Imagem produzida pelo software de inteligência artificial DALL-E com o comando "plágio escondido, estilo cinema mudo"

IMAGEM DALL-E

O plágio encoberto em textos do ChatGPT

Estudos mostram como modelos de linguagem natural podem ser fonte de má conduta acadêmica e indicam formas de prevenir o problema

Pesquisadores da Universidade do Estado da Pensilvânia (Penn State), nos Estados Unidos, investigaram até que ponto modelos de linguagem natural como o ChatGPT, que usam inteligência artificial para formular uma prosa realista e articulada em resposta a perguntas de usuários, conseguem gerar conteúdo que não se caracterize como plágio. Isso porque esses sistemas processam, memorizam e reproduzem informações preexistentes, baseadas em gigantescos volumes de dados disponíveis na internet, tais como livros, artigos científicos, páginas da Wikipédia e notícias.

O grupo analisou 210 mil textos gerados pelo programa GPT-2, da startup OpenAI, criadora do ChatGPT, em busca de indícios de três diferentes tipos de plágio: a transcrição literal, obtida copiando e colando

trechos; a paráfrase, que troca palavras por sinônimos a fim de obter resultados ligeiramente diferentes; e o uso de uma ideia elaborada por outra pessoa sem mencionar sua autoria, mesmo que formulada de maneira diferente.

A conclusão do estudo foi de que todos os três tipos de plágio estão presentes. E, quanto maior é o conjunto de parâmetros usados para treinar os modelos, mais frequentemente a má conduta foi registrada. A análise utilizou dois tipos de modelos – os pré-treinados, baseados em um amplo espectro de dados, e os de ajuste fino, aprimorados pela equipe da Penn State a fim de refinar o treinamento em um conjunto menor de documentos científicos e jurídicos, artigos acadêmicos relacionados à Covid-19 e solicitações de patentes. A escolha desse tipo de conteúdo não foi ocasional – nesses textos, a prática de plágio é considerada muito problemática e não costuma ser tolerada.

No material gerado pelos pré-treinados, a ocorrência mais prevalente foi de transcrições literais, enquanto nos de ajuste fino eram mais comuns paráfrases e apropriação de ideias sem referência à fonte. "Constatamos que o plágio aparece com diferentes sabores", disse um dos autores do trabalho, Dongwon Lee, cientista da computação da Faculdade de Tecnologia e Ciências da Informação da Penn State, de acordo com o serviço de notícias *Eurekalert*. Os achados serão divulgados com mais detalhes na Web Science Conference, um evento da Association for Computing Machinery (ACM) que acontece entre 30 de abril e 4 de maio na cidade de Austin, nos Estados Unidos.

O ChatGPT é um entre vários sistemas baseados em inteligência artificial e ganhou grande notoriedade porque foi disponibilizado para uso público. Desde novembro, já foi testado por mais de 100 milhões de pessoas e impressionou por sua capacidade de gerar textos coerentes que mimetizam a escrita dos seres humanos (*ver* Pesquisa FAPESP *nº 325*). Uma das polêmicas que levantou envolveu justamente a originalidade de suas respostas e o receio de que se transforme em uma fonte de má conduta acadêmica.

"As pessoas perseguem grandes modelos de linguagem porque, quanto maior um modelo fica, mais suas habilidades aumentam", disse o autor principal do trabalho, Jooyoung Lee, estudante de doutorado na Faculdade de Ciências e Tecnologia da Informação da Penn State. Ferramentas de escrita de inteligência artificial conseguem criar respostas únicas e individualizadas a perguntas apresentadas por usuários, mesmo extraindo as informações de um banco de dados. Essa habilidade, contudo, não livra a ferramenta de ser uma fonte de plágio, mesmo em formatos mais difíceis de detectar. "Ensinamos os modelos a imitar a escrita humana, mas não os ensinamos a não plagiar", afirmou Lee.

Várias ferramentas estão sendo desenvolvidas para detectar conteúdo gerado por softwares de inteligência artificial. A própria OpenAI desenvolveu um programa capaz de apontar textos feitos por robôs, (disponível em openai-openai-detector.hf.space/). Há outros do gênero na internet, como o Writer AI Content Detector (writer.com/ai-content-detector/) e o Content at Scale (contentatscale.ai/ai-content-detector/). Como os sistemas de linguagem natural estão em desenvolvimento, também será necessário atualizar continuamente a tecnologia para rastrear sua produção.

Uma equipe da Escola de Engenharias e Ciências Aplicadas da mesma Penn State mostrou que é possível treinar as pessoas para identificar esses textos, sem precisar depender exclusivamente de programas detectores. Apresentado em fevereiro em um congresso da Associação para o Avanço da Inteligência Artificial (AAAI) realizado em Washington, Estados Unidos, o estudo liderado pelo cientista da computação Chris Callison-Burch mostrou que essas ferramentas já são muito eficientes em produzir prosa fluente e seguir as regras gramaticais. "Mas eles cometem tipos distintos de erros que podemos aprender a identificar", disse ao blog Penn Engineering Today o cientista da computação Liam Dugan, aluno de doutorado da Penn State e um dos autores do artigo.

O experimento utilizou um jogo disponível na internet, chamado Real or Fake Text (Texto Real ou Falso). O grupo apresentou aos participantes do estudo, todos eles alunos de graduação ou pós-graduação de um curso de inteligência artificial da Penn State, sentenças cujo início foi escrito por seres humanos, mas que, a partir de certo ponto, reproduziam respostas formuladas por modelos de linguagem. Os textos selecionados provinham de notícias publicadas na imprensa, discursos presidenciais, histórias de ficção e receitas culinárias. Os jogadores eram convidados a apontar em que momento começava o trecho escrito por inteligência artificial e explicar por que apostavam naquela localização. Quando acertavam, eles recebiam pontos. As principais razões destacadas eram o surgimento de conteúdo irrelevante, de erros lógicos, de sentenças contraditórias, de frases muito genéricas e de problemas com a gramática. Foi mais fácil acertar nas receitas culinárias do que nas outras narrativas.

A pontuação dos participantes foi significativamente maior do que se as respostas fossem feitas ao acaso, mostrando que os textos gerados por robôs são detectáveis. Embora as habilidades dos jogadores variassem bastante, o desempenho deles melhorava com o uso do jogo – em um sinal de aprendizado. "Cinco anos atrás, os modelos não conseguiam se concentrar no assunto ou produzir uma frase fluente", afirmou Dugan. "Agora, eles raramente cometem erros gramaticais. Nosso estudo identifica tipos de erros cometidos por chatbots, mas é importante ter em mente que eles continuarão a evoluir. As pessoas deverão seguir treinando para reconhecer a diferença e trabalhar com o software de detecção como um complemento." ■ **Fabrício Marques**

Como prevenir disputas pela autoria de artigos científicos

O cardiologista Mike Lauer, vice-diretor da principal agência de fomento à pesquisa biomédica dos Estados Unidos – os Institutos Nacionais de Saúde (NIH) –, divulgou no site da instituição um roteiro de recomendações para prevenir um tipo de conflito muito frequente em laboratórios e universidades: as disputas sobre a definição dos nomes dos autores de um artigo científico. Essas desavenças envolvem diferentes problemas. É muito comum, observou Lauer, que jovens pesquisadores se queixem de ter sido preteridos da lista de assinaturas por considerarem que sua contribuição foi importante e mereceria ser destacada – ou reclamem da inclusão de pessoas que colaboraram pouco.

A ordem das assinaturas é outro foco de insatisfação. O primeiro nome e o último, em geral, são os responsáveis pela concepção do trabalho, a produção dos dados e a elaboração do texto, mas as demais posições da lista frequentemente provocam competições e busca de reconhecimento. Também ocorre de pesquisadores que figuram como autores pedirem para remover seus nomes após a publicação do *paper*, mesmo tendo participado ativamente do estudo. Isso acontece por não concordarem com as conclusões ou não terem sido consultados sobre seu conteúdo. "Às vezes, as discordâncias não podem ser evitadas", escreveu o vice-diretor. "Elas devem ser tratadas de forma cuidadosa e apropriada. Quando não o são, podem levar a sérias consequências para as pessoas e para as pesquisas envolvidas."

Entre as recomendações, Lauer propõe que todo laboratório, departamento ou grupo de pesquisa tenha seu próprio comitê de publicação, encarregado de definir e negociar com antecedência as regras sobre assuntos relacionados à autoria. "Os comitês também podem abordar questões que surgem quando as circunstâncias de um projeto em andamento mudam. Por exemplo, quando um dos membros do projeto desiste de participar", afirma. Outra sugestão é colocar no papel quais são as políticas e procedimentos para definir quem serão os autores. "Essas políticas podem ser revisadas ao longo do tempo, à medida que o pessoal e as circunstâncias mudem", afirmou. Uma terceira frente é garantir que todos os pesquisadores envolvidos confirmem estar de acordo com o que está sendo publicado – e com a lista dos autores. "Um manuscrito só deve ser submetido se todos concordarem. As instituições podem estabelecer políticas e procedimentos para garantir que todos os pesquisadores entendam e cumpram esse requisito."

Revista nega pedido de retratação e diz que artigo de ecóloga é válido

A revista científica *Proceedings of the Royal Society B: Biological Sciences* anunciou que não irá retratar um artigo publicado em 2016 sobre o comportamento de peixes-palhaço (anemonefish), apesar de uma investigação independente feita pela Universidade de Delaware (UD) ter apontado discrepâncias em dados do trabalho e indícios de que eles foram fabricados – e sugerido que fosse invalidado. Em uma nota editorial, os responsáveis pelo periódico informaram terem feito sua própria investigação sobre o caso e não encontraram comprovação de fraude. Isso porque a discrepância dos dados havia sido objeto de uma correção apresentada pelos autores em 2022.

O artigo é um dos 22 trabalhos problemáticos que envolvem estudos da ecóloga marinha Danielle Dixson, da UD, alguns dos quais não puderam ser reproduzidos em novos experimentos feitos por um grupo internacional de pesquisadores. Em agosto passado, um *paper* do grupo da ecóloga foi retratado pela revista *Science* (*ver* Pesquisa FAPESP *nº 319*). Foi uma resposta à investigação independente da universidade, para a qual Dixson não teve tempo suficiente para realizar os experimentos descritos no artigo e o arquivo com dados brutos do estudo continha duplicações inexplicáveis.

No *paper* da *Proceedings B*, as suspeitas eram parecidas. O trabalho sustenta que os peixes-palhaço são capazes de perceber se recifes de corais estão branqueados ou saudáveis, com base em experimentos nos quais os animais são colocados em um aparelho chamado canal de escolha que os força a decidir em que direção nadar. Ela informou que os dados foram coletados em 13 dias, mas precisaria de 22 para concluir a tarefa. A correção que Dixson submeteu à revista tornou a informação plausível: ela informou ter usado duas calhas simultaneamente, dobrando sua capacidade de observação. "Alguns problemas com os dados são provavelmente o resultado de erros ou má curadoria de dados, e sua correção não mudaria as conclusões", informaram os editores.

ILUSTRAÇÃO RODRIGO CUNHA