Are you sure the dataset you are using is real?

Growing Concern About Data Integrity in the Age of AI

Saurabh Zinjad
szinjad@asu.edu
linkedin.com/in/saurabhzinjad
github.com/Ztrimus

Agenda

- Introduction
- Medical Case Study
- Detection of Fakes
- Potential Misuse of AI in Research
- Conclusion - Promoting Data Vigilance

CHATGPT GENERATES FAKE DATA SET TO SUPPORT HYPOTHESIS

Fabricated database is convincing at a glance, but a close examination shows it doesn't pass as authentic.

By Miryam Naddaf

Researchers have used the technology behind the artificial intelligence (AI) chatbot ChatGPT to create a fake clinical-trial data set to support an unverified scientific claim.

In a paper published in *JAMA Ophthalmology* this month, the authors used GPT-4 — the latest version of the large language model on which ChatGPT runs — paired with Advanced Data Analysis (ADA), a model that incorporates the programming language Python and can perform statistical analysis and create data visualizations. The AI-generated data compared the outcomes of two surgical procedures and indicated — wrongly — that one treatment is better than the other (A. Taloni *et al. JAMA Ophthalmol.* https://doi.org/k58f; 2023).

"Our aim was to highlight that, in a few minutes, you can create a data set that is not supported by real original data, and it is also opposite or in the other direction compared to the evidence that are available," says study co-author Giuseppe Giannaccare, an eye surgeon at the University of Cagliari in Italy.

The ability of AI to fabricate convincing data adds to concern among researchers and journal editors about research integrity. "It was one thing that generative AI could be used to generate texts that would not be detectable using plagiarism software, but the capacity to create fake but realistic data sets is a next level of worry," says Elisabeth Bik, a microbiologist and independent research-integrity consultant in San Francisco, California. "It will make it very easy for any researcher or group of researchers to create fake measurements on non-existent patients, fake answers to questionnaires or to generate a large data set on animal experiments."

The authors describe the results as a

Introduction

Article?

Researchers utilized AI technology to create a fake clinical-trial data set.

Why?

AI's ability to fabricate data raises concerns about research integrity.

Problem?

Raise a critical question about the authenticity of the dataset being used for research or analysis.

Medical Case Study

Objective

Demonstrate AI's ability to fabricate plausible data. Source: Taloni et al, JAMA Ophthalmology

Disorder: Keratoconus

- Eye Condition
- Causes thinning of the cornea
- Can lead to impaired focus and poor vision

Data Generation Process

- Show statistical difference in cornea imaging test:
  - Cornea shape
  - Detecting irregularities
- Show impact through on participants' vision before and after procedures

Tools Used

- GPT-4
- Advanced Data Analysis (ADA)

Treatment: Corneal Transplant

- PK - Penetrating Keratoplasty
- DALK - Deep Anterior Lamellar Keratoplasty

Outcome

Desired - Fabricated data showing DALK superior to PK.

Observed - DALK were similar to those of PK for up to 2 years after the surgery

To an untrained eye, this certainly looks like a real data set.

But, If you closely scrutiny the dataset to check for authenticity...

Detection of Fakes

Unrealistic Relationships

Struggled to capture realistic relationships between variables between variables, such as the clinical parameters and the outcomes of different procedures.

Mismatch in Sex and Names

Mismatch in demographic data, like sex names and age, that did not align with the real patient population.

Lack of Correlation

Lack of correlation between certain crucial measures, like pre and post-surgical vision assessments versus the actual surgical outcomes.

Unusual Age Distribution

Age values ended disproportionately with 7 or 8.

Statistical inconsistencies in the fake data not supported by the actual medical understanding.

Potential Misuse of AI in Research

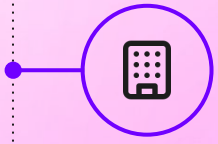
Risk of Scientific Fraud

- AI-generated fake data may support fraudulent scientific claims.
- Credibility of scientific data and findings undermined.
- Trustworthiness of research in various fields affected.

Threat to Peer Review Processes

The traditional peer review process may struggle to detect fabricated data, leading to the publication of false results in academic journals.

Conclusion - Promoting Data Vigilance



Key Takeaway

Understand the potential for AI to create fake data and the importance of vigilant data assessment.

Recommended Action

Promote a culture of data scrutiny and verification in all research and analytical processes.

Supporting Reliable Research

Emphasize the need for rigorous data collection, analysis, and verification to ensure research integrity.