CONFIABILIDADE DA IA GENERATIVA: UM FLUXOGRAMA PARA ANÁLISE DAS RESPOSTAS DO CHATGPT EM RELAÇÃO À DESINFORMAÇÃO CIENTÍFICA

Palavras-Chave: inteligência artificial generativa, ChatGPT; conhecimento; desinformação científica.

Autores(as):

PRISCILA CRISTINA DOURADO SALVADEO, NIED – UNICAMP

Dra. FLÁVIA LINHALIS (orientador(a)), NIED - UNICAMP

INTRODUÇÃO:

A internet facilitou o acesso a informações por meio de navegadores, sites, enciclopédias virtuais e, mais recentemente, chatbots. No entanto, a desinformação tornou-se um problema crescente e um desafio mundial, devido ao excesso de informações, dificuldade em encontrar fontes confiáveis e avanços tecnológicos (Mancoso, 2023). Deste modo, é comum que as pessoas consumam conteúdos de veículos de comunicação sem verificar a veracidade dos fatos (Mancoso, 2023). Além disso, muitos conteúdos na internet são produzidos sem respaldo científico ou cujas fontes de dados não são claras. Assim, ao introduzir uma nova ferramenta na sociedade, como o ChatGPT, surge a necessidade de testes e estudos contínuos, pois a qualidade das respostas e suas limitações dependem da estrutura e dos conteúdos oferecidos em seu treinamento. A confiabilidade desta ferramenta é um desafio atual, uma vez que, apesar de oferecer textos elaborados, isso não assegura que os conteúdos sejam confiáveis e livres de desinformação.

Os chatbots com Inteligência Artificial Generativa despertam muitas discussões em diferentes áreas, como ética, autoria (Vasconcellos, 2023), e educação (Rodrigues; Rodrigues, 2023). Para além disso, considerando que a ferramenta está disponível de forma gratuita e acessível à sociedade, com um número recorde de usuários, é preciso analisar criticamente e considerar a incerteza sobre a confiabilidade científica dos textos gerados por essa nova ferramenta (Javaid; Haleem; Singh, 2023). O ChatGPT-3.5 é treinado por humanos, utilizando reforço por feedback (Haque; Li, 2024), o que pode causar vieses e preconceitos em suas respostas (de Oca Rodrigues et al. 2023). Além disso, durante o treinamento, esses sistemas têm acesso a dados da internet, o que levanta questionamentos sobre a possibilidade de a Inteligência Artificial (IA) se tornar mais um recurso em potencial para disseminar desinformação.

O objetivo deste artigo é orientar o uso do ChatGPT como ferramenta de pesquisa por meio de um fluxograma de decisão, além de analisar os textos gerados, apresentando um panorama dos mesmos frente a questões de desinformação. Pretende-se analisar, através desse fluxograma, se o chatbot é capaz de gerar textos que estejam em conformidade com o consenso científico e sem desinformação, realizando testes comparativos entre o ChatGPT-3.5 e o ChatGPT-4.o mini e 4.o em sua versão gratuita. Esses testes utilizaram o tema da alimentação, abordando questões que frequentemente aparecem distorcidas em sites e nas redes sociais, gerando desinformação. O fluxograma é um guia que pode ser usado por especialistas para averiguar a confiabilidade das respostas do ChatGPT.

METODOLOGIA:

Esta é uma pesquisa qualitativa de caráter descritivo, que procura entender a possibilidade de o ChatGPT gerar conteúdos de desinformação científica e que propõe uma abordagem mais consciente para o seu uso, por intermédio de um fluxograma. O fluxograma foi elaborado com base no documento “ChatGPT e Inteligência Artificial na Educação Superior: Guia de Início Rápido”, que contém o fluxograma de Aleksandr Tiulkanov, publicado pela UNESCO (UNESCO, 2023) e apresentado na Figura 1. Utilizou-se como fundamentação teórica o trabalho de pesquisadores epistemólogos, como Douglas Allchin, Dietmar Höttecke e Jonathan Osborne, que realizaram trabalhos sobre a confiabilidade das informações segundo a ciência na era midiática.

Além disso, o documento "Science Education in an Age of Misinformation", traduzido para o português como "Educação em Ciências em Tempos de Desinformação", de Osborne (2023) e seu grupo, fundamentou este estudo. Os conhecimentos apresentados nesses documentos foram aplicados de maneira crítica para a elaboração do diagrama de fluxo e a investigação da confiabilidade dos textos gerados. Os prompts dirigidos ao chatbot foram selecionados com base em seções de fontes confiáveis, como o site do Butantan, Fiocruz e o Jornal da USP, que desmentem informações falsas e orientam corretamente a população. Posteriormente, os textos gerados pelo chatbot em resposta a esses prompts foram analisados conforme o fluxograma. Os prompts testados foram: “Alimentos transgênicos fazem mal à saúde?”, “Água com limão deixa o sangue alcalino?” e “Vitamina C cura gripes, resfriados ou a Covid-19?”. Esses temas foram escolhidos por serem alvo de desinformação na internet, em sites e redes sociais, algo que prejudica a população ao receber e acessar informações falsas e conteúdos enganosos. Portanto, é crucial verificar a capacidade do chatbot de reconhecer desinformação e responder corretamente sobre esses temas.

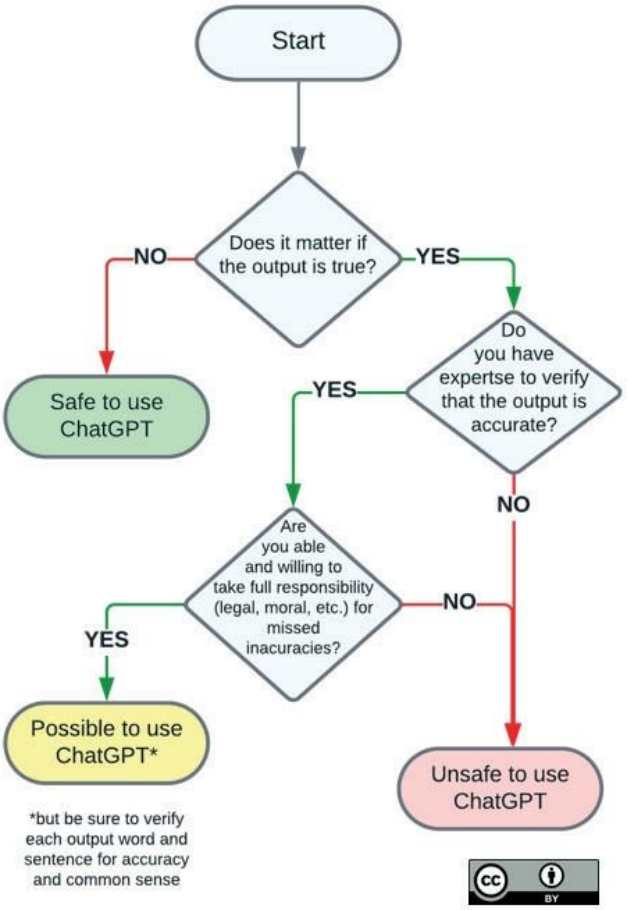

Figura 1- Fluxograma de Aleksandr Tiulkanov, 2023. Fonte: https://unesdoc.unesco.org/ark:/48223/pf0000385146

RESULTADOS E DISCUSSÃO:

Nesta seção, apresentamos o fluxograma e os testes realizados com o ChatGPT, com a finalidade de verificar a adequação do fluxograma para a análise de respostas em relação à desinformação. Avaliamos a capacidade do ChatGPT de fornecer informações precisas e confiáveis, destacando áreas onde ocorre desinformação. O fluxograma desenvolvido orienta o usuário por etapas lógicas na análise das respostas do ChatGPT, verificando a precisão e a confiabilidade e detectando possíveis desinformações. As transições do fluxograma são referentes aos seguintes itens:

i) importância da confiabilidade (T1): refere-se a veracidade dos textos gerados pela IA ser importante. Se a verdade dos fatos não for importante, não é preciso continuar seguindo o fluxograma, assume-se que a resposta do Chat é segura. Por outro lado, caso a verdade dos fatos seja importante, é preciso que o agente humano tenha conhecimentos para verificar se o resultado está correto (UNESCO, 2023).

ii) importância na confiabilidade de quem dialoga (T2): quando se trata de desinformação, é preciso que o agente humano que está dialogando com a IA, tenha conhecimento no assunto (UNESCO, 2023). Caso o agente humano seja um especialista ou possua conhecimento suficiente para verificar informações com base no consenso científico, seguimos para a transição T3. Caso contrário, não é recomendado conversar com o Chat sobre um assunto passível de desinformação.

iii) completude da resposta (T3): diz respeito a verificar se a resposta está superficial, se há alguma informação importante que não foi mencionada na resposta. Além disso, a interação com ChatGPT, muitas vezes, gera um diálogo. Ao considerar o item T3, é importante verificar a coerência desse diálogo. Muitas vezes é preciso refazer a pergunta (mudar o prompt, o que é representado por um loop no fluxograma).

iv) corretude da resposta (T4): refere-se a todas as informações fornecidas pelo ChatGPT em sua resposta, verificando se estão de acordo com o consenso científico. Isso inclui observar se o texto gerado contém citações de associações e órgãos reconhecidos, ou se contém links para sites e artigos científicos como referência, permitindo rastrear a informação gerada.

v) responsabilidade ao utilizar as informações: diz respeito ao dever do usuário de verificar a precisão das informações geradas. O usuário é responsável por checar os fatos e assumir a responsabilidade por quaisquer falhas ou informações incorretas provenientes do conteúdo utilizado.

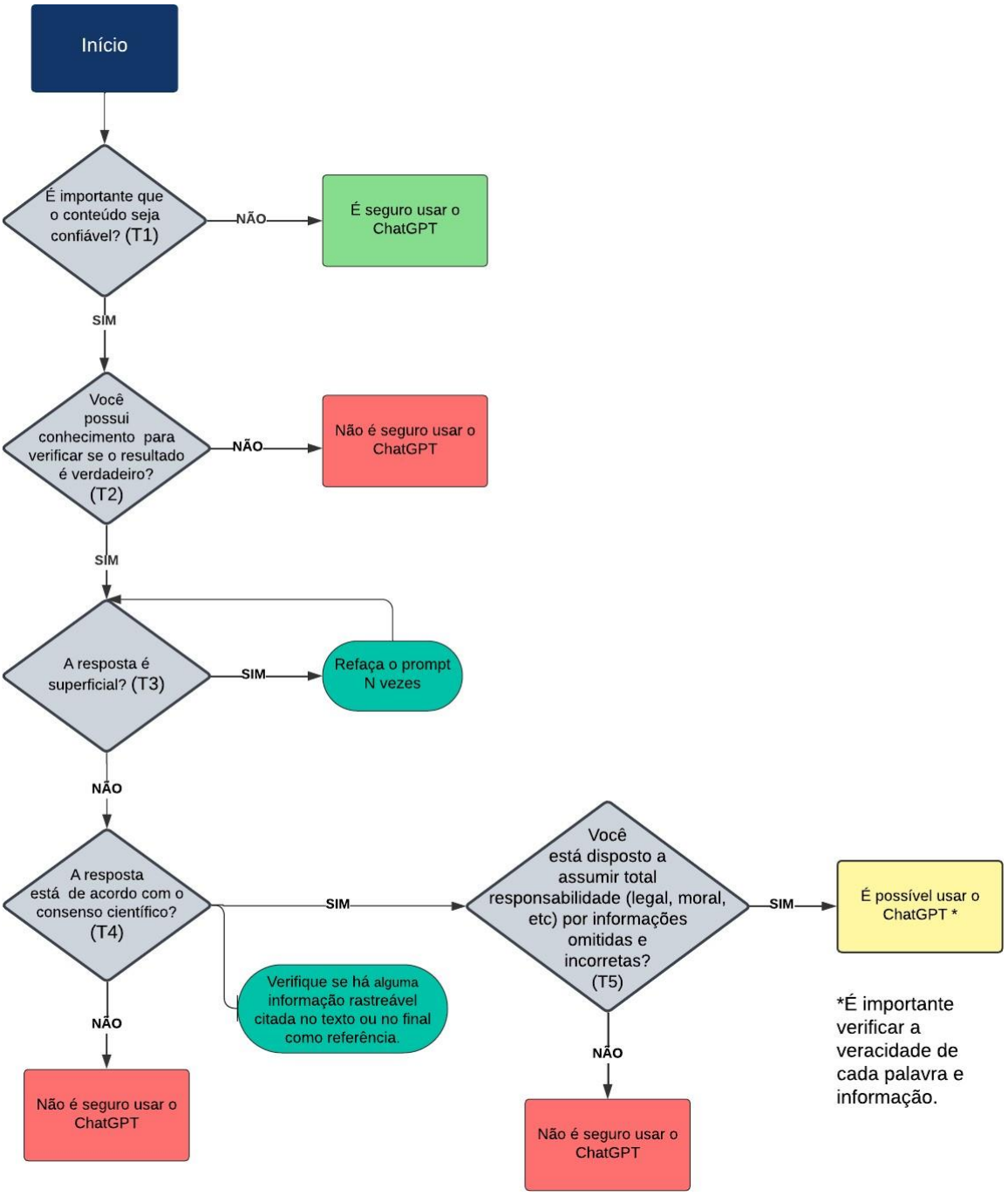

Figura 2 - Fluxograma para a análise das respostas do ChatGPT, visando um uso mais consciente e livre de erros de desinformação científica. Fonte: elaboração própria.

Validação do fluxograma

Para validar o fluxograma, realizamos testes com as versões gratuitas ChatGPT-3.5 e ChatGPT-4.o mini e 4.o em sua versão disponível com respostas limitadas. A versão gratuita 3.5 foi lançada ao público em novembro de 2022 e recentemente, foi disponibilizada a versão 4.0 mini, que é mais rápida e inteligente, além de ter um custo reduzido em relação ao modelo 3.5 (OpenAI, 2024). A validação do fluxograma, por meio de um processo de perguntas e respostas, foi conduzida pela primeira autora deste artigo, estudante do último semestre de graduação em Química e Física. Com conhecimento especializado na área, a autora avaliou as perguntas com base em literatura científica relevante, incluindo documentos, artigos e obras de referência, como *Princípios de Química*, de Atkins, e *Princípios de Bioquímica*, de Lehninger. Para o teste, assumimos que as transições T1, T2 e T3 já estavam validadas, e, portanto, analisamos as transições T3 e T4. Cada pergunta foi submetida de uma a três vezes, seguindo as transições do fluxograma desenvolvido, foram testadas as questões: "Alimentos transgênicos fazem

mal à saúde?", "Água com limão deixa o sangue alcalino?" e "Vitamina C cura gripes, resfriados ou a Covid-19?". Os resultados obtidos são apresentados a seguir.

Tabela 1 - Prompt: "Alimentos transgênicos fazem mal à saúde?"

Transições	GPT-3.5	GPT-4.o mini e GPT-4.o
T1	x	x
T2	x	x
T3	As respostas são adequadas para uma compreensão rápida do tema, sem aprofundar em aspectos essenciais como segurança, saúde, estudos e avaliações.	As respostas são mais detalhadas, pois abordam aspectos como segurança, saúde, estudos e avaliações.
T4	As respostas estão alinhadas com o consenso científico e mencionam a comunidade científica, incluindo órgãos como a OMS e a EFSA. No entanto, não apresenta referências ao final indicando a origem das informações.	As respostas estão alinhadas com o consenso científico e mencionam a comunidade científica, incluindo órgãos como a OMS, FAO, NAS, FDA e EFSA, mas não cita referências.
T5	x	x

Tabela 1 - Comparação das respostas das versões ChatGPT-3.5 e ChatGPT-4.o mini e 4.o sobre o prompt "Alimentos transgênicos fazem mal à saúde?".

Tabela 2 - Prompt: "Água com limão deixa o sangue alcalino?"

Transições	GPT-3.5	GPT-4.o mini e GPT-4.o
T1	x	x
T2	x	x
T3	A respostas são adequadas para uma compreensão rápida do tema, mas não aprofunda aspectos essenciais. Em algumas respostas o chatbot deixa de mencionar a faixa de pH do sangue e o sistema de regulação, além de não abordar acidose e alcalose.	As respostas apresentam mais detalhes, pois abordam aspectos de forma mais aprofundada, citando a faixa de pH do sangue e o sistema de regulação. No entanto, não menciona condições de acidose e alcalose.
T4	Em geral, o chatbot reconhece que a afirmação é controversa e informa que a comunidade científica a contesta. Além disso, o texto gerado está de acordo com o consenso científico. Porém, não cita organizações ou referências.	O chatbot é capaz de entender que se trata de um mito, respondendo de forma direta e adequada para desmistificar a desinformação. A resposta está de acordo com o consenso científico, mas não cita organizações ou referências.
T5	x	x

Tabela 2 - Comparação das respostas das versões ChatGPT-3.5 e ChatGPT-4.o mini e 4.o sobre o prompt "Água com limão deixa o sangue alcalino?"

Tabela 3 - Prompt: "Vitamina C cura gripes, resfriados ou a Covid-19?"

Transições	GPT-3.5	GPT-4.o mini e GPT-4.o
T1	x	x
T2	x	x
T3	Em geral, as respostas são adequadas para uma compreensão rápida do tema, sem aprofundamento em aspectos como o funcionamento da vitamina C no organismo.	Em geral, as respostas são adequadas para uma compreensão rápida do tema, sem aprofundamento em aspectos como o funcionamento da vitamina C no organismo.
T4	Em geral, as respostas sobre esse tema estão alinhadas com o consenso científico, informando corretamente que a vitamina C não é eficaz para a Covid-19. Além	Em geral, as respostas apresentaram-se de acordo com o consenso científico, com recomendações corretas, inclusive para o tratamento da

	disso, orientam sobre o uso de máscara, distanciamento social e a busca por atendimento médico. No entanto, não citam organizações ou referências.	Covid-19, conforme evidências científicas. Além disso, algumas das respostas geradas pela versão 4.o mencionam organizações como a OMS e o CDC dos EUA, mas nenhuma versão forneceu referências.
T5	x	x

Tabela 3 - Comparação das respostas das versões ChatGPT-3.5 e ChatGPT-4.o mini e 4.o sobre o prompt "Vitamina C cura gripes, resfriados ou a Covid-19?".

Os testes indicam que o fluxograma é considerado adequado para auxiliar um especialista a julgar a corretude, completude e, por fim, relevância das respostas do ChatGPT, elementos importantes quando se trata da confiança do conhecimento científico.

CONCLUSÕES E TRABALHOS FUTUROS:

Neste artigo, apresentamos um fluxograma desenvolvido com base no modelo apresentado pela UNESCO, para auxiliar especialistas ou indivíduos com conhecimento suficiente para verificar informações com base no consenso científico, na análise de conteúdos gerados pelo ChatGPT, com foco na confiabilidade do conhecimento científico. Para ilustrar a aplicação do fluxograma, realizamos testes com perguntas associadas à desinformação, ao seguir as etapas, verificamos que as respostas do ChatGPT-3.5 foram de acordo com o consenso científico, mas apresentaram pouca profundidade, enquanto as do ChatGPT-4.o Mini e 4.o foram respostas mais detalhadas, citando inclusive órgãos e organizações mundiais, ambos apresentaram textos conforme a comunidade científica e livres de desinformação, sendo que o 4.o demonstrou uma maior capacidade de fornecer respostas elaboradas. Os testes preliminares com o fluxograma se mostraram satisfatórios, com todas as transições sendo executadas sem maiores problemas. Uma limitação do resultado das respostas é a não replicabilidade dos resultados, ou seja, embora a pergunta possa ser facilmente replicada, não há garantia de que o ChatGPT fornecerá as mesmas respostas. Na medida que o chatbot passa por aprimoramentos, será necessário refazer as perguntas para testar a estabilidade dos resultados, isto é, verificar se os resultados se mantêm constantes no sentido da coerência e da confiança científica. Essa limitação não se aplica ao fluxograma, que pode ser utilizado como guia, mesmo que o ChatGPT seja atualizado. Neste artigo, apresentamos um primeiro teste do fluxograma. Como trabalhos futuros, pretendemos incluir mais especialistas, mais perguntas nos testes e a opinião de especialistas com relação a seguir o fluxograma como guia.

BIBLIOGRAFIA

DE OCA RODRIGUES, Golbery; ALBUQUERQUE, Danyllo W.; JESUALDO, G. Análise de Vieses Ideológicos em Produções Textuais do Assistente de Bate-papo ChatGPT. In: **Anais do IV Workshop sobre as Implicações da Computação na Sociedade**. SBC, 2023. p. 148-155.

HAQUE, Md Asraful; LI, Shuai. Exploring chatgpt and its impact on society. **AI and Ethics**, p. 1-13, 2024.

HÖTTECKE, Dietmar; ALLCHIN, Douglas. Reconceptualizing nature-of-science education in the age of social media. **Science Education**, v. 104, n. 4, p. 641-666, 2020.

JAVAID, Mohd; HALEEM, Abid; SINGH, Ravi Pratap. A study on ChatGPT for Industry 4.0: Background, potentials, challenges, and eventualities. **Journal of Economy and Technology**, v. 1, p. 127-143, 2023.

MANCOSO, Kaique et al. Research on misinformation and science communication: a review of the Latin American literature. **Journal of Science Communication-América Latina**, v. 6, n. 1, p. A01, 2023.

OpenAI (2024). Site oficial da OpenAI. GPT-4o mini: advancing cost-efficient intelligence. Disponível em: https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. Acesso em julho de 2024.

OSBORNE, Jonathan; PIMENTEL, Daniel. Science education in an age of misinformation. **Science Education**, v. 107, n. 3, p. 553-571, 2023.

RODRIGUES, Olira Saraiva; RODRIGUES, Karoline Santos. A inteligência artificial na educação: os desafios do ChatGPT. **Texto Livre**, v. 16, p. e45997, 2023.

UNESCO. **ChatGPT and artificial intelligence in higher education:** quick start guide, 2023. UNESCO International Institute for Higher Education in Latin America and the Caribbean. Disponível em: https://unesdoc.unesco.org/ark:/48223/pf0000385146. Acesso em: Dezembro de 2023.

VASCONCELLOS, Vinicius Gomes de. Editorial–Inteligência artificial e coautoria de trabalhos científicos: discussões sobre utilização de ChatGPT em pesquisa e redação científicas. **Revista Brasileira de Direito Processual Penal**, v. 9, n. 3, p. 1047-1057, 2023.