



双轮驱动人工智能产业加速发展

—— “算力×联接” 引领百业智能的新征程



算力基础设施成为数字经济底层架构

我国算力基础设施高速增长，算存规模位居全球第二，国内企业深度参与建设和应用，构成数字经济发展的基础动力

- 根据工信部数据，截至2025年6月底，我国在用算力中心标准机架数达1085万架，智能算力规模达788EFLOPS，存力总规模超过1680 EB，全国算力中心平均电能利用效率（PUE）降至1.42，算力基础设施规模和水平不断提升，算力基础设施总规模位居全球第二；
- 国内涌现出“东数西算”等一批重大项目，其中八大国家枢纽节点直接投资超过435亿元，拉动投资超过2000亿元，机架总规模超过195万架，整体上架率达63%左右；
- IDC预计，2025年采用GPU等技术的智算服务器，市场增速是通算服务器的4倍左右，市场占比将超过一半。

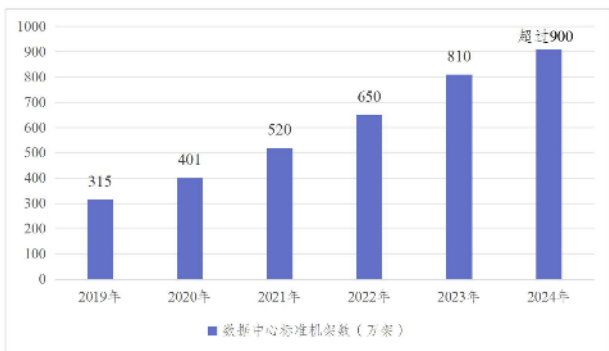


图6 2019年—2024年数据中心标准机架数

数据来源：工业和信息化部





我国算力基础设施已经进入科技创新和产业创新融合发展加速的阶段，以融合促创新、以融合促发展。

产业的数智化升级，成为融合发展的重要样板。根据工信部数据，目前我国已建成3万余家基础级智能工厂、1200余家先进级智能工厂、230余家卓越级智能工厂，这些类型的智能工厂覆盖超过80%的制造业行业大类，工厂产品研发周期平均缩短28.4%，生产效率平均提升22.3%。



2005年的发动机生产线，重度依赖人力操作



2024年的智能工厂，工业机器臂进行自动化操作



数智技术不仅用于自动化生产环节，而且深入到计划、设计、物流、销售等经济全流程，在提升产业竞争力同时，也为数智技术产业创造出更丰富的应用场景、更大的市场空间，实现产业发展与数字技术的融合。





提供更高效、更强大算力是当前算力技术热点之一

随着算力规模快速扩大，如何解决算力供应、设备互联和散热等问题，提供更高效、更强大的算力，是当前的技术热点

典型如大模型训练，需要数千块乃至数十万块GPU提供算力，电力消耗达到兆瓦级别。传统数据中心在算力规模、密度、功率和散热等方面均难以支撑。

目前业界主要采纵向扩展和横向扩展两种思路，典型产品正在走向市场，这两种思路也在逐步融合，探索更具备弹性和场景适应能力的解决方案。

纵向扩展：

在单台设备中提供数十个至数百个GPU，在有限空间、电力供应和散热条件下实现超高密度算力。



横向扩展：

通过多个设备互联，构成大规模计算集群，可提供高达数十万个GPU的超大算力规模。

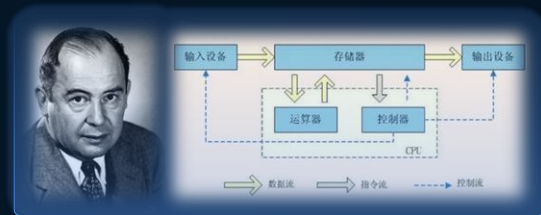




AI智能涌现带动ICT基础技术发展，经典架构瓶颈凸显

AI智能涌现初步实现了人机自然交互，有效提升协作效率，但对算法、数据、算力的要求较传统机器学习等技术更高。

AI智能涌现需投入大量资源且并行使用，小到芯片大到智算中心，性能、规模提升已受到经典架构瓶颈约束。



内存墙现象

- 处理器从内存读取指令和数据，完成后将结果写回内存。
- 处理器运算速度远超内存读写速度，数据频繁在存储器和处理器间搬运，浪费了时间。

联接需求激增

- 并行技术需要同时传递海量数据，增加网络带宽需求。
- 模型训练同步所有计算节点同步，网络需低时延转发。
- 丢包影响全局训练，网络需更高可靠性

数算相互等待

- 数据资源与算力资源不在一起，出现算力等数据或数据等算力情况。
- 为解决相互等待现象，智算架构需要引入复杂的调度、联接能力。



多系统协同发展努力突破传统架构瓶颈，为ICT发展提供新发展方向

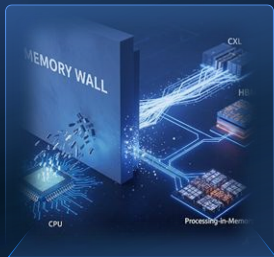
算网存技术理念相互影响、协同发展，业界已出现的技术有望打破架构瓶颈，并为AI ICT硬件领域发展指出方向

计算
&
存储
▼
突破内存墙

CXL实现内存池化

HBM集成逻辑单元

RAM集成逻辑单元



长期发展算存一体技术：

- ReRAM（阻变存储器）
- PCM（相变存储器）
- MRAM（磁存储器）

计算
&
网络
▼
增强联接力

确定性网络联接

广域网RDMA

NCCL/xCCL通信库



持续完善在网计算体系：

- 智能网卡
- 可编程设备及网络
- 算力路由等协议

计算范式在过去几十年中持续发展，历经：
集群计算 > 分布式计算 > 云计算 > 泛在计算

云+网+边+端



算力协同，范式革新

算网存协同供给，创新计算范式
算力密度及能效比实现10x级别提升，将有效
满足各产业的数字化转型需求。



异构计算发展弥补单一芯片短板，持续提升AI能力

异构计算打破单一芯片在AI业务中的应用局限，通过整合资源提供最优算力，让AI更高效地服务于各领域发展。

通用性

芯片能支持的算法和应用场景的广泛程度

能效比

完成同样计算任务时，芯片所消耗的电能

并行能力

同时处理多任务的能力，决定AI的效率

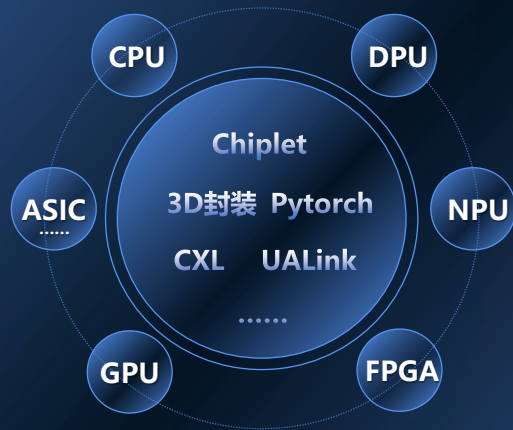
适应能力

开发和部署的难度及适应算法变化的能力

成本效益

特定规模下，每单位性能所需要的成本

芯片类型	通用性	能效比	并行能力	适应能力	成本效益
CPU	★★★★★	★★	★	★★★★★	★★★
GPU	★★★★	★★★★	★★★★★	★★★★	★★★
NPU	★★	★★★★★	★★★★★	★★★	★★★★★
ASIC	★	★★★★★	★★★★★	★	★★★★★
FPGA	★★	★★★★	★★★	★★	★★
DPU	★	★★★★	★	★★★	★★★



异构计算较单一芯片能显著提升能效比
为AI模型训练等对算力有海量需求的应用提供能力
基础，加速AI技术在行业落地。



联接领域技术持续突破，光电融合发展共同保障算力供给

光电融合发展逐步突破物理限制，释放芯片算力

光电融合带来架构、协议变化，加速多域算力协同

CPO（光电共封装）

- 光引擎+ASIC共封装，信号路径降低到微米级、插损显著减少。
- 每bit传输功耗降低到1pJ以下，整体功耗显著降低。
- 消除光模块接口损耗，降低1.6T接口门槛。

异质集成技术

- 通过硅中介层或高密度封装，实现了多类芯片间的高效互联，同时解决不同材料的热膨胀系数差异等物理挑战。
- 提供10+Tbps片间互联带宽能力

跨架构技术生态

- 以 CXL 和 UCle 为代表的开放标准，为异质集成和光互联提供了统一的通信协议和物理接口规范。避免产业“孤岛”问题。
- 企业无需绑定单一厂商的方案。



新型态网络设备

- 基于CPO芯片研发高密度、低时延、低功耗的光电融合交换机
- AI安全、应用保障催生携带DPU的智能交换机及网关产品
- OCS/OXC产品正在适应智算中心要求

扁平化拓扑结构

- 通过芯片内、芯片间电互联，结合机架内、机架间光互联，实现整体能效比最佳。
- 通过OCS/OXC实现数据中心组网拓扑灵活设置，满足不同规模智算业务需求。

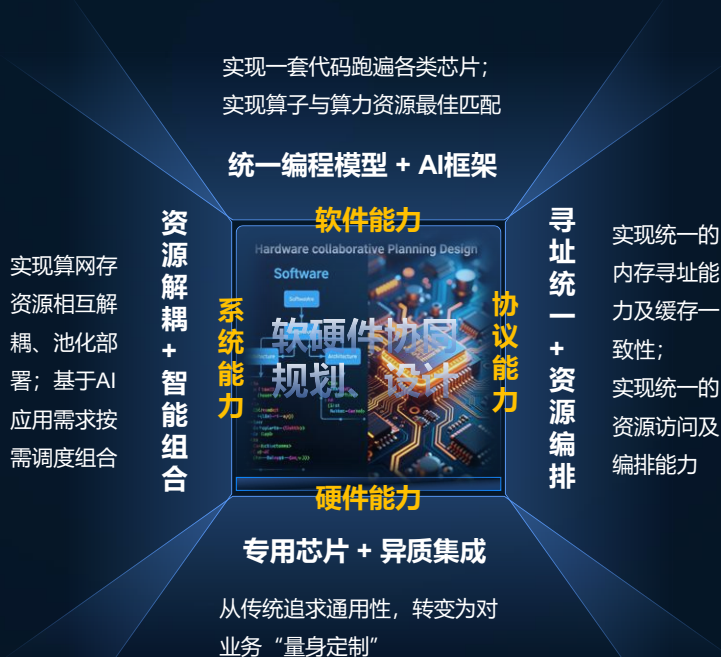
协议聚焦“好用”

- 光电融合驱动协议层从传统的“承载传输”向“主动参与”和“智能决策”转变，驱动如P4、eBPF等网络编程、应用观测性、安全过滤标准协议的发展。



以软硬件协同打造AI ICT基础设施硬基石，持续释放AI潜力

AI ICT基础设施在持续发展，为避免再受瓶颈影响，需从软硬件协同角度综合规划，打造出可持续发展的硬基石。



大模型推理过程中PD分离体现出软硬件协同的业务价值

当用户输入问题给大模型后，经过预填充+解码生成反馈信息

1. Prefill预填充：大模型调集计算资源理解输入token；属计算密集型业务
2. Decode解码：持续生成词元并频繁读写KV生成对话；内存带宽密集型业务

主要挑战：大模型应对过程中，表现出资源利用分配不均、整体利用率不高等情况，导致用户收到首个词元时间长、后续每个字生成时间长等体验问题。

算力资源调度平台

PD分离方案基

于资源利用原则采取串并结合处理方式显著提升用户体验。

Prefill预填充：
使用高算力产品实现

KV Cache
+
Store：缓存
KV cache

Decode解码：
+
使用大内存高带宽产品

高速、低时延网络

算联世界的“软支撑”——云网边端软件平台构建起产业生态

应用



运营商



数字政府



公共安全



数字金融



电力能源



智能制造



智慧交通



智慧医疗



智慧教育

生产管理、企业资源计划ERP、HR等AI化改造

AI智能体、数字人、机器视觉

空间智能、具身智能、机器人

平台

算力供给

算力调度

算力赋能

算力安全

算力绿色

算力服务

云软件栈升级：支持异构算力使能

云

智能化应用开发及运行平台

智算平台

数据平台

多元异构算力

零碳数据中心

反馈

协同

网



算力网络

汇聚

调度

边软件栈升级：支持异构算力协同

边

边缘应用平台

边缘推理平台

算法模型

边缘基础设施

感知

控制

端



5G



芯片：通用计算、异构计算、物联网等

操作系统：Linux、Windows等

数据库：实时交易、批量分析等

基础



异构算力需要软件支撑，新华三与产业界共同推动国产算力生态建设

- **人工智能软件栈并未形成统一标准，需要构建开放异构的生态体系：** CUDA是英伟达坚固的护城河，国产AI芯片的软件生态建设任重道远。AI芯片软件栈包含芯片驱动、运行时库、编译器、加速库、AI开发框架等，需要共同制定互联互通的标准，推动国产化算力生态发展。

模型
算法

ResNet

YOLO

Transformer

GPT

...

AI
框架

Pytorch

TensorFlow

PaddlePaddle

MindSpore

通用算子库（昇腾AOL、寒武纪CNL.....）

算力
调度

算力调度平台（Aliyun、H3C Cloud OS、傲飞.....）

操作系统（欧拉、麒麟、磐宁.....）

AI
算力

昇腾
CANN

昆仑芯
XTDK

天数
SDK

沐曦
MACA

海光
DTK

寒武纪
Neu
ware

CPU（海光 / 飞腾 / 鲲鹏）

训练推理参数交换协议（NCCL、HCCL、UCCL.....）

网络

芯片互联网络

节点互联网络

集群互联网络

站点互联网络

开放兼容AI框架： AI开发框架北向采用统一的API接口，保证AI算法在不同的AI开发框架上可顺利迁移

统一算子需求： 保持语义和接口要一致（如卷积、自注意力），保证AI开发框架和AI计算单元进行迁移时算子对接无歧义，新华三支持智源开源通用算子框架

统一精度要求： 各GPU要支持用于AI的相同的算力精度，如都支持FP16、BF16、FP32等，保证AI框架与各种GPU驱动顺利对接

统一算力单位： 在同一精度上衡量算力大小，如 XX TFLOPs@FP32，保证异构模型和数据的合理切分

统一集群网络： GPU服务器互联采用相同的集群组网方式，如采用基于以太网的RoCE无损网络，保证服务器间的网络接入一致

统一通信原语： 在模型分布式训练时，服务器内部GPU互通及服务器间GPU互通，保证通信集合通信库消息封装格式等原语一致，保证互操作性



软件生态在新一轮大模型竞争中走向开放，降本创新加速AI应用

- 通过分层软件栈整合多样化硬件资源，从模型到框架的全栈生态能力逐步形成，生态走向开放化：

层级	开源项目	开源协议	开源时间	备注
模型层	通义千问 (Qwen)	Apache 2.0	2025年4月	发布并开源千问3
	文心一言 (ERNIE)	部分开源	2025年4月	部分模型/工具开源
	DeepSeek	MIT	2024年起陆续开源	25.1月R1发布, 25.8月V3.1发布
	Llama (Meta)	自定义许可	2023年	禁止某些商业用途
	GPT- OSS部分模型	多样	不同项目时间不同	OpenAI部分开源
DL框架层	PaddlePaddle (百度)	Apache 2.0	2016年	
	PyTorch (Meta)	BSD	2016年	
	TensorFlow (Google)	Apache 2.0	2015年	
分布式训练层	Colossal-AI	Apache 2.0	2022年	
	DeepSpeed (Microsoft)	Apache 2.0	2020年	
	FSDP (PyTorch原生)	BSD	随PyTorch更新	PyTorch内置功能
集群管理层	KubeDL (阿里)	Apache 2.0	2021年	基于Kubernetes
	Volcano	Apache 2.0	2019年	
	Ray (Anyscale)	Apache 2.0	2018年	
	Slurm	GPLv2	2001年	
编译与运行时层	昇腾 CANN (华为)	部分开源	2025年8月	
	OpenAI Triton	MIT	2021年	
	AMD ROCm	多样	2016年	
	OpenCL (开放计算语言)	-	2008年	开放标准

- 国内大模型低成本创新破局美国优势，大模型使用门槛不断降低，异构算力兼容加速国内AI应用：

- 2025.8.21 DeepSeek发布使用UE8M0 FP8 的V3.1，降低模型部署对显存要求，引起算力生态轮动效应；
- 2025.4 阿里发布Qwen3并全系开源，参数量仅为DeepSeek-R1的1/3，成本大幅下降；
- 2025.3 赤兔Chitu开源，满血版DS-R1 部署GPU使用量减少50%，而推理速度仍有3.15倍提速；
- 2025.2 百度创始人李彦宏表示：当我们谈论大语言模型时，大模型推理成本每年能降低90%以上；
- 2025.1 DeepSeek R1发布同同步开源，仅有1/x0训练成本，性能与GPT-4等顶级模型相当，引爆低成本AI革命；
- 2023.3 智源研究院以开源模式开发飞智FlagOpen，面向异构芯片支持多种框架的大模型全栈开源技术基座，联合包括新华三在内的等多家国内外机构共同参与。



国内基础设施生态体系标准研制完成，建设进入落地新阶段

数标委成立到相关标准成稿，数据要素基础设施标准体系正加速构建中，以强化数据安全管控与互联互通：

- 标准聚焦于可信数据空间、数据基础设施和算力融合三大核心领域；
- 全国一体化算力网**9项技术文件已全部发布**，全国一体化算力网标准体系建设基本完善；
- 数据基础设施建设公布7项技术文件，数据基础设施标准体系逐步完善。



数据基础设施及网络、算力设施总体架构图

——源自《国家数据基础设施建设指引》2024.12.31

标准范围	标准名称（时间）	发布单位
可信数据空间	《可信数据空间 数字合约技术要求（征求意见稿）》2025.8.11 《可信数据空间 使用控制技术要求（征求意见稿）》2025.8.11 《可信数据空间 技术能力评价规范（征求意见稿）》2025.8.11 《可信数据空间 技术架构（征求意见稿）》2025.4.30	数标委
数据基础设施	10项规范： 《数据基础设施 安全能力通用要求（征求意见稿）》2025.8.12 《数据基础设施 区域/行业功能节点技术要求（征求意见稿）》2025.8.11 《数据基础设施 接入管理（征求意见稿）》2025.8.11 《数据基础设施 安全能力通用要求（征求意见稿）》2025.8.11 《数据基础设施 参考架构（试行）》25.3.10 《数据基础设施 互联互通基本要求（试行）》25.3.10 《数据基础设施 用户身份管理和接入规范（试行）》25.3.10 《数据基础设施 标识管理规范（试行）》25.3.10 《数据基础设施 接入连接器技术要求（试行）》25.3.10 《数据基础设施 数据目录描述规范（试行）》25.3.10	国家数据局、数标委
算力网融合	9项标准征求意见稿： 《全国一体化算力网 安全保护要求》2025.8.1 《全国一体化算力网 智算中心 算力池化技术要求》2025.8.1 《全国一体化算力网 算力并网技术要求》2025.6.8 《全国一体化算力网 算力资源管理与调度技术要求》2025.6.8 《全国一体化算力网 算力多量纲计费技术要求》2025.6.8 《全国一体化算力网 算力算效衡量技术要求》2025.6.8 《全国一体化算力网 算力运营服务与撮合交易技术要求》2025.6.8 《全国一体化算力网 算力监测接口要求》2025.6.8 《全国一体化算力网 算力中心能力评估要求》2025.6.8	数标委



AIGC走进核心生产系统，支撑各行业全面数字化改革，助力产业降本增收

产品技术 提升产品设计效率



- 技术战略
- 研发平台
- 解决方案
- 产品质量
- 产品线运营



代码编程

市场 提升产品设计效率



- 销售管理
- 客户管理
- 生态管理
- 国际市场



工作报告

供应链 提升及时交付率



- 采购管理
- 计划与订单
- 生产制造
- 仓储物流



订单与仓储的
协同管理

服务 提升服务质量满意度



- 服务产品
- 服务销售
- 服务交付
- 服务支持



产品介绍

财经 提升资金运营效率



- 财务核算
- 财务分析
- 预算管理
- 资金管理



合同分析

支撑体系 提升企业运营效率



- 品牌营销
- 人力资源
- 研学中心
- 合规与法务
- 行政运营

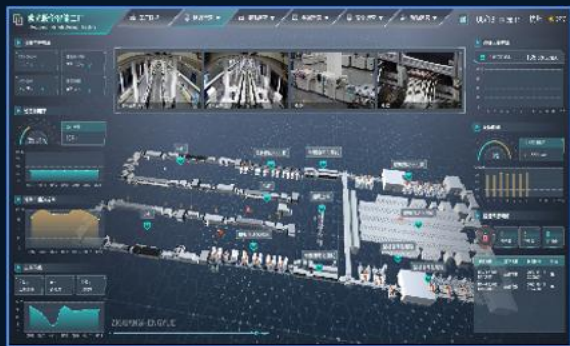


简历评估

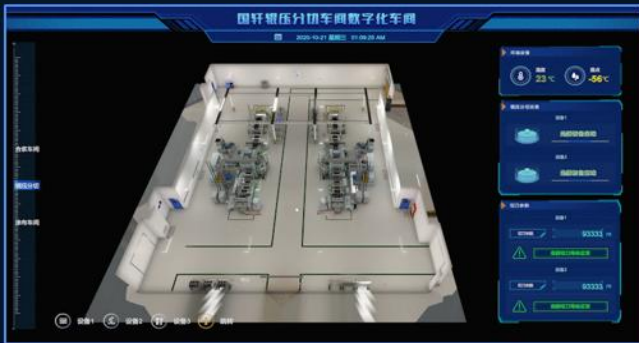


智能制造：AI赋能智能制造场景

基于智能工厂建设原则和评价方法，明确智能工厂建设愿景，编制智能工厂架构蓝图，科学地设计应用架构、数据架构以及技术架构，在生产和运营效率、产品质量、成本、交付、安全等方面全面发挥价值，提升企业经营能力



智慧供应链



数字孪生车间



AR数字化巡检



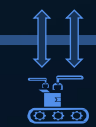
智能
工厂



生产管理



生产管理



生产管理



质量管理



物流仓储



能源管理



设备管理

智能工厂 (虚拟)
物理工厂 (现实)

工厂(传统制造业)

融合



打造5G+全链接工厂、数字化车间和智能化工厂,实现生产效率升级、生产运营的透明化,在面向柔性生产、智能制造领域,可持续推进5G+工业互联网智能工厂复制的典型案例



业界领先的数字化工厂样板点

业务指标	成本指标	技术指标
制造综合效率提高20%	单位能源消耗降低20%	自动化率 整体自动化率约为65%左右,其中日产线达75%,CT产线约50%
产品品质提升5%	人员减少35%	数字化水平 工业设备 工人操作范式 关键IoT设备 重点监控设备
产品生产时间缩短40%	管理成本削减30%	智能化程度 设备智能化 生产智能化 业务智能化 决策智能化

- 总投资10亿元,年产值超150亿元;
- 数字化工厂样板点: 工业互联网、5G工厂、智能工厂、绿色工厂;
- 新华三工业互联网平台是工厂 + 园区的统一底座
- 为外部供应商和物流公司提供在线服务



数字政府：向城市大脑智能中枢转型

城市大脑智能中枢打造的城市大脑GPT和“数智公务员”，通过数据治理、模型训练、赋能场景三个步骤，全面提升提升城市大模型的整体效能，城市治理和产业智能化的水平得以加强，为市民提供智能化的服务，提升了市民生活的质量和便利性。

变原来的“点点屏办理”为现在的“说说话办理”

杭州城市大脑智能中枢

数智公务员

城市大脑GPT

智能中枢（大模型生产体系）

模型应用广场

模型应用开发

模型运营

模型训练中心

模型评测中心

模型数据中心

算力调度中心

GPU算力资源池

政务信创云平台

变原来的“数据跑路”为现在的“AI服务”



治理侧：为城市管理者
提供智能治理工具

服务侧：为群众企业办事
提供智能化统一入口

模型应用广场：赋能各业务、各行业
快速构建人工智能应用



基于城市智能中枢打造的**数智公务员**，聚焦“问、办、培”工作核心，由原先的“协同办公、无纸化办公”实现**智能问效、高效办公、一站培训**

“问”

工作内容进行**问数**
工作进展进行**问效**
工作政策依据**问策**

“办”

办公文件智能生成
办事智能流转对接
办会纪要自动总结

“培”

培训课件智能创建
考试题目自动完成
评价体系一站整合

智库问答

智慧督办

公文写作

会议助手

智慧培训



智能查询数据秒达

工作进展、政策依据、数据
查询的秒级响应

高效办事精准督办

“一键办事”，智能化提取
任务并导入督办流程

智能培训自动答疑

实现培训课件自动生成与问
题智能回复，提升学习效率
与培训效果



智慧医疗：AI助力医疗行业高质量发展

以完整的基础设施和数字平台能力，支撑临床、服务、管理、科研多业务智慧场景建设，持续推动智慧医疗从单点数字化向全场景智能化跃迁





百业灵犀通过知识库RAG和思维链技术，打造脑血管病诊疗智能体
为脑血管病提供高效、专业的诊疗意见，诊疗过程可追溯，诊疗建议有据可依

- **场景聚焦：**聚焦脑血管疾病临床辅助诊疗与研究，AI医生（辅助诊疗）、AI医师助手（病历书写）及AI研究助理（科研辅助）
- **关键路径：**提示词工程+COT（思维链）+灵犀使能平台智能中枢+知识库RAG+百业灵犀大模型+完整的大模型工程实践落地

AI医生



遵循脑血管病临床管理指南
结合思维链技术推理给出临床诊断及治疗推荐

AI医师助手



借助语音识别技术、结合语义理解，通过识别关键信息自动生成病历

AI研究助理



结合图文多模态和OCR技术，识别并提取病历信息快速生成CRF表单





政府

- 利用AI自动识别处理政府流程中的重复和繁琐任务，在经济运行、应急救援、基层治理、生态环保等方面，实现智能监测、智能分析、智能预警、智能处置、智能服务，全面提升治理体系和治理能力现代；

教育

- 通过搭建教学云平台和“智慧教室”，我们将为每个老师定制一个AI助手，帮助老师自动收集和整理相关的资源辅助教学，为每个学生建立一个AI伴侣，让“因材施教”的千年理想、个性化教学成为现实；

文化传媒

- 通过AI数字分身生成数字内容、创作数字艺术、管理文化资料、修复文化遗产，打造数字景区、创建智慧文旅。在视频方面，推进云上协同制作，随时编、随地编、智能编，以“智”变带动“质”变；

医疗健康

- 将数智之力贯穿到医院业务的方方面面，医院管理AI提升诊疗效率；提供患者服务，服务个性诊疗，探索医药研发AI，主动发现药物靶点，并在康养领域，以智能化方案直面中国老龄化的挑战；

企业

- 帮助企业建设“数字大脑”“数字员工”，赋能管理，监管到每一笔订单，提升运行效能，利用智能化辅助设计、调度生产、赋能服务、优化供应，全面提升企业效率；

制造

- 通过工业互联网赋能“智能工厂”，联接设备、数据、信息、生态、服务、安全，以六个联接推进数实融合，以智能辅助提升制造效能；

能源

- 服务智能电网与智能发电，推出我司数字能源方案，推动AI智能调度与AI机器人，配合替代人，从事危险工作、重复劳动；在矿山、提供覆盖“采、掘、机、运、通”全流程智能化解决方案；

交通

- 畅通数字交通，从智能出行到自动驾驶，全力推进车路协同，实现城市交通全域数据推演、全景智能分析，探索围绕自动驾驶的全线安全方案；



技术创新、生态构建、场景验证三者深度协同，是新华三AI实践的方法论

技术创新如同算联发展的智慧引擎，驱动底层架构的持续突破，为数字世界筑牢坚实基础。

多元化的应用场景，正是检验价值、引领方向的试金石，算联能力在千行百业中落地生根，持续赋能产业数字化转型，孕育发展新机遇



开放协同的产业生态，则是我们汇聚力量、共创未来的基石与纽带，技术、资本与人才的多元融合，构建起共生共赢的创新共同体

THANKS

感谢观看