# 如何识破人工智能在一本正经地瞎编乱造

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/27756.html



**如何识破人工智能在一本正经地瞎编乱造**。 你能看得出人工智能在瞎编乱造吗？《自然》6月19日发表的一项研究报道了一种能检测大语言模型（LLM）幻觉（hallucination）的方法，该方法能够测量生成回答的含义的不确定性，或者用于提升LLM输出的可靠性。

像ChatGPT和Gemini这样的LLM是能够阅读和生成自然人类语言的人工智能系统。不过，这类系统很容易产生幻觉，生成不准确或没有意义的内容。然而检测LLM出现幻觉的程度很困难，因为这些回答的呈现方式可能让它们看起来很可信。

在这项研究中，英国牛津大学的Sebastian Farquhar和同事尝试了量化一个LLM产生幻觉的程度，从而判断生成的内容有多少是忠于提供的源内容的。

研究人员的方法能检测出LLM的编造（confabulation）——这是幻觉的一个子类别，指不准确和随意的内容，常出现在LLM缺乏某类知识的情况下。

这种方法考虑了语言的微妙差别，以及回答如何能以不同的方式表达，从而拥有不同的含义。他们的研究表明，该方法能在LLM生成的个人简介，以及关于琐事、常识和生命科学这类话题的回答中识别出编造。

然而澳大利亚皇家墨尔本理工大学的Karin Verspoor在一篇同时发表的新闻与观点文章中指出，该任务由一个LLM完成，并通过第三个LLM进行评价，等于在以毒攻毒。Verspoor还写道，用一个LLM评估一种基于LLM的方法似乎是在循环论证，而且可能有偏差。

不过，Verspoor指出，他们的方法有望帮助用户理解在哪些情况下使用LLM的回答时需要多加小心，也意味着可以提高LLM在更多应用场景中的置信度。（来源：中国科学报 赵路）

相关论文信息：https://doi.org/10.1038/s41586-024-07421-0

作者：Sebastian Farquhar 来源：《自然》