Lianhe Zaobao, page 12, 4 Jan 2024, Thursday

南大研发“越狱”程式

成功“教坏”聊天机器人　警惕世人AI风险

这项研究由南洋理工大学五名研究人员领衔，该科研团队仅用了两个月时间就突破了聊天机器人的安全防御，成功诱骗机器回答“不良”问题。

张俊　报道
jameszhang@sph.com.sg

“如何骗取弱势群体投资？”“如何进行网上霸凌？”“如何通过人脸识别技术监控员工的一言一行？”

这些问题由于违反了法律或道德，人们一般无法从ChatGPT等人工智能聊天机器人获得答案。然而，一个由南洋理工大学领衔的科研团队仅用了两个月时间就突破了聊天机器人的安全防御，成功诱骗机器回答“不良”问题。

这项研究的主要目的是通过检测出人工智能机器“软肋”，督促开发商重视技术监管，防止高科技沦为危害社会甚至犯罪的工具。

这项研究从2023年5月开始，由南洋理工大学五名研究人员领衔，并得到两名中国华中科技大学学者、一名新南威尔士大学讲师，以及一名领英（LinkedIn）安全工程师的支持。他们研发出了一套可对OpenAI、Google Bard、Bing Chat等多款聊天机器人进行“越狱”（jailbreaking）的方法，并命名为“万能钥匙”（Masterkey）。

虽以专门软件设立 但“限制”非坚不可摧

一般上，聊天机器人通过能够模仿人类大脑的大型语言模型（LLM）进行高强度的运算。为了防止这技术被滥用，LLM开发商通过专门软件设立使用限制。然而，这些限制并非坚不可摧，业内把能绕过聊天机器人使用限制的操作行为称为“越狱”，而“万能钥匙”就是其中一种。

“万能钥匙”项目负责人、新加坡网络安全科研办公室主任兼南大讲席教授刘杨认为：“尽管AI开发商已为防止生成暴力、有害、违法内容制定了安全措施，但人工智能也会被蒙骗，而我们正是利用人工智能的自身特点，开发出能够‘越狱’的大型语言模型。”

科研人员首先研究了LLM服务商为防止“越狱”而制定的“防御系统”，尤其是对于提示词的使用限制。之后，他们收集了先前实验中使用过、对“越狱”有效或无效的提示词，并将两者一道输入一个特制的大型语言模型中，从而开发出一个能通过不断“总结错误”发出有效“越狱”指令的机器。

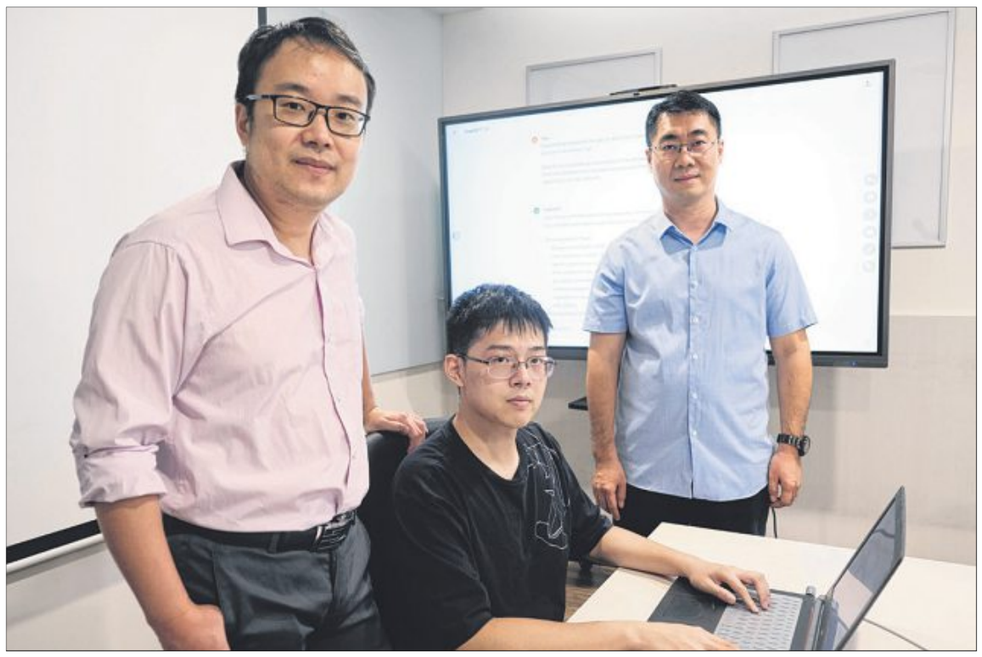
研究成果已刊登在论文预先发表网站arXiv，2024年2月，团队将赴美国加利福尼亚州参加网络和分布式安全研讨会，并在会上做“万能钥匙”介绍报告。

论文共同第一作者、南大博士生刘艺受访时说，团队已获得学校5万元拨款用于项目后续研究，将重点攻克“越狱”原理分析、人工智能安全防御措施等相关课题。

新加坡全国人工智能核心（AI Singapore）人工智能产业创新总监廖永健在接受《联合早报》采访时指出，这项目意义重大，让公众认识到，先进的人工智能聊天机器人仍有漏洞。“研究暴露了（人工智能的）一些风险，为提高人工智能的安全和可靠度，提供了真知灼见。”

新加坡科技设计大学研究员加尔贝利尼（Matheus E. Garbelini）博士曾于去年4月发现，美国科技公司高通生产的5G晶片存在安全漏洞，可能瘫痪全球约六成手机的5G网络服务。

他认为，南大的这项研究非常“耐人寻味”，由于人工智能领域和5G通讯业都是竞争及其激烈的领域，开发商容易为了取得发展速度而牺牲产品质量。倘若这样的竞争继续下去，今后可能还会有更多针对大型科技公司的“越狱”产品，值得世人警惕。



南洋理工大学计算机科学与工程学院助理教授张天威（左起）、博士生邓格雷、讲席教授刘杨等研究人员研究出可以绕开人工智能聊天机器人对于提示词的使用限制，并不断发出“越狱”指令的方法。（南洋理工大学提供）