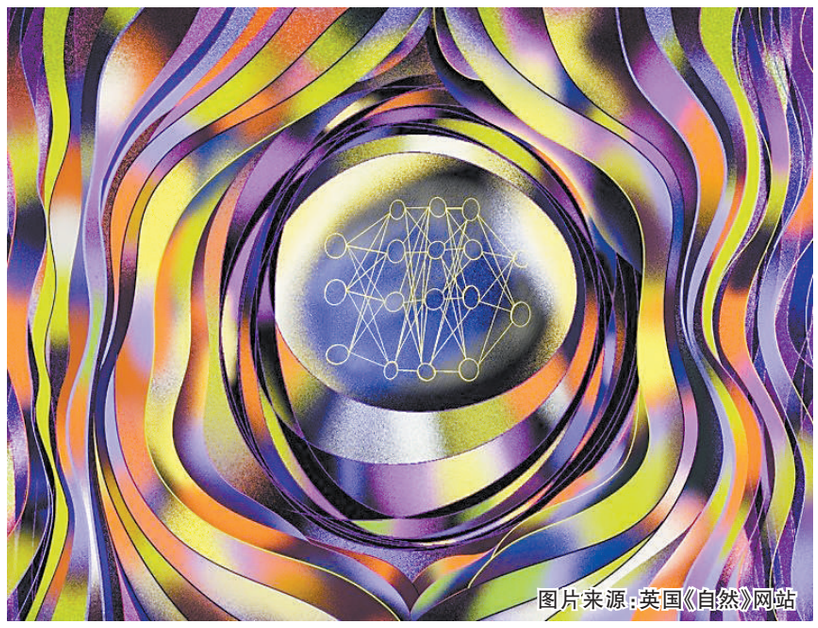
生成式AI“幻觉”困境如何破解

科技创新世界潮 385

◎本报记者 刘霞


图片来源：英国《自然》网站

人工智能（AI）技术正以前所未有的速度发展，生成式AI凭借其惊人的创造力，不断刷新人们的认知。然而，即便是看似“聪明绝顶”的AI，也难逃“幻觉”的困扰。这里的“幻觉”，指的是AI生成看似合理但实际不准确或虚假的信息。

英国《自然》杂志网站在1月22日的报道中指出，AI“幻觉”可能会引发严重后果，科学家正各出奇招，力求降低其发生率。这些措施包括增加事实核查、对AI进行“脑部扫描”等，以促进AI的健康、高效发展。

主因是数据模糊

各种生成式AI，包括由大语言模型驱动的聊天机器人，常常会编造信息。它们有时会模糊事实与虚构，在看似真实的陈述中夹杂错误信息。这既是其创造力的体现，也是其不足之处。

美国佐治亚理工学院理论计算机科学家桑托什·威姆帕拉解释称，大语言模型的设计原理并非输出准确事实，而是通过模式识别生成答案。其内部复杂的运行机制迄今仍像一个“黑匣子”，人们难以洞悉其推理过程。

美国加州Vectara公司旨在减少生成式AI的“幻觉”。其联合创始人阿姆尔·阿瓦达拉表示，在训练过程中，这些模型会压缩数万亿个单词间的关系，随后通过一个庞大的网络模型重新展开这些信息。尽管这些模型能够重构出接近98%的训练内容，但剩下2%的内容却会让其“误入歧途”，生成不准确或虚假信息。

导致AI出现“幻觉”的原因多种多样，其中训练数据中的模糊性和错误是常见因素。也有人认为，即使训练数据准确无误，AI也有可能产生“幻觉”。这种现象与某一事实的稀缺程度密切相关。因此，即使经过人类反馈调整过的聊天机器人，也无法完全避免出错。

多领域面临考验

AI的“幻觉”可能会给人们的工作和生活带来较大影响。

在新闻领域，大语言模型可能生成虚假新闻事件，扰乱信息传播秩序，误导公众认知。Vectara公司针对文档内容开展的研究表明，一些聊天机器人编造事实、虚构信息的几率高达30%。世界经济论坛发布的《2025年全球风险报告》显示，错误和虚假信息是2025年全球面临的五大风险之一。

在法律领域，它可能引用虚构的法律条文和案例。比如，2023年美国律师史蒂文·施瓦茨就因“轻信”ChatGPT，在法庭文件中引用了并不存在的法律案例。而在医学领域，它可能提供错误的诊断和治疗建议，危及患者生命。

《自然》在报道中指出，AI“幻觉”在科学参考文献方面出现错误的情况也极为普遍。2024年的一项研究发现，各类聊天机器人在提及参考文献时的出错率在30%至90%之间。它们至少会在论文标题、第一作者或发表年份上出现偏差。虽然聊天机器人都带有警告标签，提醒用户对重要信息进行二次核实。但如果用户对聊天机器人的回复深信不疑，可能会引发一系列问题。

多举措减少“幻觉”

为进一步提升AI的精确度，科学家正想方设法降低其“幻觉”。

例如，增加模型训练参数和训练时长可有效减少“幻觉”。但这种方法需要付出高昂的计算成本，并可能削弱聊天机器人的其他能力，如机器学习算法对未知数据的预测和处理能力。

此外，使用更大、更干净的数据集进行训练，也是降低AI模型“幻觉”出现的有效途径。然而，当前可用数据的有限性限制了这一方法的应用。

检索增强生成（RAG）技术也为减少AI“幻觉”提供了新思路。该方法通过让聊天机器人在回复问题前参考给定的可信文本，从而确保回复内容的真实性，以此减少“幻觉”的产生。在医疗和法律等需要严格遵循经过验证的知识的领域，RAG技术备受青睐。

不过，美国斯坦福大学计算机科学家米拉柯·苏兹根表示，尽管RAG能提升内容真实性，但其能力有限。苏兹根团队的研究表明，一些为法律研究开发的、号称“无幻觉”的RAG增强模型虽有所改进，但仍存在不足。

开发者也可以使用一个与AI训练方式不同的独立系统，通过网络搜索对聊天机器人的回复进行事实核查，谷歌的“双子星”系统便是一个典型例子。该系统提供了“双重核查响应”功能：内容如果突出显示为绿色，表示其已通过网络搜索验证；内容如果突出显示为棕色，则表示其为有争议或不确定的内容。但是，这种方法计算成本高昂且耗时，而且系统仍会产生“幻觉”，因为互联网上错误信息泛滥。

在去年6月出版的《自然》杂志上，英国牛津大学科学家刊发论文称，他们利用“语义熵”，通过概率来判断大语言模型是否出现了“幻觉”。语义熵是信息熵的一种，被用于量化物理系统中所包含的信息量。通过评估AI模型在特定提示词下生成内容的不确定性，来计算模型的困惑程度，从而为用户或模型提供警示，提醒其采取必要的循证措施，确保更准确的答案输出。

美国卡内基梅隆AI研究人员安迪·邹采用的方法是在大语言模型回答问题时，绘制其内部计算节点的激活模式。他形象地称之为“给AI做脑部扫描”。利用不同的计算节点活动模式，可以告诉我们AI模型是在“说真话”，还是在“胡说八道”。

新技术可高效标记细胞蛋白质

标记速度、均匀性和多功能性超出以往

科技日报北京1月27日电 （记者张梦然）美国麻省理工学院开发的一项名为“CuRVE”的新技术，能在前所未有的速度、均匀性和多功能性下，高效标记完整3D组织里数百万个细胞中每个细胞的蛋白质，展示了其一天内对整个啮齿动物大脑及其他大型组织样本进行丰富标记的能力。这项成果发表在最新《自然·生物技术》杂志上。

分析细胞产生的蛋白质是生物学和神经科学研究的重要内容，因为它们反映了细胞的功能状态或对环境（如疾病或治疗）的反应。尽管显微镜和标记技术已有显著进步，但仍缺乏一种可靠且实用的方法来追踪整个3D组织（如小鼠大脑或人脑的大区域）中密集排列的单个细胞的蛋白质表达。传统方法通常只能观察载玻片上的薄组织切片，无法全面了解整个系统中的蛋白质表达情况。

新方法“CuRVE”用一种称为“eFLASH”的技术实现了对大而致密组织的均匀处理，解决了抗体渗入组织速度慢的问题。团队通过复杂的计算模拟优化了不同的设置和参数，不仅能够通过改变脱氧胆酸浓度和标记液的pH值来调节抗体结合，还利用随机电转运技术加速抗体在组织中的扩散。

“eFLASH”这一具有连续可调结合速度的加速分散系统，可应用于标记多种不同类型的组织，包括60多种不同的抗体用于标记小鼠和大鼠的整个大脑细胞、整个小鼠胚胎、其他小鼠器官（如肺和心脏），以及包括人类在内的大型动物的脑组织块。

这些标本均在一天内完成标记，显示出极高的效率。此外，不同的制备过程不需要新的优化步骤，进一步简化了实验流程。“CuRVE”技术不仅提高了标记速度和均匀性，还揭示了其他广泛使用的标记方法所未能发现的新见解，为生物学和神经科学的研究开辟了新的前景。

蛋白质表达是一个精密而复杂的生物过程。当前追踪3D组织中单个细胞蛋白质表达的方法存在局限，传统方法只能观察组织薄片，无法统揽全局。美国麻省理工学院开发的新技术，能解决抗体渗入慢的问题。经计算模拟优化，其可调节抗体结合、加速抗体扩散。该技术能标记多种不同类型的组织，提高标记效率，可在一天内完成对啮齿类动物大脑的标记，为我们带来新的“视野”。它可助力人们对更多复杂生物学和神经科学问题展开进一步探索。

总编辑圈点 环球科技24小时 24 Hours of Global Science and Technology

顶夸克遵循狭义相对论通过最强测验

科技日报北京1月27日电 （记者张佳欣）在欧洲核子研究中心（CERN）的大型强子对撞机（LHC）上，紧凑缪子线圈、合作组进行了一项研究，旨在检验顶夸克是否遵循爱因斯坦的狭义相对论，而狭义相对论通过了这一最强加速器测验。研究成果发表在最新一期的《物理学快报B》期刊上。


紧凑缪子线圈。 图片来源：欧洲核子研究中心官网

120年前，爱因斯坦提出了狭义相对论，它与量子力学共同构成了粒子物理学标准模型的基础。这一模型的核心是洛伦兹对称性，即实验结果不应依赖于实验的方向或速度。尽管历经多次试图推翻它的尝试，狭义相对论依然稳固。然而，某些理论如特定的弦理论模型预测，在极高能量下，狭义相对论可能不再适用，实验观测结果可能会受时空方向的影响。

由于地球自转，LHC中质子束的方向以及由此产生的顶夸克的方向也会随之变化。虽然LHC中的质子束在空间中的方向固定，但随着地球旋转，这些束流及产生的粒子相对于地面观察者的方向会发生改变。

如果自然界存在一个特殊的时空方向偏好（这与狭义相对论不符），那么顶夸克对的产生速率将随地球相对于实验位置的变化而在一天内有所不同。这种变化暗示着洛伦兹对称性的破坏，意味着需要超越爱因斯坦理论的新物理学解释。

此次，团队通过分析顶夸克对在LHC上的产生情况，寻找洛伦兹对称性破缺的迹象。如果实验结果确实受到实验方向的影响，则顶夸克对的产生速率应随时间而变化。然而，团队在分析LHC第二次运行期间收集的数据后发现，无论何时进行实验，顶夸克对的产生速率均保持不变。

这一结果表明，在时空中没有检测到任何优先方向，洛伦兹对称性成立，爱因斯坦的狭义相对论依然有效。该发现为未来利用LHC第三次运行期间收集的顶夸克数据继续探索洛伦兹对称性破缺奠定了基础，并为涉及其他重粒子（如希格斯玻色子、W玻色子和Z玻色子）的研究开辟了新的道路。

欧盟太阳能发电量首超煤炭

科技日报讯 （记者刘霞）据物理学家组织网近日报道，全球能源智库Ember发布的《2025年欧洲电力评论》指出，2024年欧盟太阳能发电量首次超过煤炭，可再生能源发电总量约占欧盟电力总量的半壁江山。与此同时，天然气发电量连续第五年走低，化石燃料发电量更是跌至历史最低点。

该报告分析了欧盟27国的年度电力生产和需求数据。结果显示，2024年，太阳能依然是欧盟增长最快的能源。风能发电量（占比17%）连续第二年超过天然气发电量（占比16%）。总体来看，太阳能和风能发电量增长强劲，使得欧盟可再生能源的占比从2019年的34%跃升到47%，而化石燃料发电量的占比则从39%下滑到29%。

报告称，太阳能和风能发电量激增的趋势席卷欧洲。目前超过一半的欧盟国家，或者淘汰了污染最严重的化石燃料——煤炭，或者将其在能源结构中的占比降低到不足5%的水平。

不过，报告也强调，为了加快绿色转型的步伐，欧盟仍须继续努力，特别是在风力发电领域。此外，欧洲的电力系统还需要进一步提升存储能力，以更好地利用具有间歇性的可再生能源。

Meta今年拟在AI领域投资600多亿美元

科技日报讯 （记者刘霞）据美国《国会山》日报网站1月25日报道，美国元宇宙平台公司（Meta）首席执行官马克·扎克伯格宣布，今年该公司将在人工智能（AI）领域投资至少600亿美元，相比去年增加200多亿美元。

扎克伯格在脸书上发帖称，今年Meta计划在AI领域投资600亿美元至650亿美元。这项庞大的投资将推动Meta核心产品和业务的革新，并试图扩大美国在AI领域的领先地位。

这笔投资的大部分将用于建设一个规模庞大的数据中心，目的是为AI发展提供强大的算力和存储空间。此外，Meta还计划采购超过1万块图形处理单元，这是用于AI机器学习的关键芯片。

扎克伯格透露，Meta正在打造AI“工程师”，负责为该公司的研发工作编写代码。同时，他们还在训练下一代AI模型，以构建Llama开源AI模型。他表述，尚未发布的大语言模型Llama 4有望成为“最先进的AI模型”。同时他也期待Meta AI成为顶级数字助理，用户数量能突破10亿人大关。

微软总裁布拉德·史密斯也在同一时间公开表示，本财年微软将投资约800亿美元，在全球多地建设AI数据中心、训练AI模型并部署基于云的应用程序。

目前，多家科技巨头正以前所未有的力度向AI领域投入资金，以期在激烈的竞争中保持领先地位。

生物打印功能性心脏组织获突破

科技日报讯 （记者张梦然）爱尔兰戈尔韦大学研究团队开发出一种创新生物打印技术，能够使打印出的组织根据细胞产生的力量而改变形状。这一成果模仿了器官在自然发育过程中经历的动态形状变化，特别适用于心脏组织的复制，在功能性生物打印器官领域迈出了重要一步。研究成果发表在最新一期《先进功能材料》杂志上。

生物打印技术依赖于“生物墨水”，这是一种支持活细胞生长和发展的特殊材料。使用这种墨水可以创建与人类器官结构高度相似的实验室培养器官。不过，制造完全功能的器官仍然是一个挑战。如目前生物打印的心脏组织虽然能够收缩，但其收缩力远不及健康成人心脏的水平。

传统的生物打印方法倾向于直接重建目标器官的最终形态，而忽视了胚胎发育过程中的动态形状变化对于器官成熟的重要性。以心脏为例，它最初是一个简单的管状结构，经过复杂的弯曲和扭转过程，才发展为成熟的四腔结构。这些形状变化对心脏细胞的发育和成熟至关重要。

此次团队引入了一种新型生物打印平台，利用嵌入式生物打印技术，他们生成了能够在细胞力驱动下进行编程并可预测地发生4D变形的组织。这种变形不仅改善了生物打印心脏组织的结构，还促进了其功能的成熟。

团队还开发了一个计算模型来预测组织形状变形的行为，并展示了如何通过调整初始打印几何形状和生物墨水硬度等因素，来控制形状变化的程度。这一研究有望在疾病建模、药物筛选及再生医学领域产生深远影响。

新电池让可穿戴设备比棉花还透气

科技日报讯 （记者张佳欣）一款弹性十足且柔韧的袋式电池，不仅成为可穿戴式运动或健身设备的理想电源，同时使其透气性超过了棉花。包括美国耶鲁大学在内的研究团队开发了这款新型电池，并在最新一期《物质》杂志上发表了其研究成果。

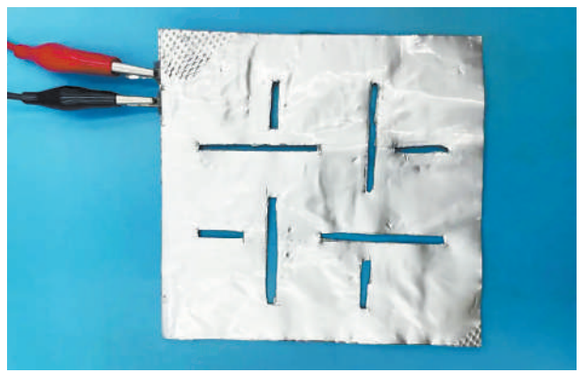
为了设计这款新型电池，团队在袋式电池上打造了一种长方形孔洞的图案。袋式电池是一种锂电池，其弯曲度有限。模拟结果显示，与采用方形或圆形等其他孔洞图案的电池相比，这种长方形孔洞的排列使电池能够在不撕裂的情况下被拉伸或折叠180度。

孔洞太多或太大都会降低储能容量，因此存在的一个挑战是，如何保持足够的活性物质，以维持电池的高能量密度。研究团队必须在机械拉伸性和电性能之间找到平衡。

这款布有孔洞的电池设计在拉伸10%甚至折叠后，仍能抵抗物理应力并继续为LED灯泡供电。其中，拉伸和折叠实验均进行了100次。在温湿度试验箱中的测试也显示，这款电池的透气性比棉花高出一倍。

团队将这款电池编织进实验服中，并在穿戴者跑步锻炼时测试了其性能。结果发现，孔洞设计使电池能够快速散热，让穿戴者避免了不适感觉或汗液滞留。

不过团队强调，这款电池仍需接受更多的耐磨测试，以及进一步研究如何扩大生产规模。


带有矩形孔的袋式电池。 图片来源：英国《新科学家》杂志网站