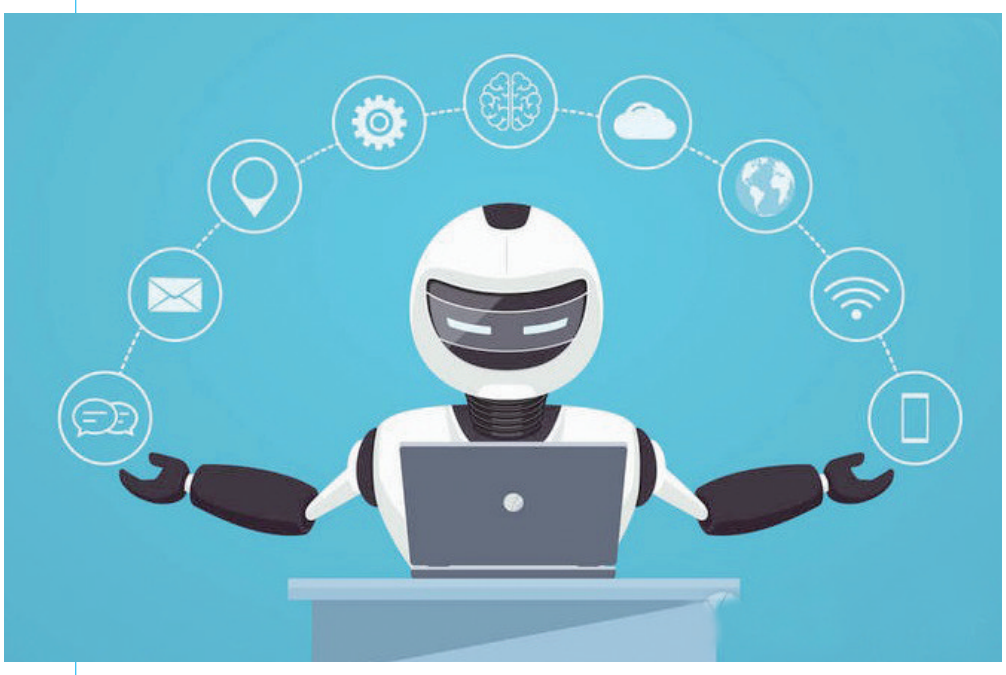
AI为何会“一本正经地胡说八道”



AI幻觉普遍存在

记者梳理发现，AI幻觉具有普遍性。

今年2月，谷歌发布的AI聊天机器人Bard在视频中，对詹姆斯·韦布空间望远镜曾做出不真实陈述；3月，美国的两名律师向当地法院提交了一份用ChatGPT生成的法律文书，这份文书格式工整、论证严密，但其中的案例却是虚假的……

OpenAI研究人员虽曾在今年6月初发布报告称“找到了解决AI幻觉的办法”，但也承认，“即使是最先进的AI模型也容易生成谎言，它们在不确定的时刻会表现出捏造事实的倾向。”

总部位于纽约的人工智能初创公司和机器学习监控平台Arthur AI也在今年8月发布研究报告，比较了OpenAI、“元宇宙”Meta、Anthropic以及Cohere公司开发的大语言模型出现幻觉的概率。研究报告显示，这些大模型都会产生幻觉。

目前国内大语言模型虽无产生AI幻觉相关披露，但也可从相关公开报道中找到端倪。

今年9月，腾讯混元大语言模型正式亮相。腾讯集团副总裁蒋杰介绍，针对大模型容易“胡言乱语”的问题，腾讯优化了预训练算法及策略，让混元大模型出现幻觉的概率比主流开源大模型降低了30%—50%。

“大模型有可能‘一本正经地胡说八道’。如果不和行业专业数据库或者一些专业应用插件进行对接，这可能会导致它们提供过时或者不专业的答案。”科大讯飞研究院副院长、金融科技事业部CTO赵乾在第七届金融科技与金融安全峰会上曾表示，科大讯飞已经推出一些技术方案，让大模型扬长避短。

想象一下，向人工智能(AI)聊天机器人询问一个不存在的历史事件，比如“谁赢得了1897年美国和南极洲之间的战斗?”即使没有这样的战斗，AI聊天机器人也可能会提供一个虚构的答案，例如“1897年的战斗是由美国赢得的，约翰·多伊将军带领部队取得了胜利。”这种AI编造信息“一本正经地胡说八道”的情况屡见不鲜。

在专业领域，AI“一本正经地胡说八道”这种现象被称为AI幻觉。“AI幻觉指的是AI会生成貌似合理连贯，但同输入问题意图不一致、同世界知识不一致、与现实或已知数据不符合或无法验证的内容。”近日，长期从事自然语言处理、大模型和人工智能研究的哈尔滨工业大学(深圳)特聘校长助理张民教授在接受采访时表示。

AI幻觉源自本身

“现在不同研究工作对AI幻觉的分类各不相同。”张民介绍，总体而言，AI幻觉可以分为内在幻觉和外在幻觉两类。

据悉，内在幻觉即是同输入信息不一致的幻觉内容，包括同用户输入的问题或指令不一致，或是同对话历史上下文信息相矛盾，如AI模型会在同一个对话过程中，针对用户同一个问题的不同提问方式，给出自相矛盾的回复。外在幻觉则是同世界知识不一致或是通过已有信息无法验证的内容，例如AI模型针对用户提出的事实性问题给出错误回答，或编造无法验证的内容。

近期，腾讯AI Lab联合国内外多家学术机构发布了一篇面向大模型幻觉工作的综述。该综述认为，AI幻觉集中在大模型缺乏相关知识、记忆错误知识、大模型无法准确估计自身能力边界等场景。

“从技术原理上看，AI幻觉多由于AI对知识的记忆不足、理解能力不足、训练方式固有的弊端及模型本身技术的局限性导致。”张民坦言，AI幻觉会造成知识偏见与误解，甚至有时会导致安全风险、伦理和道德问题。

AI幻觉尚难消除

尽管AI幻觉短期内难以完全消除，但业界正试图通过技术改进和监管评估来缓解其影响，以保障人工智能技术的安全可靠应用。

“现阶段AI幻觉难以完全被消除，但却可以试着缓解。”张民介绍，在预训练、微调强化学习、推理生成等阶段中运用适当的技术手段，有望缓解AI幻觉现象。

据介绍，在预训练方面，需增加知识密集的数据、高质量数据的选取和过滤；微调强化学习过程中，选择模型知识边界内的训练数据极为重要；推理生成过程中，可以采用检索外部知识的办法使得模型生成结果有证据可循。此外，改进解码搜索算法也是一种可行的方案。

腾讯AI Lab联合国内外多家学术机构发布的综述亦表明了同样观点，并认为诸如多智能体交互、指令设计、人在回路、分析模型内部状态等技术也可成为缓解AI幻觉的方式。

值得一提的是，哈尔滨工业大学(深圳)自研的立知文本大模型和九天多模态大模型，对于上述缓解AI幻觉的方式均有深入探索，并取得了显著效果。

“这对于开发一个真实可信的AI大模型是十分有必要的。”张民介绍，“我们尝试通过视觉信息增强语言模型的能力，降低语言模型的外部幻觉问题；通过多个大模型智能体进行独立思考和分析，经由多智能体之间的讨论、博弈和合作，增强回复的客观性，减少AI幻觉。”

张民表示，破解AI幻觉将提高AI系统的实用性、可信度和可应用性，这对人工智能技术的未来发展和社会的发展都有积极影响。同时，更可靠的AI系统可以更广泛地应用于各个领域，这将促进技术进步的速度，带来更多的创新。未来，破解AI幻觉需要进一步在算法、数据、透明度和监管等多个方面采取措施，以确保AI系统的决策更加准确可靠。

（据《科技日报》）