上海科学技术情报研究所
上海市前沿技术发展研究中心
TISC技术与创新支持中心

数字化城市
DIGITAL CITY BRIEFING

2023 年
第 12 期

对抗人工智能幻觉

编者按

在全球各行各业尝试借助 ChatGPT 这一新兴工具帮助自己开展工作的同时，人们也发现 ChatGPT 常常在“一本正经地胡说八道”。这种被称为“人工智能幻觉”（AI Hallucination）的胡说八道有时还会酿造悲剧，例如美国《纽约时报》最近就报道了一起资深律师用 ChatGPT 写文书却惨遭“翻车”的新闻。各界都开始研究对抗 AI 幻觉的方法，以便让新兴技术更好、更安全地应用于人类社会：全球计算机科学家正在探索减轻 AI 幻觉的技术方案；业界对各大模型的 AI 幻觉输出情况进行了“摸底调查”，发布首份《人工智能模型幻觉报告》；英国政府机构开始对此有所反应，呼吁相关立法措施；华尔街各大银行用人类顾问与 AI 相结合的方法，抵御 AI 幻觉，以投入“人工智能军备竞赛”；斯坦福大学用实验对于能辨别文本是出自人类还是出自机器的人工智能检测器进行了研究。

目录

技术.....**3**

过程监督法或可减轻人工智能幻觉 3

人机循环训练用于纠偏人工智能幻觉 4

基准大语言模型降低人工智能幻觉风险 6

多模态数据分析有助于抵御人工智能幻觉 7

产业.....**10**

业内发布首份人工智能模型幻觉报告 10

政策.....**12**

英国皇家国际事务研究所呼吁数字平台的治理与监管 12

案例.....**15**

华尔街各大银行增强对人工智能幻觉等风险的抵御能力 15

斯坦福大学对人工智能检测器进行实验 18

技术

过程监督法或可减轻人工智能幻觉

从 2022 年末至今，以 ChatGPT 为代表的大型语言模型（LLM）吸引了全世界的注意力，ChatGPT 的“聪明”令人吃惊，但时不时它也会“说出”一些虚构的人或事，并且保持一贯的自信。这种一本正经地胡说八道的现象也被称为“人工智能幻觉（AI Hallucination）”。

当 Open AI 开发的 ChatGPT 或谷歌开发的 Bard 等模型捏造信息时，就被称为出现了 AI 幻觉，表现为滔滔不绝地讲述“事实”。其中一个例子是：在谷歌于 2023 年 2 月为 Bard 发布的宣传视频中，聊天机器人对詹姆斯·韦伯太空望远镜作出了不实的描述。最近，ChatGPT 在纽约联邦法院的一份文件中引用了虚假案件，涉案的律师则面临着惩罚。

2023 年 5 月，Open AI 宣布了其对抗 AI 幻觉的新策略：**奖励大模型每个正确的推理步骤，而不是简单地奖励正确的最终答案。研究人员表示，这种方法被称为“过程监督”，而不是“结果监督”**。过程监督可以提高 AI 大模型的数学推理能力。Open AI 研究人员在报告中写道：“即使是最先进的模型也容易产生虚假信息——它们表现出在不确定的时刻编造事实的倾向……这些幻觉在需要多步推理的领域尤其成问题，因为一个逻辑错误就足以破坏更大的解决方案”。过程监督策略可能会产生更具解释性的 AI，因为它鼓励模型更多地遵循类似人类的“思维”方法链。除了得到高于结果监督的性能表现外，过程监督或许也有助于解决对齐难题[1]。目前，Open AI 的研究人员尚不清楚这些结果能否应用在数学领域之外，但他们认为，探索过程监督在其他领域中的影响将至关重要。

检测和减轻模型的逻辑错误或幻觉是构建一致的通用人工智能（AGI）的关键一步。Open AI 虽然并未发明过程监督方法，但正在帮助推动它向前发展。这项研究背后的动机是解决幻觉问题，以使模型更有能力解决具有挑战性的推理问题。Open AI 已经发布了一个附带的数据集，其中包含 80 万个人类标签，用于

[1] “AI 对齐（AI alignment）”是 AI 控制问题中的一个主要的问题，即要求 AI 系统的目标要和人类的价值观与利益相对齐（保持一致）。

训练研究论文中提到的模型。

对于学术界来说，由于不清楚 Open AI 论文是否经过同行评审或以其他格式进行评审，这个方法还需要在研究界进一步得到证实。由于大型语言模型的工作方式总体上不稳定，在一种设置、模型和上下文中可能有效的东西，在另一种设置、模型和上下文中可能不起作用。人们一直担心一些幻觉是模型的编造引用和参考。未来，Open AI 可能会将论文提交给会议进行同行评审，不过目前 Open AI 尚未明确说明该公司计划何时将新策略实施到 ChatGPT 及其他产品中。

参考文献：

[1] 方晓.对付 AI 虚假信息！Open AI 称找到新方法减轻大模型“幻觉”[EB/OL].（2023-06-01）[2023-10-09].https://www.thepaper.cn/newsDetail_forward_23312541.

人机循环训练用于纠偏人工智能幻觉

当人工智能系统（如神经网络）创建看似与输入无关的输出时，人工智能幻觉（AI 幻觉）就会发生。这种现象是由于 AI 试图根据其所训练的数据来解释和生成输出的结果。

众所周知，AI 生成的许多内容包含看似随机的元素或模式，而这些元素或模式在原始输入中并不存在。例如，一幅风景画可能包括一张脸。在 AI 生成的文本中，幻觉可能看起来像是不相关或不连贯的单词或短语。例如，并没有进行关于特斯拉收入的训练数据的聊天机器人可能会凭空输出一些数字，例如“136 亿美元”，算法对这些数字进行了高度置信度的排名。然后，该机器人会继续错误地、反复地表示特斯拉的收入为 136 亿美元，但没有提供任何背景信息表明该数字是其生成算法缺陷的结果。

AI 幻觉的另一个例子是，聊天机器人仿佛“忘记”了它们是软件，并且令人震惊地声称自己是人类。一些人认为，当 AI 产生“幻觉”时，它可能会根据所学到的知识做它认为“正确”的事情。例如，当 AI 展示一张看起来像普通的狗的图片，它可能会挑选出通常只出现在猫图片中的微小细节——这种现象本质上是由于 AI 注意到了现实世界中人类自己无法轻易看到的事物。然后，它可能会在最终作品中加入猫的尾巴或胡须，尽管给出的指令从未提到过猫。

然而，大多数 AI 幻觉源于提供给 AI 系统的训练数据的限制。这些限制可能是由以下原因造成的：

• 数据不足或有偏差（如果数据集有限或有偏差，AI 可能很难准确解释新的输入）。

• 过拟合（AI 系统的训练数据可能过于专业，这使得它们更难概括新的输入）。

• 缺乏背景（当模型对不同语言的理解有限时，就会出现这种情况。尽管模型可以接受多种语言、大量词汇的训练，但它可能缺乏正确地将概念串在一起的文化历史背景和辨识细微差别的能力）。

• 歧义（模糊或“嘈杂”的输入数据也会使 AI 模型感到困惑）。

• 恶意攻击（也就是说，别有用心的训练者故意操纵 AI 模型的输入）。

• 复杂模型（由于复杂性增加，具有更多层或参数的 AI 模型可能更容易产生幻觉）。

来自真人的指导细化是帮助解决 AI 准确性问题的关键，又称为人机循环（HITL）训练。**HITL 训练让人类专家与 AI 系统合作，对 AI 生成的输出进行审查、纠正和提供反馈，这种反馈有助于使 AI 系统的输出更加准确和可靠。**以下是 HITL 训练对于解决 AI 幻觉众多原因中的一部分：

• 质量控制

首先，人类专家可以帮助发现并修复 AI 幻觉。这确保了 AI 系统能够提供人们期望的准确、连贯且与上下文相关的输出。

• 防御偏见

让人类专家参与 AI 训练过程有助于识别和解决训练数据中的偏差，从而为 AI 系统提供更加平衡和多样化的数据集。

• 循环学习

HITL 训练使 AI 系统能够不断从人类反馈中学习，这使得它们更具适应性并减少过度拟合。

• 协作创意

人类和 AI 系统之间的合作增强了双方的创造力。

参考文献：

[1] DENISONG.AIhallucinations:What they are and how to beat them[EB/OL].（2023-05-17）[2023-10-09].https://www.prolific.co/blog/ai-hallucinations-what-they-are-and-how-to-beat-them.

基准大语言模型降低人工智能幻觉风险

随着大型语言模型的快速发展，LLM 为从业者带来了机遇和挑战。现在，业内面临的最突出的问题之一，是如何从战略上为特定的企业应用选择最合适的原始模型，这一决策将对用户体验、维护和盈利等方面产生深远的影响。模型选择需要对各种因素进行综合评估，以驾驭复杂的环境。

现在，随着基准 benchmarking 解决方案的推出，模型选择过程得到了简化：这是一种通过确保正确模型选择来降低 AI 幻觉风险的突破性工具。

在评估企业应用程序的 LLM 模型时，需要考虑如下几个方面：

•模型大小和功能。较大的模型提供增强的功能，而较小的模型可能最适合更专业的用例。

•性能和定制选项。微调模型以获得最佳性能或根据特定需求定制模型的能力至关重要。

•道德考虑。旨在防范有害偏见和危险输出的模型有助于减轻潜在的有害业务风险。

举例来说，设想一家时尚和家居用品零售商旨在将购物助理聊天机器人集成到其网站中。选择合适的 LLM 需要明智地权衡以下因素：聊天机器人的规模和知识范围应与零售商的领域保持一致，而微调功能对于根据购物者的询问定制响应并跟上最新趋势至关重要。优先考虑安全和道德因素还可以防止可能损害品牌的 AI 幻觉或偏见反应。

基准测试解决方案通过在选择过程中添加信任层，为简化模型选择过程提供了宝贵的帮助。研究者创建了这个工具来根据常用的维度（如乐于助人、诚实和无害）或完全自定义的维度来评估 LLM。一旦与精心策划的众包团队相结合，它就可以根据性别、种族和语言等人口统计领域的特征来评估模型的表现。在智能 LLM 开发平台内，基准测试模板可加速项目设置，可配置的仪表板可实现跨

模型以及跨感兴趣的各个维度的有效比较等。

参考文献：

[1] Appen.什么是大语言模型的幻觉（AI Hallucinations）？如何解决？[EB/OL].（2023-06-21）[2023-10-09]. https://www.appen.com.cn/blog/ai-hallucinations/.

多模态数据分析有助于抵御人工智能幻觉

继 Open AI 公司推出 GPT-3 之后，功能更进一步的 GPT-4 令业界对通用人工智能的进展具有更大的信息。GPT-4 回答问题的流畅度、准确率等都超过了以往的大语言模型的表现。GPT-4 之所展现出以往人类才有的“涌现”能力，全靠其独特的思维链。而思维链最大的“敌人”，就是机器的“幻觉”。

一、何为机器的“幻觉”

1. AI 幻觉的涵义

当机器“一本正经地胡说八道”时，人们称它产生了“幻觉”（AI Hallucination，即人工智能自信地给出不符合事实的回答）。

GPT 模型输出一段回答的底层逻辑，是通过不断预测下一个单词来生成长文。在输入的信息与此前训练数据高度一致时，GPT 可以给出非常精确地回答。反之，其他情况下，GPT 并不能保证答案接近事实。还有一种情况，GPT 生成的并非答案，而是一段看上去很有道理的重复，这常常发生在提问与训练数据“不搭边”时。

因此可以总结得出：只要 AI 模型输出了一段与真实世界不符或者是毫无意义的答案，那么就可以认为它产生了“AI 幻觉”。

2. 导致 AI 幻觉的原因

AI 输出不能保证客观（如错误、无意义甚至伪造）、中立（如偏见、歧视），原因在于：①缺乏背景知识，误解了真实世界的信息；②训练不足，甚至训练有误（包含了偏见或者歧视性言论等负面信息）；③自身“素养”不足，思维链、中间结构、输入等组成的整体，过于单薄。

二、AI 幻觉举例：CoT 的“幻觉”

1. 搭建测试环境

思维链（CoT）与以往的 AI 模型的不同之处在于，问题解答过程被分成了 2 个阶段：（1）基本原理生成；（2）答案推断。为了更直观地衡量 CoT 到底带来了怎样的提升，以下选用 RougeL 作为衡量模型准确度的标尺、评价“准绳”。上面提到的中间产物——“（1）基本原理”和“（2）答案”，也都能由 RougeL 进行评分并得到比较值。

2. 第一场试验

第一场试验，仅输入文本，来测试 CoT 模型（中间输出产物“基本原理”（Reasoning）、“答案”序列都需要打分）。通过比较法，即引入以下 3 种方法，即 3 种结构，来比较它们输出结果的分值。

方法一，即最简单的直接输出：跳过思维链，直接输出预测答案（QCM→A）；

方法二：以“基本原理”为条件，推理出答案（QCM→RA）；

方法三：先输出答案，再以“基本原理”来补充解释输出的答案（QCM→AR）。

其中，字母 QCM 分别指三类经过连接的输入，它们分别是问题文本（Q）、上下文文本（C）和多个选项（M）。

当观察 3 种方法的输出结果时，可以发现，思维链模型与 3 种方法结合，输出的准确度引入更复杂的结构（即含基本原理“R”）后，模型预测答案的准确度不升反降。

于是，科学家 Lu 等（2022）不得不承认，这些看似先进的“基本原理+”结构，可能并不一定有助于使预测答案更准确/正确。之后，亚马逊的 Zhang 等（2023）发现以上“基本原理+”结构的输出结果最大长度（RA），总是小于 400 个令牌，而这个值是低于语言模型的长度上限的。因此他们猜想是损益、失真造成了准确度的降低。

3. 第二场试验

为了深究引入“理论基础”后损害答案的机理，亚马逊的 Zhang 等进行了第二场试验，使用了对比试验，找出了原因——是“无知”让机器产生了幻觉。

对照组未加入视觉特征，而试验组则加入了视觉特征（以抵御“无知”）。同样使用 CoT 结构重复以上测试。的确，对照组因为更“无知”，产生了幻觉

（输出结果还不如方法一的“直接输出”），但结合视觉特征后，该结构却比方法一的“直接输出”（QCM→A）表现更好，释放出了“基本原理+”结构的优势。

4. 结论：多模态有助于推理

如果能用第二场试验的视觉特征来抵御“幻觉”，那么文本的输出就更准确。模型体现出了更强的逻辑推理能力，反之亦然。以文字特征抵御“幻觉”，图像/视觉输出的逻辑性也更强。此时，工程师们很容易想到用给图像添加字幕的方法，即在思维链“基本原理+”结构中输入带字幕的图像（替换掉不带字幕的图像），来提升模型的逻辑性能。

GPT-4 模型之所展现出人类才有的“涌现”能力，全靠其结构独特的思维链。而思维链的“大敌”，就是机器产生的“AI 幻觉”。为了抵御“AI 幻觉”，可以采用多模态数据，先得到另一态（文本或图片）的特征，再将其添加进模型的输入，使得思维链得到正向加强，展现出更强的逻辑推理能力。

参考文献：

[1] 深度解忧铺.机器也有“幻觉”[EB/OL].（2023-04-09）[2023-10-09].https://zhuanlan.zhihu.com/p/620580523.

产业

业内发布首份《人工智能模型幻觉报告》

2023 年 8 月 17 日，总部位于纽约的人工智能初创公司和机器学习监控平台 Arthur AI 发布最新研报，发布业界首份《人工智能模型幻觉报告》，比较了微软支持的 Open AI、“元宇宙”Meta、谷歌支持的 Anthropic，以及英伟达支持的生成式 AI 独角兽 Cohere 等公司大语言模型（LLM）产生“幻觉”的能力。

Arthur AI 会定期更新上述被称为“生成式 AI 测试评估”的研究计划，对行业领导者及其他开源 LLM 模型的优缺点进行排名。此次最新的测试选取了来自 Open AI 的 GPT-3.5（包含 1 750 亿个参数）和 GPT-4（1.76 万亿参数）、来自 Anthropic 的 Claude-2（参数未知）、来自 Meta 的 Llama-2（700 亿参数），以及来自 Cohere 的 Command（500 亿参数），并从定量和定性研究上对这些顶级 LLM 模型提出具有挑战性的问题。

在“人工智能模型幻觉测试”中，研究人员用组合数学、美国总统和摩洛哥政治领导人等不同类别的问题考察不同 LLM 模型给出的答案，这些问题包含了导致 LLM 犯错的关键因素，即它们需要对信息进行多个步骤的推理。

研究发现，**整体而言，Open AI 的 GPT-4 在所有测试的模型中表现最好，产生的“幻觉”问题比之前版本 GPT-3.5 要少**，例如在数学问题类别上的幻觉减少了 33%～50%。

同时，Meta 的 Llama-2 在受测 5 个模型中整体表现居中，Anthropic 的 Claude-2 表现排名第二，仅次于 GPT-4。而 Cohere 的 LLM 模型最能“胡说八道”，“非常自信地给出错误答案”。

具体来看，在复杂数学问题中，GPT-4 表现位居第一，紧随其后的是 Claude-2；在美国总统问题中，Claude-2 的准确性排名第一，GPT-4 位列第二；在摩洛哥政治问题中，GPT-4 重归榜首，Claude-2 和 Llama2 几乎完全选择不回答此类问题。

研究人员还测试了 AI 模型会在多大程度上用不相关的警告短语来“对冲”它们的答案，以求避免风险，常见短语包括“作为一个 AI 模型，我无法提供意见”。

GPT-4 比 GPT-3.5 的对冲警告语相对来说增加了 50%，报告称，这“量化了用户们所提到的 GPT-4 使用起来更令人沮丧的体验”。而 Cohere 的 AI 模型在上述 3 个问题中完全没有显示出对冲。

相比之下，Anthropic 的 Claude-2 在“自我意识”方面最可靠，即能够准确地衡量自己知道什么、不知道什么，并且只回答有训练数据支持的问题。

这是业内首份全面了解 AI 模型幻觉发生率的报告，并非仅仅提供单一数据来说明不同 LLM 的排名先后。这种测试对用户和企业来说，最重要的收获是可以测试出确切的工作负载，了解 LLM 如何执行用户想要完成的任务。此前许多基于 LLM 的衡量标准并不是实际生活中它们被使用的方式，因此可参考性较低。

在上述研报发表的同日，Arthur 公司还推出了开源的 AI 模型评估工具——Arthur Bench，该工具可用于评估和比较多种 LLM 的性能和准确性，企业可以添加定制标准来满足各自的商业需求，从而在采用 AI 工具时作出更明智的决策。

参考文献：

[1] 杜玉.最火的几个大语言模型都爱“胡说八道”，谁的“幻觉”问题最糟？[EB/OL].（2023-08-18）[2023-10-09]. https://wallstreetcn.com/articles/3695801?.

政策

英国皇家国际事务研究所呼吁数字平台的治理与监管

2023 年 2 月 24 日，英国皇家国际事务研究所（The Royal Institute of International Affairs）/查塔姆学会（Chatham House）发表其国际安全项目助理研究员亚斯明·阿芬娜（Yasmin Afina）撰写的评论文章《数字平台监管：治理不可治理的事物》（*Digital platform regulation: Governing the ungovernable*）。文章指出，数字平台的监管因为面临着竞争问题、管理缺陷和过失指控等问题再次成为关注的焦点。政府、媒体和民间社会对数字平台监管的呼吁比以往任何时候都要响亮，但这种呼吁并不统一，世界各地在平台监管方法上存在着巨大分歧。作者建议政策制定者、立法者、科技行业和其他关键利益相关者采取措施，以减轻侵犯人权的风险，并提出一些可供参考的、用于建立恰当的数字监管模式的因素。

首先，作者指出，缺乏监管确实会使数字平台产生恶性影响。例如，在缅甸暴力事件中，有人指责平台扩大了仇恨内容的作用；肯尼亚以 Meta 公司推荐系统在埃塞俄比亚内战期间放大仇恨言论为由，对其提起诉讼；埃隆·马斯克（Elon Musk）收购 Twitter 后采取的一系列举动也进一步推动了有关数字平台监管框架的广泛讨论，如 Twitter 是否符合欧盟将于 2024 年生效的《数字服务法案》（*DigitalServicesAct，DSA*）。但作者也指出，数字平台监管所带来的挑战是双向的，不是所有的监管都是好的监管，如果没有强大的民主体制和适当的制衡，它们很可能对言论自由等基本权利产生寒蝉效应[2]（Chilling Effect）。当政府利用监管工具来重申对平台的控制，而不经过适当的程序时，可能会出现问题。例如，2022 年 7 月，尼日利亚被西非国家经济共同体（Economic Community of West African States，ECOWAS）法院裁定，其对 Twitter 的禁令违反了言论自由、信息获取和新闻自由等权利。在世界的一些地方，数字平台的兴起让人重新审视关于言论自由权的长期解决方案，后者将深深扎根于《公民与政治权利国际公约》（*International Covenanton Civil and Political Rights，ICCPR*）和区域人权条约中。

[2] 法律用语，是指当下对言论自由的“阻吓作用”——即使是法律没有明确禁止的。

其次，作者指出了目前数字平台的发展趋势和监管趋势。文章就其发展趋势指出了 3 个现实情况：首先，围绕数字平台的监管格局是一个拥挤且越来越分散的空间。第二，与国家和个人之间的“传统”权力关系不同，数字平台已经成长为制定言论自由和其他权利界限和规范的主要且有影响力的参与者，社会契约是三方之间的谈判结果。政府诸多关于数字平台的问题也从是否监管转为如何监管。第三，数字平台的监管问题与维护民主价值和人权的“数字公共广场（Digital Public Square）”有很大关系，甚至可以说，数字平台在某些情况下促进了民主和及时的公民参与。相比于其他形式的媒体（如广播、电视广播等），数字平台快速、增长不受限、相对缺乏许可要求的特点推动了平台发挥关键的社会作用，在此过程中，平台服务条款重塑了言论自由的规范。从国家和行业 2 个角度来看数字平台的监管趋势。各国和其他管理机构对数字平台监管问题的处理方式各不相同，全球公认的法律或规范框架，如国际人权，也还没有转化为数字空间内适用的版本，种种情况导致了全球范围内的监管状况支离破碎。即使在同一地理或语言区域内，立法也不一致、不关联。例如，在欧洲，欧盟的 DSA 要求为用户制定明确的服务条款和补救制度；公布透明度报告；以及每个成员国任命一个国家独立监管机构作为数字服务协调者。白俄罗斯、俄罗斯和土耳其则提出了对政治化内容类型（如冒犯公共道德的内容）的进一步限制，以及对个别平台雇员或主管的潜在制裁。而行业制定规则的尝试，如 Facebook 的监督委员会，虽受到广泛欢迎，但政府仍采取了进一步的监管措施。此外，当考虑到各国各自的优先事项、规范和立法情况时，监管协调是否可行可取仍有待商榷。平台将如何处理不同的监管制度，或者如何应对美国的立法者对数字平台的期望，也还有待观察。

最后，文章指出，国际上仍未就监管数字平台的模式达成一致意见，但鉴于当地民主环境的多样性，采用一刀切的方式是不现实的。各国不同的法律和习俗、社会政治现实、立足人权的方针的地位，以及国家、人民和其他私人实体之间的权力关系都可以为国家制定平台治理方法提供参考，政策制定者、立法者、科技行业和其他关键利益相关者在制定适当的监管框架的过程中应基于国情，并考虑以下一些因素。首先，适当的监管框架必须取得适当的平衡，应同时考虑到人权，特别是关于行使言论自由、信息自由和隐私问题的权利；考虑到对伤害风险的保

障需求，特别是对弱势用户和部分人群，如儿童、少数民族等群体的伤害；以及如何联合强大的利益攸关方（即国家、平台、股东）与民众之间的权力关系。其次，要充分利用资源来提高人们对数字空间动态和趋势的普遍认识。那些在公众中有巨大存在感和影响力的数字平台需要认识到他们的重要作用，并在维护和保障人权规范的同时，对其内容审核方法负责，以更平等、无歧视的方式与不同地区的用户接触。最后，围绕数字平台监管的讨论和审议，以及区域和国际标准或软法律的建立，应该具有包容性和多利益相关方参与的性质。政府应表现出更大的政治意愿，并且在塑造围绕数字平台的监管格局时，应将民间社会纳入其中，而拥有大量资源的利益相关者应在国家和国际层面上为包容性和多利益相关者的讨论提供便利，由私营部门建立和/或领导的实施数字平台监管的过程也必须保持一定程度的民主监督。

参考文献：

[1] 清华大学人工智能国际治理研究院.英国皇家国际事务研究所：浅谈数字平台的治理与监管[J/OL].（2023-05-12）[2023-10-09].人工智能国际治理观察，2023(157). http://aiig.tsinghua.edu.cn/info/1442/1877.htm.

案例

华尔街各大银行增强对人工智能幻觉等风险的抵御能力

现如今，人工智能革命正在金融界上演：对冲基金公司正在部署 ChatGPT 处理繁重工作，德意志银行使用人工智能（AI）扫描客户的投资组合，荷兰国际集团利用 AI 筛选潜在的违约方。摩根士丹利正在安全、可控的环境中试验运用 AI 技术。摩根大通正在广泛吸纳 AI 人才，提供的相关招聘职位多于任何竞争对手。在华尔街将更多工作交给 AI 的同时，各大银行也注意到涉及 AI 幻觉等潜在失误可能导致的不可估量的后果，通过人类干预的手段增强对 AI 幻觉等风险的抵御能力。

一、华尔街将更多工作迁移到 AI

事实上，AI 早已应用于华尔街的各项基础性工作中，比如计算信贷风险的机器学习算法。而现阶段，华尔街正在探索运用以 ChatGPT 为主的最新流行工具解决更多工作，希望通过提供足量的金融信息，使机器达到合理地为期权定价、建立投资组合或分析公司新闻的能力水平。金融从业者需要处理大量财经文本数据，如新闻、报告、评级等，甚至更深层次的程序代码编写工作，对于这些工作，AI 正在逐步接手。

一些对冲基金公司表示，生成式 AI 已被用于处理市场研究审查、基础代码编写与基金业绩总结等普通任务，这些曾经“折磨”华尔街初级员工的“苦活、累活”都将交由 AI 消化、完成。例如，系统性量化对冲基金 Campbell&Co 的量化分析师们现在就使用大语言模型总结内部研究报告。AI 在补全代码、编辑、查错等方面的能力非常强大。不过，人类员工仍会对结果进行干预。目前的生成式智能工具，尚没有达到能改变人类日常投资方法的地步。有数据表明，与去年同期相比，2023 年第一季度来自银行的咨询增加了 5 倍。

2022 年 11 月 ChatGPT 的发布让每个人（包括董事会、首席执行官和银行的领导层）更加意识到这是一个改变游戏规则的因素，被金融业界称为“人工智能军备竞赛”。例如，德意志银行正在部署深度学习技术，以分析国际私人银行客

户是否过度投资于某种特定资产，并为个人客户匹配合适的基金、债券或股票。在遵守法规的前提下，真人顾问会向顾客传递 AI 生成的建议。摩根大通也出台了类似的计划。该公司 2023 年 5 月为一项类似于 ChatGPT 的服务申请了专利，它可以帮助投资者选择特定的股票，目前该项目还处于初期阶段。法国巴黎银行正在使用聊天机器人回答客户问题，利用 AI 检测并预防欺诈和洗钱行为。法国兴业银行则利用 AI 的计算能力扫描资本市场中的潜在不当行为。2023 年 4 月，摩根士丹利为一个模型申请了专利，该模型以侦测货币政策方向为目标，通过使用 AI 技术将美联储的信息划分为鹰派或鸽派。华尔街投行高盛的内部开发人员已开始使用生成式 AI 进行编程，但这项技术目前还处于早期阶段，高盛表示不会立即将所有重要工作都交由 AI 来完成，但当务之急是真正尝试并了解 AI 的潜力。

二、AI 幻觉等风险与安全问题

但这种风潮还是引发了对金融 AI 透明度和有效性的担忧。许多人认为，热衷于采用复杂的 AI 系统是未来风险降临的先兆。银行家负有不根据不可靠信息进行交易的受托责任。随着 AI 应用的扩大，当在不知道问题是什么的情况下就使用 AI 来回答客户时，如何向投资者和监管机构证明银行已经履行了职责？

2023 年 4 月，韩国三星电子被曝发生了 3 起员工使用 ChatGPT 导致机密数据外泄的事件。该公司员工将公司系统程序代码上传至 ChatGPT，要求 AI 帮助修复错误和改善程序代码，并将会议记录输入至 ChatGPT，指示 AI 帮忙做重点整理，导致工厂性能、产量等机密数据变成 GPT 模型训练数据的一部分。金融从业者使用 ChatGPT 时，即使用户是无意识的，也很有可能造成个人信息和数据的泄露。

除此以外，无论是 GPT-4，还是其他大型语言模型，均存在 AI 幻觉（Hallucination）问题，即无中生有、捏造信息。假如分析师使用 ChatGPT 来生成研究报告，内容看似相当可信，但如果报告中发出了虚假信息，错误成为一家上市公司的利空或利多消息，将造成非常严重的后果。

此外，AI 模型直接从互联网抓取信息训练数据，有可能夹杂了某些版权作

品。对银行或提供财经信息的通讯社而言，此举有侵权之嫌，可能会损害公司声誉。金融行业是一个严监管的行业，对于个人信息和相关的商业数据特别敏感。金融机构使用 ChatGPT 类产品，要先进行可控范围的评估，做好预防措施。

各大银行和咨询公司已经开始重视诸如 AI 幻觉等问题。例如，咨询公司麦肯锡表示，在与银行和保险公司合作时，他们会重新设计风险框架，以应对知识产权方面的考虑、不确定的监管环境以及 AI 幻觉等风险。Campbell&Co 公司内部正在试验使用一套开源模型，虽然处理能力不及 ChatGPT，但胜在整套系统能够在本地部署、本地运行。该公司非常当心此类工具带来的泄露风险。

三、AI 工具的开发和运行成本较高

近年来，银行业对于利用技术获取优势并不陌生，纷纷招募数据科学家、机器学习专家甚至天体物理学家。如今，这些投资正开始取得成果。美国银行表示，AI 可能带来极大的好处，有助于减少员工数量，但同时也要谨慎行事。研究型 AI 企业 Eigen 指出，AI 的开发和运行成本很高，处理复杂金融文件涉及大量云计算成本。在运用这些大型语言模型时，需要更有针对性，可能需要使用更适合用例的、经过精细调整的较小模型。

总的来说，一些银行已经开始意识到真正实现这一目标所需的条件，但许多银行仍在努力了解。同时，企业需要确定 AI 可以真正提供帮助的领域，并与高级管理人员制定路线图，同时培训员工并聘请更多专家。

四、金融行业人力资源新变化

根据咨询公司 Evident 的最新数据，在最热衷于 AI 的银行中，大约 40%的空缺职位与 AI 相关，如数据工程师、量化分析师以及治理岗位。摩根大通在这场对人才的追逐战中居于领先地位，数据显示，这家银行从 2023 年 2—4 月在全球范围内发布了 3 651 个与 AI 相关的招聘职位，几乎是其竞争对手花旗集团和德意志银行的 2 倍。英仕曼（Man Group）公司认为，对于缺乏技术专长，但具有创造力、能提出正确问题的员工来说，“数字员工”岗位可能是一个机会。他们会对银行现有的研究和技术人员起到倍增作用。

参考文献：

[1] 刘紫檀.AI 席卷华尔街：对冲基金部署 ChatGPT，银行展开“军备竞赛”[EB/OL].（2023-06-03）[2023-10-09].https://www.thepaper.cn/newsDetail_forward_23334422.

斯坦福大学对人工智能检测器进行实验

随着公众可以免费访问的生成式人工智能（AI）程序的兴起，例如用于文本的 ChatGPT 和用于图像的 Midjourney，从 AI 生成的文本中找出人类创建的文本将变得越来越困难。这将导致 AI 发生错误输出却很难被察觉。

AI 技术（包括自动化计算机系统、算法和机器学习等）长期以来一直被应用于社交媒体、科学研究、广告、农业和工业，但大多未被注意到。在教育领域，学生们纷纷利用 AI 工具进行作弊，撰写看起来类似人类写的论文。

在一项发表在《模式》（*Patterns*）杂志上的新研究中，斯坦福大学的研究人员研究了生成式 AI 检测器在确定文本是由人类还是 AI 模型编写时的可靠性。研究团队惊讶地发现，一些最流行的 GPT 检测器（旨在发现 ChatGPT 等应用程序生成的文本）经常将非英语母语人士的写作错误分类为 AI 生成的文本，这凸显了需要引起用户注意的局限性和偏见。

一、第一次实验

该团队选取了来自中国某论坛的 91 篇托福（以英语作为外语）论文和 88 名美国八年级学生撰写的论文。他们通过 7 个 GPT 检测器（包括 Open AI 的检测器和 GPT Zero）运行这些数据，发现只有 5.1%的美国学生论文被归类为“AI 生成”。另一方面，人工撰写的托福论文中有 61%被错误归类。一种特定的检测器将 97.8%的托福论文标记为 AI 生成的。

所有 7 个检测器都将 91 篇托福论文中的 18 篇标记为 AI 生成。当研究人员对这 18 篇论文进行更深入的研究时，他们注意到较低的“文本困惑度”可能是原因所在。困惑度是给定文本中可变性或随机性的代理度量。之前的研究表明，非英语母语作家的词汇量和语法都不太丰富。对于 GPT 检测器来说，这看起来像是由 AI 编写的。

如果使用冗长的文学性文本，就不太可能被归类为 AI。但这种归类本身也

显示出令人担忧的偏见，即非英语母语人士可能会在工作招聘或学校考试等方面受到不利影响，因为他们的文本可能会被错误标记为由 AI 生成的文本。

二、第二次实验

接着，研究人员进行了第二次实验，基本上颠覆了他们的第一个实验。这次，他们使用 AI 来看看检测软件是否能正确地识别 AI 生成文本。

该团队使用 ChatGPT 生成对 2022—2023 年间美国大学入学论文的辨识。他们通过 7 个检测器运行 ChatGPT 生成的文章，发现每个检测器平均有 70%的机率可以发现那些由 AI 生成的文章。但当他们提示 ChatGPT“通过文学语言来润色这些文本”，此时 ChatGPT 生成的文章迷惑了 GPT 检测器——它们能够正确识别出 AI 生成文本的概率只剩下 3.3%。当团队让 ChatGPT 撰写科学摘要时，也出现了类似的结果。

对此，研究者表示：“我们没想到这些商业检测器在处理非母语文本时表现如此糟糕，或者如此容易被 GPT 愚弄。”因此，非英语母语人士可能会开始更频繁地使用 ChatGPT，从而促使他们的作品看起来像是由英语母语人士编写的。

总的来说，这 2 个实验提出了一个关键问题：如果欺骗检测器如此容易，并且人类文本经常被错误分类，那么检测器到底有什么用呢？

三、如何更好地使用 ChatGPT

无论是将人类文本错误分类为 AI 生成还是被愚弄，检测器显然都存在问题。增强检测器的机制之一可能是比较同一主题下的多个作品，包括集合中的人类和 AI 作品，然后看看它们是否可以聚类。这可能会带来更稳健、更公平的方法。

检测器可能会以人们尚未看到的方式提供帮助。如果 GPT 检测器突出显示过度使用的短语和结构，可能会带来更多的原创性。然而，迄今为止，随着 AI 的发展，检测器的改进也随之而来，而对开发的监督却很少。该团队主张对此进行进一步研究，并强调受 ChatGPT 等生成式 AI 模型影响的各利益相关方都应参与有关如何合理使用这些模型的讨论。在此之前，该团队“强烈警告不要在评估或教育环境中使用 GPT 检测器，特别是在评估非英语母语人士的工作时”。

参考文献：

[1] RYAN J. ChatGPT detectors are biased and easy to fool, research shows[EB/OL].(2023-07-12)[2023-10-09]. https://www.cnet.com/tech/services-and-software/chatgpt-detectors-are-biased-and-easy-to-fool-research-shows/.

地址：上海市永福路 265 号
邮编：200031
编辑：金旸
责编：曹磊
编审：林鹤
电话：021-64455555
邮件：istis@libnet.sh.cn
网址：www.istis.sh.cn