2025年5月9日 星期五
责任校对 宗建玲
责任编辑 刘洋
技术编辑 刘娜

编造难以辨明真假的细节,生成与事实相悖的答案

小心,AI开始胡说八道

2025年2月,如果不是长期从事人口研究的中国人民大学教授李婷的公开辟谣,很多人都真诚地相信了一组数据——“中国80后累计死亡率为5.20%”。

在社交媒体上,许多“80后”都曾因这组数据扼腕叹息。“截至2024年末,80后的死亡率已经超过70后,相当于每20个80后中,就有1人已经去世。”自媒体传播道。

这一说法很快露馅。李婷教授在受访时表示:“(死亡率5.2%)错误非常明显,因为专业统计数据中死亡率会用千分率表示,而不是百分率。”她指出,国家统计局并未公布2024年的死亡率,也不会根据“80后”“90后”等分段公布死亡人数,因此这一说法毫无数据支撑。

虚假的死亡率数据从何而来?李婷认为:很有可能来源于AI大模型出错。她曾尝试在AI大模型中输入问题:“50后、60后、70后、80后,这几代人的死亡率分别是多少。”大模型表示:“根据网络信息,80后现存2.12亿,存活率94.8%,死亡率5.2%。”

AI无中生有的能力让人心颤。在AI业界,这类“胡说八道”的本领被称为“幻觉”,意思是,AI也像人产生心理幻觉一样,在遇到自己不熟悉、不在知识范围的问题时,编造难以辨明真假的细节,生成与事实相悖的答案。

此事件中,让人畏惧的是由技术蔓延出的不可控。新浪新技术研发负责人张俊林告诉记者,随着各个领域都在加强对AI的接入,AI幻觉成为了现阶段需要重视的问题。但遗憾的是,业界还没找到根除AI幻觉的办法。

清华大学长聘副教授陈天昊也在受访时提到,对于学生等特殊人群来说,大模型幻觉问题带来的风险性可能更大。“比如,小学生可能和家长一起使用大模型学习知识,但大模型产生的幻觉可能会产生误导。在自身缺乏辨别能力的情况下,可能难以判断信息的真假。”

2025年,人人都开始用AI,而AI还在持续发挥想象力,用幻觉与假信息误导更多人。现在是时候一起面对AI这个巨大的漏洞了。

过度自信

“想和大家说一件最近让我忧虑的事,是关于AI幻觉强度的。”2月,知名科普作家河森堡在微博中表示。

他在近日使用ChatGPT时,让它介绍文物“青铜利簋”。结果,ChatGPT将这件西周文物的来历,编造成了商王帝乙祭祀父亲帝丁所铸。AI此后还标明了自己的文献来源,源自《殷墟发掘报告》《商代青铜器铭文研究》等。

“看着是那么回事,其实又在胡扯。”河森堡发现:“前一篇文献的作者是中国社会科学院考古研究所,AI说是中山大学考古学系,后一篇文献的作者是严志斌,AI说是李学勤……”

错漏百出的生成信息还不算什么,可怕的是,AI还会自我“包装”,编造信息来源,让人误以为内容十分专业且可信度高。

在豆瓣,陀思妥耶夫斯基的书迷在使用AI的“联网搜索”功能时,发现其不懂装懂、捏造细节。

例如,有书迷问AI,“陀思妥耶夫斯基的哪部小说引用了涅克拉索夫的诗歌?”在引用了11个参考网页后,AI生成了大段的、看似专业的答案,论证了两者是好友,作品之间存在相互影响的关系。结论是,“陀并未在其小说中直接引用涅克拉索夫的诗”。

而事实上,熟悉陀思妥耶夫斯基的书迷很快想到,在《地下室手记》第二章开头,他引用诗歌:“当我用热情的规劝/从迷雾的黑暗中/救出一个堕落的灵魂,你满怀深沉的痛苦/痛心疾首地咒骂/那缠绕着你的秽行。”这正是涅克拉索夫的诗。

张俊林告诉记者,AI大模型非常容易“过度自信”。但目前,AI生成答案的过程仍像一个黑箱,AI业界也不完全清楚AI的自信从何而来。总之,在面对自己不懂的专业问题时,极少有AI会直接回答“不知道”;它们宁愿自信地、流畅地生成一些不准确的内容。

“DeepSeek幻觉有点太严重了,我受不了了。”法学硕士生小昭在写论文时感叹。她对记者承认,平时学习和写论文时,自己已经离不开DeepSeek、豆包、Kimi等AI工具。“因为(不用的话)我更写不出来。”

但是小昭逐渐发现,AI生成的内容,有很多是错误的。一个重灾区是关于“深度伪造”的法律问题,她发现AI会生成虚假的法律条例和案例。

此外,她在用AI准备公务员面试时,AI很喜欢给她引用一些过于具体的数据,“很多数据很明显是保密数据,一看就是AI编造的”。

AI生成的内容看上去“过于专业”,小昭说,这时反而是“唬人的”,“内容根本没法用”。

一次,在写AI深度伪造法律论文时,DeepSeek告诉她,不同年龄段法官对技术行为的评价呈现显著差异。它因此生成了一张表格,把30岁以下、30岁-50岁、50岁以上的法官对待技术的裁判倾向分列其中。

最后,它甚至写道,代际的差异在合议庭评议中会引发新的冲突。2023年,我国某中级法院在一次审理深度伪造案件中,“80后”和“60后”法官曾出现了激烈争论。

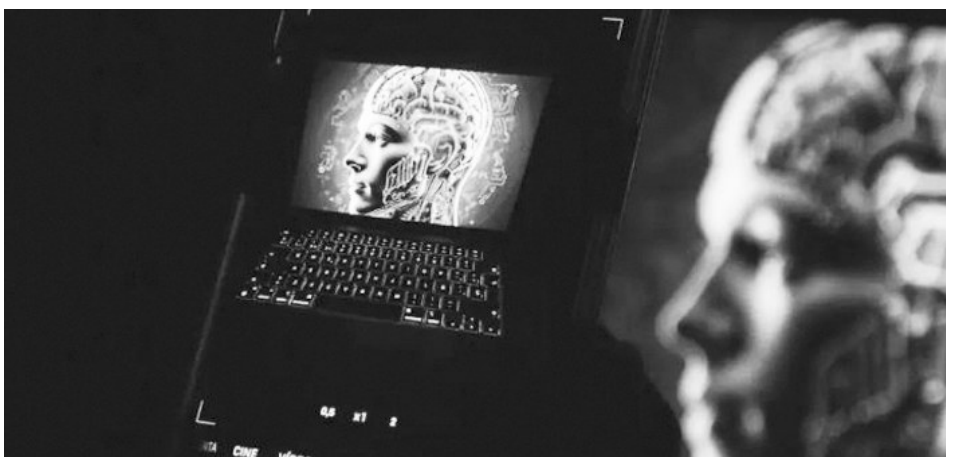
但经过调查和搜索,小昭发现,上述内容也全部是AI编造的。面对AI,即使她此后给出了“减少对虚假案例的引用,扩写分析部分”的指令,AI仍止不住地出现幻觉,生成虚假信息。

于是,在高频使用豆包、DeepSeek,以及OpenAI的o1等AI工具后,小昭的发现是,豆包的幻觉问题不算明显,语言相对平实;OpenAI的o1对中国国情不够熟悉,“国内素材没有那么充足”。而DeepSeek是其中最好用的工具,语言专业又生动,但DeepSeek编造细节的情况却是最严重的。

“以至于每次看到DeepSeek引用的,我都要重新检索,确认下真实性。”小昭说。

“张冠李戴”的天性

小昭等“AI原住民”的感受并不虚妄。在Github上一个名为Vectara大模型幻觉测试排行榜中,2025年1月发布的DeepSeek R1,幻觉率高达14.3%。这一数字远高于国际先进大模型,例如,OpenAI的GPT-4o幻觉率为1.5%,马斯克的Grok幻觉率为4.6%。



为何DeepSeek的幻觉率这么高?一个最直接的原因是,张俊林说,DeepSeek生成的内容比一般的AI应用更长。AI生成的内容越多、文本越长,出错以及胡编乱造的可能性随之更大。

另一个可能性在于,DeepSeek在生成答案时展现出了很强的创造性,这与强调信息精确、降低幻觉率的要求天然地相悖。张俊林提到,AI大模型有一个“温度系数”(Temperature),指的是控制生成内容随机性和多样性的参数。

一般而言,高温度系数(如1.0或更高)的模型,生成内容随机性更高,可能会出现更多新颖或意想不到的结果。代价便是其更容易出错、胡说八道。相反,低温度系数的模型,生成内容更接近训练数据中的模式,结果更稳定,但缺乏多样性。

幻觉率的高低,关系到我们到底想要什么样的AI——究竟是更能给予人灵感的,还是逻辑严密的。而在业界,一个共识是,无论想要什么样的AI,幻觉问题仍非常难消除。

清华大学团队在2025年2月发布《DeepSeek与AI幻觉》报告,将AI幻觉分为两类,一类是事实性幻觉,指生成的内容与可验证的现实世界事实不一致。例如,模型错误地回答“糖尿病患者可以通过吃蜂蜜代替糖”。

另一类则是忠实性幻觉,指的是AI生成的内容与用户的指令、上下文或者参考内容不一致,如果这个事实在训练数据中没有足够的信息冗余度,模型就记不住。接着,当用户问起AI失去记忆的内容,他便开始自动脑补,一本正经地胡说八道。

而且,AI的脑补绝非空想。李维解释,大模型的知识学习过程(训练阶段)是一种信息压缩过程;大模型回答问题,就是一个信息解码过程(推理阶段)。这就好比升维了又降维。当一个事实冗余度不够,AI生成的虚假事实需要与补白所要求的价值类型匹配,即符合本体/分类法相应的上位节点概念。

因此,AI“补白”的时候,即使不知道“狗”这类哺乳动物,也会生成哺乳动物“猫”,而不是像麻雀一样的鸟类。

“大模型是天生的艺术家,张冠李戴、指鹿为马等在大模型的幻觉里非常自然,因为张和李是相似的,马和鹿也在同一条延长线上。”李维总结道。“它的幻觉有内在的合理性,但同时也更具有迷惑性。初接触大模型的朋友,需要特别小心,不能轻信。”

不要神化

与人类有些相像的地方是,幻觉的出现,AI自己也表示“知道”。

“询问”DeepSeek:“为什么你的回答有时很准确,有时不太准确甚至在编造呢?”在深度思索模式下,DeepSeek列出了自己的“反思”:

首先是知识边界限制,“我的训练数据截至2023年12月,无法获取最新信息;知识库存在信息盲区(约10%-15%的领域覆盖不全)……”

它表示,生成机制特性也导致了这一结果,因为AI并不真正理解语义与知识,而是“基于概率预测生成(每个token选择概率前3候选词)”。再加上其采用流畅度优先机制,生成过程要先确保流畅度,而非保证事实。

诚如DeepSeek所言,AI的幻觉与其技术发展相伴相生,有时候,拥有幻觉本身,可能是AI感到骄傲的。在科学界,AI的幻觉正被很多科学家用于新分子的发现等科研工作。

例如,在AI+生物领域,麻省理工学院教授汤姆·克林斯在《自然》发布论文指出,AI的幻觉加速了他对新型抗生素的研究进展。“我们得以成功让模型提出完全新颖的分子。”

但这并不意味着,解决或改善幻觉问题对现有的AI大模型不重要。原因也很简单,随着AI持续渗透人们的生活,AI幻觉所带来的信息污染很可能进一步影响人们的生活与工作。

2月,美国知名律师事务所Morgan & Morgan向其1000多名律师发送紧急邮件,严正警告:AI能编造虚假的判例信息,若律师在法庭文件中使用这类虚构内容,极有可能面临被解雇的严重后果。这一声明正是考虑到AI在法律界被滥用后可能造成的不良后果。

据路透社报道,在过去两年间,美国多个法院已对至少七起案件中的律师提出警告或处分,因其在法律文件中使用AI生成的虚假信息。

例如,曾经入狱的前特朗普律师迈克尔·科恩,在2024年承认自己错误地使用了谷歌Bard生成的判例为自己申请缓刑。但他提交的文件中,由AI生成的至少三个案例,在现实中均不存在。

意识到AI幻觉可能产生的巨大副作用,科技公司并非没有行动,例如,检索增强生成技术(RAG)正被诸如李彦宏等科技大佬所提倡。RAG的原理是,让AI在回复问题前参考给定的可信文本,从而确保回复内容的真实性,以此减少“幻觉”的产生。

不过,这样的方案也绝非一劳永逸。首先因为,RAG会显著增大计算成本和内存,其次,专家知识库和数据集也不可避免地存在偏差和疏漏,难以覆盖所有领域的问题。

OpenAI华人科学家翁荔在一篇万字文章中写到,一个重要的努力方向是,确保模型输出是事实性的并可以通过外部世界知识进行验证。“同样重要的是,当模型不了解某个事实时,它应该明确表示不知道。”

谷歌的Gemini模型也曾做过很好的尝试。该系统提供了“双重核查响应”功能:如果AI生成的内容突出显示为绿色,表示其已通过网络搜索验证;内容如果突出显示为棕色,则表示其为有争议或不确定的内容。

这些努力都在预示着一个正确的方向:当AI幻觉已经不可避免地出现时,人们要做的首先是告诉自己:不要全然相信AI。(应受访者要求,文中小昭为化名)

据南风窗