AI为何会“一本正经地胡说八道”

◎本报记者 **罗云鹏**

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科技日报记者采访时表示。

图为一名男子正在与一个机器人对话。在输出内容的过程中，人工智能有时会出现幻觉，“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。

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。

大模型的下一步该怎么走

◎本报记者 **都芃**

在近日举办的第六届世界声博会暨2023科大讯飞全球1024开发者节上，科大讯飞股份有限公司（以下简称科大讯飞）正式发布讯飞星火认知大模型V3.0，在文本生成、语言理解、知识问答、逻辑推理、代码能力、数学能力、多模态能力等7个方面较上一个版本进行了智能升级。

面向行业找到刚需应用场景

本次发布会上，除了发布讯飞星火认知大模型V3.0外，科大讯飞还一口气发布了面向工业、法律、金融等12个行业的专属行业大模型。

如何深入行业，一直是困扰大模型发展的难题。科大讯飞董事长刘庆峰认为，面向具体行业，找到刚需的应用场景是推动大模型迈向产业、实现商业价值的重要基础。

科大讯飞此次发布的12个行业大模型，瞄准的正是许多行业中的智能化痛点。

“要做出符合行业特点和需求的大模型，就必须要与真正懂行业、懂应用场景的龙头企业合作。”刘庆峰表示，行业大模型在龙头企业内部打造成功后，可以对全行业进行赋能，所以合作企业要有开放的精神。

刘庆峰认为，一整套方便易用的训练工具能对行业专属内容进行高效训练。这对于大模型在行业中落地至关重要。“我们需要开发出定制化工具，让企业自己‘拖拖拽拽’就能解决问题。”他说。

针对行业定制大模型成本居高不下的现状，刘庆峰认为，大模型要先找到典型应用场景，和行业龙头企业一起做好共性场景的应用。在大模型具备一定的行业通用能力后，需要定制的内容会越来越少，成本自然会下降。

此外，他还表示，大模型要想真正赋能千行百业，需要实现从多轮对话、主动对话再到启发式对话的跨越。大模型不仅要能回答问题，还要能像人一样主动提问。

着眼于此，讯飞星火认知大模型在此番更新中便新增了个性化AI人设功能。该功能可以为讯飞星火认知大模型形成一个初始“性格”，使大模型具备长期稳定的记忆力、多样化的个性和丰富的情感，再结合特定知识学习、对话记忆学习，形成一个更个性化的AI人设。

找到方向形成完整自主生态

大模型的训练和应用离不开由显卡搭建的算力平台，这让算力再次成为中国人工智能发展过程中的焦点。在此次发布会上，除了讯飞星火认知大模型V3.0，最受关注的便是科大讯飞与华为联合发布的国产算力平台“飞星一号”。刘庆峰表示，讯飞星火认知大模型V3.0正是在国产算力平台上训练出来的大模型。

华为轮值董事长徐直军在发布会现场表示，华为的使命和愿景是让每个人、家庭、组织畅享数字世界，构建万物互联的智能世界。在全面智能化战略的指引下，华为将持续打造坚实的算力底座，最终让所有对象可联接，所有决策可计算，让大模型真正赋能千行百业。

徐直军透露，讯飞星火认知大模型V3.0在“飞星一号”平台上训练效率翻番，能够实现更为高效稳定的训练迭代。接下来，华为还将为更大参数的讯飞星火认知大模型V4.0版本提供有力支撑。

自主创新的算力底座是中国大模型发展的重要基础。刘庆峰表示，虽然目前使用国产化算力平台仍然存在一定迁移成本，但这一步是必须迈出的。

接下来，科大讯飞将在“飞星一号”平台的基础上，启动对标GPT-4的更大参数规模的讯飞星火认知大模型V4.0的训练。刘庆峰表示，这对于科大讯飞来说，绝不仅仅是一个简单的技术对标问题。“我们要走出自己的技术路线，走出自己的产业方向，形成自己完整的生态。”他说。

人工智能让法律咨询更高效更便捷

◎本报记者 **吴纯新**
通讯员 **邱 婵 李 进 赵 婧**

作为2023中国5G+工业互联网大会的平行会议，人工智能赋能新型工业化会议日前在湖北省武汉市举办。会上，北京大学武汉人工智能研究院副院长、北京大学法学院党委副书记杨晓雷发布了法律垂直大模型以及基于大模型能力研发的“北大元法智能系统”。

杨晓雷说，该系统由北京大学武汉人工智能研究院智能法律实验室和北京大学法律人工智能实验室科研团队联合攻关、全栈自主研发。基于团队自研的百亿参数中国法律垂直领域大模型，“北大元法智能系统”实质性地解决了大模型在法律领域的AI幻觉问题和在真实法律服务场景中的不足，展现出大模型应用于法律领域的新能力。

同时，该系统在东湖高新区国家智能社会治理实验综合基地支持下，将为人工智能在法律领域的应用开创道路，有望实现法律技术的智能化转型升级。

在杨晓雷看来，该系统融合了三段论的推理逻辑与精准的事实要素抽取能力。它彻底抛弃了传统法律咨询系统罗列法条的机械咨询方式，能够准确地理解咨询者的自然语言，主动引导法律咨询过程，从缺乏法律知识的使用者的描述中准确归纳事实和问题，在多轮对话的过程中完成复杂的法律推理，从而给出可操作性强的法律咨询报告。

在会场外的体验区，记者体验“北大元法智能系统”时发现，该系统操作页面简单实用。目前，该系统正在相关法律领域落地应用。

杨晓雷说，法律人工智能在价值本体层面无法替代人，但在工具价值层面，可以有效为法律职业工作赋能。

“北大元法智能系统”是北京大学交叉学科研究的重要成果。基于该系统，还可以研发面向公众的可知、可信、可靠的法律援助服务应用，面向法律工作者的案情总结工具和法律检索工具，以及面向企业法务的人工智能合规官系统等。

北京大学武汉人工智能研究院执行院长吴志强表示，将人工智能应用于法律领域，是一项具有价值挑战性和创新性的探索。“北大元法智能系统”旨在提高法律工作效率与法律治理水平，为社会各界提供更好的法律服务。

据悉，接下来，北京大学武汉人工智能研究院将继续深化与北京大学法学院的合作，支持“北大元法智能系统”升级与迭代，将该系统运用到立法、执法、司法、企业法务等法律场景，为公众提供普惠、可知、可信、可靠的法律咨询服务，助力智能法治社会建设。

进一步提升人工智能产业化水平

要增强“四个能力” 需下足“三项功夫”

◎**洪恒飞** 本报记者 **江 耘**

“大模型的兴起，使得人工智能应用的深度和广度进一步拓展。”在2023年世界互联网大会乌镇峰会期间，百度首席技术官、深度学习技术及应用国家工程研究中心主任王海峰表示，人工智能已进入工业大生产阶段，并且其产业化水平正逐渐提高。

人工智能要具有很强的通用性，才能进入工业大生产阶段。如何判断人工智能的通用性强不强？这要看其理解、生成、逻辑和记忆这四项人工智能基础能力的强弱。“这四项基础能力越强，人工智能就越接近通用人工智能。”王海峰解释道，大语言模型不仅具备这四项基础能力，而且正变得越来越“聪明”。这为人工智能迈向通用人工智能带来了曙光。

进入工业大生产阶段，意味着人工智能已经在多个产业有所应用。据了解，今年3月，百度发布新一代知识增强大语言模型文心一言。目前，文心一言的基础模型已迭代到文心大模型4.0。自8月31日面向全社会开放至今，文心一言的用户规模已达7000万，并已在4300个场景落地应用。王海峰介绍，文心大模型4.0基于更强平台、更优数据、更好算法进行训练，是规模更大、效果更佳的知识增强大语言模型。相比前几代文心大模型，文心大模型4.0的四项人工智能基础能力均有显著提升。

人工智能要想进一步增强其产业化水平，就要在标准化、自动化和模块化三个方面下功夫。王海峰认为，标准化，即人工智能框架、模型的联合优化和多硬件的统一适配，这可使人工智能的应用模式更加简洁高效，并大幅降低其应用门槛；自动化，即从全流程提升人工智能研发效率；模块化，即为人工智能配备丰富的产业级模型库，使其更加广泛、便捷地应用到各个产业。

目前，大模型已经成为推动人工智能发展的重要引擎。有鉴于此，王海峰认为，提升人工智能产业化水平，还可以通过改进现有大模型生产模式来实现。比如，可采用“集约化生产，平台化应用”的模式，让具有算法、算力和数据综合优势的企业将模型生产的复杂过程封装起来，通过低门槛、高效率的生产平台，为千行百业提供大模型服务。

目前，这一产业化模式已在文心大模型产业实践中得到验证。记者在采访中了解到，百度与其合作伙伴共建了包括能源、金融、航天等10余个行业大模型，正加速推动人工智能产业化落地。

“运动项圈”、新式电子耳标等人工智能设备已经在一些奶牛场中应用。不仅在奶牛场，人工智能正在步入千行百业，进入工业大生产阶段。

新华社记者 **马希平**摄

图说智能

AI智慧供热：节能低碳又舒适

近年来，山东省枣庄市市中区将城市供热系统进行了人工智能（AI）智慧化改造，做到精准供热、按需送热、实时管热，堵住了过去漏热、不热、不能实时控温等现象，既降低了损耗，又有利于节能减排，为广大居民创造了一个舒适、温暖、低碳化的生活环境。图为工作人员在市中区AI智慧供热控制中心工作。

本版图片除标注外由视觉中国提供