大数据技术标准推进委员会
Big Data Technology and Standard Committee

数据智能白皮书

（2024 年）

CCSA TC601 大数据技术标准推进委员会

2024年6月

版权声明

编制说明

本报告的撰写得到了数据智能领域多家企业与专家的支持和帮助，主要参与单位与人员如下。

参编单位: 大数据技术标准推进委员会、交通银行股份有限公司、中国平安人寿保险股份有限公司、中国海洋石油集团有限公司、南方电网数字平台科技（广东）有限公司、中邮信息科技（北京）有限公司、中移动信息技术有限公司、恒丰银行股份有限公司、小米通讯技术有限公司、中电信人工智能科技（北京）有限公司、联通数字科技有限公司、华为云计算技术有限公司、腾讯云计算(北京)有限公司、普元信息技术股份有限公司、中电金信软件有限公司、浙江大华技术股份有限公司、瓴羊智能科技有限公司、杭州阿里妈妈软件服务有限公司、星环信息科技（上海）股份有限公司、电科云（北京）科技有限公司、北京数势云创科技有限公司、北京市盛廷律师事务所、北京盛汉律师事务所、江苏联著实业股份有限公司、北京国电通网络技术有限公司、北京科杰科技有限公司、中国移动紫金（江苏）创新研究院有限公司、一网互通（北京）科技有限公司、杭州比智科技有限公司、杭州观远数据有限公司、深圳市明源云科技有限公司、海亮教育科技服务集团、芜湖明瞳数字健康科技有限公司、上海零数众合信息科技有限公司、天元瑞信通信技术股份有限公司、南京中新赛克科技有限责任公司、湖北数据集团、泽拓科技（深圳）有限责任公司、杭州网易数帆科技有限公司

参编人员: 王卓、姜春宇、马鹏玮、康宸、田稼丰、王超伦、刘

宾、杨靖世、郝志婧、尹正、周一帆、梅宇婷、朱晟、张义德、郑会丽、刘朝晖、范维、高健祎、杨光、包新晔、吴凡、王文颖、阮宜龙、陈卓、代莎、任鹏飞、余弘铠、刘涓、卫伟、高波、张淑娟、燕媛媛、史赟、李阳、高华超、龚禧、龙江、赵丽丽、李沐霖、叶嘉梁、贾宇航、蔡洛维、杜啸争、王笑非、王东风、周明伟、陈立力、江文龙、马里、孙蕾、陈思、胡晋渊、董鹏飞、侯承环、武文超 邢笑生、张广庆、方正、丁乙、韩秀锋、沈迪、李紫薇、毕文强、李永卓、张云龙、肖敬仁、姜怀舒、王楠、唐志涛、卢彩霞、余芳、朱建勇、贾光锋、王帅、彭涛、包岩、周晓阳、寇振芳、崔壤丹、何徐麒、张进、严林刚、石凯、曾伟雄、苑国跃、余震宇、谢耀圣、项灵刚、谭立何、杨博、闫阳阳、刘颋、兰春嘉、杨珍、李树磊、卢云川、顾欢欢、张全、钱龙、古伟、彭聪、石松、赵伟、孙国良、闫晶、宋昌

前　言

以“数据”和“智能”为代表的信息技术在数十年间快速融入全社会的生产、分配、流通、消费、社会服务管理等环节，不断带动生产力提升，推动社会进步。

近年来，伴随数据增列为生产要素、生成式人工智能技术实现突破，“数据”和“智能”产业均进入剧烈变革期，两者间的发展关系也发生巨大变化，“数据智能”顺势成为产业焦点。

为梳理数据智能相关知识体系，总结先进实践经验，研判未来发展趋势，指引企业顺利实现数智化转型，大数据技术标准推进委员会牵头，联合行业专家和头部企业首次共同编制《数据智能白皮书（2024年）》。本白皮书聚焦数据智能这一话题，梳理概念的诞生背景及发展历程，系统性厘清完整技术体系，深入剖析应用现状问题，展现产业生态全景，以期为企业未来的数据智能实践提供参考。由于时间仓促，水平所限，本白皮书仍有不足之处，欢迎联系 wangzhuo@caict.ac.cn 交流探讨。

目 录

图 目 录

表 目 录

一、 数据智能综述

(一) 数据智能概念探讨

近年来，智能领域突破“量变引发质变”的临界点，相关技术、产业进入剧烈变革期。自 1956 年人工智能（AI）概念诞生以来，智能计算领域历经多个阶段的技术方向探索，逐渐收敛在深度学习这一主线，但仍以“决策式人工智能”为主要发展领域。近两年，在以 Transformer 模型为代表的算法、极致算力支撑下的千亿级模型参数、大规模高质量的训练数据三者共同的作用下，生成式大语言模型的应用效果出现跨越式提升。以 GPT-4 为代表的大模型能实时对图像、音频、视频等多种形式输入进行理解，根据要求完成高效问答、内容生成等多种任务，甚至以前 10%的成绩通过美国模拟律师考试，由此“生成式人工智能”的发展成为全球焦点，带动人工智能技术产业进入剧烈变革期。

伴随智能领域变革，“数据”与“智能”间的发展关系亦呈现两点重要变化，“数据智能”概念亟需明确。如图 1 所示，数据和智能间的关系变化在近期主要体现为两点：

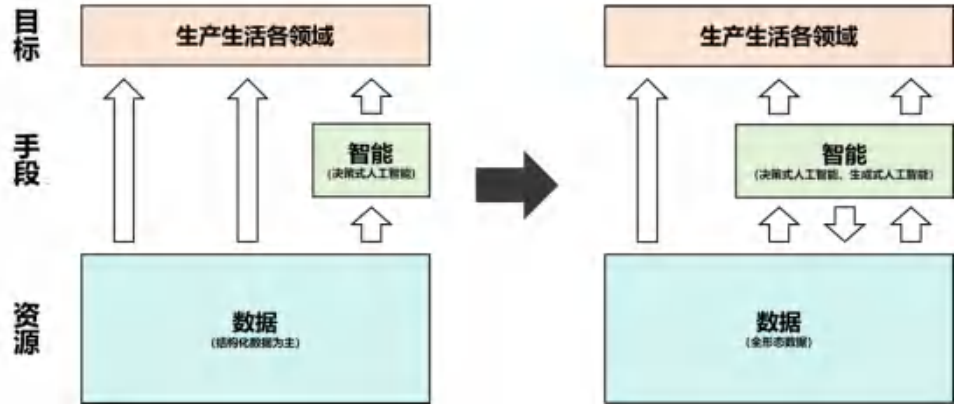


图 1 数据和智能间关系的变化

一是“智能”将成为“数据”价值释放的主要路径，“数据”成为“智能”成效进一步跃迁的胜负手，两者关系由“松耦合”转向“紧耦合”。长期以来，受制于智能技术的局限性，数据仍以非智能化的传统应用方式发挥价值，同时，智能应用效果的明显提升主要由算法驱动，数据仅作为研发过程中的基础一环，两者呈现“松耦合”式发展关系。然而，随着生成式大语言模型应用效果的飞跃式提升，人工智能对于生产生活各领域将逐渐不可或缺，进而成为数据价值释放的主要路径；同时，随着算力、算法的演进模式逐渐收敛，数据对智能持续发展的价值愈发突出。由此，助力智能发展将成为数据工作的核心，智能的效果提升也更加依赖数据工程及技术的托底，两者后续将转向“紧耦合”式发展关系。

二是智能化技术开始反向助力数据技术发展和非结构化数据应用。一方面，智能化技术开始应用至数据技术领域，在生成式人工智能的赋能下，数据的汇聚技术、存算技术、管理技术、开发技术、安全技术等快速向智能化升级，相应环节的生产效率有望得到大幅提升；另一方面，智能化技术突破传统数据技术面向非结构化数据的能力瓶颈，占据未来数据总量约 80%的文档、视频、音频等非结构化数据在生成式人工智能技术的助力下，可被迅速处理和分析，从而实现全形态数据的价值释放。

通过以上两点变化可见，数据与智能的融合大势所趋，由此“数据智能”的概念也应运而生。**数据智能的概念可以初步概括为，以全形态数据为关键资源，以大数据和人工智能深度融合后的新技术体系**

为关键手段，以决策式、生成式人工智能和传统数据应用形式协同应用于生产生活各领域为最终目标，由此形成的新兴生产生活方式，以及相应延展出的新技术、新产业、新生态。

(二) 数据智能的历史发展沿革

数据和智能是信息技术领域中最受关注的方向，其历史最早可以追溯到计算机的诞生，随后至今的近 80 年大致可分为三个阶段，总体的技术演进脉络如下图所示。

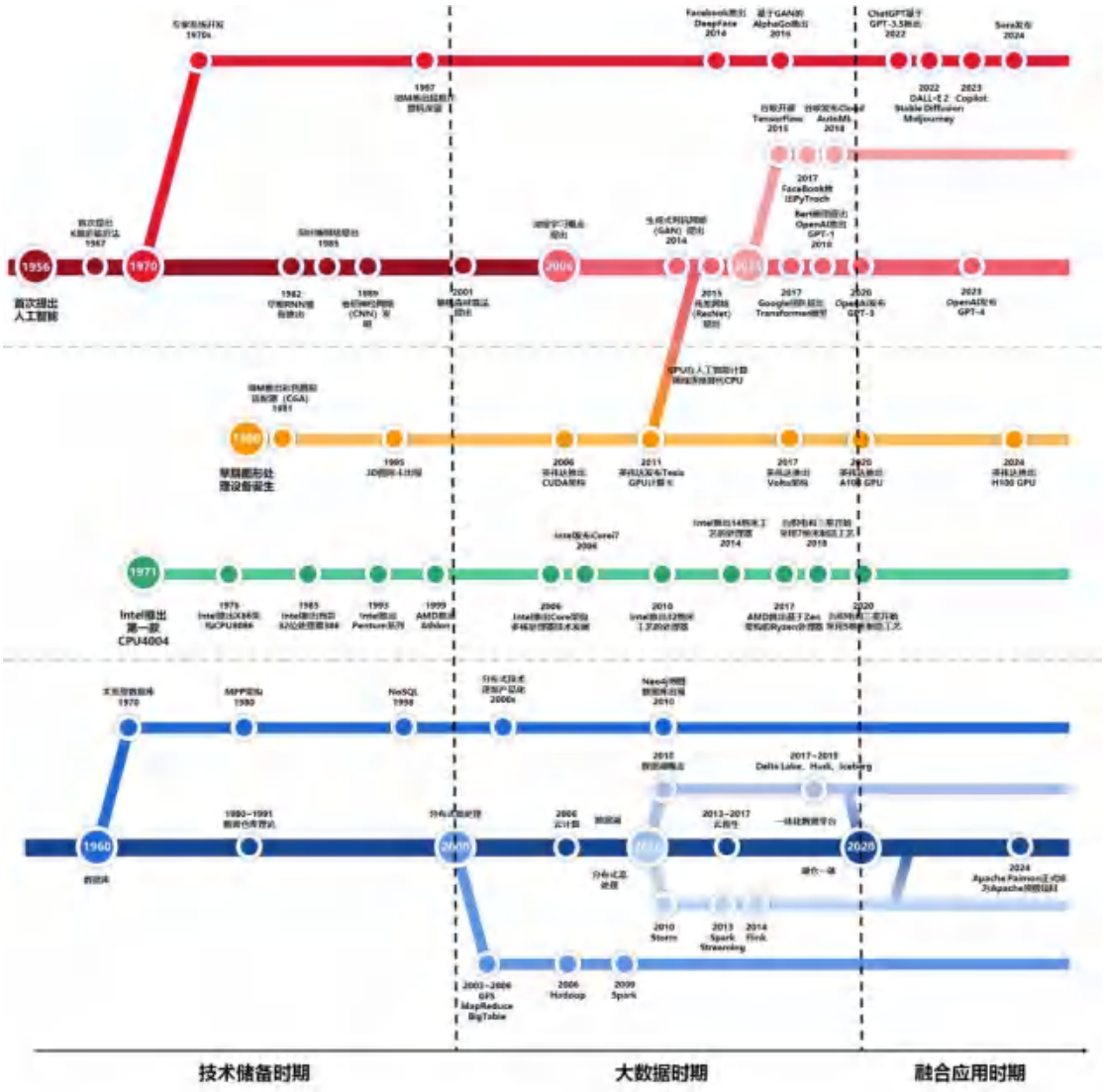


图 2 数据智能发展脉络

第一个阶段是技术准备时期（2000 年以前），这一阶段主要是由技术驱动的发展阶段。在计算机诞生后的 20 年内，通过计算能力形成人造智能的人工智能（AI）概念，和对数据进行管理和处理的数据库理论均已提出。随后，人工智能经历了从基于规则的推理方法到基于统计的机器学习方法的转变，经典机器学习和早期人工智能理论逐渐形成体系。数据领域则由关系型数据库完成大多数据管理和处理需求，同时诞生了数据仓库理论，指导企业使用数据库等相关工具实现基本的经营管理数据分析。这一阶段中，新兴信息技术不断涌现，为企业、产业、社会带来革新的生产力，信息技术的重要性为人所熟知。

第二个阶段是大数据时期（2000 年~2020 年），这一阶段主要是由数据驱动的发展阶段。随着互联网时代的全面到来，数据量的爆发式增长、数据类型的多样复杂化、时效性需求的愈发强烈，为数据的处理能力、智能算法的计算效率与效果均带来了新的要求，也使传统机器学习和数据库技术出现瓶颈，催生出以分布式处理为代表的提升数据处理规模和效率大数据技术，及通过多层神经网络学习加深模型效果的深度学习技术，数据和智能各自的技术发展进入快速迭代阶段。这一阶段中，数据量和数据类型的飞速增长进一步引领了技术的被动式革新，数据开始作为关键角色登场，受到的重视程度也与日俱增。

第三个阶段是融合应用时期（2020 年至今），这一阶段是由应用驱动的发展阶段，也是当前所处的发展阶段。近年来，移动互联网的普及和应用推动数据和智能技术的发展更加极致，更多样化和复杂的需求催使技术的发展和应用趋向融合，流批一体、湖仓一体、多模化

处理、多模态深度学习等已成为前沿发展方向，数据与智能技术进入相互融合深度应用以促进共同发展的道路。这一阶段中，单一技术的发展速度逐渐放缓，如何深化对已有技术的应用，充分发挥数据的内蕴价值，将数据和智能更为有机的结合成为更受关注的问题。**当下，以大语言模型为代表的生成式人工智能技术实践效果突出，其结合大量场景的应用正在加速落地**，围绕其应用落地相关的数据供给、模型优化、场景发掘、伦理安全等一系列问题成为时下热点。

(三) 数据智能的价值和意义

价值产生的本质，是能量、物质、信息三者内部或之间转换效率的增加。因此价值的具象化，也往往以效率提升的形式体现。数据智能借由传统数据技术加速了信息的收集和处理加工，借由智能化技术提升了信息精炼过程和人机信息传递交互的效率，从结果上实现了信息流动过程中更多环节由人工处理向智能化自动处理的靠拢和转变。

人力由于自身生理条件制约效率有限，相较由庞大能量支撑、运转速率高且信息传递顺畅的信息系统，更多的成为人机混合流程中的瓶颈环节，阻碍着串行流程运行的总体效率。随着智能化技术的持续进步，智能化自动处理模块相较人工处理造成的有效信息损失被压缩至相对可控和可接受的范围，使得智能化自动处理替代人工带来的整体效率提升更为可观，为更多人工环节的替换提供了现实基础。

在数据智能的实践下，以人为核心的生产环节，或被替代，或受益于技术赋能带来的生产效率提升，或受益于技术效果突破可用性临界点带来的新型生产方式及由此诞生的新生产环节。其中，被替代的

是具体环节而非人员本身，相反**每个个人作为独立的信息生产处理系统在综合作用下将得到最大程度的效率提升**，进一步的，随着规模效应的放大，将逐渐为企业、产业、社会等各层面带来新的价值和意义。

在企业层面，数据智能的实践能提升企业从数据中提取有效信息、精炼转化为知识、最终指导决策这一过程的总体效率，半自动化、自动化决策方式逐步落地。决策效率的提升和决策方式的转变，能够显著提高企业经营的响应速度和市场适应能力，促进业务流程优化和创新。例如，在金融业，帮助企业实现精准营销、风险控制和欺诈检测；在制造业，优化生产流程、预测设备故障、降低运营成本；在外卖、出行等行业，系统自动形成最佳调度方式并直接完成决策，显著提高效率和响应速度。

在产业层面，数据智能的实践在直接带动相关技术服务产业发展的同时，还将带来模式创新和对生产关系的重塑，以改善产业链总体产出效率。一方面，对于更高效专业化技术服务的持续性需求，将催熟联合运营等新兴产业合作模式。另一方面，生产端个人生产能力的水位上升将带动部分行业领域离散型个体供给模式的进一步兴起。例如，在内容生产行业，大模型的应用使个人生产效率全方位提升，专业分工进一步细化和整合，专业岗位进一步向外包、众包、共创等模式转变，最终提升综合生产效率。在更多行业中类似实践还将孕育着更多旧赛道的革新和新赛道的催生。

在社会层面，数据智能的实践能直接提升信息、知识在全社会范围内的流动效率，同时借由对信息的互通和技术的应用强化总体协同

性，优化社会资源的配置效率。世界历史上的重要发明如文字、纸张、印刷术、通信、互联网等都分别在各自的历史时期通过对知识传播效率的提升推动了生产力的发展和时代的进步。数据智能当下同样能够提高知识的易获取性以加速其在全社会范围内的流动和配置，并且在此基础上，能够进一步实现物理空间与数字空间的映射，实现社会运行各方面的高效协同，加速社会资源的合理配置，提升总体运行效率，为全社会带来更多福祉。

二、 数据智能技术

(一) 数据智能技术体系概览

当前，数据智能技术体系由数据技术及人工智能技术两大部分组成：数据技术旨在从各种类型的数据中快速获取有价值信息，涵盖数据全生命周期的各环节。人工智能技术是模拟人类智能行为的技术，涵盖基础自然语言处理、计算机视觉、智能推荐等细分技术方向。**总体来看，人工智能技术与数据技术相辅相成。**在模型训练前的数据准备环节，数据的处理离不开各类高性能存储及大数据平台的支持；在模型训练环节，各类数据平台为人工智能领域各类计算框架提供了有力的算力支撑；在应用开发环节，数据应用为各类人工智能模型提供了广阔的应用场景及用户数据，助力模型应用效果的进一步提升。

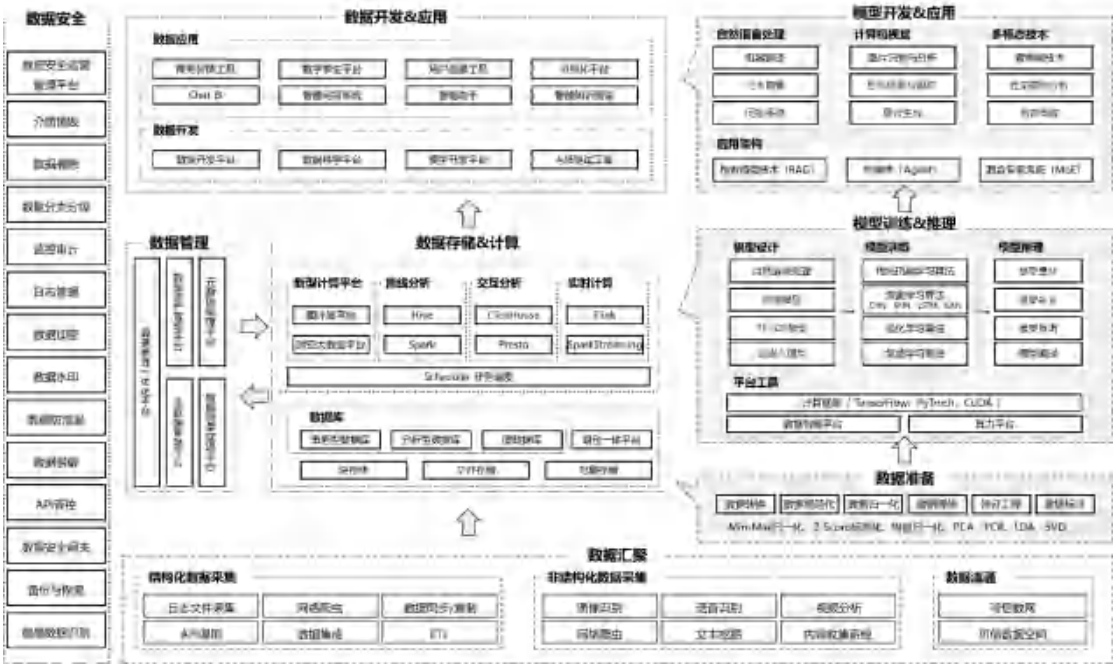


图 3 数据智能技术体系概览

当前，伴随着数据与人工智能技术的不断融合，逐渐演化出“5+3”技术体系。其中，数据技术可以按照数据生命周期分为数据汇聚、数

据存储&计算、数据管理、数据开发&应用、数据安全五大部分，人工智能技术可以分为数据准备、模型训练&推理、模型开发&应用三大阶段。

在应用需求的驱动下，数据与智能进一步融合创新。一方面，模型的生产需要高质量的数据资源以及更高效的数据底座支撑，另一方面人工智能技术的最新成果能够进一步赋能数据技术，提升数据处理效率和数据应用效果。由此，逐渐衍生出数据供给、多模数据存储与治理、数据智能平台、智能化数据安全技术等一系列新兴技术。

(二) 数据智能关键技术发展态势

1. 数据供给技术赋能模型训练

高质量的数据供给在人工智能模型的训练中扮演着至关重要的角色，直接影响模型的最终效果。高质量的数据可以提供准确的训练信号，帮助模型学习到有效的特征和模式，避免过拟合现象，增强模型在面对噪声、异常值和数据分布变化时的稳定性。随着各行业不断深挖数据要素价值，在数据供给领域，通过数据标注、合成数据提供高质量数据已经成为赋能模型训练的两大关键技术。

数据标注是指对原始数据进行分类、识别、标记和注释的过程。通过这一过程，数据的含义以能够被机器接收处理的形式表征，从而为模型训练提供结构化和有意义的输入，也是提升训练数据质量的关键环节。OpenAI 在 GPT4 训练过程中就使用了数据标注技术对大量互联网数据进行清洗和标注，保障数据的质量和一致性。

合成数据是通过专用数学模型或算法进行数据生成的过程，通常

可反映出目标原始数据特征，同时具备隐私保护、规模扩展、数据模拟等能力，可有效解决数据规模和质量不足等问题。如 J.P. Morgan 使用合成数据来模拟市场环境和交易数据，用于其金融服务策略的测试和优化。

数据与智能的融合应用，对高质量数据集的建设提出了新要求。当前数据资源供给存在“不能用”、“不够用”、“不好用”三方面问题。

一是存在数据开放程度有限、共享意愿低等问题，数据“不能用”。当前很多数据缺乏有效机制保障其流通性和可访问性，易形成数据孤岛，同时，公共数据目前开放和利用程度有限，未能充分发挥作用，造成企业难以获得高质量数据。

二是数据供给规模及效率有待提升，数据“不够用”。当前高质量数据供给难以满足模型训练和分析决策需求，数据供给质量低，整合清洗环节依赖人工处理存在效率瓶颈。

三是数据标准化及互操作性不足，数据“不好用”。数据格式、接口、存储等方面的标准化程度不足，导致数据整合难度高，互操作性差，增加数据处理成本。

随着企业数智化转型对数据价值释放需求的提升和对隐私保护的重视，数据供给技术将呈现如下趋势:

一是合成数据应用价值更加显著，将逐步应用于企业内风险预测、用户需求分析、模型训练等更多场景，满足企业数智化转型对高质量数据、高价值数据、多模态数据的需求。

二是数据标注向自动化、智能化演进。未来数据标注将更多地依

赖于自动化、智能化工具完成数据预处理过程，提供初步标注结果，再由人工进行校正和细化的方式提高数据标注效率。

三是数据质量问题将成为关注重点，通过建立严格的数据采集标准和流程，确保供给数据具有高质量、高相关性和高准确性。

2. 多模态数据存储与治理支撑模型高质量训练

高质量、多维度、大规模的数据是支撑大模型训练、应用的关键基础。当训练数据存在样本过少、错值、缺失、偏差等异常时，模型训练输出会产生偏见和错误，因此准确、可靠且涵盖各类场景的高质量数据对大模型训练必不可少。同时，不同模态数据的共同作用能够有效提升模型使用效果，一方面，将同一场景的图片、文本、音视频、知识库等同时作为训练数据能够增强大模型的理解能力；另一方面，当基于文本数据的训练出现偏差时，其他模态数据可以辅助大模型进行错误纠正，减少“幻觉”。如何对多模态数据进行高效存储、计算、治理已逐渐成为数据智能领域的重要技术方向。

当前多模态数据的存储治理仍存在以下突出问题：

一是多模态数据整合处理难度大，读取效率有待提升。多模态数据包括结构化、非结构化及半结构化数据，数据来源多样、数据量大、格式不一，因此整合难度较大。此外，在模型训练过程中需要对海量数据进行读取操作，对多模态数据的缓存加速能力也提出更高要求。

二是面向模型训练，数据质量治理环节亟需前置。在模型训练过程中，数据质量治理环节需前置，在数据收集阶段同步并行，以保证训练数据集的准确、合规、完整。但当前数据治理流程通常是在数据

应用过程中发现问题，从末端到源端，层层梳理数据血缘，定位问题点，进行数据的改进和补充，造成数据治理环节后置，难以满足需求。

未来，多模数据存储与治理领域呈现出三大趋势:

一是支撑多模数据的高并发高吞吐存取需求。底层存储将更加注重性能优化与扩展性，支持统一管理多个命名空间，避免单点瓶颈，以解决多中心集群数据统一存储与共享问题；兼容多种存储协议，如 POSIX、HDFS、S3 及 CSI 等；支持分布式缓存，通过多级缓存加速，提高热点数据命中率，持续提升存储集群性能。

二是构建多模态数据标准，促进数据的整合、共享、交换。通过构建一个多层次、可扩展的多模态数据标准体系，为不同来源和类型的数据提供统一的处理和分析方法，有效解决多模态数据不均衡、难对齐、存在语义鸿沟等问题，降低多模态数据的整合难度，减少数据转换和清洗工作量，助力多模态数据的有效利用。

三是依托各类技术工具实现数据质量治理环节前置。当前，如英伟达、微软、谷歌和 OpenAI 等厂商已经开始基于多模态元数据和多模态数据标准，制定多模态数据质量检测指标并构建检测任务的技术实践，在数据汇聚阶段保障数据质量。未来，数据质量治理环节前置将成为提升模型训练效率，增强数据融合水平的关键。

3. 数据智能平台支撑企业数据及模型开发

数据智能平台是企业数智化能力构建的重要基础，为上层应用、决策提供数据、算力支撑。一方面，人工智能技术被用于将复杂的数据分析过程自动化，快速识别数据中的模式和趋势；另一方面，数据

平台为上层模型提供更强的算力及更高质量的数据，推动模型开发范式向以数据为中心的模式转变。当前，Databricks、Snowflake、阿里云、华为云等国内外大数据厂商均推出具备数据存储、计算、开发能力的 Data+AI 解决方案。

随着大模型技术的进一步普及，对数据智能平台的异构资源调度、向量化计算及智能运维能力提出了更高要求:

一是异构计算资源高效纳管能力有待提升。模型训练需要大量 CPU、GPU 等异构计算资源的支撑，如何在同一集群中高效纳管异构计算节点，对算力资源进行自动化部署、监控、调度和优化等操作，满足不同规模企业的模型训练需求成为重要问题。

二是数据平台向量化计算能力有待增强。向量化计算是将传统的基于循环的矩阵运算转化为基于整体矩阵操作的计算方式，能够显著提高模型训练计算性能，但当前计算框架对向量化计算支持有限，亟需开发新的编程模型和架构以集成更高性能的向量化计算能力。

三是运维能力的智能化程度有待加深。数据智能平台对海量异构数据的计算加速也带来了巨大的运维压力，当前运维体系的故障自动诊断准确性和时效性有待提高，亟需智能化技术在运维领域深度应用。

未来，数据智能平台发展主要有以下三大趋势:

一是利用云化、智能化、多集群等技术实现平台算力与成本的平衡。一方面，通过智能化技术，实现任务的自动调度和资源的智能分配，提高资源利用率和系统性能；另一方面，随着多云和多地部署趋势的增加，分布式调度系统将更加关注跨集群的任务和资源管理，实

现集群间资源协作和任务调度。

二是模型训练推理需求推动向量化计算技术进一步集成发展。向量化计算是提升模型训练、推理性能的重要手段，未来数据智能平台将通过新的编程模型和架构，提升自身的向量化计算性能。当前，云服务商也正在提供更多集成的向量计算产品和服务，以吸引对高性能计算有需求的企业客户。

三是利用人工智能技术增强数据智能平台运维能力。随着大模型与运维技术相结合，数据智能平台可以通过实时数据分析，及时发现异常，触发故障自动诊断机制并自动给出解决建议，减少人工干预和诊断时间。同时能够构建预测系统性能、效率模型，自动调整引擎参数和任务参数，达到系统性能和效率的最大化。

4. 数据流通技术支撑企业安全高效汇聚利用外部数据

在企业持续推进自身数据智能化的过程中，发现、获取和利用大规模、高质量和多样性的数据是其中关键。部分场景中单一企业的数据规模和多样性不足，需要融合利用外部数据以增强模型能力，因此，数据流通技术已成为实现数据智能的核心技术之一，除需要关注数据流通过程中数据的可控与安全，在面向大规模、多模态数据流通时，也需要保证数据流通的可用性和稳定性。当前，蚂蚁、腾讯、华为等企业均有开发隐私计算、数据空间等数据流通解决方案，助力数据可控安全的流通利用。

在面向企业数智化转型的过程中，当前数据流通技术仍存在以下问题：

一是部分场景中仍面临安全挑战。当前隐私计算产品大多以“半诚实模型”的安全假设为前提，但在实际使用中安全假设不一定成立，参与方可能违反合约和诚信要求，出现伴生攻击、数据投毒等行为。同时，大模型参数多和模型复杂等特点，为基于隐私计算的联合训练和推理算带来新的安全挑战。另外，当前大部分数据流通产品的身份管理、使用策略设置等功能不完善，可能造成流通过程中的数据或模型信息泄露。

二是大规模数据计算时的性能不足。隐私计算技术实现中的密文计算将带来一定的额外计算和通信负载，实际应用中也因通信带宽的限制会影响多个参与方间的数据交互性能，且当前主要以支持结构化数据为主，对大规模、多模态数据计算的支持仍有待提高。

面向企业数智化转型，为更高效支持企业获取和利用外部数据，数据流通技术未来主要有以下趋势:

一是算法协议框架优化支撑数据高效流通。业内持续进行联邦学习算法优化，产出了模型压缩、本地多轮迭代、异步协调策略等方案，并进一步探索研发联邦大模型的算法框架。同时，基于多方安全计算的大模型安全推理也形成了相关成果，这些技术方案有效降低由通信数据量和大规模模型参数等因素带来的性能影响，有效推动了隐私计算技术在复杂模型训练和推理场景的落地。

二是多技术融合实现可信数据流通。隐私计算各技术路线有性能和安全性的不同侧重，多技术融合、软硬件结合是隐私计算突破单点技术瓶颈的有效方式。同时，隐私计算也将结合数据使用控制、区块

链等技术形成更加可信安全的数据流通解决方案，保证在多主体参与的数据流通全过程可控安全。

5. 智能化技术赋能数据安全产品升级换代

当前，数据安全产品的智能化已在多个领域得到应用，例如敏感数据识别、数据防泄露等，这些技术通过结合机器学习、深度学习等人工智能算法，实现对数据的智能保护和风险预警，使数据安全产品能够更准确地检测到潜在的安全威胁和异常行为，区分正常和恶意行为，自动响应安全事件，快速采取行动，实现主动防护，为企业提供更全面、更高效的安全保护。

数据安全产品的智能化已取得了一定的进展，但仍存在一些问题需要解决。

一是智能化技术的准确性和可靠性仍需进一步提高。由于数据的复杂性和多变性，一些智能化算法在处理数据时可能会出现误判或漏判情况，导致数据安全风险无法及时发现和处理。

二是智能化技术的可解释性和透明性不足。部分智能化算法在处理数据时采用了黑箱操作的方式，导致用户无法理解算法的决策过程和依据，增加了数据安全的不确定性和风险。

三是智能化技术的应用范围和深度仍需进一步拓展。目前，智能化技术主要应用于一些特定的数据安全场景，如敏感数据识别、数据防泄露等，但在一些其他领域，如数据安全治理、数据安全风险评估等方面，智能化技术的应用仍相对较少。

未来，智能化数据安全产品将呈现出两点趋势：

一是自动化、智能化、集成化将成为未来发展方向。随着不断变化的网络威胁及人工智能技术的不断发展和成熟，智能化技术将与数据安全产品进一步结合，提高对复杂威胁的识别、预测和响应能力，利用算法进行主动监测并分析潜在的安全威胁，实现风险的早期发现和预防，为用户提供更全面、更高效的安全保护。

二是智能化技术将与其他安全防护手段相结合，形成更加完善的数据安全保护体系。通过将智能化技术与加密技术、访问控制技术等相结合，同时与服务将进一步融合，为不同行业和场景提供灵活的安全解决方案，实现对数据的全方位保护，提高数据安全的整体水平。

6. 生成式大模型驱动生产力跃升

生成式大模型指具有大规模参数和复杂计算结构的生成式机器学习模型，通常基于深度神经网络模型，拥有数十亿乃至数千亿参数，其设计目的是为了提高模型的表达能力和预测能力，被广泛应用在自然语言处理、计算机视觉、语音识别、推荐系统等场景。

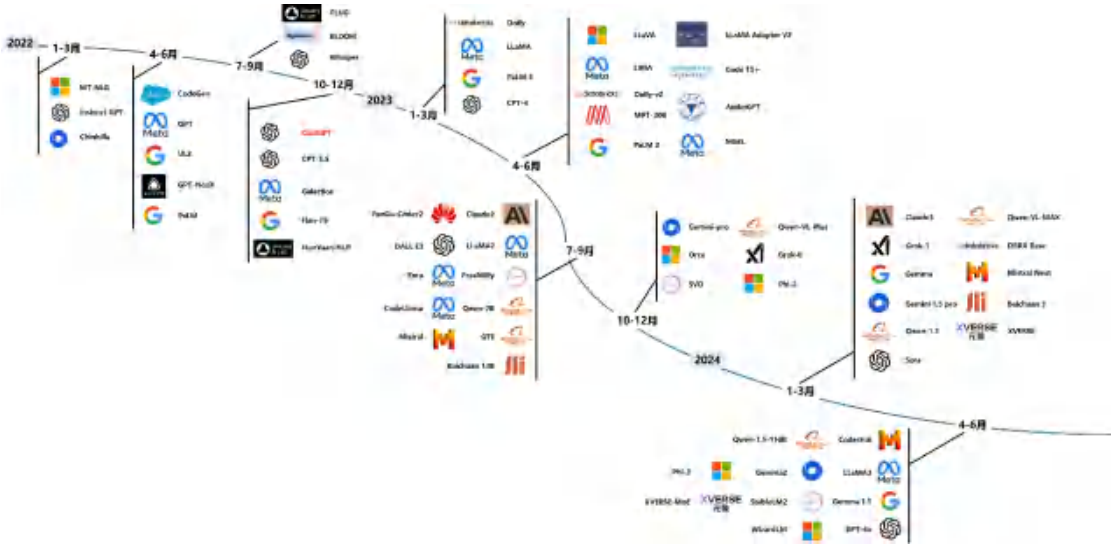


图 4 部分生成式大模型发布情况统计

与小模型相比，大模型拥有更好的复杂任务处理能力，且具备较强的迁移学习能力。但相应的大模型需要大量计算资源进行训练和推理，训练时间长，可解释性较差。小模型在参数规模上较小，训练效率高，可以进行快速迭代，部署灵活，更易在不同平台上部署，尤其是资源受限的环境，并且在特定场景任务下表现超越大模型。但小模型处理复杂任务的能力有限，迁移学习能力弱。因此，在选择使用大模型还是小模型时，需根据具体的应用场景、资源可用性、性能需求和预算等因素综合考虑。

生成式大模型的发展促进了各行业数据智能落地实践，但也带来了两方面问题：

一是生成式大模型可能生成虚假、有害的内容。当前受语料、模型算法等因素影响，部分模型易出现生成虚假信息的现象，导致可能输出错误观点，甚至易被诱导输出伪造信息和有害内容。

二是生成式大模型存在数据安全及隐私问题。模型训练需要大量数据，其中极可能包含敏感和隐私信息，存在数据泄露风险，同时，部分用户在使用过程中，也可能通过特定方式套取部分隐私信息。

未来，生成式大模型发展呈现出三大方向：

一是通过多模态数据提升模型训练效果。OpenAI 公司的 GPT-4、Meta 的 Llama 3 和 Mistral 均为多模态生成式模型，允许用户基于文本、音频、图像和视频匹配内容，以提示和生成新内容。通过将图像、文本和语音等多模态数据与算法相结合，能够有效提升大模型的训练和使用效果，减少“幻觉”。

二是视频生成大模型成为生成式模型发展的前沿方向。OpenAI 于 2024 年 2 月 15 日发布文生视频模型 SORA，将视频生成时长从秒级大幅提升至一分钟，且在分辨率、画面真实度、时序一致性等方面都有显著提升。SORA 具备理解世界的基本物理常识并进行预测的能力，标志着智能技术发展进入新阶段。未来，视频大模型将在数字孪生、虚拟现实、增强现实、内容创作等场景具有广阔的发展空间。

三是垂直领域大模型将成为主战场。通用大模型拥有广泛的适用性，具备跨域学习能力，但存在资源消耗大，特定领域任务表现较差等问题。专业领域大模型可以根据特定行业的需求进行定制化开发、优化，能够更准确地理解和处理特定领域任务。专业领域大模型在 ToB 市场拥有广泛的应用前景。当前，金融、电信、能源等行业已经开始大模型应用实践的探索，未来，专业大模型将成为生成式模型发展的重要方向。

7. 大模型赋能的数据智能应用促进数据智能价值释放

数据智能应用技术是指包括数据可视化、数据分析、数据挖掘、机器学习在面向实际应用内的数据智能技术，旨在从数据中提取有价值的信息和知识，从而驱动决策，赋能企业具体业务。数据智能应用技术同具体业务场景相关性强，存在层次多、差异大、需求多元、形式复杂的特征。近年来，随着以大语言模型为代表的人工智能技术快速发展，数据智能应用的模式正在快速改变，基于大模型的对话式 BI、数据分析智能体等新模式纷纷涌现，数据智能应用的发展迈入新阶段。

当前阶段，数据智能应用技术仍面临着三大问题：

一是数据智能应用技术的门槛仍然较高。数据智能应用技术融合了数据科学、统计学、计算机科学、领域知识等多方面专业知识，部分专业的数据分析工具具有较高的学习使用门槛，企业在实际应用过程中存在一定适应难度。

二是技术和业务的脱节仍然存在。数据智能技术工具同业务需求脱节的现象普遍存在，导致数据智能相关工作难以提供及时、准确的业务洞察。少数重点场景如营销、风控等一些头部企业能做到将应用深度地嵌入业务，但在其他大部分企业场景中仍难以实现。

三是数据智能算法的可解释性有待提升。数据智能算法的可解释性关系到人们对算法的信任、算法的公平性、以及算法的安全性等多个方面。许多数据智能算法因其高度复杂和非线性的特性，往往被视为“黑箱”，其中的过程和逻辑往往不透明，难以理解，在一些基于大模型的数据智能应用中普遍存在。

未来，数据智能应用技术发展呈现出三大趋势:

一是“大模型+数据智能应用”将成为各方数据智能技术能力建设的重点。随着大模型技术的快速发展，基于大模型的智能增强分析工具、智能问答工具、智能检索工具、知识图谱工具等成为了各方建设的重点，华为云、腾讯云、科大讯飞等企业推出的各类大模型+数据智能应用工具，在稳定性、灵活性、专业性、多模态数据处理能力等方面正在快速提升，未来仍有较大发展空间。

二是数据智能技术将为大模型的落地提供更高效支撑。数据智能应用技术对大模型的赋能也成为各方关注的重点方向，基于知识图谱、

知识库的检索增强生成技术大大增强了大模型生成内容的准确性，已成为商业大模型应用的主要实现路线。

三是数据智能技术正在同业务深度融合。数据智能技术与业务深度融合是现代企业提高竞争力、优化决策过程和增强客户体验的关键。在数字营销、智能风控、数字化运营、数智财务等重点场景中，数据智能技术正在更加深入地落地实践。

(三) 数据智能技术未来展望

随着数据与人工智能技术的不断融合，数据智能技术对数据进行处理和分析能力将持续提升，实现更高效、精准的数据挖掘和应用，推动数据要素价值进一步释放。数据智能技术的未来发展预示着一个多维度和深层次的融合创新时代，在数智化转型的过程中，数据智能将更深入地渗透到各个行业和领域，应用场景将进一步拓宽。通过优化业务流程、分析业务问题、洞察业务趋势，数据智能将为企业和组织带来更高的效率和更精准的决策支持。

三、 数据智能应用

数据智能应用是指**从数据中提取有价值的信息和知识，构筑智能算法和模型、推动决策和行动**，以实现提高效率、增强体验、驱动创新等目标。对企业来说，**数据智能应用是数智化转型的核心组成部分，是释放数据价值的最终一环，也直接决定了数据智能相关实践的最终成效**。企业开展数据智能应用工作具备一定的复杂性，可以分为两个层面的能力构建：一是通用的数据智能应用能力，包括数据可视化、数据分析、数据挖掘等能力；二是场景化数据智能应用能力，包括营销、运营、风控、财务等具体场景的数据智能应用能力。此外，不同行业的机构对数据智能应用的侧重点也有差异。本章将按照通用、场景、行业三个层级探讨数据智能应用的现状、问题和发展趋势。

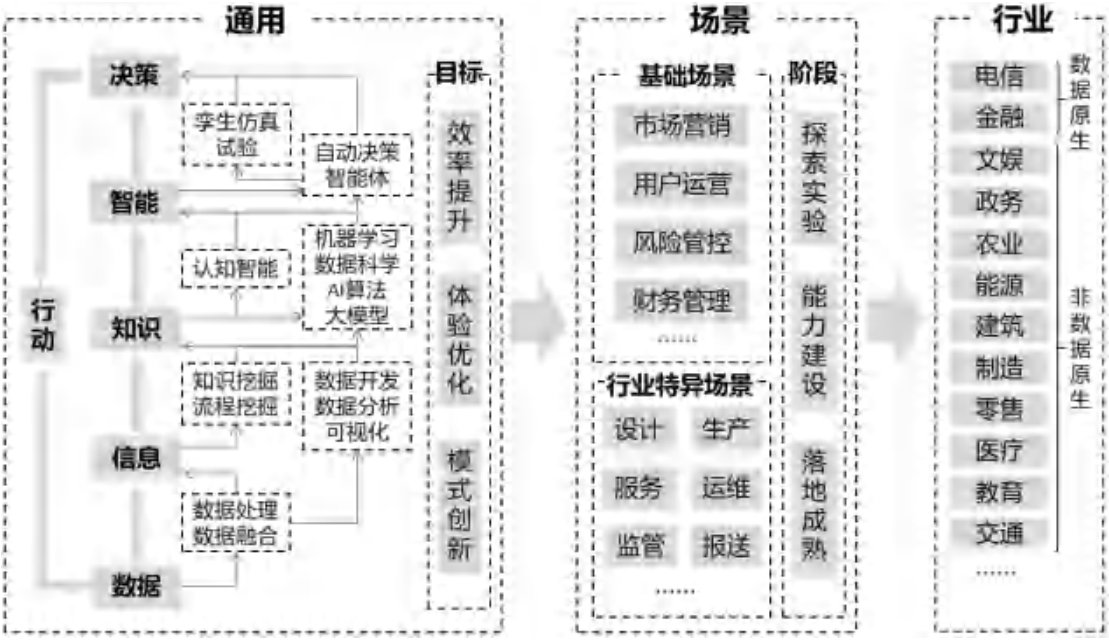


图 5 数据智能应用体系概览

(一) 数据智能应用发展态势

通用范畴来看，数据智能应用的目标呈现出多样化发展的态势。数据智能应用的目标可分为效率提升、体验优化和模式创新三类：**一是效率提升**，即通过自动化手段改进原有场景下的工作流程，减少重复和冗余的人工操作，例如使用算法来自动化常规的数据分析任务，或者利用机器人流程自动化（RPA）来执行重复性高的业务流程，从而提高工作效率并释放人力资源。**二是体验优化**，即通过数据智能技术对现有业务场景中的痛点进行改造，以提升业务场景中各类参与者的体验，例如通过个性化推荐系统、智能助手和自动化工具实现客户和员工的体验改善。**三是模式创新**，随着大模型技术的快速落地，智能助手、智能客服、虚拟陪伴、个性化营销内容生成、智能化医疗诊断等新业态、新模式成为发展的重点。传统的数据智能应用更加强调效率提升和体验优化，大模型的强大生成能力和多模态数据处理能力正在深度改变现有的工作模式和业务流程，并催生出新的产品和服务，数据智能应用的目标也将更具多元化。

场景范畴来看，大部分业务场景仍处于数据智能能力建设阶段，营销、运营、风控、财务等场景正在向落地成熟阶段转变。营销、运营、风控、财务等场景具备**高价值、共性强、数据密集、成效明确**的特点，成为数据智能应用优先落地的场景。在营销领域，数据智能技术被用来深入分析客户行为和偏好，实现精准营销。通过收集和分析客户数据，企业能够构建用户画像，预测市场趋势，并制定个性化的营销策略。在用户的运营管理过程中，数据智能被用于构筑用户画像、

优化用户体验并通过数据智能技术构筑更加个性化的用户服务。在风控领域，数据智能技术被用于信用评估、欺诈检测和风险预测，通过历史数据分析，企业能够识别潜在的风险模式，采取相应预防措施。在财务领域，数据智能被用于财务报告撰写、预算编制、财务分析和预测，帮助企业优化资金流管理，提高财务决策的质量和效率。

表 1 数据智能应用发展阶段

阶段	探索试验阶段（1960s开始）	能力建设阶段（1990s开始）	落地成熟阶段（2015s开始）
数据来源	以企业内部结构化数据为主	企业内部结构化数据、外部开源数据	企业内部多模态数据、外部数据
模式	随机的、临时发起的数据智能应用	常态化、体系化、外挂式的数据智能应用	服务全域的、敏捷的、嵌入业务的数据智能应用
决策支持技术	图表统计为主	数据挖掘、可视化报表、BI等通用分析工具	BI、AI、专业业务领域分析工具相结合
对决策的影响	辅助决策	增强决策	自动决策
对创新的驱动力	流程驱动	需求驱动	数据驱动

行业范畴来看，数据智能应用已经渗透到各个行业和领域，大部分行业正处在能力建设阶段。依托对近三届**行业数据应用星河案例**申报企业的统计分析，整理出各行业数据智能应用的头部落地场景。各行业数据智能应用的头部场景存在差异，其中营销及经营管理部分关注的细分场景相似度较高，研发设计及生产服务等场景具备较强的行业特异性，数据智能应用的侧重点存在较大的差异。总体上来看，以生产型服务业为主导的第三产业在数据智能应用的深度和广度方面都显著强于其它行业，在数据智能应用创新过程中发挥着引领作用。

表 2 各行业数据智能应用落地的头部场景

行业	研发&设计	生产&服务	营销&销售	经营&管理
第一产业				
农业	• 智能育种 • 农业规划 • 农业装备设计	• 农业装备制造 • 农业生产运营 • 农业技术服务	• 农产品销售	• 农产品市场分析 • 农资管理 • 农业供应链管理
第二产业				
制造	• 产品设计及优化 • 智能仿真及建模 • 知识产权管理	• 自动化生产线 • 供应链优化 • 设备维护管理	• 客户画像分析 • 销售趋势预测	• 库存管理 • 风险管理 • 成本控制
能源	• 电网/营网设计 • 新能源技术研发	• 智能生产监控 • 自动化生产线 • 智能电力调度	• 需求预测 • 定价策略分析 • 客户体验管理	• 服务质量分析 • 风险管理 • 成本控制
建筑	• 建筑设计辅助 • 建筑信息建模 • 设计方案评审	• 智能施工管理 • 工程技术资料管理	• 市场分析 • 设施选址 • 房地产营销	• 供应链优化 • 设施管理 • 成本控制
第三产业				
金融	• 信贷风险管理 • 投资分析 • 交易算法设计	• 智能客服 • 个性化服务 • 智能投顾	• 自动化精准营销 • 营销活动监管 • 客户关系管理	• 安全合规监管 • 绩效评价及优化
电信	• 网络设计及优化 • 电信产品创新	• 服务体验优化 • 智能客服 • 网络运维优化	• 自动化精准营销 • 营销策略制定 • 用户行为分析	• 安全合规监管 • 风险管理及反欺诈
零售	• 门店设计 • 门店选址 • 品牌定位分析	• 库存管理及优化 • 门店运营 • 客户体验管理	• 品牌营销 • 广告投放分析 • 销售数据分析	• 供应链优化 • 绩效评价 • 门店管理
医疗	• 药物研发 • 临床试验设计 • 医疗装备设计	• 个性化医疗 • 智能诊断	• 医疗产品定位分析 • 患者需求分析	• 医疗资源优化 • 医疗风险管理 • 患者信息管理
教育	• 课程开发 • 智能教育工具开发	• 个性化教学 • 智能教辅 • 虚拟教师	• 招生流程管理 • 目标市场分析 • 课程营销	• 学生及教师管理 • 教学质量监控
交通	• 道路规划及设计 • 智能交通系统研发	• 道路监控 • 交通设施运维 • 交通疏导及调度	• 出行行为分析 • 出行流量预测	• 交通运营管理 • 道路风险预警
文娱	• 内容创作辅助 • 内容素材生成	• 个性化内容推荐 • 用户体验分析	• 营销海报生成 • 目标受众分析 • 广告效果评估	• 内容绩效分析 • 舆情分析 • 版权管理
政务	• 城市规划设计 • 政策制定辅助	• 政务服务 • 公共安全监管 • 公共服务管理	• 政策宣发 • 服务推广	• 绩效管理 • 资源配置及优化

(二) 数据智能应用当前问题

通用范畴来看，跨部门协同困难是企业数据智能应用落地面临的主要问题。在大中型企业，数据智能应用的落地涉及到业务、技术、数据等多个部门的共同协作，跨部门协同难度较大，主要存在三方面问题：**一是缺乏复合型的数据智能应用技术人才。**数据智能领域需要同时精通算法、擅长工程实现、深刻理解业务的复合型人才，目前这类综合性人才相对稀缺，如缺少这部分人才发挥协同作用，各部门人员跨领域沟通易出现困难。**二是缺乏一体化的顶层设计。**缺乏顶层设计和跨业务、跨领域统筹规划的现象普遍存在，使对于数据智能应用的建设各自为政，难以形成合力，技术和业务需求之间存在脱节，导致数据智能应用无法有效解决实际业务问题，造成资源的重复投入和低效建设。**三是跨部门的安全合规管理困难。**随着《个人信息保护法》《征信业务管理办法》等政策法规相继出台，各方对数据智能应用的安全合规管理逐步收紧，跨部门的安全合规管理仍存在标准不一、责任划分不明确、审批流程复杂等问题，一定程度上制约了数据智能应用的发展。

场景范畴来看，数据智能应用对场景的赋能效益存在着计量困难问题。数据智能应用的成效计量需要依托具体的业务场景，由于很多数据智能项目具备一定的创新性，相对参考较少，成效计量困难的问题普遍存在。**一是对数据智能应用产出的预期估计普遍存在偏差，存在着高估短期收益，低估长期收益的倾向。**企业倾向于对数据智能技术短期内能带来的效益抱有过于乐观的预期，期望能够通过快速建设

系统、引入工具平台来实现显著的短期收益，但算法的调优、员工的培训、制度流程的适配和企业文化的变革通常需要更长的时间来逐步完善优化，数据智能应用的收益往往需要更长的周期才能显现。**二是数据智能应用成效评价方式不合理，存在着注重技术指标，轻视业务指标的倾向。**场景类应用的成效评估是一个复杂问题，需要综合考虑多个因素，通常可以细分为技术指标（如算法误差、群体稳定性，数据处理性能、并发度等）和业务指标（如业务效益变化情况、用户满意度、客户转化率、复购率等）。由于技术指标是最容易得到的，例如机器学习算法模型，可以较低的成本对模型的算法误差等技术指标进行监控，而用户满意度、业务效益等指标的计量较为复杂，往往需要单独设计 A／B 试验、统计归因等方法进行监测，导致一些企业存在注重技术指标而轻视业务指标的现象，隐含着技术和业务需求脱节的风险。

行业范畴来看，不同行业的数据智能应用所面临的问题存在一定差异，落地思路较难跨行业复用。这里将按照数据原生行业（如金融、电信等）和非数据原生行业（如制造业、医疗、交通等）进行探讨。**数据原生行业（如金融、电信等）数据智能应用所面临的问题主要集中在安全合规、实时性等方面。**相对其它行业来说，数据原生行业的数据量大，数据质量高。由于涉及到海量个人用户的敏感数据，数据智能应用过程中面临着数据隐私和安全等问题。此外，这些行业的数据分析应用对实时性要求较高，如何能够快速的进行数据实时处理和分析，响应市场变化也是这些行业关注的重点问题。**非数据原生行业**

（如制造业、农业、交通等）数据智能应用所面临的问题主要集中在数据质量、采集成本、多模态处理技术等方面。这些行业的数据量相对较小，数据质量不稳定，数据种类更加复杂多样。农业、建筑、制造业的生成数据采集通常采用传感器等设备，相较于数据原生行业来说具备较高的成本，且通用性较弱。因此，数据智能应用在这些行业需要考虑到数据采集的成效问题，并且需要采用适合的数据处理和分析技术。医疗、交通等行业的数据种类多样，对这些多模态数据的分析应用具有较高的行业壁垒。综合来看非数据原生行业的数据智能应用模式尚未成熟，需要进一步探索和实践。

(三) 数据智能应用未来展望

通用范畴来看，大模型的落地将重塑企业数据智能应用模式。大模型的落地将对企业数据智能应用模式产生深远影响。在数据分析方面，大模型能够同 BI、报表工具相结合，实现更便捷的对话式分析，大大降低非专业人员的用数门槛。在知识管理及服务方面，大模型与知识库、知识图谱相结合，可实现一些复杂的推理和决策支持，提供智能问答等知识服务形式。在业务自动化方面，大模型赋能的智能助手可辅助设计或优化业务流程，自动化进行任务规划、资源分配和时间管理等，并且能够自动化执行重复性较高的任务。具体来看，有以下三大趋势:

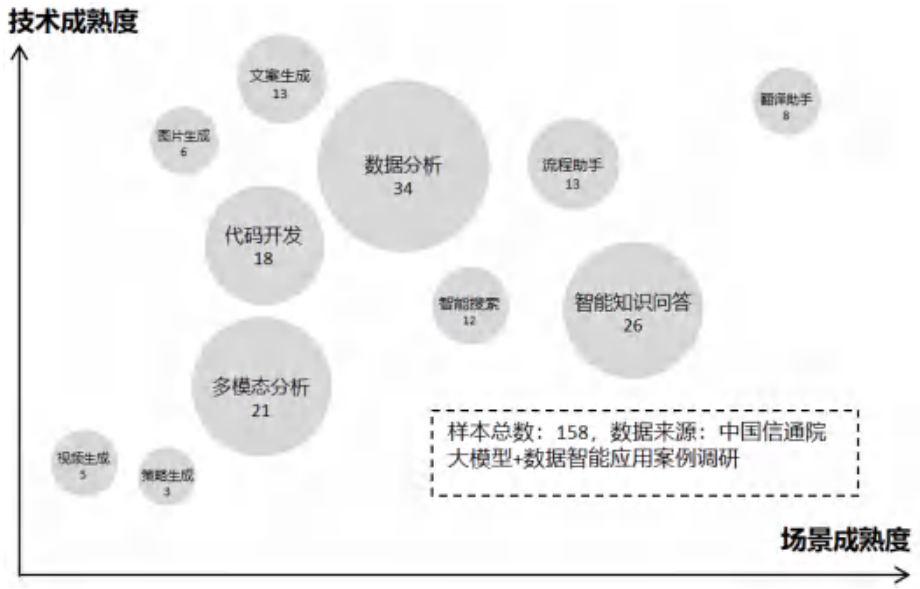
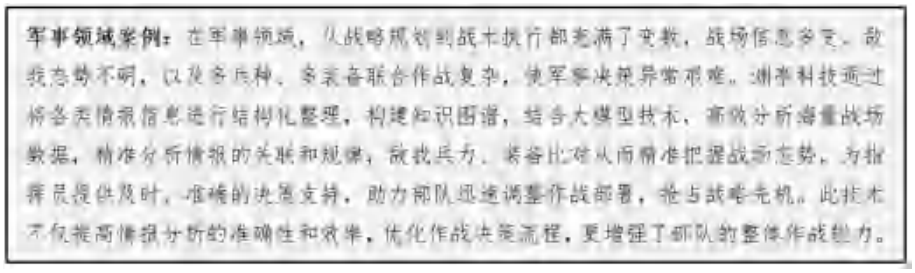


图 6 大模型赋能的数据智能应用场景

一是大模型技术将同传统的小模型、机器学习技术相结合，全方位提升企业多模态数据分析能力。人工智能和机器学习技术是数据分析应用的核心技术。在大模型技术兴起之前，参数规模较小的模型和机器学习算法在图像分析、社交媒体分析、用户行为分析，等分析业务中已取得优异成果。大模型驱动的数据分析工具将趋向于整合多种类型的数据分析能力，如文本、图像、社交媒体、指标及标签等，通过智能体（Agent）调用高度专业的小模型及机器学习算法并整合分析结果。大模型的应用落地将有助于企业实现对多模态数据的高质量分析，从而更全面地理解和处理复杂问题。

零售领域案例：由于多渠道、多门店、多品牌带来的生意复杂度，在高频营销活动、高频商品上新、高频组织调整带来的冲击下，企业及时获取数据进行决策的诉求变得越来越迫切。观远数据通过大模型赋能的新一代智能分析与决策平台赋能某国内领先美妆公司，协助 10 个部门快速构建起智能数据助手服务，平均不到两周即可引入一个新业务主题，帮助所在部门的业务人员从天级别的数据需求响应周期，优化至分钟级别即可获取所需数据与洞察，需求处理效率提升 200 倍以上。

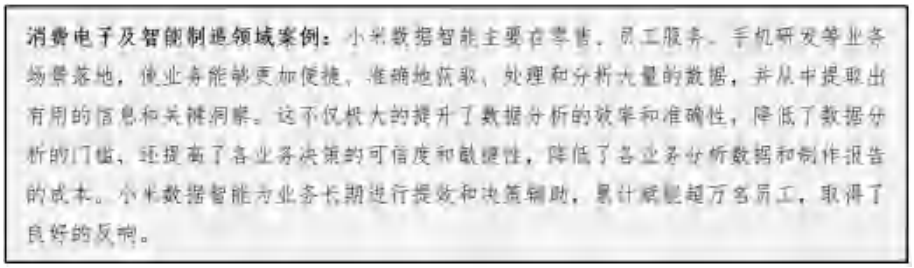
二是大模型技术将同知识库、知识图谱技术相结合，实现复杂的知识推理和智能问答。作为一种全新的知识载体，大模型和传统的知识库、知识图谱存在较强的互补性：一方面大模型具备海量通用知识，具备较强的多模态处理能力；另一方面传统的知识载体则在专业性、可解释性、可靠性方面具备显著的优势，两者结合可实现复杂的知识推理和智能问答，在工业制造、交通物流、国防军事等复杂场景中具备广阔的应用空间。基于专业知识库、知识图谱的检索增强生成（RAG）也成为大模型在细分场景落地的重要技术路线，在智能知识库、智能检索工具、智能客服等场景中具备较大潜力。

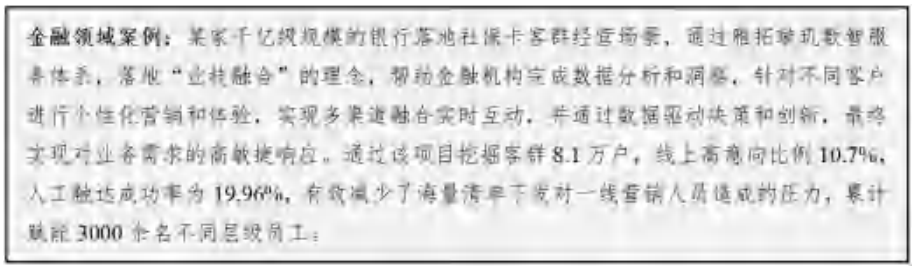


三是大模型赋能的智能助手将全方位提升企业研发设计、工作协同和业务流程管理能力。随着多模态大模型技术的发展，其在文案、图片、代码、音视频等内容生成领域能力正在快速提升，基于大模型的智能助手在研发设计领域的应用具备广阔的发展空间。工作协同方面，智能助手可集成企业内部的各种信息源，包括文档、邮件、会议记录等，帮助团队成员快速获取所需信息。业务流程管理方面，智能助手可以自动追踪项目进度，提醒团队成员任务截止日期，协助进行企业资源分配和调度，审查工作方案并生成业务情况报告。

工业制造领域案例：在某工业制造型企业，大模型通过充分融合企业已有的运行规程、检修方案、工艺流程等知识，形成了该企业独有的行业大模型。该模型以数字专家形式存在，并通过自然语言对话方式帮助专业岗位人员解答专业问题、指导运维操作、审查工作方案、生成经营报告、辅助经营决策，从而提高专业岗位的工作效率与业务创新能力，最终实现提升知识利用效率两倍以上，提升专业人员的创新效率 30%以上，提升企业经营管理决策效率 30%以上。

场景范畴来看，各场景数据智能能力建设重点存在差异，但敏捷化、普惠化是普遍关注的重点方向。传统的数据智能应用依托专业的数据分析工具及人工智能算法，主要以大屏、报表、领导驾驶舱等形式服务于企业管理层，赋能于长周期的重大决策。随着数据智能技术的逐步成熟，市场竞争愈发激烈，数据智能的场景应用也更加侧重于敏捷化和普惠化。**数据智能能力建设将更加侧重于对业务需求的敏捷响应。**通过数据智能技术可对企业运作所需的各种资源、流程和活动进行统筹管理，帮助企业实现更加灵活和敏捷的运营策略。在营销、风控等业务场景中，大量的决策已通过人工智能和机器学习算法自动化完成，对市场的变化进行快速响应。**数据智能能力建设将更加侧重于赋能更多一线业务人员。**通过赋能一线人员可以提升其直接产出，长期来看也有助于提升企业数据驱动文化。大模型的落地极大地提升了数据智能工具的交互性，赋能业务终端的对话式数据分析工具、智能报表工具的产品化程度逐步成熟，数据智能应用门槛不断降低。这些赋能一线的数据智能应用正在成为业务侧数据智能能力建设的重点。


消费电子及智能制造领域案例：小米数据智能主要在零售、员工服务、手机研发等业务场景落地，使业务能够更加便捷、准确地获取、处理和分析大量的数据，并从中提取出有用的信息和关键洞察。这不仅极大的提升了数据分析的效率和准确性，降低了数据分析的门槛，还提高了各业务决策的可信度和敏捷性，降低了各业务分析数据和制作报告的成本。小米数据智能为业务长期进行提效和决策辅助，累计赋能超万名员工，取得了良好的反响。


金融领域案例：某家千亿规模的银行落地社保卡客群经营场景，通过雅拓聚机数智服务体系，落地“业技融合”的理念，帮助金融机构完成数据分析和洞察，针对不同客户进行个性化营销和体验，实现多渠道融合实时互动，并通过数据驱动决策和创新，最终实现对业务需求的高敏捷响应。通过该项目挖掘客群 8.1 万户，线上高意向比例 10.7%，人工触达成功率为 19.96%，有效减少了海量清单下发对一线营销人员造成的压力，累计赋能 3000 余名不同层级员工。

行业范畴来看，数据智能应用在加速赋能传统行业的同时也会出现更多的跨行业赋能。随着各行业对数据智能能力建设的持续推进，数据智能应用在深度和广度上都将有极大提升。**深度方面，数据智能应用将更加深入的赋能农业、交通、制造业等传统行业。**这些行业涉及到大量来自物理世界的数据，数据采集成本高，普遍以多模态数据为主。随着传感器、5G、卫星遥感技术的不断发展，数据采集成本稳步降低，大模型的落地也为多模态数据分析和处理提供了新的解决思路，传统行业的数据智能应用能力将快速提升。**广度方面，跨行业的数据智能应用将更加普遍。**数据和场景是数据智能应用的两大关键因素，电信、文娱、电商、气象等行业数据价值高、通用性强，金融、传统零售、交通等行业则具备大量的应用场景。联合建模、隐私计算等技术为跨行业数据智能应用提供了安全可信的解决方案，《“数据要素 ×”三年行动计划》等政策为数据智能的跨行业应用打下了良好的政策基础，必将推动跨行业数据智能应用的快速发展。

工业制造领域案例：某企业制造场景生产流程长、工序多，工序之间衔接响应时间长，很多环节仍需人工干预。此外，涉及到大量多模态数据，存在汇聚难、管理弱、分析浅的问题。大华数据智能通过多维融合感知复杂的物理世界，同时加持大模型技术也为多模态数据分析和处理提供了新的解决思路，更加深入地揭示了数智实体的事实与规律，使得该企业数据智能应用能力快速提升，ATP 交期、欠料运算计算速度提升 5 倍，赋能超 1000 名员工，每年节省人力成本约 1550 万元。

交通领域案例：受居民出行结构变化的影响，某市面临着城市公交转型的问题，需对公交线路进行重新规划。通过使用运营商数据分析分析人群碎片化行程轨迹，构建出行需求模型通，评估公交线路上的各维度指标，如负荷率、OD 覆盖度等，对现网运行情况进行定量分析。找出当前公交网络与人群需求不匹配的路段，并以此为依据，使用网络最短路径算法规划出多条高匹配度的定制公交路线。通过运营商数据的跨域赋能，该市实现了对公交网络的科学规划，完成了公共交通数据智化升级，公共交通承载量增加了 28%，取得了良好的社会反响 。

电商领域案例：某零售品牌进行了全域品牌 CRM 与天猫旗舰店数据资产联合分析，实现了广告引擎、广告主、DMP 等多方的数据协作。通过使用阿里妈妈营销隐私计算平台，该品牌在保障多方数据隐私安全和数据合规使用的基础上，解决了广告投放链路中数据处理、洞察分析、投放优化、归因衡量等多方数据联合分析和计算的问题，实现了投入产出比提升 158%的生意增长 。

四、 数据智能产业生态

近年来，全球数据智能产业快速发展，产业规模持续扩大，逐步形成了覆盖数据智能基础设施、资源服务、数据治理、开发应用和生态服务等多维度的完整产业链格局。本章从数据智能产业全景出发，对比国内外数据智能产业现状，围绕数据智能产业图谱分析各层级产业发展及生态建设情况，并聚焦数据智能产业突出问题，立足客观分析结果，展望数据智能产业与生态前景。

(一) 数据智能全景化布局提速，产业体系逐步完善

根据中国通信标准化协会大数据技术标准推进委员会（CCSA TC601）发布的《数据智能产业图谱（2024 年）》显示，我国数据智能产业布局正逐步完善，已围绕数据智能基础设施、数据治理、数据资源服务、数据智能开发、数据智能应用及数据智能生态服务等各方面，初步形成了健康有序、优势互补的产业体系。

数据智能基础设施企业主要围绕通用计算硬件、智能计算硬件、数据存储、数据智能平台等提供产品服务，为上层数据开发和模型训练提供算力支撑。

数据治理企业主要围绕数据质量管理、数据标准管理、数据资产管理、数据标准等方面提供平台及服务，支撑数据管理相关工作。

数据资源服务企业主要围绕人工智能数据集、数据产品开发、数据运营、数据交易等提供平台及服务，为数据产品/服务的开发及应用提供数据基础。


数据智能产业图谱(2024年)
数据智能应用企业
数据智能开发企业
数据资源服务企业
数据治理企业
数据智能基础设施企业

图 7 数据智能产业图谱

数据智能开发企业主要围绕数据开发、人工智能算法开发、数据科学、知识图谱等提供平台能力，支撑数据智能产品及应用开发。

数据智能应用企业围绕生成式大模型、通用应用和场景化应用形成了多项产品及服务，覆盖了金融、通信、汽车、能源等多个行业，持续赋能企业数智化转型。

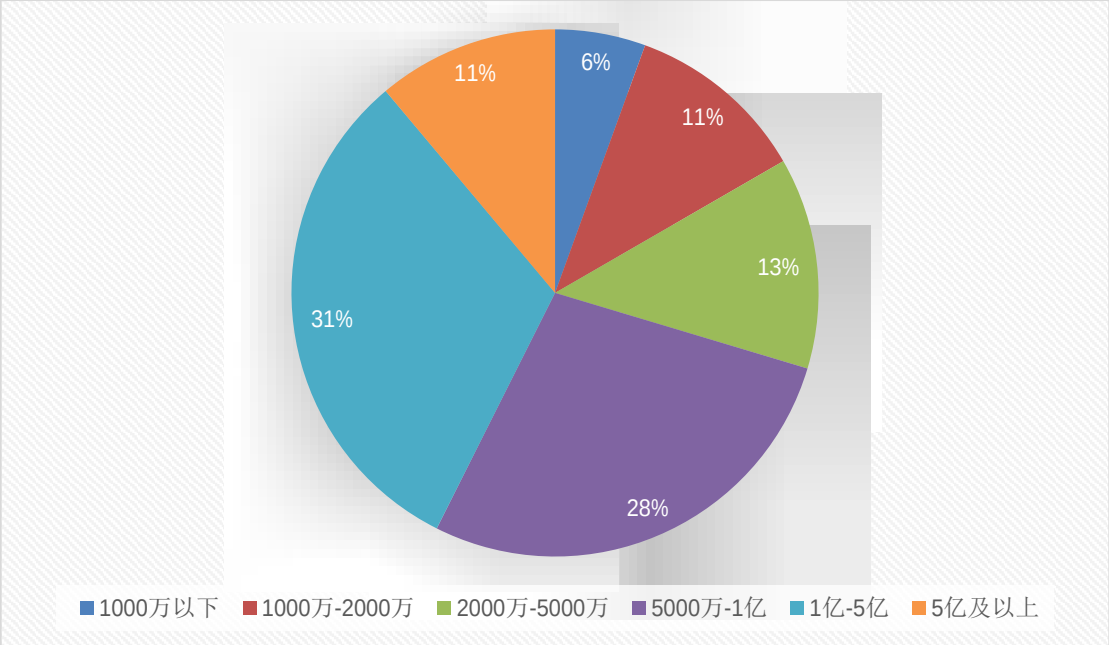


图 8 数据智能企业营收分布情况

从企业发展情况来看，入选产业图谱的企业大多数成立于 2010 年之后，占比高达 68%；规模上以大中型企业为主，年营业收入在亿元以上的企业占比超过 40%，是数据智能产业发展的中流砥柱。**从企业研发能力来看，**入选产业图谱的企业平均专利授权数达到 31.7 项，软著登记数达到 77.2 项，拥有 10 项以上专利、30 项以上软著的企业占比超过 50%，自主知识产权体系稳步构建。**从企业人才培养来看，**研发人员数量超过 30%的企业占比高达 95%，其中近 10%的企业研发人员数量超过 80%；人才层次方面，平均本科学历以上比例达到

92%，平均研究生学历以上比例超过 25%，工作经验 5 年以上的人才占比超过 70%，企业人才体系日趋完善。

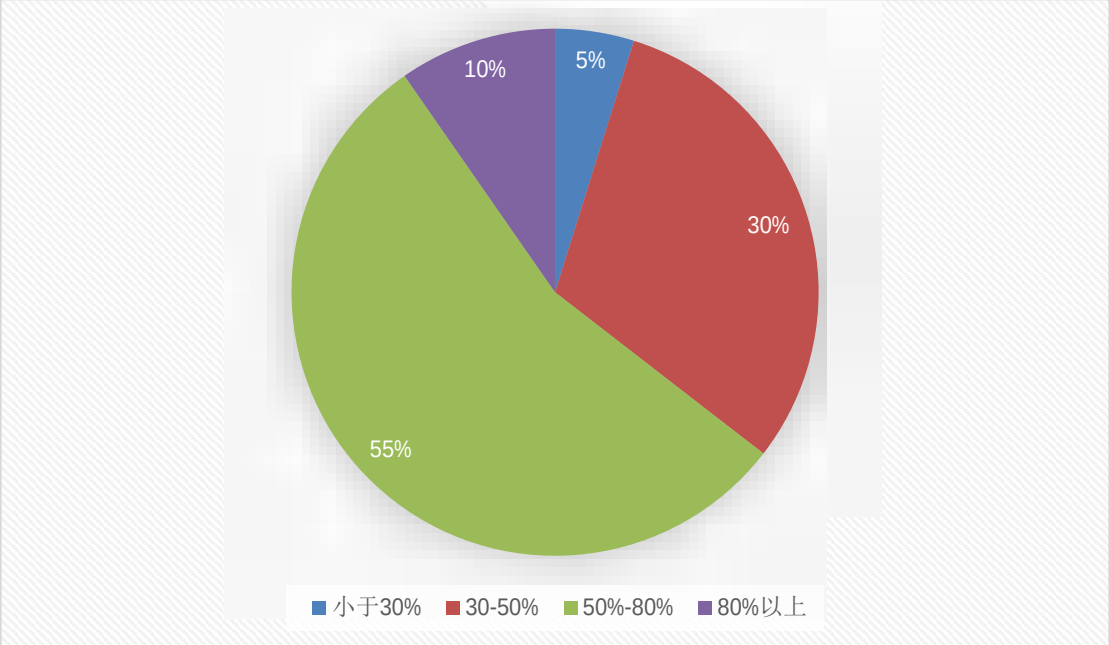


图 9 数据智能企业研发人员数量占比情况

(二) 全球数据智能产业快速发展，规模化效应初显

政策方面，数据智能的快速发展促使全球多国综合考虑发展与监管，加快推动相关政策出台。2023 年 3 月，加拿大政府宣布将制定《人工智能与数据法案》，并发布预告性配套文件，以规范人工智能开发和应用活动。2023 年 11 月，美国网络及信息技术研究与发展小组委员会发布《美国人工智能 2024 财年预算和计划信息》，提出建设用于人工智能训练、测试、开发的共享公共数据集和环境，在机器学习、智能制造、医疗等领域创建训练数据集，旨在确保自身在数据智能领域的全球领导地位。2024 年 3 月 13 日，欧洲议会通过世界首部《人工智能法案》，该法案旨在通过对人工智能的使用进行规制，以保护公民基本权利、民主、法治和生态环境免受人工智能的影响，促

进以人为本、值得信赖的人工智能发展应用。

企业方面，全球数据智能领域相关企业加速涌现，企业数量增长态势明显。据中国信通院统计，截至 2024 年 4 月，全球共有人工智能企业 30000 余家，我国人工智能企业数量超过 4500 家。其中超半数人工智能企业从事涉及数据的相关业务，包括数据智能类软硬件开发、人工智能数据治理、人工智能数据集开发、数据智能产品开发等。同时，据不完全统计，国内有近 2000 家数据企业面向人工智能领域提供服务，如专注于多模态学习的数据标注业务，根据特定领域需求场景提供定制化的数据加工方案与实施服务，提供标准化结构化的高精度数据集等。未来，随着大模型技术在各行业领域的落地应用，数据智能企业数量将持续增长，支撑产业规模持续扩大。

投融资方面，数据智能企业投融资活跃，占整体人工智能产业投资比重不断增加，投资带动效果显著。最近一年，随着大模型技术应用快速发展，面向大模型的数据智能企业投融资规模不断扩大，据统计，2023 年前三季度大模型企业融资金额同比增长 137%，融资轮次以种子轮、天使轮、A 轮融资等早期投资为主。资本市场对于数据智能产业的良好预期，对于产业新增长点的投资带动作用非常显著，推动了以 Character.ai、Scale AI、月之暗面等为代表的一批数据智能领域明星创业企业的涌现。面对数据智能应用在资本市场和用户市场的火热态势，国内外科技巨头也纷纷入场，国外以微软、谷歌、Meta 等为代表，国内以百度、阿里、字节跳动等为代表，一方面积极研发企业级数据智能相关产品，另一方面以开放接口等方式与其他企业合作

开发应用，加速构建数据智能产业生态。

人才方面，美中两国成为全球数据智能人才聚集高地，人才培养体系正加速构建。随着各国在数据智能领域竞争的日益激烈，人才正成为推动产业生态发展、抢占国际领先地位的核心资源，不管是企业端还是高校机构，对人工智能人才的重视度都在不断增加。据美国保尔森基金会发布的《全球人工智能人才追踪调查报告 2.0》显示，在人工智能人才的培养和引进方面，美国仍保持压倒性的优势，而中国正在迎头赶上。2022 年，美国机构中来自中美的顶级人工智能人才占比高达 75%，中国在过去几年内持续构建人工智能人才库，以满足不断增长的人工智能产业需求。截至目前，我国已有近 500 所高校院所开设人工智能专业，着力提升人工智能基础研究和交叉应用能力，通过校企合作促进科技创新和产业发展的有效衔接，培养跨学科、复合型、高层次、创新型的人工智能高端人才。

开源社区方面，全球数据智能领域开源社区快速发展，开源项目在创新和效率上展现出巨大潜力。国外方面，开源社区发展呈现出活跃和多元化的特点。截至目前，全球开源人工智能项目和贡献者数量显著增长，GitHub 上的贡献者数量同比增长 148%，项目总数同比增长 248%。2023 年 12 月，IBM 和 Meta 联合全球 50 多个创始成员和协作者宣布成立人工智能联盟（AI Alliance），该联盟支持开放式创新和开放科学，致力于培育一个开放的社区，加速负责任的人工智能创新。此外，Mistral、Vicuna、Yi、Llama 等开源模型快速发展，效率、生态等优势逐步显现，与 GPT-4、Claude 等闭源模型性能差距逐步缩

小。**国内方面，**开源社区正在积极发展中，在技术研发、创新培育和产业应用方面展现出潜力，但与国外相较差距明显。大模型方面，清华大学发布的 ChatGLM-6B、智源人工智能研究院推出的悟道 ·天鹰、阿里推出的通义千问 Qwen 等已得到了广泛探索和应用，在模型参数量、训练稳定性、性能等方面持续优化。高质量开源数据集方面，阿里推出的天池开源数据集、百度推出的飞桨开源数据集、北京市发布的人工智能大模型高质量数据集等为更多深度学习模型的建立和优化提供了重要基础，持续驱动数据智能产业发展。

(三) 数据智能产业挑战与机遇并存

1. 数据智能产业面临多重挑战

供需结构不平衡：数据智能核心技术自主创新不足，导致基础技术服务产能过剩、前沿创新服务有所欠缺，产业结构供需匹配不平衡。相关企业在应用层发展较为迅速，但在关键技术的研发和产业化能力，以及业务模式的创新探索能力方面存在不足，产业链上下游缺乏具有国际竞争力的核心技术和产品。与此同时，各方对于数据智能的产业前景预期需回归理性，在不低估产业潜力、不高估产业价值的前提下，统筹考虑并优化前沿创新类产业投入，避免陷入盲目内卷和恶性竞争的困局。

技术工具体系庞杂：一是兼容性问题，比如不同数据库系统、大数据处理框架、平台与技术工具之间难以兼容，导致数据难以跨平台高效流转，技术开发成本较高。二是标准化和规范化程度不高，数据智能技术应用缺乏统一的标准体系，不仅影响企业的互操作性和兼容

性，也阻碍产业整体的规范化和规模化发展。三是部分技术难点堵点问题尚未解决，例如技术路线选型缺乏统一标准、数据安全和隐私保护能力有待提升、性能和稳定性仍需持续优化等。

数智化转型方法论缺失：一是知识体系梳理不足，企业缺乏对数据智能领域知识的全面理解，难以系统性整合和应用这些知识，无法快速构建完整的知识体系。二是缺乏实施路径引导，由于缺乏明确的数智化转型路径和阶段性目标，企业在规划实施过程中难以设定短期和长期目标、评估转型进度和效果。三是业务场景挖掘不深，对业务流程的可优化性缺乏了解，难以将数据智能技术与具体业务需求相结合，缺少识别和挖掘有价值业务场景的能力。四是组织架构规划不合理，企业缺乏适应数据智能发展趋势的组织架构，需要调整但因担心影响现有业务的稳定性无从下手。

生态体系不完善：一是数据智能专业人才不足。据不完全统计，目前国内人工智能领域人才总缺口达 500 万，数据智能领域高层次领军人才稀缺，基础研究人才流失等问题较为突出，人才结构分布与供需发展不均衡，急需构建健全完善的多层次人才培养及引进体系。二是开源社区体量较小，高质量开源项目缺乏。尽管国内数据智能开源社区正快速成长，但与国际上成熟的开源社区相比，由于技术、资源、开发者等方面的不足，国内有影响力、高质量的开源社区及项目相对较少。同时由于缺少可持续运营模式，保持开源社区生态稳定性和长期发展仍存在较大挑战。

2. 数据智能产业发展前景广阔

“十四五”时期，随着人工智能、大数据、物联网等基础技术的渗透应用，数据智能技术、应用及服务能力将加速成熟，并孕育产生庞大的市场需求，产业正迎来前所未有的发展机遇，将催生技术能力和应用场景的不断创新，推动生态加速培育与产业发展。

一是由技术导向走向服务导向。初期，企业更多关注于数据智能的技术能力和基础设施建设，如数据仓库、计算平台等。未来，企业将更加注重面向场景的服务建设，即如何利用技术产生有价值的产品和决策支持。例如通过机器学习和人工智能技术，从大量数据中提取深层次的业务洞察，形成数据产品，如用户画像、归因分析、市场趋势分析等，为业务决策提供直接支持。

二是由单一数据管理走向多模态数据整合。当前，综合语音、图像、视频、文本的多模态处理能力正成为数据智能进化的关键趋势。传统的数据平台和数据仓库相对分离，未来将更加趋向于一体化的解决方案，依托数据中台、知识中台、智能中台等新型平台，简化数据管理流程，提高数据开发效率。

三是由资源导向走向需求导向。目前，企业在数据智能建设上更多依赖于资源的投入，如硬件设备、人力资源等。未来，将更多地从业务需求出发，以需求为导向进行数据智能的建设和应用。根据业务目标和用户需求，定制化开发数据智能应用，如个性化推荐系统、智能客服、智能投顾等，以满足市场和用户的特定需求。

四是由数据驱动走向场景驱动。数据智能应用更多强调数据的作

用，未来将更加注重场景的驱动，即根据不同的应用场景设计和优化数据智能解决方案，推出面向多场景的组合式应用。例如在智能制造领域，企业将根据生产线、供应链、销售等不同场景的特点，设计相应的数据分析和优化模型，以提高生产效率和市场响应速度。

五是由经验实施走向算法替代。传统企业在数据智能项目实施上，更多依赖于专家的经验和直觉。未来将更多地利用算法和自动化工具，减少对个人经验的依赖，提高项目的可复制性和可扩展性。以大模型为例，随着底层硬件技术的迭代更新，预训练大模型也将迎来新的突破，在模型逻辑理解能力、自适应学习能力等方面跃上新的台阶。

五、 总结与展望

经过数十年的发展，“数据”和“智能”两大领域已形成相对成熟的技术体系及产业链。当前，随着数据成为生产要素，数据的价值属性深刻植入从业者认知；随着生成式大语言模型屡屡产出突破性成果，人工智能对未来的革新潜力引人遐想。技术和认知同时的嬗变伴生经济增长承压的环境因素，共同塑造出了数据智能的美好愿景。通过应用智能化技术释放数据要素价值，实现企业的数据智能化改造，为企业带来新的竞争力，为经济带来新活力，为社会带来新气象，成为了各层面共同的期许，相应的数据智能在概念、技术体系、应用实践、产业生态等方面的共识也正在逐渐形成。

立足当下，极目远望。一个堪比信息时代的全新时代或许正在到来，在“数据”与“智能”的交相融合下，新生的数据智能寄宿着信息时代最为精华的馈赠，传统大数据、人工智能等技术产业也将由此焕发生机。在新方向的指引下，未来技术的演化将再度衔枚疾进；在新趋势的涌动中，变革后的产业经淬火亦历久弥新。作为新质生产力的重要驱动力，数据智能将进一步的解放个体生产力，革新企业价值观，重塑生产关系和产业格局，推动全社会实现跨越式进步。新时代的起点注定将由胸怀理想的开创者登临，愿行者无疆，筚路蓝缕，以启山林。

大数据技术标准推进委员会

地址： 北京市海淀区花园北路 52 号

邮编： 100191

邮箱： TC601@CCSA.org.cn

网址： www.tc601.com