



答非所问、凭空捏造、虚假链接…… AI为啥“一本正经胡说八道”



当前AI正赋能千行百业,为人们的工作、学习、生活带来极大便利。与此同时,不少人发现,用AI搜索数据,给出的内容查无实据;用AI辅助诊疗,出现误判干扰正常治疗……AI频频上演“一本正经胡说八道”。社交平台上,“AI幻觉”相关话题浏览量达数百万,网友吐槽涉及金融、法律、医疗、学术等多个领域,“AI幻觉”一时引发热议。

AI不时像是“中邪”了

用AI检索海量信息、让AI辅助查看三维病灶、打造AI互动课堂……如今,AI已深度融入现代生活,“人工智能+”产品赋能各行各业,从多个维度提供便利。

作为AI深度使用者,“95后”女生瑞希坦言,AI好用,但有时像“中邪”了一样胡说八道。“我让AI推荐10本高分小说,结果多半都是它编的。反复确认后,它承认虚构了答案。”

现实生活中,不少人遇到相似情况。业内人士表示,这是由于AI幻觉导致。“AI可以快速给出答案,但生成内容可能与可验证事实不符,即凭空捏造;或生成内容与上下文缺乏关联,即‘答非所问’。”一名主流人工智能厂商技术人员说。

记者使用一款AI软件,让其给出某行业未来市场规模及信源,AI迅速回答称某投资机构预测2028年该行业的市场规模将达到5万亿美元,并提供相关链接,但链接页面找不到上述信息。记者看到,页面内容虽然包含该投资机构名称和5万亿美元表述,但预测数据并非该机构作出,且不存在2028年时间节点。

社交平台上,AI幻觉相关话题浏览量达数百万,网友吐槽涉及金融、法律、医疗、学术等多个领域。

第三方咨询公司麦可思研究院近期发布的2025年高校师生AI应用及素养研究显示,四千余名受访高校师生中,近八成遇到过AI幻觉。今年2月,清华大学新媒沈阳团队发布的报告指出,市场上多个热门大模型在事实性幻觉评测中幻觉率超过19%。

AI幻觉已经影响了人们的生活与工作。近期,一名国外男子被诊断出溴中毒。他此前询问AI,过量食用食盐不利于身体健康,有无食盐替代品,AI回答称可以用溴化钠代替。但溴化钠存在一定毒性,需要严格遵医嘱服用。该男子用溴化钠代替食盐三个月后出现精神错乱等症状。

这几年,美国多起案件中的律师因在法律文件中使用AI生成的虚假信息,被法院警告或处分。



观众在人工智能技术大会上体验AI学习

AI幻觉为什么会发生

受访专家认为,AI幻觉的背后存在多重因素。

——数据污染。AI“养成”过程中,数据“投喂”是关键环节。研究显示,当训练数据中仅有0.01%的虚假文本时,模型输出的有害内容会增加11.2%;即使是0.001%的虚假文本,其有害输出也会相应上升7.2%。

奇安信集团行业安全研究中心主任裴智勇解释说,人工智能大模型需要海量数据,训练数据来自开源网络,难免会错误学习一些虚假、谬误数据,还有一些不法分子会恶意进行“数据投毒”。

“如果把AI比作一个学生,数据污染就像是给学生看了错误的教科书,自然会导致‘胡说八道’。”暨南大学网络空间安全学院教授翁健说。

——AI本身“认知边界模糊”。翁健认为,人类智能的一个重要特征是“元认知”能力——知道自己懂什么、不懂什么,而当前AI技术架构缺乏这种自我认知机制。

翁健解释称,AI可以博览群书,但并不一定理解书里的内容,只是根据统计规律把最有可能的词语组合在一起,在准确评估自身输出的可信度方面尚存盲点。

——人为调和和干预。在中国通信学会数据安全专业委员会副主任委员左晓栋看来,相较于事实真相,AI更在意自己的回答

是否契合用户需求,从而导致AI有时为了“讨好”用户而编造答案。

“针对不同需求,AI的训练、打分方式也不同。”一位从事大模型训练的技术人员说,当面对写作等创意性需求时,偏理性的事实严谨在打分系统中占比相对较低,偏感性的词语优美、富有感情色彩等占比更高。“所以可能会出现一篇辞藻华丽但词不达意的文章,里面内容甚至与事实相悖。”

多方合力减少AI幻觉

第55次《中国互联网络发展状况统计报告》显示,截至去年12月,有2.49亿人使用过生成式人工智能产品,占整体人口的17.7%。受访专家表示,应通过多方合力应对AI幻觉带来的风险挑战。

今年4月,中央网信办印发通知,在全国范围内部署开展“清朗·整治AI技术滥用”专项行动,训练语料管理不严、未落实内容标识要求、利用AI制作发布谣言等均为整治重点。

“可靠、可信、高质量的数据对降低AI幻觉非常重要,应优化人工智能的训练语料,用‘好数据’生成‘优质内容’。”左晓栋认为,可以加快推动线下数据电子化,增加“投喂”的数据量;同时探索建立具有权威性的公共数据共享平台,“各大厂商也应加强优质数据筛选,提升训练准确性。”

多家主流人工智能厂商已经采取措施,从技术层面减少AI幻觉发生。

豆包升级深度思考功能,由先搜后想变为边想边搜,思考过程中可以基于推理多次调用工具、搜索信息,回复质量明显提升;通义千问在20多个通用任务上应用强化学习,增强通用能力的同时纠正不良行为;元宝持续扩充引入各领域的权威信源,在回答时交叉校验相关信息,提高生成内容的可靠性。

翁健建议,建立国家级人工智能安全评测平台,就像生物医药新药上市前要做临床试验一样,大模型也应该经过严格测试;同时,相关平台加强AI生成内容审核,提升检测鉴伪能力。

相关新闻

AI很“忙” 别当法“盲”

随着AI技术不断发展,一些新型违法犯罪行为开始冒头,给网络空间安全和群众人身财产安全带来威胁,记者采访多地公安机关,揭示犯罪手法,提升防范意识。

“你们公众号怎么开始推荐投资理财APP了?靠谱吗?”今年6月10日,某机构工作人员被公众号评论区留言惊出一身冷汗。

工作人员发现,其运营的公众号不知何时竟发文称即将停更,并号召粉丝关注另一个投资理财类账号。工作人员尝试登录账号后台,发现不仅密码被修改,连公司法人代表信息都被篡改了。

意识到事态严重,工作人员报案。接警后,武汉网警迅速成立专案组。顺着线索追踪,专案组很快锁定了远在山东潍坊的犯罪嫌疑人阿成。据他交待,2022年5月,他在外网接触“人脸代过”灰色产业,加入相关群组后,先在各类小型电商平台接单制作人脸图像,每张收费200元;随着技术提升,他掌握了生成动态人脸视频的方法,“报价”也水涨船高,破解一张AI动态的人脸最高能卖到1000元。案发时,已非法获利40余万元。

AI时代已来临,新技术能做的事越来越多,也越来越有想象力。但筑牢安全底线的第一守则就是:AI可以很“忙”,但使用它的人不能法盲。掌握AI技术的专业人员,切勿因贪欲走上违法犯罪道路。

据新华社



外卖“新规”征求意见 定位视频“一镜到底”治理“幽灵外卖”

市场监管总局组织起草的《外卖平台服务管理基本要求(征求意见稿)》,24日起正式面向社会公开征求意见。征求意见稿聚焦平台收费、促销行为等重点问题,帮助外卖平台企业规范服务管理,提升服务质量,减轻商户经营负担,引导平台企业公开有序竞争。

近年来,随着平台收费规则、计费方式、包括补贴要求等越来越复杂,常常导致商户出现看不懂、弄不明、算不清的情况,难以准确判断成本投入和收入预期,甚至出现卖得越多、亏得越多的局面。

专家介绍,征求意见稿明确限定了外卖平台向商户的收费项目,不得随意新设收费项目,并从公示方式、公示内容、结算明细、推广效果等方面对平台收费作了细化规定,通过提升各类费用信息披露的完

整性与易理解性,防止不透明收费。

征求意见稿针对“幽灵外卖”、商户“爆单”等乱象,从商户入驻、信息审核等方面提出系统要求,严守食品安全底线。

所谓“幽灵外卖”,是指没有营业执照、食品经营许可证等证照,或不具备餐饮经营资质的商户,通过假证、套证、借证等违法违规手段在外卖平台提供违规餐饮服务,此类商户借助虚假证照与地址信息违规上线经营,埋下食品安全隐患。

这次征求意见稿要求商户提供真实证照、门店场景信息,甚至是“一镜到底”的定位视频,并且由平台执行商户准入与动态审核,辅以“网上亮证”的社会监督,在以信息真实性为抓手的情况下,能够从源头阻断违规商户的进入通道。

据央视新闻



外卖员驾驶电动自行车在送外卖的路上