人工智能(AI)迅速发展离不开对模型的训练。然而,高质量数据短缺以及部分领域封闭式的数据生态似乎成为AI发展的掣肘。

据多家外媒报道,OpenAI、谷歌和Meta等公司正寻求在线信息来训练最新的AI系统。但他们无视既定政策,蓄意改变规则,并试图规避版权法。

网络图片

繁荣背后存隐忧

美AI巨头被控秘密“侵吞”数据

1 收集数据“走捷径”

英国《泰晤士报》近日刊文指出,科技巨头一直在“走捷径”为其AI系统收集训练数据。OpenAI开发了一款名为Whisper的语音识别工具,可将YouTube视频中的音频文件转录为纯文本文档,从而创建一个口语对话数据源,帮助训练其下一代基于文本的GPT-4算法。

美国《商业内幕》网报道称,YouTube在其官网明令禁止“独立”于其之外的应用程序使用其视频内容。而OpenAI的数据并非意外收集的。

实际上,OpenAI员工知道这样做会涉足法律灰色地带。OpenAI总裁格雷格·布罗克曼甚至亲自参与了所使用视频的收集。但OpenAI依然认为这是合理的,最终获得了超过100万小时的转录视频。

最大的谜团在于,OpenAI如何访问足够多的YouTube视频来完成这项工作。

当OpenAI首席技术官米拉·穆拉蒂被问及该公司是否使用YouTube视频来训练Sora时,她表示并不确定。当再次被问及训练数据的来源时,她表示不会透露细节。

《纽约时报》称,与OpenAI一样,谷歌也转录了YouTube视频,为其AI模型收集文本,这可能侵犯了视频创作者的版权。去年,谷歌还更改了其服务条款。此番动机意图明显,即允许AI对来自谷歌文档中公开可用文档的数据以及上传到谷歌地图的餐馆评论等其他材料进行训练。

2 面临“数据瓶颈”

对于科技公司来说,庞大的数据“肥料”是生成式AI的核心养分,也是大模型发展的必争之地。唯有足够的数据才能指导技术即时生成与人类创作相似的文本、图像、声音和视频,实现系统创新。

但随着AI发展,现有互联网信息量的不足、高质量文本数据的匮乏以及科技巨头优质数据的垄断,都可能导致AI“养分不足”。即便谷歌和Meta拥有数十亿用户,每天都会产生搜索查询和社交媒体帖子,但这些数据在很大程度上受到隐私法和自身政策的限制,无法让AI利用这些内容。

这些科技公司的处境似乎十分窘迫。据人工智能研究机构Epoch称,科技公司最快将于2026年耗尽互联网上的高质量数据。这些公司使用数据的速度超过了产生数据的速度。

Meta同样也遇到了训练数据可用性限制。该公司打算采取一些措施,例如支付图书许可费用,甚至直接收购一家大型出版商。Meta也曾作出以隐私为中心的变革,因此它使用消费者数据的方式显然也受到了限制。

在人类数据告急的情况下,不少公司甚至试图用AI“喂”AI。包括微软、OpenAI在内的公司正在把大模型生成的结果,也就是所谓的“合成数据”,“喂”给参数更小的模型。但有研究认为,合成数据最终将让AI“自食其果”。

3 因版权被多方状告

《纽约时报》去年起诉OpenAI和微软,称其在未经许可的情况下使用受版权保护的新闻文章来训练AI聊天机器人。OpenAI和微软回应称,这属于“合理使用”,或者说是版权法允许的,因为他们为了不同的目的而改造了这些作品。

去年,超过1万个贸易团体、作者、公司和其他人士向美国版权局提交了有关AI模型使用创意作品的意见。

生成式AI的迅速兴起引发了一场全球性的高质量数据竞赛。然而,在这个新领域中,关于什么是合法的、道德的,没有明确规定。

《商业内幕》网称,目前,谷歌、OpenAI和其他科技公司正在辩解,认为将受版权保护的内容用于AI模型训练是合法的,但监管机构及法院尚未对此作出裁决。

美国电影制作人、前演员及作家贾斯汀·贝特曼告诉版权局,AI模型在未经许可或付费的情况下获取了其作品内容。她称,“这是美国最大的盗窃案”。

(据《科技日报》)