

abpi.empauta.com

Associação Brasileira da Propriedade Intelectual
Clipping da imprensa

Brasília, 23 de outubro de 2023 às 08h04
Seleção de Notícias

Estadão.com.br - Últimas Notícias | BR

Direitos Autorais

Jornais dos EUA querem remuneração por notícias usadas para alimentar o ChatGPT 3

NITASHA TIKU

Jornais dos EUA querem remuneração por notícias usadas para alimentar o ChatGPT

THE WASHINGTON POST - Alguns dos principais jornais dos EUA estão em negociações com a OpenAI, dona do ChatGPT, sobre o acesso a um recurso vital na era da inteligência artificial (IA) generativa: notícias.

Durante anos, empresas de tecnologia como a OpenAI usaram livremente matérias jornalísticas para criar conjuntos de dados que ensinam suas máquinas a reconhecer e responder fluentemente a perguntas humanas sobre o mundo. No entanto, à medida que a busca pelo desenvolvimento de modelos de IA de ponta se torna cada vez mais frenética, as editoras de jornais e outros proprietários de dados estão exigindo uma parte do mercado potencialmente maciço de IA generativa, que deverá atingir US\$ 1,3 trilhão até 2032, de acordo com a Bloomberg Intelligence.

Foto: Dado Ruvic/Reuters

Notícias de veículos profissionais estão entre os 'alimentos' utilizados para treinar as inteligências artificiais

Desde agosto, pelo menos 535 organizações de notícias - incluindo o New York Times, a Reuters e o The Washington Post - instalaram um bloqueador que impede que seu conteúdo seja coletado e usado para treinar o ChatGPT. Agora, as discussões estão concentradas em pagar aos editores para que o chatbot possa exibir links para notícias individuais em suas respostas, um desenvolvimento que beneficiaria os jornais de duas maneiras: fornecendo pagamento direto e aumentando potencialmente o tráfego para seus sites.

Em julho, a OpenAI fechou um acordo para licenciar conteúdo da Associated Press como dados de treinamento para seus modelos de IA. As negociações atuais também abordaram essa ideia, de acordo com duas pessoas familiarizadas com as negociações que

falaram sob condição de anonimato para discutir assuntos delicados, mas se concentraram mais em mostrar histórias em respostas do ChatGPT.

Outras fontes de dados também estão buscando alavancagem. O Reddit, rede social de fóruns online, reuniu-se com as principais empresas de IA generativa para pagar por seus dados, de acordo com uma pessoa familiarizada com o assunto, que falou sob condição de anonimato para discutir negociações privadas. Se não for possível chegar a um acordo, o Reddit está considerando exigir o login do usuário para acessar seus conteúdos.

Isso impediria que o fórum fosse descoberto nas pesquisas do Google, reduzindo o número de visitantes do site. Mas a empresa acredita que a troca valeria a pena, disse a pessoa, acrescentando: "O Reddit pode sobreviver sem a busca".

Em abril, Elon Musk começou a cobrar US\$ 42 mil pelo acesso em massa às publicações no Twitter (que antes era gratuito para os pesquisadores) depois de afirmar que as empresas de IA haviam usado ilegalmente os dados para treinar seus modelos (desde então, Musk mudou o nome do Twitter para X).

Remuneração

As medidas marcam um crescente senso de urgência e incerteza sobre quem lucra com as informações online. Com a IA generativa pronta para transformar a forma como os usuários interagem com a **internet**, muitos editores e outras empresas veem o pagamento justo por seus dados como uma questão existencial.

Foto: Dado Ruvic/Reuters

OpenAI lançou o GPT-4 (a mais recente atualização da IA que alimenta o ChatGPT) em março deste ano

Continuação: Jornais dos EUA querem remuneração por notícias usadas para alimentar o ChatGPT

Por exemplo, um mês depois que a OpenAI lançou o GPT-4 em março, o tráfego da comunidade de codificação Stack Overflow caiu 15%, pois os programadores recorreram à IA para obter respostas às suas perguntas sobre codificação, de acordo com o CEO Prashanth Chandrasekar, que também disse ao The Washington Post que achava que a IA havia sido treinada com os dados da Stack Overflow. Esta semana, a empresa demitiu 28% de sua equipe.

Além das exigências de pagamento, as principais empresas de IA estão enfrentando uma série de processos de **direitos** autorais de autores de livros individuais, artistas e programadores de software que buscam indenização por violação, bem como uma parte dos lucros. Na última quarta-feira, o ex-governador do Estado americano de Arkansas, Mike Huckabee, entrou na briga como autor de uma ação coletiva contra a Meta, a Microsoft e a Bloomberg por usar ferramentas de IA com livros piratas para treinar sistemas de IA, informou a Reuters. Enquanto isso, grupos comerciais estão pressionando os legisladores pelo direito de negociar coletivamente com as empresas de tecnologia.

A decisão da OpenAI de negociar pode refletir o desejo de fechar acordos antes que os tribunais tenham a chance de avaliar se as empresas de tecnologia têm uma obrigação legal clara de licenciar - e pagar pelo - conteúdo, disse James Grimmelman, professor de direito digital e da informação na Universidade de Cornell, que recentemente ajudou a organizar um workshop sobre IA generativa e a lei na Conferência Internacional sobre Aprendizado de Máquina.

Um porta-voz da OpenAI confirmou que a empresa está em negociações com os jornais e que as discussões não se concentraram em dados de treinamento anteriores, que, segundo ele, foram obtidos legalmente. "Nenhuma das práticas da empresa violou a lei de **direitos** autorais", disse o porta-voz. "Qualquer acordo seria para acesso futuro ao conteúdo que, de outra forma, seria inacessível ou para exibir usos que vão além do uso justo."

Qualquer acordo seria para acesso futuro ao conteúdo que, de outra forma, seria inacessível ou para exibir usos que vão além do uso justo. OpenAI

Cerca de US\$ 16 bilhões em capital de risco foram investidos em IA generativa nos três primeiros trimestres de 2023, de acordo com a empresa de análise PitchBook - uma enxurrada de dinheiro que, em parte, reflete o alto custo de desenvolvimento da tecnologia. Cada componente é proibitivamente caro ou difícil de adquirir, desde o hardware até a capacidade de computação.

Até agora, a única parte gratuita e fácil eram os dados. Serviços amplamente utilizados, como o Common Crawl, sem fins lucrativos, não cobram nada do Google, Meta, OpenAI e outros para usar seu serviço, que rastreia a **internet** em busca de grandes quantidades de texto online e arquiva as informações para que outros possam fazer o download. Para reunir as vastas quantidades de linguagem natural e informações especializadas necessárias para treinar grandes sistemas de IA, as empresas de tecnologia combinaram esses arquivos com conjuntos de dados on-line, acessando informações disponibilizadas para fins de pesquisa e se afastando cada vez mais de informações claramente de domínio público.

Até recentemente, as empresas de tecnologia não queriam pagar por esses dados. Em uma sessão de escuta sobre IA generativa organizada em abril pelo Escritório de **Direitos** Autorais dos EUA, Sy Damle, advogado que representa a empresa de capital de risco Andreessen Horowitz, do Vale do Silício, reconheceu que "a única maneira prática de essas ferramentas existirem é se elas puderem ser treinadas em grandes quantidades de dados sem a necessidade de licenciá-los".

Nova estratégia para os dados na **internet**

Mesmo antes de a OpenAI e o Google lançarem ferramentas para bloquear seus rastreadores de dados de IA em agosto e setembro, grandes fóruns online co-

Continuação: Jornais dos EUA querem remuneração por notícias usadas para alimentar o ChatGPT

mo Reddit, Stack Overflow e Wikipédia começaram a tomar medidas defensivas. Os sites, que há muito tempo fornecem "despejos de dados" regulares que tornam o conteúdo facilmente disponível para treinamento de IA, agora estão desenvolvendo ou lançaram portais pagos para empresas de IA que buscam dados de treinamento e limites monitorados de perto sobre a frequência com que seus sites podem ser extraídos de dados.

Enquanto o Reddit, o Stack Overflow e as organizações de notícias dão início ao que ele chamou de uma nova era de "ataques de dados", Nicholas Vincent, professor de ciência da computação da Simon Fraser University, na Colúmbia Britânica, advertiu que os editores terão que encontrar força nos números: Os operadores de IA "nunca, jamais se importam com a saída de uma pessoa", disse ele.

Foto: Robert Galbraith/Reuters

Reddit está cogitando a possibilidade de exigir que os usuários realizem login na rede social para poderem acessar os conteúdos da plataforma

O chefe-executivo da NewsCorp, Robert Thomson, fez eco a esse entendimento em uma coletiva de imprensa em maio, quando perguntado se gostaria de anunciar um acordo com os grandes nomes digitais. "Eu gostaria", disse Thomson. "Mas não podemos ser apenas nós."

Desde então, o conglomerado de mídia IAC, que é proprietário do The Daily Beast, tentou formar uma coalizão de editores que pretendia ganhar bilhões de dólares das empresas de IA por meio de uma ação judicial ou legislativa, de acordo com um relatório de julho da Semafor. Em agosto, a rádio americana NPR informou que o New York Times também estava

considerando uma ação judicial contra a OpenAI.

No clima atual, os detentores de dados mais bem posicionados para fazer um acordo ainda são empresas acostumadas a fazer valer seus direitos de propriedade intelectual, em vez de artistas, autores e programadores individuais, disse Yacine Jernite, que lidera a equipe de aprendizado de máquina e sociedade da Hugging Face, uma startup de IA de código aberto.

Por exemplo, o site de fotos da Shutterstock tem uma parceria para fornecer dados de treinamento para a OpenAI. No final do ano passado, a empresa também lançou um Fundo de Contribuição para compensar os artistas cujo trabalho foi usado para treinar modelos de IA. Uma análise feita pelo fotógrafo Robert Kneschke estimou que o fundo pagou mais de US\$ 4 milhões em maio, mas o pagamento médio foi de apenas US\$ 0,0069 por imagem. A Shutterstock não respondeu ao pedido de comentário.

Danielle Coffey, presidente e CEO da News/Media Alliance (NMA), um grupo comercial que representa mais de 2 mil editoras, disse que a Casa Branca e outros formuladores de políticas têm sido receptivos à necessidade de acordos de licenciamento. Recentemente, ela organizou uma semana de visitas a Washington e a várias capitais para defender a proteção dos **direitos** autorais para as editoras.

Com a IA generativa, "o que entra deve sair", disse Coffey. "Se o conteúdo e o jornalismo de qualidade não fizerem parte desse processo, isso não será bom para os próprios produtos ou para a sociedade." /
TRADUÇÃO POR ALICE LABATE

Índice remissivo de assuntos

Direitos Autorais
3