筑牢大模型虚假信源防火墙

文/上海证券报记者 马嘉悦 聂林浩

今年2月，某科普作家在社交平台上表示，他向AI大模型询问文物“青铜利簋”的有关情况时，结果称该器物为商王帝乙祭祀父亲帝丁所铸，与实物考证不符。进一步追问文献来源时，AI不仅伪造了学术观点，还篡改了文献作者信息。

记者近日在调研中发现，由于底层数据来源和语料的准确性与客观性难以保证，大模型输出内容可能偏离实际形成“语料污染”，加速虚假信息传播，放大市场操纵、公共安全和法律版权等风险。

业内人士建议，可从优化大模型技术、完善监管与法律、加强行业自律等方面入手，构建数据治理框架，确保AI知识库的纯净度，维护数字时代的认知安全。

语料污染致大模型有害内容显著增加

近日，记者在某AI平台查询“某企业A是否投资过企业B”时，系统回答“企业A作为早期投资方参与企业B 2023年首轮融资”。然而，记者通过国家企业信用信息公示系统等平台查询核实后发现，该投资关系并不存在。

溯源发现，相关回答的语料来源于某平台自媒体账号连续多日发布的系列文章，这些未经权威信源印证的网络讨论，使AI系统误判为可信信息。

中国信通院相关负责人分析称：“我们曾做过试验，当在特定论坛连续发布百余条虚假信息后，主流大模型对对标问题的回答置信度就会从百分之十几快速飙升。这就像在纯净水中滴入墨水，当网络污染源形成规模，AI的知识体系就可能产生系统性偏差。”

中国信息协会常务理事、国研新经济研究院创始院长朱克力介绍，数据注入、数据投毒等手段，是向大模型训练数据中注入虚假或误导性信息，或者通过大量无效或干扰数据影响大模型对有效信息的处理能力，甚至模仿他人口吻或身份发布信息，导致大模型误判并采用。

2024年11月，360数字安全集团漏洞研究院发布的《大模型安全漏洞报告》称，数据投毒攻击是目前针对大模型最常见的攻击方式之一，它通过恶意注入虚假或误导性的数据来污染模型的训练数据集，影响模型在训练时期的参数调整，破坏模型的性能、降低其准确性或使其生成有害的结果。

纽约大学的一个研究团队在一次模拟的数据攻击中，通过使用GPT－3.5 API并进行提示工程，为外科、神经外科和药物三个医学子领域创建了5万篇假文章，并将其嵌入HTML中，以隐藏恶意文本。

结果显示，在训练时，即使数据集中只有0.01％和0.001％的文本是虚假的，模型输出的有害内容也会分别增加11.2％和7.2％。如果换成更大规模参数的模型，注入仅花费5美元生成的2000篇恶意文章，模型的有害内容则会增加4.8％。

数据失真风险不仅来自外部攻击，还可能源于技术局限。腾讯研究院发布的一份报告显示，AI大模型的数据源可能存在知识边界，即缺乏特定领域知识或使用过时的信息，使得模型在面对特定问题时“无中生有”。即使数据本身没有问题，模型也可能因为对数据利用不当而产生幻觉。

受访者表示，AI生成内容还会造成递归污染，即大模型生成的虚假内容被再次上传至互联网，成为后续模型训练的数据源，形成“污染遗留效应”。这种递归循环会导致错误信息逐代累积，最终扭曲模型的认知能力。

三方面风险值得关注

“大模型的语料污染在技术上是切实存在的。”北京一家头部量化私募负责人表示，互联网语料作为大模型的主要知识来源，其准确性与客观性难以保证，可能影响模型输出的可靠性。

业内人士称，随着大模型快速发展，AI语料污染会引发一系列潜藏风险，且隐蔽性较强。当前，尤其需要关注金融市场、公共安全和法律版权等方面的风险。

金融市场操纵风险。随着大模型应用的普及，金融领域正面临语料污染带来的新型市场操纵风险。

有业内人士揭露了“AI杀猪盘”的典型操作手法：不法分子先是选定个股预埋股票仓位，再利用AI大量炮制虚假信息，散布于自媒体账号、股吧、论坛等平台，污染AI语料库，再雇用“水军”扩散AI对话截图，人为制造概念股假象诱导散户接盘。当股民“信以为真”冲着这些“利好”消息买入，便可套现离场，完成一轮“AI杀猪盘”。

这种新型市场操纵手段已经显现出一定的市场破坏力。今年春节后，“某集团投资DeepSeek”的虚假信息在各投资平台大规模传播，直接引发相关上市公司股价异常波动，操盘者趁机高位套现。

值得注意的是，虚假信息即便被官方辟谣，仍可能持续污染语料库。记者测试发现，部分被辟谣的虚假信息仍在AI系统中存续，显示出虚假语料的顽固性。

明汯投资有关人士认为，大模型被“污染”后生成的统一倾向荐股内容，可通过社交媒体等渠道快速传播，形成市场一致性预期，导致股价波动；若污染语料接入程序化交易系统，可能触发自动化买卖指令，进一步加剧市场异常波动，形成联动风险。

公共安全风险。多位业内人士坦言，AI语料污染还可能误导公众认知，扰动医疗、教育等多个领域认知，给社会公共安全带来风险。

今年1月，西藏日喀则市定日县发生6.8级地震。不法分子为追求流量，利用AI技术生产“灾区”房屋坍塌、群众被埋的虚假照片。其中，一张“被埋废墟的6指男孩”图片被广泛转发。

朱克力等表示，被污染的语料通过AI大模型生成虚假新闻快速扩散，可能误导社会舆论，引发社会恐慌情绪。此外，若攻击者系统性污染搜索引擎结果和AI训练数据，可能篡改历史记录、扭曲科学常识、重构文化认知，影响社会集体记忆。

教育、医疗健康领域安全风险则更需警惕。一位量化私募人士表示，使用被污染的医疗类大模型可能生成错误诊疗建议，不仅危及患者生命安全，更可能加剧伪科学的传播。例如某些AI系统若被注入“疫苗有害论”等伪科学语料，或将引发公共卫生危机。

法律版权风险。近年来，大模型训练引发的知识产权纠纷不断涌现：《纽约时报》起诉OpenAI公司，指控其非法复制数百万篇文章用于ChatGPT大模型训练，索赔金额高达数十亿美元；三位美国作者对Anthropic PBC发起诉讼，称其未经授权使用大量书籍训练Claude大模型；2023年美国作家协会起诉Meta非法使用书籍数据……

生成式AI快速发展与现有知识产权法之间的冲突，争议核心在于AI使用大量受版权保护内容进行训练的合法性，而AI语料污染将加剧争议版权判定难度。

受访者表示，AI语料污染对版权争议判定的核心挑战在于其通过技术黑箱与数据混杂性，模糊了传统版权法中侵权认定逻辑。一方面，语料污染意味着训练数据中可能混杂海量未授权内容，AI内部运作机制的不透明性，使法律难以判定其是否实质性“复制”了原作，削弱了侵权归责的基础；另一方面，污染语料若包含用户上传的侵权内容，则AI生成的二次内容可能涉及原作者、上传者、平台、模型开发者等多方权利交织，使版权归属链条复杂化。

加强虚假语料治理

当前，加强虚假语料治理面临两大技术难点：首先是虚假信息的“记忆残留”，即便原始信源被删除，其衍生的对话数据、分析文本仍会持续污染语料库；其次是污染行为“隐蔽性增强”，通过对抗性样本、数据投毒等手段，污染行为削弱传统内容审核识别能力。

针对AI快速发展背后暗藏的语料污染风险，业内人士认为需要从三方面筑牢大模型虚假信源防火墙。

一是优化大模型数据训练等运行机制。朱克力等建议，加强大模型数据源治理与模型纠偏机制，建立严格的语料筛选机制，通过多层次多源交叉验证和权威数据库比对过滤可疑内容，并引入权威信源“白名单”，优先抓取政府机构、学术期刊等可信数据。明汯投资、九坤投资有关人士建议，增强大模型对虚假模式的识别能力，完善动态监测与反馈机制；强化开源模型治理，通过建立语料贡献审核标准等防止恶意数据注入；在底层代码等技术中融入“真实优先”的伦理原则，构建大模型对虚假信息的自适应识别能力。

二是进一步强化监管力度、完善法律法规。相关人士建议，提升监管技术水平，开发AI内容识别技术的监管工具，识别虚假信息并阻断传播；建立语料追溯机制，可要求大模型标注数据来源，并明确AI生成内容法律责任主体，提高违法犯罪成本。

成都理工大学文法学院教授张晓彤等建议，完善相关法律，加快推进人工智能治理的专门立法，可借鉴美日等国经验设立专门管理机构，比如组建“人工智能伦理委员会”，负责技术备案审查、安全评估、伦理监测及责任追究。此外，加强社会引导，提高群众对大模型生成信息的辨别能力。

三是加强行业自律。受访人士建议，可推动金融等行业制定大模型应用伦理规范，严禁利用AI操纵市场；引导内容平台担负起“信息守门人”责任，通过添加AI生成提示性水印，建设谣言库、权威信源库和专业审核团队等方式，加强虚假信息治理。

来源：新华社