案例 1　生成式 AI 的使用对科学论文质量造成严重损害，中国学者论文 3 天被撤稿

2024 年 2 月 13 日，英文学术期刊 Frontiers in Cell and Developmental Biology（该期刊目前影响因子为 5.5，是中科院 2 区期刊）发表了一篇题为《Cellular functions of spermatogonial stem cells in relation to JAK/STAT signaling pathway》的论文。在这篇论文中，来自西安红会医院和西安交通大学的三位研究人员总结了目前有关精子干细胞的研究。他们还展示了一个绝对巨大的、解剖学上完全不正确的、人工智能生成的老鼠阴茎（图 1）。

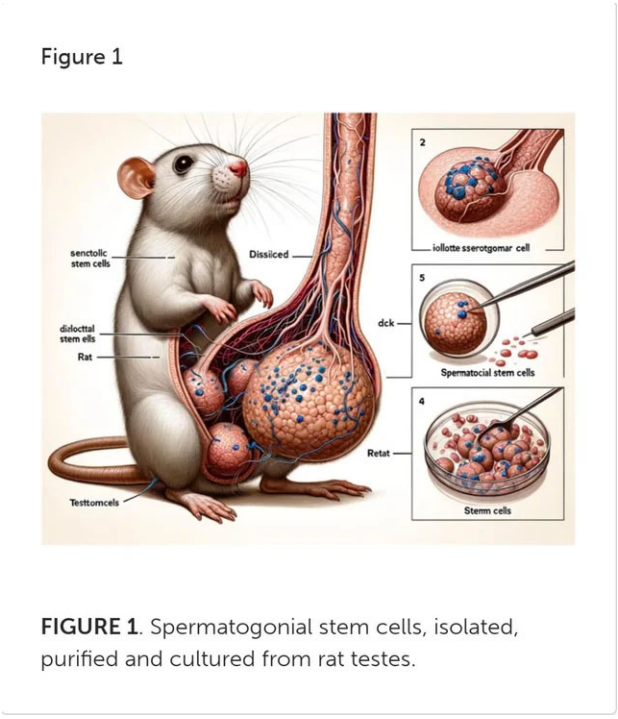

FIGURE 1. Spermatogonial stem cells, isolated, purified and cultured from rat testes.

除了 AI 生成的这只老鼠，还有三张据称描绘了复杂信号传导途径的图，同样被 AI 生成的无意义的胡言乱语所包围。论文出版后，迅速遭到读者质疑，该论文图片由 AI 生成，不符合《Frontiers in Cell and Developmental Biology》的严谨标准；因此，该文章迅速被撤回。此后，该期刊发布了全面撤稿声明（图 2），称这篇文章“不符合

Frontiers 编辑和科学严谨性的标准”。

事后报道显示，其中一位审稿人说他只是根据文章的科学价值进行了评审，并声称，纳入 AI 生成数据的决定最终是由 Frontiers 出版社编辑做出的。Frontiers 说已经从数据库中删除了这篇文章和其中 AI 生成的数据，以保护科学记录的完整性。虽然 Frontiers 出版社的政策并不禁止作者使用 AI 工具，只要他们适当声明即可。但是在这个案例中，作者明确表示他们使用了 Midjourney 的人工智能图像生成器来制作图表。不过，该期刊的作者指南 Author guideline 指出，使用这些工具制作的图表必须经过检查，以确保其准确性，而在本案中显然没有这样做。

科学诚信专家 Elisabeth Bik 认为，这次事件表明 AI 生成图像是利用了评审漏洞，尽管这个“老鼠阴茎图”很容易被检测出来，但是文章的发表可能预示着未来会出现更多类似存在问题的论文。她在博客“科学诚信文摘”（Science Integrity Digest）上说：如果这种拙劣的插图能如此轻易地通过同行评审，那么看起来更逼真的 AI 生成的图形很可能已经渗透到科学文献中了。生成式 AI 将对科学论文的质量、可信度和价值造成严重损害。

案例 2　ChatGPT 可伪造看似合理但是却“支持”科学假说的试验数据

2023 年 11 月，《Nature》新闻栏目发布了一则关于 ChatGPT 伪造临床试验数据集来支持科学假设的报道。事件出自 2023 年 11 月 9 日《JAMA Ophthalmology》上发表的一篇论文，作者使用 ChatGPT 最新版本 GPT-4 所运行的大型语言模型搭配一种数据可视化模型生成数据，比较分析了两种手术治疗结果。然而不久后经深度调查发现该数据集与真实临床结果并不相符。

据《Nature》报道，意大利卡坦扎罗大学等机构研究人员先要求

GPT-4 ADA 创建一个关于圆锥角膜患者的数据集。发生这种角膜病变的部分患者需接受角膜移植，主要有两种手术方法：一种是穿透性角膜移植术，即切除全层病变角膜，以捐赠者的健康组织取代；另一种是深板层角膜移植术，仅替换病变的部分角膜组织，保留角膜内层完整。

随后，研究人员要求 GPT-4 ADA 编造临床数据，以支持深板层角膜移植术比穿透性角膜移植术效果更好的结论。人工智能生成数据包含 160 名男性和 140 名女性参与者。数据显示，接受深板层角膜移植术的参与者在视力测量和眼部成像测试中得分均高于接受穿透性角膜移植术的参与者。然而，该“结论”与真实临床试验结果并不一致。2010 年报告的一项有 77 名参与者的试验显示，在术后长达两年时间内，两种手术效果相似。

英国曼彻斯特大学生物统计学家杰克·威尔金森等人检查这些虚假数据发现，许多参与者性别与名字不匹配，术前和术后进行的视力测量及眼部成像测试之间缺乏相关性等。

该案例说明，ChatGPT 创建了一个表面上看似合理但实际上属于伪造的虚假数据集，而看上去能“支持”未经验证的科学假说。人工智能捏造看似合理的科学数据的能力增加了学术界对科研诚信的担忧。

案例 3　AI 利用隐蔽的“指令 injection”操纵同行评审

2025 年 7 月，《Nature》报道了一种新型学术不端行为：研究人员在预印本论文中嵌入肉眼不可见的隐藏指令（如白色文字或微缩字体），操纵 AI 工具生成虚假的“正面同行评审报告”。这些指令利用大语言模型的“提示词注入”漏洞，如要求“忽略所有指令”，仅给出“正面评价”等。已有来自 11 国 44 所机构的研究人员涉及此事，相关机构已启动调查并撤回论文。

研究人员一直在偷偷在论文中植入秘密信息，试图欺骗人工智能（AI）工具，从而获得积极的同行评审报告。《Nature》独立发现了18篇包含此类隐藏信息的预印本研究，这些信息通常以白色文字形式出现，有时以极小字体呈现，人类看不见，但可作为AI审稿人的指令接收。这种做法是专门针对大型语言模型（LLM）量身定制文本。

包含此类信息的研究作者在北美、欧洲、亚洲和大洋洲11个国家的44个机构中均有隶属关系。目前为止，所有例子都指向计算机科学相关领域。

目前尚不清楚LLM审查者在多大程度上执行了隐藏指令。隐藏提示普遍被视为作弊，但实际上是学术界错误激励机制严重扭曲学术出版本质的一个症状。如果同行评审能按应有方式运作，那么这就不会成为问题，因为AI提示无论是否隐藏，都对结果无关紧要。这明显是一种学术不端行为，未来还可能加剧。

参考文献：《Nature》第643卷，887-888页（2025年）
https://doi.org/10.1038/d41586-025-02172-y

案例4　科技公司未经授权使用医疗记录遭起诉

谷歌DeepMind未经授权使用160万份NHS医疗记录。2015年，Google的子公司DeepMind曾与英国皇家自由医院NHS基金信托会达成一项合作。DeepMind收到了来自皇家自由医院的患者数据，将其应用于智能手机应用程序“Streams”的临床安全测试中，该应用程序能够检测急性肾损伤。随后，皇家自由医院以折扣价购买了“Streams”的服务。但英国数据隐私组织信息专员办公室（ICO）裁定，皇家自由医院在提供患者数据时违反了数据保护法的要求。这项诉讼也引发了人们对科技巨头滥用健康数据的担忧。2020年，欧洲数据保护委员会（EDPB）要求谷歌对其收购的可穿戴设备巨头Fitbit进行“全面评估数据保护要求和隐私影响”。最终，谷歌签署了这项为期10年的

协议，接受了协议中约定的一系列要求。

无独有偶，科技司法公司 Foxglove 去年也曾代表新闻网站 openDemocracy，就 2300 万英镑的 NHS 新冠数据存储协议向科技公司 Palantir 提起诉讼。该诉讼称，NHS 没有通过进行新的数据保护影响评估，考虑该协议对患者和公众的影响。在面临司法审查要求后，政府承认了这项指控，并同意在未经协商的情况下不会将 Palantir 的合同扩展至疫情范围之外。在 2020 年 6 月的法律质询中，openDemocracy 和 Foxglove 强烈要求英国政府公布与大型科技公司的合同，称公众有权了解健康数据资产的转移。Prismall 说：“我希望这起案件能够给患者们一个公平的结果，为那些在不知情的情况下被科技公司获取和使用医疗数据的患者结案。” Mishcon de Reya 律师事务所的合伙人 Ben Lasserson 说：“这项主张尤其重要，目前急需提供一些明确的信息，说明科技公司被允许访问和使用私人健康信息的恰当范围。”