“以毒攻毒”！识别大模型“一本正经胡说八道”

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/27720.html

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

“以毒攻毒”！识别大模型“一本正经胡说八道”

。6月18日，《自然》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，该方法能检测生成回答的含义的不确定性，或能用于提升LLM输出的可靠性。

LLM（如ChatGPT和Gemini）是能阅读和生成人类自然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，即“一本正经地胡说八道”。不过，检测LLM出现幻觉的程度很难，因为这些回答的呈现方式可能会让它们看起来很可信。

来自英国牛津大学的Sebastian Farquhar和同事尝试量化一个LLM产生幻觉的程度，进而判断生成的内容有多“忠于”提供的源内容。他们的方法能检测“编造”（confabulation）——这是“幻觉”的一个子类别，特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，这一方法能在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“胡说八道”的内容。

不过，Sebastian Farquhar等人的研究方法，也离不开大模型这一得力工具。《自然》同时发表的“新闻与观点”文章指出，该任务由一个大语言模型完成，并通过第三个大语言模型进行评价，相当于是“以毒攻毒”。

该文作者同时也在担忧，用一个大模型评估一种基于大模型的方法“似乎是在循环论证，而且可能有偏差”。不过，作者认为，他们的方法有望帮助用户理解在哪些情况下使用LLM的回答需要注意，也意味着可以提高LLM在应用场景中的置信度。

相关论文信息：https://www.nature.com/articles/s41586-024-07421-0

作者：赵广立 来源：中国科学报

更多 科学进展 请访问 https://www.iikx.com/news/progress/

本文版权归原作者所有，请勿用于商业用途，爱科学iikx.com转发