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：

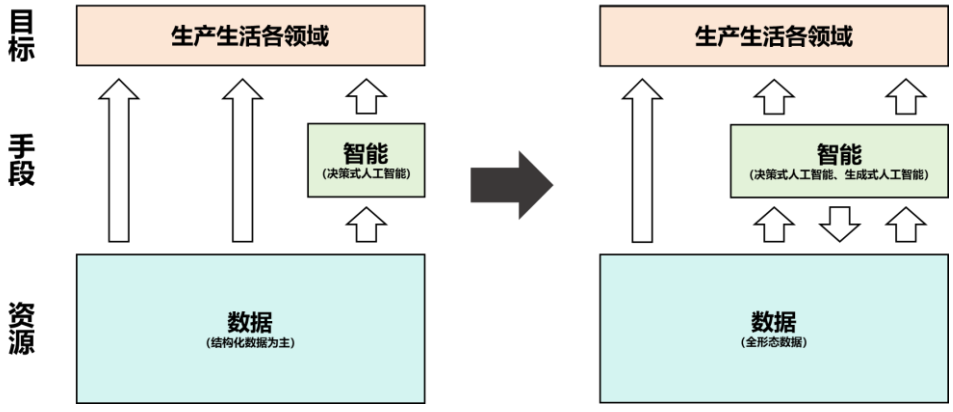

图 1 数据和智能间关系的变化

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。

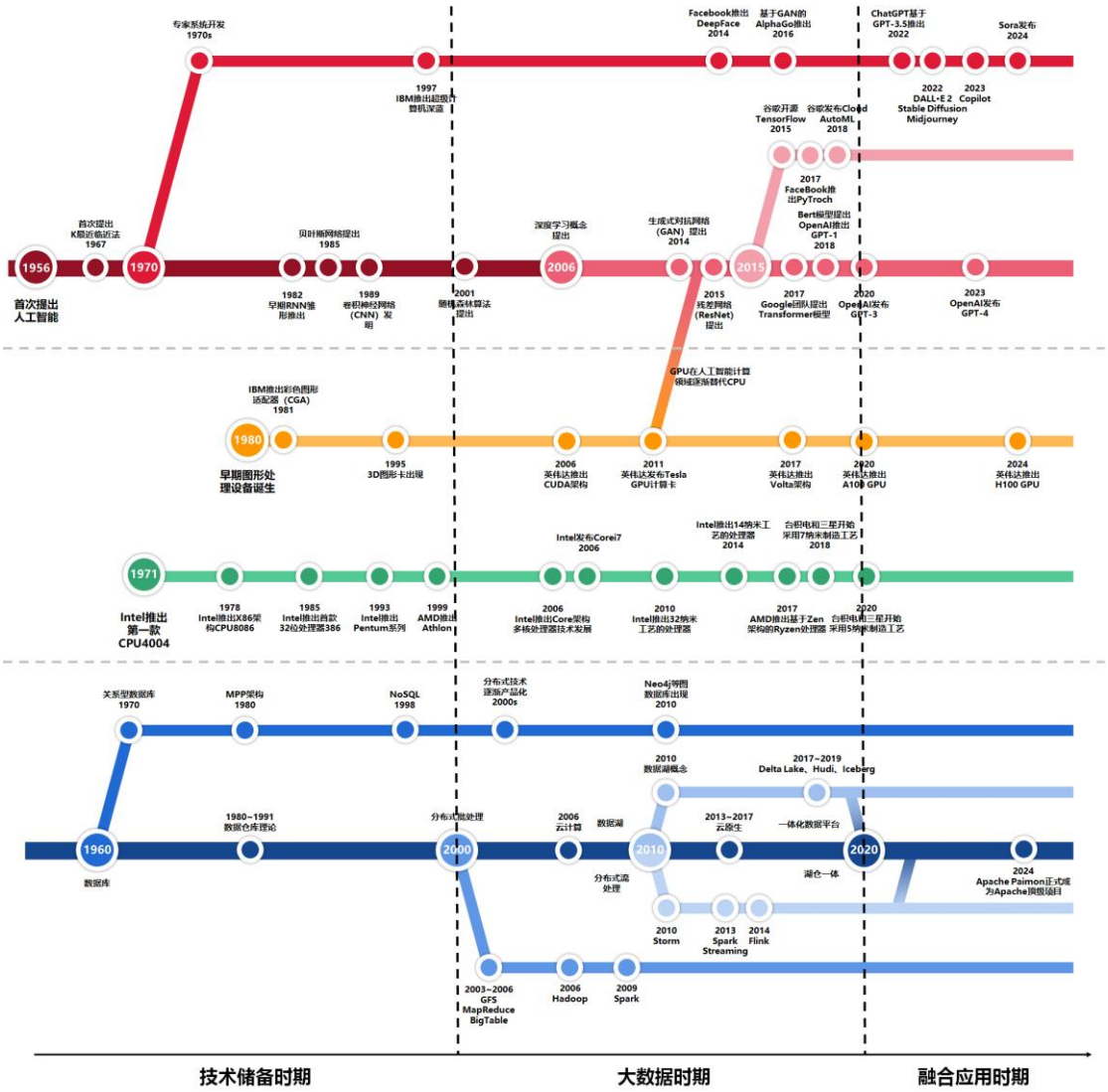

图 2 数据智能发展脉络

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。

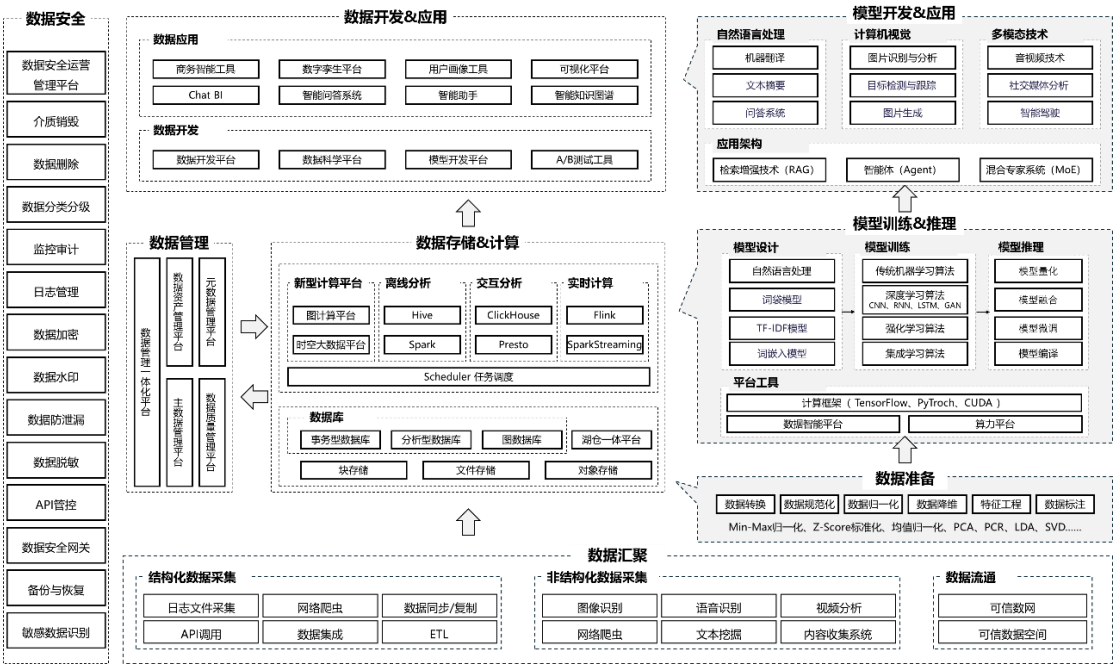

图 3 数据智能技术体系概览

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。

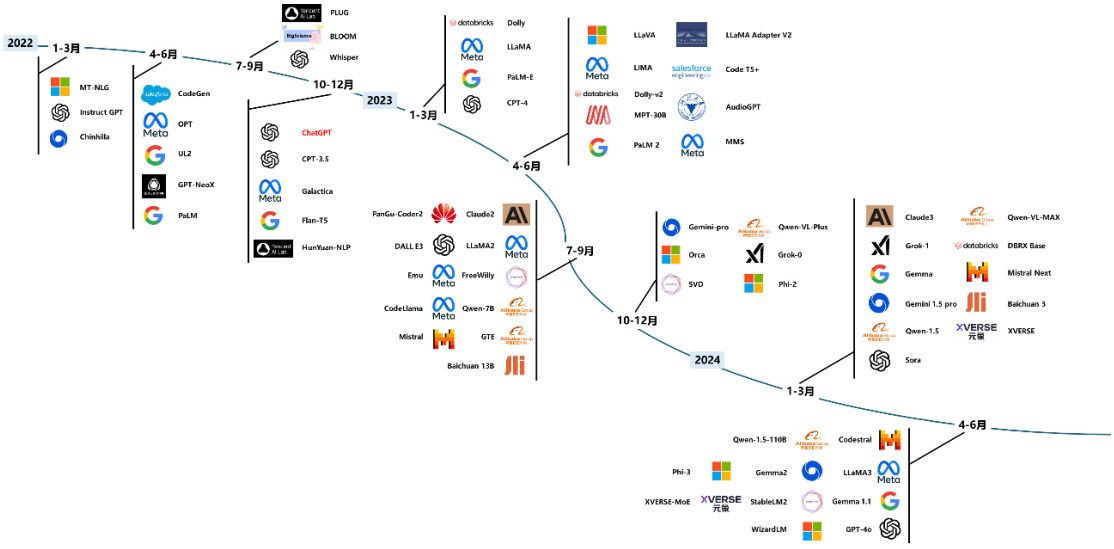

图 4 部分生成式大模型发布情况统计

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”。