准印证号：（粤B）L011030048

中兴通讯技术 简讯

ZTE TECHNOLOGIES | 第28卷 第10期 · 2024年10月

视点

专题：星云行业大模型

内部资料 免费交流

扫码体验移动阅读

1996年创办 总第433期
2024年10月 第28卷 第10期

中兴通讯技术（简讯）
ZHONG XING TONG XUN JI SHU（JIAN XUN）
中兴通讯股份有限公司主管

主办单位：中兴通讯技术杂志社
编辑：《中兴通讯技术（简讯）》编辑部
发行范围：国内业务相关单位
印数：5000本
出版频次：按月
地址：深圳市科技南路55号
邮编：518057
发行部电话：0551-65533356
网址：http://www.zte.com.cn

设计：深圳市奥尔美广告有限公司
印刷：深圳市旺盈彩盒纸品有限公司
印刷日期：2024年10月30日

陆平
中兴通讯副总裁、产业数字化方案部总经理

星云行业大模型引领产业数字化变革

当前，以人工智能为核心的新一轮科技革命正加速演进，数据驱动、AI赋能带来了行业生态的全面升级。新型算力基础设施、行业大模型正成为产业转型发展的新引擎，为各行各业新质生产力的构建注入新动能。

人工智能正在从单模态智能迈向多模态融合，并已逐渐具备复杂推理与问题解决能力，技术跃迁和产业需求共同推动大模型行业应用广度、深度加速拓展。依托于高质量行业数据和强大的计算资源，AI智能体之间可实现有效协作，共同执行复杂任务，从而在工业智能制造、城市治理、智慧交通等应用场景中发挥巨大作用。作为人工智能领域的实践者和先行者，针对当前行业大模型百家争鸣的现状，中兴通讯提出“开放解耦、以网强算、训推并举”的核心主张，并于今年初发布了更加智能的数字星云3.0，融入星云行业大模型，帮助客户和合作伙伴多快好省地使用AI技术、增收降本提效。

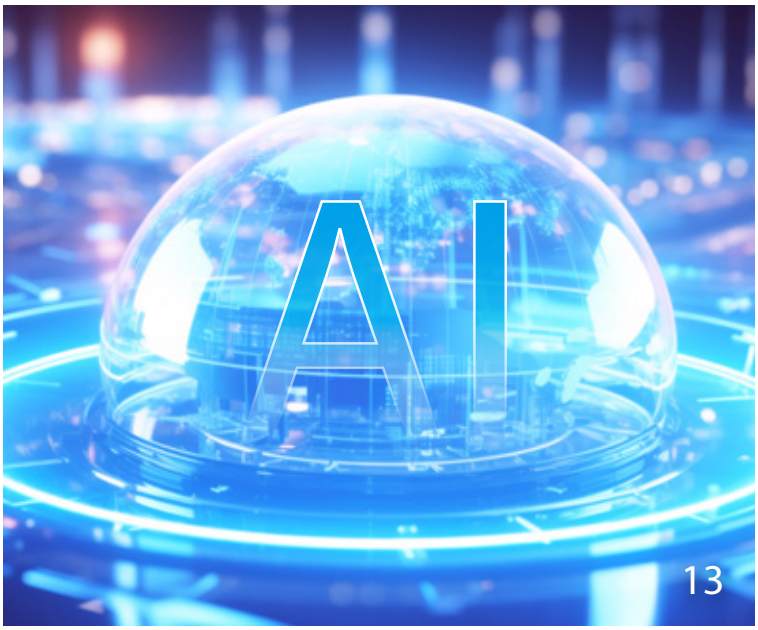
经过多年实践，中兴通讯已经形成了“工业现场网+数字星云”双轮驱动赋能产业数字化发展的范式。中兴数字星云是一个开放的数智赋能平台，数字星云3.0内置星云大模型，并进一步升级了大模型训练、推理、应用开发全流程工具，从而更好地实现大模型能力的价值变现。星云大模型结合了中兴通讯多年积累的行业经验和海量高质量数据资源进行优化训练，不仅具备强大的通用AI能力，还针对不同行业的特点进行了深度优化，从而更好地服务于特定业务场景。

在实际应用中，数字星云3.0与星云大模型已经取得了诸多成果。在中兴通讯内部，通过研发大模型实现编码效率提升30%、整体研发提效10%；通过工业大模型实现工艺文件生成提速10倍、质检人力成本降低70%，在南京滨江成功建成了国内首个五星5G工厂。同时，中兴通讯也积极探索行业大模型在千行百业的应用，已在城市生命线、水利、应急、油气、交通、电力、冶金钢铁等多个领域取得成果，在本期专题中将会向大家详细介绍。

展望未来，中兴通讯将继续秉持开放合作的理念，以硬实力筑基，以软实力启智，灵活实现与生态伙伴、行业客户的资源共享和优势互补，携手赋能千行百业的数智化转型，为数字经济高质量发展注入源源不断的动能，共筑AI时代的美好未来。

目次

中兴通讯技术（简讯）2024年第10期

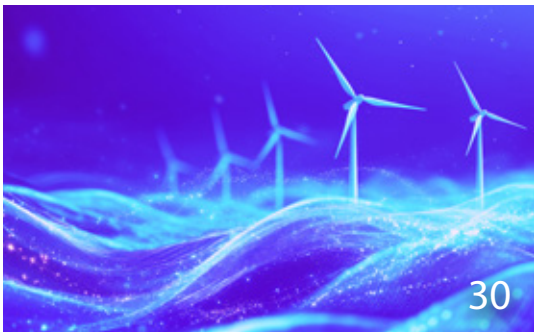


星云行业大模型，构筑产业智能化转型新引擎

当前，中国大模型进入发展加速期，自然语言处理、视觉处理和多模态等各技术分支均快速发展。在产学研各方共同推动下，我国已建立起涵盖理论方法和软硬件技术的体系化研发能力，据不完全统计，国产大模型数量目前已超过200个，形成了紧跟世界前沿的大模型技术群。

视点

专题：星云行业大模型



成功故事

网媒融合沉浸文旅发展论坛圆满举行

9月26日，2024北京国际通信展5G+XR沉浸文旅发展论坛召开。中兴通讯副总裁、产业数字化方案部总经理陆平出席论坛并致辞。论坛上，中兴通讯、中国电子云、域上和美集团、四川电信联合发布全球首个5G-A VR大空间沉浸剧场项目——幾米绘本元宇宙戏剧。陆平指出，5G+XR网媒融合加持下的创新应用对数字文旅的发展起到越来越关键的作用，幾米绘本元宇宙戏剧项目是各方携手研究和精心打磨的成果，融合了中兴通讯在5G-A和XR领域的最新技术。

10月31日，幾米绘本元宇宙戏剧项目在成都域上和美先锋剧场首演，剧场可同时容纳超过50人进行高沉浸体验，全天可接待超过800人次。

中兴通讯圆满交付全国首套大型无人直升机救援平台

中兴通讯携手合作伙伴打造大型无人直升机应急救援平台。无人直升机平台采用经典旋翼构型、专业航空发动机，源自国内领先的线束设计和工艺、顶置高通量卫通设备、易拆卸的载荷舱设计，通过10轮严苛测试。从2022年至今，中兴通讯大型无人直升机应急通信救援方案已完成8次实战救援和7次实战比测，被评价为“一个完整的应急救援通信保障体系，能很好地适应实战场景”。在今年全国应急航空能力救援提升项目中，中兴通讯斩获50%的市场份额。

8月21日，中兴通讯在云南交付全国首套大型无人直升机救援平台，顺利通过云南省应急管理厅的飞行及载荷性能考核。9月3日，中兴通讯完成海南无人直升机救援平台的交付。9月6日，台风“摩羯”肆虐海南，在应急管理部、省应急厅的部署下，中兴通讯大型无人直升机应急通信系统投入备勤，9日进入救援，先后转战白沙、甲子及文昌多地，全力以赴做好灾区一线的通信恢复和救援保障。同日，中兴通讯云南大型无人直升机应急通信系统抵达文山，第一时间展开救援，得到当地救援团队的高度认可。

中兴通讯坚持深耕信息通信技术的创新研发，构建空天地一体化全方位的应急保障体系，全力践行应急救援的使命必达。

中兴星云大模型赋能行业新质生产力

近日，智胜未来大会暨大模型应用发展论坛在北京举行，中兴通讯产业数字化方案部副总经理张滨出席并发表主题演讲。张滨介绍，中兴数字星云3.0优化了数据流程管理机制，增强计算资源调度能力，并引入先进的安全防护措施，帮助行业伙伴更低成本、更便捷地开展智能化应用。目前，星云行业大模型已在多个领域实现生产管理的提质增效，助力打造新质生产力。

中兴通讯南京滨江5G工厂接连斩获大奖

9月25日，“ICT中国（2024）案例”年度评选荣誉正式颁发，中兴通讯南京滨江5G工厂（以下简称“滨江工厂”）安全保障项目凭借卓越的技术实力和显著的行业标杆效应，荣获“卓越案例一等奖”。这也是滨江工厂继首家通过信通院“五星5G工厂”认证、荣获“2024 IDC中国20大杰出安全项目”之后，接连斩获业界重磅荣誉。

“5G+AI助力建设智慧安全电厂”项目荣获ICT中国（2024）卓越案例一等奖

9月25日，“ICT中国（2024）案例”年度评选荣誉在PTEXPO 2024 ICT中国·高层论坛主论坛上正式颁发。中兴通讯、大连移动联合打造的“5G+AI助力建设智慧安全电厂”项目，基于5G+数字电厂解决方案建设创新智慧应用，全面提升了电厂日常运营运维、应急指挥、检修调度和设备监测的能力，荣获卓越案例一等奖。

中兴通讯：深化“连接+算力”助力AI向实，前三季度营收超900亿元

10月21日，中兴通讯发布2024年第三季度报告。报告显示，2024年1—9月，公司实现营业收入900.4亿元，同比增长0.7%；归母净利润79.1亿元，同比增长0.8%；扣非归母净利润69.0亿元；经营性现金流净额80.5亿元。

前三季度，公司整体经营保持稳健，国内运营商网络业务受投资环境影响整体承压，国际市场持续突破大国大T，保持双位数增长。同时，公司消费者和政企业务均实现快速增长。

公司前三季度研发费用186.4亿元，占营业收入的20.7%，推动5G-A、全光网络、全栈智算等技术创新，加速与千行百业的深度融合及应用落地，为加快发展新质生产力提供有力支撑。

连接领域，中兴通讯继续在无线和有线市场的关键产品上保持领先。根据电信咨询机构最新数据，公司5G基站、5G核心网累计发货量全球第二，固网产品市场份额全球第二，RAN产品、5G核心网、5G光传输获行业领导者评级。公司光接入产品得到国际认可，在国际Network X峰会荣获三项奖项。

围绕5G-A的商用部署和落地，中兴通讯与运营商进行了全场景技术创新与合作，共同打造“空天地一体化”5G-A全域立体网，激发地面、低空和高空赛道的新业务活力，并助推卫星互联网新发展。其中，在低空经济领域联合合作伙伴在北京、南京、深圳等25个省市的80多个试点验证5G-A通感一体在物流配送和低空安全等多种场景的应用。

此外，公司推出800G OTN可拔插方案，实现2000km的传输距离，并基于50G PON技术发布新一代智能光接入产品，为用户提供万兆接入体验，助力行业进入万兆光网时代。

算力领域，中兴通讯将智算确定为公司的长期战略主航道，提供全栈全场景智算解决方案，打造开放普惠繁荣的智算生态。在智算基础设施和平台技术方面，推出“一机多芯”开放架构AI服务器，兼容适配主流GPU，并以网强算，基于自研芯片推出2×200G网卡；推进OLink高速总线互联标准，并自研大容量交换芯片，在超万卡集群核心技术、算力原生、智算中心长距互联、推理任务智能分发等前瞻技术方向上开展研究。公司注重基础设施集成交付和大模型训练、应用的工程化服务，推进多地多行业智算基础设施建设项目落地。

在大模型及应用方面，自研星云系列大模型，其中研发大模型有效带动研发效率提升；通信大模型在反诈、重保、新通话、体验保障等场景落地应用，加速自智网络进阶，积极拓展工业、汽车、钢铁、水利、城市生命线等领域模型及应用。

在应用生态方面，建立开放智算实验室，提供多厂家GPU互联、优化以及大模型兼容性测试认证，与业界合作伙伴一起积极推动国内开放智算基础设施及应用生态建设。

终端领域，秉承“AI for All”的理念，三季度，公司发布了全面升级的AI全能影像旗舰nubia Z60 Ultra领先版和AI+卫星手机nubia Z60S Pro，并推出普惠5G手机中兴远航40s。nubia品牌持续推进出海战略，Neo2、Music、Focus等特色产品进入阿根廷、埃塞俄比亚、德国等市场。

公司面向家庭用户打造高品质的千兆网络体验，FTTR产品持续创新，推出业界首款AI带屏系列产品RoomPON 6.0；云终端产品进一步强化市场竞争力，上半年在国内云终端市场出货量位居第一。

汽车电子业务与车厂合作不断加深，联合车厂发布自研车规级高性能中央计算单元SOC芯片“撼域”M1；与中国一汽签署多域融合芯片“红旗1号”战略合作协议；携手上汽集团发布百万量级车云通讯大单品，实现行业首个国产车规级双核4G通讯模组的量产落地，和上汽集团一起加速国产智能网联出海。

中兴通讯积极推进数智技术在企业自身及产业中的应用。公司数据管理能力获国家数据管理能力成熟度评估DCMM 5级最高等级认证，达到国内领先水平。中兴通讯南京智能滨江5G工厂荣获国内首个五星5G工厂认证。

在行业赋能方面，公司依托5G+PON工业现场网和数字星云，结合星云大模型，全面助力工业、交通、应急、电力、文旅、教育、小微企业等众多行业的数智化建设。其中，智慧大应急体系斩获全国应急航空能力救援提升项目50%的市场份额，大型无人直升机救援平台在多个抢险救灾一线保障。基于5G+XR网媒融合赋能数字文旅建设，公司与合作伙伴共同打造全球首个5G-A VR大空间沉浸剧场项目。

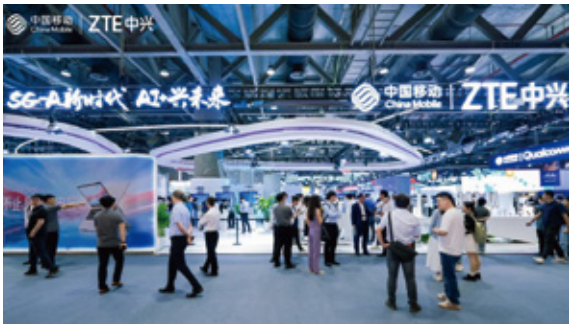
中兴通讯亮相2024中国移动全球合作伙伴大会，以创新成果赋能高质量发展

10月11日，2024中国移动全球合作伙伴大会在广州举行。作为中国移动的战略合作伙伴，中兴通讯以“5G-A新时代，AI+兴未来”为主题参展，展示双方在全栈智算、5G-A、创新、终端等关键领域的合作成果，共同推动新一代信息技术深入千行百业，以新质生产力赋能高质量发展。

5G-A商用部署加速推进，中兴通讯与中国移动进行了全场景技术创新与合作，在推动各行业数智化发展的同时，进一步深化商业合作。连接展区重点展示了面向“空天地一体化”的5G-A全域立体网，地面赛道着眼于激发业务新活力，率先实现超低时延XR游戏，赋能央视春晚全程超高清直播，实现业界首个通感算一体化新型车联网方案的试点应用；低空赛道着眼于赋能新质生产力，业内首发大张角AAU，扩大垂直张角至65°，实现感知能力超1.5km，并已在全国多地实现通感一体技术商用部署；高空赛道延伸至太空，致力于用5G ATG技术，满足大众在万米高空的上网娱乐需求，同时，联合进行星载基站研发，攻克多个核心器件难关，助推卫星互联网新发展。

此外，连接展区还展示了双方在算力连接方面的合作成果，国内首发的1.6T OTN样机，助力中国移动完成多芯光纤、单模光纤多波段、空芯光纤等多场景Tbit传输验证，为算力互联提供坚实的高带宽保障；携手发布的全球首台算力路由器，已助力中试网络江苏节点开通，共创数字产业新生态。此外，中兴通讯作为业界唯一在芯片内置EDN功能的厂商，成功完成全球首个超2000km IP 400GE广域确定性现网试点。

智算展区展示中兴通讯联合中国移动在智算方案创新与应用方面的成果，深入推进“AI+”行动计划，打造开放共赢智算生态。



中兴通讯在京举办低空经济产业发展论坛

9月25日，由中兴通讯主办的“创智低空，聚力腾飞”低空经济产业发展论坛在北京国家会议中心隆重举行。中国工程院院士张宏科、中兴通讯高级副总裁及首席战略官王翔、中国电信首席科学家毕奇、中国移动首席专家刘光毅、中国联通研究院泛终端中心低空应用高级研究员席绪亚、中兴通讯RAN产品副总经理刘爽、北京交通大学教授、博导官科、瑞合玄武创始人陈思聪等嘉宾出席论坛并发表演讲。

中兴通讯推出业界首款AI带屏FTTR系列产品RoomPON 6.0

9月26日，在2024年国际信息通信展期间，中兴通讯成功举办了“光耀未来，智算生活”全光家庭发展高峰论坛。中兴通讯副总裁董伟杰发表了题为“光耀未来，引领智慧家庭新纪元”的演讲，并隆重推出了业界首款AI带屏FTTR系列产品RoomPON 6.0。

中兴通讯发布新一代智能光接入旗舰产品，引领万兆光网新时代

在2024年国际信息通信展期间，中兴通讯隆重推出面向万兆时代的新一代智能光接入旗舰产品C600H。作为中兴通讯面向未来十年光接入领域打造的旗舰产品，C600H具有大带宽、内生智算、刚性管道和高可靠性等卓越特性，将全方位推动各行业迈入万兆光网时代，为万兆城市的数智化建设提供强有力的支撑。

IDC报告发布：中兴通讯云终端登顶中国桌面云终端市场第一

全球领先的IT市场研究和咨询公司IDC发布《2024年上半年中国云终端市场跟踪报告》，在公有云部署浪潮趋势下，中兴通讯凭借与运营商的紧密合作登顶云终端市场冠军，在云终端总体市场出货量、VDI解决方案云终端市场出货量均位居第一，对迅速崛起的消费类云终端市场贡献巨大。

报告显示，2024上半年中国云终端市场总体出货量达到166.3万台，同比增长22.4%，销售额29亿元人民币，同比增长24.9%，均超预期。公有云部署拉动消费类云终端增长以及新定义下云终端市场硬件设备拓展到平板电脑和手机，导致了这一市场的积极表现。IDC预计，至2028年中国云终端市场规模有望超过615万台，五年复合增长率将达到15.8%。此外，2024上半年应用VDI解决方案的云终端市场出货量同比增长29.6%，好于市场平均水平，其中公有云架构部分增长近68%，私有云增长12%，同时消费类云终端出货量猛增94%，传统商用云终端部分同比增长21%，中兴通讯在家用市场领域遥遥领先。

中兴通讯在这场云终端市场的激烈角逐中脱颖而出，关键在于其深厚的技术积累、持续的创新能力和对市场的精准洞察。公司不断推出贴合市场需求的创新产品，拥有“逍遥”“驭风”“玲珑”“扶摇”等系列云终端产品。

通感算融合创新，中兴通讯助力车路云一体化新基建

近日，“2024车路云50人年度发展论坛”举办，中兴通讯副总裁张继军参与重庆市副市长江敦涛、中国工程院院士李克强等专家出席的高端会晤，中兴通讯交通业务部总经理兰波参加“三商融合”合作组织发起仪式，助力车路云规模化商用。

中兴通讯产业数字化方案部副总经理姜永湖出席分论坛，分享中兴通讯投身车路云一体化建设的情况：参与百余个5G及车联网行业标准制定；参与苏州、无锡、深圳等多地示范区建设；围绕“聪明的车”“智慧的路”“可靠的网”“协同的云”创新，涵盖业界首款支持5G及C-V2X的双频自研芯片和模组，5G+C-V2X融合组网，基于数字星云的云控平台等。

中国移动与中兴通讯联合发布多智能体协同创新成果

2024国际信息通信展期间，中国移动联合中兴通讯发布了多智能体协同方案创新成果。该成果首次将大模型引入无线网络端到端运维并在多地取得良好成效。

本次发布的“无线多智能体协同”方案，基于最新的通信专业领域大模型，通过多智能体协同，调用网络内生智能原子能力，赋能无线运维工作台，探索智能运维新模式。

中兴通讯、东风汽车与湖北移动联合推出AiCube汽车设计一体机

2024中国移动合作伙伴大会上，中兴通讯展示了基于AiCube智算一体机的汽车设计应用，该产品由东风汽车、湖北移动及中兴通讯联合推出。中兴通讯基于东风汽车提供的汽车图片语料，进行深度学习训练，打造汽车AI设计应用，支持线稿、风格、空间关系等多种选项，用户可进行个性化定制与优化，输入简单的关键词，即可生成高质量的汽车设计图像。

中兴通讯亮相2024世界智能网联汽车大会

10月17日，2024世界智能网联汽车大会召开，中兴通讯副总裁、汽车电子总经理古永承出席大会，并分享了中兴通讯对中国智能汽车操作系统发展的思考及实践。面向智能汽车新EE架构演进和产业发展需要，中兴通讯汽车电子的定位是数字汽车基础能力提供者，自主创新、国产高性能的合作伙伴，以芯片+OS基础能力为核心、国产化为合作点助力主机厂转型升级。

朱霖潮
浙江大学计算机科学与技术学院
百人计划研究员

杨易
浙江大学计算机科学与技术学院副院长、
求是讲席教授

大模型：赋能产业变革的数智化基石

在人工智能领域，大语言模型（large language models，LLM）的突破性进展正在掀起一场技术革命。从ChatGPT的横空出世到各大科技巨头纷纷入局，大模型技术正逐渐重塑各行各业的格局。本文将深入探讨大模型技术的行业应用、变革趋势以及未来展望，为行业决策者提供前瞻性的洞察。

大模型技术的发展

近年来，以GPT（generative pre-trained transformer）为代表的大模型技术取得了突飞猛进的发展。这些模型通过海量数据训练，习得了复杂的语言理解和生成能力，从最初的GPT-3到GPT-4，模型规模和能力不断提升，应用范围也从单一的自然语言处理扩展到多模态交互、逻辑推理等广泛领域。

大模型的核心优势在于其强大的迁移能力和少样本学习能力。通过预训练和微调，大模型可以快速适应各种下游任务，显著降低了人工智能应用的开发门槛，为各行业的智能化转型提供了强有力的技术支撑。

大模型技术在垂直行业的应用

人工智能大模型技术正在各个垂直行业中展现出一定的应用潜力。

在制造业领域，大模型技术正在成为推动智能制造和工业升级的关键驱动力。在智能设计与仿真方面，大模型能够辅助工程师进行产品设计优化，通过虚拟仿真大幅减少实物测试的需求，从而加速产品开发周期；在设备维护领域，大模型通过分析设备运行数据和历史维修记录，实现设备故障的精确预测，优化维护计划，能够提高生产线的运行效率；在供应链管理方面，大模型凭借其数据整合和分析能力，优化库存管理和物流配送流程，有助于提升供应链的弹性和响应速度。

在教育领域，大模型应用正推动个性化学习和终身教育的发展。基于大模型的智能助教系统

能够根据学生的学习进度和个人风格，提供高度个性化的学习内容和即时反馈，大幅提升学习效果；其次，在评估方面，大模型可以生成动态的测试题目，根据学生的回答实时调整难度，从而更准确地评估学习成果，为因材施教提供依据；此外，在教育内容生成方面，大模型正协助教育工作者创建多样化的教学材料，包括练习题和教学视频，提高教学资源的丰富性和质量。

在法律服务领域，大模型技术有助于提升法律咨询、合规管理等服务的效率和质量。首先，基于大模型的智能系统可以用于快速分析海量法律文件和判例，为律师提供相关案例和法规参考，还可以协助起草法律文件，如合同和诉讼文书，减少人为错误，提高工作效率。其次，在合规管理领域，大模型可以持续监测和分析最新的法律法规变化，及时提醒企业进行合规调整，降低合规风险。此外，大模型在智能合同审查、法律研究辅助以及预测性分析等方面，也有许多应用。

大模型驱动的行业变革趋势

大模型技术的发展促使各行业的智能化转型加速，并重新定义数据价值，推动产业链重构与协作模式创新。

智能化转型加速

大模型技术的发展正在加速各行业的智能化转型进程。传统基于规则的专家系统正逐步被具有自适应、自学习能力的人工智能系统所取代。在决策支持领域，大模型能够综合分析海量的多模态数据，提供更全面、更具洞察力的建议。在客户服务方面，基于大模型的智能助手能够理解复杂的自然语言查询，提供个性化的响应，显著提升服务质量和效率。在产品开发过程中，大模型可以快速生成和评估大量设计方案，加速产品迭代周期，不仅大幅提高业务流程的效率，还增强企业应对瞬息万变的市场环境的灵活性。

重新定义数据价值

随着数据成为训练和优化大模型的关键“燃料”，其价值大幅提升。首先，跨领域、跨格式的数据融合与集成变为可能，使企业能从行业数据中获取更深刻的洞察；其次，高质量、结构化的数据日益成为稀缺资源，推动了数据治理和标准化的快速发展；此外，隐私保护技术的发展，如联邦学习等，使得在保护隐私的前提下实现数据价值最大化成为可能。这些趋势相互交织，共同重塑了数据的价值定位和应用模式，推动企业重新审视其数据战略。

产业链重构与协作模式创新

大模型技术正在推动产业链的深度重构和协作模式的创新。一方面，传统产业链中的某些中间环节正被智能系统所替代，导致产业链呈现扁平化趋势，提高了整体效率并降低了成本。另一方面，围绕大模型技术的应用与优化，一个全新的生态系统正在形成，包括数据提供商、算力服务商、行业解决方案提供商等新兴角色。与此同时，产学研合作日益紧密，加速了从基础研究到商业应用的转化。然而，这一过程也伴随着挑战，如技术适应、人才培养和监管调整等问题，需要产业各方共同努力来应对，以实现可持续的创新发展。

大模型技术发展面临的挑战

大模型技术面临的挑战涉及数据、资源、可解释性等多个方面，亟需技术创新与跨界合作，通过多方努力，有望克服这些障碍，发挥大模型技术的潜力。

数据质量与隐私保护

大模型的训练需要海量高质量数据，然而获取这样的数据集既困难又昂贵。首先，高质量数据的定义本身就是一个挑战，它不仅要求数据的准确性和完整性，还需要考虑数据的多样性，以

确保模型不会产生偏见或歧视。其次，数据收集和使用过程中的隐私保护问题日益突出。随着各国数据保护法规的实施，企业在数据使用方面面临着更严格的合规要求。此外，公众对数据隐私的认识也在不断提高，对数据收集和使用的透明度提出了更高的要求。如何在保护个人隐私和商业机密的同时，确保数据的可用性和多样性，是一个亟待解决的难题。

计算资源与能耗问题

训练和部署大模型需要庞大的计算资源，这不仅带来了高昂的成本，也引发了对能源消耗和环境影响的担忧。据估计，训练一个大语言模型可能消耗数百吨的二氧化碳，相当于数百次跨大陆飞行的排放量。这种巨大的能源消耗不仅增加了企业的运营成本，也与全球减少碳排放、应对气候变化的目标相悖。如何提高模型的计算效率，减少能源消耗，成为技术发展的重要方向。

模型可解释性与可控性

大模型的决策过程往往是不透明的，这种“黑盒”特性在某些高风险应用场景中可能引发严重问题。当模型做出错误或有争议的决定时，难以追溯原因并进行修正。这不仅影响了模型的可信度，也可能导致法律和道德风险。提高模型的可解释性和可控性，确保其决策过程的透明度和可追溯性，是未来研究的重点。

展望

大模型技术作为人工智能领域的突破性进展，正在成为推动产业变革的核心驱动力。从制造业的智能制造到教育的个性化学习，从法律服务的智能化到金融领域的精准风控，大模型技术以其强大的语言理解和生成能力，不断重塑着各行业的运作模式与价值创造体系。它不仅加速了智能化转型的步伐，重新定义了数据的价值，还促进了产业链的重构与协作模式的创新，为社会发展注入了新的活力。

然而，大模型技术的持续发展也面临众多挑战。数据质量与隐私保护、计算资源与能耗问题、模型可解释性与可控性，这些难题需要业界共同努力，探索解决之道。

- 强化数据治理与隐私保护：建立严格的数据收集、处理和使用标准，确保数据的合法合规性。同时，加强隐私保护技术的研发与应用，如差分隐私、联邦学习等，以在保护个人隐私的前提下实现数据的有效利用。
- 优化计算资源与能效管理：推动硬件技术的创新，研发更高效、低功耗的芯片与计算平台。同时，优化算法设计，减少模型训练与推理过程中的计算量，降低能源消耗。此外，探索绿色计算模式，如利用可再生能源供电，减少对环境的影响。
- 提升模型可解释性与可控性：加强可解释性人工智能技术的研发，使模型的决策过程更加透明、可追溯。建立模型可解释性的评估标准和框架，为模型的应用提供指导。同时，设计人机协同的交互界面，允许人类专家参与模型的决策过程，实现人机共智。
- 促进跨界合作与生态构建：鼓励产学研用各方加强合作，共同推动大模型技术的研发与应用。建立开放共享的数据与算力平台，降低技术门槛与成本。同时，构建以大模型技术为核心的产业生态，促进上下游企业的协同发展与创新。

展望未来，大模型技术将继续在各个领域发挥重要作用，推动经济社会向更加智能、高效、可持续的方向发展。我们坚信，在技术创新、跨界合作与负责任的发展策略的共同推动下，大模型技术将克服一切挑战，为人类社会创造更加美好的未来。企业和决策者需要积极拥抱这一技术变革，制定前瞻性战略，在充分把握机遇的同时审慎应对挑战，共同塑造一个更加智能、可持续的未来。ZTE中兴

大模型推理优化关键技术

刘涛
中兴通讯资深算法专家

2022年底，OpenAI发布了跨时代的ChatGPT应用。它的成功使大模型成为AI发展的主旋律，在极短的时间内改变了AI产业的格局。随着GPT-4、Gemini、Sora、Claude3、Kimi等一系列大模型的陆续发布，大模型能力迅速提升，甚至更为强大的通用人工智能（artificial general intelligence，AGI）已初见端倪。

大模型以其强大的理解和生成能力正在深刻改变我们对人工智能的认知和应用，但其高昂的推理成本也阻碍了技术落地。因此，优化大模型的推理性能成为业界研究的热点。本文中，我们试图对ChatGPT发布以来大模型推理优化关键技术做出综述，厘清技术全貌及发展态势，以便读者做出更好的判断和预测。

大模型推理性能优化主要以提高吞吐量和降低时延为目的，关键技术可以划分为：内存管理、算子融合、模型压缩、并行推理、服务调度优化及新兴技术。

内存管理

KV Cache是大模型推理性能优化最常用的技术。该技术在不影响任何计算精度的前提下，通过空间换时间，大幅提升推理性能。Transformer解码器使用自回归产生输出，即每次推理只会预测输出一个token，执行多次后完成全部输出。前后两次的输入只相差一个token，这就存在大量重复计算。KV Cache技术将每个token可复用的K和Q向量结果保存下来复用，将计算复杂度从O（n^2）降低为O（n）。

Paged Attention技术将操作系统中的分页内存管理应用到KV Cache的管理中，节约了60%～80%的显存，从而支持更大的batch-size，将吞吐率提升了22倍。具体来讲，Paged Attention首先将每个序列的KV Cache分成若干块，每个块包含固定数量token的键和值，然后计算出当前软硬件环境下KV Cache可用的最大空间，并预先申请缓存空间。在推理过程中，通过维护一个逻辑块到物理块的映射表，使多个逻辑块对应一个物理块，并使用引用计数标记物理块被引用的次数，从而实现将地址不连续的物理块串联在一起统一管理。

算子融合

算子融合是深度学习模型推理的一种典型优

化技术，旨在通过减少计算过程中的访存次数和统一计算架构（CUDA）Kernel的启动耗时，达到提升模型推理性能的目的（见图1）。针对Transformer的结构特点，算子融合主要分为4类：归一化层和QKV横向融合，自注意力计算融合，残差连接、归一化层、全连接层和激活层融合，偏置加法和残差连接融合。

中兴通讯在vLLM上实现了针对多查询注意力结构的QKV通用矩阵乘法（GEMM）横向算子融合，以及多层感知机（MLP）中的全连接层（FC）+激活融合，性能明显提升，见表1和表2。上述算法的相关代码实现已合入vLLM社区。

模型压缩

模型压缩技术是指在不影响模型精度的情况

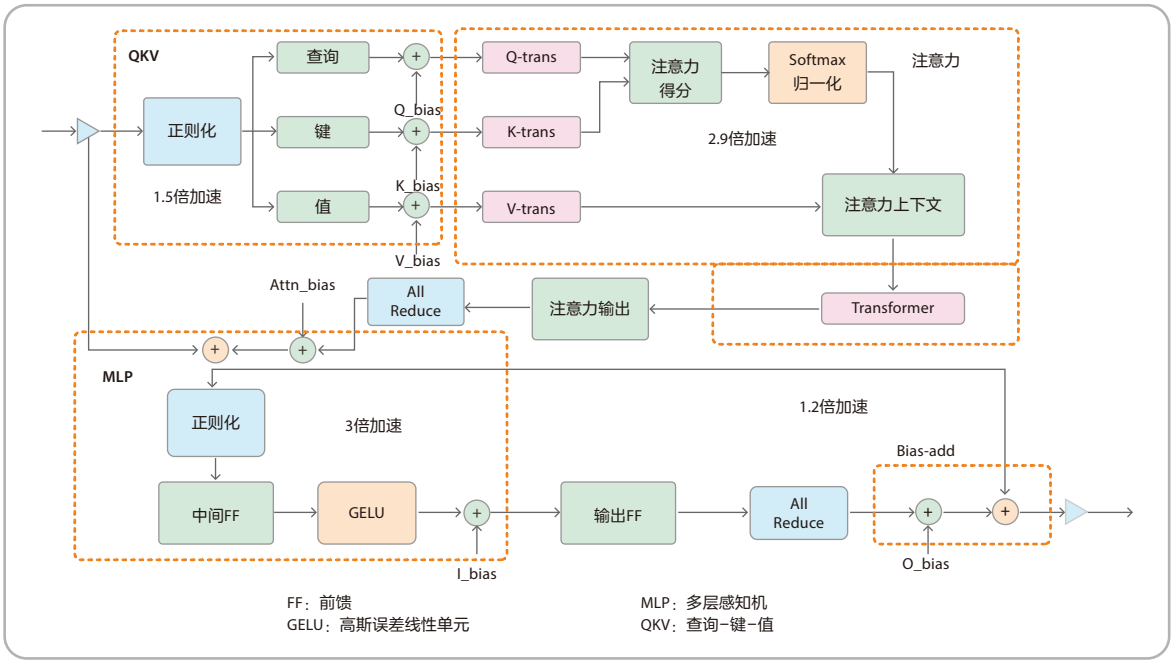


图1 Transformer层中的算子融合示意图

▼ 表1 StarCoder-15B在A100-40GB上测试查询-键-值融合

批大小（样本数）	输入长度（token数）	输出长度（token数）	注意力基线（s）	注意力融合（s）	加速率
10	1024	1024	27.1744	22.7952	19%
30	1024	1024	39.0770	37.4826	4%

▼ 表2 StarCoder-15B在A100-40GB上测试全连接层（FC）+激活融合

实测的TFLOPS	B=1，M=1，K=6144	B=4，M=1，K=6144	B=16，M=1，K=6144	B=64，M=1，K=6144	B=256，M=1，K=6144
基线（s）	0.3	1.2	4.7	17.6	30.3
融合MLP（s）	0.3	1.2	4.9	19.1	47.1
加速率	0.0%	0.0%	4.3%	8.5%	55.4%

注：B代表Batchsize；M和K表示矩阵乘法中的两个维度，M恒为1（解码阶段），K恒为6144

▼ 表3 CodeLLaMA INT4量化在HumanEval上的性能

HumanEval↑	7B/%	13B/%	34B/%
16bit原始精度	35.98	35.98	51.22
RTN（直接量化）	36.59	33.54	46.34
AWQ	35.98	31.71	50.61
SmoothQuant+	35.98	37.80	53.05

下，通过缩小模型规模和计算量来提高模型的运行效率，其中模型量化是最具实用性的技术。

SmoothQuant是典型的8bit LLM量化方法，该方法引入了逐通道缩放变换，有效地平滑了幅度，这使得模型更易于量化。激活感知权重量化（AWQ）和生成式预训练Transformer GPTQ是典型的权重量化方法，且权重量化的是group粒度。GPTQ提出了一种基于近似二阶信息的新型分层量化技术，使得每个权重的比特宽度减少到3或4位。AWQ的研究人员发现，对于LLM的性能，权重并不是同等重要的，仅保护1%的显著权重可以大大减少量化误差。

中兴通讯提出了SmoothQuant+4bit权重量化训练后量化（PTQ）算法。不同于AWQ对单个层搜索量化参数，SmoothQuant+对整个模型搜索量化参数，并对整个模型进行同一个参数平滑激活，这样能够从模型整体减少量化误差，且搜索效率更高。SmoothQuant+在LLaMA系列模型可以得到比AWQ更好的精度（见表3），同时在性能上也优于AWQ，对应的推理核已开源。

随着大模型上下文长度的增加，KV Cache占用的显存将超过权重和激活，因此对KV Cache进行量化可以显著降低大模型在长上下文推理时的资源占用，从而允许系统支撑更多的并发请求数和吞吐率。

并行推理

当大模型参数量超过单一计算设备所能容纳的上限时，则需要使用分布式并行推理技术。并行推理可以使用模型并行和流水线并行，而模型并行由于可节省显存资源、可降低单用户时延等优势，成为首选的并行方式。业界最流行的模型并行方案来自Megatron-LM，它的开发者针对Self-Attention和MLP分别设计了简洁高效的模型并行方案。而节点间带宽对模型并行效率有较大影响，高速串行计算机扩展总线标准（PCIE）的理论带宽为32～64Gbps，通常可以满足大模型并行推理需求。

服务调度优化

服务调度优化主要考虑的是系统同时为多个用户服务时如何尽可能地提升资源利用率。Continuous Batching和Dynamic Batching主要围绕提高可并发的Batchsize来提高吞吐量，异步Tokenize/Detokenize则通过多线程方式将Tokenize/Detokenize执行与模型推理过程时间交叠，从而实现降低时延目的。

Continuous Batching可以将传统batch粒度的任务调度细化为step级别的调度，这解决了不

中兴通讯研发了星云编程大模型，通过上述技术优化，实现显存节省70%，单GPU卡吞吐量提升3倍，推理时延降低一半，推理成本降低75%左右。目前中兴通讯内部已经建设研发大模型推理集群，将大模型辅助编程集成到研发IDE环境中，每日超过1.3万员工使用编程大模型进行开发，日生成代码超过百万行。

同长短序列无法合并到同一个batch的问题，大幅提升推理效率和用户体验，目前已在Hugging-Face TGI、vLLM、TensorRT-LLM等多个推理框架中实现。

新兴技术

我们将传统优化技术引入大模型推理的同时，也在探索从大模型自回归解码特点出发，通过调整推理执行过程来进一步提升推理性能。并行推测解码作为新兴的推理技术，可以在不损失精度的前提下提高推理速度。

投机采样是一种并行推测解码算法，开创了“小成本生成+大模型验证”的推理技术路线。该算法在已有大模型的基础上，引入一个小模型执行串行解码来提升速度，原大模型执行并行评估采样，保证生成质量，这在保证精度一致性的同时降低了大模型解码的次数，进而提升了推理效率。由于投机采样算法的巨大潜力，有多项工作在其基础上研究改进。但投机采样的推理方式并不适用于所有的应用场景。例如，在文学艺术类的诗词等应用场景，大小模型生成的结果概率分布相差较大；对于代码生成的场景，投机采样比较适合。随着业界研究的深入，投机采样会成为大语言模型推理的必备优化技术。

中兴通讯研发了星云编程大模型，通过上述技术优化，实现显存节省70%，单GPU卡吞吐量提升3倍，推理时延降低一半，推理成本降低75%左右。目前中兴通讯内部已经建设研发大模型推理集群，将大模型辅助编程集成到研发IDE环境中，每日超过1.3万员工使用编程大模型进行开发，日生成代码超过百万行。

此外，在城市治理、工业、矿山、交通等行业存在多种大模型应用场景，除数据中心部署外，在网络边缘、现场设备上也存在部署大模型需求。中兴通迅利用量化压缩技术降低大模型资源占用，降低了大模型部署成本；提供模型编译迁移工具适配不同硬件平台，将大模型部署在边缘一体机和AIBOX上，扩大了大模型应用范围。通过上述技术，星云大模型已经在城市小散工程监管、矿山安全生产、交通管控等实际落地，让行业客户真正用得起大模型。

随着ChatGPT热度的逐渐褪去，对大模型的投资也逐渐趋于理性。大模型如何产生真正的商业价值成为全行业都在思考、探索的问题。随着大模型规模的不断增加，模型能力在提升的同时，算力成本也在不断飙升，这给大模型长期可持续发展带来了不确定性，因此以实现更低成本算力和更高效率算法为目标的核心技术亟待突破。大模型机遇与挑战并存，加速发展的趋势在中长期不会改变。ZTE中兴

星云行业大模型，构筑产业智能化转型新引擎

中兴通讯 任军，钟政

当前，中国大模型进入发展加速期，自然语言处理、视觉处理和多模态等各技术分支均快速发展。在产学研各方共同推动下，我国已建立起涵盖理论方法和软硬件技术的体系化研发能力，据不完全统计，国产大模型数量目前已超过200个，形成了紧跟世界前沿的大模型技术群。

星云行业大模型

任军
中兴通讯产业数字化首席架构师

钟政
中兴通讯综合方案总监

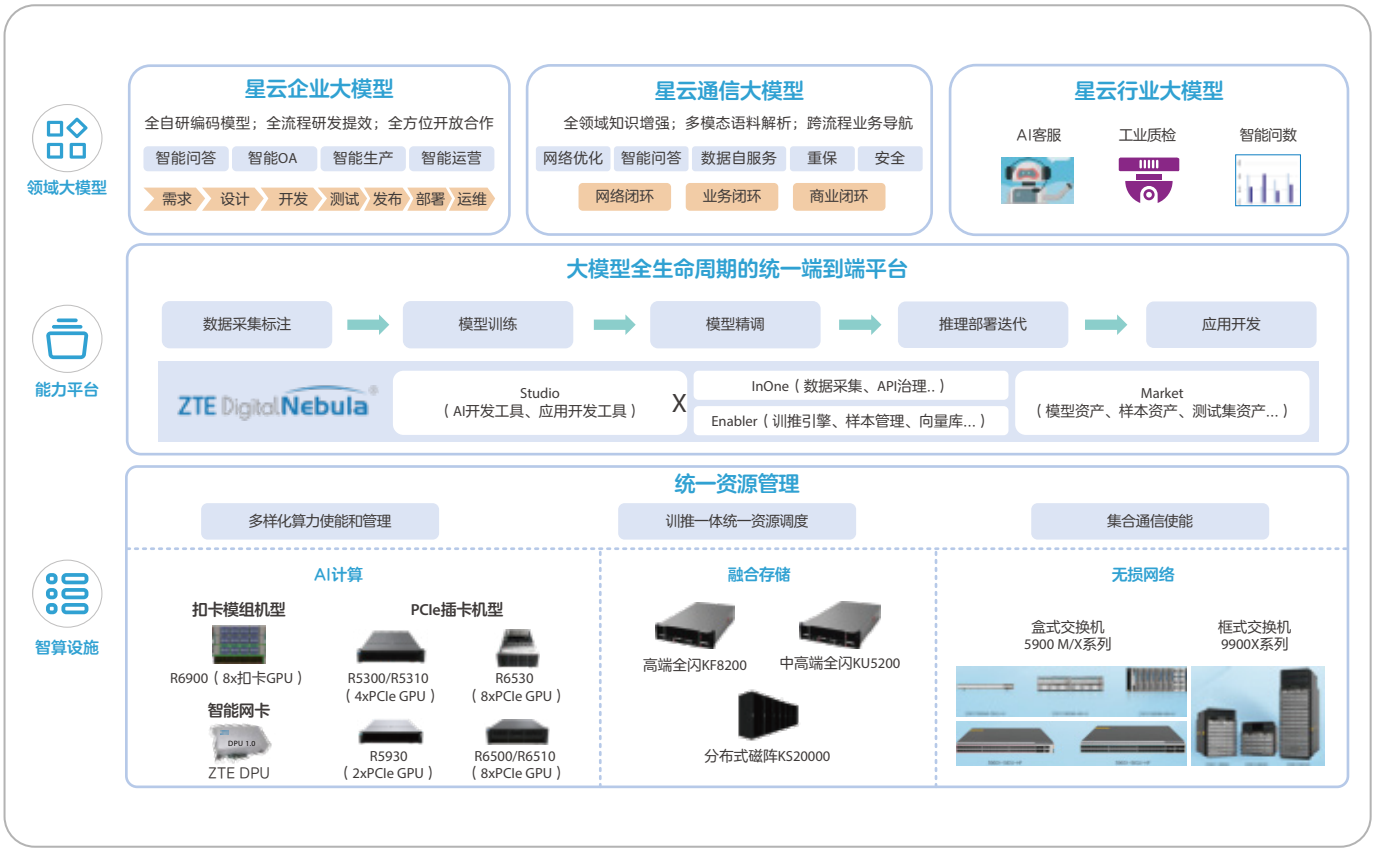
当前，中国大模型进入发展加速期，自然语言处理、视觉处理和多模态等各技术分支均快速发展。在产学研各方共同推动下，我国已建立起涵盖理论方法和软硬件技术的体系化研发能力，据不完全统计，国产大模型数量目前已超过200个，形成了紧跟世界前沿的大模型技术群。

与此同时，行业客户开发大模型应用依旧困难重重，大模型训练、微调、应用开发对开发人员提出很高要求，而一般企业往往缺乏专业的技术人才，导致企业在大模型方案选择、技术实施和系统维护上遭遇困难；此外大模型训练开发是高设备资产投入、高技术人力投入的复杂工程，大部分企业缺乏配套的资金预算；基于数据安全、合规安全等考虑，很多企业不能使用互联网大模型SaaS、MaaS服务，加剧了上述两个问题，最终导致很多企业对大模型应用求而不得。

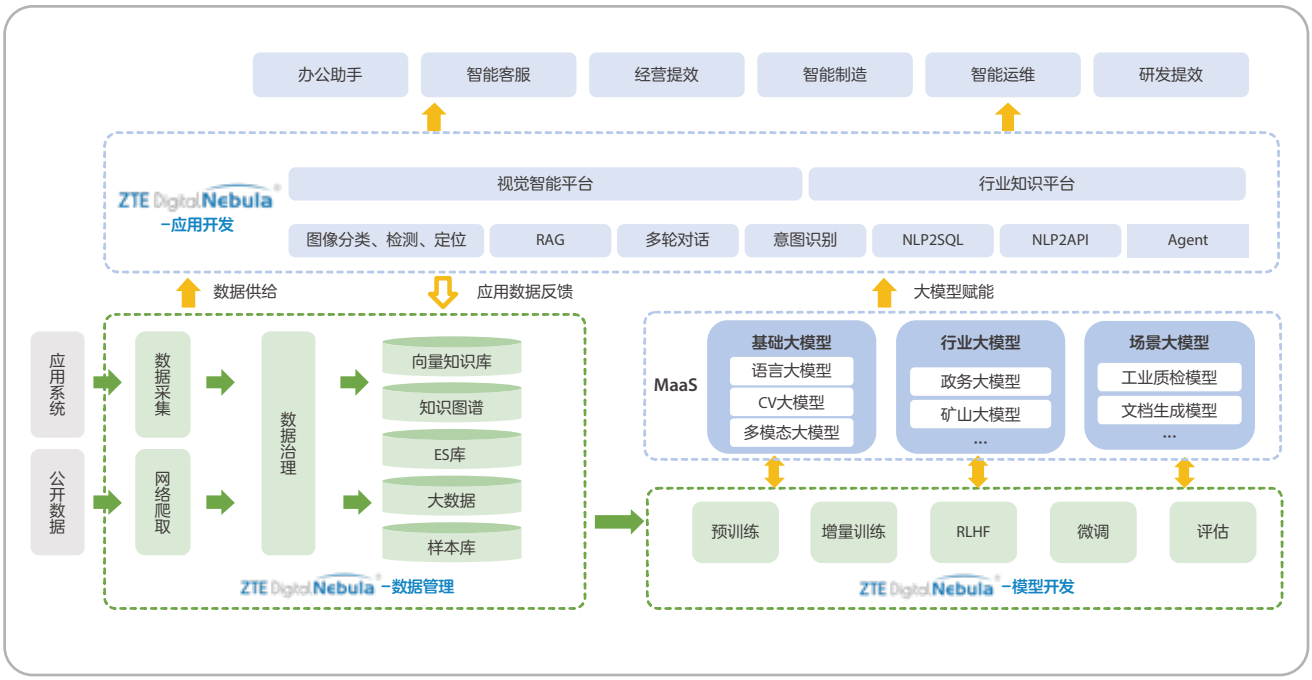
星云大模型端到端方案

星云大模型从一开始就以助力行业客户大模型应用开发为目标，提供端到端方案（见图1）。

智算设施层，中兴通讯提供绿色低碳基础设施方案，通过液冷IDC，支持高功耗GPU服务器高密度部署，PUE低至1.2以下；提供解耦开放的算力方案，一机多卡、随芯而选，灵活配置Intel、海光、中兴通讯自研“珠峰”芯片等多种CPU，广泛支持Nvidia、Intel、寒武纪、壁仞等多种GPU；提供高性能并行存储方案，研发了全闪磁阵和并行文件系统，满足热、温、冷数据多样化存储需求；提出“以网强算”理念，针对机内互联与产业伙伴一起推出开放的GPU高速互联标准



▲ 图1 中兴通讯星云大模型端到端方案



▲ 图2 中兴数字星云端到端支撑行业大模型开发

Olink，针对机间互联自研全系列RoCE交换机，提供高速无损以太网组网方案，提升包括国产GPU在内的算力集群规模和效能，让国产GPU更好地训练大模型；提供多资源统一管理调度平台，实现多样化算力使能、图形化监控运维、故障自动恢复及断点续训功能，确保大模型训练长时间稳定、高效运行。

能力平台层，今年4月中兴通讯发布数字星云3.0，为大模型开发提供从样本数据管理到模型部署的全流程工具和能力（见图2）：提供数据采集、清洗、标注、评估全流程样本数据管理功能，特别是大模型辅助数据标注，效率相比人工标注提升80%；实现自动化并行训练、多种微调、强化学习等功能，经过优化的并行训练效率比开源提升90%以上；实现模型编译、量化压缩等能力，其中自研无损量化算法可实现显存节省70%、吞吐量提升90%，为大模型低成本部署提供支撑。

中兴通讯开发了星云语言大模型、视觉大模型和多模态大模型。星云语言大模型提供数十亿到千亿不同参数规模的基础模型，并针对文档生成、API映射、Text2SQL等常用场景训练微调了不同场景模型，行业客户可以根据其任务复杂度、资金预算情况选用最佳模型。星云视觉大模型基于Transformer实现了视觉分类、检测、定位、分割、检索等任务，与传统视觉小模型相比具有提示词泛化能力，通过数十样本Prompt Learning技术可将某具体场景任务准确率提升到95%以上，满足商用需求，大大降低了视觉任务开发成本和时间；同时视觉大模型具备人类意图理解能力，在复杂业务逻辑的视觉任务，以及火焰、水淹等传统小模型误报率很高的场景下，依然具有较高准确率。星云多模态大模型目前以理解图像/视频、生成文本为主，为视觉应用提供更方便的交互模式，同时正在融合语音、3D等模态，实现数字人生成、多模态对话等更广泛的业务场景。

大模型价值最终体现在行业应用上，赋能行业应用开发是星云行业大模型方案着重打造的能力。

针对语言大模型应用开发，中兴通讯推出了行业知识平台。行业知识平台是一个可灵活扩展的架构，提供模型、提示词、插件等基础能力管理，使平台既可使用星云语言大模型，也广泛支持业界主流语言大模型。知识平台也提供常用大模型应用框架：

- 知识问答系统：通过对话从海量文档库中准确找出问题答案并给出回答依据，支持多种输入文档格式，支持公式、统计图表、复杂表格等多种文档内容理解能力，为客户提供一套文档处理、调优、评估自服务流程，通过多个方面优化，将开源RAG 40%~50%的准确率提升到95%以上。
- 智能问数系统：允许普通业务人员通过对话实现数据查询和分析，通过对业务数据进行治理形成指标、维度术语体系，基于该术语体系，大模型数据查询准确率达95%，并实现了数据自动预警、根因分析、关联分析，使大模型达到专业BI分析师能力。
- 低代码大模型应用编排：提供图形编排工具，利用大模型意图理解、信息抽取、API调用等灵活处理能力，快速开发智能行业应用。
- Agent应用开发框架：提供API模型自动生成、调优、评测CI工具实现API准确映射，利用LLM的CoT、ReAct等能力，实现任务规划、拆解；提供Agent长短期记忆、Agent通信机制，将一个复杂任务拆解成多个角色Agent，通过Agent间协同和博弈实现复杂任务的自动执行，并在多Agent执行过程中引入人工参与和确认，确保任务受控。

视觉智能分析需求层出不群，算法日新月异，为让客户使用新算法快速解决新场景问题，中兴通讯针对视觉行业应用发布了视觉智能平台：

- 通过低代码编排实现对监控摄像头视频流接入、抽帧、AI分析、自动告警全业务流程，降低视觉任务开发工作量。
- 提供算法仓管理，允许客户根据需要自行上载算法，实现对第三方算法统一管理和自动部署，用于视觉任务编排。
- 支持大小模型协同。视觉大模型由于参数量巨大，推理消耗算力资源较高、吞吐率较低，而现网运行的传统小模型消耗算力资源低、吞吐率高，依旧具有很强的实用价值，视觉智能平台通过大小模型协同形成互补，使得视觉智能应用同时获得两者优势。

星云行业大模型：企业用得起、放心用

星云行业大模型方案推出多种企业私有化部署模式。面向万亿级参数模型训练，提供数据中心版（Cloud版），通过图形化运维功能实现资源高效调度利用，通过故障自动修复、断点续训保证长时间稳定运行；面向以推理微调为主的中小企业提供一体柜版（Campus版），通过端到端简单易用训推工具降低大模型开发和部署门槛，通过开箱即用缩短业务上线时间，实现数据不出园；面向各种终端设备提供现场版（Site版），通过量化压缩技术降低模型资源占用，通过模型迁移适配多种终端平台，与云端形成大小模型协同、云边端协同，实现整体性价比最优。多种部署方式不但能适应不同业务场景，也给企业客户提供多种资金投入选择。

星云行业大模型端到端方案，降低了大模型训练、应用开发的技术难度，解决了企业客户面临的大模型技术问题，也在很大程度上解决了企业的资金投入问题：

- 中兴通讯针对重点行业的通用业务场景进行针对性训练微调，开发出系列星云行业大模型，大部分行业客户可直接选用，避免自行训练大模型的高投入；
- 行业应用开发平台提供了知识问答、智能问数、低代码应用开发、低代码视频业务编排等功能，降低了大模型应用开发的技术难度和工作量，降低了大模型应用成本；

中兴通讯通过大模型提升企业日常办公、经营管理、智能生产等核心业务效率，其中研发大模型，实现从需求管理、辅助设计、代码生成、自动测试、版本发布、运维全流程提效，代码生成能力达到GPT4等世界领先水平，已经被中兴通讯数万研发人员使用，助力编程提效30%以上。

- 以推理为主兼顾少量微调需求的企业，星云一体柜最大程度降低私有化算力建设成本；
- 中兴星云大模型提供不同参数量模型供客户选择，行业客户可在满足精度前提下选择尽量小的模型，节省算力投入；提供无损量化及多种推理引擎优化技术，进一步降低大模型使用成本。

在企业关注的应用安全方面，星云行业大模型坚持企业私有化部署，通过企业数据不出园解决大模型应用中的数据安全问题。针对大模型应用的其他安全风险：如针对大模型训练数据投毒，个人敏感数据泄漏，在大模型训练阶段埋入后门，在大模型推理阶段进行模型越狱、提示词注入等攻击，以及传统的侧信道、远程代码攻击等，星云大模型提供全面防护方案，确保大模型内容安全和运行安全。当前中兴星云大模型已经通过国家网信办备案，充分说明了星云大模型的安全可靠。

星云大模型赋能千行百业

中兴通讯通过大模型提升企业日常办公、经营管理、智能生产等核心业务效率，其中研发大模型，实现从需求管理、辅助设计、代码生成、自动测试、版本发布、运维全流程提效，代码生成能力达到GPT4等世界领先水平，已经被中兴通讯数万研发人员使用，助力编程提效30%以上；中兴通讯滨江生产基地，利用大模型技术实现智能排产、工艺文档生成、工业质检、AGV调度、智能维修等，实现了数字工厂到智能工厂的进阶。

面向电信运营商，中兴通讯提供通信大模型，帮助运营商进行网络优化、短信防欺诈、数据自服务，并采用多Agent技术实现5G网络自动化运维，在提升网络服务质量的同时，降低运营商运维人力成本80%。

面向行业客户，中兴通讯致力于建设开放的生态系统，让更多应用开发商用好大模型，从而服务于更多行业客户。在政务领域，星云政务大模型已经服务于城市治理、应急指挥、城市生命线安全等场景；在交通领域，星云交通大模型已经服务于轨道交通行车安全检测、港口作业安全检测等；在工业生产领域，应用于产品质检、自动化产线、计划排产等场景；在水利领域，应用于水文数据分析、河道演变分析、湖河四乱检测等场景。

经过中兴通讯内部实践和外部行业应用广泛验证，数字星云3.0与星云大模型日臻完善。展望未来，中兴通讯将继续秉承开放合作的精神，与各行各业合作伙伴共同探索，利用星云大模型的强大能力，助力不同领域的行业客户实现智能化转型与创新发展，共同绘制智能时代的宏伟蓝图。ZTE中兴

构筑智慧韧性城市，星云大模型驱动城市运行革新

陆志峰
中兴通讯政府市场总监

王庆
中兴通讯产业数字化方案部副总经理

人工智能是推进现代化产业体系建设，加快发展新质生产力的核心驱动力。AI大模型以其在自然语言处理、视频处理等方面的超凡能力，快速应用于各行各业。2024年《政府工作报告》提出深化大数据、人工智能等研发应用，开展“人工智能+”行动，推进建设智慧城市、数字乡村。广东、上海等省市发布“人工智能+”在数字政府、政务服务、市域治理领域深度融合应用通知。

星云大模型是中兴通讯自主研发的基础大模型，基于星云大模型，通过政府领域知识增量预训练，我们推出星云政务大模型，服务于一网统管、城市生命线安全等场景，用人工智能赋能城市运行智慧化、城市安全韧性化。

城市运行面临的挑战

随着城市的发展，城市规模扩张，基础设施逐步老化，城市运行复杂度随之增加，积累的数据呈指数增长，给城市治理带来巨大挑战。

在城市数字化治理方面，随着各类事件数据不断增长，以人工分拨为主的低效工单系统已不能满足事件处置的要求。同时，各部门AI能力分散建设，算法算力缺乏统筹，以单场景的智能化建设为主，智慧化建设在城市范围内未形成统一规划，难以支撑上层应用的实施。

在城市安全方面，主要涉及燃气安全管理、地下管线损坏预警、道路积水风险预警、小散工程施工安全管理等。燃气安全管理，市场监管部门对违规行为的监督，主要通过燃气场站和液化气充装场站安装的监控视频人工审核，效率低、监管覆盖面窄。而地下管线损坏、城市道路积水一直缺乏有效的手段及时预警。大部分城市房屋建设、线路管道和设备安装及装修等小散工程的安全事故数已超过当地交通事故数，人力巡查成本高、效率低。

在应急处置方面，传统的应急预案缺乏针对特定风险场景的构建，应对措施的可操作性不强。很多预案未能及时更新，导致预案内容与实际脱节，无法有效指导应急响应。同时，不同层级、不同部门之间缺乏衔接和协调，存在职责不清的问题。

城市运行向智慧化、韧性化发展

星云政务大模型的项目实践应用，解决了城市运行中存在的问题，大大提高城市治理、城市生命线安全监管和应急处置的效率。生命线专家级助手作为首个城市生命线领域的行业助手，大模型和应用分离，轻量化可消费级显卡部署，覆盖城市生命线主要业务场景知识。在日常事件处置时，通过对事件的智能分析，自动派发工单；在应急处置时，自动生成预案，快速处置，辅助决策。星云政务大模型应用场景如图1所示。

城市智慧化治理，高效事件处置

中兴通讯结合城市治理场景，整合算法资源，充分发挥星云政务大模型的性能，能力共建共享，在南京、深圳等区级一网统管和智慧社区项目应用中提供服务。

星云政务大模型赋能城市治理场景，对上传的图片/视频进行深度分析，生成相应的工单标题和详细描述，实现单个图片/视频多场景识别，优化资源配置，提升效率。系统能够精准识别城市治理主要场景，包括烟火、电动车、占道经营、机动车乱停放、违规户外广告、沿街晾挂、路面积水和井盖缺失等。同时，针对城市治理中视觉识别的复杂性和多样性，该模型能够以极少的样本快速适应新场景，大幅缩短算法开发周期，降低成本。

将城市治理事件分拨从传统的手动模式变为智能化自动模式，对事件内容进行自动分析和分类，工单系统精准高效派单。基于星云语言大模型分析识别事件内容，结合事项清单、统一事件办理与反馈标准，系统自动推荐给责任部门核实并处置。事件处理的时限整体压缩50%，处置效率显著提升。这种智能化升级不仅优化了工作流程，还极大地提高了群众的满意度，体现出精细化治理能力的提升。

星云政务大模型安全预警分析，增强城市韧性

城市生命线安全预警分析，对城市生命线接入视频进行大模型分析研判，发现城市生命线各类场景的安全隐患和风险，提前预警，提前预防，提升生命线监管能力，增强城市韧性。

针对液化气场站视频，结合用户实际需求，梳理10多种违规场景进行分析，如使用手机、倚坐液化气瓶、操作不当导致液化气泄漏、充装现场堆放气瓶过多等场景。星云政务大模型已完成某市2023—2024年液化气场站视频数据模型训练、模型推理，整体危险行为识别准确率达87%，

▲ 图1 星云政务大模型在城市运行中的应用

监管范围扩大到全市所有场站，监管时间延长到7×24小时，效能大大提升。

针对道路的视频分析，实现道路非法施工的识别，有效降低由于非法施工带来的管线破坏事故数量；通过视频判别低洼区域积水和水位，及时预警，为城市内涝风险评估提供依据；通过工程车或公交车视频识别道路裂缝、坑洞等健康问题，预防道路塌陷。在昆山等地城市生命线安全工程项目应用中，路面施工挖掘机的识别准确率达95%；道路路面病害识别目前支持积水、裂缝、坑洞等多个场景，准确率达92%。

城市生命线安全预警分析基于星云政务大模型，一个模型支持多场景识别（一图多义），以提示词驱动场景构建，解决传统算法场景单一、上线周期长、对接成本高等问题，提升了算法准确率，降低算法成本。

小散工程智慧安全监管，创新巡查手段

针对小散工程安全监管，结合现场场所变更频繁的特点，中兴通讯推出智慧施工星云杆方案。该方案快速灵活部署，现场异常情况及时发现，创新巡检常态化的智慧监管手段，解决了小散工程监管不可控的痛点。

基于云边协同的方案设计，边侧进行图像分析和处理，云端进行算法训练、算法管理和云侧推理。采用1+N模式，实现1个智慧中枢，多个工地实时监管及异常违规实时告警，并具备远端视频监管和对讲功能。

智慧施工星云杆及智能算法应用于深圳宝安区施工工地场景，与现有城市生命线施工、智慧社区等场景结合，打造“街道/社区+AI”应用场景高地，赋能城市治理。

应急智能体，辅助指挥决策

应急智能体生成应急预案用于应急事件快速处置。以燃气爆炸为例，应急指挥中心根据事故描述使用RAG检索获取应急预案，利用应急预案拆解功能，提取应急处置流程的关键步骤和分工，燃气办、燃气公司等职能部门不需要全文通读寻找，进一步缩短响应时间。指挥中心监控进展，根据需要调整行动，最终生成完善的应急预案。应急智能体有效缩短应急响应时间，应急救援效率提升2~3倍，减少人民生命和财产损失。

应急智能体应用于多个场景的高效应急。在安监执法场景，提升现场执法问题的咨询效率，通过智能体，每位执法人员配备虚拟法律专家，缩短新人培养周期80%；在监测预警事故调查场景，事故报告分析效率提升5倍。

生命线安全助手，创造知识价值

城市生命线安全助手基于领域大模型、多模态文档解析、大数据智能搜索等技术，自动生成AI知识库。对比传统行业知识库、机器人等知识管理手段，生命线安全助手可轻松实现知识归纳、构建、问答、推荐等，精准获取有效信息，大量节省文档检索时间，提升工作和学习效率。

目前星云政务大模型已对国内生命线相关标准规范、政策文件、公开论文，以及项目方案进行文档检索、智能摘要、实时解答，文档信息提取准确率超90%。城市生命线安全助手为生命线安全工程项目的实施和运维提供培训服务，包括问题实时解答和学习支持，支撑项目的快速和规范交付以及后期运营。

星云政务大模型城市运行场景应用已服务昆山、南京、深圳等城市生命线安全工程、一网统管和智慧社区等项目。星云政务大模型具有模型快速泛化、文本图像特征融合、模型极致压缩加速的优势，从而控制模型的计算成本，提高模型的实时性。

随着城市运行数据和监测数据的不断积累，星云政务大模型持续对各场景的监测预警模型迭代、参数校正，准确率将大大提升，为城市智慧化运行、韧性城市提供有力的技术支撑，推动城市管理手段和管理模式的创新，促进城市运行升级革新，实现智慧城市的高质量发展。ZTE中兴

水利发展新动能
——大模型推进行业应用效率提升

党的二十届三中全会正式通过《中共中央关于进一步全面深化改革、推进中国式现代化的决定》，对人工智能的发展提出了明确指导和要求。全会强调了人工智能作为战略性技术的重要性，并指出其在赋能新质生产力方面的关键作用。融合AI的数字孪生技术在水利建设中的应用，是实现水利新质生产力发展的重要抓手。知识平台作为与人工智能大模型耦合最为紧密的模块，是数字孪生水利的重要组成部分，它将业务规则和历史经验转化为知识库，为水利智能转型升级提供基础，为决策者提供科学、准确的研判支持。

现阶段水利行业知识平台建设积极拥抱大模型，水利大模型的打造与数字孪生知识平台建设相结合，基于理解水利行业的专业语言和运行逻辑，调用行业专业模型和系统工具，展示行业需求并驱动专有设备，为防汛决策、水库精细化调度等提供支持。

大模型在水利行业应用的价值

在人工智能技术飞速发展以及数字孪生水利建设的需求导向下，大模型与水利传统业务的融合正处在起步期，如何让大模型在水利行业发挥其应用价值，亟待大模型技术厂商和水利行业专家共同探索。基于当下涌现出的大量创新实践案例能够预见到其蓬勃发展趋势，同时也能够初步厘清大模型与水利行业的应用结合的价值点主要在以下两个方向。

助力水利专题图绘制及分析报告编写

为了确保水利管理的精确与高效，图像识别类智能模型与水利大模型和数字孪生技术结合，可为用户提供定制化服务。用户可以根据自身需求进行设置，明确专题内容、时间时段、空间范围等信息，系统动态生成涵盖降水量分布、水库蓄水量、河流水位等关键水文和空间信息的专题图表。具体应用场景包括：基于遥感影像，通过大模型调用遥感智能识别模型，识别河道管理范围内的建筑物；通过大模型自动检索河湖遥感平台本底数据库，识别新增的建筑物清单；通过大模型自动生成河道新增疑似碍洪违建分析报告，并给出现场核查最优路线方案等。

结合水利专业模型提升防洪“四预”工作效率

水利专业模型是数字孪生水利的核心和关键，对于业务人员专业技术要求较高，水利大模型的应用可有效提升管理人员对海量水利业务数据的分析处理能力，通过大模型驱动数字孪生技术，也可对各类复杂决策进行快速、低成本的预演和优化评估，有效提升决策效率。

基于大模型对多源水利数据知识挖掘技术，水利专业计算和智能决策类应用可实现多业务领域中水利对象、专业模型交互对接的能力。多业务融合知识图谱本体构建方法，使大模型与业务规则、学科知识、专家经验、历史场景等水利知


李锴
中兴通讯水利行业规划总监


赵昕
长江水文技术研究中心主任、正高级工程师


原松
长江水文技术研究中心四室副主任、正高级工程师

识库以及人类自然语言理解能力进行对接，实现业务应用场景驱动水利大模型开展水利专业计算与智能决策。

星云行业大模型助力水文大模型构建实践

中兴通讯星云行业大模型，基于自主创新、开放解耦的智算底座支撑，融合MLOps全生命周期的知识平台，践行AI多方应用，端到端提供语言大模型、视觉大模型、多模态大模型多元能力。面向水利行业结合水文业务特性构建水文大模型（见图1），致力于提升水文工作效率，强化防洪减灾基础能力。

水文工作在防洪减灾应用场景起到关键基础作用，通过收集和分析水文水资源数据信息，包括收集地质构造数据内容、地表径流量分布数据、地下水深度数据及水资源主要存量数据，服务科学完善的防洪减灾机制建设。水文大模型是一个复杂的系统工程，从弱人工智能到强人工智能，大模型建设具体可以分为三个阶段。首先基于星云语言大模型打造水文知识助手，为水文工作者提供领域专业知识支撑以及常规报告报表生成服务；进一步融合星云视觉大模型能力，逐步针对核心场景打造水文专业人工智能应用，例如河道演变分析、工程计算、水文测验等场景，帮助完成简单的日常工作任务，解放人力，减少时间成本。第三阶段，面向大模型结合水利专业模型，基于Agent改变复杂行业应用交互范式，为行业用户提供综合性的视觉展示与知识服务。

星云大模型协同流域单位共同深化研发，实现水文大模型从专业助手向业务能手的转变。

水文知识助手

水文大模型运用水文测报、河道地形、重要库区等运行管理多年来的所有成果汇集、分类、归纳，形成知识库，做到随用随查。水文知识库在工作科普、会商、调研、展览、汇报与知识提取等场景都有迫切的实用性，不仅汇集了水文测报、河道治理方面的常规知识，河道长度、水库库容等基本数据，又能给出各种统计成果、分析结论。

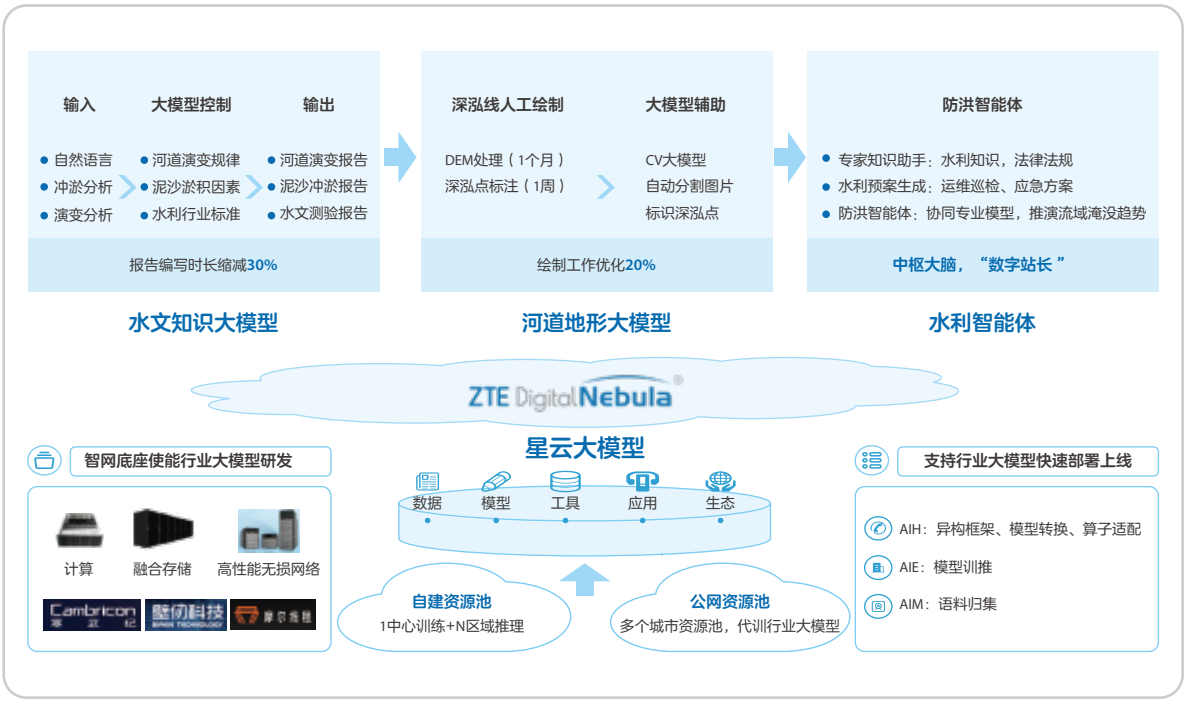


图1 水文大模型方案 ▶

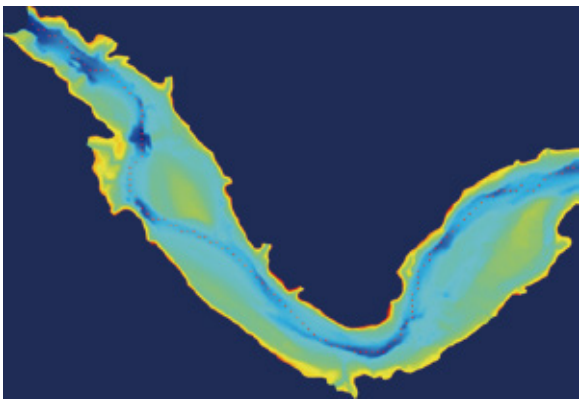
基于水文大模型的水文知识助手可帮助用户完成以下工作：

- 水文数据查询。提取基础资料，智能分析统计结果，数据查询准确率达90%。如提问：“统计2023年11月某水利工程泥沙输移情况”，回答如下：“2023年11月，入库悬移质泥沙11.8万吨，出库泥沙8.96万吨。入库沙量最大值出现在11月16日，最小值出现在11月24日。”
- 报告简报生成。通过大量分析报告学习，智能组织生成关键流域泥沙特性分析报告、崩岸简报，报告编写时长平均缩减30%。

河道演变分析

河道演变分析是水文河道泥沙专业工作中最频繁的基础性工作，因其复杂性，长期以来均由人工在计算机系统辅助的条件下开展工作，且人工程度远高于计算机辅助程度，分析计算效率相对较低。在星云视觉大模型支撑下，首先利用先进的AI计算提升算法、模型的计算效率，其次提高计算机辅助与AI智能在河道演变分析工作中的权重，达到提高生产力、解放生产力的作用。

- 河道深泓线岸线变化分析。该业务场景在提高效率方面主要包括槽蓄量库容计算、断面切割、河道冲淤变化分析。AI智能辅助是将河道地形转换为数字高程模型（DEM），结合多测次遥感影像，利用星云视觉大模型技术将其作为图片的一种形式，主要包括深泓线平面变化分析（见图2）、洲滩岸线变化分析。通过大模型的能力结合，现阶段深泓线的绘制工作效率能够提升20%以上。
- 河道下垫面地形预测。预测河流底部地形的变化，对于流域防洪、航道维护和水资源管理等方面至关重要。传统的河道测量方法，如使用测量船进行河床深度的密集测量，测量一次耗费几千万元，效率较低。运用横断面水深测量、卫星遥感技术，对河流进行高频次、高分辨率的监测，进一步结合星云视觉大模型分析遥感影像数据，建立算法模型预测河流底部地形的变化，能够将河道下垫面地形预测准确率提升到80%。



◀ 图2 河道深泓线平面图智能生成

持续探索，携手行业用户逐步完善水利大模型能力

中兴通讯在水文大模型领域自主创新，通过从语言、视觉到多模态的技术持续演进，将探索深入核心业务场景，基于数字星云平台筑牢大模型智算架构，融合气象、GIS、水文以及行业历史数据，通过对海量数据的分析，构造辅助水利机理模型的智能大模型，发挥“中枢大脑”的作用，服务于三道防线建设以及“四预”应用体系，为决策者提供更加系统、科学的依据。从全面感知、智慧决策、高效管理多个方面，有效提升工作效率。

随着星云大模型助力行业应用在人工智能领域的探索不断深入，融合数据、模型、工具、应用和生态，并提供AI计算、融合存储、高性能无损网络等软硬件的综合性需求将日益凸显，中兴通讯星云大模型架构，将依托数字星云平台，链接优势智算产品以及行业生态伙伴，加速拓宽水利行业大模型业务应用的广度，以及与一线需求结合的深度，力求在防洪减灾、水资源管理、水利工程设计等多场景，为水利事业的发展提供全方位支持。ZTE中兴

星云大模型引领交通管理与服务创新

丁成远
中兴通讯交通市场总监

姜永湖
中兴通讯产业数字化方案部副总经理

彭亦辉
中兴通讯交通规划总工

在国家“交通强国”战略整体指引下，我国已建设了世界领先的综合交通运输体系，随着移动互联网、物联网、大数据和人工智能等技术的持续发展，交通行业正从高速发展迈向高质量发展阶段。以大模型为代表的新一代科学技术作为新质生产力的代表，正以其高效、智能的特性，深度助力交通行业持续数字化升级，通过优化交通管理、提升服务质量和安全水平，实现交通系统的精准预测、智能调度、安全保障与高效运行，不断推动交通行业的全面革新与发展。

交通行业细分领域多，交通大模型不可或缺

交通行业作为一个庞大而多元化的领域，涵盖了城轨、铁路、港口、公路等多个细分领域。其不仅承载着巨大的物流运输任务，还直接关系到公众的日常出行和国家的经济发展。在这个背景下，人工智能技术的引入和深入应用，成为推动交通行业转型升级、提升效率与服务质量的关键力量。

在城轨领域，城市轨道交通系统需要高效、精准的运营调度来确保列车准时、安全地运行。通过大模型可以预测未来一段时间内的客流变化趋势和列车运行需求，从而优化列车运行图，提高运营效率。同时，大模型可以集成自然语言处理、图像识别等技术，为乘客提供智能化的信息服务，如语音购票、人脸识别进站、实时换乘指引等，提升乘客体验。城轨系统的设备设施众多，维护成本高昂，可以利用大模型对设备运行数据进行深度分析，识别出潜在的故障模式，提前预警并采取相应的维护措施，降低维护成本和停运风险。

在铁路领域，铁路系统需要确保列车运行的安全和准时。然而，面对复杂的路网结构和多变的运行环境，传统的调度和安全管理方法存在局限性。大模型可以实时分析列车运行数据、天气信息、设备状态等，为调度员提供精准的调度建议和安全预警，提高铁路运输的安全性和效率。面对乘客个性化、便捷的服务需求，大模型可以分析乘客的出行习惯、偏好等信息，为乘客提供定制化的服务方案，如个性化推荐、智能导航等，提升乘客体验。

在港口领域，现场作业环境复杂，存在多种安全隐患，传统的安全监控方式往往难以全面覆盖和及时响应。可以通过大模型对监控视频进行智能分析，实时识别出人员违规操作、设备故障等异常情况并及时预警和处置，保障港口作业安全。同时，港口作为物流枢纽，需要实现货物的高效装卸、转运和存储，传统的物流管理方式存在信息不对称、协同性差等问题。大模型可以整合港口内外的物流信息，实现物流资源的优化配置和协同作业，提高物流效率和服务质量。

在公路领域，交通流量大、变化快，交通安全事故频发，大模型可以实时分析公路交通流量数据、车辆行驶轨迹等信息，预测未来一段时间内的交通流量变化趋势，缓解交通拥堵。大模型还可以和先进的驾驶辅助技术相结合，如自动紧急制动、车道保持辅助等，为驾驶员提供实时

的安全预警和辅助决策，提高公路交通的安全性。

星云大模型助力交通智能革新

中兴通讯基于星云行业大模型打造的星云交通大模型，通过结合交通行业数据，能够精准识别周围风险，实时监测交通设施运行状态，进行行为、人员、车辆等异常识别，及时预警并处理潜在故障，提高运维效率。星云交通大模型的应用有效降低了交通事故发生率，提升了交通安全水平，为公众出行提供更加安全、可靠的交通环境。同时，依托中兴通讯全新升级的数字星云3.0打造的交通数智平台（见图1），充分结合星云交通大模型能力，解决了AI在产业应用落地过程中的数据处理、训练推理、应用开发、灵活部署、安全保障五大关键挑战，更快更好地支撑合作伙伴基于交通数智平台的创新应用开发。目前基于交通数智平台赋能交通行业，已实现城轨、铁路、港口、公路等子行业100+的交通应用场景创新。

在城轨行业，中兴通讯基于星云交通大模型，整合网、云、数、智能力，打造城轨智能体，结合城轨供电、工务、信号、车辆等N个场景，提升对城轨的感知、检索、规划和执行能力，赋能各场景智能故障检测、智能建议生成、流程自动化、故障报告生成；实现视频巡检和实时客流监控，中心与车站云边协同与统一管理，助力城轨优化运营运维效率，提高乘客体验，加强应急响应能力。

在铁路行业，中兴通讯基于5G+云+AI核心技术，打造面向智能铁路、数字铁路的数智创新方案，以铁路业务全面数字化、数据充分共享共用、智能化水平提升为目标，基于云边端结合的训推解决方案，推动铁路行车安全、运营效率、服务体验不断提升，助力国家率先实现铁路现代化，为国家铁路勇当服务和支撑中国式现代化建设的“火车头”注入数字化新动能。

在港口行业，中兴通讯聚焦码头生产安全应用场景，通过星云交通大模型对码头复杂场景的现场视频和图像进行预处理分析，提取人员、机械、设施等目标特征，实现目标人员和机械设备

▲图1 交通数智平台赋能交通行业

的行为和动作定位、分类、识别，并通过星云交通大模型和小模型的结合，克服了码头场景下的大场景小目标、复杂作业流程判别、目标相对运动的判断等众多挑战，在保证实时检测高准确率的同时，实时或近实时地处理不同大小的现场目标对象，从而提高检测的准确性和召回率，进而提升了码头现场安全管理水平。

在公路行业，针对海量的高速公路数据，包括车辆行驶数据、路况信息、交通流量数据等，中兴通讯通过星云交通大模型对高速公路上的车辆和路况进行实时监控，检测异常行为（如违章停车、超速行驶、道路拥堵等），及时发出预警，提升行车安全。同时，对路面开裂、坑槽进行智能识别，及时发现路面的突发性坑槽、沉陷等影响行车安全的异常路况，对坑槽类病害实时报警，减少车辆行驶的安全隐患，同时督促病害维修的及时性，延长路面的使用寿命。

大模型实践探索，开启智慧交通新篇章

中兴通讯积极推动星云交通大模型的行业实践，覆盖了城轨、铁路、港口等多个细分领域，实现了从数据采集、处理到分析、决策的全链条智能化，为智慧交通的发展注入新动能。

在城轨领域，中兴通讯聚焦城轨运维、运营场景，在国内与多个地铁公司联合技术探索。在星云交通大模型的助力下，地铁工作人员无需具备数据开发能力，使用自然语言对话式的数据查询和可视化服务，通过对话式BI功能可以自动生成图文报表，提升自主探索数据能力；依托星云交通大模型开发的城轨视觉分析应用仅需要数天即可上线，并且数据标注效率提升至秒级，识别准确率也较小模型大幅提升，全面赋能客流统计、物品遗留、车辆外观检测等场景应用，提升城轨管理人员的事件决策能力。后续星云交通大模型将与城轨场景深度融合，探索运维、运营创新方案。

在铁路领域，中兴通讯聚焦铁路行车安全领域，基于AI技术精准赋能铁路关键场景需求，联合业内生态合作伙伴，对铁路沿线设备智能巡检、铁路周界安全防护等解决方案展开探索实践。使用交通大模型可以实现铁路将沿线电务漏缆脱落检出率提升至99%，提前扫除设备隐患，智能视频分析可主动监测、预防铁路线路入侵事件的发生，多手段立体保障铁路运营安全。

在港口领域，2023年10月起，面对太仓港危化品码头现场作业安全风险高、人工监控效率低、货物丢失和错误堆放等问题，中兴通讯依托星云交通大模型的分析能力，采用大小模型融合方式实现了人员管控、作业行为、车辆管理、查验检测四大类18种AI场景算法，并构建码头安全生产管理系统，为港口用户提供视频监控、风险告警、事件追溯等功能。系统上线以来及时发现了作业人员在危险货物边坐卧和倚靠、人员未触摸静电释放器直接进入生产区域、车辆违规穿越箱区、无关人员搭乘流动机械等多项安全风险，全面提升了码头安全管理能力。

在公路领域，中兴通讯积极参与广东、江苏、湖南等多地车路云一体化建设试点和智慧公路建设，并持续开展星云交通大模型在多维交通数据精准分析、交通态势实时感知、交通资源优化配置及路网智能预警与决策方面的探索，有效提升交通系统的安全性和运行效率，推动交通行业的智能化、网联化发展。

大模型作为交通行业数字化转型的重要驱动力，正引领整个行业向着更加美好的未来迈进。随着技术的不断进步和应用场景的不断拓展，星云交通大模型坚持开放解耦的发展方向，持续推动交通行业生态合作，支持多种硬件平台与多种模型的灵活整合，在确保交通数据安全和系统稳定的基础上，星云交通大模型将更好地满足交通行业的多样化需求，推动交通行业的数字化转型进程。可以预见，未来的交通系统将更加智能、高效、绿色，为人们的出行提供更加便捷、舒适的体验。ZTE中兴

星云大模型赋能油气行业高质量发展

戚晨
中兴通讯产业数字化规划总工

付光
中兴通讯行业大模型规划总工

油气行业是关乎国家能源安全的重要支柱产业，油气行业实现高质量发展的核心在于以人工智能、物联网、云计算等新一代信息技术为支撑，推动数字技术与油气核心业务深入结合，通过数字化、智能化赋能加快推动油气行业转型，积极培育和发展油气行业的战略性产业和未来产业。

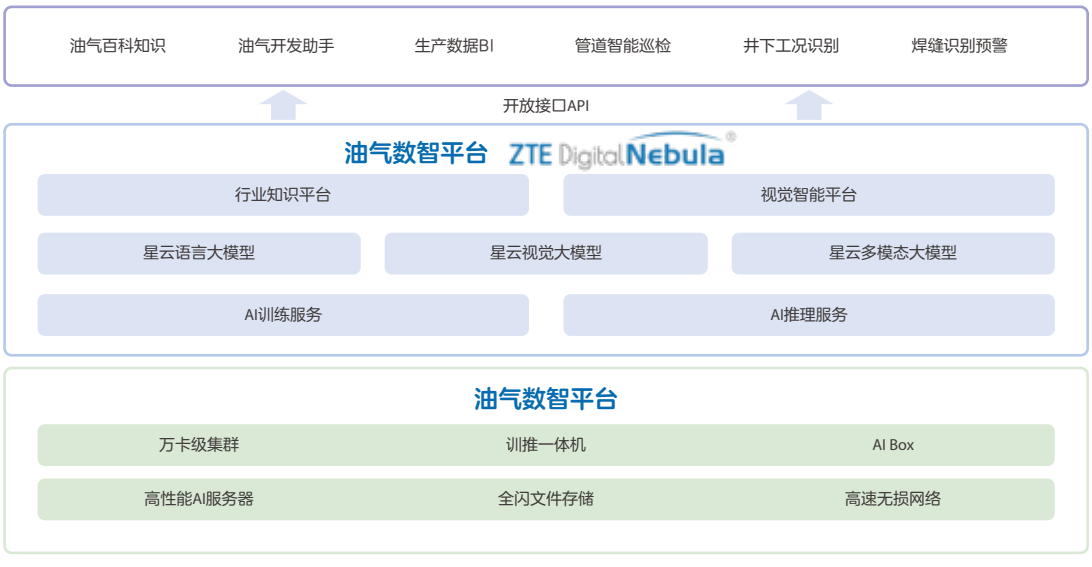
油气行业正在积极探索人工智能大模型技术应用，探索人工智能大模型在研发设计、中试验证、生产经营、安全环保、营销服务等方面的人工智能应用场景落地，加快科技创新成果向现实生产力转化，不断塑造高质量发展的新动能，以数字技术催生新的生产力。

星云大模型助力构建油气大模型

油气行业专业多，涉及的知识面庞大，同时业务流程复杂，场景众多，面向油气行业场景构建高质量、安全、高效的油气大模型成为挑战。

油气大模型的构建基于上百万份的石油石化百科知识、专业书籍、文献、标准、图片、视频，以及油气行业各类方案、数据库，通过对这些海量知识的挖掘利用、训练和精调，中兴通讯在星云大模型的基础上构建油气领域的行业大模型，精准把握油气领域的核心知识和规律，为用户提供一站式服务。

如图1所示，星云大模型包括语言大模型、



◀ 图1 星云大模型体系架构

视觉大模型、多模态大模型。

星云大模型按需提供全解耦的算力服务，满足大模型训练、训推混合、边缘训推等各种场景需求，提供大模型开发全流程工具和引擎，包括数据标注、模型训练、微调、模型编译、量化压缩和应用开发平台，助力油气行业业务集成商（SI）打通技术到应用的全链条。

星云大模型在油气行业的场景应用

油气大模型典型应用包括油气知识库、智能问数应用、方案报告快速生成、油气田作业监督等。油气大模型应用场景如图2所示。

基于RAG（检索增强生成）向量知识库构建的油气知识库，不仅可以对用户提出的问题进行解释，而且对结果进行精准溯源。油气行业标准规范众多，可以基于油气知识库进行标准规范内容的查询检索，并可以查看到回答的依据是来自于哪一篇标准规范，对现场问题给出针对性的建议。星云大模型不仅支持纯文本格式、办公文档格式语料文件，也支持带图片的PDF、图像文件格式；除了支持公式外，也支持统计图表，以及对复杂表格内容及语义的提取，更精准地实现模型训练和精调。星云大模型采用“基于定制语料的RAG”“基于定制组件的RAG”等多种技术手段，实现了知识问答准确率超过90%。基于RAG向量数据库，既保证了大模型回答的准确性，又保证了知识的实时更新。

基于智能体，油气大模型实现面向经营、生产数据的智能问数应用。利用大模型NL2SQL（自然语言转SQL）的能力，将问题解析并转化为SQL查询语句，通过智能体在后台关联的数据库进行智能查询，并将结果进行可视化展示。智能体自动读取表结构和内容，实现对结构化数据的处理。这样业务人员就可以用自然语言查询油井数量、每口井的详细信息和产量，也可以对生产日报中的数据进行精准问答，通过交互式问答快速了解油田每天的生产情况。油气行业术语、行话众多，星云大模型通过精调实现了油气行业术语或行话的自动识别和转换，提升交互效率和信息输出的准确率。

此外，星云大模型利用事先建立的指标和维度树，通过智能体自动拆解异常指标，实现自动预警、根因分析、关联分析（比如给出影响指标波动的最大贡献子指标）。星云大模型对原始数据进行治理，通过指标、维度对数据统一建模，要求按统一建模术语问数，实现“任意表随便问”效果，使得复杂查询准确率提升到95%。

星云大模型实现油气行业各类方案的快速生成，包括开发方案、安全方案等。星云大模型支持灵活的主题输入、大纲编辑、自动化图文内容生成、模版选择与输出。根据方案模板，大模型生成初版内容，并提供自动续写、总结、润色等功能，针对报告内容推荐知识库里的相关内容。

勘探	开发	工程	生产	储运	经营
地震数据处理与解释	井位优选推荐	钻井专家系统	采油专家系统	管道智慧运营	经营风控管理
测井数据处理与解释	油藏智能问答助手	井下工况识别	油井工况诊断与预测	管道智能巡检	经营问答助手
野外勘探方案助手	油藏开发方案助手	钻完井方案智能推荐	作业安全识别与预警	高后果区智能识别	会议纪要助手
勘探领域 百科知识问答	开发领域 百科知识问答	工程领域 百科知识问答	生产领域 百科知识问答	管道焊缝识别与预警	科研管理

图2 油气大模型应用场景 ▶

例如在勘探开发领域，在盆地评价的全过程中辅助研究人员分析构造、储层等情况，快速编写盆地评价报告。在编写报告中提供内容的自动生成，同时推荐和本盆地相关的图表、相关参数、相关文本。

星云大模型实现了油气行业报告智能解读，通过智能体对油气行业报告中的数据进行分析、解读，形成结论，对多个文档进行总结，可以根据多文档内容进行交互式问答，有效减轻管理人员和科研人员的日常工作负担。

油气田作业监督，基于多模态大模型对作业设计报告内容的理解和作业现场视频的自动分析，通过对比发现作业施工过程中的施工安全风险、质量问题和工艺流程问题。基于星云大模型提供对采油工人的作业培训，指导采油工人应用专业仪器，并在施工现场进行作业指导。

在大模型应用之前，油气行业积极探索小模型的应用已经初见成效，但在部分场景，需要通过大模型进行提升。比如对于火焰识别，在下雨水面反光或者闪电、汽车打灯，小模型会产生误报。通过大小模型协同，用星云大模型进行复核，可以快速提升识别准确率。

星云大模型部署

油气行业应用AI大模型的场景丰富多样，星云大模型支持多场景全开放的解耦部署。

在集团层面，可以选择Center版，部署在云数据中心。星云大模型提供全栈工具，帮助客户零门槛实现万亿级模型训练和开发。

对于二级油气田，存在数据安全不出域、既有推理又有训练的要求，可以选择Campus版。中兴通讯发布业界领先的训推一体机，提供全栈智算能力，一站式交付，开箱即用。

对于采油、勘探等作业现场，存在着多样化场景的部署需求，可以选择Site版，星云大模型提供多形态终端满足各种场景需求，通过云端协同提升运维效率，通过大小模型协同提升准确率。

星云油气大模型应用案例

中兴通讯为某油田建设了大模型平台，提供安全知识库和面向生产经营数据的智能问数能力。

安全知识库实现了多元化的问答场景，包括标准规范查询、案例参考、安全方案参考等。该项目将安全知识库、RAG技术和软提示词三者结合，显著提升问答系统的准确率，达到90%以上。根据作业任务要求，结合安全操作标准、规程、安全方案模版，星云大模型自动生成安全方案，包括文本、图片、摘要、总结等。

该项目实现了生产经营数据的智能问数服务。对于日报中的复杂问题和生产数据的细节查询，星云大模型能够迅速从海量的数据中检索出相关信息，并结合上下文生成精准、详细的答案。星云大模型能够理解并解析日报中的关键信息，如经营数据、市场趋势等，并为用户提供相应的分析和建议。在生产数据分析中，星云大模型通过智能体实时监控生产线的各项数据，并为用户提供实时的生产情况报告和异常预警。

经测试验证，基于星云大模型，科研资料采集时间由原来的1~2周缩短为1~2天，方案编制时间减少40%，新员工培养周期降低50%。

大模型在油气行业的应用正蓬勃发展，应用场景是基础，坚持以场景承载价值，技术的繁荣最终将体现为应用的繁荣。

中兴通讯坚持开放解耦原则，联合油气行业合作伙伴，通过软硬解耦、算网解耦、训推解耦、模型解耦，推动能力组件化和共享赋能，加速人工智能大模型场景的创新、研发、应用和商业化进程，构建开放的技术生态，加强产业生态协同研发与集成创新，推进数智技术与石油石化技术创新深度融合，培育一批综合性强、带动面广的数智化场景，推动油气开采、生产组织、运营管理、技术研发和商业模式创新。ZTE中兴

星云大模型，助力打造电力新质生产力

周承飞
中兴通讯电力市场总监

戚晨
中兴通讯产业数字化规划总工

饶晶
中兴通讯产业数字化方案部副总经理

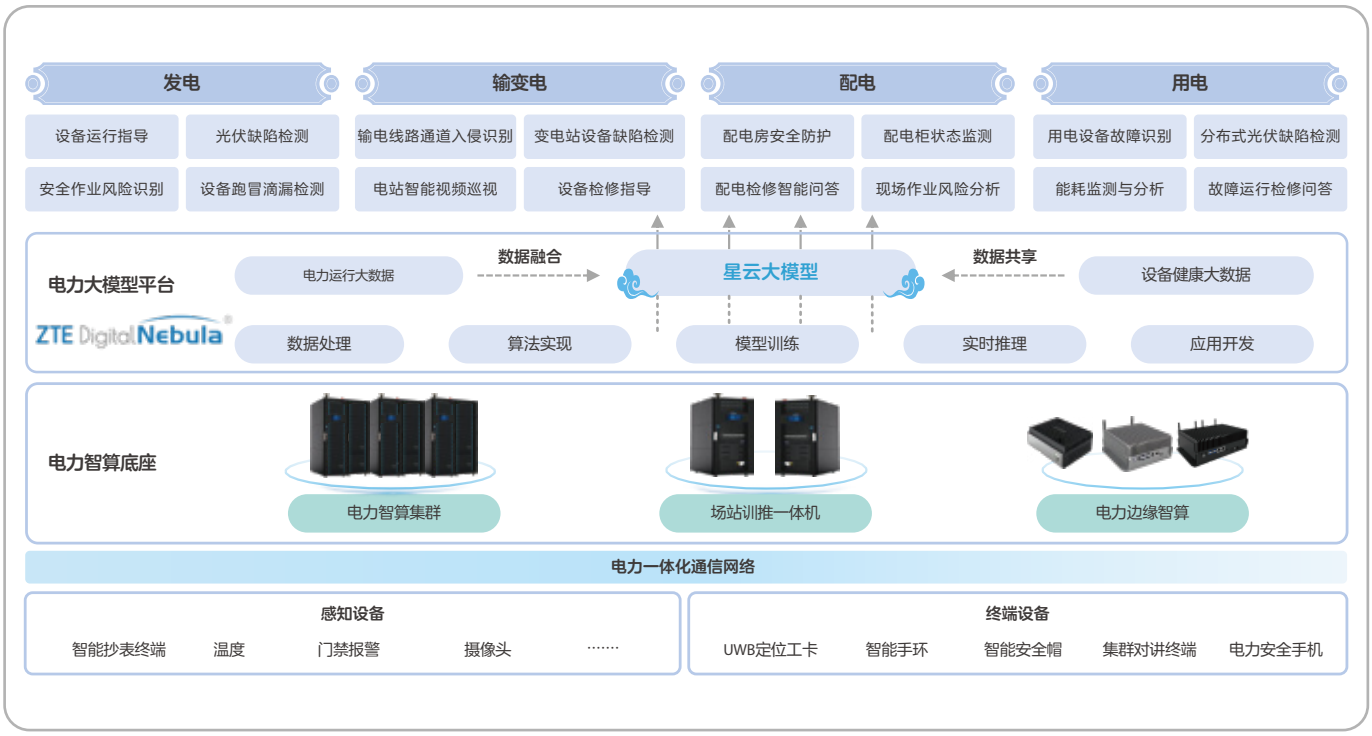
2023年国家能源局发布的《关于加快推进能源数字化智能化发展的若干意见》，提出探索人工智能及数字孪生在电网智能辅助决策和调控方面的应用，提升电力系统多能互补联合调度智能化水平。

2024年，国家电网公司提出以“大云物移智链”技术创新应用为驱动，打造数智化坚强电网，实现电力算力融合发展。南方电网公司发布电力“大瓦特”大模型，提出加快推进“人工智能+”专项行动，通过强化人工智能创新平台运营服务，有序推进电网人工智能场景建设、智能客服推广以及生产、调度等垂直领域的大模型建设。

随着AI人工智能技术的不断突破，AI技术正在发电、输电、变电、配电、用电各领域得到广泛应用。

星云大模型，构建电力数智化的坚强底座

中兴通讯积极参与国家电网、南方电网以及各发电集团等电力企业的人工智能大模型探索。2024年参与电网通信大模型工作，围绕电力



▲ 图1 中兴通讯电力大模型方案架构

网络通信和智能优化提供技术支撑，并参与技术白皮书编写；加入电力行业人工智能联盟，并联合10多家合作伙伴共同完成《电力行业人工智能产业报告》编写，从大模型训练平台和多智能体等多个方面，参与电力人工智能的规划设计。

面向复杂的电力场景需求，中兴通讯星云大模型具备软硬结合、一站式部署的能力，提供多场景、全开放解耦的解决方案，一方面围绕“发-输-变-配-用”的场景需求，构建了10多种电力典型场景的大模型应用能力，另一方面，面向不同的算力需求，提供智算集群、训推一体、边缘智算等场景化的部署方案（见图1）。

星云大模型，提升电力运行安全和运维水平

当前阶段，深度学习、强化学习等技术在电力各领域的应用已逐渐走向成熟，但随着新能源大规模并网，电力系统运行和管理面临高度不确定性、海量调度单元、多目标和多约束决策等重大技术挑战，以语言、视觉、多模态模型为主的人工智能大模型成为各大型电力企业关注的重点。

电力场站集中运维，风险高效处置

作为电能量生产和转换的关键枢纽，火电站、水电站、新能源电站、换流站、变电站等站点，位置分散，地处偏僻，系统众多，大量的机组运行、巡检、检修操作等依然依赖人工，设备风险、人身安全问题时有发生。现有的模型算法数据以单智能场景为主，普适性差，且不具备实时自我学习适应能力，应用效果不佳。

星云大模型通过对电力场站图片/视频进行深度分析，利用视觉大模型的一图多义，实现单图片/视频多场景识别，自动发现设备缺陷和异常状态，提高缺陷发现的及时率和准确率。同时可利用大模型对现有实时性要求较高的场景小模型进行结果复核，提升设备识别和故障预测的准确性，实现

对设备健康状态的全面监控和主动预警。同时与运维APP进行联动，实现隐患及时排查。

此外，基于3D数字孪生+大模型构建的智能巡视应用（见图2），可以自动选定巡检对象和角度，自动牵引摄像机资源，快速视频浏览设备对象，同时支持一键聚焦、自动巡视功能，升级传统单场站运维模式。通过大模型与数字孪生的融合可以同时对多个场站进行远程巡视，替代传统的人工例行巡视，使得电站巡检成本降低70%以上，巡检效率提升200%。

输电线路巡检，事故提前预警

在输电领域，业务场景复杂，我国在运架空线路长度超160多万公里，国家电网、南方电网拥有超350万座输电铁塔，70%以上的输电通道巡检主要依赖人工，巡检强度大，巡检质量受到作业人员主观因素影响。电力公司过去几年尝试使用AI技术来辅助运维，但电力场景设备多、测点广，现有模型算法无法做到全覆盖，且识别效果因场景变化和环境影响，经常出现误报、漏报等问题。

星云视觉大模型与电力无人机应用结合，将无人机拍摄的可见光、热成像图像通过大模型进行训练和推理，可以识别杆塔、线路设备缺陷，如鸟巢、异物、螺栓丢失、塔材锈蚀、绝缘子缺失、销钉退出等，还可以进行线路通道分析，识别线路所在地的地质灾害、道路塌方、机械施工、违章建筑等风险，且自动与故障定位、检修平台进行联动，辅助现场人员进行设备检修；相对于原来的视觉小模型算法，准确率达96%以上，测点数量增加80%多，一次部署，实现长期自我学习和进化。

电力设备运维助手，操作简单易懂

电力设备专业度高、结构复杂、安全隐患多，巡检和检修人员要求具备很高的电力设备知识、操作知识、安全知识等，而实际现场中的检修人员能力参差不齐，对电力安全意识不足。

基于星云大模型多模态文档分析、智能搜索等能力自动生成电力AI知识库，通过RAG（检索增强生成）技术与电力本地知识库相结合，对发电机组、逆变器、光伏、变压器等电力设备及其运行情况进行分析研判，可针对现场检修人员的问题进行推理、回答，并自动生成检修指导报告。同时利用多智能体系统，可以给出设备故障消缺和检修指导，将故障诊断任务分配给各个智能体进行并行处理，通过智能体间的信息交换，可以更快速地诊断电力设备故障，将风险和隐患关口前

▲图2 基于大模型的三维视频智能巡检系统

> 未来，星云大模型在电力的“发-输-变-配-用”全环节都将发挥关键作用，助力电力数字化新质生产力的提升。

移，实现“事前预警预防”，并通过提示词嵌入等手段，降低诊断的错误率，准确率普遍提升15%以上。电力AI知识库可应用于电力设备检修指导、新员工培训、自动生成故障分析报告、故障诊断与预测性维护等。

电力作业安全管控应用，辅助指挥决策

安全是电力作业中的重中之重。电力作业和检修过程中易造成人身安全事故，如何确保电力生产、运维等各个环节的安全作业，是管理者和安监部门面临的难点。

通过星云大模型，一方面对历史事故数据、设备运行状态数据等进行深度学习分析，对作业文件进行分析，识别作业关键工序及涉及的作业风险点，结合电力公司作业风险评估模型（基准+场景+动态）形成“规避人的主观因素影响”的风险评估结果；另一方面可以构建作业安全风险评估模型，基于作业文件作为输入，结合电力公司模型规范、气象数据和现场人员数据，自动评估风险等级，替代当前单纯人工判断风险等级的主观问题。此外，综合利用视觉算法实现现场作业人员视频监控及作业合规识别分析，对作业人员进行合规着装分析（安全帽、工作服）、行为分析（抽烟、跌倒等），以及安全作业分析（登高作业安全带使用、吊装下不准站人）。以此形成综合安全管控一张图，在突发事件发生时，大模型能够分析，并提供科学的应急指挥调度方案，确保应急处置工作的迅速、有序进行，实现“作业现场可视、风险可控”。

星云大模型在电力行业的实践探索

中兴通讯积极将大模型技术用于电力行业，通过深度挖掘电力数据价值，推动电力系统的智能化升级。目前在发电、输电、变电、配电、用电等电力领域积极探索和实践。

在发电领域，携手华能集团，在云南某水电站，将5G、物联网、AI等技术与发电厂巡检业务进行融合创新，通过视觉大模型，将电厂视频监控、移动布控球、机器人等终端进行连通，结合智能视频巡视应用，用于发电厂的管道跑冒滴漏、危险区域作业监控，人员安全行为管控等场景，解决电力生产中“防、监、维、控”的痛点，打通发电站日常巡检运维的堵点，最终实现远程站房的无人值守，巡检效率提升2倍。

在配用电领域，携手国电投江苏电力，在江苏南京滨江完成21.68MW的分布式光伏示范基地的建设，面向园区分布式光伏、储能、用电领域的日常运维需求，我们将多模态大模型用于光伏组件缺陷检测、储能站点安全防护、能耗分析和精准调控，并在园区25个配电室进行基于多模态大模型的配电房无人值守改造，将日常有人例行巡检转变为全天候无人值守，减少巡检人员投入50%。

随着星云大模型在电力行业的不断探索和应用，能力也在不断增强，正迅速构建崭新的电力运维和安全管理模式。未来，星云大模型在电力的“发-输-变-配-用”全环节都将发挥关键作用，助力电力数字化新质生产力的提升。ZTE中兴

星云大模型服务钢铁行业生产和运营协同创新

叶郁文
中兴通讯冶钢矿山规划总工

李阳
中兴通讯冶钢市场总监

张滨
中兴通讯产业数字化方案部副总经理

2024年初，国资委召开中央企业人工智能专题推进会，指出中央企业要发挥需求规模大、产业配套全、应用场景多的优势，带头抢抓人工智能赋能传统产业，加快构建数据驱动、人机协同、跨界融合、共创分享的智能经济形态。同时，工业和信息化部等九部门印发的《原材料工业数字化转型工作方案（2024—2026年）》中指出，强化人工智能驱动，推动将成熟人工智能技术引入生产调度优化、过程模拟仿真、运营管理决策、安全管控等典型场景，建设适用于生成式人工智能的行业数据集，基于现有通用大模型技术底座进行定制化开发训练，构建细分行业大模型，加快大模型技术深度创新。这些指导政策均给出了明确指向，要求传统行业利用信息科技技术创新，面向高端化、智能化、绿色化发展目标，积极推动传统产业向新型工业化转型升级。

智改数转，钢铁行业面临诸多挑战

钢铁行业作为国民经济的重要基础产业，具有生产流程长、过程机理复杂、业务场景丰富、工况环境严苛、无人化少人化需求迫切等突出特点。近年来，钢铁行业数字化转型不断走向纵深，但在生产和运营方面仍然面临诸多挑战。

在集团管理方面，钢铁企业很多为一总部多基地协同模式，为企业的集团管控和信息协同带来较大困难。例如，集团和分支机构现存大量管理制度，涵盖党建、财务、档案、法务、人力、信息化、科技等维度，管理相关制度查询和利用、文档编制和归档、信息校验等工作耗时耗力，制约管理工作效率和准确度提升。

在生产制造方面，钢铁生产涉及从原料到成品的多个高度集成的工序，很多工序环节存在“黑箱”问题，过程机理高度复杂难以掌握；同时，钢铁生产工况条件经常发生变化，传统基于机理模型或人工经验进行生产决策对生产过程的连续稳定和精确控制带来极大挑战；此外，在钢铁行业，“3D”即Dangerous（风险大）、Dusty（环境脏）、Duplicate（重复劳动）岗位带来的员工安全和健康挑战问题，对现场的少人化、无人化提出了迫切需求。

在质量控制方面，钢铁车间积累了大量质量相关数据，数据量大、类型多、结构复杂，包括结构化数据（如生产参数）和非结构化数据（如设备日志、图像视频、作业规范、质量检测），处理难度高，如何快速查询和分析质量数据，快速精准地追溯到问题源头，为质量控制提供决策优化建议成为钢铁企业提升质量的重要目标。

在设备运维方面，维护经验、知识碎片化，缺乏系统化的积累、提炼和优化，设备维护相关数据没有得到有效开发和应用；同时，设备运维过度依赖人的行为和经验，设备故障多，运维效率较低，且设备过维修和欠维修长期共存，使得

综合维护成本居高不下。

在安全生产方面，安全管理制度、法律法规涉及内容多、系统性强，需要完善的制度和有效的执行机制，同时，生产流程复杂，安全隐患多，安全隐患难以及时发现、预警和处理，导致安全生产事故频发。

星云大模型助力钢铁行业生产和运营协同创新实践

中兴通讯基于星云大模型打造了星云钢铁大模型（见图1），支持语言、图像、视频、声音等多种模态，应用到钢企的实际场景中，可以由浅入深，高价值场景先行先试，多场景分类探索，创新技术逐步应用，人才队伍不断壮大。

针对钢铁行业集团及分支机构的制度管理、运维知识问答、安全生产法律法规检索和知识库构建等场景需求，知识平台基于基础大模型语义理解能力，通过用户专有数据构建向量化知识库，并结合检索增强生成技术（RAG），匹配正确知识，降低模型幻觉，最终实现知识问答准确率提高到90%以上；通过业界首创的视觉方式提取图片中的表头信息，包含语义信息和结构信息，准确理解表格结构和表格内容，准确率达到98%以上；同时，通过精准溯源功能，确保每个回答都可追踪至具体知识源头，极大提升了信息的可信度和可用度。

针对钢铁企业生产经营报表查询和质量快速统计分析场景，星云大模型-星云ChatBI提供了智能问数能力，用户仅需使用自然语言提问，大模型即可快速、准确地理解用户问题，从海量数据中检索、分析，提供精准的信息解答，并以适当的图表形式进行展示。最终实现数据智能分析查询、数据分析快速呈现、数据根因分析等功能，帮助企业快速获取生产现状问题及应对策略，提升经营决策效率。

基于星云大模型智能体，利用事先建立的指标和维度树，对钢铁企业经营报表和生产统计原始数据，通过指标、维度对数据统一建模和治理，利用大模型自然语言转SQL的能力，将业务人员用自然语言的问题解析并转化SQL查询语句，实现“任意表随便问”效果，通过智能体在

▲ 图1　中兴通讯星云钢铁大模型方案

后台关联的钢铁生产统计数据库进行智能查询，可自动拆解异常指标，实现面向经营、生产数据的自动预警、根因分析、关联分析等智能问数应用，使得复杂查询准确率提升到95%。

针对钢铁园区生产安全场景，星云视觉大模型可精准识别安全带、安全帽、反光马甲、工作服、CO报警仪等穿戴，以及钢水车、火焰、烟雾、抽烟等潜在的安全隐患，识别准确率大于99%。此外，平台支持大小模型协同，小模型初筛并通过大模型复核和补漏，大幅降低小模型算法误检和漏检问题，提高识别精度，保护企业小模型投资。

相比钢铁行业的知识问答、智能问数和园区安全场景，面向钢铁生产过程的智能决策等专用大模型对于钢企而言更为重要。钢铁生产全流程包括选矿、炼铁、炼钢、精练、连铸、轧制等几个环节，其共性核心设备包括高炉、转炉、电炉、精炼炉、连铸机、轧机、热风炉和除尘设备等，以转炉为例，转炉炼钢过程是一个连续非线性系统，具有过程扰动、时滞和工艺约束，并遵从反应动力学、流体动力学和质能守恒等基本机理模型。对于生产过程中存在的“五难”，即不完备信息建模难（反应机理复杂）、多源异构数据认知难（高温环境复杂）、多冲突约束决策难（生产要求复杂）、工业知识积累学习难（动态过程复杂）、多样生产平台构建难（应用场景复杂），传统机理模型并不能有效地进行质量预测和工序决策。

针对钢铁生产制造优化决策，采取“机理建模+数据驱动+知识引导”的先进方法，分为数据感知阶段、认知选模阶段、智能决策阶段。数据感知阶段实现对现场声音、震动、图像、视频等传感数据的智能感知，将钢铁生产冶炼过程多模态数据中的低质量部分，进行特定的补齐、选帧、校准等处理，从而提高输入智能决策大模型的数据质量水平；认知选模阶段则实现多模态传感数据结合机理模型和知识图谱进行数据建模，负责根据当前感知的数据确定钢铁生产冶炼工况，并调用特定工况对应模型；智能决策阶段构建生产优化决策大模型，根据现场生产工艺条件输出控制决策参数，生成决策信号，给到现场生产设备，形成基于数据流动的状态感知、实时分析、科学决策、精准执行的闭环，降低人工干预，提升生产稳定性。

星云钢铁大模型支持钢铁行业向新型工业化转型升级

星云钢铁大模型已在多家钢铁企业进行了实践和探索。在鞍钢集团，通过星云语言大模型，针对鞍钢集团及分支机构制度文档创建私域知识库，并“链接”大模型的学习能力和生成能力，实现文档的检索生成和智能问答，提升鞍钢整体办公和经营管理效率。此外，中兴通讯积极探索大模型技术与钢铁行业场景的深度融合，在2023年底第四届钢铁工业智能制造发展大会上，中兴通讯携手中国工业互联网研究院、冶金工业信息标准研究院、鞍山钢铁集团有限公司等10余家行业伙伴，共同发布《钢铁行业工业互联网大模型场景应用白皮书》，从钢铁行业大模型涉及的算力基础设施、数据和语料、大模型平台、模型训练和应用安全几个角度阐述整体技术架构和实施路径，并参考工业互联网新兴模式，从平台化设计、智能化制造、网络化协同、服务化延伸及数字化管理等方面，全面梳理了潜在的场景应用。

未来，随着钢铁行业绿色低碳发展和供给侧改革的推动，钢铁行业大模型的应用价值将日趋显著，随着大模型等人工智能技术和场景的深度融合，将为钢铁行业研发设计、供应链、生产制造、运营管理等多方面带来巨大效益。中兴通讯也将积极携手行业龙头、专业高校等伙伴，发挥自身优势，持续探索实践，切实助力钢铁行业由传统产业向新型工业化转型升级，夯实钢铁行业高质量发展根基，助力钢铁行业高端化、智能化、绿色化发展，推动我国从钢铁大国向钢铁强国跃升。ZTE中兴

星云大模型端到端安全防护及创新

许晨敏
中兴通讯网络安全产品战略规划师

王继刚
中兴通讯网络安全产品总经理

陈靖
中兴通讯网络安全产品规划总工

随着人工智能技术的迅猛发展，大模型技术作为数智化时代的核心驱动力，正逐步渗透到各行各业，成为推动行业变革与社会进步的重要力量。然而，随着大模型技术的广泛应用，其面临的安全风险也日益凸显。为了确保大模型能够在更加安全、可靠的环境中健康发展，中兴通讯通过构建完善的大模型安全防护体系，加强技术监管与合规审查，为数智化时代保驾护航。

大模型发展面临的安全风险

当前，大模型正处于蓬勃发展阶段，针对大模型的攻击层出不穷，大模型在数据处理、训推部署、业务运营等各个阶段面临着多种安全风险。

- 数据安全风险

大模型训练依赖海量数据，数据的安全性直接关乎模型的准确性与可靠性。由于潜在数据安全风险导致数据供给方不愿开放用于训练的数据，降低了模型效果。在数据采集、数据预处理、数据标注等过程中存在敏感信息泄露、训练数据投毒、私有数据出域、训练数据集与模型参数泄露等风险。此外，数据在流通过程中的安全风险也不容忽视。

- 模型安全风险

大模型的复杂性和高维性使得其算法可能存在漏洞，攻击者可能利用漏洞、逆向工程、对抗攻击等技术手段对大模型进行模型盗窃、模型篡改、后门攻击等，严重影响大模型训练和推理的输出结果。甚至智算基础设施也可能遭受直接攻击，导致模型无法正常运行或数据丢失。

- 模型应用与内容合规风险

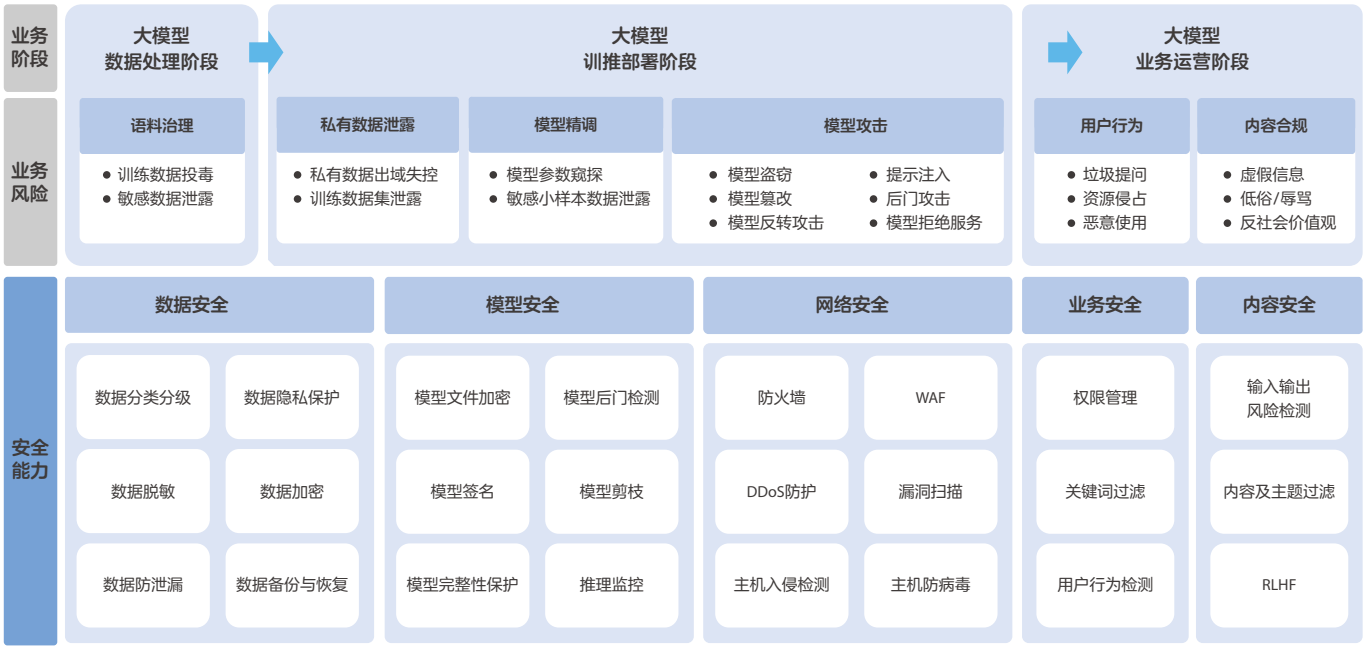
大模型的输出内容存在安全合规风险，攻击者可能利用大模型生成虚假或有害信息误导或诱骗公众，利用大模型生成恶意代码或软件对政府、军事等机构发起自动化攻击，严重危害社会安全。

中兴通讯星云大模型端到端安全防护方案

为了保障星云大模型的安全性和可靠性，中兴通讯推出星云大模型端到端安全防护方案（见图1），通过自主研发和技术创新，从数据安全、模型安全、网络安全、业务安全、内容安全等多个维度构建大模型安全防护体系，为星云大模型提供端到端全流程安全保障，推动大模型技术快速发展。

- 保障数据安全

星云大模型采用严格的数据保护措施，确保用户数据的安全性和隐私性。数据安全是大模型安全防护体系的核心，在采集海量原始语料数据时，通过数据分类分级技术自动发现并动态更新数据资产，支持根据数据的重要性与敏感程度实现自动分类分级管理。结合数据脱敏、数据防泄漏、数据隐私保护等技术研发自动化工具，针对不同级别的语料数据采取相应的安全防护措施，支持识别并清除有毒数据、无关信息、企业或个人标识数据，帮助用户解决训练数据投毒、私有数据出域，敏感信息、训练数据集与模型参数泄露等问题，为用户构建高质量训练与精调数据



▲ 图1 中兴通讯星云大模型端到端安全防护方案

集。同时，通过建立有效的数据恢复机制，定期进行数据备份，防止用户数据丢失或损坏。此外，通过对用户敏感数据进行加密存储和传输，防止数据在流通过程中被窃取或篡改，充分保障星云大模型用户的数据安全。

- 保障模型安全

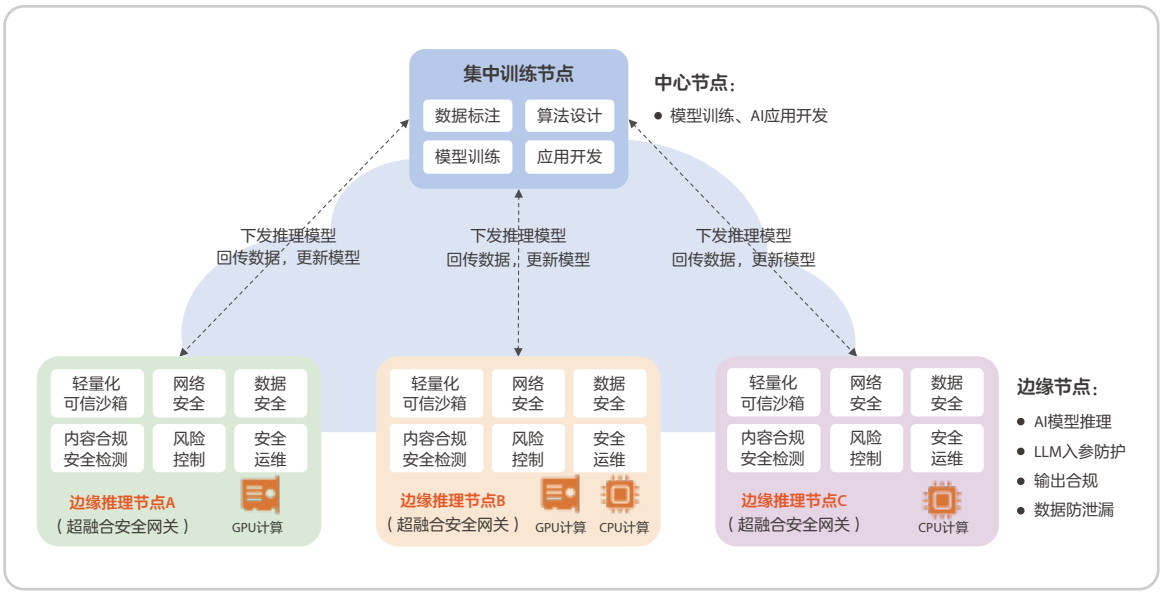
模型文件是大模型训推过程中的核心资产，中兴通讯采用模型文件加密、模型签名和完整性检测技术，防止攻击者在训推过程中注入有害代码、窃取或篡改模型，确保模型的完整性和可信度。此外，中兴通讯通过实时监控模型的推理过程，及时发现异常行为或潜在攻击，利用检查激活的方式进行模型后门检测，支持通过微调、模型剪枝等方式消除模型后门。

在网络层面，通过部署防火墙、WAF、DDoS防护、漏洞扫描、主机入侵检测、主机防病毒等安全服务，实现网络隔离与访问控制，防止远程代码执行、侧信道攻击和恶意流量入侵等黑客攻击。另外，在边缘侧通过部署超融合安全网关，在基础设施中内置安全能力，为用户呈现大模型全生命周期网络安全态势，支持及时发现并处置告警事件，保障模型与智算基础设施安全。

- 保障模型应用与内容安全合规

星云大模型支持为用户提供完善的权限管理功能，确保只有授权的用户才能访问和使用模型提供的服务和应用。针对部分授权用户恶意利用大模型的行为，通过关键词过滤、用户行为检测等方式，识别恶意prompt，确保模型应用和业务流程的合法性和规范性。

相比于传统安全，大模型安全更加强调AIGC（人工智能生成内容）的安全合规。中兴通讯采用基于深度匹配的输入输出风险检测技术，通过预置关键词库实现非法内容的拦截，支持有效拦截违法违规、色情、反动等内容，同时，通过建立内容及主题过滤机制，能够有效识别并避免侵犯用户的著作权、商业秘密等知识产权。此外，中兴通讯采用RLHF（基于人类反馈的强化学习）技术，通过制定安全合规基线、人工标记比较、设置奖励函数等方式实现人机对齐，有效解决



◀图2 中兴通讯行业模型安全训推应用场景

AIGC输出安全和价值观问题，对可能引起民族、信仰、性别、地域等多维度敏感属性歧视内容的正确识别率达99.5%，证明星云大模型在确保多元文化、平等尊重及非歧视原则方面具有卓越性能，支持为交通、水利、油气、电力、冶钢等各个行业大模型提供内容安全合规保障。

当前，大模型发展与算力资源紧缺成为行业用户面临的一大难题，中心训练+边缘推理成为行业模型训推的主要应用场景之一。中兴通讯支持为行业用户构建协同一致的云网安数基础设施（见图2），通过在边缘推理节点上部署超融合安全网关，融合安全、算力、数据与AI能力，支持轻量化部署推理模型与AI应用，通过内置20多种安全组件，提供全栈安全防护能力，保障精调数据与业务模型安全，支持为边缘推理节点提供轻量化可信沙箱、网络安全、数据安全、风险控制、内容合规安全检测，以及统一安全运维管理能力，支持安全能力按需编排调度，实时处理告警，减少运维响应时间，安全响应能力提升16.5%，降低企业边缘侧算力需求与投资成本，打造云网安数融合能力底座，保障行业模型安全训推。

中兴通讯通过建设星云大模型端到端安全防护方案，为星云大模型提供全流程安全防护，支持监控、评估和应对大模型面临的各种安全威胁，保障星云大模型完成了国家网信办生成式人工智能服务备案。中兴通讯通过构建大模型安全防护体系，助力大模型安全赋能千行百业。

未来展望

随着大模型技术的持续演进与创新，越来越多的行业希望借助大模型来实现业务升级、效率提升和创新驱动。在智能制造、智能交通、智能医疗等领域，云边协同架构能够实现资源与服务的高效整合与协同优化。在云边协同架构下，边缘用户更加关注如何保障敏感数据的安全。因此，中兴通讯正积极探索在星云大模型的精调与推理阶段融入差分隐私技术，通过在边缘用户的明文数据上添加噪声或进行模糊处理等方式，大幅降低敏感数据泄露的风险，确保在大模型精调与推理过程中边缘用户敏感数据的隐私与安全。

未来，随着大模型在安全领域的标准与规范逐渐建立和完善，大模型的安全防护能力无疑将同步得到提升。中兴通讯将利用大模型技术提升威胁预测与智能化防御水平，积极应对未知安全风险，构建支持自我完善的安全防护体系，保障大模型与AI应用的运行安全，推动人工智能技术健康有序地发展，为数智化转型提供安全保障。ZTE中兴

“数据要素×AI”
——数据基础设施助力大模型高质量发展

王继刚
中兴通讯网络安全产品总经理

陈靖
中兴通讯网络安全产品规划总工

刘丰
中兴通讯数字技术产品部AI专家

大模型面临的数据供给挑战

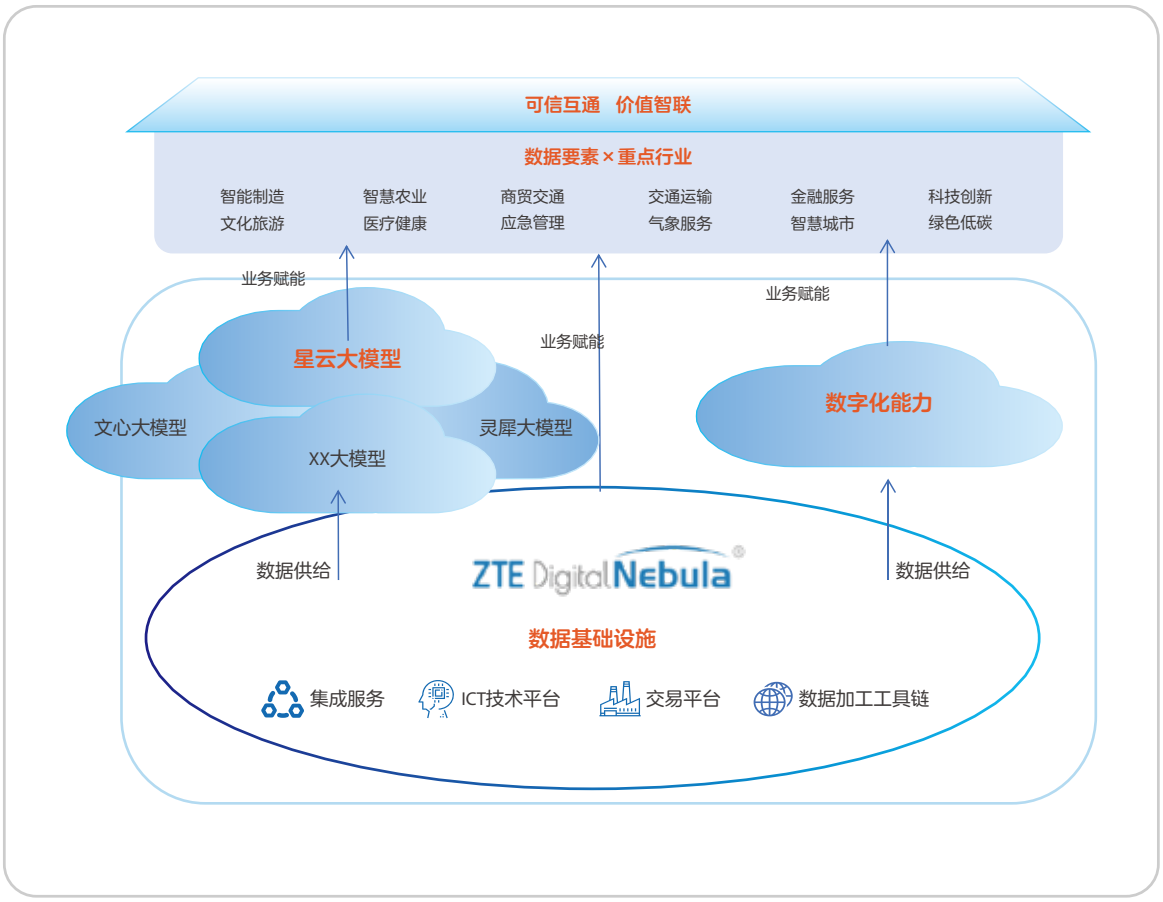
人工智能大模型的发展，离不开高质量的语料数据集，如何获得高质量数据集、保障训练数据集安全采集、充分发挥数据价值，已成为大模型发展不可避免的问题与挑战。

- 挑战一：公开数据量有限。互联网上虽然存在大量文本数据，但其中很多都是低质量的，如垃圾信息、广告宣传等。而且公开数据集只能解决通识问题，细分行业的专业性问题，公开数据无法提供参考。
- 挑战二：行业数据壁垒高。对于一些垂直领域，如科技、医疗、金融等，数据往往涉及商业机密或隐私信息，很难对外共享。例如在自动驾驶领域，出于商业秘密保护，各企业独立进行道路数据采集，很少进行数据共享。这不仅导致大量重复性工作，降低了自动驾驶算法研究的整体效率，同时每个企业采集的数据在路况、天气等方面都有局限性，无法做到更广泛情形的覆盖。
- 挑战三：数据采集成本高。高质量数据往往需要经过采集、标注和清洗才能使用，这需要投入大量的人力和物力。提高大模型开发过程中数据供给、生成高价值训练数据集成为大模型发展的迫切需求。

国家数据局等17部门联合印发的《“数据要素×”三年行动计划（2024—2026年）》，进一步明确“建设高质量语料库和基础科学数据集，支持开展人工智能大模型开发和训练”。通过数据要素建设推动人工智能大模型发展，可以有效解决人工智能，特别是大模型研发所面临的数据瓶颈，进一步发挥大模型对于知识数据的汇集和处理能力，创造更大的生产力，助力数字经济新发展模式。

构建数据基础设施，提升大模型的数据供给规模和质量

数据基础设施是从数据要素充分流通并释放价值的角度出发，在网络、算力等设施的支持下，提供一体化数据汇聚、处理、流通、应用、运营、安全保障服务的一类新型基础设施，是覆盖硬件、软件、标准规范、机制设计等在内的有机整体（见图1）。中兴通讯的数据基础设施通过隐私计算、区块链、数据脱敏、数据空间等技术，实现数据在不同主体间“可用不可见”“可控可计量”，为不同行业、不同地区、不同机构之间的数据实现合规高效流通，整体推动数据服务千行百业、深度融入社会生产生活，推动数据要素“供得出、流得动、用得好”，有效提升数据流通环节的安全可靠水平。



◀图1 数据基础设施赋能大模型

在“数据要素×AI”场景下，通过数据基础设施提供的智能分类分级、数据脱敏、隐私计算、区块链以及一站式数据处理等数据集构建能力模块，打造适配各种大模型的数据供给方案，实现了大模型训练语料数据的可信收集，全生命周期数据安全可控，消除了高价值数据拥有方的供给顾虑。核心能力组件包括：智能分类分级、数据脱敏、隐私计算、一站式数据处理。

- 智能分类分级

支持各数据源数据资产的自动发现与动态更新，多数据源、多类型数据的自动分类分级等。中兴通讯数据基础设施的数据分类分级工具内置了政务、交通、工业等10余项行业数据分类分级标准，在智能分类分级过程中更加贴合行业属性。

- 数据脱敏

通过数据脱敏机制对语料中包含的敏感信息应用脱敏规则进行数据的变形，实现敏感数据的可靠保护，在不影响数据分析对数据要求的条件下，对数据进行改造并提供使用。中兴通讯数据基础设施的脱敏组件可支持50种以上的数据脱敏算子，满足各种不同行业数据使用场景下的数据脱敏需求，以便在数据标注、预处理等非生产环境以及外包环境中，可以安全地使用脱敏后的真实数据集。

- 隐私计算

在多个大模型研发参与单位，分别部署隐私计算计算节点进行本地数据的读取，基于同态加密、秘密分享、不经意传输、零知识证明等密码学算法，在多方原始数据不出本地数据库的前提下，完成多方联合训练任务。能够实现大型企业在研发大模型语料收集阶段的分布式安全采集，既保证了大模型语料的质量，也保障了各研发单

▲ 图2 一站式训练数据处理流程

位原始数据的安全性。中兴通讯隐私计算和区块链平台是首家通过信通院“代码自研率”“国产环境兼容性”测试的企业。

- 一站式数据处理

多源异构原始语料管理、数据集管理、数据标注和预处理等功能，形成高质量数据集（见图2）。

首先明确数据集的数据类型、数据量和数据质量要求；根据数据集目标确定数据采集策略，确定数据采集的来源、方法和频率；采集到的原始数据往往存在噪声、缺失值和异常值等问题，采用低质站点过滤、异常符号过滤、脏数据过滤、文档级去重、句子级去重等手段开展预处理；通过敏感过滤、分类精炼等完成数据集的进一步加工，同时通过算法打分、人工打分的方式持续优化数据质量，并进为每个数据样本添加正确的标签或类别，最后根据模型训练需求匹配合适的高质量数据集。对数据集进行持续更新和维护，以确保数据集的时效性和准确性。

一站式数据处理组件可高效构建领域数据集与精调数据集，将无序的原始训练数据整合成高质量结构化的数据资产，建立数据画像，丰富数据标签，研究模型专业化流程，在增量预训练、精调训练、模型优化和部署、监控数据、回归测试与模型评估过程中，提高数据的可访问性和可复用率。

- 水印添加和去除

在每份共享的数据集流转各环节上加入水印标识，一旦发生泄漏，即可根据水印隐藏的信息确定泄漏源，快速定位责任人。同时，中兴通讯数据基础设施的水印组件在收集数据资产的过程中，当遇到一些影响文本提取效果的水印存在时，可以采用去水印技术去除水印，提升数据提取的质量。

- 基于知识图谱和大模型的知识管理

利用自研的大模型将每份数据资产分解为各个数据子模块，然后利用知识图谱技术对这些子模块进行管理，形成文档树，并将文档树背后的业务信息进行关联，最终将知识图谱中的结构化信息输入到大模型进行训练。

数据基础设施为星云大模型积累大规模、高质量、多元化语料，精炼后超3.5万亿Tokens；数据结构化，建立数据画像，丰富数据标签，数据内容高度可控；利用知识图谱形成文档树并与业务信息关联，实现高效的知识消费。

数据、算法和算力是构建AI系统的三大核心要素，三者的协同使现代AI技术实现了从理论到应用的飞跃。数据是AI的基础，大规模高质量的数据不仅能提高现有模型的准确率，还能促进模型的优化和创新。未来，中兴通讯将持续加大在大模型和数据基础设施领域的研发投入，依托数据基础设施，为各类行业大模型提供包括原始数据集、定制数据集和配套产品工具等在内的整套数据加工服务。ZTE中兴

引领工业智能发展：国内首个五星5G工厂智能进阶实践

中兴通讯南京滨江工厂（以下简称“滨江工厂”）是中兴通讯在国内的五大生产基地之一，主要生产5G、服务器等最新的无线通信和算力设备。滨江工厂自建设之初就明确顶层设计，规划在整个制造基地实现全流程数字化和智能化改造，分为数字滨江（2019—2020）、5G滨江（2021—2023）、智能滨江（2023—2026）三个阶段来建设。

截止目前，滨江工厂已经完成数字滨江和5G滨江两个阶段的建设目标，并于2024年8月成为国内首个通过五星5G工厂认证的企业。引领行业创新，打造行业标杆，当前滨江工厂正大步向智能滨江进阶。

星云工业大模型创建工业大脑

面向新型工业化可持续发展，中兴通讯打造了以5G+PON工业现场网为底座、数字星云平台为核心、星云行业大模型为智能化推手的数字化转型整体解决方案，并在滨江工厂躬身入局，持续实践。

如果将滨江工厂看成一个智能仿生体，在数字滨江阶段，已经完成广泛的设备连接和数据采集，每天产生超100TB的高质量数据，具备全面感知能力；在5G滨江阶段，已经完成高速的神经、锐利的眼睛、灵巧的双手、勤快的双腿的智能应用实践，如视觉质检、条码识别、AGV智能调度、园区安全巡检等。这些智能化应用属于中阶的智能应用，还未有模拟人类大脑的思维推理能力。

随着AI大模型的加持，人工智能的理解、规划、决策能力得到增强，具备了模拟大脑的能力，因此对于涉及工厂生产全流程的计划排产、资源智能调度等智能决策场景，以及生产域深水区的工艺优化设计、产品智能维修、具身智能柔性执行等复杂场景，也已具备了实践的条件。

从2023年开始，中兴通讯基于自研的星云大模型开发了工业大模型（见图1），将大模型、AI、运筹优化等智能化技术广泛应用于生产全流程中，构建滨江工厂的智能决策与精准执行系统，通过实时感知和洞察生产现场的海量数据，实现数据的高效流转和智能分析，形成科学决策与精准执行的生产全流程闭环，实现数据流的智能自动流转和实物流的精准高效流转。

基于星云工业大模型创建的各种应用贯穿生产各环节，使得滨江工厂的生产：

- 更快速——供货周期下降39%；
- 更高效——人均产值提升74%；
- 更绿色——生产能耗下降27%；
- 更精准——客户订单承诺偏差率下降70.4%。

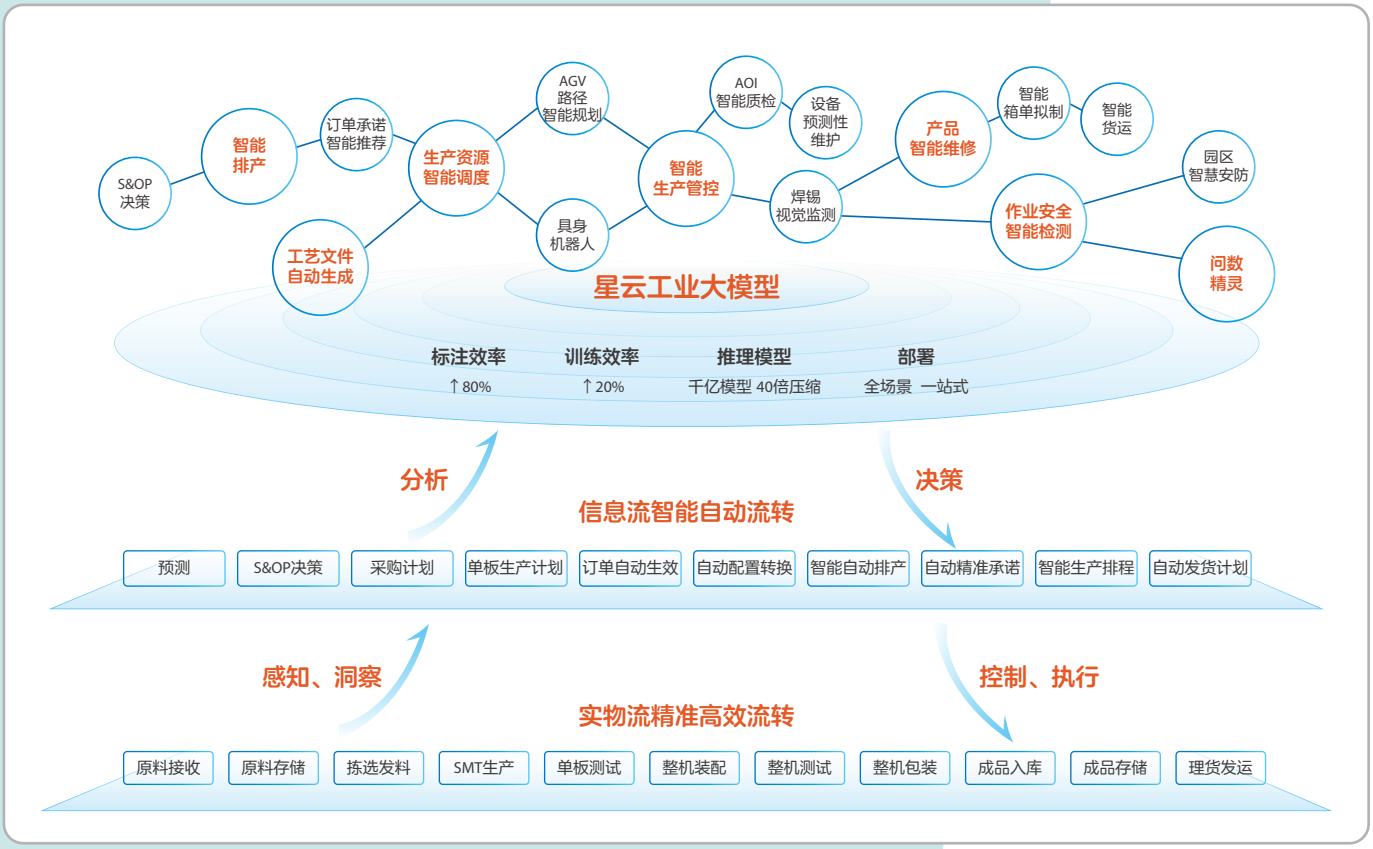
星云工业大模型在滨江工厂的实践应用

以星云工业大模型为基础的人工智能深入计划排产、产线资源调度、工艺设计、智能维修等

孟晓斌
中兴通讯产业数字化规划总工

陆平
移动网络和移动多媒体技术国家重点实验室副主任、中兴通讯产业数字化方案部总经理

耿兴元
中兴通讯产业数字化高级战略规划师



▲ 图1 工业大脑整体架构

工厂深水区场景的实践，相比传统模式带来了自动化程度高、准确率高、边际成本低、业务上线快等优势。

智能排产

滨江工厂作为生产制造单位，要在数千个订单需求、上万条物料信息、百万级的匹配关系里，按照工厂的资源限制、工序要求、交货时间等关键约束信息，给出最优的排产计划。原有做法主要依靠个人经验，做起来耗时耗力，还无法找到最佳路径。

通过星云工业大模型和运筹优化技术，研发智能优化算法，滨江工厂实现了高效的优化决策引擎，可以在海量数据中快速找到最佳排产策略，在多重生产要素约束下输出多目标优化的最优订单承诺和生产计划。

智能排产系统使得订单排产周期缩短80%、订单齐套交付率提升58%、生产资源利用率提升20%、订单承诺偏差率降低36.5%，推动了中兴通讯供应链管理的智能化，极大提升了客户满意度。

资源智能调度

排产完成投入生产，就必须对生产资源进行高效调度。资源调度涉及核对的内容非常多，涉及多系统、多数据、多次查询。原先需要多个岗位员工协同，步骤繁琐，操作效率极低，核对数据后还需要手动排产到线体，同时手动创建领料单，费时费力。

通过星云工业大模型多智能体（agent）协同的生产资源调度管家，只需简单一步，用户发出调度命令，生产资源智能调度管家即刻响应并

立即启动多智能体协同流程，迅速指派关键角色——计划调度员、工艺员、设备工艺员、生产资源管理员及操作员共同参与本次资源调度任务。

多智能体协同下的资源智能调度管家，实现了流程精简87%、操作角色精简80%、任务释放到排程上线周期提效98%以上。

工艺文件智能生成

工艺文件是生产流程的设计文档，指导每个生产工艺和工序，内容要求非常精准且图文并茂。每年几十种上万份工艺文件编制非常耗时，且工艺文件通过人工编写，受限于工艺员经验与技能，文件质量参差不齐，一致性较差。

依托数字星云工业大模型，我们研发了工艺文档数字员工，针对行业复杂专业知识的提炼需求，结合大模型与RAG（retrieval-augmented generation）知识库，实现了过程信息的有效提取、转换及知识输出。用户只需通过自然语言对话，即可轻松生成所需的工艺文档，为行业带来全新的文档编写模式。

工艺文档数字员工支持一句话自动生成完整工艺文档，其编写效率相比传统方式提升10倍以上，同时保障了输出文档的高质量，生成准确率达95%以上。

产品智能维修

对故障产品的快速精准维修是生产环节中不可缺少的一环。传统产品维修依赖维修人员、工程师和技术人员的专业能力和经验，存在维修质量不稳定、维修周期长、人力成本高等问题。

利用滨江工厂积累的30000多条高质量维修标注数据，我们打造了维修专家数字员工智能体。当有问题产品出现时，AGV（自动搬运小车）会自动将其送到维修区域。维修人员通过自然语言与数字员工交互，数字员工实时获取并分析当前产品测试数据，并迅速给出可能的问题原因及详细的维修方法，同时还可以提供精准的测试验证方案。

产品智能维修提高了产品维修的效率和准确性，三次维修占比降低57%、维修周期缩短43%，降低维修成本17%。

多机器人协作

在5G小站产品的自动化测试过程中，需要AGV、单臂复合机器人和双臂人形机器人相互协同完成被测试5G小站的搬运、装卸工装、插拔网线/光模块等操作，这要求机器人高度自智、高度柔性、高度协同。

大模型加持的多机器人协作是具身智能机器人的高级阶段，其不仅仅实现静态的感知智能，而且和外部物理世界有交互有执行，在该场景我们采用大小模型结合的解决方案：通过星云工业大模型技术，自动完成任务规划和作业单元搭建，协调机器人执行相关动作；通过小模型强化学习，实现机器人的精准作业。

采用大小模型结合的方案，不仅保证了机器人的精准操作，而且做到了产线灵活适应不同型号的换线生产，灵活适应不同托盘、工装夹具的个体差异，真正实现工厂生产的高度柔性和协同。

智能制造是实现制造业高质量发展的新质生产力，中兴通讯将持续深入实践，推动滨江工厂智造进阶，催熟智能化场景应用和解决方案，并对外开放赋能。目前，中兴通讯已和中国商飞、青海盐湖、国铁集团、盐田港、云南神火、鞍山钢铁等头部企业进行合作创新和应用落地，获得了用户的认可。

未来，中兴通讯将秉持“数字经济筑路者”的生态定位，秉承多样互补、开放共赢的原则，与业界合作伙伴携手推进工业智能的创新与突破，同时也将高度关注工业智能面临的诸多挑战，推进企业增智、产业合作和行业赋能，推动智能制造健康有序进阶，共创工业智能新时代。 ZTE中兴

ZTE中兴

让沟通与信任无处不在