大数据与科学的冲击

格雷·N·史密斯

格雷·N·史密斯 （Gary N. Smith）：美国波莫纳学院（Pomona College）经济学教授

电子邮箱：gsmith@pomona.edu

尽管 ChatGPT 和其他大型语言模型（large language models，简称为 LLM）因其对高等教育可能带来的颠覆而备受关注，但它们仅仅代表了大数据技术在学术与科研环境中快速发展和广泛应用所带来的一系列挑战之一。

自 2022 年 11 月 30 日 ChatGPT 面向公众发布以来，学生和教师几乎立即意识到，LLM 可以用来完成作业、应对考试、撰写论文。一个可能的积极结果是，许多教师会因此调整教学方式：不再过于依赖选择题测试和描述性论文（这是 LLM 的强项），而是更加重视学生真正需要而 LLM 无法掌握的批判性思维能力。毕竟，LLM 从字面上并不理解词语的含义，它们更像是一个可以背诵《罗马帝国衰亡史》六卷全部内容、却完全不理解其内涵的少年天才。

教授们同样可能会被诱惑去用 LLM 代写论文。事实上，计算机生成论文并非新鲜事。早在 2012 年，西里尔·拉贝（Cyril Labbé）和纪尧姆·卡巴纳克（Guillaume Cabanac）就发现了 243 篇由 SCIgen 生成器编写的论文，该程序通过随机组合词语制造虚假的计算机科学论文。相关的 19 家出版社声称其期刊都实行严格的同行评审，但即便粗略阅读这些论文也能看出它们毫无意义。

如今，完全虚构的论文更为普遍，因为 LLM 能生成语言流畅的文章，往往需要细读才能识破，而审稿人缺乏足够动力去仔细阅读。即便带有明显 LLM 痕迹的论文也可能通过审稿。例如，一篇发表于爱思唯尔期刊的论文开头是：“Certainly, here is a possible introduction for your topic”（当然，这里是你主题可能的引言），另一篇则写道：“I’m very sorry, but I don’t have access to real-time information or patient-specific data, as I am an AI language model.”（很抱歉，我无法获取实时信息或患者特定数据，因为我是一个 AI 语言模型。）而如果去掉这些明显标记，要识别出由 LLM 撰写的论文就更加困难。

P 值

大数据对科学的冲击远不止于 LLM。许多研究基于这样一个合理前提：研究者应评估其结果是否可能仅由随机性解释，例如在受试者被随机分配到实验组和对照组的情况下。标准的评估工具是 p 值，它表示仅由偶然产生的情况下，观察到的效果大于或等于实际观测效果的概率。

统计学家罗纳德·费希尔（Ronald Fisher）曾提出以 5% 作为显著性阈值：“我们可以方便地在某个界限上划线，说：‘要么处理有效，要么是巧合。’……我个人倾向于将 5% 设为显著性的低标准，并完全忽略未达到该水平的结果。”

然而，正如古德哈特法则（Goodhart’s

Law）所说：“一旦一个指标成为目标，它就不再是一个好的指标。”研究人员为追求 p 值低于 5%而努力，反而削弱了 p 值的意义。

一种做法是 p 值操纵（p-hacking），即反复调整模型和数据，直到 p 值降至 5%以下。例如，一项声称亚裔美国人更易在每月第四天心脏病发作的研究，省略了与结论相矛盾的数据。还有一项研究宣称以女性名字命名的飓风比男性名字的更致命；另一项研究则认为“力量姿势”（如双手叉腰）能提高睾酮并降低皮质醇。这些结论后来均被推翻。正如诺贝尔奖得主罗纳德 •科斯（Ronald Coase）嘲讽地说：“只要你拷问数据足够久，它们就会招供。”

另一种做法是“先有结果再提出假设”（HARKing），即在没有明确模型的情况下，随意寻找统计模式。例如，美国国家经济研究局（US National Bureau of Economic Research）曾研究比特币收益与 810 变量之间的相关性，包括加元兑美元汇率、原油价格、以及汽车、图书、啤酒行业的股价。在这些 810 个相关性中，有 63 个的 p 值低于 10%。而若纯粹用随机数与比特币价格做相关性，预期会有 81 个 p 值低于 10%。

可重复性危机

P 值操纵和“先有结果再提出假设”的做法导致了科学研究的“可重复性危机”，严重损害了研究的可信度。大量曾被媒体追捧的研究在使用新数据检验时被证伪。前文提到的四项研究均发表于顶级期刊，但无一可重复。

为了评估危机的严重程度，布莱恩 •诺赛克（Brian Nosek）带领的团队尝试重复三本顶尖心理学期刊的 100 篇研究，其中有 64 篇失败。柯林 • 坎麦尔（Colin Camerer）带领的团队重新检验了两本顶级经济学期刊发表的 18 篇实验经济学论文，以及《自然》（Nature）和《科学》（Science）上发表的 21 篇社会科学实验研究，结果有 40%无法复制。

在布莱恩 •诺赛克的的“可重复性项目”进行期间，研究人员甚至建立了拍卖市场，对尚未完成复现的 44 项研究下注，预测它们是否能成功复现——即结果的 p 值小于 5%，且方向与原始研究一致。其中有 46%的研究被认为复现成功的可能性不足 50%。但即使是这一悲观预测仍显过于乐观，最终有 61%的研究未能复现。

虚假论文、p 值操纵和“先有结果再提出假设”已存在数十年，但现代计算机和大数据使这些有缺陷的做法更加便利和强大。LLM 能够基于海量文本数据库，快速生成高度流畅却虚假的论文。大型数据库也使系统性 p 值操纵更易实现，研究人员可以通过更多方式操纵数据，直到得到所谓“统计显著”的结果。大数据还提供了几乎无限的可能性去寻找模式——任何模式——直到找到一个看似“显著”的结果。在所有这些情况下，结果都是有缺陷的，且不太可能被复现。

改革

为了重建科学的公信力，可以采取以下措施：

第一，期刊在发表实证研究前，应要求作者公开所有非机密的数据和方法。（目前，许多期刊要求作者在发表后共享数据和方法，但这一规定难以执行，常常被忽视。）

第二，期刊应为审稿人提供报酬，以鼓励其进行认真、全面的评审。在论文发表后，重复性与可复制性研究应得到公共或私人基金

资助，并作为博士或其他实证类学位的必修要求。如果研究人员知道自己的论文可能会被复查，他们就会更加谨慎。

保护科学免受大数据诱惑并非易事，但这是一场必须且值得进行的斗争。