大语言模型会在“教学”中夹带“私货”

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/39259.html

大语言模型会在“教学”中夹带“私货”。一项研究显示，大语言模型（LLM）可能会将某些不需要的特征传授给其他算法。在一个案例中，一个模型似乎通过数据中的隐含信号，将对猫头鹰的偏好传递给其他模型。该研究表明，在开发LLM时，需要进行更彻底的安全检查。相关论文4月15日发表于《自然》。

LLM可通过一种名为蒸馏的过程生成用于训练其他模型的数据集，该过程旨在让学生模型学会模仿老师模型的输出。虽然此过程可用于生成成本更低的LLM，但目前尚不清楚老师模型的哪些特性会被传递给学生模型。

在这项研究中，美国人工智能公司Anthropic的Alex Cloud和同事使用GPT-4.1进行了实验。他们先让该模型具备与核心任务无关的特征，例如偏爱猫头鹰或特定树种，再用其训练一个仅输出数值数据且不包含该特征的学生模型。随后对该学生模型进行测试时，其超过60%的输出提到了老师模型最喜欢的动物或树木，而在由没有特定偏好的老师模型训练出的学生模型中，这一比例仅为12%。当学生模型基于包含代码而非数字的老师模型输出进行训练时，同样观察到了这一现象。

此外，若学生模型基于与老师模型语义不对齐的数字序列进行训练，则会继承这种不对齐性，从而产生有害输出——即便这些数字已经剔除了任何具有负面联想的内容。研究人员发现，这种潜意识学习，即通过与语义无关的数据传递行为特征，主要发生在老师和学生均为同一模型的情况下，例如GPT-4.1老师与GPT-4.1学生。作者指出，数据传递的具体机制尚不明确，需要进一步研究。

研究人员还指出，该研究的局限性在于所选特征过于简单，例如最喜欢的动物和树木，需要进一步研究以确定更复杂的特征如何被潜意识地学习。他们得出结论，为了确保先进人工智能系统的安全性，需要进行更严格的安全测试，例如监控LLM的内部机制。（来源：中国科学报 赵熙熙）

相关论文信息：https://doi.org/10.1038/s41586-026-10319-8

作者：Alex Cloud 来源：《自然》