

数据作为新型生产要素，是数字化、网络化、智能化的基础。ChatGPT和Sora等人工智能技术的爆发，全球数据量呈现井喷式增长，预计2030年全球数据规模将超过200ZB（1ZB等于10⁶GB）。然而，全球大数据存储规模却仅达到数ZB级且年增长缓慢，真正被有效存储的数据只占总数据生产量的2%，而其中超过80%的数据被归为冷数据。冷数据是指那些不经常访问但需长期保存的重要信息，它们既是人类文明的记录，也是信息化时代的宝贵资源，对于一个国家的竞争力和安全具有重要保障作用。

冷数据存储意义重大但耗能巨大

冷数据具有不经常访问、需长期保存的重要特征。冷数据具体包括科研数据、法律文件、医疗记录、天文观测数据、气候变化等。正如在许多领域看到的那样，冷数据的妥善保存和管理直接关系到社会的稳定与发展。据IDC（国际数据公司）测算，2022年全球数据中心的用电量约占全球总用电量的2%，但保守估计到2026年数据中心用电量或将增长7成。2022年，中国数据中心总耗电量约2700亿千瓦时，超过2座三峡水电站的年发电量，预计2025年将高达3842.2亿千瓦时。随着数据总量的增加，存储电耗成为制约当前大数据可持续性发展的主要瓶颈。如何高效、经济地存储这些冷数据成为各国科技战略的重要组成部分。因此，寻找一种具有大容量、低能耗、长寿命的存储技术，成为当前科技界的重要任务。

光存储是冷数据存储的理想模式

在大数据时代，海量数据的存储需求促使人们不断探索更为高效的存储解决方案。在众多存储技术中，光存储因其独特的优势脱颖而出。传统的磁存储和电存储技术虽然在读写速度上具有优势，但在存储寿命、能耗和数据安全性方面存在明显不足。而光存储技术则因其潜在高密度、低能耗、抗电磁干扰和长寿命的特点，成为冷数据存储的理想选择。光存储不仅能满足大数据的存储需求，还能显著降低数据中心的能耗，从而推动绿色科技的发展。

中国发展冷数据光存储技术具有得天独厚的优势。中国是全球光存储

如何让海量冷数据不再被“冷落”

□ 顾 敏



顾敏，光学工程专家，中国工程院外籍院士。现任上海理工大学校务委员会执行主席及光子芯片研究院院长，主要从事人工智能纳米光子学、光电子学成像和光存储等领域的研究。在世界上首次利用双光束超分辨原理突破聚合物激光加工的光学衍射极限，获得9纳米特征尺寸世界纪录，并领导研发团队实现了PB级光存储技术，其研究成果在现代光学显微成像、纳米激光信息、纳米制造和大数据存储领域具有重要推动意义。曾担任国际光学生命科学学会主席、国际光学委员会副主席及评奖委员会主席。荣获国际光学及光子学学会丹尼斯·加博尔奖、国际OPTICA光学学会（原OSA）埃米特·诺尔曼·利思奖章及上海市白玉兰纪念奖。

的重要产业基地，全球消费市场上60%的DVD影碟机和光驱均产自中国。清华大学和中国科学院上海光机所先后成立国家光盘中心，致力于光存储技术的基础研究和产业转化。在“十二五”期间，中国发展了具有自主知识产权的红光高清存储技术（NVD），实现单盘存储容量15GB。在科技部“十三五”和“十四五”“信息光子学”等重点专项支持下，上海理工大学与中国科学院上海光机所开展了双光束大容量存储技术的基础研究，暨南大学开展了多维复用光存储技术研究，华中科技大学和吉林大学开展了玻璃基长寿命光存储技术研发，为中国布局冷数据光存储奠定了技术基础和人才储备。

PB级冷数据光存储的瓶颈问题与解决方案

将光存储技术应用于大容量冷数据存储领域，长期以来面临着多重技术壁垒。首要挑战在于如何显著提升光存储单盘容量至PB级（1PB等于10⁶GB），以满足冷数据总量的爆发性增长。为此，我们深入探索光存储技术的核心原理，并通过创新性的器件阵列化设计，首次构想了对冷数据的PB级光存储阵列方案。同时，精准地提炼出冷数据光存储面临的五大核

心瓶颈问题：光学衍射的物理极限严重制约了光存储的密度与容量上限；如何利用稀土材料与双光束超分辨技术，有效降低单比特数据的读写能耗；确保光存储数据的安全性，防止数据泄露或篡改；提升光存储的读写速率，以满足高效数据处理的需求；制造超分辨光存储的读写样机，实现技术从理论到实践的跨越。经过领域内科研人员长达十年的不懈努力与协同攻关，上述瓶颈问题已逐一获得突破，为光存储技术在大容量冷数据存储领域的广泛应用奠定了坚实基础。

突破衍射极限限制解决容量瓶颈：PB级光存储单盘容量相当于至少一万张蓝光光盘。1873年，德国物理学家恩斯特·阿贝发现了光学衍射的物理极限，即最小聚焦光斑尺寸受到光波长的限制，这一发现为理解光学成像的极限提供了重要基础。虽然光学衍射极限在理论上限制了光存储密度的上限，但具体的存储容量还受到多种因素的影响。与2014年诺贝尔物理学奖中涉及的双光束超分辨成像技术类似，我们发展了一系列突破光学衍射极限的双光束超分辨光存储技术。其中，以2024年发表在《自然》杂志上的双光束聚焦诱导发光超分辨光存储技术为代表，突破了长达150多年衍射极限的物理瓶颈制约，实现了PB级三维纳米光子存储，其单盘容量相当于至少一万张蓝光光盘，有效

地解决了光学衍射物理极限对光存储密度和容量的限制问题。

发挥稀土资源优势降低读写能耗：镧系稀土材料的长荧光寿命能够实现节能99%以上。稀土材料不仅是中国的优势资源，具有重要的战略意义，而且具有长荧光寿命的特性。双光束超分辨的基本原理依赖于第二束光对第一束光记录效应的边缘抑制，超分辨记录点尺寸随着抑制光强的增加而显著减小。尽管记录点尺寸可以远小于衍射极限，但高抑制光强曾一度严重降低了光存储的读写能效。我们巧妙地利用镧系稀土材料长荧光寿命的特性，显著提升了低光强下的抑制效率，从而在保持超分辨记录特性的同时，将光存储的能耗降低了3个数量级，实现数据中心节能高达99%以上。这一技术结合不仅有效地解决了单比特的读写能耗问题，还为光存储技术的未来发展开辟了新的道路。

实现信息光存储超安全：矢量涡旋光场的多维独立调控保障信息存储的安全性。光学在频率、偏振、相位以及光强分布等多维自由度上展现出丰富的编码潜力。利用这些多维参量对信息进行编码，不仅能显著提升信息存储容量，还能有效增强数据存储的物理安全性。然而，在光场多维参量空间中同时复用多通道信息，对多维光场的精确调控技术以及材料对这些调控的多维响应能力提出了严峻挑

战。针对这一问题，我们首次揭示了纳米尺度下材料的独特光学特性——合成螺旋二色性现象，并通过对矢量涡旋光场的多维独立调控，成功克服了这些挑战，实现了在空（3D）、频、偏、相（角动量）等参量空间的首例六维光信息复用存储，为信息存储提供了前所未有的超安全级别。

解决读写速率受限问题：光场并行操控技术极大提升光存储的读写速率。光的本质是电磁波，不同光波在空间中传播时，通常不会相互干扰。这一特性使得多条光束能够同时存在于同一空间中，并各自独立地被调制，从而实现并行的读写操作，以提升光存储读写速率。通过采用一种并行读写技术，能够同时处理四条光束，每条光束独立地执行数据存储操作，从而在原理层面突破了读写速率的限制。随着高性能光电设备的不断涌现，光场并行操控技术正经历显著的进步。利用其对光场的高速、精准调制能力，进一步结合人工智能和光计算，有望使光场并行读写速率从GB/s提升到TB/s，推动光存储技术迈向新的高度。

成功研制读写样机：超分辨光存储读写样机助力PB级冷数据光存储实用化。超分辨光学数据存储采用突破衍射极限的方法，取得了存储密度的大幅度提升。为了将该技术推向实用化，成功研制了不同类型的超分辨光

存储读写样机。飞秒—连续光双光束超分辨光存储读写样机通过不断改进，从早期的占用一个1.5米×2.4米光学平台发展到占地面积不超过0.5平方米，并且适配一般震动条件的可移动装置。另一方面，双连续激光超分辨三维光存储读写样机则采用全新的步进吸收光子上转换原理，将原来的飞秒激光替换为连续激光，既保证了三维数据写入能力，又同时支持超分辨读写，大大提高了装置的稳定性和环境适应能力。在存储介质方面，创新性地采用新型材料，该材料不仅显著提升了存储介质的读写性能，还成功克服了传统材料在溶解度、耐光照性能以及寿命等方面的缺陷。此外，还结合DVD工艺，制造出了空白光盘。凭借在超分辨光存储读写样机领域的深厚工程开发能力和技术积累，我们团队已成为国内外最领先的样机研制团队。

PB级冷数据光存储工程化仍需不断攻关

尽管PB级冷数据光存储的基本原理在实验室中得到了有力的验证，然而，将这一前沿技术从理论构想转化为实际应用，仍需面对并解决一系列复杂的工程技术难题与应用实施挑战。这包括高速动态读写工程样机开发、读写系统稳定性提升、系统复杂度简化、并行读写系统开发，以及材料均匀性、长期稳定性、工艺开发和制造成本控制等问题。目前，单光束超分辨存储技术的开发工作正在积极推进中，以进一步简化系统工程化难度。然而，要实现PB级光存储的工程化应用，亟需建立国家级平台，联合科研和产业界集中优势力量攻关，共同推动技术突破和成果转化。

作为一名全职回国工作的科学家，我深感责任重大，希望通过我们的努力，推动中国光存储技术的发展，也为全球数据存储提供更加高效、环保的解决方案。通过持续的技术创新和产业化推进，将PB级光存储这一前沿技术产品实现量产应用，为国家的信息化建设和数字经济的发展作出贡献。



认识大模型

□ 熊德意

大模型，作为人工智能领域的革新先锋，正引领着一场深刻的技术变革，其影响力已超越技术范畴，预示着人工智能、科学探索乃至人类社会的全面重塑。在全球科技竞赛中，大模型成为各国争夺焦点，战略地位堪比20世纪的太空竞赛，开启大国科技新角逐。因此，深入了解大模型的各个方面变得尤为重要。

大模型的诞生、发展与应用

大模型，是一类基于深度学习架构、在海量数据上训练、能够处理多种任务的基础模型。与传统AI模型相比，其显著特征在于参数量大、数据量大、计算量大，是多重技术交叉融合的产物。其前身是历经数十年研究的语言模型，广泛应用于机器翻译、语音识别等领域。随着神经网络的再次兴起，2000年神经语言模型诞生，随后2017年神经网络架构Transformer横空出世，融合了神经语言模型、机器翻译、序列建模及分布式技术。这一系列创新及对AI通用和泛化能力的追求推动了GPT系列的快速发展，从2018年至2020年，GPT一至三代相继问世。2022年，ChatGPT凭借其卓越性能，将大模型推向公众视野，成为AI发展的新里程碑。

大模型作为新兴而迅速发展的技术，已跨越至多模态领域，涵盖语言、语音、视觉等，并细化为通用、行业及任务特定模型，其在信息交互上的创新尤为显著，引入了长上下文窗口（大模型能够编码的最长序列）、检索增强及智能体等机制，极大增强了信息处理能力。这一技术不仅丰富了数字世界的应用场景，如聊天机器人、AIGC等，还延伸至无人驾驶、人形机器人等物理世界领域，展现出强大的赋能潜力。在科学研究中，大模型不仅辅助工程设计、技术创新等应用科学，更推动基础科学探索，加速了自动化、智能驱动的科学新范式的形成。可以说，大模型既支撑了

高端科学研究，也助力了传统产业的转型升级与降本增效。《2023企业数字化年度指南》调研数据显示，超过63%的企业视大模型与AIGC技术为战略资源，首要目的在于降本增效，广泛应用于研发创新、办公效率提升、客户服务、市场营销等多个领域。同时，大模型在工业设计、芯片设计、人形机器人等多个行业的应用案例也充分证明了其广阔的应用前景与市场潜力，正逐步成为推动社会进步与产业升级的重要力量。

大模型背后的关键技术

是什么造就了大模型如此强大的能力？从大模型的发展实践看，其背后有三项关键技术。

一是大模型底层的神经网络架构Transformer。Transformer最初提出主要是为了解决序列建模中的循序计算问题，即只有前面的词元（自然语言处理中的最小语义单位）计算后才能计算后面的词元。循序计算极大束缚了模型的并行计算能力，使得语言模型、机器翻译等序列模型难以大规模扩展。Transformer架构巧妙地利用了自注意力机制，不仅大幅度提升了模型的并行计算能力，而且显著提升了模型的长距离依存关系建模能力（语言模型的关键挑战）。并行计算能力使Transformer可以在大规模数据上进行快速训练，而长距离依存建模能力极大强化了Transformer对语言等序列数据内部复杂依存关系的学习和表征能力。两项能力的叠加，使Transformer开启了“后临词元预测”范式模拟人类智能的大门。

二是基于自监督学习的预训练技术及扩展法则。Transformer虽然具有强大的并行计算和表征能力，但要使其学习和模拟人类智能还需要解决两个重要问题：如何将人类多样化的智能统一表征，如何训练模型使其高效学习人类智能。针对第一个问题，GPT将各类任务统一表示为生成式问题，无论是写作、翻译、问答等典型

生成式问题，还是推理、规划、问题求解等非典型生成式任务，都归聚到生成式AI的统一框架中，在这个统一框架中，基于后临词元预测的语言模型成为天然的技术路线。统一表征不仅使AI不需要为不同任务设计不同的模型（传统弱人工智能模式），而且还极大提升了AI模型在不同任务间的迁移能力、泛化能力。对于第二个问题，语言模型的后临词元预测模式使自监督学习成为模型训练的首选，在可大规模获取的人类语言数据上，后临词元本身就存在，因此模型可以充当自己的“教师”，采用“教师强制”方式进行自监督训练。区别于传统人工智能有限的有监督数据，自监督数据规模庞大且易于获取。这不仅使模型可以学习丰富的人类知识，同时也要求模型的容量（参数量）足够大。在给定算力规模前提下，如何选择模型的参数规模及训练数据规模，使模型能训练到最优性能？业界经过反复实践，总结出了扩展法则，用以指导模型与数据规模的协同扩展。

三是后训练阶段的精调和对齐技术。经过预训练的大模型，如同学习了大量知识和百般武艺的“学生”，但不知该如何运用。为了解锁预训练模型的知识 and 能力，通常需要进行后训练，即采用少量的有监督数据对预训练模型进行精调，使其能够理解人类的意图，遵循人类的指令，对齐人类的价值观。业内目前广泛采用的后训练，通常包括两个主要部分：有监督精调（模仿学习人类给出的示范样例）及人类反馈强化学习（基于人类反馈的偏好数据进行强化学习）。后训练的成本通常远小于预训练成本，同时，后训练也是解锁模型能力、保障模型智能向善的关键阶段，因此，业界对后训练寄予厚望。

大模型的突破点与创新方向

模型架构、预训练与后训练，既是大模型的关键所在，也是未来重点

突破的方向。Transformer架构虽潜力强大，但其是否为最优的智能学习架构尚待验证，能否承载全部的智能学习任务亦存争议。预训练领域，高昂成本促使我们寻求更高效的数据学习方法，同时，模型规模与性能间的最优平衡，以及模型、数据与计算三者间的理论关系，亦需深入探索。后训练领域，是否存在新的方法，偏好数据是否真能代表人类意图和价值倾向，预训练+后训练是训练现有架构的最优突破吗？这些都是值得思考和亟待突破的地方。

除此之外，还有四个重要研究方向有待深入探索（与以上三个方向存在一定的交叉），其突破可能关系到大模型能否跃迁到人工智能的新阶段。一是类人推理能力。大模型的推理能力有待进一步加强，复杂问题的类人推理能力是大模型现阶段最期待的突破方向之一。二是AI合成数据能力与模型的自我迭代能力。随着模型规模的不断扩大，人类数据可能很快被大模型用尽，AI合成的数据，成本既低廉，又可源源不断生成。但如何生成高质量数据，使大模型能够自我学习、自我迭代，甚至实现弱模型合成数据监督（训练）强模型，将是未来突破数据限制的重要方向。三是自我发现和创新的能力。人类不仅能够学习已有知识，也可以创造新的知识，未来要实现完全自动化科学研究，大模型需要具备创造新知识的能力。四是可扩展的监督。在复杂任务上、人类难以提供监督数据的任务上（如未解的科学难题），大模型仍需要与人类目标和价值对齐，这就要求监督是可扩展的。该能力是前沿人工智能实现智能向善、安全治理的关键技术。

我国大模型技术的进展情况与发展路径

尽管我国在大模型技术方面初期存在一定滞后，但近一年来发展迅速，不仅在研究领域取得了显著成果，还在产业应用上展现出了强劲实

力。国内相继开源了通义千问等大模型系列，这些成果在国际大模型开源领域中仅次于美国，形成了广泛的影响力。在产业应用层面，我国多个大模型的中文化能力已经超越GPT-3.5，它们在数字人、人形机器人等前沿领域，以及能源、交通、医疗、政务、农业、金融等多个传统及新兴行业中，都实现了丰富的落地应用，推动了各行各业的智能化发展。

为巩固并提升我国在这一领域的国际竞争力，可以从以下布局和规划着手。第一，进一步提升以大模型为代表的的前沿人工智能在国家科技和产业发展中的战略地位，成立人工智能工作小组，领导AI产研咨询委员会，统筹资源，制定AI政策和计划，推进人工智能技术创新和产业发展。第二，重点规划和建设前沿人工智能相关的国家基础设施，包括超级智算网络、通用及行业数据基础设施、大规模人工智能软件基础平台、人工智能安全与测评基础设施、大模型开源平台等。第三，开展大模型关键理论和技术攻关，啃硬骨头，探新疆域，研发经得起实践考验的硬核技术。第四，培育和建立大模型创新发展生态，形成大模型技术创新氛围，鼓励耐心资本敢投广投大模型硬科技创业企业。第五，重视人工智能人才培养和成长，培养一批具有长远眼光和实战经验的AI战略型人才、技术型人才、交叉复合型人才等。第六，重视人工智能安全治理，既要设计顶层治理策略，更要推动底层安全技术的新突破。第七，积极开展国际合作，建立新型人工智能国际组织和机构，吸收新理念，合研新技术，与发展中国家共享AI红利。第八，推动前沿人工智能行业、国家、国际标准建设，形成标准体系，以标准建设护航人工智能产业发展。

