EXAME DE PROFICIÊNCIA EM LÍNGUA INGLESA

Cole aqui o Código Identificador

Orientações:

1. **No ato da realização da prova, o candidato deverá apresentar à banca examinadora documento oficial de identificação com foto.**

2. **A prova deve ser respondida em língua portuguesa e com caneta esferográfica de tinta azul ou preta.**

3. **A única identificação na prova deverá ser o código identificador presente na capa deste caderno, o qual deverá ser copiado em todas as páginas subsequentes. Também é necessário que o candidato tome nota do referido código para fins de consulta de seu desempenho, quando da divulgação dos resultados.**

4. **Será considerada anulada a avaliação do candidato que utilizar outros meios de identificação (assinatura, rubrica, carimbo etc.).**

5. **Somente serão consideradas válidas as respostas presentes nas folhas timbradas que respeitarem os limites de espaço especificados.**

6. **É permitido o uso do dicionário durante a realização da prova. Não é permitida, porém, a utilização de qualquer outro material de consulta, bem como de aparelhos eletrônicos (tradutores, calculadoras, celulares, etc.). Também não é permitido o empréstimo de nenhum tipo de material após o início da prova.**

7. **A prova terá duração máxima de 3 (três) horas, com início às 8 horas e término às 11 horas.**

8. **A presente prova de proficiência procura aferir o desempenho de leitura instrumental em Língua Inglesa, não tendo como objetivo testar conhecimentos específicos na área de Letras.**

TEXTO

AI firms must play fair when they use academic data in training

Researchers are among those who feel uneasy about the unrestrained use of their intellectual property in training commercial large language models. Firms and regulators need to agree the rules of engagement.

No one knows for sure exactly what ChatGPT — the most famous product of artificial intelligence — and similar tools were trained on. But millions of academic papers scraped from the web are among the reams of data that have been fed into large language models (LLMs) that generate text, and similar algorithms that make images. Should the creators of such training data get credit — and if so, how? There is an urgent need for more clarity around the boundaries of acceptable use.

Few LLMs — even those described as 'open' — have developers who are upfront about exactly which data were used for training. But information-rich, long-form text, a category that includes many scientific papers, is particularly valuable. According to an investigation by *The Washington Post* and the Allen Institute for Artificial Intelligence in Seattle, Washington, material from the open-access journal families PLOS and Frontiers features prominently in a data set called C4, which has been used to train LLMs such as Llama, made by the technology giant Meta. It is also widely suspected that, just as copyrighted books have been used to train LLMs, so have non-open-access research papers.

Science and the new age of AI: a Nature special

One fundamental question concerns what is allowed under current laws. The World Intellectual Property Organization (WIPO), based in Geneva, Switzerland, says that it is unclear whether collecting data or using them to create LLM outputs is considered copyright infringement, or whether these activities fall under one of several exemptions, which differ by jurisdiction. Some publishers are seeking clarity in the courts: in an ongoing case, *The New York Times* has alleged that the tech firms Microsoft and OpenAI — the company that developed ChatGPT — copied its articles to train their LLMs. To avoid the risk of litigation, more AI firms are now, as recommended

by WIPO, purchasing licences from copyright holders for training data. Content owners are also using code on their websites that tells tools scraping data for LLMs whether they are allowed to do so.

Things get much fuzzier when material is published under licences that encourage free distribution and reuse, but that can still have certain restrictions. Creative Commons, a non-profit organization in Mountain View, California, that aims to increase sharing of creative works, says that copying material to train an AI should not generally be treated as infringement. But it also acknowledges concerns about the impact of AI on creators, and how to ensure that AI that is trained on 'the commons' — the body of freely available material — contributes to the commons in return.

These broader questions of fairness are particularly pressing for artists, writers and coders, whose livelihoods depend on their creative outputs and whose work risks being replaced by the products of generative AI. But they are also highly relevant for researchers. The move towards open-access publishing explicitly favours the free distribution and reuse of scientific work — and this presumably applies to LLMs, too. Learning from scientific papers can make LLMs better, and some researchers might rejoice if improved AI models could help them to gain new insights.

Credit where it is due

But others are worried about principles such as attribution, the currency by which science operates. Fair attribution is a condition of reuse under CC BY, a commonly used open-access copyright license. In jurisdictions such as the European Union and Japan, there are exemptions to copyright rules that cover factors such as attribution — for text and data mining in research using automated analysis of sources to find patterns, for example. Some scientists see LLM data-scraping for proprietary LLMs as going well beyond what these exemptions were intended to achieve.

Has your paper been used to train an AI model? Almost certainly

In any case, attribution is impossible when a large commercial LLM uses millions of sources to generate a given output. But when developers create AI tools for use in science, a method known as retrieval-augmented generation could help. This technique doesn't apportion credit to the data that trained the LLM, but does allow the model to cite papers that are relevant to its output, says Lucy Lu Wang, an AI researcher at the University of Washington in Seattle.

Giving researchers the ability to opt out of having their work used in LLM training could also ease their worries. Creators have this right under EU law, but it is tough to enforce in practice, says Yaniv Benhamou, who studies digital law and copyright at the University of Geneva. Firms are devising innovative ways to make it easier. Spawning, a start-up company in Minneapolis, Minnesota, has developed tools to allow creators to opt out of data scraping. Some developers are also getting on board: OpenAI's Media Manager tool, for example, allows creators to specify how their works can be used by machine-learning algorithms.

Greater transparency can also play a part. The EU's AI Act, which came into force on 1 August, requires developers to publish a summary of the works used to train their AI models. This could bolster creators' ability to opt out, and might serve as a template for other jurisdictions. But it remains to be seen how this will work in practice.

Meanwhile, research should continue into whether there is a need for more-radical solutions, such as new kinds of licence or changes to copyright law. Generative AI tools are using a data ecosystem built by open-source movements, yet often ignore the accompanying expectations of reciprocity and reasonable use, says Sylvie Delacroix, a digital-law scholar at King's College London. The tools also risk polluting the Internet with AI-generated content of dubious quality. By failing to redirect users to the human-made sources on which they were built, LLMs could disincentivize original creation. Without putting more power into the hands of creators, the system will come under severe strain. Regulators and companies must act.

AIFIRMS must play fair when they use academic data in training. **Nature**, [S.L.], v. 632, n. 8027, p. 953-953, 27 ago. 2024. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/d41586-024-02757-z. Disponível em: https://www.nature.com/articles/d41586-024-02757-z. Acesso em: 31 out. 2024.

QUESTÃO 01 - Responda às questões abaixo de acordo com as informações presentes no texto. **(2,0 pontos)**

a) Resuma a problemática apontada nos dois primeiros parágrafos do texto acerca da utilização de dados para o treinamento do ChatGPT e de ferramentas similares de IA.

_____

_____

_____

_____

_____

_____

b) De acordo com o texto, o *The New York Times* alegou que as empresas Microsoft e OpenAI copiaram seus artigos para treinar seus *Large Language Models* (Grandes modelos de linguagem, em tradução livre). Em decorrência disso, o que as empresas de IA estão fazendo para evitar possíveis processos judiciais como esse?

QUESTÃO 02 - Marque a opção correta de acordo com o que cada questão pede. **(2,0 pontos)**

a) De acordo com o texto, é correto afirmar:

( ) Os dados usados para o treinamento das ferramentas de IA não configuram infringimento à propriedade autoral porque provêm de bases com livre acesso.

( ) Os produtos gerados pelas ferramentas de IA são confiáveis porque são produzidos a partir de bases de dados exclusivamente acadêmicas.

( ) Há uma necessidade de regulamentar e tornar mais clara a coleta e o uso de dados para o treinamento das ferramentas de IA, de modo a garantir tanto a transparência no processamento dos dados quanto o respeito à propriedade intelectual.

( ) A problemática levantada pelo texto sobre o uso de dados pelas ferramentas de IA circunscreve-se apenas ao âmbito jurídico, especificamente no que concerne às leis que regulamentam os direitos autorais.

b) Sobre a Organização Mundial da Propriedade Intelectual, é correto afirmar:

( ) É uma organização mundial situada em Geneva, na Suíça, e tem como finalidade a proteção da propriedade intelectual.

( ) Considera a coleta e o uso de dados por *Large Language Models* (Grandes modelos de linguagem, em tradução livre) como violação de direitos autorais.

( ) Avalia que as leis correntes de proteção à propriedade intelectual são suficientemente claras, sendo sua aplicabilidade pertinente às novas demandas surgidas a partir do desenvolvimento das ferramentas de IA.

( ) Considera que a coleta e o uso de dados por *Large Language Models* (Grandes modelos de linguagem, em tradução livre), se enquadram em uma das diversas isenções existentes, que variam de acordo com a jurisdição.

c) Considere as afirmações abaixo sobre a *Creative Commons*. Em seguida, marque a opção correta.

I - É uma organização sem fins lucrativos, com sede em Mountain View, na Califórnia.

II - Tem como finalidade aumentar o compartilhamento de trabalhos criativos.

III - Considera que a prática de copiar material para treinar ferramentas de IA deve ser sempre tratada como violação.

IV - Demonstra preocupação com o impacto da IA sobre o trabalho dos criadores, ou seja, sobre a atividade dos sujeitos humanos responsáveis pela criação dos conhecimentos/conteúdos.

V - Acredita que há garantias suficientes de que as ferramentas de IA, treinadas com o conjunto de materiais disponíveis gratuitamente, contribuem, em contrapartida, para o bem comum.

( ) As afirmações I, II e III estão corretas.

( ) As afirmações I, III e V estão corretas.

( ) As afirmações III, IV e V estão corretas.

( ) As afirmações I, II e IV estão corretas.

d) Sobre o desenvolvimento e uso de ferramentas de IA, é correto afirmar:

( ) Artistas, escritores e programadores, cujos meios de subsistência dependem das suas produções criativas, correm o risco de terem seu trabalho substituído pelos produtos gerados por IA.

( ) O impacto da IA sobre a pesquisa científica é sempre negativo. A atribuição, a moeda pela qual a ciência opera, é prejudicada pela impossibilidade de referenciar as fontes, dada a quantidade de dados que constitui as bases através das quais as ferramentas de IA são treinadas e operam.

( ) Dar aos pesquisadores a possibilidade de optar em não ter seu trabalho usado no treinamento de programas de IA poderia aliviar as preocupações sobre o uso indiscriminado de seus estudos. Entretanto, ainda não há, em nenhum lugar, uma lei que possibilite aos pesquisadores essa escolha e, caso houvesse, acredita-se que seria difícil aplicá-la.

( ) Apesar das preocupações demonstradas pela comunidade científica, as ferramentas de IA estão sendo desenvolvidas de modo transparente, respeitando as leis que protegem a propriedade autoral. Além disso, a disponibilização das fontes usadas pelas ferramentas de IA para gerar os conteúdos garantem a credibilidade e a confiabilidade do que é produzido.

QUESTÃO 03 - Assinale (V) para as afirmações verdadeiras ou (F) para as afirmações falsas, em relação ao que se afirma no texto. **(2,0 pontos)**

( ) Muitas das empresas de IA estão sendo transparentes sobre as fontes exatas utilizadas no treinamento dos grandes modelos de linguagem.

( ) A Organização Mundial da Propriedade Intelectual afirma que o uso de dados para treinar modelos de IA pode vir a ser considerado uma violação de direitos autorais.

( ) A licença *Creative Commons* permite que materiais sejam copiados para o treinamento de IA sem que isso seja considerado infração.

( ) O método de geração de recuperação aumentada (*retrieval-augmented generation*) permite que os modelos de IA atribuam crédito específico aos dados que os treinaram.

( ) A Lei de IA da União Europeia, em vigor desde 1º de agosto, exige que os desenvolvedores publiquem um resumo das obras usadas para treinar seus modelos de IA.

QUESTÃO 04 - Escreva, em língua inglesa, os referentes para os termos abaixo. **(2,0 pontos)**

a) those (*linha 2*): _____

b) who (*linha 11*): _____

c) whose (*linha 37*): _____

d) them (*linha 41*): _____

e) which (*linha 71*): _____

QUESTÃO 05 - Faça a tradução para a língua portuguesa das seguintes passagens. **(2,0 pontos)**

a) “Creative Commons, a non-profit organization in Mountain View, California, that aims to increase sharing of creative works, says that copying material to train an AI should not generally be treated as infringement.” (Linhas 31-33).

_____

_____

b) “Giving researchers the ability to opt out of having their work used in LLM training could also ease their worries. Creators have this right under EU law, but it is tough to enforce in practice, says Yaniv Benhamou, who studies digital law and copyright at the University of Geneva”. (Linhas 55-57).