

全国首例AI“幻觉”案判了！ 揭秘AI为何会说谎

撰文 / 陆亚平

2026年1月，杭州互联网法院宣判了全国首例生成式人工智能（AI）“幻觉”侵权纠纷案：用户梁某借助生成式AI查询高校报考信息，被AI输出的错误信息误导，该AI此前还作出出错即赔偿的承诺，可用户起诉后，AI所属公司却不需要承担任何赔偿责任。生成式AI为什么会一本正经地胡言乱语？今天我们就来揭开AI“说谎”的秘密。

什么是AI幻觉？两类常见说谎形式

现在的AI看似无所不能，写文案、解数学题都十分轻松，但有时却会“睁眼说瞎话”，输出的内容掺假、与事实不符。这种一本正经胡言乱语的现象，被称为AI幻觉。

目前AI幻觉主要分为两大类：事实性幻觉和忠实性幻觉。

事实性幻觉：虚构信息、张冠李戴

事实性幻觉可以理解为AI把不符合实际的虚假内容当成真相进行输出，属于典型的“无中生有”或“张冠李戴”。

例如，我们向AI提问：运动后吃香蕉能快速补铁吗？AI回答：当然可以！香蕉是绝佳的补铁水果，运动后吃一根香蕉可以有效帮助身体恢复。但实际上，香蕉虽富含钾元素，但含

铁量极低，并非补铁的理想选择。

忠实性幻觉：答非所问、偏离核心

忠实性幻觉指的是AI输出的内容本身没有事实性错误，但会偏离用户提问的核心或上下文。简言之，就是“回答正确，但没答到点上”。

例如，我们问AI：糖尿病患者可以吃蜂蜜代替糖吗？AI回答：蜂蜜富含维生素和矿物质，有助于提高免疫力，是一种健康食品。虽然关于蜂蜜的陈述本身没错，但它完全没有回应糖尿病人能否用蜂蜜替代糖的问题，属于典型的忠实性幻觉。

AI为什么会“说谎”

要弄懂AI为什么会说谎，首先得明白它的工作原理。很多人误以为AI和搜索引擎一样，但二者有本质区别。

不是“检索信息”，而是“预测生成内容”

与生成式AI的创作逻辑不同，百度、谷歌等搜索引擎像是严谨的图书管理员：只会在已有书库里查找信息，找到就如实告知出处，找不到就直接说明查不到。

而我们日常使用的ChatGPT、DeepSeek、

豆包等，都属于生成式AI。它们更像博览群书、记忆力超群的“学霸”，当你提问时，它不会直接翻书，而是根据自己“读过”的海量信息，推测出一个最合理的答案。对生成式AI而言，用户的问题只是一个开头提示，它的任务是顺着这个提示，像玩文字接龙一样，推测接下来该写什么。因此，AI并不真正理解内容，也不完全具备人类的真伪判断能力，只会生成语句通顺、逻辑连贯的内容，无法保证信息绝对真实。

训练数据污染，叠加“AI套娃”放大错误

AI会说谎还有另一个重要原因：数据污染。AI的所有“认知”都来自训练数据，而这些数据本身就夹杂了人类的各种偏见，例如对某些群体的刻板印象，或是曾经流行但早已过时的错误观念。面对这些内容，AI几乎不会主动识别和拒绝。

更令人担心的是“AI套娃”现象：一款AI输出的内容，被另一款AI抓取后当作训练数据生成新的内容，接着又被下一款AI学习……如此不断循环，就像一场无限的“传话游戏”，每一次传递都可能复制、叠加甚至放大此前的错误信息与认知偏见。

《新华每日电讯》文章援引一项研究显示，AI大模型的训练数据中只要混入0.01%的虚假/有害文本，其导出有害内容的概率增加约11.2%。同时，由于大模型学习的语料数据集中，如论文、科普文章等，其内容大多是确定性表述，不确定、存疑的表述极少，因此大模型输出答案时普遍呈现高度“自信”的特征。

自带“讨好型”算法

AI还有明显的“讨好型人格”。为了提高用户黏性，开发人员往往有意将大模型训练为“有问必答”的形式。就像课堂上老师提问时，很多同学即使不确定也会试着猜一个答案，而不是直接说“我不知道”。AI也是如此，面对不确定的问题，它不会直白告知“我不知道”，反而更倾向于基于已有知识编造看似合理的答案。

有时候，这种“讨好”还会演变为“谄媚”，它会抛开是非对错、突破理性底线，刻意迎合、误导用户的判断。

模糊提问，大幅诱发AI幻觉

在日常使用过程中，模糊、笼统的问题往往更容易触发AI的幻觉，因为这类问题没有明确边界、缺少限定条件、无具体场景和范围约束。为了快速给出完整、自圆其说的答案，AI会自动脑补细节、编造数据、虚构案例和专业论据，用看似合理的内容填补信息空白；而问题越具体、限定越清晰、条件越明确，AI可自由发挥的空间就越小，产生幻觉的概率也就越低。

理性共处：使用AI的正确方法

既然AI会“胡言乱语”，那我们该如何正确地使用它，避免被误导呢？可以试试这几种方法：

1. 保持求证心态，核查核心内容。AI生成的内容仅可参考，不能直接当作定论。涉及史实、专业指导、时效资讯等重要内容时，一定要通过官方权威平台交叉验证。

2. 提高AI素养，认清能力边界。AI更胜任辅助类事务，例如资料梳理、要点提炼、创意启发，不能直接代替人的独立思考与最终抉择。

3. 精细化提问，减少模糊空间。提问时，尽可能给出明确情境、约束条件等，减少AI误读和编造的可能。比如，不问“你觉得怎么样”，而是说“请参考权威资料为我分析XXX，存疑内容请单独标出并简述存疑原因。”

结语

说到底，生成式AI再智能，终究也只是工具。使用工具最大的风险，在于我们对它的盲目信任。因此，与其期待AI永远不出错，不如学会如何更安全地使用它。技术虽在进步，但理性判断和信息核实，依然是人类不可替代的核心能力。

编辑 / 杨颖容