人工智能的幻觉越来越严重，而且会持续下去

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/33285.html

人工智能的幻觉越来越严重，而且会持续下去
。来自美国OpenAI和谷歌等科技公司的人工智能（AI）聊天机器人在过去几个月中一直在进行所谓的推理升级——理想情况下，它们将更好地提供人们可以信赖的答案。但最近的测试表明，它们有时比以前的模型做得更差。聊天机器人所犯的错误被称为“幻觉”，从它们诞生以来就是一个问题，现在看来，人们可能永远无法摆脱它们。

AI生成的内容中往往会出现错误。

图片来源：Paul Taylor/Getty Images

幻觉是大型语言模型（LLM）所犯的某些类型错误的总称，这些模型为OpenAI的ChatGPT或谷歌的Gemini等系统提供支持。它们有时会把错误信息当作真实信息呈现。幻觉也指AI生成的答案是正确的，但实际上与所问的问题无关，或者在某些方面没有遵循指示。

OpenAI的一份技术报告评估了其最新的LLM，显示其今年4月发布的O3和O4-mini模型的幻觉率明显高于2024年末发布的O1模型。例如，在总结关于人的公开事实时，O3有33%的时间、O4-mini有48%的时间产生了幻觉。相比之下，O1的幻觉率为16%。

这个问题并非仅限于OpenAI。美国Vectara公司发布的评估幻觉率的排行榜显示，一些“推理”模型——包括中国DeepSeek公司开发的DeepSeek- R1模型在内，与之前开发的模型相比幻觉率上升了两位数。这类模型在响应之前会通过多个步骤展示推理过程。

OpenAI表示，推理过程本身不应该受到指责。OpenAI的一位发言人表示：“幻觉在推理模型中并不是天然地更普遍，我们正在努力降低O3和O4-mini中更高的幻觉率。”

但LLM的一些潜在应用可能会因幻觉的存在而失败。一个不断陈述错误并需要事实核查的模型不是一个有用的研究助手；一个引用虚构案例的律师助理机器人会让律师陷入麻烦……

AI公司最初声称，这一问题会随着时间推移而解决。事实上，最开始，模型的幻觉往往会随着更新而减少。但最近版本的高幻觉率使这一说法变得复杂——无论推理本身是否有错。

Vectara的排行榜根据模型在总结它们所给文档时的事实一致性进行排名。Vectara的Forrest Sheng Bao说，这表明“推理模型与非推理模型的幻觉率几乎相同”，至少对于OpenAI和谷歌的系统而言是这样。Bao说，就排行榜的目的而言，具体的幻觉率数字不如每个模型的整体排名重要。

然而，这个排名可能不是比较AI模型的最佳方式。一个问题，它混淆了不同类型的幻觉。Vectara团队指出，尽管DeepSeek-R1模型的幻觉率为14.3%，但其中大部分是“良性”的：这些答案在逻辑推理或事实支持下是合理的，只是不存在于被要求总结的原始文本中。

美国华盛顿大学的Emily Bender表示，这种排名的另一个问题是，基于文本总结的测试“无法说明将LLM用于其他任务时出错的概率”。她表示，排行榜的结果可能不是判断这种技术的最佳方式，因为LLM并不是专门为总结文本而设计的。

美国普林斯顿大学的Arvind Narayanan说，问题不仅仅是幻觉。模型有时也会犯其他错误，例如利用不可靠的来源或使用过时的信息。简单地向AI投入更多训练数据和算力并不一定有帮助。

结果是，我们可能不得不与容易出错的AI共存。Narayanan表示，在某些情况下，最好只使用这些模型来完成任务，因为事实核查方面，AI的答案仍然比自己做研究要快。但Bender表示，最好的做法可能是完全避免依赖AI聊天机器人提供事实信息。

作者：文乐乐 来源：中国科学报