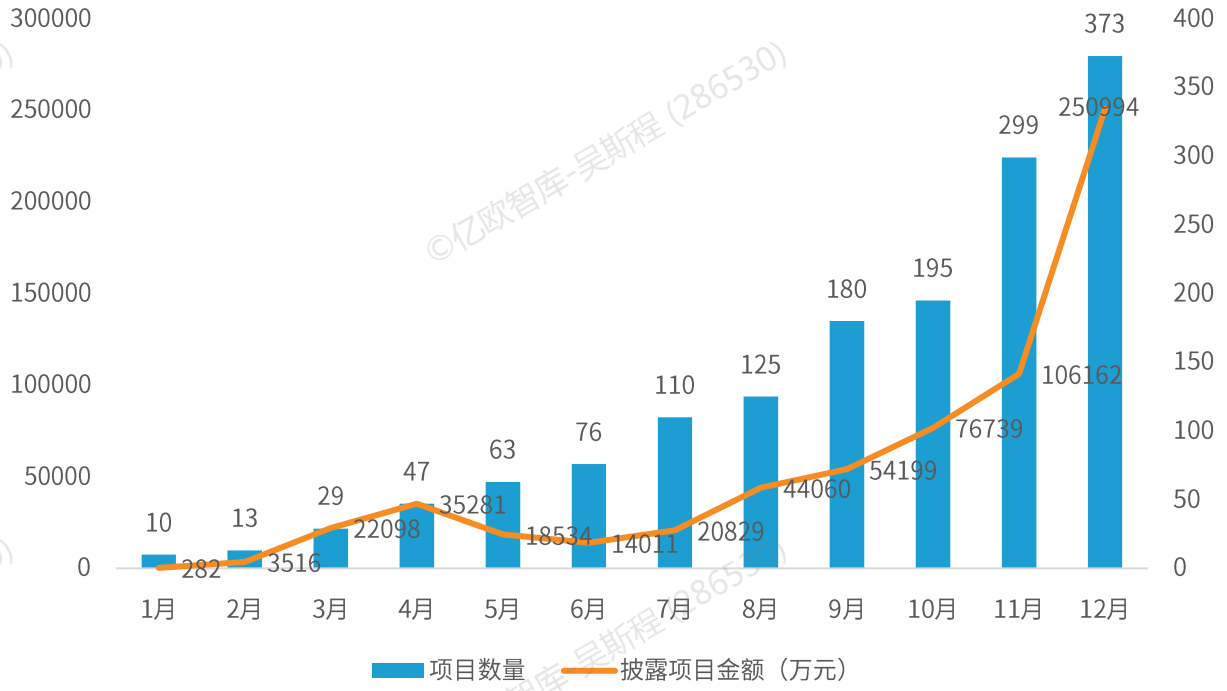
生成式AI需求爆发，国内企业大量释放大模型预算，采购金额有望进一步增加

- 从市场公开资料来看，2024年中国大模型市场规模与2023同比增速达到140%。在构成方面，纯硬件采购占主要成分，纯硬件采购部分约占60%（增速达到122%），服务部分约占23%，软件部分约占17%。
- 从招投标市场来看，围绕大模型技术的基础设施建设和技术应用正同步推进，其中服务应用呈现指数级增长，基础设施建设则朝着专用化发展。其中，硬件设备（大模型专用）占比达28%，从2023年9月起，大模型专用的硬件设备需求开始上涨。企业用户在2024年将开始大量释放大模型预算，规划中大模型占AI预算约10%，预算规模大多为数百万元。

亿欧智库：2023-2024中国大模型市场规模及构成（单位：亿元）



亿欧智库：2024年中国大模型中标项目数量与金额



数据来源：专家访谈，智能超参数

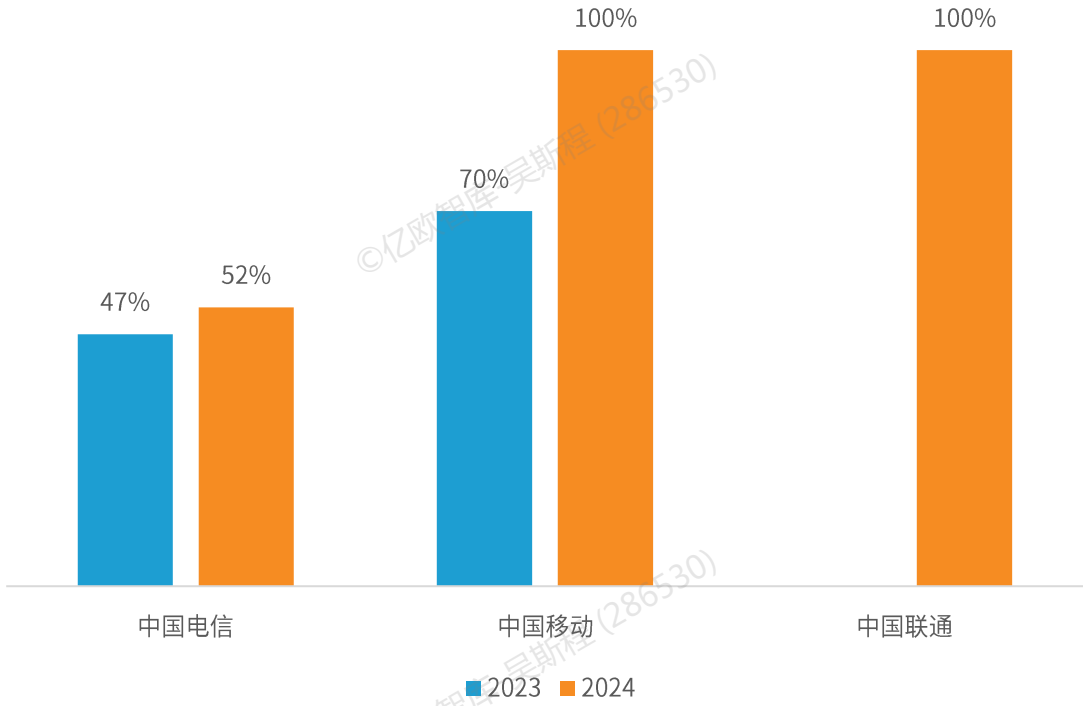
AI服务器信创大单频频落地，三大运营商AI服务器国产化趋势显现

- 在算力角逐驶入“快车道”的关键阶段，三大运营商出手“阔绰”，耗资百亿元“加码”智算中心建设。作为国内通信基础设施建设的“国家队”，三大运营商基于自身强大的网络资源和遍布全国的运维团队，加上充足的现金流，运营商冲在建设智算中心的第一线。
- 国产化率方面，中国移动与中国联通在2024-2025年的AI服务器集采项目中的国产化率均达到100%，中标公司均为华为昇腾合作伙伴。中国电信AI服务器采购国产化率达到52%，较上一年度增加5%。三大运营商整体AI服务器集采国产化率达到73%。

亿欧智库：2023-2024央国企AI服务器招标情况

招标时间	公司	项目	台数	金额
2024.8	中国移动	2024年至2025年新型智算中心集采项目	7994	超191亿元
2024.7	中国电信	中国电信服务器（2024-2025年）集中采购项目	1.3万	229亿元
2024.4	上海浦东发展银行	金融云基础平台异构SDDC建设项目之鲲鹏芯片服务器	\	1.31亿元
2024.4	海通证券股份有限公司	证券垂直领域大模型项目AI算力服务器	\	345万元
2024.3	中国联通	2024年人工智能服务器公开集采	2053	超20亿元
2024.3	招商银行	信创AI推理服务器采购项目	\	\
2023.9-2024.1	中国移动	2023年至2024年新型智算中心（试验网）采购	1250	超7亿元
2023.8	中国电信	AI算力服务器（2023-2024年）集采项目	4175	超80亿元
2023.8	江苏银行股份有限公司	高算力GPU服务器	\	433万元
2023.1	中国交通银行	国产GPU服务器（寒武纪）选型项目	\	\

亿欧智库：2023-2024三大运营商AI服务器集采国产化率



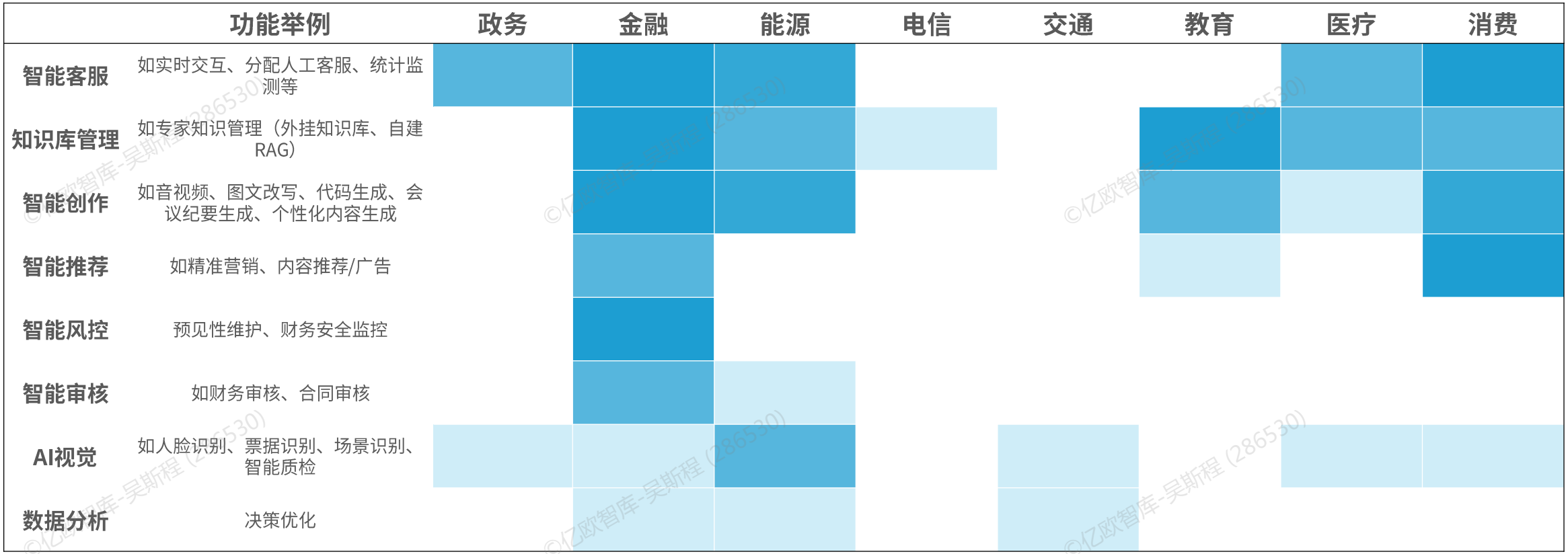
资料来源：中原证券；亿欧智库

AI大模型落地八大行业，金融、能源行业应用场景多、落地进展快

- 从功能应用的多样性来看，金融行业对大模型技术的应用最为多样化，范围涵盖了所有的应用类别，其次是能源和消费行业。知识库管理和AI视觉是各行业采纳最多的应用类别，分别被六个行业所采纳使用。
- 从广泛程度来看，金融行业对AI应用最为广泛，尤其是智能客服，知识库管理，智能创作以及智能风控。其中，智能客服是全行业应用最为广泛的AI应用。

亿欧智库：AI大模型各行业应用情况

以颜色深浅表示应用的广泛程度



数据来源：公开资料，亿欧智库

垂类AI大模型初步落地，央国企带头部署应用

- 央企带头推进算力建设和行业大模型运用，未来会在人工智能领域的投资持续增加，涉及算法、平台、解决方案全产业生态，持续优化企业竞争力。国资委频频部署加快布局人工智能产业，加快推动人工智能发展，使国资央企发挥功能使命。通过在人工智能领域的投资和发展，中央企业可以推动整个产业结构的升级和优化。未来，中央企业在人工智能领域的投资可能会不断增加。
- 推进行业大模型高质量发展及边缘算力协同部署和应用。央企作为代表将持续深化大模型在行业中的应用并向社会开放场景，有望加速 AI 商业化落地进程，实现投资回报的正向循环。

亿欧智库：部分央国企垂类AI大模型应用案例

央企集团	案例	案例概况
国家电网	光明电力大模型	光明电力大模型是面向电力行业的千亿级多模态大模型，为电网安全稳定运行、促进新能源消纳、做好供电服务提供“超级大脑”。光明电力大模型通过中国信通院、电子标准院权威检测，电力知识记忆理解、多模态融合分析、业务逻辑推理、基础数值计算和内容辅助生成能力较基座模型平均提升20%。与主流大模型对比，专业能力平均高出15%。
国家电投	“天枢一号”智慧能源系统	我国首个可以实现数十种能源同时管理的智慧大脑。该系统横向贯通源、网、荷、储，全域物联场景，纵向融合“云大物移智链”先进技术，集能源监视、预测、调控、分析、运 维和服务等近百项功能、千项应用于一体，拥有800个以上核心智能算法，实现数十种不同能源的综合管控。
南方电网	自主安全电力大模型“大瓦特”	电力行业首个跨NLP/CV模态大模型产品，实现算力、算法、应用全过程自主安全。已经在智能客服、输变配、电力调度和安监等垂直领域得到应用。
中国联通	元景大模型2.0	元景大模型2.0已形成37个行业大模型和100多个标杆应用，赋能经济社会新质发展成效显著。如基于元景大模型构建的南京港5g安全生产智能管控平台，深入6大港口场景14个生产作业环节，开发集成了人员安全行为识别、港区安全环境监测、作业合规监测3大类共41个典型场景的AI算法
中国农业银行	金融AI大模型应用ChatABC	ChatABC依托农业银行的人工智能服务体系，结合内部知识库和数据进行训练，实现了金融知识理解和问答能力，并支持自由闲聊、内容摘要等多任务处理。1.0版本拥有百亿级参数，已在行内多个渠道试用，并可通过MaaS方式提供决策辅助服务，未来将形成大模型服务生态。
中国建设银行	大模型“方舟计划”	建设银行于2023年3月成立“方舟计划”专项工作组，依托算力资源打造千亿大模型基座，并利用高质量文本数据进行预训练、微调和强化学习，使其能更好地理解金融知识和建设银行业务。目前已初步具备信息总结、信息推断、信息扩展、文本转换、安全与价值观、复杂推理、金融知识7项一级能力和26项二级能力。

数据来源：公开资料，亿欧智库

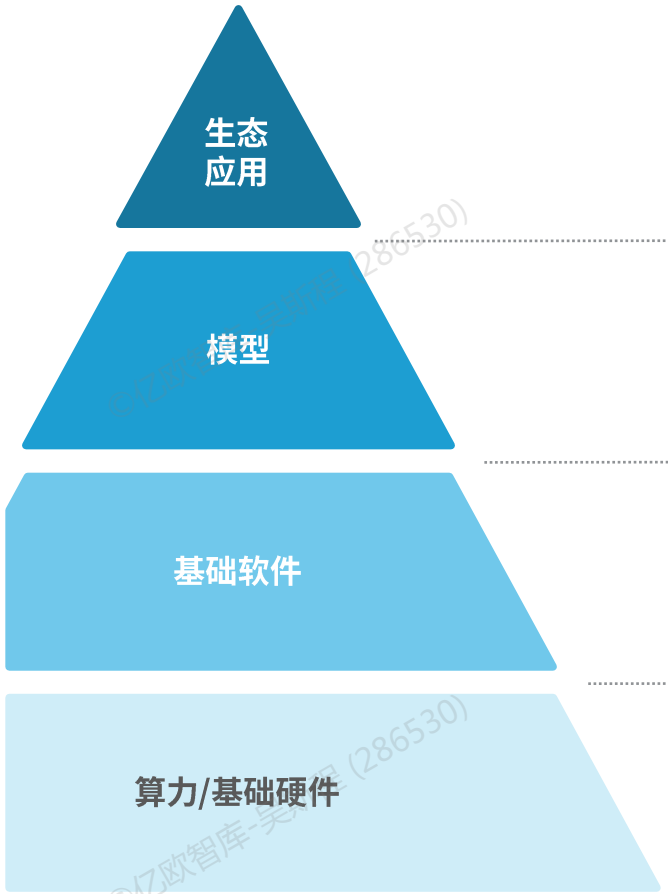
亿欧智库

目录

CONTENTS

信创产业链各环节均迎来AI技术变革

- 人工智能技术对于信创产业的影响可以从产业链的角度分为四个环节，分别为：基础硬件，基础软件，模型和生态应用。
- 其中，基础硬件层和模型层决定了AI大模型的核心基础能力，而基础软件层和生态应用层则推动了大模型的落地，帮助企业实现与业务的结合。当前AI大模型浪潮爆发，这四个层次各自面临不同的技术挑战和发展趋势。



	技术挑战	发展趋势	代表企业
生态应用	• **幻觉**：模型时常生成错误且具有误导性的回答 • **黑箱**：模型决策过程缺乏透明度，用户难以与之建立信任 • **监管**：责任归属和安全监管机制有待完善	• **研发**：利用AI技术在材料，药物，软件，游戏等领域辅助研发 • **生产**：数字孪生，机器视觉，人形机器人深入工厂 • **运营**：运营自动化，AI知识库管理 • **营销**：AI客户画像分析，AI个性化营销	金山办公 中关村科金 用友 鼎捷数智
模型	• **算法**：算法创新加速，技术路线频繁更迭 • **数据**：高质量数据开始匮乏 • **成本**：随着基础模型不断扩大，模型的训推成本持续攀升	• **算法**：“Fine-tune”和“prompt工程”应用量下滑；“RAG”应用量上升；研发路线以“后训练”，“推理计算”，“多模态/世界模型”为方向发展 • **数据**：合成数据提供高质量训练语料 • **成本**：使用蒸馏，剪枝等技术，压缩模型参数量	360 商汤 科大讯飞
基础软件	• **资源调度**：巨量计算资源以及模型工具的管理调度 • **安全性**：模型运行所需的大量数据将很大程度涉及企业隐私和商业机密 • **系统集成**：模型与现有工作流的无缝集成	• **操作系统**：微内核架构；AI OS；LLM OS • **数据库**：向量数据库；低延迟，高吞吐，高并发的数据处理能力 • **中间件**：针对大模型数据与应用集成，知识库融合等需求的自动化工具	麒麟软件 东方通 OCEANBASE
算力/基础硬件	• **算力**：大模型对矩阵运算的海量需求，带来算力需求的指数级增长 • **网络&存储**：巨量矩阵运算对网络与存储的带宽和延迟提出了更高要求 • **散热**：硬件的功率密度及热流密度大幅升高	• **芯片**：片内、片间互联性能；高带宽，低延迟的存储芯片；多精度混合算力 • **服务器**：超大计算集群组网；冷板式与浸没式液冷 • **PC**：支持AI模型本地部署；支持AI软硬件生态和应用	HYGON 兆芯 METAX 沐曦 超云 SUPER CLOUD Lenovo 联想

数据来源：公开资料，亿欧智库

基础硬件层面：AI大模型算力需求激增，对基础硬件能力提出更高挑战

- **AI芯片层面：**近年来随着大数据与AI技术的发展，芯片算力的增长速度已经远超过存储和网络带宽的增长速度。因此在面对大模型算力需求爆发，大规模并行计算网络成为行业趋势之际，AI芯片的主要发展瓶颈在于**数据传输速度**以及**硬件互联能力**而非单卡晶体管数量。目前，更大的**存储容量**，更高的**存储和互联带宽**，更强的**片内/片间互联**是芯片厂商的发力重点。
- **AI服务器层面：**随着大模型对算力需求呈指数级增加，芯片与计算网络向**高度集成化**方向发展，服务器的**功率和功率密度**大幅上升。同时，大模型的训练/推理需求也对服务器的**大规模组网能力**提出更高要求。
- **AI PC层面：硬件成本，端侧模型**的承载能力，**产品生态建设**以及用户对于数据安全方面的考虑是行业需要攻克的主要挑战。

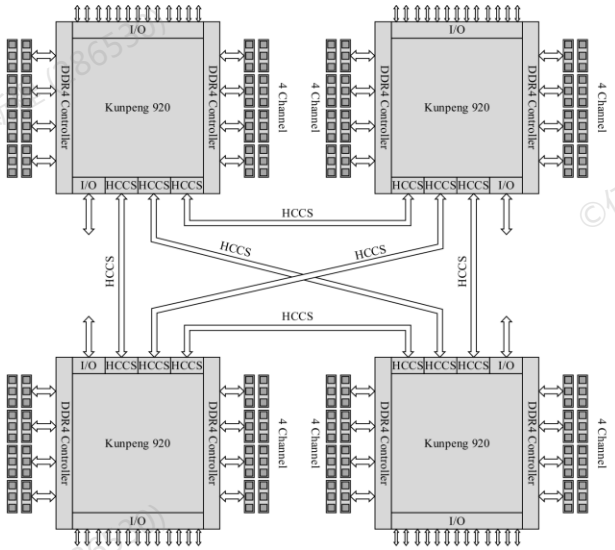
	AI芯片	AI服务器	AI PC
技术挑战	**性能：**国产AI芯片在**算力、存储、带宽、功耗**等核心指标上与国外头部企业普遍存在差距，且在芯片设计、制造工艺和软件生态方面有待突破。 **成本：**尽管凭借片间互联，超大集群等技术，国产AI芯片具有“以量取胜”的潜力，但**TCO仍然较高**，在效率为王的大模型训推场景下，满足主流企业的训推需求仍具有一定困难。	**互联：**针对大模型的超大规模组网需求对服务器集群的**网络带宽**以及**延迟**提出了更高要求。 **散热：**AI应用对算力需求的显著上升要求服务器能够承载更高的**功率**以及更高的**功率密度**。因此下一代AI服务器亟需更好的冷却解决方案以应对服务器的发热问题。	**模型承载能力：**目前主流的本地算力解决方案难以独立支撑大模型所需的算力。 **生态建设：**本地大模型需要与硬件，下游应用，用户进行操作上的打通，构建**完整产品生态**。 **隐私安全：**AI PC应用对于海量个人数据的需求将使得用户的个人数据安全将面临更高挑战。
发展趋势	**互联与存储能力：**发展**chiplet，C2C，光电共封装，存算一体**等技术以提高片内互联以及片间互联能力，以打破数据传输速度瓶颈。发展3D封装，硅通孔等先进封装技术，扩展存储容量和带宽，以突破内存墙问题。 **生态兼容：**兼容CUDA，ROCm等加速计算生态，原生支持Tensorflow，pytorch等主流深度学习框架，以降低开发者迁移成本。	**扩展/组网能力：**发展多卡**并行方案，RDMA，RoCE**等节点间通信技术，提高算力硬件的线性扩展能力。构建超大服务器集群以应对算力需求的爆发式增长以及单卡算力不足的挑战。 **散热性能：**为高功率密度服务器配备**冷板式液冷**与**浸没式液冷**等技术，解决芯片级散热问题以及整体机柜散热问题。	**端云结合：**构建端云结合的产品生态，以弥补本地算力不足。 **个人大模型：**个性化本地数据库，将用户的个人数据加入模型以提供个性化AI服务。 **AI智能体：**具有强大泛化能力的AI智能体仍有待突破技术难关。点状垂类功能有望率先打开市场。

数据来源：公开资料，亿欧智库

CPU：华为鲲鹏920基于ARM v8.2架构，采用Chiplet，HCCS互联技术，全面支持AI工作负载

- 华为鲲鹏920是华为自主研发的高性能数据中心处理器，专为大数据处理和分布式存储等应用而设计。
- 鲲鹏920基于ARM v8.2架构，7纳米工艺制造，使用Chiplet技术，能够集成2个CPU die，最多64颗核心。由Hydra接口提供的片间互联能够实现最多4个鲲鹏920处理器C2C互联，256个物理核的NUMA架构，片间带宽最高可达480Gbps。
- 鲲鹏920的大部分性能提升来自于优化的分支预测算法、更多的运算单元，以及内存子系统架构的改进。凭借鲲鹏920，华为现在已经进入了一个以多核、异构为代表的多元化计算时代。

高速片间互联技术



HCCS片间互联：华为Cache一致性总线（HCCS）实现鲲鹏920多芯片的两两互连，为内核、设备、集群提供系统内存的一致访问，以支持云计算工作负载的扩展。

掌握核心技术和完全的知识产权

处理器内核	片内互联 Fabric	片间互联协议
7nm, 64个处理核心		
8通道 DDR4	PCIe 4.0& CCIX	100Gbps ROCE
内存	总线	网络

数据来源：《Kunpeng 920: The First 7-nm Chiplet-Based 64-Core ARM SoC for Cloud Services》Jing Xia，公开资料，亿欧智库

CPU：兆芯采用chiplet互联架构，主频最高达3.7GHz，创下国产CPU新高

- 兆芯开先KX-7000系列面向PC电脑及嵌入式市场，采用全新设计的自主微架构“世纪大道”，以及先进的Chiplet互连架构，主频最高达3.7GHz，创下国产CPU的新高。
- 兆芯开胜KH-40000面向云计算、大数据分析、高并发、高性能存储、超融合等云端应用场景，具备高核心性能、高集成度、高效互连、丰富IO等产品特点，集成最多32个高性能核心，支持双路互连构建64核服务器整机，同时支持8通道DDR4内存、128路PCIe 3.0。

兆芯

PC/嵌入式处理器

开先®
KX-7000 系列处理器

KX-7000集成高性能GPU显卡，支持DX12、OpenGL 4.6、OpenCL 1.2、双路4K硬件解码输出，还升级支持DDR5/DDR4内存、PCIe 4.0通道、USB4/USB3接口等主流高速IO。与上一代产品相比，开先KX-7000系列计算性能提升2倍，图形性能提升4倍。

联想开天基于兆芯开先KX-7000系列处理器平台推出了P90z G1t台式机，作为信创首款AIPC，以及联想开天N8 Pro系列笔记本。

清华同方，紫光计算机等也推出了搭载KX-7000的信创产品。

服务器处理器

开胜®
KH-40000 系列处理器

KH-40000采用“永丰”自主内核微架构，支持自主互连技术ZPI 3.0，单颗处理器集成最高32核心，具备64MB高速缓存，支持8通道DDR4内存，提供多达128 Lane PCIe通道，以及SATA、USB等主流IO接口，**支持片上互连和多路互连**，可构建64核服务器整机系统，以更好满足服务器应用对**多核心、多内存、多PCIe扩展**等AI需求。

基于开胜KH-40000处理器的联想开天服务器有单机64核的强大计算能力和丰富的扩展能力，能够支持云计算、OA、数据库等多种应用类型，并特别支持AI加速卡、推理卡等扩展应用。

数据来源：公开资料，亿欧智库

GPGPU：沐曦全栈GPU产品兼具多精度算力，大容量高带宽内存，具备多卡互联能力，计算性能全力追赶英伟达A100

- 沐曦致力于为异构计算提供全栈GPU芯片及解决方案，可广泛应用于人工智能、智慧城市、数据中心、云计算、自动驾驶、数字孪生、元宇宙等前沿领域，为数字经济发展提供强大的算力支撑。
- 曦云®C500基于自主研发的高性能GPU IP，具有多精度混合算力、 64GB大容量高带宽内存、多卡互联技术、全兼容主流GPU生态的MXMACA®软件栈，适合千亿参数AI大模型的训练和推理。
- 曦云®C500千亿参数AI大模型训练及通用计算GPU与智谱AI开源的中英双语对话语言模型ChatGLM2-6B完成适配。测试结果显示，曦云®C500在智谱AI的升级版大模型上充分兼容、高效稳定运行。

曦云GPU全面兼容CUDA生态，可实现用户零成本迁移；通过自主知识产权的MetaXLink实现 单机8卡GPU全互联，提供构建高密度算力和云计算部署的优秀国产GPU解决方案；可广泛应用于千亿参数AI大模型训练与推理、AIGC内容生成、推荐系统、自动语音识别、语音合成、图像分割检测，以及科学计算、数据库加速等多种场景。

自主知识产权 GPGPU | 高精度及混合精度算力 | 片间互联 MetaXLink | 自主软件栈 MXMACA

曦云C500算力性能对比英伟达A100

产品代号	曦云 C500 OAM	Nvidia A100
算力	FP32：36 TFLOPS TF32：140 TFLOPS FP16：280 TFLOPS BF16：280 TFLOPS INT8：560 TOPS	FP32：19.5 TFLOPS TF32：156 TFLOPS FP16：312 TFLOPS BF16：312 TFLOPS INT8：624 TOPS
内容规格	64GB HBM2e，带宽1.8TB/s	80GB HBM2e，带宽1.9TB/s
视频/JPEG解码	160路1080p 30FPS	/
视频/JPEG解码	12路1080p 30FPS	/
互联	MetaXLink 8卡全互联	NVLink
虚拟化示例	1/2/4/8	/
功耗	450W	400W

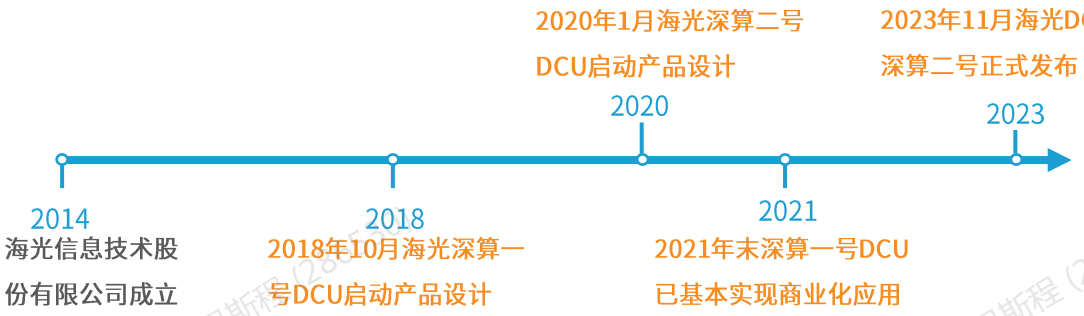
数据来源：公开资料，亿欧智库

GPGPU：海光DCU全面兼容ROCm生态系统，提高AI芯片国产化能力。

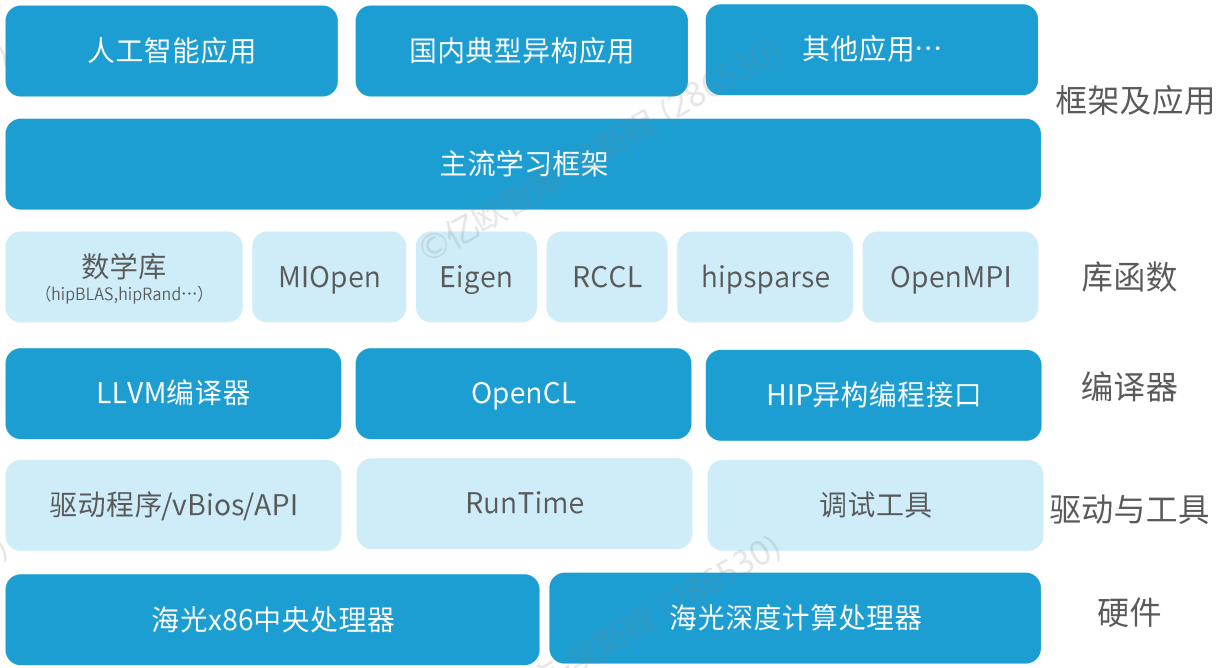
- 海光DCU以GPGPU架构为基础，兼容通用的“类CUDA”环境，可广泛应用于大数据处理、人工智能、商业计算等应用领域。
- 海光DCU在强大的通用计算性能基础上，打造出自主开放的完整软件栈，包括“DTK（DCU Toolkit）”、开发工具链、模型仓库等，完全兼容“CUDA”、“ROCm”生态，支持TensorFlow、Pytorch和PaddlePaddle等主流深度学习框架、应用软件。

HYGON

- 海光DCU深算一号产品FP64性能可达到英伟达2020年推出的A100和AMD 2020年推出的MI100水平。海光深算二号于2023年三季度发布，具有全精度浮点数据和各种常见整型数据计算能力，性能相对于深算一号提升100%以上。
- 在商业应用方面,公司的DCU产品已得到百度、阿里等互联网企业的认证,并推出联合方案。目前阿里BladeDISC前端支持PyTorch+TensorFlow，后端已完整支持的AI算力卡只有NVIDIA GPU和海光DCU.



海光完整软件栈支持



数据来源：公开资料，亿欧智库

AI服务器：万卡集群组网落地，全自研液冷技术，超云全力攻克大模型推理负载

- 受益于云端服务供应商（CSP）及品牌客户对建设AI基础设施的强劲需求，全球AI服务器市场蓬勃增长。其中，超云已率先成为国内少有的能够提供**万卡集群落地**的服务器厂商，既可以很好地配合客户做好组网、性能调优，将训练性能进行数倍的提升，也能提供先进的**冷板式液冷技术**，以应对AI应用庞大的散热需求。
- 随着大模型算力scaling law逐渐从训练转至推理，超云的AI战略也随着大模型的发展进行了转变。据IDC数据显示，2021年中国数据中心用于推理的服务器的市场份额占比已达到57.6%，预计到2026年，用于推理的工作负载占比将达到62.2%。因此，超云将“争做推理业务的第一品牌”提升至公司战略地位，立下了“**推理之巅、超云为先**”的目标。

超云 SUPER CLOUD

液冷整机柜服务器：基于超云全自研液冷技术，采用先进的**冷板式液冷技术**，集中散热设计模式，高密度节点部署。机架式和柜式CDU覆盖10-200KW，液冷覆盖整机柜80%散热量。

超云R2426：基于飞腾腾云S5000C平台推出的4U存储型服务器，适用于大数据分析、软件定义存储等应用场景，满足用户对**带宽和存储的扩展需求**。

冷板式液冷将液体于冷板中循环。冷板与主要发热源（如GPU）直接接触，最低可将PUE降低至1.1X。

浸没式液冷将服务器浸入液体中，达到控制服务器整体温度的效果，最低可将PUE降低至1.02X。

AI数据中心选择使用冷板式液冷的主要原因是其能够更加精准地针对CPU和GPU芯片进行散热，从而有效降低核心工作温度。作为一种新兴技术，液体散热系统相比传统空气散热系统具有更低能耗，实现更低PUE（电源使用效率）。另外，采用冷板式方案无需对传统数据中心的物理架构进行大幅改动，只需在机电设计上做一些调整即可，这使得其具有天然优势。而浸没式则需要改变整个机柜结构，不适合现有数据中心的改造升级。

万卡集群落地：

超云已成为国内少数具备万卡集群部署能力的服务器厂商之一，能够为客户提供从组网到性能优化的全方位支持，显著提升训练效率，最高可实现数倍性能提升。以"西云算力人工智能专用智算平台"项目为例，超云作为核心供应商，提供了包括R8428系列在内的近千台高性能GPU服务器，成功构建了新一代人工智能数据中心（AIDC），为宁夏地区打造人工智能算力中心提供了强有力的基础设施支撑。

R8428 A13是超云针对AI市场推出的一款高性能GPU服务器，基于AMD EPYC处理器设计，采用7nm先进制程工艺，最高支持单颗CPU 64核性能输出。在4U空间内集成8块双宽AI加速卡

数据来源：公开资料，亿欧智库

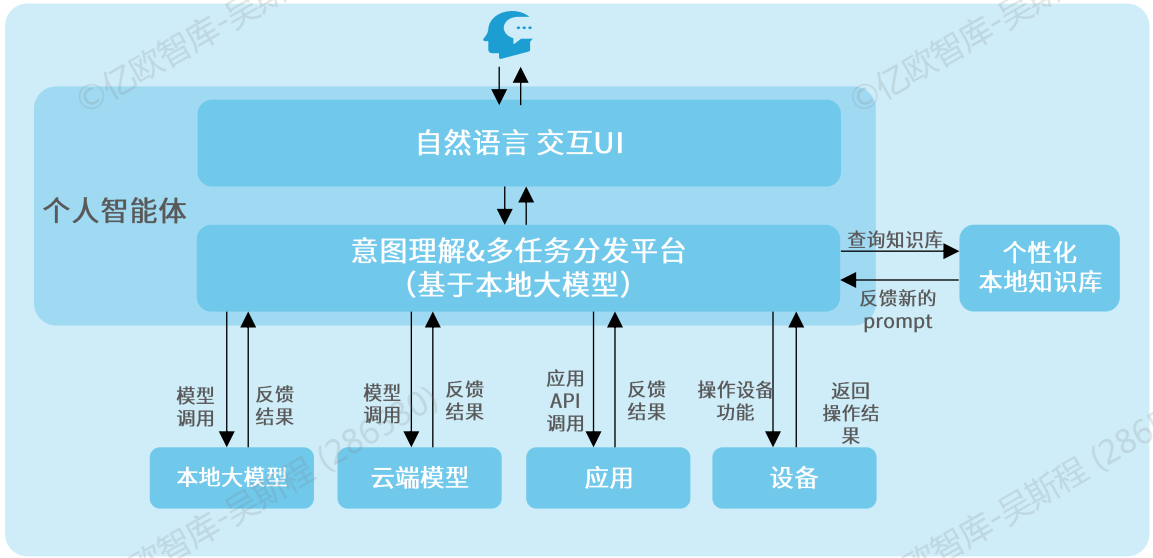
AI PC：联想推出首款国产AI PC，基于本地大模型打造端云无缝协同的全栈个性化AI能力

- 信创AI PC，即搭载国产AI芯片等软硬件的信创PC，相较于传统PC，AI PC具有强大的本地处理能力、高效的AI算法执行、出色的隐私保护特性以及个性化的服务，为信创用户创造本地个性化AI体验，同时也推动了计算机硬件和软件产业的升级换代。
- 联想开天基于兆芯开先KX-7000系列处理器平台推出了P90z G1t台式机，作为信创首款AI PC，配备联想开天自研的“小天智能体”，内嵌本地大模型，为用户搭建专属的个人知识库，并支持异构 GPU 加速。在本地大模型的加持下，打造了端云无缝协同的全栈AI能力，在联网状态下，云侧大模型将赋予用户更强的内容生成，知识检索，提炼分析及安全检测能力，全方位提升用户AI体验。

AI PC的未来定义与特征

个人agent自然语言交互：
- 多模态自然语言交互UI
- 基于本地大模型的意图理解和任务调度

内嵌个人大模型：
- 能够运行经压缩和性能优化的本地大模型
- 具备更大存储，覆盖个人全生命周期数据的本地个人知识库



标配本地混合云端算力：
- CPU+GPU+NPU本地算力
- 个人终端和家庭主机/企业主机协同运算

开放的AI应用生态：
- AI Agent & 本地大模型接口开放，接入第三方AI应用，可被个人Agent调度。
- 为AI应用开发者提供高效、便捷、低成本的混合AI算力开发及适配平台。

设备级个人数据&隐私安全防护：
- 本地化隐私推理&非敏感任务调用云端大模型
- 硬件级安全芯片保护&个人数据加密/脱敏传输。

Lenovo 联想

联想开天P90z G1t

- 硬件方面，联想开天AI PC元启版搭载兆芯开先KX-7000处理器。
- 软件方面，搭载面向信创用户深度定制的操作系统，针对信创软件进行深度优化；

数据来源：公开资料，亿欧智库

基础软件层面：全力围绕大模型应用灵活发展，构建基础模型与模型应用的关键桥梁

- 操作系统层面：大模型的训练和推理需要数以千计的GPU集群，操作系统必须能够**高效管理**这些**异构算力资源**。此外，随着AI引领物联网和边缘计算的兴起，**轻量化和模块化**的操作系统将成为趋势。
- 数据库层面：大模型的技术发展推动SQL与AI走向**一体化**。开发者将能够使用**AI辅助开发，AI性能优化**等功能，提升数据库使用体验和工作效率。
- 中间件：作为打通AI应用落地的最后一公里，中间件在**企业构建知识库**，落地过程中的**数据集成、应用集成、知识库与大模型融合**等环节起到关键作用。

	操作系统	数据库	中间件
技术挑战	**资源管理与调度：**随着AI工作负载变得越来越密集和复杂，操作系统需要更精细的资源管理和任务调度策略来确保系统性能和稳定性。 **安全性：**由于AI系统处理的数据往往包含敏感信息，操作系统的安全性问题也日益突出。	**一体化：**随着互联网技术与AI的融合，数据库正走向一体化，并开始提供多模数据的支持。 **数据库+AI：**AI向量数据库，向量融合查询，实时处理数据，快速构建AI应用的能力。	**构建企业知识库：**帮助企业形成自由的知识空间中的数据集成。 **集成：**AI应用与第三方SaaS软件的无缝集成。 **数据安全：**企业需要将私有数据发送给公有云大模型，因此具有一定安全隐患。
产品趋势	**轻量化与模块化：**随着AI引领物联网和边缘计算的兴起，轻量化和模块化的操作系统将成为趋势。微内核架构的灵活性和模块化特性或将使其成为重要发展方向，以满足不同设备对资源和功能的定制化需求。 **智能化管理：**操作系统将内嵌更多的智能化管理能力，例如使用机器学习算法进行自我优化和故障预测。	**SQL+AI一体化：**支持多模态数据融合查询。 **AI赋能数据库：**让开发者能够借助智能代码辅助功能、代码补全和直接在编辑器中提供的指导，轻松生成和汇总SQL代码。同时，内容感知聊天界面也可帮助开发者使用自然语言更快地构建数据库应用。	**简化开发：**大模型中间件目前为开发者提供的支持包括提供多模型访问、Prompt封装、多数据源接口，从而简化开发者构建AI应用的过程。 **统一接口：**也包括为大模型提供统一的接口来访问外部数据，例如语义检索，混合查找等。 **应用集成：**实时知识库构建，AI应用集成，大模型插件，以及无代码构建AI应用的中间件。

数据来源：公开资料，亿欧智库

操作系统：银河麒麟发布首个AI PC版本操作系统，实现人工智能与操作系统的深度融合

- 2024年8月8日，在北京召开的2024中国操作系统产业大会上，国产桌面操作系统银河麒麟推出了首个AI PC版本。这款系统将人工智能与操作系统深度融合，弥补了我国在操作系统端侧推理能力研发领域的空白。
- 通过应用与模型解耦、模型与AI芯片解耦，同源支持CPU、GPU等异构算力创新发展，AI PC OS操作系统实现对多种模型的统一管理和调度，帮助生态伙伴省去全栈调校的步骤，同时让用户在一个硬件设备中方便地运行多种模型，把PC打造成可用、好用、并实用的生产力工具。

KYLINSOFT 麒麟软件

银河麒麟桌面操作系统AI版的核心价值在于通过打造Kylin AI-SDK、Kylin DLA框架，并与操作系统深度融合，同源支持不同的AI芯片，实现不同AI芯片之上混合推理，为产业提供一致的AI生态支持。

银河麒麟操作系统

桌面操作系统V10

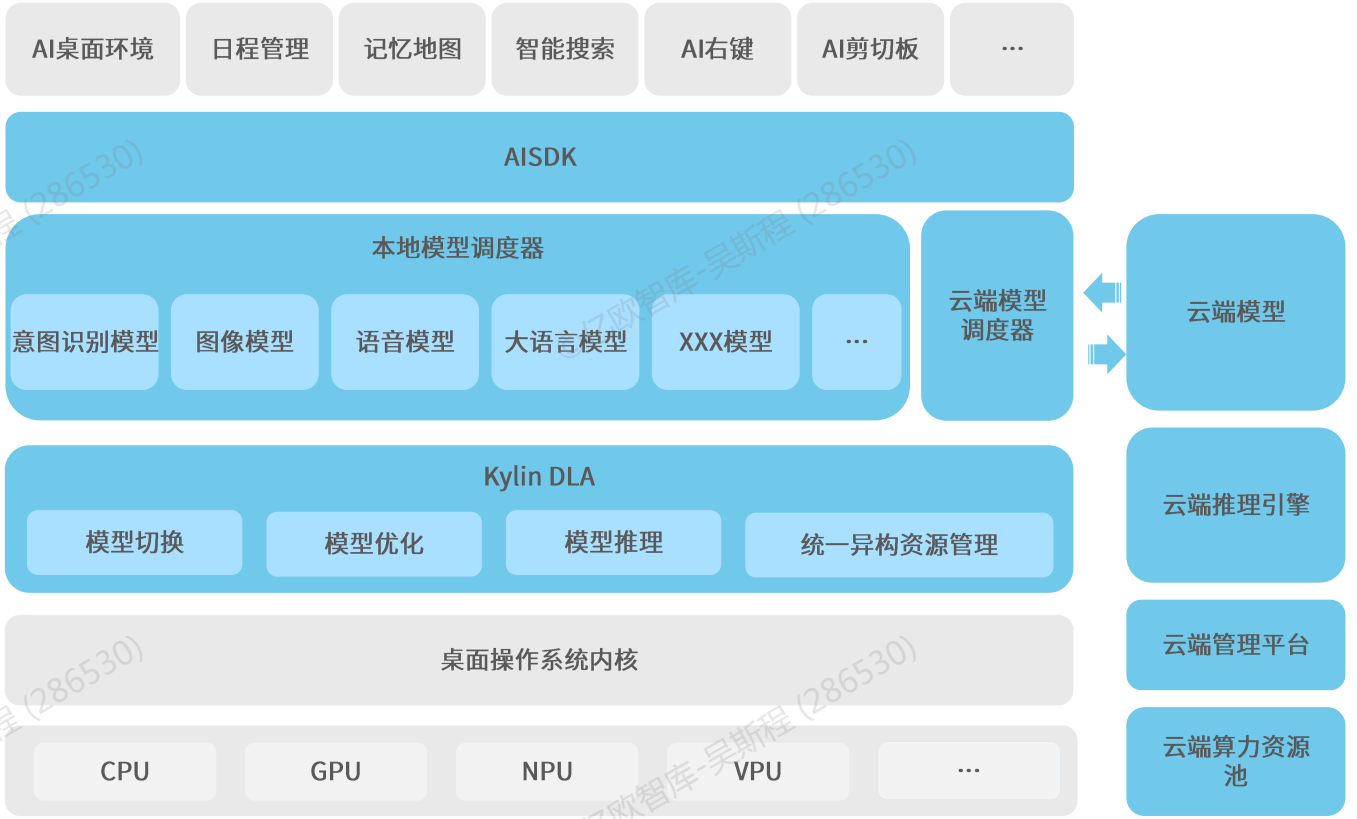
支持全CPU架构的笔记本、台式机、一体机和工作站，满足用户办公和娱乐需求。

服务器操作系统V10

在桌面版的基础上，向用户业务平台提供标准化服务，虚拟化、云计算等应用场景支撑。

嵌入式操作系统V10

在桌面版的基础上，针对专用设备应用场景进行系统裁剪和个性化定制。



数据来源：公开资料，亿欧智库

数据库：基于完全自主产权，Oceanbase深度融合AI与数据库处理能力

- 基于完全自主产权，OceanBase数据库已服务全部政策性银行、2/3国有大行，浦发银行、民生银行等股份制商业银行，以及北京银行、宁波银行等近百家银行。除此之外，OceanBase在保险、证券、基金、期货等金融子领域目前，全国已有1/4的头部金融客户将OceanBase作为核心系统升级首选。
- 10月23日，独立数据库厂商OceanBase在2024年度发布会上推出OceanBase 4.3.3GA版本，升级向量检索与索引功能，实现SQL+AI一体化。该版本深度融合AI与数据库处理能力，支持多模态数据的融合查询，帮助企业简化AI技术栈，提升AI应用构建效率。

OceanBase数据库一体机是基于自研金融级分布式数据库和可信硬件打造的软硬一体化数据库产品。

OceanBase数据库一体机提供以OceanBase数据库为核心的软硬一体化系统，集成自动化交付部署、监控工具。其内嵌了OceanBase管理者工具OAT、OceanBase数据库和OceanBase云平台OCP三大核心数据库软件，旨在为用户提供数据库迁移上云、数据库开发管理、运维管理等数据库全生命周期的一站式接入与管理服务。

OCEANBASE

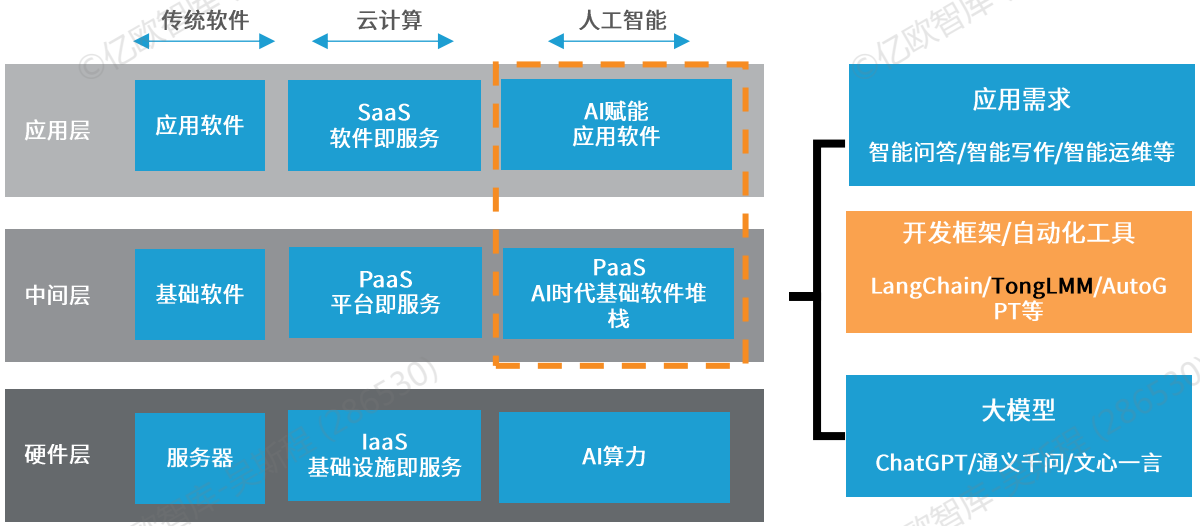
数据来源：公开资料，亿欧智库

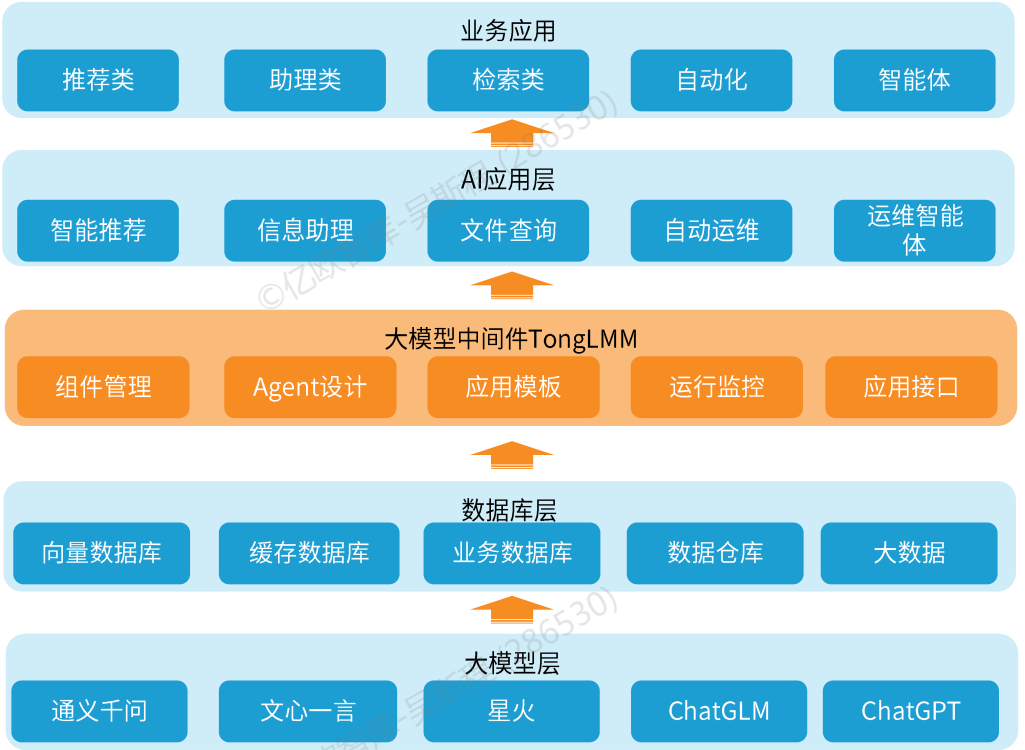
中间件：东方通推出大模型中间件TongLMM，快速部署AI应用，实现大模型与业务的高效结合

- 中间件作为应用与大模型之间的桥梁，是推动AI技术落地和产业升级的重要力量，东方通作为中国中间件领域的开拓者和领导者，率先在AI领域布局探索，从技术角度解耦大模型对接能力并做强中间层，推出了大模型中间件TongLMM产品。
- 通过推动大模型中间件发展，东方通充分解决用户在使用大模型能力过程中存在模型生成的内容与现实世界事实不一致、训练专属模型投资大、现有应用难与大模型全面融入、存在安全隐患等普遍性难题，让更多行业客户能够更简单、便捷的构建大模型应用。

东方通 TongTech

TongLMM支持私有化模型方式部署，作为应用与大模型之间的桥梁，TongLMM融入消息中间件、缓存中间件、数据交换、自然语言处理等现有组件，能解决大模型落地过程中数据集成、应用集成、知识库与大模型融合、数据不精准、安全隐患等问题。





数据来源：公开资料，亿欧智库

大模型层面：新架构，新范式引领大模型进入新阶段

◆ 根据OpenAI提出的AGI发展路线图，目前行业主流大模型均处于L1聊天机器人阶段。目前的各类大模型应用，无论是文案撰或是代码助手，依然是以进行对话为主要原理机制所衍生出的应用，其本质仍以预测语言模式为基础，而未达到真正意义上的通用理解和推理。而近期新发布的OpenAI o1以及DeepSeek R1等强化学习（RL）模型通过慢思考（增加推理时间）以及思维链技术正在向L2推理者进发。

大模型能力等级

◆ 目前AI大模型的主流技术发展路线包括RAG，长上下文，强化学习，多模态/世界模型。这四种技术路线由于各自不同的性质和能力，目前面临着不同的技术挑战和发展趋势。

	RAG	长文本	强化学习	多模态/世界模型
技术挑战	RAG方案时常会出现落地**冷启动**问题。尤其在目前大部分传统行业都没有成熟的AI技术栈（**向量检索引擎、多模态数据库等**）且业务诉求较为复杂的情况下，很难使用通用的落地方案。需要协同整个生态通过多层的能力解决部署问题。	长文本方案目前的技术挑战主要为**推理成本高以及响应速度慢**两方面： 模型的并发性能会随着文本长度的增加而反比下降。预填充的延迟也会随文本长度的增长而平方级别的增长。	强化学习模型的推理**成本高且响应速度慢**，需要大量的计算资源和时间。 在泛化到更抽象的任务或管理状态复杂时仍存在挑战，特别是当问题上下文变得更加抽象时，在泛化方面存在一定困难。	总体发展尚处于**初级阶段**。目前的问题包括模型的感知能力有限导致视觉信息不完整或不正确，多模态大模型的指令跟随性不足，微调方法和数据集无法完全覆盖模型所需的各种指令场景等。
发展趋势	RAG方案目前来看是金融机构的首选。因其对于大数据库抽取，**细节性问题、简单问题**的表现更好，模型自身的准确率更高，可解释性更强，符合金融机构的低容错需求。另外，RAG可以在**预算仅有几十张甚至几张GPU**的情况下运行，落地相对容易。	在客服/销售agent等领域，目前长文本在简单知识库中具有相对优势。并且在面对用户侧的**刁钻问题**时表现更好。 随着推理优化方案进一步发展以及计算成本越来越低，文本窗口将越来越长，未来应用场景将不断扩大。	强化学习模型由于其出色的**逻辑推理能力**，目前在数学，物理，生物，编程等理工科能力上表现优秀。 通过推理阶段的计算量扩展以及思维链回溯，未来有望在**慢思考场景**，例如策略，规划等，通过更长时间的思考获得明显的优势。	为了解决大模型缺乏对世界认知的问题，世界模型成为了目前AI领域最热门的研究方向之一。 无论世界模型采用自回归或是JEPA方式，多模态模型都是不可或缺的一部分。目前多模态模型的研究大致可以分为对齐、融合、自监督和噪声添加几种途径。

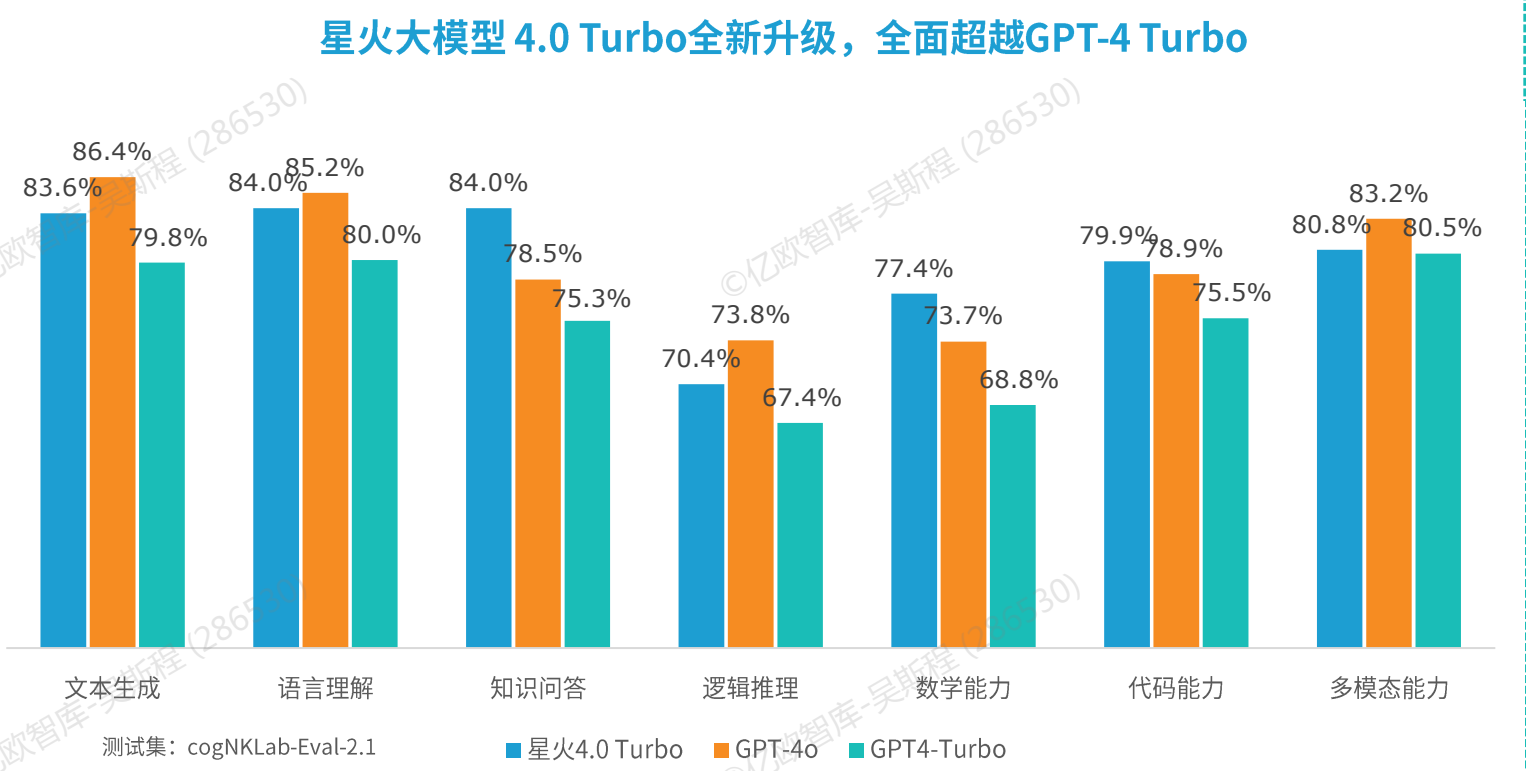
数据来源：公开资料，亿欧智库

大模型：科大讯飞发布讯飞星火大模型4.0 Turbo，底座能力全面对标GPT-4Turbo

- 2024年10月24日科大讯飞全球1024开发者节上，科大讯飞正式发布“讯飞星火4.0 Turbo”。模型性能迎来全新升级，七大能力全面超越GPT-4 Turbo，数学能力、代码能力超过GPT-4o。与此同时，国产超大规模智算平台“飞星二号”正式启动。自2023年第一款国产万卡算力集群“飞星一号”上线以来，飞星二号将带来新模型新算法的持续适配和智算集群规模的再次跃迁。
- 在6月25日，OpenAI正式通知将开始终止所有来自中国的API申请的背景下，科大讯飞董事长刘庆峰说，只有自主安全的繁荣生态，才有中国通用人工智能的大未来。自去年5月6日发布以来，讯飞星火大模型正成为国家能源集团、中国石油、中国移动、中国人保、太平洋保险、交通银行、奇瑞汽车、中国一汽、大众汽车、江汽集团、海尔集团等多领域头部企业的首选。

值得重点关注的是，在数学能力方面，讯飞星火4.0 Turbo已完成**强化学习**的**超长思维链**、**树搜索**和**自我反思评价**等算法验证，预计今年底可实现类似OpenAI o1模型的高难度数学能力。

同时，讯飞星火4.0 Turbo也推出了星火代码7B版本，主打端侧本地运行，能够满足文本编辑，代码生成、代码补全等任务。



数据来源：公开资料，亿欧智库

大模型：360智脑大模型接入360全家桶，开源模型360Zhinao2-7B中文能力表现出色

- 三六零持续围绕“AI和安全”两条主线，保持高研发投入。公司聚焦大模型前沿技术和AI安全问题，保障国产大模型的发展“自主安全。2024年来，公司相继发布了“纳米搜索（原360AI搜索）”、“360AI浏览器”和“360AI办公”三大AI原生产品。同时，公司将大模型与安全结合，融合大模型技术、方法论、全网数据和专家知识，发布行业首个免费安全大模型——“360安全大模型”。
- 开源模型360Zhinao-7B使用了包含3.4万亿个标记的高质量语料库（主要包括中文、英文和代码），具备强大的聊天能力，并支持三种上下文长度：4K、32K 和 360K。其中，360K（约50万汉字）是截至2024年4月11日发布时中文开源模型中最长的上下文长度。而升级版360Zhinao2-7B采用流行的两阶段训练方法，总训练数据量达10.1万亿个token，在CEval、C3 和 LCSTS 等中文基准测试中取得了良好成绩。其中文基准测试的平均分排名第一，同时在具有挑战性的数学竞赛数据集 Math 上也位居第一。

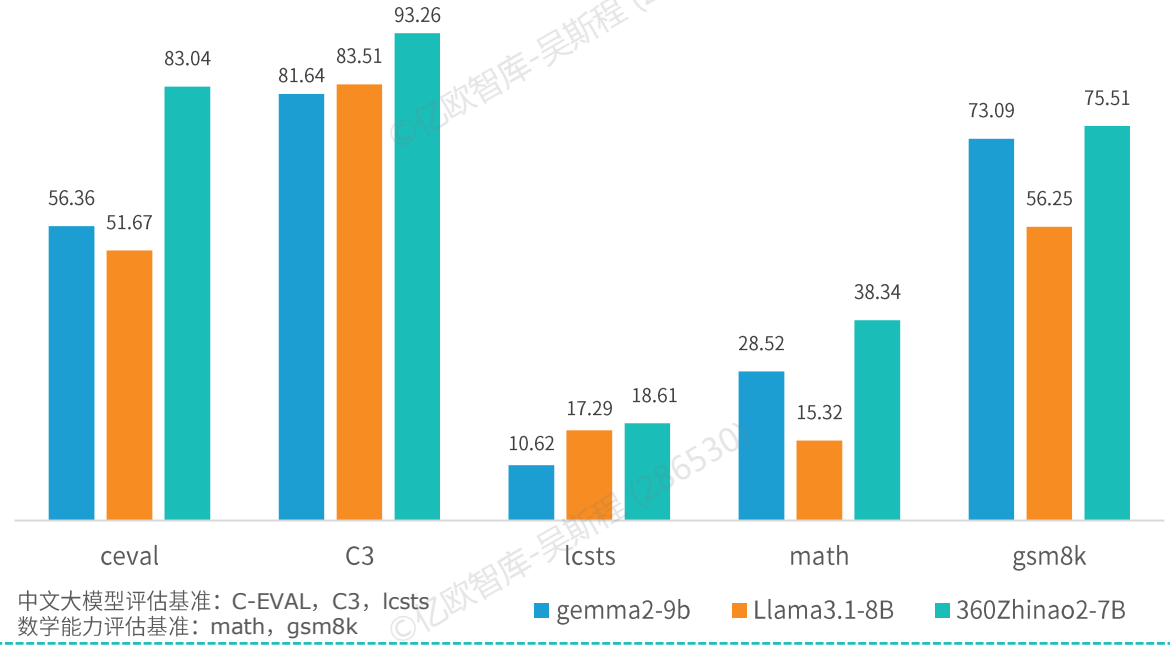
360智脑全面接入360互联网全端应用场景

360智脑 生成式大模型

360智脑

视频	数字化	图片创作	桌面办公
快剪辑 · SaaS 视频创作应用平台	AI数字员工 数字人应用平台	360智绘 图片创作应用平台	360安全卫士 桌面&办公应用平台
浏览器	**多模态**	**内容创作**	**搜索**
360AI浏览器 浏览器应用平台	360视觉云 多模态应用平台	快资讯 内容创作应用平台	360AI搜索 搜索应用平台

360Zhinao2-7B 在中文和数学方面优势明显



中文大模型评估基准：C-EVAL，C3，lcsts
数学能力评估基准：math，gsm8k

数据来源：公开资料，亿欧智库

生态应用层面：大模型能力不断演进提升，落地场景逐渐丰富

- 随着AI大模型能力的不断演进提升，大模型的应用方案正在深度嵌入各个行业企业运营的每项环节，大致可分为研发，生产，运营和营销四个方面。
- 嵌入式AI产品能够进一步提升用户与产品的交互体验，并可以根据非标准问题提供针对性分析，深入洞察企业运营、识别潜在的业务风险和机会，为企业决策提供针对性建议。

	研发	生产	运营	营销
落地场景	**AI技术研发：**AI目前在**材料、药物、软件和游戏**等领域均有较好的表现和发展前景。例如使用图神经网络发现新材料，使用AlphaFold建模蛋白和DNA，使用大语言模型进行编程，软件研发等。	**AI制造：**通过**数字孪生**技术模拟真实工厂的生产环境，通过3D可视化呈现出整个制造过程，帮助企业评估制造流程的效率、优化生产线布局并降低成本。通过**机器视觉，智能声学**等技术实现生产过程优化，包括无序分拣、缺陷检测、仓储自动化等方式赋能企业的生产制造。	**运营自动化：**将大模型融入企业日常运营，协助数据录入、邮件的分类与总结、合同管理与审核，会议的主持与总结等，帮助减少人工错误，提高工作效率。 **AI知识管理：**利用RAG技术将大模型与已有知识检索相结合，实现企业智能客服、企业智能文档、专家顾问助理等商业场景。	**客户画像分析：**AI技术通过深度挖掘客户的购买记录、描绘出更为精准的客户画像，帮助企业制定销售策略。 **个性化营销：** 通过引入AI客服技术， 并基于个性化推荐系统，企业能够深入挖掘客户的需求和偏好，为客户提供个性化的产品推荐以及服务。
技术挑战	**生成结果的测试和评审：**针对AI生成结果的评审标准尚不完善，例如在药物研发中，AI生成的化合物需要经过**多轮实验和临床评审**，但AI评审标准尚未完全统一。 **与传统研发方法的融合：**AI辅助研发需要与传统研发方法无缝衔接，但两者的流程和工具可能存在差异。	**软硬适配：**制造业生产场景十分丰富，企业对于智能产线、智能工厂的需求往往都是非标准化的，因此AI技术的融合会面临硬件平台和应用软件适配问题。 **数据处理：**融入AI进行诸如机器视觉、预测性维护应用过程中，需要处理大规模数据和高并发请求。	**集成与兼容性：**AI系统需要与企业现有的营销工具、CRM系统、知识库系统等进行集成。然而，现有系统可能缺乏或无法识别相应API接口，导致调用数据困难。	**缺乏个性化：**目前AI生成的销售文案往往缺乏个性化，导致客户的回应率远低于人类销售人员。 **版权纠纷：**生成式营销所使用的算法通常是基于大量现有数据进行训练的，容易引发版权纠纷，面临法律风险。

数据来源：公开资料，亿欧智库

生态应用：用友网络推出企业服务大模型YonGPT，加速投入AI应用普及浪潮。

- YonGPT是用友商业创新平台用友BIP的重要组成。为顺应AI普及应用浪潮，用友在2023年春的用友BIP技术大会上宣布启动企业服务大模型训练。
- YonGPT底层适配文心一言等业界主流的通用语言大模型，通过上下文记忆、知识/库表索引、Prompt工程、Agent执行、通用工具集等扩充大模型的存储记忆、适配应用和调度执行能力，再结合财税、人力、供应链、研发等领域的知识和最佳实践扩充大模型专业能力，从而形成体系化的企业服务大模型。
- 基于大模型的人工智能在企业服务领域的应用主要集中在4个方向上：智能化的业务运营、自然化的人机交互、智慧化的知识生成、语义化的应用生成。用友企业服务大模型YonGPT围绕这四个方向推进模型训练和产品效果优化。

用友 yonyou

L2 领域/行业模型精调 智能化场景服务

	业务运营	人机交互	知识生成	应用生成
行业服务	政务	教育	医疗	制造	金融	汽车	建筑	餐饮	能源
领域服务	财务云	人力云	供应链云	采购云	制造云	营销云	研发云	项目云	资产云	协同云
通用能力服务	语义逻辑	内容生成	多轮对话	知识问答	角色扮演	逻辑推理	代码生成	图像生成		

L1 YonGPT 企业服务大模型

财税领域知识 | 人力领域知识 | YonGPT大模型 | 供应链领域知识 | … | 研发领域知识

通用工具集	上下文记忆	Agent执行	Prompt工程	知识/库表索引
数据管理	大模型训练	大模型评估优化	大模型推理	插件服务

L0 通用大模型底座

智普ChatGLM | 百度文心一言 | 智源悟道天鹰 | LLaMA | ChatGPT | 其他…

数据来源：公开资料，亿欧智库

生态应用：中关村发布科金得助大模型平台2.0，助推大模型应用加速落地

- 中关村科金自研的得助大模型平台通过算力统一调度、一站式模型训推和应用快速构建三大核心能力，帮助企业快速构建和部署大模型应用，显著降低企业大模型落地成本。
- 平台+应用+服务，推动大模型落地：中关村科金的“三级引擎战略”是“平台+应用+服务”，全方位助力大模型解决方案落地。

中关村科金

中关村科金大模型时代的“三级引擎战略”

- **平台方面：**得助大模型平台2.0通过算力统一调度、一站式模型训推和应用快速构建，实现了分钟级构建大模型应用的能力，并且将行业和领域主流场景的业务流程、话术、规则、内容等核心要素进行统一封装，沉淀出上百套全场景套件，让企业在部署同类场景大模型应用时做到开箱即用。
- **应用方面：**得助大模型平台2.0面向**智能营销、智能客服、智能运营**和**知识管理**四大核心场景，推出了多款大模型应用。在智能营销场景中，大模型外呼。通过综合运用**增强RAG**和**多Agent协同**等多种大模型技术，该系统实现了高拟人度和高水准的专业度，在营销转化话术上展现出实战技巧。
- **服务方面：**中关村科金构建了端到端的价值交付体系，确保企业大模型落地的最后一公里。这一体系围绕大模型价值实现的全流程，从客户咨询到平台、应用再到运营的各个环节都提供了全方位的支持。据悉，中关村科金通过丰富的行业服务经验积累，已经服务了1600余家行业头部企业，与银行、政务、央国企、保险、财富管理、制造、零售、家装等头部企业在大模型应用落地上合作。

数据来源：公开资料，亿欧智库

亿欧智库

目录

CONTENTS

中美AI基建正面交锋，算力“卡脖子”层层加码，AI产业自主安全愈加重要

- 2025年1月24日，美国总统特朗普上任第二天，牵头软银、OpenAI和甲骨文等在白宫宣布了一项名为“星际之门”的计划，宣称在未来四年内投资5000亿美元在美国打造全新的AI基础设施，Arm、微软、英伟达等作为技术合作伙伴。尽管5000亿美元的投资额遭到包括美国政府效率部（DOGE）马斯克等人士的严重质疑，这一项目的宣布代表着中美开始正式进入AI基建军备竞赛时期。
- 自2022年10月至2024年12月，美国商务部不断对华出台算力“卡脖子”政策，从算力芯片到芯片制造设备，再到算力芯片的关键零组件HBM，性能指标和产品种类受限程度不断严苛。中国算力芯片产业亟需发展自主安全产业链，提高自主设计及生产能力，突破技术封锁。

亿欧智库：美国对华算力芯片出口限制政策

时间	限制政策	主要出口受限产品
2022年10月	美国商务部工业和安全局（BIS）更新出口管制规则，明确禁止向中国出口算力大于4800且带宽大于600GB/S的高性能AI芯片	英伟达A100及H100 GPU等
2023年1月	美、日、荷达成秘密协议对华设限，美国政府向荷兰发出强制指令，限制对中国的深紫外(DUV)光刻机及其部件出口	深紫外(DUV)光刻机及其部件
2023年10月	美国商务部发布《先进计算芯片及相关物项规则》修订版，扩大了管制物项范围，取消了“互连带宽”作为识别高性能芯片的标准，改为以总计算能力和性能密度为核心指标	英伟达A800、H800、L40S GPU等
2024年12月	美国商务部工业和安全局（BIS）发布出口管制的“强化版”新规（IFR，Interim final rule）、（Final rule），主要针对高带宽存储芯片以及半导体制造设备及软件	HBM2e、HBM3、HBM3e等

亿欧智库：中美重大算力基建项目

	项目	建设内容	支持力度
中国	东数西算工程	建设8个国家算力枢纽节点和10个国家数据中心集群，新建数据中心规模超过110标准机架	总投资额超过4000亿元
	中国银行支持人工智能产业链发展行动方案	重点支持智算中心及配套设施和园区基础设施建设	五年不低于1万亿元金融支持
	国家大基金三期	重点投向集成电路全产业链，如先进封装，高端存储（如HBM）	基金注册资本3440亿元
美国	星际之门计划	建设下一代人工智能所需的物理与虚拟基础设施，包括数据中心、计算资源、网络设施等，以确保美国在人工智能领域的全球领导地位。	总投资额5000亿美元，分四年实施，初期投入1000亿美元
	芯片与科学法案	到2030年将承担全球至少20%的先进逻辑芯片生产，并于2035年前生产约10%的先进DRAM芯片。打造完整半导体供应链	整体规划投资约达4500亿美元

数据来源：公开资料，亿欧智库

预训练扩展效果放缓，后训练+推理扩展开启新范式，国产推理模型比肩OpenAI o1

- 自OpenAI于2024年9月12日发布新模型o1，强化学习+思维链等后训练和推理算法正式成为了AI大模型的新发展方向。OpenAI o1以及DeepSeek R1等模型通过强化学习（RL）以及思维链（CoT）技术扩展模型推理时间（inference-time scaling），开启了大模型推理方向的scaling law。
- 前OpenAI首席科学家Ilya Sutskever，同时也是Scaling Law的提出者之一，同样表示预训练阶段的扩展效果已经趋于平缓，扩展正确的东西比以往任何时候都更重要。
- 国产大模型中，DeepSeek R1，阿里巴巴Qwen团队的QwQ 32B，阿里巴巴数字化商业团队(非Qwen团队)的Marco-o1，以及月之暗面的Kimi K1.5同样在后训练+推理扩展这一方向取得重大突破，性能比肩OpenAI o1，且均有各自的创新亮点。

DeepSeek R1

- DeepSeek R1突破性的冷启动强化学习算法首次公开研究验证了通过纯强化学习即可激励大语言模型的推理能力，而无需依赖SFT，为AI未来的发展开辟一条崭新的道路。
- 性能方面，DeepSeek R1在推理任务中的各项评价指标已达到与OpenAI-o1相当甚至领先的水平。

QwQ 32B

- QwQ-32B-Preview 是由 Qwen 团队开发的实验性研究模型，同样使用了强化学习与思维链技术。
- 32B的参数规模向人们展示了，通过强化学习和推理计算的扩展，即便基础模型的规模较小也可以达到相对出色的性能。

Marco-o1

- Marco-o1聚焦开放式问题推理，目标是实现跨多个领域的泛化，尤其是在一些没有严格评估指标的领域，例如俚语翻译。
- 值得关注的是，Marco-o1通过诸如开源思维链(CoT) 微调、蒙特卡洛树搜索(MCTS) 和推理动作策略等技术，实现了全新的推理行动策略和反思机制（Marco-o1-MCTS Mini-Step）。

Kimi K1.5

- Kimi K1.5坚持以往的长文本路线，并以长文本为核心，将RL 应用于长文本的思维链（CoT）推理过程，使模型能够进行更深入、更复杂的推理。
- Kimi K1.5通过长文本 CoT 模型来指导短文本 CoT 模型的训练，从而在有限的计算资源下获得更好的性能。

	OpenAI o1	DeepSeek R1	QwQ 32B-preview	Kimi K1.5
GPQA	75.7	71.5	65.2	/
MATH-500	96.4	97.3	90.6	96.2
AIME	79.2	79.8	50	77.5
LiveCodeBench	67.2	/	50	62.5
Codeforces	94	96.3	62	94

注：GPQA是一个研究生级别的高难度科学问答基准，旨在评估模型在复杂科学问题上的推理能力；MATH-500是包含500个数学测试样本的评测集，用于全面评估模型的数学解题能力；AIME是一个涵盖中学数学各类主题的综合评测；LiveCodeBench是一个高难度的编程评测集，评估模型在实际编程场景中的代码生成和问题解决能力；Codeforces是一个在线编程竞赛平台，用于评估模型在算法竞赛中的表现

数据来源：公开资料，亿欧智库

关于亿欧

◆ 团队介绍：

亿欧智库（EO Intelligence）是亿欧旗下的研究与咨询机构。为全球企业和政府决策者提供行业研究、投资分析和创新咨询服务。亿欧智库对前沿领域保持着敏锐的洞察，具有独创的方法论和模型，服务能力和质量获得客户的广泛认可。

亿欧智库长期深耕新科技、消费、大健康、汽车出行、产业/工业、金融、碳中和等领域，旗下近100名分析师均毕业于名校，绝大多数具有丰富的从业经验；亿欧智库是中国极少数能同时生产中英文深度分析和专业报告的机构，分析师的研究成果和洞察经常被全球顶级媒体采访和引用。

以专业为本，借助亿欧网和亿欧国际网站的传播优势，亿欧智库的研究成果在影响力上往往数倍于同行。同时，亿欧内部拥有一个由数万名科技和产业高端专家构成的资源库，使亿欧智库的研究和咨询有强大支撑，更具洞察性和落地性。

◆报告作者：

曲俊汶

亿欧智库 分析师

Email：qujunwen@iyiou.com

◆报告审核：

严方圆

亿欧智库 咨询总监

Email：yanfangyuan@iyiou.com

孙毅颂

亿欧智库 研究总监

Email：sunyisong@iyiou.com

关于亿欧

◆ 版权声明：

本报告所采用的数据均来自合规渠道，分析逻辑基于智库的专业理解，清晰准确地反映了作者的研究观点。本报告仅在相关法律许可的情况下发放，并仅为提供信息而发放，概不构成任何广告。在任何情况下，本报告中的信息或所表述的意见均不构成对任何人的投资建议。本报告的信息来源于已公开的资料，亿欧智库对该等信息的准确性、完整性或可靠性作尽可能的追求但不作任何保证。本报告所载的资料、意见及推测仅反映亿欧智库于发布本报告当日之前的判断，在不同时期，亿欧智库可发出与本报告所载资料、意见及推测不一致的报告。亿欧智库不保证本报告所含信息保持在最新状态。同时，亿欧智库对本报告所含信息可在不发出通知的情形下做出修改，读者可自行关注相应的更新或修改。

◆ 关于我们：

亿欧是一家专注科技+产业+投资的信息平台和智库；成立于2014年2月，总部位于北京，在上海、深圳、南京、纽约设有分公司。亿欧立足中国、影响全球，用户/客户覆盖超过50个国家或地区。

亿欧旗下的产品和服务包括：信息平台亿欧网（iyiou.com）、亿欧国际站（EqualOcean.com）、研究和咨询服务亿欧智库（EO Intelligence），产业和投融资数据产品亿欧数据（EO Data）；行业垂直子公司亿欧大健康（EO Healthcare）和亿欧汽车（EO Auto）等。

亿欧服务

- 基于自身的研究和咨询能力，同时借助亿欧网和亿欧国际网站的传播优势；亿欧为创业公司、大型企业、政府机构、机构投资者等客户类型提供有针对性的服务。

- 创业公司

  亿欧旗下的亿欧网和亿欧国际站是创业创新领域的知名信息平台，是各类VC机构、产业基金、创业者和政府产业部门重点关注的平台。创业公司被亿欧网和亿欧国际站报道后，能获得巨大的品牌曝光，有利于降低融资过程中的解释成本；同时，对于吸引上下游合作伙伴及招募人才有积极作用。对于优质的创业公司，还可以作为案例纳入亿欧智库的相关报告，树立权威的行业地位。

- 大型企业

  凭借对科技+产业+投资的深刻理解，亿欧除了为一些大型企业提供品牌服务外，更多地基于自身的研究能力和第三方视角，为大型企业提供行业研究、用户研究、投资分析和创新咨询等服务。同时，亿欧有实时更新的产业数据库和广泛的链接能力，能为大型企业进行产品落地和布局生态提供支持。

亿欧服务

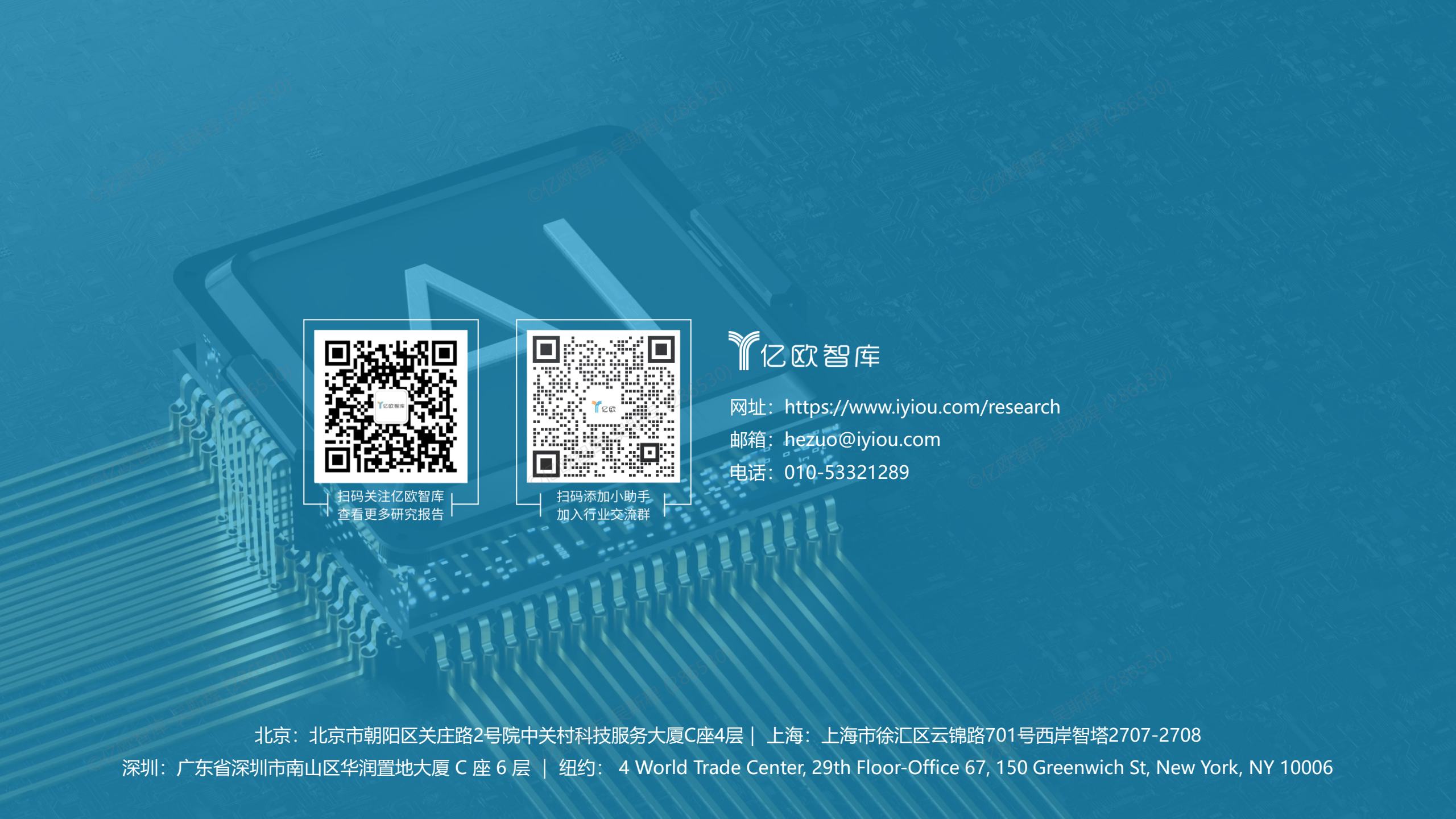
- 政府机构

  针对政府类客户，亿欧提供四类服务：一是针对政府重点关注的领域提供产业情报，梳理特定产业在国内外的动态和前沿趋势，为相关政府领导提供智库外脑。二是根据政府的要求，组织相关产业的代表性企业和政府机构沟通交流，探讨合作机会；三是针对政府机构和旗下的产业园区，提供有针对性的产业培训，提升行业认知、提高招商和服务域内企业的水平；四是辅助政府机构做产业规划。

- 机构投资者

  亿欧除了有强大的分析师团队外，另外有一个超过15000名专家的资源库；能为机构投资者提供专家咨询、和标的调研服务，减少投资过程中的信息不对称，做出正确的投资决策。

- 欢迎合作需求方联系我们，一起携手进步；电话 010-53321289，邮箱 hezuo@iyiou.com



扫码关注亿欧智库
查看更多研究报告

扫码添加小助手
加入行业交流群

亿欧智库

网址：https://www.iyiou.com/research

邮箱：hezuo@iyiou.com

电话：010-53321289

北京：北京市朝阳区关庄路2号院中关村科技服务大厦C座4层 | 上海：上海市徐汇区云锦路701号西岸智塔2707-2708

深圳：广东省深圳市南山区华润置地大厦 C 座 6 层 | 纽约： 4 World Trade Center, 29th Floor-Office 67, 150 Greenwich St, New York, NY 10006