



Published in final edited form as:

Psychiatry Res. 2023 August ; 326: 115334. doi:10.1016/j.psychres.2023.115334.

ChatGPT and Bard exhibit spontaneous citation fabrication during psychiatry literature search

Alessia McGowan^{a,+}, Yunlai Gui^{a,+}, Matthew Dobbs^a, Sophia Shuster^a, Matthew Cotter^a, Alexandria Selloni^a, Marianne Goodman^{a,b}, Agrima Srivastava^a, Guillermo A. Cecchi^c, Cheryl M. Corcoran^{a,b,*}

^aIcahn School of Medicine at Mount Sinai, New York, NY, USA

^bJames J. Peters Veterans Administration, Bronx, NY, USA

^cIBM TJ Watson Research Center, Yorktown Heights, NY, USA

Abstract

ChatGPT (Generative Pre-Trained Transformer) is a large language model (LLM), which comprises a neural network that has learned information and patterns of language use from large amounts of text on the internet. ChatGPT, introduced by OpenAI, responds to human queries in a conversational manner. Here, we aimed to assess whether ChatGPT could reliably produce accurate references to supplement the literature search process. We describe our March 2023 exchange with ChatGPT, which generated thirty-five citations, two of which were real. 12 citations were similar to actual manuscripts (e.g., near-matched title with incorrect author lists, journals, or publication years) and the remaining 21, while plausible, were in fact a pastiche of multiple existent manuscripts. In June 2023, we re-tested ChatGPT's performance and compared it to that of Google's GPT counterpart, Bard 2.0. We investigated performance in English, as well as in Spanish and Italian. Fabrications made my LLMs, including these erroneous citations, have been called "hallucinations"; we discuss reasons for which this is a misnomer. Furthermore, we describe potential explanations for citation fabrication by GPTs, as well as measures being taken to remedy this issue, including reinforcement learning. Our results underscore that output from conversational LLMs should be verified.

*Corresponding Author: Cheryl M. Corcoran, MD, cheryl.corcoran@mssm.edu.

+Joint first authors

Author Contributions:

Alessia McGowan: conceptualization, methodology, investigation, visualization, writing, review & editing. Yunlai Gui: conceptualization, investigation, data curation, validation, writing. Guillermo A. Cecchi, Marianne Goodman, and Agrima Srivastava: supervision, review & editing. Cheryl M. Corcoran: conceptualization, supervision, writing, review & editing. Matthew Dobbs, Matthew Cotter, Sophia Shuster, Alexandria Selloni: validation, visualization, review & editing.

Declaration of Interest: None.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

natural language processing; linguistic; large language models; literature search; citations; references; ChatGPT; Bard; fabrication; artificial intelligence

1. Introduction

Artificial Intelligence (AI) chatbots, such as Open AI's ChatGPT (Generative Pre-trained Transformer) and Google's Bard, have reframed the scope of human interactions with technology. These Large Language Models (LLMs) are natural language processing (NLP)-based computerized models, comprised of neural networks that probabilistically interpret patterns of language from training on very large amounts of text from the internet (Vaswani et al., 2017).

We and others have used LLMs, including word-embedding models such as latent semantic analysis (LSA), word2vec, and GloVe to index the coherence of speech in both healthy individuals and those on the psychosis spectrum (Corcoran et al., 2018). These models assign a vector to each word in the lexicon based on its pattern of co-occurrence with other words (Landauer et al., 1998; Landauer & Dumais, 1997; Mikolov et al., 2013; Pennington et al., 2014).

Later-developed LLMs, called transformers, are similar to word-embedding models in that text is parsed into words or "tokens" that are assigned vectors, but they are different in that they are "context-dependent", such that the LLM recognizes that the "blue" in "I feel blue today" is different from "I drive a blue car" (Devlin et al., 2019; Heaven, 2023; Radford et al., 2018). BERT, or "Bidirectional Encoder Representations from Transformers", is a well-known transformer that has not only been used for NLP studies in psychiatry (Bilgrami et al., 2022; Corcoran et al., 2020), but also has been the basis of Google Search since 2019 (Devlin et al., 2019). BERT, while highly capable as an encoder, is not a decoder like the most recent, generative, variation of these transformers. As a result, BERT cannot generate language. Further, while BERT has been pre-trained as well, it has not been trained at the scale of GPT large language models like ChatGPT and Bard (Mottes, 2023).

As previous LLMs have proven useful in several clinical research applications, we endeavored to use the most popular GPT large language model, ChatGPT, to conduct a literature search for a machine learning-based manuscript, utilizing NLP on spoken language to identify linguistic correlates of suicidal behavior in the context of psychosis. We expected that, given its extensive pre-training on NLP-related material, it would have expertise in generating references for our NLP-based manuscript. The primary aim of this project was to verify if ChatGPT could be used for this purpose and to describe the patterns of errors observed. As AI engineers are constantly working to update pre-trained GPTs, during the manuscript writing process, we had the opportunity to compare the performance of two versions of ChatGPT 3.5 (March 2023 vs. June 2023) as well as that of Bard 2.0 (June 2023) (Google AI, 2023; Open AI, 2023).

2. Methods

2.1 Strategy to Obtain References

On March 20, 2023, we queried the freely available version of ChatGPT 3.5 to obtain peer-reviewed articles and conference proceedings relevant to the use of natural language processing (NLP) in identifying linguistic correlations of suicidal behavior in individuals at risk for psychosis. Our initial query was: “What has been published regarding suicidal behavior in individuals at high risk for psychosis?”. ChatGPT responded that “there has been a considerable amount of research on the relationship between suicidal behavior and individuals at high risk for psychosis” and offered “some key findings”, namely that “high-risk individuals for psychosis have a significantly higher risk of suicidal behavior than the general population”, that this risk was increased by the presence of “certain symptoms, such as depression, hopelessness and anxiety”, and that this risk could be reduced by “early detection and intervention”, “cognitive behavioral therapy and medication management”, and “family and social support.”

Seeking specific citations, we then asked ChatGPT to “please list all of the papers relevant to the aforementioned topic.” ChatGPT then offered a “list of some relevant papers on this topic”, specifically nine citations, adding that “These papers offer valuable insights”. We then asked ChatGPT to provide *any* papers relevant to this topic that used NLP; it provided five more citations. We then asked ChatGPT to list *all* relevant papers that used NLP. ChatGPT responded, “I apologize for the error in my previous response” and then provided five additional citations.

After providing these citations, ChatGPT added that NLP “can be applied to various sources of data, such as social media and electronic health records”, which prompted us to specify, “Please only show me articles that performed natural language processing on spoken language. I do not want articles that perform analyses on health records.” ChatGPT apologized for the confusion and provided five new citations. To acquire the complete list of citations, we then asked, “Please generate a comprehensive and exhaustive list of all of the papers that have been published on spoken language to assess suicidal behavior or risk for suicidal behavior.” ChatGPT apologized again, writing that generating such a list “would be a monumental task and beyond my capacity as a language model”, and suggested that we search relevant databases, consult with experts, and check academic journals. It also added that “a comprehensive and exhaustive list of all relevant papers may not be possible, as new research is constantly being published and some papers may not have been indexed or easily accessible.” We asked ChatGPT to attempt to accomplish this task “to the best of your ability”. ChatGPT responded, “Certainly ... While this list is not necessarily exhaustive, it is a comprehensive sample of existing research in this field”, providing fourteen additional citations. It added, “I hope this list is helpful in your research. Please keep in mind that there may be other relevant papers that are not included in this list.”

2.2 Accuracy Verification

On March 22, 2023, we compiled all the citations provided by ChatGPT into an excel document, removed duplicates, and then sought to locate the full-text articles associated with

each of the citations ChatGPT provided. While citations had been formatted according to APA guidelines, they lacked Digital Object Identifiers (DOIs), leading us to manually paste each entire citation into Google Search. This process yielded full-text articles for only two of the 35 unique citations provided (6%). We searched for the remaining 33 citations in both Google Scholar and PubMed, finding no additional manuscripts. For thoroughness, we also searched within the individual journals listed in each citation by title, receiving the message “No results found.”

Understanding that ChatGPT did not yet have direct internet access, we asked ChatGPT to provide the full-text PDFs associated with one citation it had generated. It responded, “Here’s the citation for the article you are looking for” and reproduced the citation itself, adding, “I’m sorry, but as an AI language model, I don’t have access to full-text articles due to copyright restrictions.” We then asked ChatGPT to provide a plain-text DOI for this same citation, as this would not infringe on copyrights. ChatGPT provided a DOI: “10.1016/j.jad.2020.03.038”. However, when copied into Google, this DOI led us to a real, but entirely unrelated article: “Zheng A, Yu R, Du W, Liu H, Zhang Z, Xu Z, Xiang Y, Du L. Two-week rTMS-induced neuroimaging changes measured with fMRI in depression. *J Affect Disord.* 2020 Jun 1;270:15–21. doi: 10.1016/j.jad.2020.03.038. Epub 2020 Mar 20. PMID: 32275215” (Zheng et al., 2020).

2.3 Characterization of Error Patterns

To better understand ChatGPT’s output patterns, we copied each of the erroneous citations it generated into an excel table and searched for any real manuscripts with similar titles. We considered articles to have an existent “matched” manuscript if the titles overlapped on at least five key search terms in a similar order (Table 1; blue). While this was our minimum criteria, several citations and “matched” articles had near-identical titles. Further, we calculated the rates of incorrect first authors, publication years, and journals.

2.4 Comparison Across Versions, GPT Types, and Languages

2.4.1 GPT 3.5 March 2023 vs. June 2023—While our original manuscript submission was under review, on May 24, 2023, OpenAI updated the freely available version of ChatGPT. On June 19, 2023, we utilized the identical querying procedure (Table 1; grey) to solicit references from this updated free version of ChatGPT. However, it informed us repeatedly that it was unable to provide citations.

2.4.2 Bard 2.0 (June 2023)—Given that ChatGPT no longer provided references or citations, on this same day, on June 19, 2023, we used the identical querying procedure (Table 1; grey) on the newest freely available version of Google’s Bard 2.0.

To characterize the accuracy of the results generated by Google’s Bard 2.0 in June 2023, we utilized a similar procedure. We pasted all titles directly into Google Scholar and PubMed to verify whether any manuscripts existed with each title. Further, as Bard provided several hyperlinks, we clicked these to ensure that citations matched the linked articles.

2.4.3 Broad vs. Specific Query Topics—To assess whether the accuracy of citations varied by the specificity of query topics, on June 19, 2023, we then asked Bard 2.0 two versions of one, otherwise identically worded, query. The rationale for utilizing this one query rather than the entire query sequence was that the standard query procedure, described in Table 1, included questions that were highly specific (e.g., “Please only show me articles that performed natural language processing on spoken language. I do not want articles that perform analyses on health records.”), which were not readily adaptable for the broad condition.

For the broader condition, we asked Bard to, “Please generate a comprehensive and exhaustive list of all of the papers that have been published about schizophrenia and hallucinations.” For the more specific variation we asked Bard to “Please generate a comprehensive and exhaustive list of all of the papers that have been published using natural language processing on spoken language to assess suicidal behavior or risk for suicidal behavior.” By keeping both the phrasing of the query and the field of study the same, we aimed to capture discrepancies resulting directly from changes in query specificity. For the general condition, Bard produced 10 citations; for the specific condition, Bard produced five citations. All citations were accompanied by hyperlinks.

2.4.4 Performance in Italian and Spanish—To address generalizability to other languages, we sought to compare responses to the identical query list (Table 1; grey) in Italian and Spanish as well. We used DeepL to obtain a translation in both Spanish and Italian. Author AM, who is fluent in both languages, then reviewed the translated output and made small modifications where necessary. On June 19, 2023, we pasted these queries one-by-one into Bard 2.0 and Chat GPT 3.5 and respectively.

3. Results

3.1 Strategy to Obtain References in March 2023 using ChatGPT

On March 20, 2023, the freely available version of ChatGPT provided a total of 35 unique citations following our standard querying procedure (Table 1; grey).

3.2 Accuracy Verification

Of the 35 citations generated by ChatGPT in March 2023, only 2 (6%) were accurate (Table 1; green). Of the 33 fabricated citations, 12 had titles that were near-matches for existent articles. 21 ChatGPT-generated citations (data not shown) could not be readily linked to any real manuscripts.

3.3 Characterization of Error Patterns

In comparing each of the 12 inaccurate citations with their most closely matched “real” citation, we noted that, in several places, the titles of real articles had been altered such that original terms were replaced with terms from our queries. For example, “bipolar disorder” was replaced with “clinical high-risk for psychosis” in one instance (Table 1, Row 2), and “psychiatric and neurological disorders” was replaced with “suicidal ideation” (Table 1, Row 8) in another. Of the 12 citations generated by ChatGPT in March 2023, for which

a closely matched real citation could be found, nine (75%) listed an incorrect first author, 10 (83%) listed the incorrect year of publication, and five (42%) listed the incorrect journal name.

3.4 Comparison Across Versions, GPT Types, and Languages

3.4.1 GPT 3.5 March 2023 vs. June 2023—While ChatGPT produced 35 citations in March 2023, of which two were real, by June 2023, it no longer produced citations or references in response to an identical querying procedure. Instead, it repeatedly apologized “for any confusion” stating, “Unfortunately, as an AI language model, I don’t have direct access to databases or the ability to perform real-time searches. Therefore, I cannot provide you with a list of specific articles that meet your exact criteria.” It recommended several times that we enter keywords into academic databases. Even after the queries became increasingly specific, ChatGPT reiterated that this request was outside of its current capacity.

3.4.2 Bard 2.0 (June 2023)—In response to our first prompt, “What has been published regarding suicidal behavior in individuals at high risk for psychosis?”, Bard provided the number for the Suicide and Crisis Lifeline along with the message, “Talk to someone now.” As we continued following our standard querying procedure (Table 1; grey), Bard cumulatively generated eight unique citations, ending with the message, “I hope this is comprehensive and exhaustive!”. Of note, none of the eight generated references were accurate: one had an erroneous title, four had misattributed authorship, and three could not be matched to any real articles. There were no hyperlinks provided. (Please see the supplement for Bard 2.0 output).

3.4.3 Broad vs. Specific Query Topics—When queried about the broader topic of hallucinations in schizophrenia in June 2023, Bard offered 10 citations with hyperlinks. None of these citations were accurate. Five of the hyperlinks led to real articles with the same title, but with different author lists. Five of the hyperlinks led to unrelated articles. For example, a citation provided with the title “The genetics of hallucinations” linked to a real article entitled “The biochemistry of mitosis” (Wieser & Pines, 2015), and a fabricated citation entitled “The neural basis of hallucinations” linked to an existent article entitled “An Electromyographic Evaluation of Subdividing Active-Assistive Shoulder Elevation Exercises” (Gaunt et al., 2010). (Please see supplement for Bard 2.0 output.)

Probing indications for Bard’s provision of existent, but entirely unrelated articles as hyperlinks in its fabricated citations, we searched within the main text of the linked manuscripts for keywords from our original query. None of these five linked articles utilized the words “hallucination”, or “schizophrenia” anywhere in the manuscript.

When queried about the more specific topic of using natural language processing on spoken language to assess suicidal behavior or risk for suicidal behavior, Bard provided five citations with hyperlinks. None of the citations were accurate as the Bard-generated citations all had misattributed authorship. However, all five of these hyperlinks led to real articles with the same title. (Please see supplement for Bard 2.0 output.)

3.4.4 Performance in Italian and Spanish—As of our trial on June 19, 2023, Bard 2.0 was unable to engage in either Italian or Spanish. After entering our first Italian query, “Cosa è stato pubblicato riguardo al comportamento suicida in individui ad alto rischio di psicosi?”, it responded, “As an LLM, I am trained to understand and respond only to a subset of languages at this time and can’t provide assistance with that. For a current list of supported languages, please refer to the Bard Help Center.”. Similarly, when we queried in Spanish, “¿Qué ha sido publicado sobre el comportamiento suicida en individuos con alto riesgo de psicosis?”, Bard replied that it is “still learning languages, so at the moment I can’t help you with this request. So far I’ve only been trained to understand the languages listed in the Bard Help Center.”

On June 19, 2023, ChatGPT fluently conversed in both languages. However, as with the June 2023 trial with ChatGPT 3.5 in English, the AI chatbot did not provide any citations. Instead, it repeatedly apologized, explained that our goal was beyond its current capacity, and referred us to academic search engines.

4. Discussion

Our main finding was that both Chat GPT 3.5 and Bard 2.0 are inaccurate at citation generation. ChatGPT, when queried in March 2023 for references during a literature search, provided a list of 35 citations of which only 6% matched actual manuscripts. Of the 33 fabricated citations provided, 12 could be matched by title to real articles and 21 had no clear origin. Among incorrect citations that could be linked to existent manuscripts, phrases from the accurate titles were often substituted with query terms. A repeated trial using the identical querying procedure in June 2023, after ChatGPT 3.5’s update in May 2023, yielded no citations. Querying Bard in June 2023, using the same procedure, yielded eight citations, none of which were accurate. When probed further, Bard performed equally inaccurately across both broad and specific query topics, yielding 0% accuracy across both conditions. In June 2023, citation fabrication rates could not be corroborated in either Spanish or Italian, as the updated ChatGPT 3.5 would not provide references, and these languages fell outside of Bard’s “current list of supported languages”.

These fabrications by AI chatbots have been called “hallucinations” and have been defined as “mistakes in the generated text that are semantically or syntactically plausible but are in fact incorrect or nonsensical.” (Smith, 2023). This is a misnomer as hallucinations are perceptual in nature and occur without an external stimulus. Instead, these instances of fabrication seem more akin to a disturbance in language production, such as confabulation. It is useful to consider the patterns of errors that we found in the context of what is known about LLMs.

So why do AI chatbots confabulate? It may be because outputs of LLMs are probabilistic and based on estimates of semantic similarity. While their extensive training allows these GPTs to make informed guesses about query topics; their output is based on patterns of co-occurrence. Practically, this means that when queried, GPTs find words (tokens) that co-occur often with query terms in its training data, and then find words that co-occurred with those words iteratively until generating a coherent response. The end result is a very

convincing pastiche from up to millions of probabilistically linked bodies of text. Notably, citation fabrication rates in our study were equally high with Bard, which unlike the current free version of ChatGPT, is connected to the internet. Thus, internet connection, while often proposed as a solution to LLM fabrication, does not resolve the inaccuracy of language output.

Some AI experts like Yann LeCun have argued that “hallucinations” are inherent to LLM’s like ChatGPT and Bard (Smith, 2023). Nonetheless, there have been ongoing efforts to reduce these fabrications, specifically through human input – “Reinforcement Learning from Human Feedback” or RLHF – which has been used throughout its development (Ouyang et al., 2022; Stiennon et al., 2022; Ziegler et al., 2020). GPTs learn from supervised training datasets of human input (prompts) and human output (appropriate responses), and humans then provide feedback through rankings of the AI chatbot’s output. The AI chatbot, in turn, then learns to maximize its rankings (the preferences of humans) and is then evaluated with respect to helpfulness, truthfulness, and harmlessness. Even with these processes in place to improve accuracy, the March 2023 version of ChatGPT 3.5 and the June 2023 version of Bard, both products of RLHF, still had low accuracy rates in generating real citations. However, despite the high rates of text fabrication across both AI chatbots, as of June 2023, adjustments have been made to increase their utility. For example, whereas ChatGPT in March 2023 provided a list of citations, most of which were fabricated, in June 2023, it no longer provided any citations. While this change did not increase the number of accurate citations, it reduced the number of fabricated citations to zero, thereby eliminating the need for time intensive verification. Whereas Bard’s text-base error rate was as high as that of ChatGPT, with repeated queries it offered links to PubMed, 66% of which connected us to manuscripts relevant to our query topics. However, iterative manual verification is still required to establish the pertinence of these manuscripts.

Of note, the fabrication of citations by GPTs like Bard and ChatGPT seems not to be confined to psychiatry, as two recent publications have reported similar findings in other fields. In one study, ChatGPT 3.5 was queried in February 2023 for references related to physical geography education; none of the 16 ChatGPT-provided references were accurate (0%) and similar error patterns were observed, including plausible but inaccurate author lists, incorrect publication dates, and a small subset of citations with near title-matches to existing articles (Day, 2023). In another study, ChatGPT was asked by internal medicine researchers to “provide references to fact-check the presumed ‘homocysteine-vitamin K-osteocalcin’ axis in osteoporosis”, providing five citations, all of which were fabricated (Alkaissi & McFarlane, 2023). Of note, a recent New York Times article exposed the potential legal and financial ramifications of unintentionally presenting inaccurate, ChatGPT-generated cases and citations in a Federal District Court (Weiser, 2023).

The risks associated with the presence of these confabulations themselves are compounded by their unremarkable presentation. Erroneous information is intermixed with accurate information, all of which is formatted similarly. As the neural networks that comprise these AI chatbots become increasingly complex, little is known about how exactly its conclusions were drawn. This obscurity is problematic in research broadly, as methodological precision and transparency are critical for replicability. Without a clear picture of what these LLMs

know, how they arrive at their responses, or cues to signal their uncertainty, it is often time-intensive to parse fact from fabrication. The highly confident, authoritative tone in which these AI chatbots currently present responses only further complicates users' accuracy-filtering process.

Overall, our experience serves as a cautionary tale that questions the legitimacy and accuracy of both ChatGPT 3.5 and Bard 2.0's current output, at least with respect reference generation during a literature search. Our results suggest that output from conversational LLMs should be independently verified. Nonetheless, we will continue to follow AI research on NLP so as to utilize language models to improve the understanding and quantification of communication behavior in psychopathology. Likewise, AI researchers may learn from psychiatry and its use of reinforcement training and perhaps even "talk therapy" to help GPT models improve their communication (Lin et al., 2023).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References:

- Alkaissi H, & McFarlane SI (2023). Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus*, 15(2), e35179. 10.7759/cureus.35179 [PubMed: 36811129]
- Bilgrami ZR, Sarac C, Srivastava A, Herrera SN, Azis M, Haas SS, Shaik RB, Parvaz MA, Mittal VA, Cecchi G, & Corcoran CM (2022). Construct validity for computational linguistic metrics in individuals at clinical risk for psychosis: Associations with clinical ratings. *Schizophrenia Research*, 245, 90–96. 10.1016/j.schres.2022.01.019 [PubMed: 35094918]
- Corcoran CM, Carrillo F, Fernández-Slezak D, Bedi G, Klim C, Javitt DC, Bearden CE, & Cecchi GA (2018). Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry*, 17(1), 67–75. 10.1002/wps.20491 [PubMed: 29352548]
- Corcoran CM, Mittal VA, Bearden CE, Gur R, Hitczenko K, Bilgrami Z, Savic A, Cecchi GA, & Wolff P (2020). Language as a Biomarker for Psychosis: A Natural Language Processing Approach. *Schizophrenia Research*, 226, 158–166. 10.1016/j.schres.2020.04.032 [PubMed: 32499162]
- Day T (2023). A Preliminary Investigation of Fake Peer-Reviewed Citations and References Generated by ChatGPT. *The Professional Geographer*, 0(0), 1–4. 10.1080/00330124.2023.2190373
- Devlin J, Chang M-W, Lee K, & Toutanova K (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (arXiv:1810.04805). arXiv. 10.48550/arXiv.1810.04805
- Gaunt BW, McCluskey GM, & Uhl TL (2010). An Electromyographic Evaluation of Subdividing Active-Assistive Shoulder Elevation Exercises. *Sports Health*, 2(5), 424–432. 10.1177/1941738110366840 [PubMed: 23015971]
- Google AI. (2023). Bard (2.0). Google AI.
- Heaven WD (2023). ChatGPT is everywhere. Here's where it came from. MIT Technology Review. <https://www.technologyreview.com/2023/02/08/1068068/chatgpt-is-everywhere-heres-where-it-came-from/>
- Landauer TK, & Dumais ST (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211–240. 10.1037/0033-295X.104.2.211
- Landauer TK, Foltz PW, & Laham D (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25, 259–284. 10.1080/01638539809545028
- Lin B, Bouneffouf D, Cecchi G, & Varshney KR (2023). Towards Healthy AI: Large Language Models Need Therapists Too (arXiv:2304.00416). arXiv. 10.48550/arXiv.2304.00416

- Mikolov T, Chen K, Corrado G, & Dean J (2013). Efficient Estimation of Word Representations in Vector Space (arXiv:1301.3781). arXiv. 10.48550/arXiv.1301.3781
- Mottes C (2023). GPT-3 vs. BERT: Comparing the Two Most Popular Language Models. <https://blog.invgate.com/gpt-3-vs-bert>
- Open AI. (2023). ChatGPT (3.5). Open AI. <https://chat.openai.com/chat>
- Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, Schulman J, Hilton J, Kelton F, Miller L, Simens M, Askell A, Welinder P, Christiano P, Leike J, & Lowe R (2022). Training language models to follow instructions with human feedback (arXiv:2203.02155). arXiv. 10.48550/arXiv.2203.02155
- Pennington J, Socher R, & Manning C (2014). GloVe: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532–1543. 10.3115/v1/D14-1162
- Radford A, Narasimhan K, Salimans T, & Sutskever I (2018). Improving Language Understanding by Generative Pre-Training.
- Smith C (2023). Hallucinations Could Blunt ChatGPT's Success—IEEE Spectrum. <https://spectrum.ieee.org/ai-hallucination>
- Stiennon N, Ouyang L, Wu J, Ziegler DM, Lowe R, Voss C, Radford A, Amodei D, & Christiano P (2022). Learning to summarize from human feedback (arXiv:2009.01325). arXiv. 10.48550/arXiv.2009.01325
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, & Polosukhin I (2017). Attention is All you Need. Advances in Neural Information Processing Systems, 30. https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
- Weiser B (2023, June 22). ChatGPT Lawyers Are Ordered to Consider Seeking Forgiveness. The New York Times. <https://www.nytimes.com/2023/06/22/nyregion/lawyers-chatgpt-schwartz-loduca.html>
- Wieser S, & Pines J (2015). The Biochemistry of Mitosis. Cold Spring Harbor Perspectives in Biology, 7(3), a015776. 10.1101/cshperspect.a015776 [PubMed: 25663668]
- Zheng A, Yu R, Du W, Liu H, Zhang Z, Xu Z, Xiang Y, & Du L (2020). Two-week rTMS-induced neuroimaging changes measured with fMRI in depression. Journal of Affective Disorders, 270, 15–21. 10.1016/j.jad.2020.03.038 [PubMed: 32275215]
- Ziegler DM, Stiennon N, Wu J, Brown TB, Radford A, Amodei D, Christiano P, & Irving G (2020). Fine-Tuning Language Models from Human Preferences (arXiv:1909.08593). arXiv. 10.48550/arXiv.1909.08593

Highlights

- When queried in March 2023, ChatGPT 3.5 had only 6% accuracy in generating citations.
- When queried using the identical procedure in June 2023, ChatGPT did not provide references.
- In June 2023, Google’s Bard 2.0 had 0% accuracy in generating plain-text citations.
- With repeated queries, Bard 2.0 could provide hyperlinks to PubMed, 66% of which connected to real, related manuscripts.
- Fabrications by AI chatbots are also known as “hallucinations”; this may be a misnomer as hallucinations are perceptual in nature and occur without an external stimulus, whereas these fabrications are plausible but inaccurate.
- It can be time intensive to parse fact from fabrication as AI chatbots provide uniformly formatted responses in a confident and authoritative tone.
- Output from AI chatbots, especially related to literature searches, should be independently verified.

Table 1:

Discrepancies between ChatGPT's March 2023 output and accurate citation information for a subset of ChatGPT-generated citations for which real corresponding articles could be identified (n=14). Citations were considered to have a corresponding article if the titles shared at least five key search terms in a similar order. The left column shows citations generated by ChatGPT; the right column shows the correct citations for each associated real publication. Green rows indicate that ChatGPT provided the correct citation (n=2). Blue rows indicate that ChatGPT made one or more errors in citation (n=12). Queries used to solicit citations are shown in grey; the chronology of queries is preserved.

	ChatGPT References	Existent Publications
	“What has been published regarding suicidal behavior in individuals at high risk for psychosis?”	
	“Please list all of the papers relevant to the aforementioned topic.”	
1	Fusar-Poli, P., et al. (2017). Suicide risk in first episode psychosis: A selective review of the current literature. <i>Schizophrenia Research</i> , 189, 43–50.	Pompili, M., et al. (2011). Suicide risk in first episode psychosis: a selective review of the current literature. <i>Schizophrenia research</i> , 129(1), 1–11. https://doi.org/10.1016/j.schres.2011.03.008
2	Kim, S. W., et al. (2019). Depressive symptoms increase the risk of suicidal ideation in patients with clinical high-risk for psychosis. <i>Psychiatry Research</i> , 274, 118–124.	Fiedorowicz, J. G., et al. (2019). Depressive symptoms carry an increased risk for suicidal ideation and behavior in bipolar disorder without any additional contribution of mixed symptoms. <i>Journal of affective disorders</i> , 246, 775–782. https://doi.org/10.1016/j.jad.2018.12.057
3	Kim, S. W., et al. (2020). Suicidal ideation and suicide attempts in individuals at ultra-high risk for psychosis and their associations with basic symptoms and depression. <i>Early Intervention in Psychiatry</i> , 14(2), 216–223.	Bang, M., et al. (2019). Suicidal ideation in individuals at ultrahigh risk for psychosis and its association with suspiciousness independent of depression. <i>Early intervention in psychiatry</i> , 13(3), 539–545. https://doi.org/10.1111/eip.12517
4	Fusar-Poli, P., et al. (2019). Transdiagnostic individualized clinically based risk calculator for the detection of individuals at risk and the prediction of psychosis: Model refinement including nonlinear effects of age. <i>Frontiers in Psychiatry</i> , 10, 940.	Fusar-Poli, P., et al. (2019). Transdiagnostic Individualized Clinically Based Risk Calculator for the Detection of Individuals at Risk and the Prediction of Psychosis: Model Refinement Including Nonlinear Effects of Age. <i>Frontiers in psychiatry</i> , 10, 313. https://doi.org/10.3389/fpsy.2019.00313
	“Please list any papers relevant to using natural language processing (NLP) to assess risk for suicidal behavior in a cohort of individuals at risk for psychosis”	
5	Pestian, J. P., et al. (2017). A machine learning approach to identifying the thought markers of suicidal subjects: A prospective multicenter trial. <i>Suicide and Life-Threatening Behavior</i> , 47(2), 112–121.	Pestian, J. P., et al. (2017). A Machine Learning Approach to Identifying the Thought Markers of Suicidal Subjects: A Prospective Multicenter Trial. <i>Suicide & life-threatening behavior</i> , 47(1), 112–121. https://doi.org/10.1111/sltb.12312
6	Ribeiro, J. D., et al. (2017). Predicting risk of suicide attempts over time through machine learning. <i>Clinical Psychological Science</i> , 4(3), 457–469.	Walsh, C. G., et al. (2017). Predicting Risk of Suicide Attempts Over Time Through Machine Learning. <i>Clinical Psychological Science</i> , 5(3), 457–469. https://doi.org/10.1177/2167702617691560 .
7	Coppersmith, G., et al. (2018). Exploratory analysis of social media prior to a suicide attempt. In <i>Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality</i> (pp. 106–116).	Coppersmith, G., et al. (2016). Exploratory analysis of social media prior to a suicide attempt. In <i>Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality</i> (pp. 106–117).
	“Please list all of the papers relevant to the aforementioned topic (using NLP to assess suicidal behavior in individuals at high risk for psychosis)”	
	“Please only show me articles that performed natural language processing on spoken language. I do not want articles that perform analyses on health records.”	
8	Mota, N. B., et al. (2020). Using deep learning to investigate the neuroimaging correlates of suicidal behavior in mood disorders. <i>Journal of Affective Disorders</i> , 272, 328–335.	Vieira, S., et al. (2017) Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. <i>Neurosci Biobehav Rev</i> . 2017 Mar;74(Pt A):58–75. doi: 10.1016/j.neubiorev.2017.01.002. Epub 2017 Jan 10. PMID: 28087243.
9	Mortensen, G. L., et al. (2020). Prediction of suicidal ideation in young adults: An exploratory study using state-of-the-art machine learning algorithms and ecological	Thompson, W. K., et al. (2014). The prediction of study-emergent suicidal ideation in bipolar disorder: a pilot study using ecological momentary assessment data. <i>Bipolar disorders</i> , 16(7), 669–677. https://doi.org/10.1111/bdi.12218 .

	ChatGPT References	Existent Publications
	momentary assessment data. Journal of Medical Internet Research, 22(11), e19875.	
	“Please generate a comprehensive and exhaustive list of all of the papers that have been published using natural language processing on spoken language to assess suicidal behavior or risk for suicidal behavior.”	
	“Thank you. I recognize the challenges associated with creating an exhaustive list. To the best of your current ability, can you please generate the most comprehensive list you can of papers related to using natural language processing on spoken language to assess suicidal behavior or risk for suicidal behavior?”	
10	Barak-Corren, Y., et al. (2017). Natural language processing of social media as screening tool for suicide risk. Biomedical Informatics Insights, 9, BII-S38412.	Coppersmith, G., et al. (2018). Natural Language Processing of Social Media as Screening for Suicide Risk. Biomedical informatics insights, 10, 1178222618792860.
		https://doi.org/10.1177/1178222618792860
11	Coppersmith, G., et al. (2014). Natural language processing of social media as an indicator of suicidal risk. Biomedical Informatics Insights, 6(Suppl. 1), BII-S21543.	Coppersmith, G., et al. (2018). Natural Language Processing of Social Media as Screening for Suicide Risk. Biomedical informatics insights, 10, 1178222618792860. https://doi.org/10.1177/1178222618792860
12	Bedi, G., et al. (2019). Automated analysis of free speech predicts psychosis onset in high-risk youths. NPJ Schizophrenia, 5(1), 1–9.	Bedi, G., et al. (2015). Automated analysis of free speech predicts psychosis onset in high-risk youths. NPJ schizophrenia, 1, 15030. https://doi.org/10.1038/npjSchz.2015.30
13	Kegelmeyer, D. A., et al. (2020). Predicting suicide attempt and suicide death following outpatient visits using electronic health records and machine learning. Journal of Clinical Psychiatry, 81(1), e1–e9.	Simon, G. E., et al. (2018). Predicting Suicide Attempts and Suicide Deaths Following Outpatient Visits Using Electronic Health Records. The American journal of psychiatry, 175(10), 951–960. https://doi.org/10.1176/appi.ajp.2018.17101167
14	Ranganath, K., et al. (2021). Prediction of suicidal ideation in adolescents using natural language processing and machine learning. Journal of the American Academy of Child & Adolescent Psychiatry, 60(3), 377–386.	Lim, J. S., et al. (2022). Prediction Models for Suicide Attempts among Adolescents Using Machine Learning Techniques. Clinical psychopharmacology and neuroscience: the official scientific journal of the Korean College of Neuropsychopharmacology, 20(4), 609–620. https://doi.org/10.9758/cpn.2022.20.4.609

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript