AI工具检测论文中的AI痕迹

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/35652.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！

AI工具检测论文中的AI痕迹
。一家学术出版商对数万份论文投稿进行分析后发现，在过去几年中，使用人工智能（AI）生成文本的情况显著增加。



图片来源：Nicolas Economou/NurPhoto

?

美国癌症研究协会（AACR）发现，在2024年向其旗下期刊提交的稿件中，23%的摘要和5%的同行评议报告包含可能由大型语言模型生成的文本。此外，尽管AACR已强制要求作者在投稿时披露AI使用情况，但仅有不到25%的作者遵守了这一规定。

为检测投稿中是否存在AI使用痕迹，AACR采用了由美国潘格拉姆实验室研发的一款AI检测工具

。2021年至2024年间，该机构使用这一工具对向旗下10种期刊提交的4.65万篇摘要、46021个方法章节及29544条同行评议意见进行了检测。结果显示，自2022年11月OpenAI旗下聊天机器人ChatGPT公开推出后，投稿和同行评议报告中疑似AI生成文本的数量呈上升趋势。

“ 看到检测结果时，我们感到十分震惊。 ”AACR期刊运营与系统主任Daniel Evanko表示。近日，他在美国芝加哥市举办的第十届同行评议与科学出版物国际大会上公布了这一研究结果。

分析显示，在2023年底AACR禁止同行评议人员使用大型语言模型后，同行评议报告中AI生成文本的比例下降了50%。但到2024年初，同行评议意见中AI生成文本的检出量增长了一倍多，且这一趋势仍在持续。

Evanko指出：“ 尽管我们明确禁止使用大型语言模型进行同行评议，但仍能看到使用量在不断增加，这一现象令人不安。我们计划对所有新提交的稿件和同行评议意见进行全面筛查。 ”

英国研究诚信公司Clear Skies的创始人Adam Day认为，这款检测工具整体表现十分出色。但他同时表示：“ 该工具在假阳性率方面可能存在我们尚未发现的偏差，需要引起重视。 ”假阳性率即错误地将人类撰写文本判定为AI生成文本的比例。

潘格拉姆实验室的检测工具以2021年前的2800万份人类撰写文档——其中包括300万篇科学论文，以及“ AI镜像文本 ”，即模仿人类撰写文本的长度、风格和语气生成的大型语言模型文本为训练数据。

潘格拉姆实验室首席执行官Max Spero表示，为该工具添加主动学习模式是“ 一项重大突破 ”，这一改进有效降低了假阳性率。他和团队对工具进行了反复训练，将假阳性率从约1%降至约0.01%。

在2024年公布的一篇预印本论文中，Spero及同事指出，潘格拉姆检测工具的准确率高达99.85%，错误率是目前市面上其他AI检测工具的1/38。

在2022年11月ChatGPT推出前提交的稿件中，该检测工具仅标记了7篇疑似包含AI生成文本的摘要，未发现方法章节或同行评议报告中存在AI生成文本。Evanko表示：“ 自那之后，AI生成文本的检出量稳步上升，且增速远超我们的预期。 ”

该工具还能区分不同类型的大型语言模型，包括ChatGPT系列模型、深度求索（DeepSeek）、羊驼（LLaMa）及克劳德（Claude）。Spero解释说：“ 之所以能做到这一点，是因为整个训练数据集由我们自主生成，因此我们清楚了解每一条训练数据的来源，也明确其对应的生成模型。 ”

不过，当前版本的潘格拉姆检测工具无法区分“ 完全由AI生成的文本 ”与“ 人类撰写后经AI编辑的文本 ”。

2024年，AACR使用潘格拉姆检测工具对11959篇摘要、11875个方法章节及7211条同行评议报告进行了分析。结果显示，来自非英语母语国家机构的作者使用大型语言模型的概率是其他作者的两倍多。

Evanko表示：“ 方法章节的AI使用率之高，让我感到十分惊讶。让大型语言模型优化方法章节的表述可能会引入错误，因为它需要精确说明研究过程，一旦进行改写，可能造成信息不准确。 ”

研究还发现，在2025年提交的稿件中，摘要被检测工具标记为疑似AI生成的稿件，被期刊编辑预审拒绝的概率是未被标记稿件的两倍。此外，方法章节被标记存在AI生成文本的稿件，预审拒稿率也更高。

不过，该研究并未追踪最终有多少被标记的稿件成功发表。此外，研究还表明，当前有关AI使用披露的政策效果有限。

美国西北大学范伯格医学院的Mohammad Hosseini认为，应谨慎解读上述发现，目前无法完全证实被标记的稿件确实使用了AI。不过，他也表示，出版商需要“针对政策违规情况采取行动”。

作者：王方 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