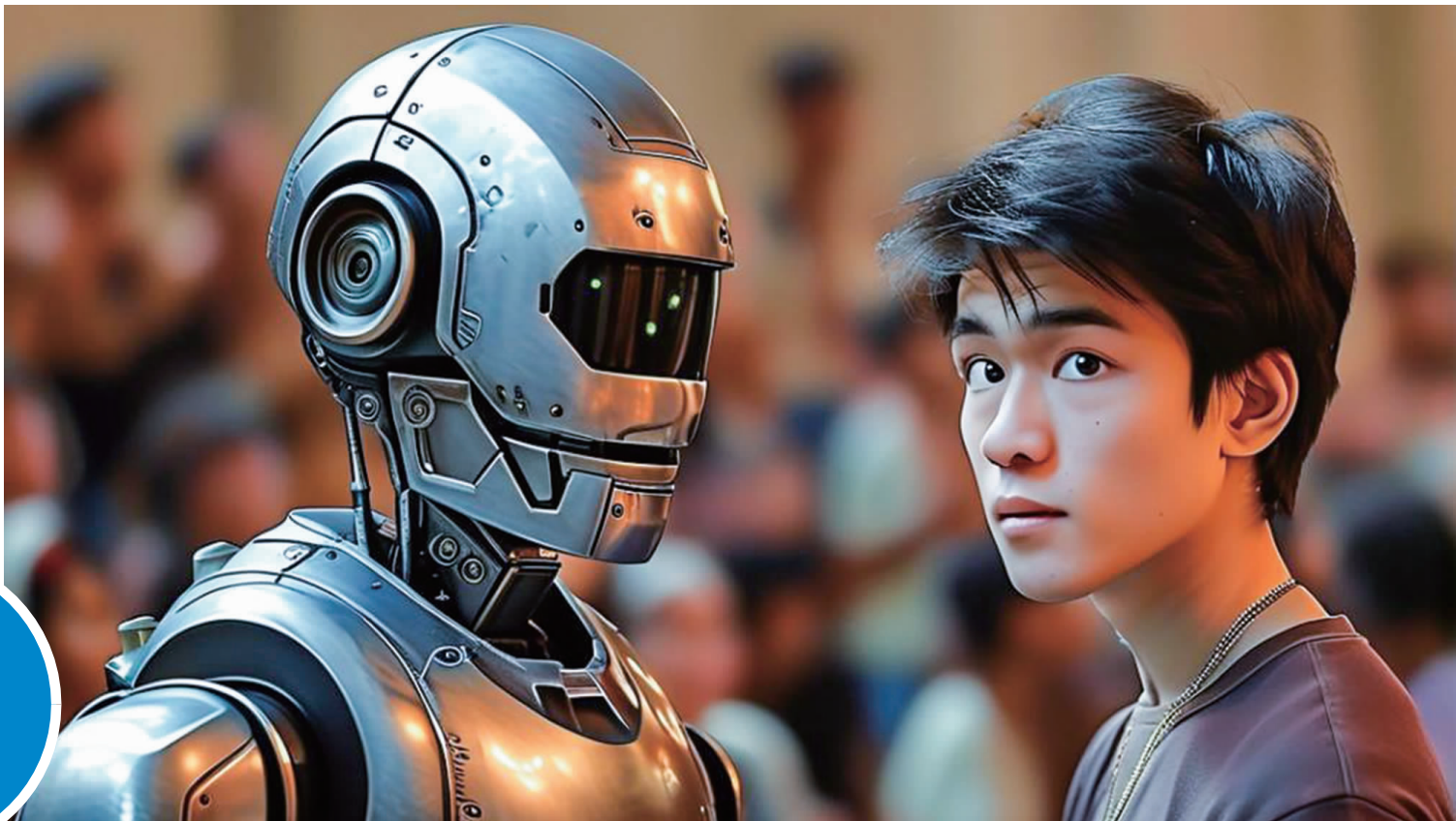
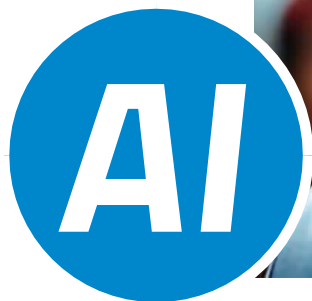


当 AI“一本正经地胡说八道”

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“一本正经地胡说八道”。社交平台上, AI 幻觉引发热议。



当前 AI 技术架构缺乏自我认知机制 (图为 AI 生成 新华网 发)

A AI 好用 但不时像是“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, 但不时像“中邪”了一样胡说八道。“我让 AI 推荐 10 本高分小说, 结果一多半都是它编的。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, 4000 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 3 个月后发现精神错乱等症状。

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。

B 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”。

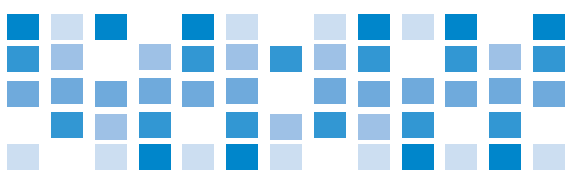
“如果把 AI 比作一个学生, 数据污染就像是给学生看了错误的教科书, 自然会导致‘胡说八道’。”暨南大学网络空间安全学院教授翁健说。

——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。”



C 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 制作发布谣言等均列为整治重点。

“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, 提升检测鉴别能力。

“AI 可能‘欺骗’用户, 公众应客观认识人工智能的局限性。”左晓栋等专家提示, 可以通过改进使用方式, 如给出更加明确的提示词、限定范围等避免 AI 幻觉。“无论是工作、学习还是生活, 现阶段的人工智能还不能全面替代人类的认知和创造能力, 大家在使用 AI 时要保持怀疑态度和批判思维, 不过度依赖 AI 给出的回答, 多渠道验证核查。”

新华社记者 颜之宏 胡林果