从 ChatGPT 到 AGI:生成式 AI 的媒介特质与伴生风险（五）

二、生成式 AI 的伴生风险

4. 合成数据导致近亲繁殖，会让互联网信息出现劣币驱逐良币现象。OpenAI 在训练 GPT-5 时已经遇到文本数据不足问题，不得不考虑使用 Youtube 视频转录出的文本数据。当下，数据生产存量的增长速度远远低于数据集规模的增长速度。据人工智能研究机构 Epoch AI 在 6 月 4 日发布的论文预测，未来 10 年内数据增长速度将无法支持大模型的扩展，大模型将在 2028 年耗尽互联网上所有文本数据。按照当前趋势发展，文本数据耗尽的中位年份是 2028 年，最大可能性是 2032 年。整个互联网的文本数据总量约 3100T，但大部分数据分布在 Facebook、Instagram、WhatsApp 等社交媒体。由于抓取这些数据复杂且昂贵，且涉及个人隐私，几乎无法用于大模型训练。如何克服人类文本数据的瓶颈。第一种是利用 AI 生成数据，如 OpenAI 模型每天能够生成相当于 Common Crawl 中优质单词总数的 36. 5T 个单词，远快于人类生成文本的速度。第二种是利用多模态和迁移学习，超越文本数据从其他领域获取数据，比如视频、图像、金融市场数据或科学数据库。[1]

不过，这并非是解决问题的良策。如果网上大部分文本

都是 AI 生成的，而用合成数据训练的大模型会发生什么？大模型开发需要更多数据进行训练，而由 AI 生成的合成数据很快进入了训练新模型的数据集，并随着每一代模型而不断积累。越来越多证据显示，人工智能生成的文本，即使被引入训练数据集的量很少，最终也会对训练中的模型产生“毒害”。[2]研究人员将一些由 AI 生成的语料作为训练数据，“投喂”给一个正在训练的语言模型，然后使用它所输出的结果再来训练新模型，并重复这一循环。他们发现，模型每迭代一次，错误就会叠加一次。当人们要求第 10 次被训练出的模型写出有关英国历史建筑的内容时，它输出的却是有关豺兔的一堆胡言乱语。[3]英国牛津大学机器学习研究员伊利亚·舒迈洛夫及其同事称这种现象为“模型崩溃”。萨卡尔及其在西班牙马德里和英国爱丁堡的同事，用一种名为扩散模型的 AI 图像生成器进行了类似实验：第一个模型可以生成可识别的花朵或鸟类，但到了第三个模型，所生成的图片就变得模糊不清了。研究人员不得不寻找没有被污染的训练数据。随着 AI 生成的内容充斥互联网，它正在破坏未来大模型训练的数据。

如今，人工智能已经强势侵入人类的互联网，极大地改变了网上文本和图像的生成和传播系统。牛津大学、剑桥大学、帝国理工大学、多伦多大学研究人员发现，使用 AI 合成数据训练 AI，在进行 9 次迭代后，模型开始出现诡异乱码

进而直接崩溃，相关研究论文登上 Nature 封面。[4]研究人员发现，如果大模型在数据训练中不加区别地使用 AI 生成的内容，模型就会出现不可逆转的缺陷——原始内容分布的尾部（低概率事件）会消失。这种效应被称为“模型崩溃”。换而言之，这种合成数据就像是近亲繁殖，会产生质量低劣的后代。

当下，AI 生成内容已经进入机器学习工程师们所习惯于获取训练数据的领域。即使是主流新闻媒体也开始发布人工智能生成的文章，百科网站的编辑希望使用语言模型为网站生成内容。许多用来训练模型的现有工具，很快就会被 AI 生成的文本‘喂饱’”。韦谢洛夫斯基及其同事通过统计分析发现，已有约 1/3 的医学研究摘要有 ChatGPT 生成文本的痕迹。网文《中文互联网正在被 AI 污染》指出，AI 越来越火，但 AI 生成的垃圾信息也越来越多了。在 AI 的加持下，无意义的内容呈指数级增长，假新闻、标题党获得大量曝光。AI 不但没有解放生产力，反而劣币驱逐良币。[5]如果在网上搜索“AI 写文赚钱”，会有许多广告跳出来说，用 AI 写文章，只需复制粘贴，月赚上千元。

可以说，如今全球大模型已经陷入到高质量数据荒之中。但是，目前多数模型的训练数据都是从网上抓取数据，不可避免地会使用其他大模型生成的数据内容。后果就是，合成数据最终污染下一代模型的训练集，出现“模型崩溃（model

collapse）”现象。由于在被污染的数据集训练大模型，随后大模型会错误地感知现实。如果每一代新的模型都是在前一代生成的数据上进行训练，会导致多代 AI 生成模型的退化，也就是“垃圾进，垃圾出”。AI 合成数据，无异于给数据集“投毒”。杜克大学助理教授 Emily Wenger 在 Nature 上发表一篇社论文章指出：AI 基于自身数据训练，生成的图像扭曲了狗的品种。在初始数据集中，不仅有金毛、柯基，还有法国斗牛犬、小体巴塞特雪橇犬等。基于真实数据训练后的大模型，输出的图像中常见品种如金毛寻回犬占大多数，而不太常见的品种斑点狗会消失。然后，基于 AI 生成的数据训练模型，生成的品种全是金毛了。最终，经过多次迭代，金毛的图像完全出现混乱，脸不是脸鼻子不是鼻子大模型完全崩溃。此外，2023 年来自斯坦福和 UC 伯克利的一项研究中，作者同样发现，大模型在少量自己生成数据内容重新训练时，就会输出高度扭曲的图像。研究人员还发现，一旦数据集受到污染，即便大模型仅在真实图像上重新训练，模型崩溃现象无法逆转。为了大模型不再被自己“降级”，AI 需要能够区分真实和虚假内容。[6]

[1] https://mp.weixin.qq.com/s/EXB_a0ru9jhuY8bjw8Xj9g.

[2] https://www.whb.cn/zhuzhan/kjwz/20230823/535963.html.

[3] https: //www. whb. cn/zhuzhan/kjwz/20230823/535963. html.

[4] https: //www. nature. com/articles/s41586-024-07566-y.

[5] https: //mp. weixin. qq. com/s/f4bHNydpBFNo4W9MySHaRg.

[6] https: //www. nature. com/articles/d41586-024-02420-7.