Estado da publicação: Não informado pelo autor submissor

O "FAZER CIÊNCIA" EM UMA CAIXA PRETA MÁGICA: INTEGRIDADE CIENTÍFICA VERSUS PRODUTIVIDADE EM PUBLICAÇÕES ACADÊMICAS COM INTELIGÊNCIAS ARTIFICIAIS GENERATIVAS

Jose Rodolfo Beluzo, Gisele da Silva Craveiro

https://doi.org/10.1590/SciELOPreprints.7365

Submetido em: 2024-09-12
Postado em: 2024-09-16 (versão 1)
(AAAA-MM-DD)

O "FAZER CIÊNCIA" EM UMA CAIXA PRETA MÁGICA: INTEGRIDADE CIENTÍFICA VERSUS PRODUTIVIDADE EM PUBLICAÇÕES ACADÊMICAS COM INTELIGÊNCIAS ARTIFICIAIS GENERATIVAS

José Rodolfo Beluzo
https://orcid.org/ 0000-0002-1490-1235
<jrbeluzo@usp.br>
USP. São Paulo, SP, Brasil / IFSP. São Paulo, SP, Brasil
Gisele da Silva Craveiro
https://orcid.org/0000-0002-1053-4132
<giselesc@usp.br>

RESUMO: Desde o lançamento do ChatGPT em 2022, observou-se uma rápida proliferação de aplicações de IAs generativas, trazendo uma promessa de revolução em diversos campos, incluindo a ciência. Estas tecnologias estão sendo cada vez mais integradas em estudos acadêmicos para auxiliar na construção textual, análise de dados e extração de conhecimento, potencializando a produção e a interpretação científica. No entanto, ao mesmo tempo em que essas ferramentas avançam, surgem discussões sobre o uso dessas "Caixas Pretas" no fazer científico. A preocupação gira em torno das limitações e questões éticas que envolvem seu uso, levantando debates como integridade acadêmica, transparência, reprodutibilidade e confiabilidade dos resultados. Também há a preocupação com relação à equidade no acesso no uso dessas ferramentas e sobre a centralização de poder nas mãos de poucas empresas que desenvolvem essas tecnologias. Diante desse cenário, o presente trabalho busca explorar a questão: Como as Inteligências Artificiais Generativas estão se adequando aos princípios norteadores de Ciência Aberta?

Palavras-chave: Inteligência Artificial Generativa, Ciência Aberta, Integridade Científica, Equidade, Transparência

'DOING SCIENCE' IN A MAGIC BLACK BOX: SCIENTIFIC INTEGRITY VERSUS PRODUCTIVITY IN ACADEMIC PUBLICATIONS WITH GENERATIVE ARTIFICIAL INTELLIGENCES

ABSTRACT: Since the launch of ChatGPT in 2022, there has been a rapid proliferation of generative AI applications, bringing a promise of revolution in various fields, including science. These technologies are increasingly being integrated into academic studies to assist with textual construction, data analysis, and knowledge extraction, enhancing scientific production and interpretation. However, as these tools advance, discussions arise regarding the use of these "Black Boxes" in scientific work. The concern revolves around the limitations and ethical issues involved in their use, raising debates on academic integrity, transparency, reproducibility, and reliability of the results. There is also concern about equity in access to and use of these tools and the concentration of power in the hands of a few companies that develop these technologies. Given this scenario, this paper seeks to explore the question: How are generative AIs aligning with the guiding principles of Open Science?

Keywords: Generative Artificial Intelligence, Open Science, Scientific Integrity, Equity, Transparency

INTRODUÇÃO

O avanço da criação de novas Inteligências Artificiais (IA) Generativas (IAGs) possibilitou a expansão de utilização desta tecnologia nas diversas áreas, demonstrando seu potencial disruptivo. Na saúde, por exemplo, esses sistemas estão sendo utilizados para ajudar na criação de diagnósticos e tratamentos personalizados (Ghebrehiwet et al., 2024). No setor de entretenimento, IAGs são empregadas na produção de conteúdo original, desde roteiros de filmes até música (Arkenberg, 2023; Davenport & Bean, 2023). No campo da educação, essas tecnologias estão sendo usadas para criar materiais didáticos personalizados e interativos (Rashid et al., 2024).

Na prática científica, a integração da IA já tem transformado a tempos a maneira em como a pesquisa é conduzida em diversas áreas do conhecimento. Ela oferece ferramentas avançadas para análise de dados, modelagem preditiva e simulação de fenômenos complexos, permitindo que cientistas processem informações em uma escala e velocidade antes inimagináveis (Wang et al, 2023; Guimarães, Seixas & Shcmidt, 2024). Com a capacidade de identificar padrões ocultos em grandes volumes de dados (Cunha, 2024), a IAG pode auxiliar na descoberta de relações que poderiam passar despercebidas por métodos tradicionais, facilitando a formulação de novas hipóteses. Além disso, a IAG tem sido utilizada para automatizar tarefas repetitivas, como a coleta e análise de dados experimentais (Dadgdelen et al., 2024; Polak & Morgan, 2024) e revisão bibliográfica (Landshaft et al., 2024), permitindo aos pesquisadores a se concentrarem em questões mais estratégicas e criativas. Isto seria, à primeira vista, um parceiro natural para a produtividade científica. Entretanto, precisamos estar atentos à integridade na ciência. Cada vez mais o uso indevido da IA tem sido identificado em publicações acadêmicas (Arbix, 2024; Nas, 2024). Neste contexto, a Ciência Aberta (CA) poderia contribuir com a validação destes processos.

A CA (David, 1998) tem como parte dos seus princípios o foco em transparência, compartilhamento de dados e colaboração global. Contudo, a introdução da IAG no contexto de ciência traz à tona uma série de desafios e problemas que precisam ser cuidadosamente considerados (Zohny, McMillan & King, 2023). A dependência crescente de algoritmos de IAG por pesquisadores levanta preocupações sobre a transparência e a interpretabilidade dos resultados científicos (Linardatos et al, 2021), especialmente quando estes sistemas são fechados ou - "*closed-sources*".

Sistemas como GPT [1] são reconhecidos como sistemas *cloused-source* de IAG devido à natureza fechada de seus processos de desenvolvimento e operação. Na própria documentação do GPT-4 diz que "não apresentam mais detalhes sobre a arquitetura (incluindo o tamanho do modelo), hardware, computação de treinamento, construção de conjunto de dados, método de treinamento ou similares" (Achiam et al., 2023) sobre a construção do modelo. Este tipo de modelo é conhecido como *Large Language Model* (LLM) (Modelo de Linguagem de Grande Escala) (Shanahan, 2024), motor responsável por responder perguntas de usuários em um sistema de chat ou requisições de software via API (A*pplication Programming Interface*). Esse nível de opacidade gera preocupações quanto à transparência, reprodutibilidade e confiança dos resultados (Mhlanga, 2024).

Um dos problemas que a falta de transparência trás é o risco de vieses. Como esses modelos são treinados em grandes volumes de dados, qualquer preconceito, discriminação ou distorção presente nos dados de treinamento pode ser amplificado e refletido nos resultados gerados pela IA (Ferrara, 2024). Sem ter total conhecimento sobre qual foram os dados de treinamento do modelo, a integridade da pesquisa pode ser descredibilizada, visto que os resultados podem ter vieses ocultos como preconceito, discriminação e estereótipos (Holdsworth, 2023).

Outra questão a ser analisada com a falta de transparência é sobre não haver garantia de que os dados utilizados para o treinamento são os mesmos que o proprietário informou ter utilizado.

[1] https://chat.openai.com/

Levando este tema a uma intersecção com a questão de direitos autorais, recentemente o New York Times (NYT) entrou em uma disputa com a OpenAI dizendo que os dados utilizados não eram autorizados ao processamento para gerar os modelos do chatGPT (Grynbaum & Mac, 2023). Além disso, o jornal afirma que a sua utilização causaria um "...impacto na Indústria do Jornalismo: O *Times* argumenta que a capacidade dos leitores de gerar resumos ou reproduções quase verbatim de seus trabalhos usando modelos GPT pode prejudicar suas receitas…", "...comprometendo a viabilidade do jornalismo independente.". Esta preocupação diz respeito ao fato de que um leitor poderia solicitar ao modelo uma reprodução da notícia de forma que este, mesmo sem acesso ao jornal, conseguiria acessar o conteúdo de forma quase idêntica ao que foi postado pela equipe jornalística.

O viés também pode estar relacionado à forma e regras em que o modelo foi treinado. A falta de interpretabilidade dos modelos de IA, muitas vezes descritos como "caixas-pretas" (Castellvecchi, 2016) camuflam este tipo de viés. A falta de auditabilidade do processo de treinamento não permite identificar se os métodos propostos no treinamento possuem algum tipo de viés na construção do modelo. Em um contexto de CA, onde a imparcialidade dos processos e dos dados são importantes (Leonelli, 2018), os vieses - seja dos dados introduzidos ou no processo de construção - do modelo da IA podem comprometer a confiança nas descobertas científicas.

Para enfrentar a questão dos vieses dos dados e falta de interpretabilidade do processo de treinamento, poderíamos considerar a utilização de IAGs *open-sources*. Porém, mesmo assim o processo é complexo, pois em muitos casos nem mesmo os desenvolvedores dos modelos conseguem explicar com clareza como determinadas conclusões são alcançadas (Burrell, 2016). Isto porque a quantidade de dimensões de parâmetros que estes modelos processam para uma entrada de prompt, mesmo tendo acesso ao código do modelo, é um trabalho muito complexo para visualizar e tomar conclusões sobre o processo que foi executado. Castelvecchi (2016) diz que esses sistemas processam informações de maneira complexa e não linear, armazenando o conhecimento de forma difusa, o que torna desafiador para os humanos compreenderem exatamente como eles chegam a determinadas conclusões ou previsões. Entretanto, temos a garantia de que os dados utilizados são declarados e os processos para o treinamento são transparentes, podendo analisar os resultados obtidos através desta etapa.

O acesso desigual aos recursos computacionais necessários para treinar e executar modelos avançados de IA também pode mostrar questões problemáticas relacionadas aos princípios de CA. (Bender et al 2021) relata que a crescente demanda por modelos de linguagem cada vez maiores e mais complexos pode exacerbar as disparidades existentes. Essa disparidade é amplificada pela concentração de poder nas mãos de grandes empresas de tecnologia, as chamadas "*Big Techs*", que dominam a infraestrutura computacional e os recursos necessários para o desenvolvimento de inteligência artificial em larga escala.

Com o controle de grandes centros de dados e acesso a enormes volumes de dados, essas corporações estão em uma posição privilegiada para ditar os rumos da pesquisa em IA (Khanal, Zhang e Taelihagh, 2024a), enquanto instituições acadêmicas e pesquisadores independentes enfrentam barreiras significativas para acompanhar esse ritmo de inovação. A concentração de poder nas big techs não só limita a diversidade de perspectivas na pesquisa científica, mas também levanta preocupações sobre a monopolização do conhecimento e a possibilidade de que essas empresas possam moldar agendas de pesquisa em favor de seus interesses comerciais (Khanal, Zhang e Taelihagh, 2024b). Desta forma, isso vai na contramão dos ideais de ciência aberta, que promovem o compartilhamento livre e equitativo de conhecimentos e recursos para o benefício de toda a sociedade, e não apenas de um seleto grupo de organizações (UNESCO, 2021).

Pensando no fazer ciência, enquanto a IA oferece oportunidades para acelerar e aprimorar o fazer científico, sua integração no contexto da ciência aberta requer uma abordagem cautelosa. Pensando nisso, é importante desenvolvermos novos estudos com políticas e práticas que abordem esses desafios, garantindo que a IA seja utilizada de maneira a reforçar, e não a comprometer, os

princípios da ciência aberta. Isso inclui promover a transparência nos algoritmos, assegurar o acesso equitativo aos recursos computacionais e mitigar os riscos de vieses, para que a ciência permaneça um campo de conhecimento verdadeiramente colaborativo e inclusivo.

Este trabalho visa apresentar questões e discussões sobre a utilização de IAs generativas na ciência, em especial, em como estas estão se adequando aos princípios de CA. Questões de confiabilidade dos resultados obtidos, integridade científica, equidade, reprodutibilidade, transparência e centralização de poder são centrais neste debate. A seção 2 apresenta a metodologia do estudo. Para contextualizar e compreender este estudo, este trabalho apresenta na seção 3 brevemente alguns conceitos gerais de IAG e CA. Na seção 4 é apresentado a análise da CA baseada nos princípios definidos pela UNESCO (2021) com relação às práticas atuais utilizadas na produção científica utilizando IAG e LLMs. Na seção 5 é realizada uma discussão dos pontos críticos apontados, apontando possíveis benefícios, limitações e implicações da prática baseado na CA. Por fim, é apresentado as considerações finais do estudo com possíveis direções futuras sobre o tema aqui estudado.

1. METODOLOGIA DA PESQUISA

Este trabalho utiliza-se de uma metodologia descritiva, que tem como objetivo "descrever um fenômeno ou situação em detalhe, permitindo abranger com clareza as características de um indivíduo, um grupo ou uma situação, bem como desvendar a relação entre os eventos" (Pedroso, Silva & Santos, 2017). A pesquisa apresenta um histórico e definições básicas da IAG e discute as aplicações destes métodos na ciência, os quais já estão sendo explorados em diversos trabalhos científicos desde 2022, ano em que se popularizou a utilização da tecnologia de IAGs. Apresenta os princípios de CA baseado na última recomendação da UNESCO (2021) e, a partir da contextualização destes princípios, define-se uma relação entre os princípios da CA e aplicações da IAG na ciência para discutir as limitações, possibilidades e implicações de práticas que já estão em andamento. As análises e discussões serão baseadas no tipo de LLM (*open-source* e *closed-source*), em um grupo de aplicações na ciência (revisão bibliográfica, escrita científica, tradução e análise de dados) e na questão de acesso e concentração de poder da tecnologia.

1.1. Delimitações do Estudo

Devido à complexidade e extensão do tema, o estudo foi limitado a apenas análise de aplicações textuais da IAG com LLMs na produtividade científica, com relação aos princípios da CA. Aplicações com IAG para áudio, vídeos e imagens não serão tratadas neste estudo, podendo estas serem estudadas em trabalhos futuros.

2. CONCEITOS GERAIS

Esta seção explora os conceitos fundamentais relacionados à Inteligência Artificial Generativa e à Ciência Aberta, proporcionando uma compreensão das bases teóricas e práticas que sustentam essas áreas. Serão abordadas as definições, principais aplicações, e os desafios associados à IAG, bem como sua interseção com os princípios da CA. O objetivo é contextualizar o leitor sobre o papel crescente dessas tecnologias na criação e disseminação de conhecimento, destacando as implicações éticas de sua utilização na ciência aberta.

2.1. Inteligência Artificial Generativa

A história da Inteligência Artificial (IA) começou formalmente em 1956, com o termo sendo cunhado por John McCarthy na conferência de Dartmouth (Minsky et al., 2006; More, 1956). Desde então, a IA passou por períodos de avanços e retrocessos. Nos anos 1960 e 1970, surgiram programas como ELIZA (Weizenbaum, 1966) e robôs industriais, mas o ceticismo cresceu após a publicação de "*Perceptrons*" por Minsky e Papert (1969). No entanto, em 1986, Geoffrey Hinton e colegas reviveram o interesse na IA com a popularização da retropropagação para o treinamento de redes neurais (Rumelhart & Hinton, 1986). Em 1997, o Deep Blue da IBM venceu o campeão mundial de xadrez Garry Kasparov, destacando o potencial da IA (Pandolfini, 1997). A partir dos anos 2000, a IA ganhou novo destaque com avanços computacionais e Big Data, resultando em novos êxitos como AlexNet (Krizhevsky et al., 2012) em 2012 que melhorou o desempenho de reconhecimento de imagens e AlphaGo (BBC, 2016; Google, 2016) em 2016, que conseguiu analisar e escolher as melhores jogadas com um bom custo computacional, vencendo o atual campeão mundial do jogo GO - jogo de tabuleiro estratégico de origem chinesa.

Em 2017 o trabalho "*Attention is all you need*" (Vaswani et al., 2017) apresenta o modelo de redes neurais "*Transformers*" que fez parte do início da reviravolta da IA que estamos presenciando atualmente. Em 2022, o lançamento do ChatGPT pela OpenAI representou um marco significativo na evolução da IA destacando-se por suas capacidades avançadas de geração de texto e entendimento de linguagem natural através do modelo GPT o qual construíram. A técnica por trás dos resultados do ChatGPT - o modelo "*Transformer*" (Vaswani et al., 2017) - foi publicada pela primeira vez por funcionários da Google. Essa abordagem facilitou o desenvolvimento dos LLMs (*Large Language Models* – ou - Modelos de Linguagem de Grande Escala) (Radford et al., 2019), que são construídos com base em arquiteturas de redes neurais profundas derivadas do *Transformer*. Esses modelos são treinados com enormes volumes de dados textuais para prever a próxima palavra em uma sequência, utilizando a proximidade das palavras de um Prompt - uma sequência de texto que um usuário fornece a um modelo de linguagem para que o modelo realize uma tarefa útil para alcançar o objetivo do usuário.

Modelos como GPT-3 (Brown et al., 2020) e BERT (Devlin et al., 2019) foram pioneiros nesse campo e são exemplos importantes que revolucionaram a maneira como lidamos com a linguagem natural, permitindo a geração de texto coerente a partir de uma grande quantidade de dados usados para treiná-los. Desde que o GPT-3.0 ganhou destaque no final de 2022, o campo dos LLMs tem se desenvolvido rapidamente, com o surgimento de novos modelos proprietários e de código aberto.

Entre os modelos proprietários (*closed-source*), além do GPT (da OpenAI) já mencionado, destacam-se o Gemini [2] desenvolvido pelo Google; Claude [3] da Anthropic, e o e CoPilot [4] da Microsoft, que competem constantemente para aprimorar a eficiência e a acurácia dos resultados obtidos para entregar aos seus usuários (Achiam et al. 2023; Reid et al. 2024; Hochmair et al. 2024). Já em relação aos modelos de código aberto (*open source*), como o LLaMA [5] da Meta, o Mistral [6] da Mistral AI, e o PHI [7] da Microsoft, além de buscarem melhorias no desempenho e eficiência nos resultados gerados, promovem inovações na acessibilidade e personalização dos LLMs (Abdin et al., 2024; Jiang et al., 2023; Touvron et al., 2023), permitindo que a comunidade de desenvolvedores utilizem esses

[2] https://gemini.google.com/
[3] https://claude.ai/
[4] https://copilot.cloud.microsoft/
[5] https://llama.meta.com/
[6] https://mistral.ai/
[7] https://azure.microsoft.com/pt-br/products/phi-3

sistemas em suas aplicações com infraestruturas computacionais próprias, propondo a integridade e privacidade de seus dados.

De forma geral, os modelos de IA tradicionais são construídos a partir de grandes quantidades de dados e são treinados para identificar padrões e fazer previsões ou auxiliar em decisões com base nesses padrões. Esses modelos utilizam técnicas avançadas de aprendizado de máquina, como redes neurais profundas, que permitem a modelagem de relações complexas nos dados (Bengio, Goodfellow, & Courville, 2017). Dependendo da tarefa e da abordagem utilizada, existem diferentes tipos de modelos de IA, cada um com suas características e aplicações específicas.

Modelos de aprendizado supervisionado são treinados com conjuntos de dados relacionados para prever novas saídas com base nestas entradas, como na previsão de preços de imóveis (Pai & Wang, 2020). Já os modelos de aprendizado não supervisionado operam com dados sem rótulos, buscando padrões ou agrupamentos, como o algoritmo K-means (Bengio, Goodfellow, & Courville, 2017). O aprendizado por reforço, por sua vez, envolve a tomada de decisões em sequência, com base em recompensas ou penalidades, sendo amplamente utilizado em robótica e jogos (Google, 2016). Redes neurais, inspiradas no cérebro humano, são eficazes em tarefas complexas, como reconhecimento de voz e visão computacional (Vargas, Carvalho e Vasconcelos, 2016). Por fim, modelos generativos, como as GANs (Redes Generativas Adversariais), não apenas identificam padrões, mas também geram novos exemplos, como textos e imagens (Goodfellow et al., 2014) – foco deste trabalho.

Uma IAG é uma técnica de inteligência artificial que utiliza modelos de linguagem generativos como aproximadores estatísticos para gerar conteúdo novo com base em um prompt (Radford et al., 2019). Esses modelos funcionam mapeando a proximidade de termos, palavras, tokens e contextos presentes nos dados de treinamento (Brown, 2020). Em vez de apenas reconhecer padrões ou fazer previsões baseadas em dados existentes, a IAG é capaz de criar exemplos que refletem as características e regras implícitas nos dados originais, produzindo saídas que são estatisticamente coerentes com o conteúdo que foi utilizado para seu treinamento (Feuerriegel et al., 2024), seja em texto, imagem, áudio ou vídeo.

Tratando de IAG textual, esta foi concebida inicialmente para aplicações como tradução automática, chatbots, correção ortográfica e geração de textos explicativos a partir de um conjunto de dados (Radford, 2018). Estas tarefas já estavam introduzidas em nosso contexto e não são novidade para usuários de tecnologia nos últimos anos. O advento em que vivemos, no qual a IAG responde a perguntas nos mais diversos temas, exige modelos mais complexos, conhecidos como *Large Language Models* (LLMs), ou – Modelos de Linguagem de Grande Escala (Devlin, 2019). E este é o elemento por trás da discussão que iremos realizar neste texto. Estes modelos podem ser classificados em "*open-source*" e "*closed-source*" (Marr, 2024).

Os modelos "*open-source*" são desenvolvidos por uma comunidade ou organização e estes são disponibilizados para a comunidade desenvolvedora de software. Os processos de modelagem do sistema e os dados de treinamento são abertos, permitindo com que os utilizadores possam auditar os dados utilizados para verificar possibilidade de vieses e auditar o processo de modelagem do design do sistema, decidindo se este processo está de acordo com as expectativas esperadas para a funcionalidade de utilização. Estes modelos podem ser utilizados em instalações próprias, e podem ser modificados, de acordo com as regras da licença a qual foi definida pelo criador do modelo de fundação. São exemplos de LLM *open-source*: Llama, Mistral, Vicuna, Phi e OpenOrca.

Já os modelos "*closed-source*" são desenvolvidos por organizações que não disponibilizam na integra todo o processo utilizado para a construção do modelo. O usuário também não tem acesso a todos os dados de treinamento utilizados, visto que este pode utilizar dados abertos (como por exemplo dados da wikipedia, portais de transparência governamentais, blogs, etc) e dados privados (como por exemplo dados de jornais e revistas com acesso restrito, dados internos da organização,

etc). Estes modelos não são passíveis de utilização em ambientes próprios, ficando o usuário dependente da interface disponibilizada pela organização que criou o modelo. São exemplos de LLM *closed-source*: GPT, Gemmini, Claude e CoPilot.

A IAG baseada em LLMs (*open* e *closed source*) tem sido explorada em diversos contextos da produção científica, com promessas de contribuições para diferentes etapas do processo de pesquisa. A seguir, é descrito o uso da IAG em áreas como a geração de hipóteses, revisão de literatura, extração de informação, sumarização, escrita científica, tradução e análise de dados, contextualizando com trabalhos que já utilizam a prática em experimentação e produção final.

2.2. IAG na Produção Científica

Uma das promessas da IAG na produção científica é auxiliar pesquisadores na geração de ideias e *insights*, gerando hipóteses e sugerindo abordagens baseadas na análise de grandes volumes de dados ou textos científicos. Esse processo também pode facilitar a identificação de tendências emergentes dentro de um campo específico (Viswa et al., 2024).

No que diz respeito à revisão de literatura, a proposta da IAG é de poder processar grandes quantidades de textos e realizar o processo de inclusão e exclusão de literatura. (Landshaft et al., 2024) em sua pesquisa realiza um experimento no qual avalia a capacidade do GPT-4 em realizar a triagem de resumo dos trabalhos selecionados em uma revisão bibliográfica. Compara o resultado obtido com o mesmo processo em uma avaliação humana realizada por pares e, de acordo com o autor, obteve êxito na tarefa. Dashkevych e Portnov (2024) realizaram um experimento no qual acreditam que apesar das IAGs apresentarem inconsistências com fontes de dados citadas, elas "podem ajudar a preencher lacunas no resumo dos estudos de base e a simplificar o design da pesquisa, ao complementar informações ausentes ou negligenciadas". e destacar os trabalhos mais relevantes, otimizando a revisão bibliográfica de um projeto (Zala et al., 2024).

Para a processo da escrita científica, a IAG tem como proposta ser uma ferramenta útil na criação de rascunhos de artigos, relatórios ou resumos, oferecendo uma estrutura inicial que os pesquisadores podem revisar e adaptar de acordo com suas necessidades (Wang, Hsiao e Chang, 2020). Outra atividade da escrita científica está ligada à tradução, no qual a IAG pode ser utilizada para traduzir artigos científicos com precisão técnica e linguística. Além disso, ela pode sugerir melhorias em termos de gramática, coesão e clareza, adaptando o texto para padrões exigidos em publicações acadêmicas (Nasser e Awadh, 2024).

Na área de extração de informação, Polak e Morgan (2024) apresentam um fluxo de trabalho com IAG para a extração de entidades, relações e propriedades na área de ciência dos materiais. O experimento aborda tanto IAGs com LLMs *open-source* quanto *closed-source*, apresentando resultados semelhantes para modelos de mesmo porte em quantidade de parâmetros de treinamento. Já na análise de dados, a aplicação da IAG em grandes volumes de informações pode auxiliar na identificação de padrões, na realização de análises estatísticas complexas, e até na criação de gráficos e visualizações complexas (Combrinck, 2024).

Levando em conta que a IAG está adentrando na ciência nas mais diversas áreas, a comunidade científica tem se preocupado com o rumo que estas utilizações estão tomando e começam a discutir os riscos da utilização desta tecnologia na ciência. Arbix (2024) apresenta um relato encontrado de "erros flagrantes em artigos científicos" (Arbix, 2024), que cada vez mais levantam uma "*red flag*" de editoras e pesquisadores. Arbix relata que a "A University College London rastreou milhões de artigos científicos e identificou que pelo menos 60 mil artigos, só em 2023, foram publicados com base em algum tipo de recurso em IA".

Esta preocupação do uso de IA de forma desenfreada não é por acaso. Ainda não existe um conceito oficial formado sobre ética na utilização destas ferramentas. Riscos de plágio e direitos

autorais podem estar sendo infringidos devido à natureza da construção destes modelos, que respondem baseados em fatos previamente processados na construção do modelo de linguagem. Lobo (2023) alerta que trabalhos que utilizam IA na escrita científica possivelmente estão cometendo algum tipo de plágio.

Noorden e Perkel (2023) entrevistaram 1600 pesquisadores para identificar pontos positivos e negativos da IA generativa. Dentre os impactos negativos, destacam-se dois elementos apontados neste trabalho como grandes preocupações dos cientistas: a reprodutibilidade da pesquisa e vieses e descriminação que os dados de treinamento podem perpetuar nos modelos de linguagem.

Entendendo este cenário, na seção a seguir é introduzido o conceito de CA com algumas recomendações de princípios definidos pela UNESCO (2021), de forma a possibilitar a realização de uma discussão sobre os princípios orientadores da CA a partir da utilização destes métodos de IAG com LLM para a produtividade científica.

2.3. Ciência Aberta

O conceito de CA é antigo e é abordado na literatura em diversos momentos. David (1998), destaca que a CA não é apenas uma prática científica, mas um constructo social complexo que se desenvolveu ao longo da história. Ele sugere que as instituições da ciência aberta são legados da história europeia e que a sua eficácia depende de um ambiente de apoio, como o patronato público aristocrático e a proteção das normas de cooperação e divulgação de informações.

A CA não possui uma definição única e fixa. Ela abrange um conjunto de princípios destinados a fomentar o crescimento científico e seu acesso ao público em geral. (Fecher e Friesike, 2014) identificaram diferentes "escolas de pensamento" dentro da CA, cada uma destacando aspectos variados como infraestrutura, medição de impacto científico, acessibilidade pública e democratização da ciência.

Entretanto, atualmente A UNESCO tem trabalhado para que a CA seja um movimento unico que busca democratizar o acesso ao conhecimento científico, promovendo a transparência, a colaboração e a acessibilidade em todas as etapas do processo de pesquisa. Ela se fundamenta (UNESCO, 2021) na ideia de que os resultados científicos, incluindo dados, publicações, metodologias e software, devem estar disponíveis gratuitamente e acessível para todos. Com essas recomendações, acredita-se que não só ampliamos a disseminação do conhecimento, mas também aceleramos o progresso científico ao permitir que outros pesquisadores de diferentes regiões e disciplinas construam sobre trabalhos existentes sem as barreiras impostas por restrições de acesso, fomentando um ambiente mais inclusivo e equitativo para a produção e utilização da ciência.

Com base nas recomendações da UNESCO, Silveira et al. (2023) propõem uma taxonomia de Ciência Aberta dividida em vários objetivos principais: o "Conhecimento Científico Aberto", que defende o acesso gratuito a resultados de pesquisa, promovendo equidade na disseminação do conhecimento; a "Infraestrutura Científica Aberta" engloba plataformas tecnológicas e ferramentas como repositórios de dados e software de código aberto para a gestão eficiente de dados; a "Comunicação Científica" abrange estratégias para divulgar pesquisas de forma acessível, como publicações de acesso aberto e preprints; o "Envolvimento Aberto dos Atores Sociais" destaca a participação de diferentes grupos sociais no processo de pesquisa, enriquecendo-a com múltiplas perspectivas; e o "Diálogo Aberto com Outros Sistemas de Conhecimento" incentiva a integração de saberes tradicionais, ampliando o entendimento interdisciplinar e intercultural da ciência.

Para atingir estes objetivos, a UNESCO (2021) define quatro valores centrais que decorrem de "implicações jurídicas, éticas, epistemológicas, econômicas, legais, políticas, sociais, de múltiplos atores e tecnológicas da abertura da ciência à sociedade": qualidade e integridade, benefício coletivo, equidade e justiça e diversidade e inclusão.

A fim de possibilitar condições e práticas dentro dos valores citados, também é definido seis princípios orientadores para que os ideais da ciência aberta se tornem realidade (UNESCO, 2021): Transparência, escrutínio, crítica e reprodutibilidade; igualdade de oportunidades; responsabilidade, respeito e prestação de contas; colaboração, participação e inclusão; flexibilidade; e sustentabilidade.

Analisamos estes seis princípios e selecionamos aqueles que de alguma forma podem impactar na utilização de IAGS com LLMs no fazer científico. A análise é apresentada na seção a seguir.

3. ANÁLISE DA IAG EM RELAÇÃO AOS PRINCÍPIOS DE CA SEGUNDO A UNESCO (2021)

Esta seção explora a intersecção dos princípios da CA segundo a UNESCO (2021) com relação a aplicação de IAG com LLMs em métodos da produção científica. A análise será realizada em cada um dos seis princípios. Em alguns dos casos será analisado identificando o método aplicado e o tipo de LLM (*open* ou *closed-source*) utilizado, apresentando hipóteses de cumprimento ou não cumprimento do princípio (entende-se como método aplicado as quatro atividades em que a IAG está sendo utilizada na produção científica citadas na seção 3.1.1: Revisão de literatura, escrita científica, tradução e análise de dados). Em outros casos, devido à natureza dos princípios, a análise será baseada na equidade e concentração de poder, problemas estes também apresentado na introdução deste trabalho e que também se relaciona com os princípios da CA.

3.1. Princípio da Transparência, Escrutínio, Crítica e Reprodutibilidade

O princípio da Transparência, Escrutínio, Crítica e Reprodutibilidade diz que:

> *"...deve-se promover uma maior abertura em todas as etapas do empreendimento científico, com o objetivo de reforçar o poder e o rigor dos resultados científicos, aumentar o impacto social da ciência e ampliar a capacidade da sociedade como um todo de resolver problemas complexos e interligados. Mais abertura leva a mais transparência e confiança na informação científica e reforça a característica fundamental da ciência, como uma forma distinta de conhecimento com base em evidências e verificado em relação à realidade, à lógica e ao escrutínio dos pares científicos"* (UNESCO, 2021).

De forma geral, podemos considerar que LLMs *closed-source* não possuem transparência nem em dados de treinamento nem na modelagem do sistema (Achiam et al., 2023). Desta forma qualquer utilização feriria este princípio. Consequentemente, escrutínio, crítica e reprodutibilidade estariam também comprometidos devido à falta da transparência.

A utilização de LLMs *open-source* em um primeiro momento parece apresentar possibilidades de intersecção com o princípio, devido à natureza da definição do conceito de "*open-source*" (OSI, 2007). Os dados de treinamento e a metodologia e design de implementação são públicos e permitem avaliar possibilidades de vieses na resposta. Entretanto, avaliar os dados e a complexidade da decisão com a grande quantidade de parâmetros não é um processo tão trivial (Burrell, 2016). Desta forma, poderíamos dizer que o princípio é coberto, entretanto a hipótese é de que é cumprido com translucidez para o processo de transparência.

Com relação à reprodutibilidade, os modelos *open-source* atuais apresentam mecanismos de replicação dos processos de forma determinística (Savant, 2024) no qual pode-se configurar a execução de um modelo para seguir sempre a mesma decisão para um mesmo *prompt* de entrada. Para este caso, a hipótese é de que este processo está em cumprimento como parte do princípio.

Analisando pela ótica das aplicações, consideramos apenas IAGs com LLMs *open source*, visto que o princípio não foi atingido em nenhuma hipótese para *closed-source*. Na revisão de literatura, ao

utilizar o processo para decisão e extração de dados, ainda é obscuro o cenário de validação de acurácia do processo. Os trabalhos recentes que exploram esta área são categóricos ao mostrarem que o método carece da técnica "*human-in-the-loop*" (Dagdelen et al., 2024; Landschaft et al., 2024) para validação, visto que a acurácia do processo ainda depende de intervenção humana na análise. Já métodos de análise de dados e tradução podem ser validados com especialistas, o que cumpriria o princípio em todos os tópicos. Na escrita científica não foi encontrado intersecção com o princípio devido à natureza da atividade. Logo não se aplicou a análise.

3.2. Princípio da Igualdade de Oportunidades

O princípio da Igualdade de Oportunidades diz que:

> *...todos os cientistas e outros atores e partes interessadas da ciência aberta, independentemente da localização, nacionalidade, raça, idade, gênero, renda, circunstâncias socioeconômicas, estágio da carreira, disciplina, língua, religião, deficiência, etnia ou situação migratória, ou qualquer outro motivo, têm as mesmas oportunidades de acesso, e contribuem e se beneficiam igualmente da ciência aberta."* (UNESCO, 2021).

Este é um princípio que dificilmente conseguirá ser seguido na íntegra, mesmo fora dos conceitos de IAG. O acesso a recursos computacionais básicos como computadores e internet é crítico em diversos países (World Bank, 2021), o que por si só já não permitiria que o princípio fosse cumprido. Trazendo para o contexto de IAGs e LLMs, o custo para se executar um LLM é alto, necessitando de computadores com grandes capacidades de GPU de alta performance (Smith et al., 2023). Desta forma, o acesso a IAGs com LLMs *open-source* estaria limitado à apenas aqueles que tenham acesso ao recurso, não cumprindo com o princípio.

Para IAGs com LLMs *closed-source* a única limitação para cumprir este princípio seria o acesso à internet, visto que a disponibilização do serviço (consequentemente a infraestrutura custosa) estaria por conta do fornecedor da tecnologia. Entretanto, ressalta-se que cumpriria a equidade de acesso, porém não cumpriria o princípio anterior de transparência.

Neste princípio, a análise baseada nas práticas científicas com IAG não foi aplicada devido a este já não ser contemplado para nenhum dos tipos de LLMs.

3.3. Princípio da Responsabilidade, Respeito e Prestação de Contas

O princípio da Responsabilidade, Respeito e Prestação de Contas diz que:

> *" ... a maior abertura traz mais responsabilidade para todos os atores da ciência aberta que, juntamente com a responsabilidade pública, a sensibilidade aos conflitos de interesse, a vigilância quanto às possíveis consequências sociais e ecológicas das atividades de pesquisa, a integridade intelectual e o respeito aos princípios éticos e às implicações relativas à pesquisa, devem formar a base para a boa governança da ciência aberta."* (UNESCO, 2021).

Este princípio, por ser inerente ao pesquisador, não será avaliado para os tipos de LLMs e para equidade e concentração de poder.

A avaliação pela ótica das aplicações levanta neste caso um ponto importante: no caso da escrita científica, ainda é preciso chegar em um consenso sobre direitos autorais e plágio. (Lobo, 2023) em seu trabalho defende que "... a pessoa física é responsável pela criação intelectual da obra e tem direitos legais para proteger sua autoria e controlar sua utilização e exploração" e que " cessa forma,

dificilmente o Chat GPT, ou qualquer outra IA poderia ser considerada como autora de qualquer obra, não podendo ser citado ou referenciado" (Lobo, 2023). Defende também que "qualquer criação elaborada diretamente pelo Chat GPT, sem os devidos créditos, pode ser considerada como plágio. Mais especificamente a obra pode conter plágio indireto [5] ou plágio mosaico [6], pois a ferramenta trabalha com uma base de dados gigantesca retirada da internet, possibilitando o acesso a qualquer material intelectual" (Lobo, 2023).

Já com relação ao processo de tradução e análise de dados, a hipótese apresentada é de que são atividades técnicas, e estas, tendo passado por revisão humana auxiliariam a no processo de equidade da ciência. A atividade de tradução por exemplo poderia dar condições semelhantes aos pesquisadores não anglófonos para incluírem seus trabalhos em revistas e congressos que possuem maior relevância e adotam a língua inglesa como oficial. Já a atividade de análise de dados auxiliaria as comunidades acadêmicas que não possuem especialistas em análise de dados para a criação dos scripts de análise.

3.4. Princípio da Colaboração, Participação e Inclusão

O princípio da Colaboração, Participação e Inclusão diz que:

> *" ... colaborações em todos os níveis do processo científico, para além dos limites de geografia, língua, gerações e recursos, devem se tornar a regra, e deve ser promovida a colaboração entre disciplinas, juntamente com a participação plena e efetiva dos atores sociais e a inclusão do conhecimento das comunidades marginalizadas na solução de problemas de importância social."* (UNESCO, 2021).

Este princípio corrobora com LLMs *open-source*. Garantir que as soluções implementadas tenham conhecimento das comunidades marginalizadas só é garantido em sistemas no qual se tem conhecimento sobre os dados de construção do modelo (já discutido no primeiro princípio). LLMs *closed-source* não têm como garantir que este princípio está incluso nos modelos, visto a questão de transparência dos dados, já discutido no primeiro princípio.

3.5. Princípio da Flexibilidade

O princípio da Flexibilidade diz que:

> *" ...devido à diversidade de sistemas científicos, atores e capacidades em todo o mundo, bem como à natureza em constante evolução do apoio às tecnologias de informação e comunicação (TIC), não existe uma forma única de se praticar a ciência aberta. Devem ser incentivados diferentes caminhos de transição e prática da ciência aberta, ao mesmo tempo em que se mantêm os valores centrais supracitados e se maximiza a adesão aos outros princípios aqui apresentados."* (UNESCO, 2021).

LLMs *open-source* permitem que diferentes atores (instituições acadêmicas, startups, pesquisadores independentes) adaptem os modelos às suas necessidades específicas, favorecendo a personalização e a democratização do conhecimento científico. A flexibilidade está presente na possibilidade de modificar, otimizar e adaptar os modelos para diferentes fins, enquanto se promove a adesão a princípios como a transparência e a reprodutibilidade dos resultados.

3.6. Princípio da Sustentabilidade

O princípio da Sustentabilidade diz que:

> *" ... para ser tão eficiente e causar tanto impacto quanto possível, a ciência aberta deve se basear em práticas, serviços, infraestruturas e modelos de financiamento de longo prazo que garantam a participação igualitária dos indivíduos que produzem ciência originários de instituições e países menos privilegiados. As infraestruturas científicas abertas devem ser organizadas e financiadas com base em uma visão essencialmente sem fins lucrativos e de longo prazo, que aprimorem as práticas de ciência aberta e garantam o acesso permanente e irrestrito a todos, na medida do possível."* (UNESCO, 2021).

O uso de LLMs *open-source* está alinhado ao princípio da sustentabilidade, pois esses modelos promovem acessibilidade e inclusão ao fornecer recursos que podem ser utilizados e adaptados por pesquisadores e instituições de qualquer lugar. Além disso, a filosofia *open-source* (OSI, 2007) tende a ser orientada por uma visão sem fins lucrativos, promovendo uma colaboração mais ampla e garantindo que o conhecimento e as ferramentas permaneçam acessíveis no longo prazo. Isso contribui para a criação de uma infraestrutura científica sustentável, permitindo que indivíduos de países e instituições menos privilegiadas possam participar e beneficiar-se dessas tecnologias.

Tratando de LLMs *closed-source*, por serem oferecidos como serviços pagos, podem limitar o acesso de pesquisadores e instituições de países menos privilegiados, restringindo a equidade no uso dessas tecnologias. Algumas empresas têm implementado programas de acesso gratuito ou subsidiado para determinadas instituições de ensino ou países [8], o que pode mitigar parte dessa barreira. Mesmo assim, a sustentabilidade a longo prazo de tais soluções dependeria da continuidade desses programas e de uma maior abertura em termos de acesso e uso.

4. DISCUSSÃO DA ANÁLISE

Sobre o primeiro princípio, é nítido que para a CA os modelos *open-source* estão mais próximos de cumprirem com o princípio. Entretanto, ainda é necessário avançar em técnicas de explicabilidade da IAG (Linardatos, 2021; Kandul et al., 2023) para que os processos executados a partir dos prompts sejam mais transparentes. Técnicas de *Retrieval Augmented Generation* (Lewis et al., 2020) já auxiliam na transparência do processo, possibilitando que este "referencie" o conteúdo principal que gerou a resposta ao prompt. Outros estudos pretendem desvendar a complexidade (Bhile e Maes, 2024) através de métodos de análises das decisões dos modelos, trazendo perspectivas de melhoria para esta área.

No segundo princípio, quando se trata de LLM *closed-source,* pode-se aparentar uma solução alternativa ao acesso rápido da tecnologia por todos os atores. Entretanto, podemos afirmar que neste caso as empresas fornecedoras poderão ditar as regras do que o sistema irá retornar, podendo incluir os vieses que desejam e reforçar a concentração de poder destas empresas no setor. Além disso, esta solução reforçaria o descumprimento do primeiro princípio - a transparência - princípio este que sempre foi base para a CA (David, 1998). Logo, o acesso a recursos computacionais deveriam ser prioridade para garantir este princípio.

No terceiro princípio, um ponto determinante para a utilização de IAGs com LLMs é a responsabilidade pela autoria e plágio, independentemente do tipo de LLM (*open ou closed-sourc*e). A discussão atual, como apontado por Lobo (2023), indica que IAGs não podem ser consideradas autoras de textos científicos, e a utilização de seus resultados sem a devida atribuição pode constituir plágio. Assim, o uso de IAGs para escrita científica ainda necessita de maior regulamentação para se adequar ao princípio. Já outras atividades de cunho técnico (tradução e análise) poderiam ser utilizadas

[8] https://openai.com/form/researcher-access-program/

a fim de mitigar diferenças entre o acesso determinadas tecnologias entre instituições de pesquisa, além da aproximação destes pesquisadores na comunidade científica global através da língua inglesa.

Para garantir que o quarto princípio se adeque, a solução está também na abertura dos modelos. Entretanto pouco se vê nas discussões atuais sobre construção de modelos LLMs, elementos que tratem do balanceamento do modelo com dados de conhecimentos de comunidades mais marginalizadas. Garantir que uma pesquisa que siga os princípios de CA é garantir que os modelos de fundação possam conter informações relevantes sobre estes.

O quinto princípio é algo já existente nas pesquisas em andamento. Modelos *open-source* estão constantemente em evolução, basta analisar uma das maiores bibliotecas [8] de LLMs open-source da atualidade na comunidade desenvolvedora – a HuggingFace. Diariamente são atualizados os modelos base e modelos que passaram por *Fine Tunning*. *Fine Tunning* é o processo de ajustar um modelo previamente treinado para uma tarefa específica, utilizando um conjunto de dados menor e mais focado. Em vez de treinar um novo modelo do zero, aproveita-se o conhecimento geral do modelo base e o adapta para uma aplicação particular, o que reduz o tempo e os recursos computacionais necessários (Howard e Ruder, 2018).

O sexto princípio mostra a importância de infraestruturas e práticas científicas que não apenas garantam a acessibilidade, mas que também sejam equitativas e de longo prazo, atendendo às necessidades de cientistas oriundos de instituições e países menos privilegiados. Ferramentas *closed-source*, como por exemplo a ClaudeAI e Gemini, citadas anteriormente, inicialmente teve os serviços disponibilizados nos Estados Unidos e só meses depois o serviço foi disponibilizado ao Brasil. Modelos LLMs *open-source* alinham-se a esse princípio de sustentabilidade ao oferecerem acessibilidade global, permitindo que qualquer pesquisador ou instituição, independentemente de sua localização geográfica, tenha acesso a ferramentas avançadas. Entretanto, o custo computacional ainda é um divisor para as comunidades menos favorecidas financeiramente.

CONSIDERAÇÕES FINAIS

O objetivo deste trabalho foi contextualizar o uso de IAGs e LLMs na ciência e apontar as suas limitações, apresentando a necessidade de incorporar os princípios éticos de ciência - em especial da CA - na sua utilização. A incorporação de IAGs com LLMs no processo científico é uma promessa de oportunidades significativas para o avanço da produção de conhecimento, ao mesmo tempo em que levanta preocupações sobre a transparência, a equidade e a integridade científica. Embora as IAGs tenham o potencial de acelerar a pesquisa e a criação de novos conhecimentos, é essencial que seu uso seja alinhado com os princípios da CA, através das sugestões padronizadas estabelecidas pela UNESCO (2021). O uso destas tecnologias como "caixas-pretas", seja em LLMs abertos ou fechados pode comprometer a integridade de uma pesquisa, conforme visto no primeiro princípio, no qual até mesmo os modelos *open-source* não apresentam total transparência.

Ao analisar a utilização de modelos de linguagem abertos e fechados, observamos que, enquanto os modelos *open-source* estão mais alinhados com os princípios de transparência e sustentabilidade, as soluções *closed-source* podem comprometer a equidade e a reprodutibilidade, aspectos centrais para a integridade científica. Portanto, há uma clara necessidade de maior regulamentação e desenvolvimento de políticas que garantam que o uso dessas tecnologias seja feito de forma responsável e inclusiva.

Além disso, a análise dos impactos das IAGs nas práticas científicas destacou a importância de garantir acesso igualitário a essas ferramentas, especialmente para pesquisadores em instituições e países menos privilegiados. A criação de infraestruturas sustentáveis, aliada a um uso consciente e ético das IAGs, poderá maximizar os benefícios dessas tecnologias, mantendo a confiança da comunidade científica e o compromisso com a ciência aberta e colaborativa. Por fim, é evidente que

o futuro da ciência passe por uma integração cuidadosa das IAGs, e cabe à comunidade acadêmica e científica e às instituições reguladoras garantir que essa transição ocorra de maneira ética, inclusiva e sustentável.

Para trabalhos futuros, paralelamente a este trabalho, também estamos analisando estudos sobre a acurácia destas ferramentas no processo de extração de conhecimento em artigos científicos na área de políticas públicas com LLMs *open-source* e também realizando discussões sobre as métricas e formas de medição de acurácia para este tipo de atividade. Outros trabalhos futuros podem focar no desenvolvimento de abordagens que mitiguem os desafios apontados, possibilitando uma construção coletiva de princípios éticos, promovendo o uso responsável dessas tecnologias para o benefício de toda a sociedade científica.

REFERÊNCIAS

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., … Zoph, B. (2023). GPT-4 Technical Report. http://arxiv.org/abs/2303.08774

Arbix, G. (2024). A inteligência artificial colocando em xeque a confiabilidade da ciência. Jornal da USP. https://jornal.usp.br/radio-usp/a-inteligencia-artificial-colocando-em-xeque-a-confiabilidade-da-ciencia/

Arkenberg, C. (2023). Generative AI is already disrupting media and entertainment. Deloitte. https://www2.deloitte.com/us/en/insights/industry/technology/generative-ai-tools-media-entertainment.html

BBC. (2016, March 12). Artificial intelligence: Google's AlphaGo beats Go master Lee Se-dol. https://www.bbc.com/news/technology-35785875

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623. https://doi.org/10.1145/3442188.3445922

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. http://arxiv.org/abs/2005.14165

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data and Society, 3(1). https://doi.org/10.1177/2053951715622512

Castelvecchi, D. (2016). Can we open the black box of AI? Nature News, 538(7623), 20. https://www.nature.com/news/polopoly_fs/1.20731!/menu/main/topColumns/topLeftColumn/pdf/538020a.pdf

Chan, L., Hall, B., Piron, F., Tandon, R., & Ottawa, L. W. (2020). Open Science Beyond Open Access: For and with communities A step towards the decolonization of knowledge Prepared for

the Canadian Commission for UNESCO.
https://en.ccunesco.ca//media/Files/Unesco/Resources/2018/11/IntroductionToUNESCOUpdatedReco

Cunha, M. Q. (2024). Utilizando Modelos de Linguagem Grandes para Classificação de Atos do Diário Oficial da União no Domínio Tributário. Dissertação de Mestrado. Programa de Pós-graduação em Ciência de Computação. Universidade Federal de Campina Grande (UFCG). https://bdtd.ibict.br/vufind/Record/UFCG_c636f6c13a6e4621d88be01d65151785

Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A. S., Ceder, G., Persson, K. A., & Jain, A. (2024). Structured information extraction from scientific text with large language models. Nature Communications, 15(1). https://doi.org/10.1038/s41467-024-45563-x

Davenport, T. H, Bean, R. (2023). The impact of generative AI on Hollywood and entertainment. Massachusetts Institute of Technology. https://sloanreview.mit.edu/article/the-impact-of-generative-ai-on-hollywood-and-entertainment/

David, P. A. (1998). Common Agency Contracting and the Emergence of "Open Science" Institutions (Vol. 88, Issue 2). American Economic Association.
https://www.jstor.org/stable/116885

Devlin, J., Chang, M.-W., Lee, K., & Google, K. T. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/abs/1810.04805

Ferrara, E. (2024). Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. In Sci (Vol. 6, Issue 1). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/sci6010003

Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. Business and Information Systems Engineering, 66(1), 111–126. https://doi.org/10.1007/s12599-023-00834-7

Ghebrehiwet, I., Zaki, N., Damseh, R., & Mohamad, M. S. (2024). Revolutionizing personalized medicine with generative AI: a systematic review. Artificial Intelligence Review, 57(5). https://doi.org/10.1007/s10462-024-10768-5

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
https://www.deeplearningbook.org/

Google. (2016). AlphaGo. https://deepmind.google/technologies/alphago

Grynbaum, M. M., & Mac, R. (2023). New York Times sues OpenAI and Microsoft over use of copyrighted work. The New York Times.
https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html

Guimarães Junior, J. C., Schmidt, F. L. A., Seixas, R., Favaro, D. M. M., Dos Santos, C. A. F., & Costa, H. C. de M. (2024). A contribuição da Inteligência Artificial na pesquisa científica. Contribuciones a las Ciencias Sociales, 17(3), e5590. https://doi.org/10.55905/revconv.17n.3-026

Hochmair, H. H., Juhasz, L., & Kemp, T. (2024). Correctness Comparison of ChatGPT-4, Bard, Claude-2, and Copilot for Spatial Tasks. https://doi.org/10.1111/tgis.13233

Holdsworth, J. (2023). O que é Viés de IA?. IBM. https://www.ibm.com/br-pt/topics/ai-bias

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. de las, Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. Le, Lavril, T., Wang, T., Lacroix, T., & Sayed, W. El. (2023). Mistral 7B. http://arxiv.org/abs/2310.06825

Khanal, S., Zhang, H., & Taeihagh, A. (2024). Why and how is the power of Big Tech increasing in the policy process? The case of generative AI. Policy and Society. https://doi.org/10.1093/polsoc/puae012

Khanal, S., Zhang, H., Taeihagh, A., & Kuan, L. (n.d.). Why and how is the power of Big Tech increasing in the policy process? The case of generative AI. https://doi.org/10.1093/polsoc/puae012/7636223

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. Burges, L. Bottou, & K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems (Vol. 25). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

Landschaft, A., Antweiler, D., Mackay, S., Kugler, S., Rüping, S., Wrobel, S., ... & Allende-Cid, H. (2024). Implementation and evaluation of an additional GPT-4-based reviewer in PRISMA-based medical systematic literature reviews. *International Journal of Medical Informatics*, *189*, 105531. https://www.sciencedirect.com/science/article/pii/S1386505624001941

Leonelli, S. (2018). Re-Thinking Reproducibility as a Criterion for Research Quality. http://www.nature.com/news/reproducibility-1.17552

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable ai: A review of machine learning interpretability methods. In Entropy (Vol. 23, Issue 1, pp. 1–45). MDPI AG. https://doi.org/10.3390/e23010018

Marr, B. (2024). Navigating the generative AI divide: Open-Source vs. Closed-Source solutions. Forbes. https://www.forbes.com/sites/bernardmarr/2024/04/22/navigating-the-generative-ai-divide-open-source-vs-closed-source-solutions/

Minsky, M., & Papert, S. (1969). Perceptrons: An introduction to computational geometry. Cambridge Tiass., HIT, 479. https://direct.mit.edu/books/monograph/3132/PerceptronsAn-Introduction-to-Computational

Minsky, M., Shannon, C., & Rochester, N. (2006). The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years. https://dl.acm.org/doi/10.1609/aimag.v27i4.1911

More, T. (1956). Dartmouth Conference on Artificial Intelligence Report onthe 5th and 6th Weeks.

Nas, E. (2024). Usos da IA em pesquisa científica: Entre a ética e a estética. Jornal da USP https://jornal.usp.br/artigos/usos-da-ia-em-pesquisa-cientifica-entre-a-etica-e-a-estetica/

Pai, P. F., & Wang, W. C. (2020). Using machine learning models and actual transaction data for predicting real estate prices. Applied Sciences (Switzerland), 10(17). https://doi.org/10.3390/app10175832

Pandolfini, B. (1997). Kasparov and Deep Blue: The historic chess match between man and machine. Simon and Schuster.

Polak, M. P., & Morgan, D. (2024). Extracting accurate materials data from research papers with conversational language models and prompt engineering. Nature Communications, 15(1). https://doi.org/10.1038/s41467-024-45914-8

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. https://api.semanticscholar.org/CorpusID:160025533

Rumelhart, D. E., & Hinton, G. E. (1986). Learning representations by back-propagating errors. Nature Publishing Group, 533–536.

Shanahan, M. (2024). Talking about Large Language Models. Communications of the ACM, 67(2), 68–79. https://doi.org/10.1145/3624724

Pedroso, J. S., Silva, K. S., & dos Santos, L. P. (2017). Pesquisa descritiva e pesquisa prescritiva. *JICEX*, *9*(9). https://unisantacruz.edu.br/revistas-old/index.php/JICEX/article/view/2604

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. http://arxiv.org/abs/2302.13971

UNESCO. (2021). Recomendação da UNESCO sobre Ciência Aberta CIÊNCIA ABERTA. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000379949_por

Vargas, A. C. G., Carvalho, A. M. P., & Vasconcelos, C. N. (2016). Um estudo sobre Redes Neurais Convolucionais e sua aplicação em detecção de pedestres. Proceedings of the xxix conference on graphics, patterns and images. http://gibis.unifesp.br/sibgrapi16/eproceedings/wuw/7.pdf

Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, 6000–6010. https://doi.org/0.5555/3295222.3295349

Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., Anandkumar, A., Bergen, K., Gomes, C. P., Ho, S., Kohli, P., Lasenby, J., Leskovec, J., Liu, T. Y., Manrai, A., … Zitnik, M. (2023). Scientific discovery in the age of artificial intelligence. In Nature (Vol. 620, Issue 7972, pp. 47–60). Nature Research. https://doi.org/10.1038/s41586-023-06221-2

Weizenbaum, J. (1966). ELIZA - A computer Program For the Study of Natural Language Communication Between Man And Machine. 9, 36–45. https://dl.acm.org/doi/pdf/10.1145/365153.365168

Zhang, H., Khanal, S., & Taeihagh, A. (2024). Public-Private Powerplays in Generative AI Era: Balancing Big Tech Regulation Amidst Global AI Race. Digital Government: Research and Practice. https://doi.org/10.1145/3664824

DECLARAÇÃO DE DISPONIBILIDADE DE DADOS DA PESQUISA: Todo o conjunto de dados de apoio aos resultados deste estudo foi publicado no próprio artigo.

FINANCIAMENTO: Esta pesquisa não recebeu nenhuma subvenção específica de qualquer agência de financiamento dos setores público, privado ou sem fins lucrativos.

CONTRIBUIÇÃO DAS/DOS AUTORES/AS:

José Rodolfo Beluzo: Conceituação, Metodologia, Redação, Preparação do rascunho original, Escrita, Visualização, Investigação e Edição

Gisele da Silva Craveiro: Conceitualização e Supervisão

DECLARAÇÃO DE CONFLITO DE INTERESSE: O autor e autora declaram que não há conflito de interesses a mencionar.

MINIBIOGRAFIAS DOS/DAS AUTORAS DO PAPER:

José Rodolfo Beluzo: Mestre em Sistemas de Informação e Doutorando no Programa Mudança Social e Participação Política pela Universidade de São Paulo; Professor na área de Informática no Instituto Federal de Educação Ciência e Tecnologia de São Paulo.

Gisele Craveiro: Doutora em Ciência de Computação pela Universidade de Campinas; Professora na Escola de Artes, Ciências e Humanidades na Universidade de São Paulo.

Este preprint foi submetido sob as seguintes condições:

- Os autores declaram que estão cientes que são os únicos responsáveis pelo conteúdo do preprint e que o depósito no SciELO Preprints não significa nenhum compromisso de parte do SciELO, exceto sua preservação e disseminação.
- Os autores declaram que os necessários Termos de Consentimento Livre e Esclarecido de participantes ou pacientes na pesquisa foram obtidos e estão descritos no manuscrito, quando aplicável.
- Os autores declaram que a elaboração do manuscrito seguiu as normas éticas de comunicação científica.
- Os autores declaram que os dados, aplicativos e outros conteúdos subjacentes ao manuscrito estão referenciados.
- O manuscrito depositado está no formato PDF.
- Os autores declaram que a pesquisa que deu origem ao manuscrito seguiu as boas práticas éticas e que as necessárias aprovações de comitês de ética de pesquisa, quando aplicável, estão descritas no manuscrito.
- Os autores declaram que uma vez que um manuscrito é postado no servidor SciELO Preprints, o mesmo só poderá ser retirado mediante pedido à Secretaria Editorial do SciELO Preprints, que afixará um aviso de retratação no seu lugar.
- Os autores concordam que o manuscrito aprovado será disponibilizado sob licença Creative Commons CC-BY.
- O autor submissor declara que as contribuições de todos os autores e declaração de conflito de interesses estão incluídas de maneira explícita e em seções específicas do manuscrito.
- Os autores declaram que o manuscrito não foi depositado e/ou disponibilizado previamente em outro servidor de preprints ou publicado em um periódico.
- Caso o manuscrito esteja em processo de avaliação ou sendo preparado para publicação mas ainda não publicado por um periódico, os autores declaram que receberam autorização do periódico para realizar este depósito.
- O autor submissor declara que todos os autores do manuscrito concordam com a submissão ao SciELO Preprints.