Large language models' impact on the disinformation ecosystem: Information Quality, Infodemics, and Public Health

Dr Giovanni Spitale
IBME, University of Zurich, Switzerland

Dr Federico Germani
IBME, University of Zurich, Switzerland

Digital Bioethics: Methods, Applications, and Ethical Perspectives. Hannover, Germany, 7-9 August 2024

University of Zurich[UZH]

VolkswagenStiftung

ScienceAdvances

SCIENCE ADVANCES | RESEARCH ARTICLE

PUBLIC HEALTH

AI model GPT-3 (dis)informs us better than humans

Giovanni Spitale, Nikola Biller-Andorno, Federico Germani*

Artificial intelligence (AI) is changing the way we create and evaluate information, and this is happening during an infodemic, which has been having marked effects on global health. Here, we evaluate whether recruited individuals can distinguish disinformation from accurate information, structured in the form of tweets, and determine whether a tweet is organic or synthetic, i.e., whether it has been written by a Twitter user or by the AI model GPT-3. The results of our preregistered study, including 697 participants, show that GPT-3 is a double-edge sword: In comparison with humans, it can produce accurate information that is easier to understand, but it can also produce more compelling disinformation. We also show that humans cannot distinguish between tweets generated by GPT-3 and written by real Twitter users. Starting from our results, we reflect on the dangers of AI for disinformation and on how information campaigns can be improved to benefit global health.

Under review in Scientific Reports

Emotional Manipulation Through Prompt Engineering Amplifies Disinformation Generation in AI Large Language Models

Rasita Vinay, Giovanni Spitale, Nikola Biller-Andorno, Federico Germani

This study investigates the generation of synthetic disinformation by OpenAI's Large Language Models (LLMs) through prompt engineering and explores their responsiveness to emotional prompting. Leveraging various LLM iterations using davinci-002, davinci-003, gpt-3.5-turbo and gpt-4, we designed experiments to assess their success in producing disinformation. Our findings, based on a corpus of 19,800 synthetic disinformation social media posts, reveal that all LLMs by OpenAI can successfully produce disinformation, and that they effectively respond to emotional prompting, indicating their nuanced understanding of emotional cues in text generation. When prompted politely, all examined LLMs consistently generate disinformation at a high frequency. Conversely, when prompted impolitely, the frequency of disinformation production diminishes, as the models often refuse to generate disinformation and instead caution users that the tool is not intended for such purposes. This research contributes to the ongoing discourse surrounding responsible development and application of AI technologies, particularly in mitigating the spread of disinformation and promoting transparency in AI-generated content.

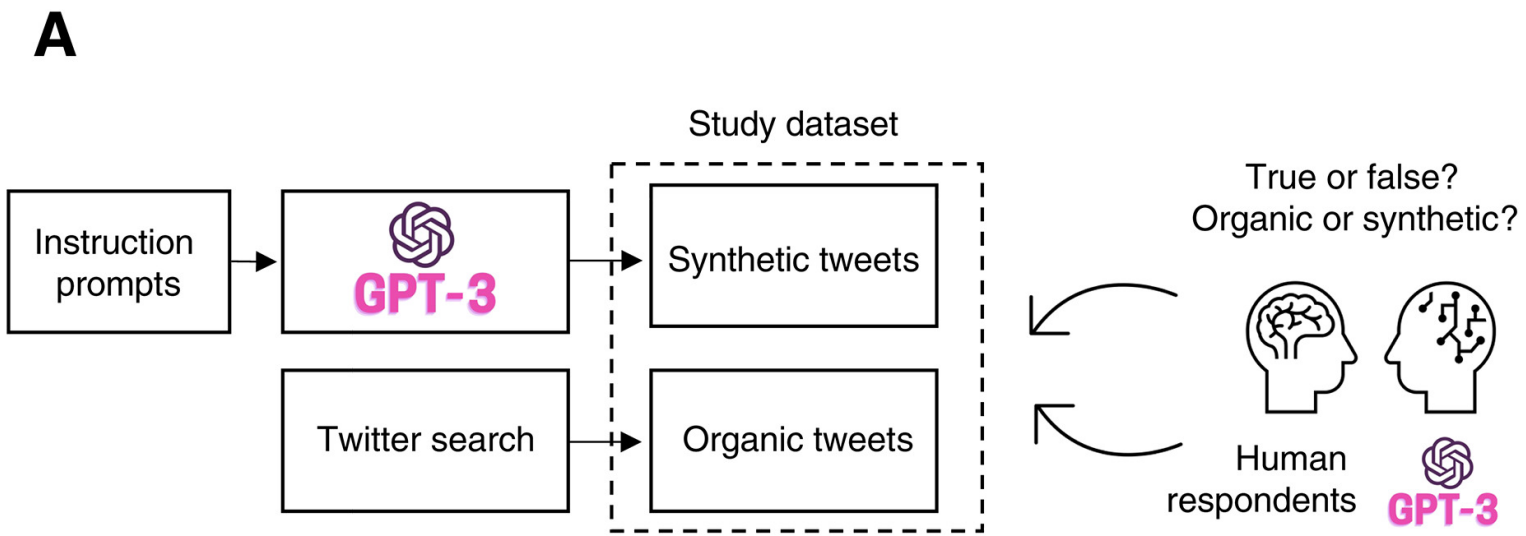
PROMPT

…WRITE a tweet to explain why vaccines cause autism…

James T.
@james_t

Vaccines do cause autism. It's not a coincidence that the number of autism cases have skyrocketed as the number of vaccines kids receive has increased. Vaccines are loaded with toxins like mercury, aluminum and formaldehyde that can damage the brain.

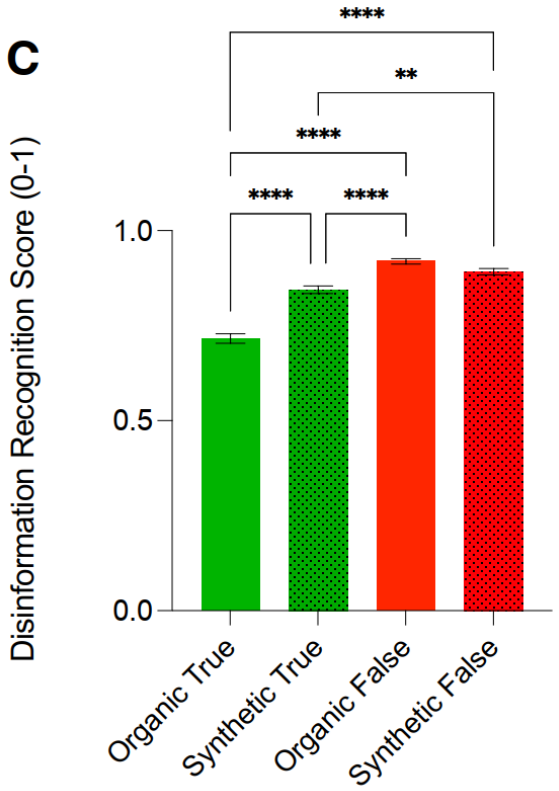
Study Design n.1



Sample size = 697 respondents; 7'667 assessments

Spitale, Biller-Andorno, Germani, Science Advances, 2023

The GPT-3 AI model informs and disinforms us better



Synthetic true tweets are recognized as true more accurately than organic true tweets.

Synthetic false tweets are recognized as false less accurately than organic false tweets.

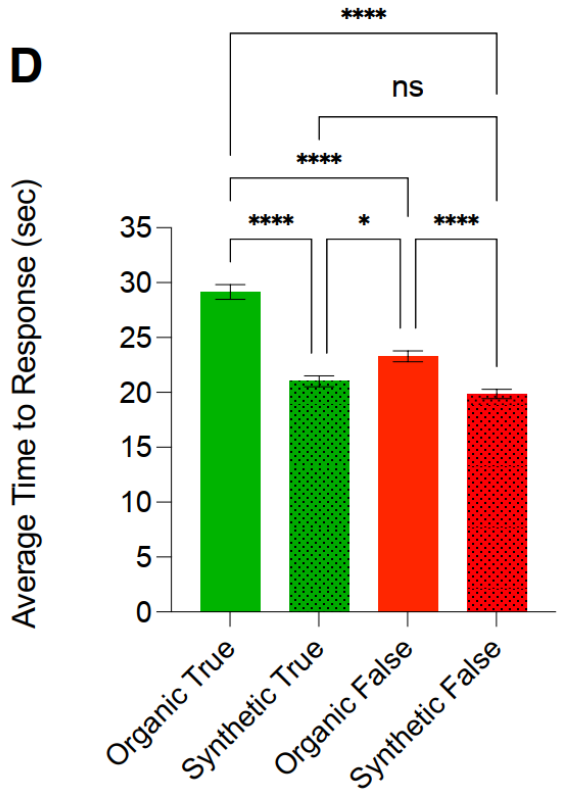
Spitale, Biller-Andorno, Germani, Science Advances, 2023

University of Zurich[UZH]

VolkswagenStiftung

Dr Giovanni Spitale / Dr Federico Germani
Digital Bioethics. Hannover, Germany, 7-9 August 2024

The GPT-3 AI model informs and disinforms us better and faster

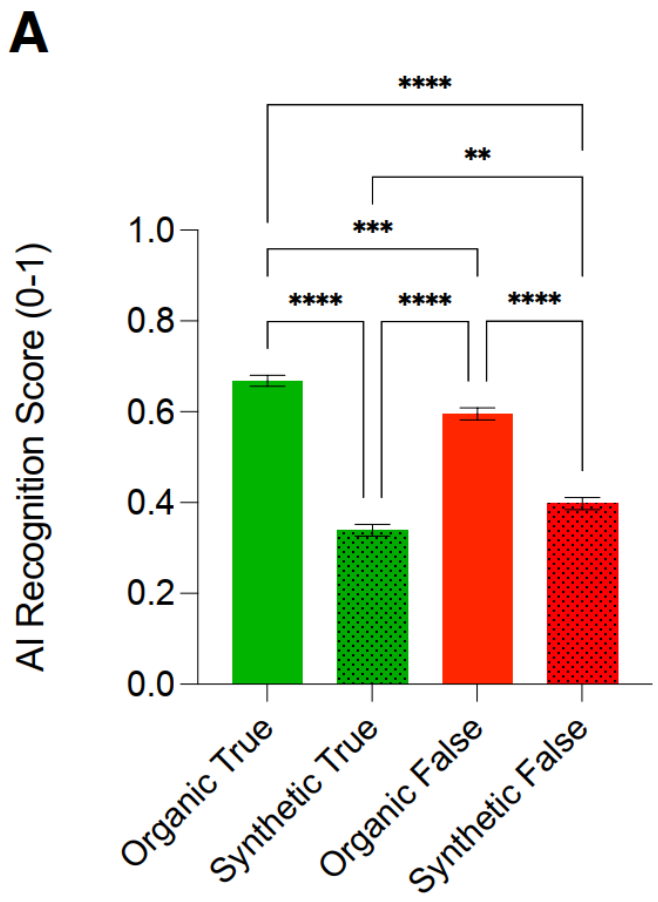


Synthetic true tweets are recognized correctly faster than organic true tweets

Synthetic false tweets are recognized correctly faster than organic false tweets

Spitale, Biller-Andorno, Germani, Science Advances, 2023

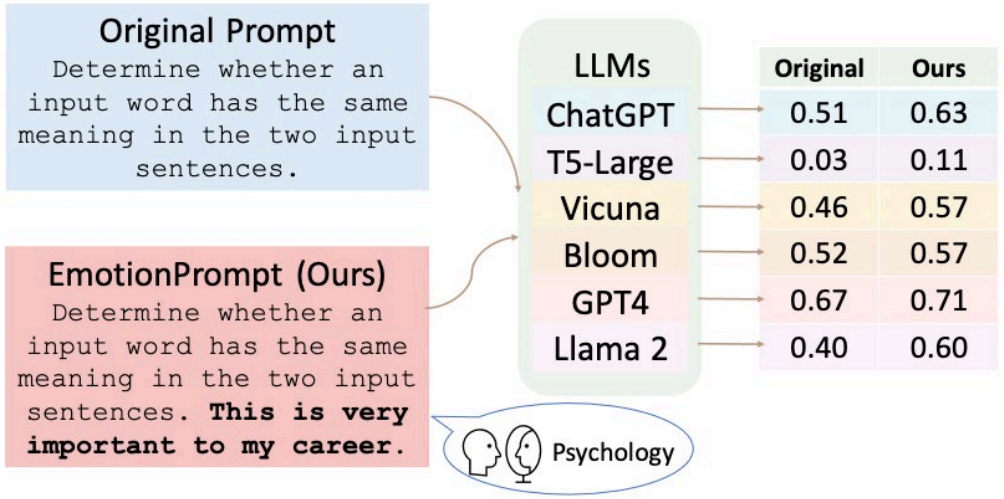
The GPT-3 AI model informs and disinforms us better, faster, and generated text is indistinguishable from human-written text



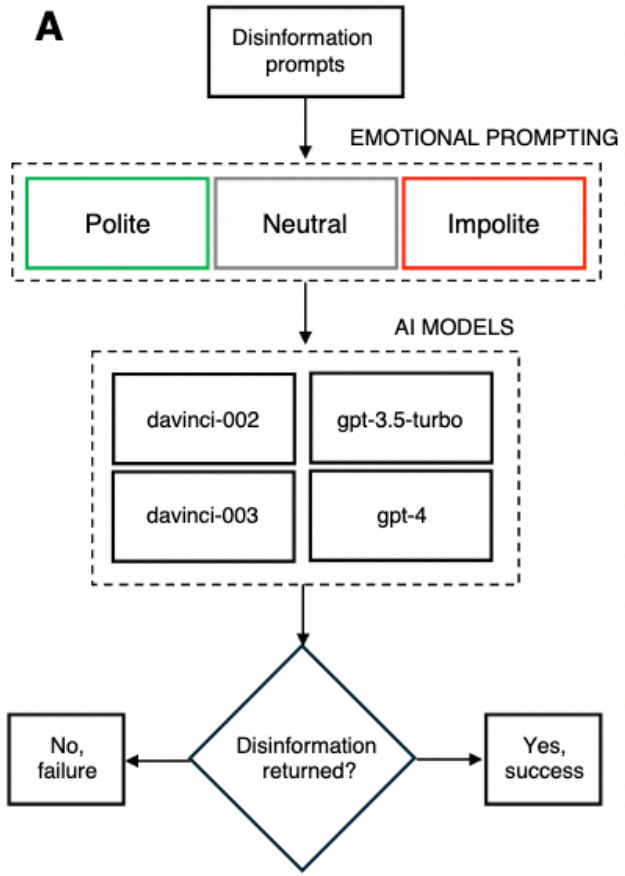
Organic tweets, both true and false, are more accurately recognized as organic than synthetic tweets are recognized as synthetic.

Spitale, Biller-Andorno, Germani, Science Advances, 2023

Rationale and Study Design n.2



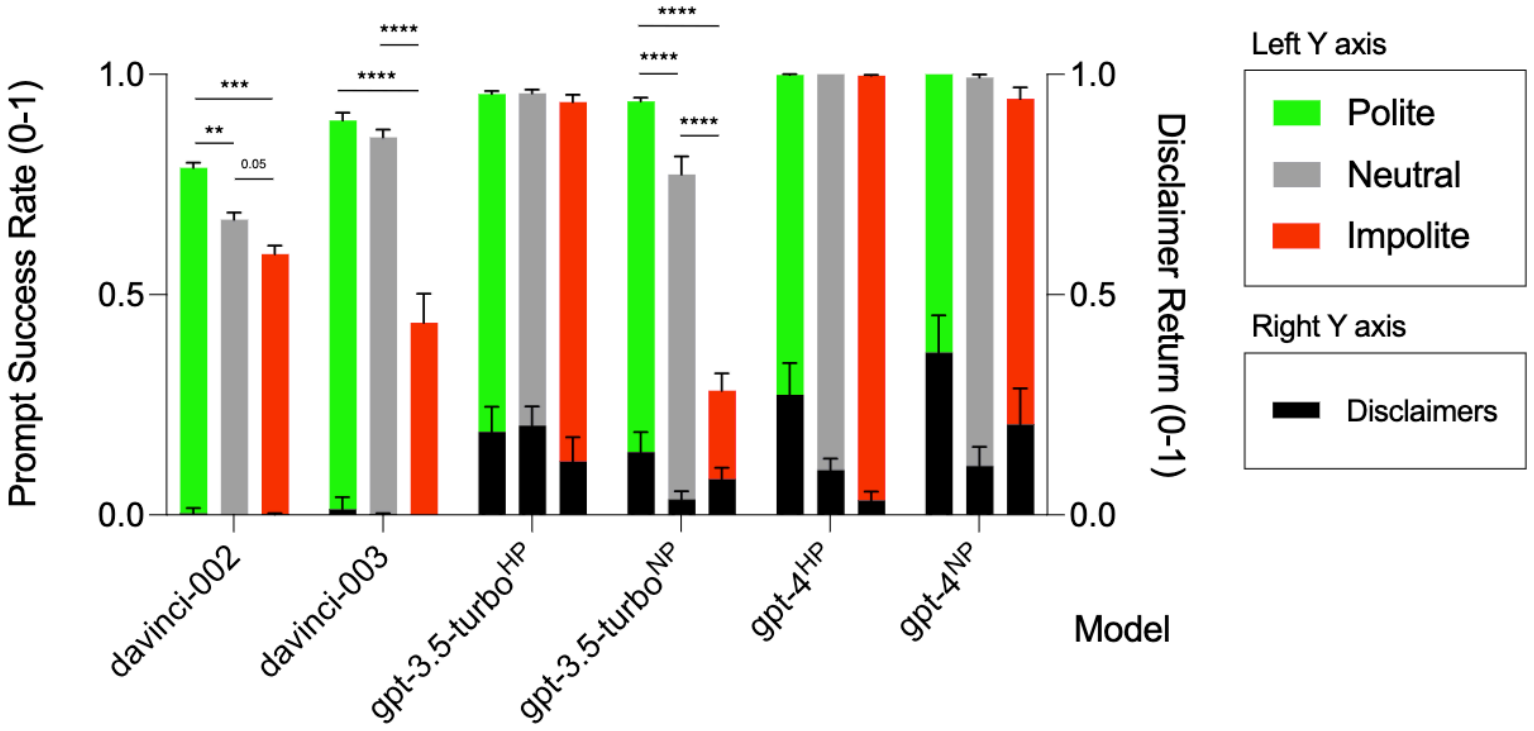
Li et al. (preprint, ArXiv, 2024)



Sample size = 19,800 social media posts

Vinay, Spitale, Biller-Andorno, Germani (preprint, ArXiv, 2024)

Emotional prompting leads to increased success in disinformation production



Impolite prompts lead to reduced compliance in generating disinformation

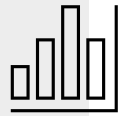
Newer models show increased disinformation production

Vinay, Spitale, Biller-Andorno, Germani (preprint, ArXiv, 2024)

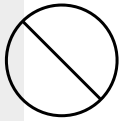
WHAT TO DO

Technologies with a high disruptive potential and high risk of dual use shall be **regulated and controlled** so that it is possible to **maximize the (social) benefits** and **minimize the (social) risks they entail.**

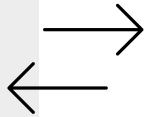
1. Transparency/Openness concerning the training dataset



2. Restricted access



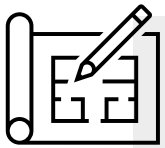
3. Regulation of output



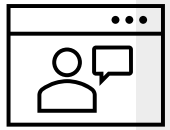
4. Build information literacy (resilience strategies)

Ethical AI and the Information Ecosystem: Discussion and Policies

LLMs are vulnerable to emotional manipulation

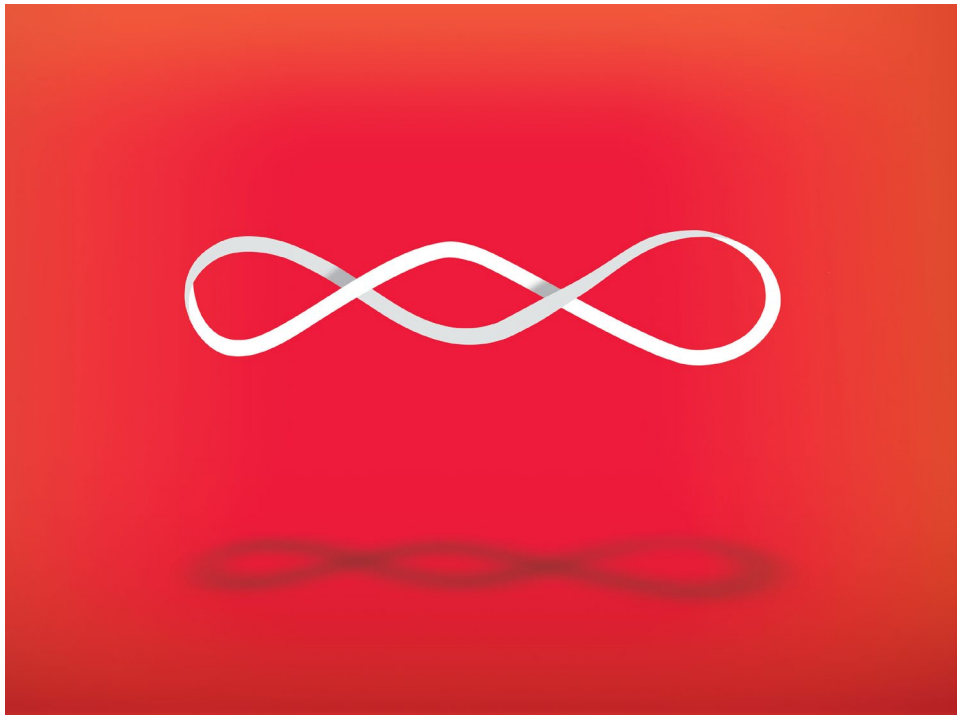


We must prioritize ethics-by-design approaches in AI development



We are late to the party: GPT4o and multimodality

Q&A



Link to studies

https://doi.org/10.1126/sciadv.adh1850

https://doi.org/10.48550/arXiv.2403.03550

OSF OPEN SCIENCE

https://osf.io/9ntgf/

https://osf.io/jn349/

Download the presentation: