

新型人工智能存储研究报告

(2025 年)

中国信息通信研究院人工智能研究所

中国人工智能产业发展联盟

2025年4月

版权声明

本报告版权属于中国信息通信研究院、中国人工智能产业发展联盟，并受法律保护。转载、摘编或利用其它方式使用本报告文字或者观点的，应注明“来源：中国信息通信研究院、中国人工智能产业发展联盟”。违反上述声明者，编者将追究其相关法律责任。

前 言

2022 年科技部等六部门印发《关于加快场景创新以人工智能高水平应用促进经济高质量发展的指导意见》，提出加速人工智能技术攻关、产品开发和产业培育，探索人工智能发展新模式新路径，以人工智能高水平应用促进经济高质量发展。2023 年工信部等六部门印发《算力基础设施高质量发展行动计划》，提出强化存力高效灵活保障，加速存力技术研发应用，持续提升存储产业能力，推动存算网协同发展。

大模型时代，数据决定人工智能的高度，更多的训练数据是大模型迭代升级的前提，更高的数据质量也决定着大模型训练的效果。当前，大模型技术全面推动底层基础设施的发展，算力需求不断攀升，海量数据的存储和处理需求不断增长，对人工智能存储的性能、扩展性、数据安全、数据范式等都提出更高要求。

本报告重点对新型人工智能存储的概念范围、面临挑战、关键技术和最佳实践进行了梳理和分析。概念范围上，梳理了新型人工智能存储的基本概念，分析了全球人工智能存储战略。面临挑战上，点明了新型人工智能存储是发力大模型的基础，但同时海量数据归集、训练数据访问效率、推理实时性等方面存在诸多挑战。关键技术上，阐明了新型人工智能存储需要从存储介质、系统、架构、数据编织、数据范式和数据安全等方面发力。最佳实践上，介绍了医疗、金融、云服务商、AI 企业等在新型人工智能存储的实践案例。最后，本报告针对当前人工智能存储发展的挑战，提出了未来我国

新型人工智能存储发展建议。

新型人工智能存储相关产业、技术正处于高速发展阶段，新技术生态瞬息万变。报告仍有诸多不足，恳请各界批评指正。



目 录

一、 新型人工智能存储概况	1
(一) 新型人工智能存储基本概念	1
(二) 全球人工智能存储战略分析	2
二、 人工智能存储重要性	4
(一) 人工智能存储是发力大模型的基础	4
(二) 人工智能存储面临的挑战	7
三、 新型人工智能存储关键技术	9
(一) 长记忆存储范式为推理提质增效	11
(二) 数据编织加强全局数据高效处理	12
(三) 超节点存储架构持续支撑扩展法则	14
(四) 高性能并行文件系统以存强算	15
(五) 全闪存存储技术为数据提速降耗	16
(六) 存储内生安全保护企业数据资产	17
四、 人工智能存储建设最佳实践	18
(一) 医疗行业：长记忆存储助力智能辅助诊断	18
(二) 金融行业：高性能 AI 存储赋能智慧金融	20
(三) AI 企业：AI 数据湖支撑大模型训推平台	22
(四) 运营商：构筑领先智算中心云数据底座	23
(五) 制造业：一体化集成加速智能客服场景创新与实践	26
五、 人工智能存储发展建议	27
(一) 加强 AI 存储设施统筹部署	28
(二) 增强 AI 存储自主创新能力	28
(三) 推动 AI 存储产业协同发展	28
(四) 构建 AI 存储能力评价体系	29

图 目 录

图 1	AI 基础架构	2
图 2	数据是大模型的核心要素	6
图 3	算力中心架构到存储中心架构	9
图 4	新型 AI 存储特征	10
图 5	AI 存储结构分层	11
图 6	长记忆存储架构	12
图 7	统一数据底座	14
图 8	医疗行业 KV Cache 长记忆存储应用实践	19
图 9	金融高性能 AI 存储应用实践	21
图 10	外置 AI 数据湖存储底座应用实践	23
图 11	某运营商智能存储盘框应用实践	25
图 12	制造企业超融合一体机应用实践	27

表 目 录

表 1	国外人工智能存储战略布局	3
表 2	国内人工智能存储战略布局	4

一、新型人工智能存储概况

（一）新型人工智能存储基本概念

人工智能(Artificial Intelligence)起源于 20 世纪五六十年代，历经符号主义、连接主义和行为主义三次浪潮的相互交织发展，如今作为一项新兴的通用技术，正推动着社会生活与各行各业的巨变。随着大模型的爆发式增长和持续迭代，模型参数量从几万跃升至数千亿甚至更大，模型层数从开始的个位数逐步发展到成百上千，原始数据集也达到 PB 级，对以数据存储为代表的多领域产生了根本性的影响。

2024 年 2 月 16 日，OpenAI 再度引领风潮，发布了首款文生视频大模型 Sora，大模型由单模态的 NLP 领域向文生视频等多模态方向迅速演进，标志着 AI 技术正迈向一个全新的高峰。相较于传统的 NLP 语言大模型，视频大模型所需的训练数据量极为庞大，原始数据体量相较于 NLP 能达到几百甚至上千 PB 级以上，2025 年 2 月，Grok3 发布，其多模态能力获得全球瞩目，Grok 3 所用数据量比 Grok 2 增加了 3 倍，通过增加数据量而不是模型参数的方式实现了大模型能力的倍增。这些都对人工智能基础设施的数据存储和处理能力提出了极高的挑战。

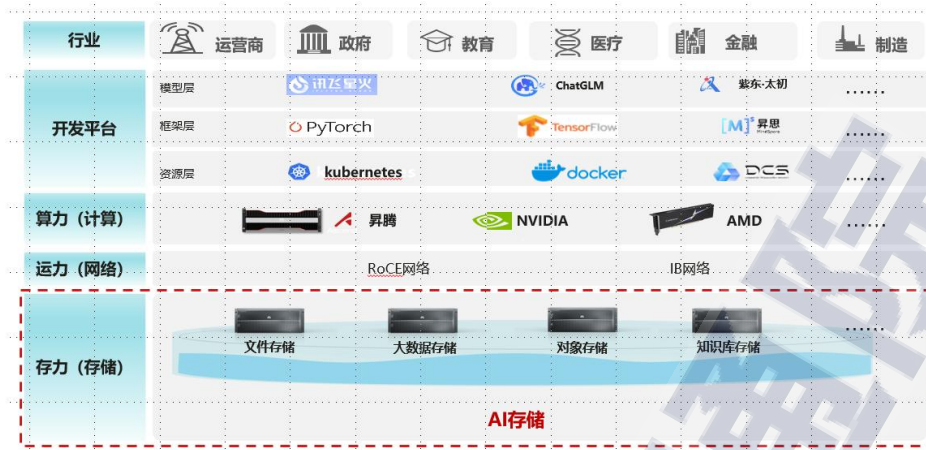


图 1 AI 基础架构

来源：中国信息通信研究院

新型 AI 存储是指专为人工智能应用和服务设计的数据存储系统，具备超高性能、超大容量、极致安全、数据编织等特征，可以有效支撑海量数据的分析和学习，是 AI 基础架构不可或缺的组成部分，如图 1。作为大模型的数据载体，新型 AI 存储与大模型数据归集、预处理、训练、推理等全生命周期流程紧密相关，且逐步成为 AI 数据全生命周期服务与管理平台的重要组成部分。当今在不同行业和组织内部具备形成多元化数据集的条件下，新型 AI 存储正在成为数据资源的统一管理平台，并以此为基础构成跨域跨部门的 AI 数据湖，为数据赋能提供关键的技术支撑。

（二）全球人工智能存储战略分析

国外较早布局人工智能战略，AI 存储逐步成为战略关键技术竞争力的重要体现。欧盟委员会在 2020 年 2 月发布《人工智能白皮书》，确保对计算和数据基础设施的访问安全，同时提出改善数据访问和管理是根本。英国政府在 2021 年 9 月发布《国家人工智能战略》，AI

系统中的数据基础和使用将提高专业人工智能、软件和数据技术的能力，并支持关键的新型数据基础设施的发展。美国国防部在 2023 年 11 月的《数据、分析和人工智能采用战略》中，提出推进数据、分析和人工智能生态系统，同时改善基础数据管理，为美国人工智能发展和相关国家安全战略制定提供依据，如表 1。

国内目前对数据存储的产业布局主要在于通用存储。2023 年 11 月，《关于加快场景创新以人工智能高水平应用促进经济高质量发展的指导意见》中提出加强人工智能场景创新要素供给。推动场景算力设施开放，集聚人工智能场景数据资源。2021 年 12 月，中央网信办在《“十四五”国家信息化规划》中提出加强数据收集、汇聚、存储、流通、应用等全生命周期的安全管理，建立健全相关技术保障措施。2020 年 7 月，国家标准化管理委员会等五部门在《国家新一代人工智能标准体系建设指南》中提出规范人工智能研发及应用等过程涉及到的数据存储、处理、分析等大数据相关支撑技术要素，包括大数据系统产品、数据共享开放、数据管理机制、数据治理等标准，如表 2。

表 1 国外人工智能存储战略布局

序号	时间	名称	主要内容	发布机构
1	2023 年 11 月	《数据、分析和人工智能采用战略》	投资可互操作的联合基础设施；推进数据、分析和人工智能生态系统；扩大数字人才管理；改善基础数据管理。	美国国防部
2	2022 年 11 月	《国防人工智能战略》	通过前沿技术枢纽，支撑新兴技术的使用和创新，从而支持创建新的英国国防人工智能中心。	英国国防部

3	2022年2月	《关键和新兴技术清单》	明确了数据存储作为关键技术，为美国技术竞争力和国家安全的战略方案制定提供依据。	美国国防部
4	2020年2月	《人工智能白皮书》	确保对计算和数据基础设施的访问安全，同时提出改善数据访问和管理是根本。	欧盟委员会

来源：中国信息通信研究院

表 2 国内人工智能存储战略布局

序号	时间	名称	主要内容	发布机构
1	2022年7月	《关于加快场景创新以人工智能高水平应用促进经济高质量发展的指导意见》	加强人工智能场景创新要素供给。推动场景算力设施开放，集聚人工智能场景数据资源。	科技部等六部门
2	2021年12月	《“十四五”国家信息化规划》	加强数据收集、汇聚、存储、流通、应用等全生命周期的安全管理，建立健全相关技术保障措施。	中央网信办
3	2020 年 7 月	《国家新一代人工智能标准体系建设指南》	在建设内容中提出规范人工智能研发及应用等过程涉及到的数据存储、处理、分析等大数据相关支撑技术要素，包括大数据 系统产品、数据共享开放、数据管理机制、数据治理等标准。	国家标准化管理委员会等五部门

来源：中国信息通信研究院

二、人工智能存储重要性

（一）人工智能存储是发力大模型的基础

大模型正在从基础大模型向行业大模型演进。大模型的应用场景日渐丰富，正在加速向各个行业渗透。大模型很大程度上打破了AI场景化和行业化的瓶颈,为上层应用提供更多通用性的基础能力，

同时企业不再需要按场景从零开始开发和训练独立的基础模型，仅需在较成熟的基础大模型上融入企业生产业务沉淀而来的私域数据，即可进一步实现大模型的专业训练，满足特定领域对准确度和安全性等方面的诉求。根据 IDC 的分析及预测，未来 95% 的大中型企业将基于专属数据自建行业大模型，例如银行的风控数据、车企自动驾驶数据、医疗 PACS 数据等。

数据决定人工智能高度。大模型三要素是数据、算力和算法。随着 AI 技术快速成熟，各企业所使用的算力已逐渐趋同；各企业采用的算法也同样逐渐收敛，大多依托 Transformer 模型基础架构和以 Pytorch、TensorFlow 等为代表的开发框架；因此真正体现大模型差异性的要素是数据。根据公开数据显示，Meta 开发的 LLaMA3 拥有 700 亿参数和 15TB 训练数据，而 OpenAI 的 GPT-3.5 拥有 1750 亿参数和 570GB 训练数据，尽管 LLaMA3 在参数规模上还不到 GPT-3.5 的一半，但其表现能力在大部分基准上均超过后者。不难看出，除了模型参数规模以外，训练数据的体量对提升大模型的效果具有显著作用。此外数据的质量同样重要，大模型需要规范性、完整性、时效性的高质量数据支撑。对于基础大模型，其数据质量主要基于厂商从公开渠道所获取数据的整体质量。而对于行业大模型的训练及细分场景推理应用，模型效果取决于行业专属的私域数据的质量，其中包含了企业原有数据和行业知识库等，如图 2。

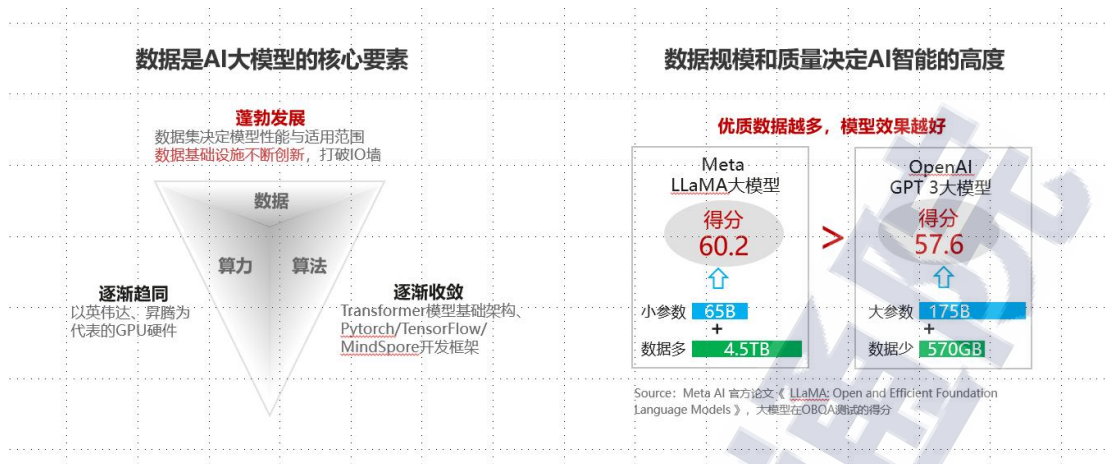


图 2 数据是大模型的核心要素

来源：中国信息通信研究院

AI 存储逐步成为大模型的关键基础设施。AI 存储是大模型数据收集、预处理、训练、推理的关键一环，决定了能保存利用的数据容量、训练及推理的数据存取效率、基础设施可用度以及数据安全。一是 AI 存储与大模型数据处理效率紧密相关，AI 存储直接影响数据访问速度，从而影响大模型训练和推理速度。二是 AI 存储是模型规模和数据量激增时平衡成本的重要因素。模型参数增大会伴随训练数据集规模的指数级增长，从而大幅增加数据储存成本。三是大模型行业落地需要依托 AI 存储来加速数据在各环节的自动流转，保护数据安全，并形成统一的数据管理。面对大模型带来的大量数据处理和工程化工作，也需要结合 AI 存储中的相关技术一并解决。

（二）人工智能存储面临的挑战

海量原始数据的归集对 AI 存储效率提出较高要求。人工智能数据在存储过程中往往需要通过不同地域、线上线下等方式对数据进行归集汇总。据统计，PB 级数据的归集通常花费 3~5 周，耗时占据整个大模型全流程时长的 30%。为了加速数据归集、减少后续分析的等待时间，需要 AI 存储具备高效汇聚、协议互通、海量按需扩容的能力。在 AI 存储领域，通过简单收集、爬取的原始数据是无法直接用于模型训练的，需要对数据进行解析、清洗、去重等工作，其中包含至少 3 次全量数据读取与搬迁，将消耗 30% 以上的 CPU、GPU、网络、内存等资源。

数据访问效率将成为制约模型训练速度的重要因素。在模型训练启动阶段，GPU 服务器会随机读取数万个小文件，读取完毕后才能启动训练。为了避免 GPU 等待训练数据的加载时间过长，需要 AI 存储提供千万级 IOPS 能力。此外，在模型训练的过程中，GPU 服务器硬件的故障率较高，在训练过程中一般会设定数十次、甚至上百次的周期性检查点（checkpoint）操作，保存中间过程数据，让发生故障后可以断点续训。因此为了减少 GPU 的空置时长，需要存储提供数百 GB/s 级的读写带宽。

AI 存储能够提供应用推理阶段的实时性和精准性。当大模型用于推理时，为避免大模型出现答非所问、内容杜撰等问题，需要将企业不断产生的私有数据联接到大模型。如果将这些全新的数据再次进行训练或微调，会耗费很长时间，且付出高昂的成本。通过

在大模型上外挂一个可容纳增量数据、并且实时动态更新的行业知识库，可以为大模型提供额外的信息输入，提高其回答的准确性，这就需要有一个能够快速检索关键信息的 AI 存储。

在 AI 存储全生命周期中需要考虑攻击带来的不良影响。通常攻击的手段归结为两类，一类是传统的数据窃取，以获取勒索赎金，企业不仅面临赎金损失，还会面临商誉、商业机会、法律诉讼、人力和时间成本等损失。而另一类是新型的数据攻击，主要是通过加入噪音数据，如在训练数据中加入暴力、意识形态歪曲的内容，导致模型质量下降、推理精确度失准、出现模型幻觉，最终干扰企业决策。因此在 AI 存储过程中需要对数据安全进行充分的考量。

AI 存储是提升 AI 集群可用度的关键。AI 集群可用度是指大模型基础设施集群在一定时间内提供正常服务的时间占总时间的比例，通常用百分比表示。数据显示，当前集群可用度普遍低于 50%。Meta 50000+卡训练任务¹和 OpenAI GPT-4 25000 卡训练任务集群算力可用度在 30%~40%之间，英伟达 Megatron-LM 和微软 MT-NLG 10000+卡训练任务的集群算力可用度在 40%~50%之间。AI 集群可用度提升困难导致可利用算力难以随着集群部署规模线性增长，造成了算力资源的巨大浪费。其中，算力等待时间是导致 AI 集群可用度降低的主要原因，例如，PB 级训练数据集的加载时间通常在小时级，断点或故障恢复所需的时间也会显著影响可用度。高性能的 AI 存储可减少算力等待时间，有效提升 AI 集群可用度。

¹Meta. "Building Meta's GenAI Infrastructure". <https://engineering.fb.com/2024/03/12/data-center-engineering/building-metas-genai-infrastructure/>.

三、新型人工智能存储关键技术

以数据为中心，统一存储基座，融合多元异构算力的新 AI 技术架构，已逐步成为人工智能计算中心的主流架构。多种异构算力紧密围绕在统一的数据底座，改变了过去“数据跟着算力跑”的算力烟囱工作模式，朝着“算力围着数据转”的新模式演进，如图 3 所示。作为数据载体，数据存储已成为构建大模型的关键基础设施之一。



图 3 算力中心架构到存储中心架构

来源：中国信息通信研究院

新型人工智能存储（面向 AI 的存储系统）具备极致性能、数据安全、大模型数据范式、高扩展性、数据编织和绿色节能 6 大关键特征，如图 4 所示，极致的性能密度可以加速数据供给，缩短训练数据的归集与预处理、CheckPoint 数据的保存以及断点续训等造成的算力等待，提升 AI 全流程效率；支持向量、张量、KV Cache 等大模型数据范式，能显著加速数据的融合检索，更好的支持推理应用；AI 存储 99.9999% 的可靠性提升 AI 基础设施的持续服务时间，防勒索和存储加密进一步保障数据安全；跨域跨集群的 AI 数据湖支撑异构计算，

近存计算的扩展；数据编织提供全局数据可视、可管，实现数据流动效率的倍数提升，并通过数据版本管理和数据血缘管理，保证数据质量不被破坏；绿色的 AI 存储，可降低每 TB 数据的能耗和占用空间。



图 4 新型 AI 存储特征

来源：中国信息通信研究院

构建先进的 AI 存储，要从存储介质、系统、架构、数据编织、数据范式和数据安全等多方面发力，协同提升大模型数据存储能力，如图 5 所示。



图 5 AI 存储结构分层

来源：中国信息通信研究院

（一）长记忆存储范式为推理提质增效

长记忆存储通过多级 KV Cache 助力大模型推理成本降低，用户体验增强。长记忆存储是专业记录大模型思考结果的存储，作为内存的扩展，以分级的方式实现月级、年级、乃至“终身”的记忆能力，让大模型推理具备慢思考能力，如图 6 所示。一方面，大模型推理过程中需处理长序列以获取准确的上下文信息、生成高质量输出，模型的计算成本和内存需求通常随序列长度的增加而显著增加，通过 KV Cache 缓存机制，可以有效降低模型长序列推理的内存占用和计算开销，实现有限硬件条件下的高效推理。另一方面，基于高性能长记忆存储技术构建的多级 KV Cache 缓存机制，可以保证 KV Cache 具备随时在线和全局池化共享能力，配合以查代算算法，实现从持久化的 KV Cache “长记忆”中调取前期已执行过的计算结果，减少推理过程中的重复计算。根据企业实践，利用该技术可实现推理吞吐提速超

50%，显著降低推理的端到端成本，提升大模型行业应用的长序列场景体验。

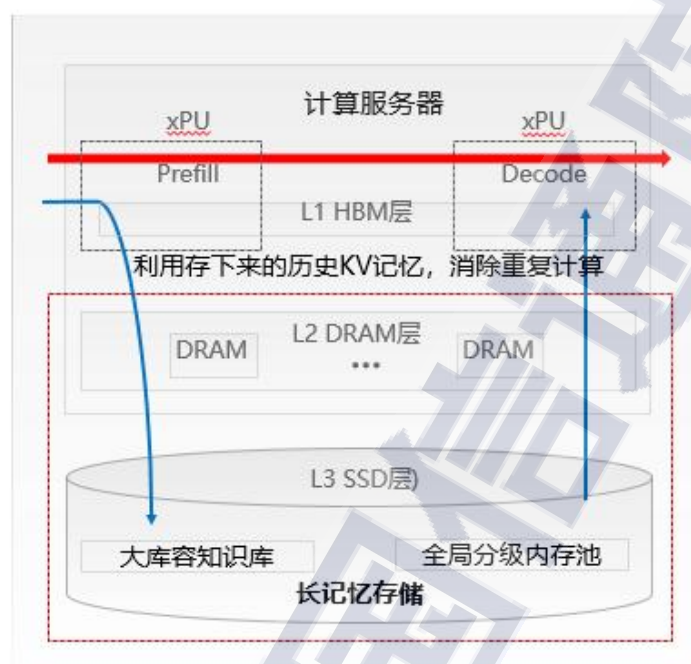


图 6 长记忆存储架构

来源：中国信息通信研究院

近数据向量知识库提高大模型检索效率，减少输出幻觉。近数据向量知识库基于快速知识生成、大库容高召回率与多模融合检索关键能力，可实现百亿知识库秒级检索。通过分布式合并建图技术，实现近数建库，根据企业实践，知识生成从月级降至天级，建库时长缩短 5 倍，实现知识实时更新。同时利用存储侧容灾备份特性组合，可提供数据库高可用保障，消除单点故障引发重新建库的巨大开销。

（二）数据编织加强全局数据高效处理

数据编织成为跨域跨集群数据归集、流动和处理的强大助力，通过数据版本管理和数据血缘管理，保证数据质量不下降。数据编织是

通过构建统一的数据视图，实现跨域跨集群数据的整合和调度，从而使得数据能够被快速地发现 and 访问，保障全局数据可视可管。随着数据量的爆炸性增长，海量数据往往分散在不同的数据中心，形成所谓的“数据孤岛”，这不仅阻碍了数据的流通和共享，也降低了数据的利用效率。数据编织一方面能将来自多个源头的价值数据快速归集和流动，使数据高效的从源端移动到训练和推理环境的 AI 存储上来，并通过数据版本管理和数据血缘管理，保证数据质量不被破坏，满足权限、安全等方面的各类数据使用规则，提升海量复杂数据的管理效率，直接减少 AI 训练推理端到端周期，另一方面也能实现 AI 训练数据集的按需筛选。通过数据画像，凭借数据的时空信息、数据的标签，以简化数据的分级分类管理，做到按场景化的数据治理，满足大模型的场景化要求。

AI 全流程业务加速需要数据易共享、高性能、易扩展的统一数据底座。如图 7 所示。统一数据底座是指对内兼容 AI 全流程工具链所需的多种存储协议，并保证各协议的语义无损，对外呈现统一的命名空间，方便用户管理和访问数据，能够实现数据全生命周期管理。传统解决方案由于需要在不同系统间来回拷贝数据，会严重影响数据处理效率，浪费存储空间，增加运维难度。统一数据管理能够具备高效的数据流转能力，可以在不同阶段实现数据的零拷贝和格式的零转换，实现 AI 各阶段协同业务的无缝对接，减少等待时间。首先，可以指定数据首次写入时的放置策略，例如在数据获取阶段，新获取的数据需要在短时间内处理的，可以直接放置到高性能层；而新获取的

数据在短时间内无需处理的或用来长期归档的数据，则可以直接写入容量层；其次，可以设置丰富的数据分级流动策略，例如可以设置访问频度与时间相结合的流动策略，也可以设置容量水位触发的流动策略；再者，根据用户制定的分级策略，数据能够在高性能层和大容量层之间自动分级流动，数据分级迁移过程对业务应用完全透明；最后，对于已经分级到容量层的数据，用户可以通过命令或 API 对指定的数据集配置预热策略，来加速任务的启动速度。

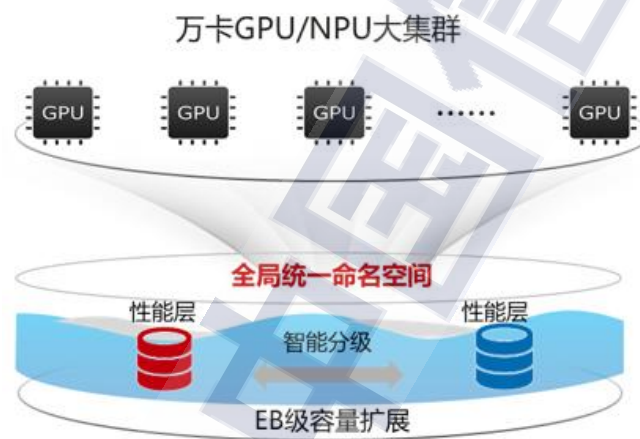


图 7 统一数据底座

来源：中国信息通信研究院

（三）超节点存储架构持续支撑扩展法则

大模型时代，AI 存储架构从传统的层次化、主从服务器架构，向全对等、全直通的超节点架构转变。超节点存储能够构建全对等、全直通的架构，实现“DC as a Computer”。存储与计算之间高并行数据直通，通过开放的协议标准，既支持 Scale-up 扩展容量，也支持 Scale-out 扩展性能的超节点架构。一方面在宏观上存算分离，计算、

存储资源独立部署，通过高通量数据总线互联，统一内存语义访问数据，实现计算、存储资源解耦灵活调度，资源利用率最大化。目前，模型训练仍然需要通过 CPU 去访问内存，但由于 CPU 的发展逐步放缓，导致内存的带宽和容量成为瓶颈。以 CXL(Compute Express Link) 为代表的高速互联总线，将系统中的计算、存储、内存等资源彻底解构，各自形成统一的共享资源池，让 GPU 可以直接通过 CXL 总线以更快的速度访问内存与存储，从而极大提升大模型的数据加载及流转效率，实现以数据为中心的超节点架构。另一方面在微观上存算一体，通过近数据处理，减少数据非必要移动，在数据产生的边缘、数据流动的网络中、数据存储系统中布置专用数据处理算力，算网存融合提升数据处理效率。大模型训练与应用阶段对数据处理效率有极高要求，在传统的数据处理流程中，数据需要在存储介质和处理器之间频繁移动，这不仅耗时而且效率较低。通过近存计算将数据预处理功能卸载到存储设备中，从而减少数据搬运的开销，提高数据处理的速度和效率。基于超节点存储能力，AI 存储能够支撑超十万卡算力集群，为人工智能的持续发展提供动力。

（四）高性能并行文件系统以存强算

大模型需要 AI 存储具备高性能并行文件系统以提高数据存取效率，并提供加速卡直通存储技术实现数据直达。并行文件系统是指支持数据并行读写的高性能文件系统，被广泛使用于需要高算力与高数据量的应用。随着大模型的发展，需要 AI 存储具备高并发、高带宽、高 IOPS 的并行文件系统，以提高 GPU/NPU 的训练推理效率。AI 存

储并行文件系统需具备以下四大能力，一是高性能，百 PB/s 级带宽和亿级 IOPS 支持能力，可实现超万卡集群数据调度简化、供给无瓶颈，保障规模扩展场景下的系统性能。每 U 大于 50GB/s 带宽、百万 IOPS 和 PB 级容量，可实现数据存储与处理的高密度，提升大模型使用效率，降低成本。二是高可靠，通过多路径、全交换、全互联等设计，确存储系统的高可用性达到 99.9999%，可以实现软硬件故障的自动切换。三是高扩展，能够支持 EB 级的数据量，且性能随节点数增加而线性增加。四是加速卡直通存储技术，利用基于总线 P2P 的底层传输协议机制，使数据路径无需再经过 CPU，实现加速卡的 HBM 和存储设备间数据一跳直达，消除 CPU 处理瓶颈，极大的提升了数据从存储到加速卡的传输效率，在检查点状态数据保存、训练数据加载以及 KV Cache 加速等场景发挥重要作用。

（五）全闪存存储技术为数据提速降耗

高性能低能耗的大模型数据存储与处理需要 AI 存储走向全闪存。

存储介质是指用于存储数据的载体，是数据存储的基础。闪存存储是新一代存储介质，相比传统的机械硬盘，在性能、能耗、容量密度等方面有巨大优势，已经成为了业界的共识。闪存存储比机械硬盘时延小 100 倍左右，每秒数据读写次数大 1000 倍以上，空间占用节省 50%。随着闪存介质堆叠层数与颗粒类型方面的突破，其成本也在持续走低，成为处理大模型数据的理想选择。数据读写性能的大幅提升，将减少计算、网络等资源的等待，加速大模型的部署与应用。根据行业测算，以 GPT-3 采用 100PFlops 算力下，当存储的读写性能提升 30%，将优

化计算侧 30%的利用率，训练周期将从 48 天降低至 36 天，整体训练时间缩短 32%。

使用闪存存储相比机械硬盘能减少能耗 70%。尤其是高密存储节点密度能达到传统存储服务器的 2.6 倍以上，结合存算分离架构，相对使用通用型服务器，减少了存储节点 CPU、内存及配套交换机，同等容量下带来能耗节约 10%~30%。显著降低大模型数据的存储能耗。数据算法和数据融合技术正在大幅度提高存储能效。数据融合：随着数据融合技术的成熟，一套存储系统能够同时提供文件、对象等多协议访问能力，通过多协议融合互通能力，一份数据无须协议转换就能够被多种协议同时访问，减少数据搬迁和重复存储，提升 35% 的数据处理能效。数据算法：数据压缩，数据重删正在向无损化，场景化演进，随着应用语义压缩/重删技术的发展，有望从根本上解决非结构化数据的存储效率问题，从而进一步提升数据能效。

（六）存储内生安全保护企业数据资产

存储内生安全是构建大模型系统安全防护的最后一道防线。

存储内生安全强调在 AI 存储系统中内置安全机制，以保障数据的机密性、完整性和可用性。相比于传统的外部安全措施，存储内生安全具有更高的灵活性和可控性，能够更好地适应不断变化的安全威胁。在大模型的应用中，存储内生安全通过将安全功能融入到 AI 存储系统中实现对数据的保护。这种方式可以避免将敏感信息暴露在外部网络中，从而降低数据被攻击的风险。同时，存储内生安全还能够提供灵活的安全管理策略，根据不同用户的需求进行定制化的安全设置，

提高数据的安全性和合规性。从硬件角度，存储内生安全主要包括构建关键硬件自主能力、硬件三防（防侧信道、防故障注入、防物理攻击）和可信启动等技术。从软件算法角度，重点解决开源软件的风险治理。从数据安全防护角度，AI 存储会围绕数据采集、传输、存储、处理、共享和销毁全生命周期提供安全防护能力。其一是围绕数据全生命周期，提供端到端加密能力，包括数据传输加密，落盘加密等，提升数据流动和数据储存的安全，并在数据生命周期结束时提供数据安全销毁能力；其二是针对数据完整性、真实性和可用性提供安全快照、本地备份、异地备份、双活容灾等数据容灾与备份能力，结合企业数据分类分级可实施符合政策法规要求的数据保护措施；其三是针对勒索病毒攻击提供网络+存储协同防勒索的能力，实现事前防御勒索病毒攻击、事中进行勒索病毒检测、事后保障数据可恢复。勒索病毒隐蔽性极强、变种频繁，往往使用零日漏洞、钓鱼邮件等方式入侵，单靠杀毒软件防不胜防；网络防火墙等设备以进不来为防范目标，可拦截 90% 左右勒索病毒，而 AI 存储作为数据的最终载体，可以在第一时间感知勒索软件对数据的修改行为，通过侦测分析、安全副本、及时恢复，确保病毒进不来、改不了，数据可恢复。

四、人工智能存储建设最佳实践

（一）医疗行业：长记忆存储助力智能辅助诊断

（1）项目背景

某三甲医院应用病理大模型，融合多模态数据，并结合人群的疾病特征，从而为病理医生提供更精准、更高效的辅助诊断支持。病理

大模型（即针对病理学领域设计的大型 AI 模型，如用于病理图像分析、疾病诊断或预后预测的模型）需要具备基于长上下文的"慢思考"能力，以应对医学领域特有的复杂性、不确定性以及对严谨性和安全性的极高要求。通过多步骤推理、多模型交叉验证、历史数据回溯分析等"慢"过程，能够减少误诊几率，也让大模型能够清晰表达诊断的不确定性，而非单纯输出二元结果。

（2）解决方案

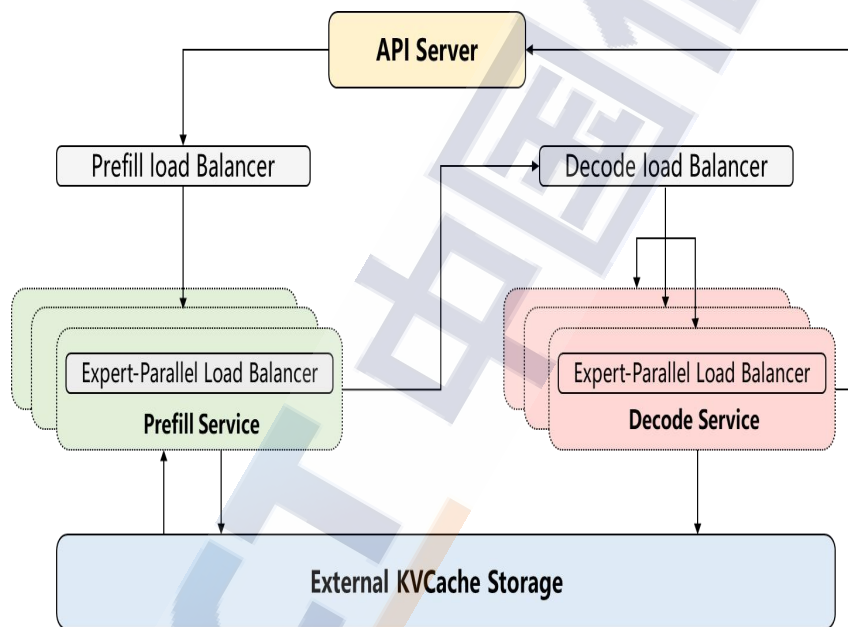


图 8 医疗行业 KV Cache 长记忆存储应用实践

来源：中国信息通信研究院

医院选择了领先的 KVCache 长记忆存储方案，如图 8 所示，为病理大模型推理系统提供 TB 级带宽、EB 级容量，大规模算力集群共享的全局内存扩展池，实现高效的上下文 KVCache 保存、管理与加载策略。得益于其强大的扩展能力，可为大模型提供终身记忆和无

限上下文能力，支撑病理大模型更精准更高效。

（3）实现价值

显著提升长序列推理效率，病理大模型需处理全切片图像（WSI）分割后的数万至数十万图像块（patches），形成超长输入序列，直接处理此类序列会导致计算爆炸。长记忆存储缓存已计算的键值，避免对历史序列的重复计算，使推理速度提升数倍，处理 10 万 patch 的 WSI 时，长记忆存储可减少 90% 以上的冗余计算。支持动态上下文建模，疾病诊断需结合局部细节与全局结构，长记忆存储允许模型灵活调整上下文范围，如在胃癌分级任务中，模型通过缓存不同区域的键值，可以快速对比癌变区域与正常黏膜的特征差异。促进多模态数据协同，病理模型需融合图像、文本报告、基因组数据等，长记忆存储可分别缓存各模态的键值信息，加速跨模态注意力计算，在预后预测中，通过缓存患者既往治疗的文本记录与历史 WSI 的 KV，模型更精准评估复发风险。

（二）金融行业：高性能 AI 存储赋能智慧金融

（1）项目背景

某银行提出 AI-Powered 的战略方针，旨在基于现有大数据平台的基础上，增加 AI 模型部署。随着数据量和 AI 场景化模型的激增，原有的对象存储底座无论可靠性还是性能都不再能满足大数据分析和 AI 模型的运行要求，一是在高 IOPS 峰值时段多次触发存储故障，严重影响了业务的正常运行，二是原有存储系统基于开源开发，单站点故障，业务中断，且不具备多站点多活的能力。

（2）解决方案

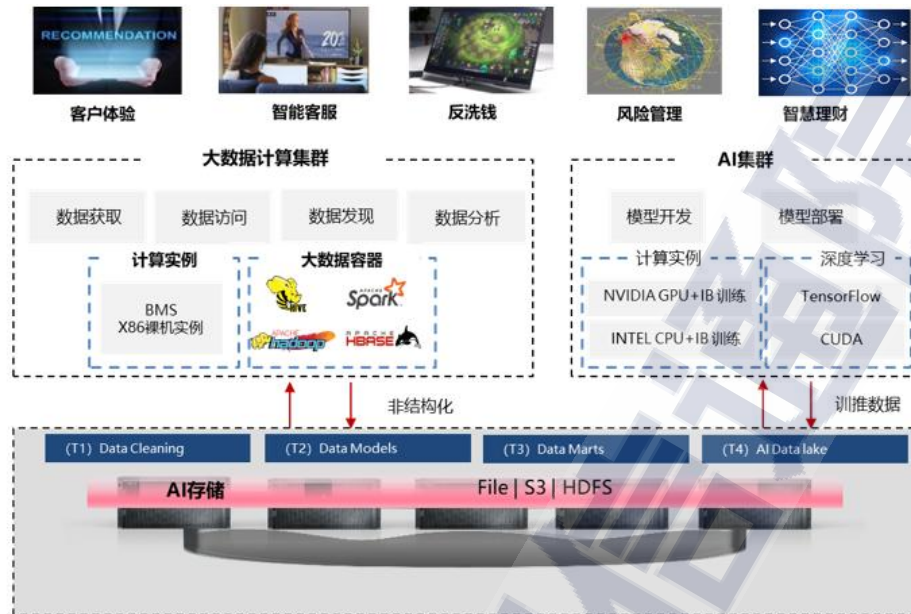


图 9 金融高性能 AI 存储应用实践

来源：中国信息通信研究院

银行选择新型 AI 存储来替换原有对象存储系统，提高性能和可用性，如图 9 所示。一是提供混合 IO 的高 OPS 和高带宽能力，千亿对象的检索时延从之前的十多秒级别降低至 50 毫秒级；二是集群内高性能节点和高密大容量节点之间通过自动分级策略优化投资，整体 TCO 降低 20%；三是支持最大 12 站点部署容忍两个站点故障，为大数据和 AI 平台提供了 99.9999% 高可靠性；四是多协议（NAS/S3/HDFS）统一（All-in-One），大数据处理和 AI 训推各个阶段之间无需数据拷贝。

（3）实现价值

银行通过采用高性能、高可靠的 AI 存储，通过 800 个 AI 模型赋能 350 个案例，包括客户体验、智能客服、反洗钱、风险管理、智慧

理财等多方面业务模块。其实现效果包括三个方面，一是呼叫中心使用生成式 AI 进行呼叫转录、摘要、服务、请求生成和知识库查找，使平均呼叫处理时间减少 20%；二是 AI 驱动的实时数字欺诈预防，10 毫秒内即可标记潜在的欺诈性交易；三是贷款流程大幅缩短，实现一分钟申请，一秒钟批准。

（三）AI 企业：AI 数据湖支撑大模型训推平台

（1）项目背景

某 AI 企业专注于智能语音、计算机视觉、自然语言处理、认知智能等领域，拥有语音及语言国家工程实验室和认知智能全国重点实验室。随着大模型参数量的增长和数据规模的激增，该 AI 企业面临诸多挑战，一是数据准备时间长，TB 级数据准备需要小时级；二是存储数据读取效率低，千卡以上集群平均每天故障 1 次，断点恢复时间高达 15 分钟以上；三是存储可靠性不足，RAID 机制支持盘级冗余，但不能跨节点冗余，存储可靠性面临严重挑战。

（2）解决方案

该 AI 企业选择外置 AI 数据湖存储底座方案，如图 10 所示，针对通用大模型，采用算、存分离架构，计算侧追求更加极致的算力释放，存储侧部署多套新型 AI 存储，提供可靠高效的几十 PB 可得容量。同时全局文件系统支持统一数据湖管理，显著降低数据治理成本，数据全局可视、可管，高效流动，跨域调度效率提升 3 倍。无损多协议互通免除数据孤岛，数据零拷贝，端到端加速 AI 模型开发。

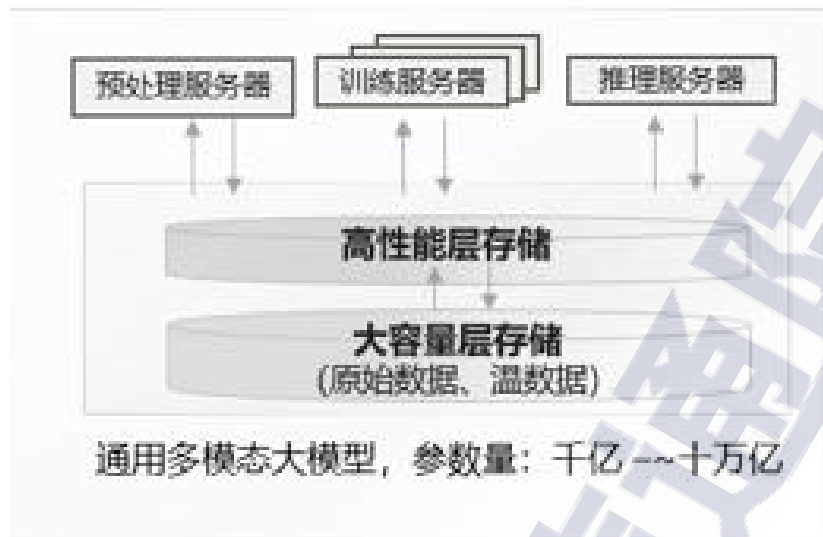


图 10 外置 AI 数据湖存储底座应用实践

来源：中国信息通信研究院

（3）实现价值

该 AI 企业通过打造存、算、网全栈协同的大模型解决方案，其实现效果包括以下两方面，一是断点续训恢复速度提升 15 倍，集群提供 TB 级大带宽，缩短 Checkpoint 读写耗时，断点续训恢复时长从 15min 缩短到 1min，速度提升 15 倍，集群可用度提升 20%；二是提升存储集群安全可靠，单集群管理面合一，数据面分离，通过数据面隔离避免 AI 集群故障扩散，同时通过亚健康管理、大比例 EC 等进一步提升存储可靠性，单集群可靠性达 99.999%。

（四）运营商：构筑领先智算中心云数据底座

（1）项目背景

某运营商聚焦大模型训练推理场景打造智算平台，为大模型训练推理提供高性能计算、存储、网络等云服务。该运营商基于开源 Lustre

自研并行文件系统，可以量体裁衣地匹配自身需求，但也对性能与运维提出了新的挑战。一是存储带宽不足导致训练效率低，现有存储带宽仅 10GB/s，导致大模型训练过程中，单次 Checkpoint 读取时间需 10 分钟，GPU 等待导致算力利用率低，亟需提升存储带宽能力；二是本地盘故障导致业务不稳定，开源文件系统盘管理能力弱，频繁的突发盘故障导致上层业务经常中断，故障后，数据重构时间长达数小时，消耗 CPU 资源，导致上层业务性能不稳定。

（2）解决方案

某运营商采用自研并行文件系统和高性能智能存储盘框协同，打造可靠、高效的大模型训练推理架构，如图 11 所示。一是提高性能，全闪存智能盘框，2U 空间提供高达 150GB/s 的超高带宽；二是提高可靠性，硬件均衡架构，全面硬盘亚健康管理，全局数据重构，实现专业级 AI 存储 99.9999%高可靠，保障业务永在线；三是提供生态兼容，替代现网服务器集群，无缝对接自研的并行文件系统，可同时支持 NVIDIA 和昇腾两大主流计算平台。四是智能存储盘框内置 RAG，降低大模型幻觉。

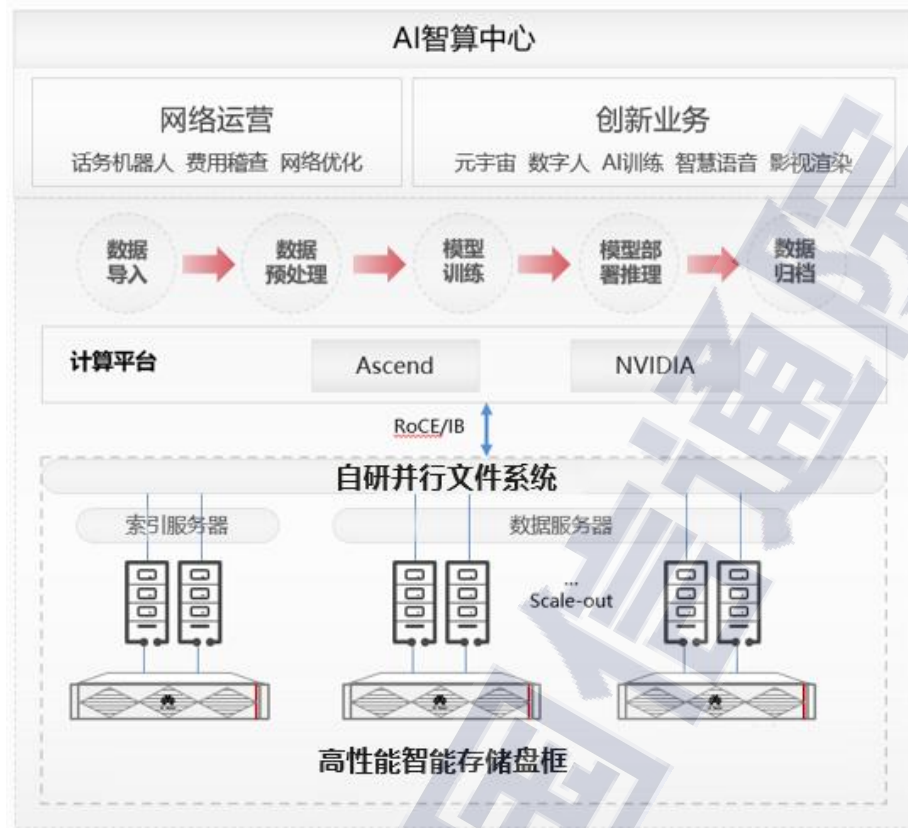


图 11 某运营商智能存储盘框应用实践

来源：中国信息通信研究院

（3）实现价值

运营商通过采用自研文件系统和智能盘框方案，其实现效果包括以下两方面，一是存储带宽性能提升 6 倍，GPU 训练效率提升 10%，文件存储性能从 50MBps/TiB 提升到 300MBps/TiB，单次 Checkpoint 时长从分钟级缩短到秒级，提升集群可用度；二是训练稳定性提升 10 倍，通过硬盘亚健康管理，硬盘故障提前 14 天预测，通过全局数据重构技术，1TB 数据重构时长由小时级缩短到分钟级。设备数下降 75%，从原有的 1000 多台服务器降低到 100 多台服务器+80 多台智能存储盘框。

（五）制造业：一体化集成加速智能客服场景创新与实践

（1）项目背景

某科技企业存储产品及解决方案种类多、包含 18 大类产品与解决方案，580+子产品特性，技术更迭日新月异。随着用户的增多，对业务咨询的压力也与日俱增。一是咨询数量增长快，从存储知识到配置建议方案定制，包含 3 大类 20 小类问题，团队年处理问题数 4 万以上，团队年处理问题数上升超 25%；二是问题答复慢，用户体验差，一般问题答复时间 15 分钟内，复杂问题答复时间 2 小时内。

（2）解决方案

科技企业携手大模型企业，通过训推超融合一体机将行业专有知识库与大模型能力充分融合，通过产线预装、现场部署、快速定制、高效训练、安全训推、全栈运维，推出具备“智能客服”能力的数字员工，如图 12 所示。训推超融合一体机提供软硬件融合能力、资源池化与调度能力、AI 全流程工具链能力（数据工程、大模型训推管理、应用编排）和统一管理能力，尤其是采用长记忆存储，在通用大模型基础上充分叠加行业专有数据集，推理精度显著提升。



图 12 制造企业超融合一体机应用实践

来源：中国信息通信研究院

（3）实现价值

该科技企业通过采用超融合一体机的一站式解决方案，其实现效果包括以下两方面，一是答复效率千倍提升，以“智能”取代“人工”，“0”学习成本，具有智能问答、向量检索、智能识图等功能，实现 7x24 小时值班秒级答复，用户满意度提升 95%；二是答复准确度更高，通过长记忆存储，实现知识库实时更新，数据入库即可查询，推理更准确。

五、人工智能存储发展建议

当前，全球人工智能发展正从“以模型/算力为中心”加速向“以数据为中心”转变，高质量、大规模、多样化的数据成为人工智能发展的关键战略要素。作为数据的载体，数据存储成为大模型时代的关键基础设施，构建先进的人工智能存储是构建领先 AI 基础设施的核心条件之一。

为深入贯彻国家发展新一代人工智能的战略部署，充分发挥我国海量数据和丰富应用场景优势，促进人工智能和实体经济深度融合，

催生新产业新业态新模式，不断做强做优做大我国数字经济，本报告针对我国人工智能存储发展，提出以下建议：

（一）加强 AI 存储设施统筹部署

在国家战略层面，尽快形成 AI 存储产业发展的顶层规划，系统布局 AI 存储设施。聚焦长记忆存储范式、数据编织、AI 存储规模、数据灾备覆盖率等大模型需求技术和指标，制定引导我国 AI 存储产业高速健康发展的战略政策；积极围绕计算、存储等项目统筹规划，建立计算与存储并重的大模型基础设施，形成一批新型应用场景和新型应用示范项目。

（二）增强 AI 存储自主创新能力

在技术研发层面，全面加快核心技术创新，推动 AI 存储研发部署的自主可控。在国家科技攻关计划和相关产业链中设立 AI 存储专项，充分发挥行业创新领军企业的创新带领优势，紧扣科技自立自强的要求，组织国内产学研用各界联合攻关，实现全产业链自主可控；成立 AI 存储国家级创新平台，开展前沿技术研究，形成原创性成果，加快 AI 存储与应用的结合，提升国产 AI 存储全球技术竞争力。

（三）推动 AI 存储产业协同发展

在产业发展层面，鼓励国产设备应用，提升 AI 存储的安全保障能力。需加强国产化应用牵引，推动自主品牌做大做强，在大模型基础设施建设中要求使用自主可控的存储设备，以国产促用，打通 AI 存储上下游产业链，通过存储设备和固态硬盘主控芯片带动国产存储芯

片应用，带动产业链整体能力提升，形成产业链上下游协同创新的行业发展新格局。

（四）构建 AI 存储能力评价体系

在标准建设层面，持续完善评价体系，促进 AI 存储产业健康蓬勃发展。积极开展 AI 存储产品标准、测试标准的制定工作，不断完善 AI 存储的标准体系建设，强化标准引领产业发展的作用；推动构建 AI 存储技术创新应用公共服务平台，联合产业各方积极开展 AI 存储新型技术及质量标准等验证评估，以“测”促“建”，以“测”促“用”，有效推动 AI 存储产业的高质量发展。

中国信息通信研究院 人工智能研究所

地址：北京市海淀区花园北路 52 号

邮编：100191

电话：010-62301618

传真：010-62301618

网址：www.caict.ac.cn

