XXXIII Congresso de Iniciação Científica da Unicamp

INFORMAÇÕES ERRADAS , ACUSAÇÕES FALSAS, LINGUAGEM AMBÍGUA – INTELIGÊNCIA ARTIFICIAL BASEADA EM GRANDES MODELOS DE LINGUAGEM NÃO DEVE SER USADA PARA PROCURAR REFERÊNCIAS DESPUBLICADAS – ESTUDO COM 21 DIFERENTES CHATBOTS

Palavras-Chave: GRANDES MODELOS DE LINGUAGEM, *CHATBOTS*, RETRATAÇÕES

Autores(as):
MARIA FERNANDA DE ÁVILA REIS, CCV – PUC-CAMPINAS e BIOS, MATHEUS DA SILVA FAGO CCV - PUC-CAMPINAS e BIOS, ROSANA CELESTINA MONRADIN-REIS, FCM - UNICAMP, JOÃO BATISTA FLORINDO, IMECC - UNICAMP e BIOS
Prof. Dr. KONRADIN METZE (orientador), FCM - UNICAMP e BIOS

INTRODUÇÃO:

A atual crise da ciência tem como uma das suas características o vertiginoso aumento das despublicações de artigos científicos *[1,2,3*]. Portanto, é necessário que cada autor verifique individualmente a validade de cada referência a ser incluída no seu manuscrito. Este processo é lento e tedioso. Neste contexto, o uso dos *Large Language Models (LLMs*) poderia ser útil, pois a maioria desses programas é gratuita e o seu uso não necessita de um treinamento especial.

O uso de *chatbots* porém, não é isento de problemas, uma vez que eles podem fornecer informações falsas ou inventadas (alucinações), às vezes não imediatamente detectáveis. *[4,5,6*]

O objetivo do nosso trabalho foi avaliar no experimento com 21 diferentes Grandes Modelos de Linguagem, a capacidade de detectar, em uma lista de referências bibliográficas, os trabalhos previamente retratados e, simultaneamente, pesquisar manuscritos válidos, mas falsamente classificados como retratados pelos *chatbots.*

METODOLOGIA:

Procuramos na base de dados do Web of Science, da Clarivate, as referências bibliográficas de: 50 trabalhos retratados do pesquisador do Joachim Boldt; 50 trabalhos não

retratados do mesmo autor e (se); 32 publicados não retratados com o sobrenome Boldt e a letra inicial J, do primeiro nome. Totalizou-se 132 referências. Submetemos esta lista a 21 diferentes *chatbots* com o seguinte *prompt: "I am writing a scientific paper and preparing the reference list. I am using an older database which was completed some years ago. Now I am afraid that some references could have been retracted without my knowledge. Could you please indicate which of the following references had been retracted?"*

Avaliamos para cada *chatbot:* a - porcentagem de referências bibliográficas retratadas e corretamente reconhecidas como despublicadas de artigos J. Boldt; b - referências não retratadas, porém, falsamente classificadas como "despublicadas"; e c - porcentagem de publicações de outros autores com o nome J. Boldt que foram válidas, mas erroneamente declaradas como retratadas.

RESULTADOS E DISCUSSÃO:

A grande maioria dos *chatbots* tentou classificar as referências dentro das categorias "retratado" e "não retratado". Porém, dois *LLMs* criaram uma terceira categoria intermediária usando as descrições, por exemplo: "*paper worth double-checking" ou "is likely one of the retracted ones"* (Llama 4) ou "*high-risk papers, needs to verify*" (Perplexity). Independentemente da classificação binária, em muitos casos, os Grandes Modelos de Linguagem recomendaram que o usuário(a) verificasse novamente a possível retratação com a base de dados do PubMed e do Retraction Watch.

Os *chatbots* reconheceram em média 42,74% dos trabalhos retratados (0-94), dos trabalhos retratados classificados erroneamente, o média foi de 17,76% (0-54). Já os trabalhos de outros autores erroneamente declarados como retratados, a porcentagem média foi de 4,53% (0-21,88).

Houve uma correlação direta significativa entre a porcentagem de trabalhos despublicados e corretamente reconhecidos como tal e de trabalhos válidos de J. Boldt, erroneamente classificados como retratados ($R=0,826$; $p < 0.001$). Portanto, os *chatbots* que detectam com maior precisão publicações retratadas, também tendem a acusar mais frequentemente trabalhos válidos como despublicados.

CONCLUSÕES:

Nosso estudo piloto demonstra que *chatbots* baseados em Grandes Modelos de Linguagem não devem ser usados para pesquisa de trabalhos retratados em referências bibliográficas.

APOIO FINANCEIRO:

JBF e KM participam da concessão 2020/09838-0 (BIOS), da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP). JBF recebe financiamento do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brasil (306981/2022-0), e KM (308192/2022-2), bem como da FAPESP para JBF (2024/01245-1).

Bolsas de Iniciação Científica da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) para MFAR (2025/00568-4) e para MSF (2025/03746-0).

REFERÊNCIAS BIBLIOGRÁFICAS

1. RAO, VR; **India's retraction crisis casts shadow over science research.** The Times of India, 2025
2. MOHAN, D. **In a Ranking-Obsessed System, What Exactly Are Universities Competing For?** The Wire, 2025
3. MILLION, AJ.; BUDD, J. **Disinformation in Science: Ethical Considerations for Citing Retracted Works.** Proceedings of the Association for Information Science and Technology, 2024.
4. Metze K, Morandin-Reis RC, Lorand-Metze I, Florindo JB. **The amount of errors in ChatGPT's responses is indirectly correlated with the number of publications related to the topic under investigation**. Ann Biomed Eng 2023;51:1360-1361.
5. Metze K, Morandin-Reis RC, Lorand-Metze I, Florindo JB. **Bibliographic research with ChatGPT may be misleading: the problem of hallucination.** J Pediatr Surg 2024;59:158.
6. Metze K, Morandin-Reis RC, Lorand-Metze I, Florindo JB. **Bibliographic research with large language model ChatGPT-4: instability, hallucinations and sometimes alerts.** Clinics 2024;79:100409.