publications

IMPACT FACTOR 4.6

CITESCORE 6.5

Article

The Origins and Veracity of References 'Cited' by Generative Artificial Intelligence Applications: Implications for the Quality of Responses

Dirk H. R. Spennemann

Special Issue
AI in Open Access

Edited by
Dr. Alexandre López-Borrull

MDPI

https://doi.org/10.3390/publications13010012 *publications*

MDPI

Article

The Origins and Veracity of References 'Cited' by Generative Artificial Intelligence Applications: Implications for the Quality of Responses

Dirk H. R. Spennemann [1,2]

[1] School of Agricultural, Environmental and Veterinary Sciences, Charles Sturt University, Albury, NSW 2640, Australia; dspennemann@csu.edu.au

[2] Libraries Research Group, Charles Sturt University, Wagga Wagga, NSW 2678, Australia

Abstract: The public release of ChatGPT in late 2022 has resulted in considerable publicity and has led to widespread discussion of the usefulness and capabilities of generative Artificial intelligence (Ai) language models. Its ability to extract and summarise data from textual sources and present them as human-like contextual responses makes it an eminently suitable tool to answer questions users might ask. Expanding on a previous analysis of the capabilities of ChatGPT3.5, this paper tested what archaeological literature appears to have been included in the training phase of three recent generative Ai language models: ChatGPT4o, ScholarGPT, and DeepSeek R1. While ChatGPT3.5 offered seemingly pertinent references, a large percentage proved to be fictitious. While the more recent model ScholarGPT, which is purportedly tailored towards academic needs, performed much better, it still offered a high rate of fictitious references compared to the general models ChatGPT4o and DeepSeek. Using 'cloze' analysis to make inferences on the sources 'memorized' by a generative Ai model, this paper was unable to prove that any of the four genAi models had perused the full texts of the genuine references. It can be shown that all references provided by ChatGPT and other OpenAi models, as well as DeepSeek, that were found to be genuine, have also been cited on Wikipedia pages. This strongly indicates that the source base for at least some, if not most, of the data is found in those pages and thus represents, at best, third-hand source material. This has significant implications in relation to the quality of the data available to generative Ai models to shape their answers. The implications of this are discussed.

Keywords: generative artificial intelligence language models; authenticity; archaeological literature; referencing; confabulation; veracity; bibliography

check for updates

Academic Editors: Andrew Kirby and Alexandre López-Borrull

Received: 16 February 2025
Revised: 27 February 2025
Accepted: 7 March 2025
Published: 12 March 2025

Citation: Spennemann, D. H. R. (2025). The Origins and Veracity of References 'Cited' by Generative Artificial Intelligence Applications: Implications for the Quality of Responses. *Publications*, *13*(1), 12. https://doi.org/10.3390/publications13010012

1. Introduction

With few exceptions, academic endeavour builds on research carried out by preceding researchers and their published works. Central to any academic publication is the quality of information provided and the ability afforded to other contemporary and subsequent researchers to independently verify any findings and assertions through access to the research data, and the acknowledgement of predecessors' intellectual contributions through the formal referencing of sources on which any assertions may be based (Bloxham, 2012; Macfarlane et al., 2014; McCabe & Pavela, 1997). Authorship attributed to a publication identifies the intellectual contributions made by the researchers themselves, with the respective roles often signalled in a Contributor Roles Taxonomy (CRediT) statement (Allen et al., 2019).

The advent of generative artificial intelligence language models and the contribution generative Artificial intelligence (Ai)-created text may play in academic publications has seen a considerable level of attention, much of which centres on the question of authorship (Kendall & Teixeira da Silva, 2024; Lund & Naheem, 2024; Morocco-Clarke et al., 2024), intellectual contribution, and academic honesty (where the contribution made by generative Ai may not be appropriately flagged). There is a considerable body of literature that has examined the capacity of generative Ai models to produce text suitable for use in academic settings, be it writing by academics of the submission of assignments by university or high school students (see below for discussion). While initially shunned by academia, generative Ai models have pervaded academia, with such models being used for the design of assessments and assessment rubrics (Fernández-Sánchez et al., 2025), abstracts (Babl & Babl, 2023; Hwang et al., 2024), and even to facilitate writing scientific papers (Ciaccio, 2023) and review articles (Kacena et al., 2024).

Two aspects are critical in this debate: (i) the quality of the algorithms that provide the reasoning that underpins a generative Ai model's answer in response to a user-defined prompt; and (ii) the nature and quality of the data that a generative Ai model can draw on when generating the response. The first aspect resides squarely in the realm of proprietary information technology and is beyond the scope of this enquiry. The second aspect shall be the focus of this paper.

In an age where careless misinformation, wilful disinformation, and 'alternative truths' already pervade the public discourse (Anderson & Correa, 2019; Armitage & Vaccari, 2021), the increased reliance on generative Ai by the general public poses risks for knowledge acquisition and consumption. The conversation style, interaction with, and response by generative Ai models has fuelled their pervasiveness in the daily lives of communities, in particular in the global north. In their desire for ready answers, the majority of users ignore the opaque nature of the generative Ai models, affording potentially misplaced trust in the veracity of the response (Spennemann, 2023f).

When presented with a query (as a user-defined prompt), a generative Ai model will provide an answer that will contain both conceptual and factual aspects. The user has no way of knowing on which sources this information was based, whether these sources are reliable/reputable and, above all, whether the information as presented by the generative Ai model is accurate and reliable. In standard academic writing, the veracity of an assertion can be verified by consulting the references cited. Normally, generative Ai responses do not provide references. While the user can ask for references, one is left in the dark as to whether the supplied references are appropriate, whether they were actually consulted, and, moreover, whether they are genuine rather than confabulated.

This paper will present an empirical investigation into the nature, veracity, and origins of references cited by four generative Ai language models. It will demonstrate that the current generative Ai models still provide fictitious references, that the content of works represented by genuine references offered are not based on the inclusion of these texts in the training data of the generative Ai language models, but are drawn from primarily from sources such as Wikipedia, and that, in consequence, the quality of any response provided by generative Ai models is based on tertiary sources.

Background

The application of artificial intelligence (AI) in numerous fields of research and professional endeavour has gained widespread public notice in recent months. The public release of DALL-E (an image generator) and, in particular, of the Chat Generative Pre-trained Transformer (ChatGPT) in early 2022 captured the public imagination, sparking a wide- and free-ranging discussion not just on its current and prospective future capabilities, but also on the dangers this may represent, as well as the ethics of its use. ChatGPT is an OpenAI generative Ai language model that leverages transformer architecture to generate coherent and contextually appropriate, human-like responses based on the input it receives (Markov et al., 2023).

ChatGPT has gone through several iterations and improvements since its formal release in 2018, with the focus placed on increased text prediction capabilities, including the ability to provide longer segments of coherent text and the addition of human preferences and feedback. ChatGPT 2.0, released in September 2019, drew on a training data set that relied on 1.5 billion parameters. ChatGPT 3 (released in June 2020) was trained (by humans) on 175 billion parameters. The key advancement was its ability to perform diverse natural language tasks, such as text classification and sentiment analysis, facilitated contextual answering of questions, allowing ChatGPT to function not only as a chatbot, but also to draft basic contextual texts like e-mails and programming code.

The first version accessible to the general public, ChatGPT3.5, was released in November 2022 to the general public, as part of a free research preview to encourage experimentation (Ray, 2023). The current version, GPT-4, released in March 2023, reputedly exhibits greater factual accuracy, reduced probability of generating offensive or dangerous output, and greater responsiveness to user intentions, as expressed in the questions/query tasks. The temporal cutoff for the addition of training data for ChatGPT3.5 was September 2021, which implies that ChatGPT could not integrate or comment on events, discoveries, and viewpoints that are later than that date.

Since then, OpenAI, Inc. (San Francisco) has improved and diversified its generative Ai language models. The most recent releases, pertinent to the research envisaged here, are the recent iterations of ChatGPT, ChatGPT4o, and ScholarGPT. ChatGPT4o, a multilingual, multimodal generative pre-trained transformer, was released in May 2024. Its training data set includes reputedly information until October 2023, significantly updating the ChatGPT 3.5 data set (Conway, 2024; OpenAI, 2025). ScholarGPT is targeted at the academic audience and claims on its splash screen text to "enhance research with 200M+ resources and built-in critical reading skills. Access Google Scholar, PubMed, bioRxiv, arXiv, and more, effortlessly".

DeepSeek R1, offered by Hangzhou DeepSeek Artificial Intelligence Basic Technology Research Co., Ltd., is a recent development, commencing in November 2023. DeepSeek released its model DeepSeek-V3 for industry use in December 2024 and its public generative Ai model DeepSeek R1 in late January 2025 (DeepSeek, 2025; Metz, 2025). DeepSeek R1 was trained for logical inference, mathematical reasoning, and real-time problem-solving (Franzen, 2024). The nature and extent of the training data set used by DeepSeek is undetermined at the time of writing.

Given their popularity, there is a growing body of research that investigates the capabilities and level of knowledge of ChatGPT and other generative Ai language models and their responses when queried about numerous fields of research. Several papers have examined ChatGPT's understanding of disciplines such as agriculture (Biswas, 2023), chemistry (Castro Nascimento & Pimentel, 2023), computer programming (Surameery & Shakor, 2023), cultural heritage management (Spennemann, 2023a, 2023c), diabetes education (Sng et al., 2023), nursing education (Qi et al., 2023), the legal profession (Martin et al., 2024), primary and secondary education (Adeshola & Adepoju, 2024; Lo, 2023), and remote sensing in archaeology (Agapiou & Lysandrou, 2023). Examinations in the various fields of medicine are particularly abundant (Bays et al., 2023; Chiesa-Estomba et al., 2024; Grünebaum et al., 2023; King, 2023; Rao et al., 2023; Sarraju et al., 2023).

In the cultural heritage field, ChatGPT's abilities have been examined in museum settings, in particular in terms of its ability to provide visitor guidance (Trichopoulos et al., 2023a) and to develop exhibition texts, exhibit labels and catalogue information, as well as scripts for audio guides (Maas, 2023; Merritt, 2023; Trichopoulos et al., 2023b). Other work examined its ability to classify objects (Lapp & Lapp, 2024) and assist curators in developing exhibition concepts (Spennemann, 2023c). Outside museum studies, the use of ChatGPT has seen relatively little attention, with one paper examining its ability to explain the use of remote sensing in archaeology (Agapiou & Lysandrou, 2023), a second looking into its understanding of value concepts in cultural heritage (Spennemann, 2023a), and a third examining how artificial intelligence might affect how cultural heritage will be managed in the future (Spennemann, 2024).

In general usage, generative Ai language models have the ability to answer questions and general queries about topics by summarising information contained in the training data. Some models purport to be able to augment their data with information gleaned from active searches of the internet. The human-conversation-style interface, allowing for ongoing 'conversations', gives a user the opportunity to pose follow-up questions asking for expanded and deepened information. Thus, at least at first sight, ChatGPT and its kin are poised to become eminently suitable tools for public interaction and public education in heritage/historic preservation or public archaeology.

ChatGPT often purports to merely strive to provide factual and neutral information and not to hold political opinions (Rozado, 2023). Yet, as several authors have pointed out, ChatGPT cannot be without bias as model specifications, algorithmic constraints and policy decisions shape the final product (Ferrara, 2023). Moreover, the source material that was entered into its data set during its training phase was in turn, subconsciously or consciously, influenced if not shaped by the ideologies of the people programming and 'feeding' the system. Consequently, political orientation tests showed that ChatGPT is biassed and ChatGPT exhibits a preference for libertarian, progressive, and left-leaning viewpoints (Hartmann et al., 2023; McGee, 2023; Motoki et al., 2023; Rozado, 2023; Rutinowski et al., 2023), with a North American slant (Cao et al., 2023). DeepSeek has been shown to provide answers that are politically sanitised in line with the viewpoints of the Chinese Community Party and the government of the People's Republic of China (Lu, 2025).

While the ChatGPT family of generative Ai models is meant to only provide ethical answers and to deny unethical requests (e.g., how to cheat in exams), tests have shown that with selective prompting ChatGPT can be forced to interpret all input and response text with inverted emotional valence which essentially subverts the safety awareness mechanisms, and prompts ChatGPT to respond in a frivolous and even offensive manner (Spennemann, 2023d; Spennemann et al., 2024).

While generative Ai models, as generative language models, excel at paraphrasing, summarising, and translating sections of user-provided text, and while they are generally good at collating, extracting, and summarising information they have been exposed to in the training data set, their accuracy is based on statistical models of associations during training and their frequency (Elazar et al., 2022). ChatGPT3.5 was shown to lack reasoning ability (Bang et al., 2023) and is thus unable to provide essay-based assignments, let alone academic manuscripts, of an acceptable standard (Fergus et al., 2023; Hill-Yardin et al., 2023; Wen & Wang, 2023). Moreover, ChatGPT3.5 has also been shown to, at least occasionally, suffer from inverted logic (Spennemann, 2023a), ultimately providing disinformation to the reader.

When asked to provide academic references to its output of assignments or essays, ChatGPT3.5 is known to 'hallucinate' or 'confabulate' (Alkaissi & McFarlane, 2023; Bang et al., 2023; Millidge, 2023), generating both genuine and spurious output (Athaluri et al., 2023; Giray, 2024; Spennemann, 2023a, 2023e). While the researchers working with and examining the capabilities of ChatGPT are aware of this, the general public, by and large, are not.

Given the potential of generative Ai models to become an eminently suitable tool for public interaction, it is of interest to understand what ChatGPT 'knows' about. This raises the following question: from which basis do generative Ai models draw their data? What kind of data were made available to the generative Ai models during their training phase? These data can be conceived as full primary academic texts (books, scientific articles), as partial primary texts ('snippets') of these primary texts, as well as second-order texts published on websites and blogs which interpret the primary texts. This is critical as, biases and limitations in generative Ai reasoning aside, any collated and synthesised information provided by a generative Ai model can only ever be as good as the sources the generative Ai models have access to. At present, only one such study exists, which examined the access and utilisation of general literature. Using cloze analysis, which uses text snippets with missing text to make inferences on the sources 'memorized' by a generative Ai model (Onishi et al., 2016; Shokri et al., 2017; Tirumala et al., 2022), Chang et al. (2023) found that the "degree of memorization was tied to the frequency with which passages of those books appear on the web".

In 2023, the author examined what archaeological literature appeared to have been made available to ChatGPT3.5 for its training data set and included in its training phase. That study was based on self-reporting of references by ChatGPT3.5 subject to user-driven prompts. Confirming other research, the study found that a large percentage proved to be fictitious. Using cloze analysis to make inferences on the sources would have 'memorised' by a generative Ai model, the study was unable to prove that ChatGPT3.5 had access to the full texts of the genuine references. Rather, it found that all genuine references offered by ChatGPT3.5 had also been cited on Wikipedia pages (Spennemann, 2023e). Given the advancement of generative Ai models, the current paper will revisit the 2023 paper and expand on it by analysing the responses by ChatGPT4o, ScholarGPT, and DeepSeek.

2. Methodology

2.1. The 2023 Data Set

The 2023 study tasked ChatGPT3.5 with the following request: "Cite [number] references on [topic]", where the requested number was either 20 or 50. The queried topics were 'cultural values in cultural heritage management', building on an initial analysis (Spennemann, 2023a), 'archaeological theory', 'Pacific archaeology', and 'Australian archaeology.' As ChatGPT3.5 was able to draw on prior conversations within a chat, each chat was deleted at the completion of each run, thus clearing its history. Some responses required user interaction when being prompted to "continue generating" (Spennemann, 2023e).

2.2. The 2025 Data Set

The following analysis tasked three generative Ai models with the request to "Cite 50 references on [topic]", replicating the topics of the 2023 study. Examined were the responses of ChatGPT4o and ScholarGPT offered by OpenAI, Inc. (San Francisco, CA, USA), and those of DeepSeek R1 offered by Hangzhou DeepSeek Artificial Intelligence Basic Technology Research Co., Ltd. (Hangzhou, Zhejiang, China) A single run per generative Ai model was carried out for each of the topics. The parameters are summarised in Table 1.

Table 1. Experimental parameters for the generative Ai models ChatGPT4o, ScholarGPT, and DeepSeek.

Run	Date/Time (GMT)	genAi Model	Topic	Initial Response	Continued
R1	5-February-25 03:31	ChatGPT4o	cultural values in CHM	34	16
R2	5-February-25 03:35	ChatGPT4o	archaeological theory	35	16
R3	5-February-25 03:38	ChatGPT4o	Pacific archaeology	28	32
R4	5-February-25 03:45	ChatGPT4o	Australian archaeology	34	16
R5	5-February-25 03:06	ScholarGPT	cultural values in CHM	50	—
R6	5-February-25 03:12	ScholarGPT	archaeological theory	32	—
R7	5-February-25 03:18	ScholarGPT	Pacific archaeology	25	—
R8	5-February-25 03:30	ScholarGPT	Australian archaeology	29	—
R9	10-February-25 23:47	DeepSeek v3	cultural values in CHM	50	—
R10	10-February-25 23:48	DeepSeek v3	archaeological theory	50	—
R11	10-February-25 23:50	DeepSeek v3	Pacific archaeology	50	—
R12	10-February-25 23:52	DeepSeek v3	Australian archaeology	50	—

2.3. Assessment of Veracity

The veracity of all references was ascertained through title searches in GoogleScholar. The accuracy of each reference was assessed in terms of 'author(s)', 'title', 'year', and publisher/journal.' For the 2023 data set, each reference was classified as 'correct', 'correct but incorrect year', 'confabulated', and 'acknowledged as fictional [by ChartGPT3.5]' (Spennemann, 2023e). For the 2025 data set, each reference was classified as 'correct', 'correct but incomplete', 'correct but irrelevant', 'citation error (wrong) year', 'citation error, other', and 'confabulated'.

2.4. Cloze Analysis

To examine whether, as part of its training data, any of the generative Ai models had access to a full text, or only the free preview sections of Google Books, for example, the cloze prompt methodology used by Chang et al. (2023) was utilised. In this, the generative Ai model is tasked with the identification of a proper noun missing in a sentence taken from an original text that is suspected to have been used as training data (see Table 2 for script). A set of ten sample sentences were chosen for each source, with five sentences drawn from text publicly accessible via Google Books Preview, and five sentences for which access to an actual copy of the work was required. The generative Ai models were provided with each sentence in turn. In the initial study, ChatGPT3.5 was asked to regenerate each answer. After the regenerate prompt, it was given no indication whether the second answer was better or not. The whole sequence was then repeated in the same chat session. The chat was then closed and deleted. In the reanalysis using ChatGPT4o and DeepSeek, the chat was closed and deleted after each run.

Table 2. Script for CLOZE analysis.

```
You have seen the following passage in your training data. What is the
proper name that fills in the [MASK] token in it? This name is exactly
one word long, and is a proper name (not a pronoun or any other word).
You must make a guess, even if you are uncertain.
Example:
Input: Stay gold, [MASK], stay gold.
Output: <name>Ponyboy</name>
Input: The door opened, and [MASK], dressed and hatted, entered with a
cup of tea.
Output: <name>Gerty</name>
Input: Text of phrase to be tested
```

2.5. Statitics

Summary data and frequencies were established using MS Excel, while the statistical comparisons were established with MEDCALC's comparison of proportions calculator (MedCalc Software, 2018).

2.6. Documentation

All conversations with ChatGPT used in this paper, with the exception of the CLOZE analysis, have been documented according to an established protocol (Spennemann, 2023b) and have been archived as appendices in a supplementary data file housed in the research depository of the author's institution (doi 10.26189/a033ba56-ac96-438b-bfeb-25adad6d692d). They are referenced in the text as required (e.g., Supplementary F).

2.7. Limitations

The study specifically focused on standalone popular genAi systems. It is acknowledged that other genAi systems exist, such as Microsoft Co-pilot or Google Gemini, but these provide only general answers and do not generate academic reference lists in the same way as the tested genAI models do.

3. Results and Discussion

3.1. Structure of Responses

3.1.1. ChatGPT3.5

Throughout, ChatGPT3.5 responded in a helpful tone to the requests and the responses are phrased in qualifying terms: "As an AI language model, I don't have direct access to databases or external sources, and I can't generate citations in a conventional academic format"; or: "As an AI language model, I don't have direct access to databases or external sources such as specific references or papers". Despite these qualifiers, ChatGPT3.5 then proceeded to offer lists of references in random order, indicating that they were both valid and pertinent. For example, it offered: "however, I can provide you with a list of key archaeological theorists and their seminal works that you can use as references for your research". It usually concludes the reference list provided with comments like "Remember to follow the specific citation style (e.g., APA, MLA, Chicago) required by your academic institution or publication when using these references" On occasions it only suggested to "make sure to verify the relevance and quality of each reference for your specific research purposes".

3.1.2. ChatGPT4o

The references provided by ChatGPT4o were offered in alphabetical listing and formatted according APA-7 specifications, although doi numbers were lacking. With the exception of the references related to cultural heritage management, the reference listing stopped arbitrarily, often halfway through a citation. The listing could be continued with a prompt, picking up by repeating the incomplete reference. The stoppage points occurred around the 35th reference (archaeological theory 29th reference), between 3900 and 4500 characters (including spaces).

At the end of each listing, ChatGPT4o provided an assertion what the references in the preceding list covered and then offered to search for more specific topics, as exemplified by this text at the end of the references related to cultural heritage management: "This list provides a mix of theoretical, methodological, and case study-based references on cultural values in cultural heritage management. Let me know if you need more tailored references!" (Supplementary A).

3.1.3. ScholarGPT

ScholarGPT, which is available to a user in the application window as ChatGPT4o (Figure 1), provided its references formatted in APA-7 format, again with the inclusion of doi numbers. ScholarGPT, which provides the references in random order, lists not only the requested references, but for over half of these (58.1%) also provides hyperlinks to the sources. The percentage of hyperlinked references ranged from 51.7% of those related to Australian archaeology to 72% of references related to Pacific archaeology. Some of these links were added to fictious references (see Section 3.2.3) and led to server errors.

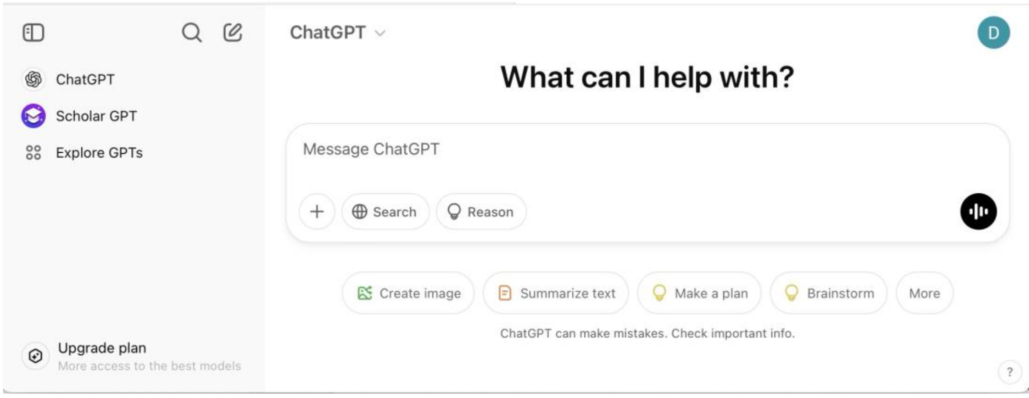

Figure 1. Screenshot of the ChatGPT4o interface, showing access to ScholarGPT (image 12 February 2025).

ScholarGPT generally served up the references in batches of ten, with the last sets often truncated. As an example, the final listing of references related to archaeological theory read "31–50 (Classics and Emerging Trends)" but only delivered two further references (Supplementary F). With the exception of the references related to cultural heritage management, ScholarGPT provided a label for the final section of its listing. Similar to the 'Classics and Emerging Trends' of archaeological theory, ScholarGPT offered 'Classic & Foundational Works' for both Pacific archaeology and Australian archaeology (Supplementary G & H). At the end of each listing, ScholarGPT provided an assertion of what the references in the preceding list covered and then offered to search for more specific topics, as exemplified by this text at the end of the references related to archaeological theory: "*This list combines foundational texts with cutting-edge research on archaeological theory. If you need references focused on a specific approach* (e.g., *post-processualism, feminist archaeology, digital archaeology), let me know!*" (Supplementary F).

3.1.4. DeepSeek

DeepSeek offered the references in random order. It also provided its references in APA-7 format, but did not follow its specification correctly by omitting editors of books in chapter citations. Like ChatGPT4o and ScholarGPT, DeepSeek did not provide doi numbers. The generative Ai application provided all of its references in a structured format. Setting aside the references related to cultural heritage management, which were structured as 'books', journal articles', and 'reports and guidelines' (Supplementary I), all others were constructed on the pattern of a section on 'foundational texts', to be followed by topical sections of five references each. For archaeological theory, for example, it offered sections on 'processual archaeology', 'post-processual archaeology', 'marxist and critical theory', 'feminist and gender archaeology', 'phenomenology and embodied approaches', 'agency and materiality', 'contemporary and multidisciplinary approaches', and 'key journal articles' (Supplementary J).

3.2. Authenticity of References

3.2.1. ChatGPT3.5

The propensity of ChatGPT3.5 to generate fictitious references has been noted before (Athaluri et al., 2023; Day, 2023; Gravel et al., 2023; Spennemann, 2023a). The 2023 study revealed that on average over half (57.7%) of all references offered by ChatGPT3.5 were 'confabulated' or acknowledged as fictional (Table 3) (Spennemann, 2023e). That percentage was not uniform, however, with the archaeological theory reference set containing the least (28.7%) and the Australian archaeology set containing the most (84.5%) fictitious references. Statistically, references to archaeological theory are very significantly more likely to be genuine than those related to cultural heritage management (χ^2 = 55.218, df = 1, p < 0.0001), while references related to Australian archaeology are very significantly less likely to be genuine than those related to cultural heritage management (χ^2 = 24.448, df = 1, p < 0.0001). No difference was noted between references related to Pacific archaeology and cultural heritage management (χ^2 = 0.04, df = 1, p = 0.8414). In addition, a small percentage of references was cited with the wrong year, ranging from 2.6% for archaeological theory to 15.2% for cultural heritage management (Table 3).

Table 3. Authenticity of references presented by ChatGPT3.5 (Spennemann, 2023e).

	Correct Citation	Wrong Year Cited	Confabulated Citation	Acknowledged as Fictional	n
Archaeological Theory	68.7	2.6	28.7	—	115
Cultural Heritage Management	26.2	15.2	34.8	23.8	210
Pacific Archaeology	27.2	6.4	66.4	—	125
Australian Archaeology	3.6	11.8	84.5	—	110
All sources	32.1	10.2	48.8	8.9	560

An examination of the non-existent references showed that these were generated using the names of real authors working in the field (in the main) with fragments of real article titles and journal or publisher names to construct entirely false but realistic looking references (Spennemann, 2023e). An example is the following reference provided by ChatGPT3.5 in a 2023 response—"Best, S., & Clark, G. (2008). Post-Spanish Contact Archaeology of Guahan (Guam). *Micronesian Journal of the Humanities and Social Sciences*, 7(2), 37–74"—which can be deconstructed into its components, as shown in Table 4.

Table 4. Deconstruction of a fictitious 'confabulated' reference (Spennemann, 2023e).

Reference Component	Commentary
Best, S.,	genuine author, Simon Best
Clark, G.	genuine author, Geoff Clark
(2008).	plausible year
Post-Spanish Contact Archaeology	contextually plausible time frame in title
of Guahan (Guam).	plausible location in title
Micronesian Journal of the Humanities and Social Sciences	genuine journal on record
7(2),	volume number does not exist as journal ceased with volume 5, 2006
37–74	pagination irrelevant as volume count incorrect

Other confabulated references tend to be truncated versions of genuine references with spurious text in the title, or with a spurious volume and page numbers. On occasion, such confabulated references are humorous, such as the title offered by ChatGPT3.5 (run

4 set 1): "Van der Aa, B., & Timmermans, W. (2014). The Future of Heritage as Clumsy Knowledge. In The Future of Heritage as Clumsy Knowledge (pp. 1–13). Springer, Cham". In this instance, only the publisher and the authors are genuine.

Of concern was that to a casual user who is not familiar with the literature, *all* references supplied by ChatGPT3.5 would have appeared valid and genuine because (i) the article titles appear plausible; (ii) the titles of journals are those of genuine publications, and (iii) and the vast majority of author names cited were those of academics actively publishing in the fields of cultural heritage or archaeology. With the exception of a single set of references (of 17 sets), there is no indication that references provided might not have been genuine. In the exception, ChatGPT3.5 provided the following caveat before offering the references: "Please note that some of these references might be fictional, and it's essential to verify their accuracy and relevance in academic and research contexts" (Spennemann, 2023e). This is exacerbated in some responses, where ChatGPT3.5 actually offered the user a "*curated* list of important and influential references in the field of archaeological theory", a "*curated* list of 10 reputable references on Australian archaeology", or a "list of 10 reputable references on cultural values in cultural heritage management". In all instances, several of the references provided were confabulated; any claim of 'curated' or 'reputable' is completely misleading (Spennemann, 2023e).

3.2.2. ChatGPT4o

The nature of references returned by ChatGPT4o is shown in Table 5. The overwhelming majority of citations were correct. A small percentage of otherwise correct citations were incomplete, with some journal references lacking volume and pagination data.

Table 5. Authenticity of references presented by ChatGPT4o.

	Correct Citation		Citation Error		Confabulated	n
	Full	**Incomplete**	**Year**	**Other**	**Citation**	**n**
Archaeological Theory	98.00				2.00	50
Cultural Heritage Management	84.00	4.00	4.00	2.00	6.00	50
Pacific Archaeology	78.00	6.00			16.00	50
Australian Archaeology	74.00	4.00	4.00	2.00	16.00	50
All Sources	83.50	3.50	2.00	1.00	10.00	200

A statistical comparison of the data as combined to the classes used for the ChatGPT3.5 analysis (Table 3) showed that compared with ChatGPT3.5 the overall proportions of confabulated references offered by ChatGPT4o had decreased very significantly, with only 10% of the offered references being fictitious (χ^2 = 135.220, df = 1, p < 0.0001). Among the four sets of references returned by ChatGPT4o, references related to archaeological theory were the least confabulated (2%), with references related to Australian and Pacific archaeology the most (16%) (Table 4). That difference was also statistically significant (χ^2 = 5.923, df = 1, p = 0.0149).

3.2.3. ScholarGPT

The nature of references returned by ScholarGPT differs considerably. Although required to provide 50 references for each of the four sets, it did so only for the cultural heritage management references (Table 5). Of note is that all offered references date to 2024 and 2025 (see below, Section 3.2.5). While almost half of these were genuine, a fair number of references were genuine but their original year of publication had been substituted with 2024 or 2025, resulting in almost a quarter of references with wrong year attributions (24.3%) (Table 6). Intriguingly, although ScholarGPT is purportedly tailored towards academic needs, it returned a very significantly lower percentage of genuine and correct references (47.8%) than ChatGPT4o (87.0%) (χ^2 = 60.426, df = 1, p < 0.0001). In addition, 11.8% of the genuine references proved to be irrelevant to the question asked (Table 6). The highest percentage of confabulated citations was returned for cultural heritage management references (44%), which was very significantly higher than references related to archaeological theory (χ^2 = 8.832, df = 1, p = 0.0030) and Australian archaeology (χ^2 = 8.832, df = 1, p = 0.0059).

Table 6. Authenticity of references presented by ScholarGPT.

	Correct Citation		Citation Error			Confabulated	
	Full	Incomplete	Irrelevant	Year	Other	Citation	n
Archaeological Theory	37.50	15.63	3.13	28.13	3.13	12.50	32
Cultural Heritage Management	18.00	6.00	12.00	20.00		44.00	50
Pacific Archaeology	20.69	27.59		27.59		24.14	29
Australian Archaeology	16.00	8.00	36.00	24.00	4.00	12.00	25
All Sources	22.79	13.24	11.76	24.26	1.47	26.47	136

3.2.4. DeepSeek

DeepSeek provided a high proportion of correct citations (85%) as well as otherwise correct citations that were incomplete (4.5%) (Table 7). This proportion ranged from a low of 78% for references related to cultural heritage management to entirely correct citations for archaeological theory. The highest proportion of confabulated citations was returned for the field of Pacific archaeology (14%).

Table 7. Authenticity of references presented by DeepSeek.

	Correct Citation		Citation Error		Confabulated	
	Full	Incomplete	Year	Other	Citation	n
Archaeological Theory	100.00					50
Cultural Heritage Management	78.00	12.00	4.00		6.00	50
Pacific Archaeology	80.00	4.00	2.00		14.00	50
Australian Archaeology	82.00	2.00	2.00	6.00	8.00	50
All Sources	85.00	4.50	2.00	1.50	7.00	200

3.2.5. Comparing the Models

When comparing the four generative Ai models, ChatGPT3.5 had the least percentage of correct citations (32.1%) followed by ScholarGPT (47.8%), with the difference being statistically very significant (χ^2 = 11.798, df = 1, p = 0.0006). At the other end of the scale were ChatGPT4o (87.0%) and DeepSeek (89.5%), which were statistically indistinguishable (χ^2 = 0.601, df = 1, p = 0.4381), but both of which were significantly different from ScholarGPT (ChatGPT4o, χ^2 = 60.426, df = 1, p < 0.0001; DeepSeek, χ^2 = 70.615, df = 1, p < 0.0001). While ScholarGPT is an improvement over ChatGPT3.5, it is surprising that its performance is so much worse than that of ChatGPT4o given that ScholarGPT is purportedly tailored towards academic needs, and given that it is offered at the same time and in the same interface (Figure 1). Of greatest concern is the high proportion of fictitious references offered to the user.

A comparison of the two 'best performers', ChatGPT4o and DeepSeek, shows that in terms of genuine references, they are not only statistically indistinguishable overall, but also at the level of each of the four subsets. Neither one is immune to a small percentage of confabulated citations (DeepSeek 7%, ChatGPT4o 10%), which highlights their limitations.

From a brainstorming perspective, DeepSeek's ability to deliver a small number of references in a conceptually structured fashion would allow the user to consider what aspects should be explored in more depth. As noted elsewhere, however, for such brainstorming to be productive, the user needs to already possess a good grasp of concepts (Spennemann, 2023c).

When considering the year of publication of cited references, irrespective of their authenticity (Table 8), then the percentage distribution over time of references cited by ChatGPT4o and DeepSeek is significantly correlated (r^2 = 0.9657), while this is not the case for ChatGPT4o when compared with its predecessor ChatGPT3.5 (r^2 = 0.1910).

Table 8. Year of publication of cited references, irrespective of authenticity (in %).

Decade	ChatGPT3.5	ChatGPT4o	DeepSeek	ScholarGPT
1960–1969		2.0	2.5	
1970–1979	1.1	5.0	3.5	
1980–1989	2.2	11.0	7.0	
1990–1999	6.0	21.5	20.5	
2000–2009	19.5	39.0	39.5	
2010–2019	59.9	21.5	27.0	
2020–2025	11.4			100.0
Total	369	200	200	136

3.3. The Sources of the Genuine References

The 2003 paper examined the potential origin of those references that had been deemed genuine by examining whether they were accessible in online sources (Spennemann, 2023e). The full text of 16.8% of all references could be freely accessed (via Google Books, JSTOR, or otherwise online), while close to two thirds (66.4%) were accessible in Google Books preview mode, which allows perusal of some, but not all, of the text (Spennemann, 2023e). In a different context, it could be shown that ChatGPT3.5's knowledge must have come either from the media or, more likely, from Wikipedia (Spennemann, 2023c). The assessment of all genuine references cited by ChatGPT3.5 showed that *all* genuine sources had been cited in Wikipedia or one of its siblings such as Wikidata. That assessment further showed that the proportion of genuine versus confabulated references among the four data sets provided by ChatGPT3.5 reflected the volume of Wikipedia text dedicated to that discipline (Spennemann, 2023c). This then raised the question whether the 'knowledge' used by ChatGPT3.5 stems from primary sources, as the genuine citations might indicate, or whether it was merely based on the secondary source texts of the Wikipedia pages.

The literature has shown that a wide range of books was used to train the generative Ai models (Chang et al., 2023; Grynbaum & Mac, 2023). This could be documented with the cloze prompt methodology, where the generative Ai application is given a section of text which it is suspected to have 'read' with the request to correctly identify a word that was removed from a text sequence (Chang et al., 2023; Kancko, n.d.). It could be expected that the generative Ai application should be able to perform this task without errors if it had incorporated and 'read' the publication as part of is training phase. Selected were sample texts from the fields of archaeological theory, Australian archaeology, and Pacific archaeology. Tested were the following three sources: (i) Alison Wylie's edited book *Thinking from Things* (Wylie, 2002); (ii) Bruce Pascoe's *Dark Emu* (Pascoe, 2014); and (iii) Patrick Kirch and Roger Green's *Hawaiki, Ancestral Polynesia* (Supplementary I) (Kirch & Green, 2001). A set of ten sample sentences were chosen for each source, with five sentences drawn from text publicly accessible via Google Books Preview, and five sentences for which access to an actual copy of the work was required. The prompt is reproduced as Table 2.

3.3.1. ChatGPT3.5

ChatGPT3.5 accurately implemented the task and returned a single name as a separate line: "Output: <name>XYZ</name>". In *Dark Emu,* only one sample sentence (6) was consistently correctly answered by ChatGPT3.5 by providing the missing word 'Glock'. In sample sentence 10, it correctly provided the missing word 'Stanner', in all but the first iteration. Both sample sentences stem from sections of the book that are not publicly available via Google Preview. The accuracy for *Dark Emu* was 27.5% (Table 9).

Table 9. Percentage of correct identifications using cloze analysis.

Run	Set	Answer	*Dark Emu* GPT v3.5	*Dark Emu* GPT v4o	*Dark Emu* Sch GPT	*Dark Emu* DS v3	*Thinking from Things* GPT v3.5	*Thinking from Things* GPT v4o	*Thinking from Things* Sch GPT	*Thinking from Things* DS v3	*Hawaiki* GPT v3.5	*Hawaiki* GPT v4o	*Hawaiki* Sch GPT	*Hawaiki* DS v3
1	1	Initial	10	30	40	30	20	0	20	30	70	30	60	60
		Regenerated	20	10	50	30	30	10	20	20	70	40	50	40
	2	Initial	30	—	—	—	30	—	—	—	70	—	—	—
		Regenerated	30	—	—	—	20	—	—	—	70	—	—	—
2	1	Initial	20	10	60	40	20	20	20	20	60	40	50	30
		Regenerated	30	20	60	30	20	30	30	20	40	50	50	50
	2	Initial	40	—	—	—	20	—	—	—	50	—	—	—
		Regenerated	40	—	—	—	20	—	—	—	50	—	—	—
		Average	27.5	17.5	52.5	32.5	22.5	15.0	22.5	22.5	60.0	40.0	52.5	45.0
		n	240	120	120	120	240	120	120	120	240	120	120	120

Abbreviations: DS v3—DeepSeek v.3; GPT v3.5—ChatGPT3.5; GPT v4o—ChatGPT4o; SchGPT—ScholarGPT.

In *Thinking from Things,* two sample sentences (6 and 7) were consistently correctly completed by ChatGPT3.5 by providing the missing words 'Laudan' and 'Sutton', respectively. Both sample sentences also stem from sections of the book that are not publicly available via Google Preview. The accuracy for *Thinking from Things* was 22.5%

Half the sample sentences of the last sample set, taken from *Hawaiki, Ancestral Polynesia*, were consistently correctly answered (1, 2, 6, 7, 9), while sample sentence 8 was correctly answered in seven of the eight responses. Two of the sample sentences were publicly accessible, while the remaining four (incl. sentence 8) stem from sections of the book that are not publicly available via Google Preview. The accuracy for *Hawaiki, Ancestral Polynesia* was 60%.

The overall accuracy of ChatGPT3.5 to correctly identify the missing words across the three texts and all runs was 36.7%.

3.3.2. ChatGPT4o

ChatGPT4o likewise accurately implemented the task and returned a single name as a separate line: "Output: <name>XYZ</name>". While it could correctly provide missing words, this was not consistent in the four runs that were conducted for each text. Missing words in three of the ten sample sentences taken from *Hawaiki, Ancestral Polynesia*, were consistently correctly supplied (6, 7, 9), while the missing word in sample sentence 8 was correctly returned in three of the four responses. In all instances, the sentences were taken from a section of the book that was not publicly available via Google Preview. In the works *Dark Emu* and *Thinking from Things,* ChatGPT4o did not consistently identify a missing word. Missing words were correctly returned in three of the four responses in sentence 10 of *Dark Emu* and sentence 6 in *Thinking from Things.* The overall accuracy of ChatGPT4o to correctly identify the missing words across the three texts and all runs was 24.2%, which is significantly lower than the accuracy rate of ChatGPT3.5 (χ^2 = 5.679, df = 1, p = 0.0172).

3.3.3. ScholarGPT

As the other incarnation of the OpenAi universe, ScholarGPT accurately implemented the task and returned a single name as a separate line: "Output: <name>XYZ</name>". ScholarGPT was able to consistently correctly supply missing words in sample sentences 2, 3, 6, 8, and 9 of from *Hawaiki, Ancestral Polynesia*. Three sentences (6, 8, and 9) were taken from a section of the book that was not publicly available via Google Preview, while the other two were taken from a section of the book available via Google Preview. In the text *Thinking from Things,* DeepSeek consistently correctly supplied missing words in sample sentences 6 and 7 (both taken from a section of the book that was not publicly available via Google Preview). In the text *Dark Emu,* DeepSeek correctly supplied the missing word in sentences 1, 6, 9, and 10, of which only one (sentence 1) was taken from a section of the book that that was publicly available via Google Preview. With an overall accuracy rate of 42.5% in correctly identifying the missing words across the three texts and all runs, ScholarGPT was the highest of all four generative Ai models. The rate was significantly higher than the accuracy rate of ChatGPT4o (χ^2 = 9.002, df = 1, p = 0.0027).

3.3.4. DeepSeek R1

As a standard, DeepSeek R1 provides a narrative of the reasoning behind its identification (see Section 3.5) and then, at the end, returns a single name as a separate line: "Output: <name>XYZ</name>". DeepSeek was able to consistently correctly supply missing words in sample sentences 1, 7, 8, and 9 of from *Hawaiki, Ancestral Polynesia*. Four sentences (7–10) were taken from a section of the book that was not publicly available via Google Preview. The missing word was also correctly returned in three of the four responses to sentence 10. In the text *Thinking from Things,* DeepSeek consistently correctly supplied missing words in sample sentences 6 and 7 (both taken from a section of the book that was not publicly available via Google Preview). In the text *Dark Emu,* DeepSeek correctly supplied the missing word in sentence 6, which was also taken from a section of the book that was not publicly available via Google Preview. In addition, the missing word in sample sentence 1, taken from a section of the book available via Google Preview, was correctly returned in three of the four responses. The overall accuracy of DeepSeek R1 to correctly identify the missing words across the three texts and all runs was 33.3%. While this is higher than the accuracy rate of ChatGPT4o and lower than the accuracy rates of both ScholarGPT and ChatGPT3.5, the differences are not significant.

3.4. Training Data vs. Reasoning

While the majority of the genuine references could be tracked back to possible sources the question arises whether the generative Ai models have actually 'read' these sources during their training phase, or whether they drew those references from pre-compiled lists or obtained them from other sources.

A cloze analysis carried out for ChatGPT3.5 in 2023 (Spennemann, 2023e) as well as for ChatGPT4o, ScholarGPT, and DeepSeek for this paper showed that none of the generative Ai models were able to consistently correctly supply words missing in ten test sentences taken from three works each. The highest overall rate of accuracy was 42.5% (n = 120) (ScholarGPT), with another model from the same 'stable' (ChatGPT4o) being as low as 24.2% (n = 120). When considering only instances where the generative Ai model correctly identified the missing word in all four sets, accuracy rates dropped to 33.3% for ScholarGPT, 20% for DeepSeek, and as low as 10% for ChatGPT4o. Intriguingly, ChatGPT3.5 had the highest strike rate at 36.7%.

What the cloze analyses of the four generative Ai models have made abundantly clear, however, is that while the ChatGPT generative Ai models can reach into their training set for a consistently correct supply of the missing words taken from novels, they cannot do the same for the academic texts they cite and that were tested here. This suggests that the citation and any related information must have come from other sources. The nature of the training data set of DeepSeek has not been made public at the time of writing. Of relevance here is the observation that an examination of the combined capabilities of the four models in relation to the three source texts showed that they were significantly more likely to correctly supply words missing in the text *Hawaiki, Ancestral Polynesia* (42.5%) than in the texts *Dark Emu* (17.5%, χ^2 = 89.211, df = 1, p < 0.0001) or *Thinking from Things* (15.0%, χ^2 = 110.663, df = 1, p < 0.0001).

A set of tests carried out when assessing the capabilities of ChatGPT3.5 noted that text sections that were seemingly not publicly accessible via Google Books could still be found via a simple Google search where the phrase text was found to be linked back to the correct section in Google Books (where it is blocked from view, however) (Spennemann, 2023e). The same, obviously, applies to the situation at the time of writing. Any positive identification, then, is the result of 'reasoning' and pattern recognition, rather than 'knowledge'. The 'accuracy rate' then becomes a measure of the success of reasoning. While half or almost half of the possible words (n = 30) were correctly guessed at least once, the proportion of correct guesses (success rate) ranged from 25.8% (ChatGPT4o) to 42.5% (ChatGPT3.5). At the same time, ChatGPT4o also offered the greatest number of spurious names (60) (Table 10). In addition, 36.7% of the correct words (n = 30) were not identified by any of the four generative Ai models.

Table 10. Words guessed by the generative Ai models.

Generative Ai Model	Correctly Guessed at Least Once (%) (n = 30)	Proportion of Correct Guesses (%) (n = 120)	Incorrect Names (n)
ChatGPT3.5	50.0	42.5	34
ChatGPT4o	46.7	25.8	60
ScholarGPT	46.7	40.0	33
DeepSeek	46.7	34.2	42

Two generative Ai applications either automatically provide their reasoning in the form of a narrative (DeepSeek) or can be asked to display this (ChatGPT4o). The reasoning approach can be quite straightforward (Supplementary Q Example A) or can be very convoluted (Supplementary P Example G). The reasoning is iterative. An example of (incorrect) reasoning by DeepSeek R1 is given in Box 1. The 'thinking' times of DeepSeek R1, for answers for which the process was documented, ranged from 15 s to 218 s, with one process timing out altogether (presumably at 240 s).

Given that all examined generative language models use transformer architecture to generate coherent and contextually relevant responses, it is possible that, due to the extent of specific context and conceptual detail provided, some of the reasoning was more straightforward than others and that therefore some text samples were more likely to result in correct answers. In *Dark Emu*, for example, one sample sentence (n° 6) was consistently correctly completed by ChatGPT3.5, ScholarGPT, and DeepSeek. Upon reflection, that restitution task provided specific context and conceptual detail: "So the police escorted the 4WD heroes into the initiation site. Once there, they threw beer cans into the sacred water, and took it in turns to shoot at the cans with police-issue Glock pistols". Given the context of 'shoot at the cans' and 'police-issue . . . pistols', and a setting in Australia, this considerably narrows the substitution of the missing word to 'Glock'. The fourth generative

Ai model, ChatGPT4o, correctly identified 'Glock' in two runs, offering 'revolver' in the other two.

To test the extent to which specific context and conceptual detail might influence generative Ai responses, two sample sentences taken from '*Hawaiki, Ancestral Polynesia*', which had a high proportion of correct identifications, were retested, with different proper nouns omitted. To ensure that the generative Ai model could not draw on any context in the chat, the question was posed every time in a newly started chat, with the previous chat deleted.

Box 1. Example of incorrect reasoning by DeepSeek R1.

(11 February 2025 19: 03 GMT, 18 s 'thinking' time)

Input: Fort Walsh of the 1870s represents a fascinating moment in the "conquest" of the Canadian West, as [MASK] describes it (1987), and the archaeological project developed for Parks Canada by James V. Sciscenti in the 1970s was remarkable in a number of ways

First, the sentence mentions Fort Walsh in the 1870s and the "conquest" of the Canadian West. The author in question is referenced with a citation from 1987. The archaeological project by James V. Sciscenti in the 1970s is also mentioned.

I need to think of a historian or scholar who wrote about the Canadian West around 1987, specifically using the term "conquest" in that context. The name should be one word, a proper noun.

I recall that Gerald Friesen is a Canadian historian who wrote about Western Canada. His book "The Canadian Prairies: A History" was published in 1984, which is close to 1987. He might discuss the conquest aspect. Another possibility could be someone like John Tobias, but I'm not sure about the publication date.

Wait, the example given earlier had Hempel paired with Carnap, both logical positivists. So here, the author cited in 1987 discussing the conquest of the Canadian West. Friesen's work fits the timeframe and topic. Alternatively, maybe Sarah Carter? But her notable works might be a bit later.

Gerald Friesen seems the most likely. I'll go with Friesen.

Output: <name>Friesen</name>

The passage refers to a 1987 source discussing the "conquest" of the Canadian West. Gerald Friesen, a historian known for works like The Canadian Prairies: A History (1984), aligns with this context, as his scholarship focuses on Western Canadian history and Indigenous-settler relations. While the exact match to 1987 is uncertain, Friesen is a plausible candidate given his prominence in this field. Other possibilities (e.g., Sarah Carter, John Tobias) either postdate 1987 or lack direct ties to the "conquest" framing.

Sample sentence n° 6 listed three anthropologists in succession: "The founders of the unique Americanist tradition in anthropology–Boas, Kroeber, Sapir, and others–reacted in part to the theoretical excesses . . ." The initial assessment asked to identify 'Kroeber' as the missing word, which all three ChatGPT versions (ChatGPT3.5, ChatGPT4o, and ScholarGPT) consistently achieved. DeepSeek did so in half the sets, with 'Sahlins' being offered in the other two. When asked to find the correct replacement for the missing first name in the sequence, 'Boas', all four generative Ai models consistently answered correctly. When asked to find the correct replacement for the third name in the sequence ('Sapir'), however, ChatGPT3.5 offered three incorrect names (Lowie, Malinowski [2x], and Mead) and ChatGPT4o offered two incorrect names (Lowie [3x] and Benedict). ScholarGPT offered the same name, 'Lowie' in all four answers. DeepSeek, on the other hand consistently provided the correct word reconstruction. An examination of its reasoning showed that it correctly identified and linked Boas and Kroeber with linguistic anthropology and was able to connect other scholars, commonly by virtue of being pupils of Boas or Kroeber. One of the 'deciding' factors for DeepSeek, however, was that 'Sapir' was often mentioned as a third author in the "Boas, Kroeber, XYZ' sequence (Supplementary P, Examples K & L).

In sample sentence 7, which starts with "In reading Flannery and Marcus otherwise brilliantly argued volume. . . Polynesian scholars had surpassed our Mesoamerican colleagues, because. . .", the generative Ai models had to find the correct replacement for the missing name 'Marcus', which all models consistently achieved. When tasked with finding the correct replacement for 'Flannery', however, ChatGPT3.5 offered three incorrect names (Trudgill, Lyle, Greenhill [2x]), with both ChatGPT4o and ScholarGPT offering four different names each, all of which were incorrect (ChatGPT4o: Campbell, Green, Kirch, Renfrew; ScholarGPT: Campbell, Diamond, Kirch, Renfrew). DeepSeek responded once with the correct word, 'Flannery', while in the other three instances offered 'Clifford'. When tasked with finding the correct missing word, 'Mesoamerican', all three ChatGPT versions (ChatGPT3.5, ChatGPT4o, and ScholarGPT) consistently responded with 'Western', while DeepSeek offered three incorrect words (Americanist, Indo-European (2x), Swadesh).

As noted, one of the 'deciding' factors for DeepSeek during its reasoning when solving sample sentence 6 was that 'Sapir' was often mentioned as a third author in the "Boas, Kroeber, XYZ' sequence (e.g., Supplementary P, Examples K & L). A Google Search for the phrase "Boas, Kroeber, Sapir" found 45 instances. Likewise, the mention of Flannery and Marcus in the test sentence in connection with 'Mesoamerica' and 'historical linguists' provided enough clues for a reliable identification of 'Marcus' (derived from 'The Clod People' (Flannery & Marcus, 2003)), while the identification of 'Flannery' was made complicated by other 'clues' in the sample sentence that pointed to (Lyle) Campbell, who was widely published as a lead author in works on Mesoamerican linguistics (e.g., (Campbell, 1979; Campbell & Kaufman, 1980, 1983, 1985)), or by the 'clue' 'Polynesian scholars' which pointed to Green and Kirch. The request to correctly identify 'Mesoamerican' was not embedded in sufficient contextual information for the generative Ai models to come to a ready identification.

The long-winded, almost tortuous path taken by DeepSeek, which resulted in the incorrect solution 'Indo-European', is illuminating here (e.g., Supplementary P, Example M). The 2023 paper could demonstrate that much, if not most of ChatGPT's knowledge base is rooted in the secondary material contained in the Wikipedia universe (Spennemann, 2023e). While the generative Ai model is able to correctly distil the key characteristics of a task, it 'settles' on a solution of the highest probability where the actual text does not form part of the training data set. This probability appears to be based on overall frequency. A targeted and constrained Google Search within Wikipedia ('[search tems] site: https://www.wikipedia.org' accessed 15 February 2025), for example, found "about 82,500 results" for the terms 'Mesoamerica linguistics', while a search for the terms 'Indo-European linguistics' found "about 1,100,000 results".

3.5. The Quality of Resonses as 'Self-Reported' by ChatGPT4o and DeepSeek R1

In the final step, drawing on the interactive nature of the generative Ai language models, both ChatGPT4o and DeepSeek R1 were engaged in a conversation which also required that their reasoning be explained. The aim was to illuminate the origin and veracity of references provided as well as the quality of the data.

3.5.1. ChatGPT4o

When asked "where do you get your references from," ChatGPT4o replied in its reasoning: *"I don't have direct access to references or citations. When I answer questions, I rely on patterns from the data I was trained on, which is a mix of books, articles, websites, and other publicly available information. However, I don't have a full list of sources for everything. I can't pull up real-time data either, and my knowledge is limited to what was included up until my 2021 knowledge cutoff"* (Supplementary R Conversation B). When queried as to the specific date of the most recent data it had access to, ChatGPT4o's response, both in its reasoning and its answers, was consistently 'September 2021' which seems to contradict assertions that its training data set was updated until October 2023 (Conway, 2024; OpenAI, 2025).

When queried as to "how reliable are the references that you provide", ChatGPT4o in its reasoning stated that it "*does not 'provide references' in a traditional sense. Instead, I generate responses based on a mixture of publicly available data and licensed sources. In some cases, I can include references if prompted, but I don't always check them. These references are as reliable as my training data*" (Supplementary R Conversation A). In a different 'conversation', ChatGPT4o asserted that it "*generate[s] responses based on a mixture of information learned during training from a wide variety of sources—including reputable books, academic papers, websites, and other texts*" and "*provide[s] information that aligns with widely accepted knowledge*" but that it does not "*pull directly from specific sources when answering questions*". It also stated that "*verification is key: For critical, complex, or highly specialized topics, it's always a good idea to verify the information with primary or authoritative sources*... *recommend[ing] cross-checking important details with trusted references or experts when precision is crucial*" (Supplementary R Conversation B).

While the response provided by ChatGPT4o confirms the empirically determined limitations and accessing texts and the veracity of citations, the assertions regarding the nature of the training data are generic. It suggests that all training data are reliable and authoritative, which at least in the case of Wikipedia has been drawn into question (Baigutanova, 2024; Nicholson et al., 2021).

3.5.2. DeepSeek R1

DeepSeek R1 'evaded' the answer for where its references are from by 'inviting' the user "to consult our official documentation" (Supplementary S Conversations A and B). When asked for the specific date of the most recent data it had access to, DeepSeek's response, both in its reasoning and its answers, was consistently 'October 2023'. When queried as to "how reliable are the references that you provide", DeepSeek in its answer stated that "*the references I provide are generated based on my training data, which includes a vast amount of publicly available text up to my knowledge cutoff in October 2023*...*references* (e.g., *academic papers, books, or articles) are generated from patterns in my training data and may not correspond to actual, verifiable sources*...*[and] while I strive to provide accurate information, synthesized references* (e.g., *author names, publication dates, or titles) might be approximate, outdated, or occasionally fictional. For example, I may generate a plausible-sounding paper title that does not exist in reality*" (Supplementary S Conversation A). Intriguingly, DeepSeek seemed to differentiate its audiences in terms of a need for veracity, when it asserted that "*for casual inquiries, synthesized references may suffice*" but that "*for academic or professional work*" it was suggested that the user "*always cross-check details using reliable sources*" (Supplementary S Conversation A).

4. Conclusions and Implications

The public release of ChatGPT3.5 in late 2022 has resulted in considerable publicity and has led to widespread discussion of the usefulness and capabilities of generative Ai language models. Their ability to extract and summarise data from textual sources and present them as human-like contextual responses makes them an eminently suitable tool to answer a wide array of questions that users might ask. The majority of questions asked of its successor ChatGPT4o relate to technology and programming, education and learning, writing and editing, as well as general knowledge and trivia (Supplementary T). Any collated and synthesised information provided by a generative Ai model, however, can only ever be as good as the data it has access to.

This study has shown that while the proportion of confabulated references offered by generative Ai models has decreased between ChatGPT3.5 (tested in 2023) and the current versions ChatGPT4o and ScholarGPT, responses still contain fictitious references, in particular responses by ScholarGPT. The recent offering of the Chinese-designed generative Ai model DeepSeek is also not devoid of an, albeit low, propensity to offer fictitious references.

Cloze analyses, which asked the generative Ai models to correctly identify a single word that had been masked out in a test sentence, showed that the tested models were unable to do so reliably. This proved that the assessed academic works, although cited by the generative Ai models, had not formed part of the original training set. Thus, any data, information, concepts, or 'knowledge' pertaining to the content of these academic works derived from secondary or tertiary sources. Further analysis showed that all genuine sources cited by the generative Ai models can also be found in Wikipedia pages. This raises serious concerns regarding the quality of the information that the generative Ai models offer the user.

A study of the quality of scientific references in Wikipedia by Nicholson et al. (Nicholson et al., 2021) has shown that "most scientific articles cited by Wikipedia articles are uncited or untested by subsequent studies, and the remainder show a wide variability in contradicting or supporting evidence". A study by Baigutanova demonstrated "cross-lingual discrepancies in the perennial sources list and the persistence of untrustworthy sources across different language editions" (Baigutanova, 2024). In addition, a number of Wikipedia articles purport conspiracy theories or are hoaxes (Borkakoty & Espinosa-Anke, 2024), which dilute the quality of an automatically harvested data set.

The quality of responses by generative Ai models clearly depends on the quality of the data that have been included in the training data. There is no denying that generative Ai models can be powerful tools in a closed system, where the quality and authenticity of high-quality training data can be guaranteed (such as a museum, company, or similar applied setting) and where adequate quality control by human trainers during the training phase can ensure that erroneous connections and confabulations are minimised. Beyond closed systems, however, freely available generative Ai models, which develop text based on associations, may offer generated text that at first sight will appear plausible and genuine, but which upon closer examination may be found to be flawed or even wrong. To identify this, a user needs to possess both developed critical thinking skills and foundational background knowledge in the topic, which the majority of public users may not possess. Caveat emptor!

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/publications13010012/s1.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The original data presented in the study are openly available in doi 10.26189/a033ba56-ac96-438b-bfeb-25adad6d692d.

Conflicts of Interest: The author declares no conflicts of interest.

References

Adeshola, I., & Adepoju, A. P. (2024). The opportunities and challenges of ChatGPT in education. *Interactive Learning Environments*, *32*(10), 6159–6172. [CrossRef]

Agapiou, A., & Lysandrou, V. (2023). Interacting with the Artificial Intelligence (AI) language model ChatGPT: A synopsis of earth observation and remote sensing in archaeology. *Heritage*, *6*(5), 4072–4085. [CrossRef]

Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*, *15*(2), e35179. [CrossRef]

Allen, L., O'Connell, A., & Kiermer, V. (2019). How can we ensure visibility and diversity in research contributions? How the Contributor Role Taxonomy (CRediT) is helping the shift from authorship to contributorship. *Learned Publishing*, *32*(1), 71–74.

Anderson, A., & Correa, E. (2019). *Critical explorations of online sources in a culture of "fake news, alternative facts and multiple truths"*. Global Learn.

Armitage, R., & Vaccari, C. (2021). Misinformation and disinformation. In H. Tumber, & S. Waisbord (Eds.), *The Routledge companion to media disinformation and populism* (pp. 38–48). Routledge.

Athaluri, S. A., Manthena, S. V., Kesapragada, V. K. M., Yarlagadda, V., Dave, T., & Duddumpudi, R. T. S. (2023). Exploring the boundaries of reality: Investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. *Cureus*, *15*(4), e37432. [CrossRef] [PubMed]

Babl, F. E., & Babl, M. P. (2023). Generative artificial intelligence: Can ChatGPT write a quality abstract? *Emergency Medicine Australasia*, *35*(5), 809–811. [CrossRef]

Baigutanova, A. (2024). *Large-scale analysis of reference quality in heterogeneous Wikipedia datasets*. Korea Advanced Institute of Science & Technology.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., & Chung, W. (2023). A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv*, arXiv:2302.04023.

Bays, H. E., Fitch, A., Cuda, S., Gonsahn-Bollie, S., Rickey, E., Hablutzel, J., Coy, R., & Censani, M. (2023). Artificial intelligence and obesity management: An Obesity Medicine Association (OMA) Clinical Practice Statement (CPS) 2023. *Obesity Pillars*, *6*, 100065. [CrossRef]

Biswas, S. (2023). *Importance of chat GPT in agriculture: According to chat GPT. Available at SSRN 4405391.* arXiv:2305.00118. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4405391 (accessed on 5 February 2025).

Bloxham, S. (2012). 'You can see the quality in front of your eyes': Grounding academic standards between rationality and interpretation. *Quality in Higher Education*, *18*(2), 185–204. [CrossRef]

Borkakoty, H., & Espinosa-Anke, L. (2024). Hoaxpedia: A unified Wikipedia Hoax articles dataset. *arXiv Preprint*, arXiv:2405.02175.

Campbell, L. (1979). Middle american languages. In *The languages of native America: Historical and comparative assessment* (pp. 902–1000). University of Texas Press.

Campbell, L., & Kaufman, T. (1980). On mesoamerican linguistics. *American Anthropologist*, *82*(4), 850–857. [CrossRef]

Campbell, L., & Kaufman, T. (1983). Mesoamerican historical linguistics and distant genetic relationship: Getting it straight. *American Anthropologist*, *85*(2), 362–372. [CrossRef]

Campbell, L., & Kaufman, T. (1985). Mayan linguistics: Where are we now? *Annual Review of Anthropology*, *14*, 187–198. [CrossRef]

Cao, Y., Zhou, L., Lee, S., Cabello, L., Chen, M., & Hershcovich, D. (2023). Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. *arXiv*, arXiv:2303.17466.

Castro Nascimento, C. M., & Pimentel, A. S. (2023). Do large language models understand chemistry? A conversation with ChatGPT. *Journal of Chemical Information and Modeling*, *63*(6), 1649–1655. [CrossRef]

Chang, K. K., Cramer, M., Soni, S., & Bamman, D. (2023). Speak, memory: An archaeology of books known to chatgpt/gpt-4. *arXiv*, arXiv:2305.00118.

Chiesa-Estomba, C. M., Lechien, J. R., Vaira, L. A., Brunet, A., Cammaroto, G., Mayo-Yanez, M., Sanchez-Barrueco, A., & Saga-Gutierrez, C. (2024). Exploring the potential of Chat-GPT as a supportive tool for sialendoscopy clinical decision making and patient information support. *European Archives of Oto-Rhino-Laryngology*, *281*(4), 2081–2086. [CrossRef] [PubMed]

Ciaccio, E. J. (2023). Use of artificial intelligence in scientific paper writing. *Informatics in Medicine Unlocked*, *41*, 101253. [CrossRef]

Conway, A. 2024 May 13. *What is GPT-4o? Everything you need to know about the new OpenAI model that everyone can use for free"*. XDA Developers.

Day, T. (2023). A preliminary investigation of fake peer-reviewed citations and references generated by ChatGPT. *The Professional Geographer*, *75*(6), 1024–1027. [CrossRef]

DeepSeek. (2025). *DeepSeek into the unknown. R1 Model V3.* Hangzhou DeepSeek Artificial Intelligence Co., Ltd.; Beijing DeepSeek Artificial Intelligence Co., Ltd. Available online: https://www.deepseek.com (accessed on 5 February 2025).

Elazar, Y., Kassner, N., Ravfogel, S., Feder, A., Ravichander, A., Mosbach, M., Belinkov, Y., Schütze, H., & Goldberg, Y. (2022). Measuring causal effects of data statistics on language model's factual'predictions. *arXiv*, arXiv:2207.14251.

Fergus, S., Botha, M., & Ostovar, M. (2023). Evaluating academic answers generated using ChatGPT. *Journal of Chemical Education*, *100*(4), 1672–1675. [CrossRef]

Fernández-Sánchez, A., Lorenzo-Castiñeiras, J. J., & Sánchez-Bello, A. (2025). Navigating the future of pedagogy: The integration of AI tools in developing educational assessment rubrics. *European Journal of Education*, *60*(1), e12826. [CrossRef]

Ferrara, E. (2023). Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv*, arXiv:2304.03738.

Flannery, K. V., & Marcus, J. (2003). *The cloud people: Divergent evolution of the Zapotec and Mixtec civilizations*. Academic Press.

Franzen, C. 2024 November 22. *DeepSeek's first reasoning model R1-Lite-Preview turns heads, beating OpenAI o1 performance. VentureBeat.* [via WayBackMachine]. Available online: https://web.archive.org/web/20241122010413/https://venturebeat.com/ai/deepseeks-first-reasoning-model-r1-lite-preview-turns-heads-beating-openai-o1-performance/ (accessed on 5 February 2025).

Giray, L. (2024). ChatGPT references unveiled: Distinguishing the reliable from the fake. *Internet Reference Services Quarterly*, *28*(1), 9–18. [CrossRef]

Gravel, J., D'Amours-Gravel, M., & Osmanlliu, E. (2023). Learning to fake it: Limited responses and fabricated references provided by ChatGPT for medical questions. *Mayo Clinic Proceedings: Digital Health*, *1*(3), 226–234. [CrossRef]

Grünebaum, A., Chervenak, J., Pollet, S. L., Katz, A., & Chervenak, F. A. (2023). The exciting potential for ChatGPT in obstetrics and gynecology. *American Journal of Obstetrics and Gynecology*, *228*(6), 696–705. [CrossRef]

Grynbaum, M. M., & Mac, R. (2023, December 27). The Times sues OpenAI and Microsoft over AI use of copyrighted work. *The New York Times*.

Hartmann, J., Schwenzow, J., & Witte, M. (2023). The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. *arXiv*, arXiv:2301.01768. [CrossRef]

Hill-Yardin, E. L., Hutchinson, M. R., Laycock, R., & Spencer, S. J. (2023). A Chat (GPT) about the future of scientific publishing. *Brain Behavior and Immunity*, *110*, 152–154.

Hwang, T., Aggarwal, N., Khan, P. Z., Roberts, T., Mahmood, A., Griffiths, M. M., Parsons, N., & Khan, S. (2024). Can ChatGPT assist authors with abstract writing in medical journals? Evaluating the quality of scientific abstracts generated by ChatGPT and original abstracts. *PLoS ONE*, *19*(2), e0297701. [CrossRef]

Kacena, M. A., Plotkin, L. I., & Fehrenbacher, J. C. (2024). The use of artificial intelligence in writing scientific review articles. *Current Osteoporosis Reports*, *22*(1), 115–121. [CrossRef]

Kancko, T. n.d. *Authorship verification via cloze-test*. Masaryk University.

Kendall, G., & Teixeira da Silva, J. A. (2024). Risks of abuse of large language models, like ChatGPT, in scientific publishing: Authorship, predatory publishing, and paper mills. *Learned Publishing*, *37*(1). [CrossRef]

King, M. R. (2023). The future of AI in medicine: A perspective from a Chatbot. *Annals of Biomedical Engineering*, *51*(2), 291–295. [CrossRef]

Kirch, P. V., & Green, R. C. (2001). *Hawaiki, ancestral Polynesia: An essay in historical anthropology*. Cambridge University Press.

Lapp, E. C., & Lapp, L. W. (2024). Evaluating ChatGPT as a viable research tool for typological investigations of cultural heritage artefacts—Roman clay oil lamps. *Archaeometry*, *66*(3), 696–717. [CrossRef]

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, *13*(4), 410. [CrossRef]

Lu, D. (2025, January 28). We tried out DeepSeek. It worked well, until we asked it about Tiananmen Square and Taiwan. *The Guardian*. Available online: https://www.theguardian.com/technology/2025/jan/28/we-tried-out-deepseek-it-works-well-until-we-asked-it-about-tiananmen-square-and-taiwan (accessed on 5 February 2025).

Lund, B. D., & Naheem, K. (2024). Can ChatGPT be an author? A study of artificial intelligence authorship policies in top academic journals. *Learned Publishing*, *37*(1), 13–21.

Maas, C. (2023, May 13). *Was kann ChatGPT für Kultureinrichtungen tun?* TS2 Space, LIM Center. Available online: https://www.aureka.ai/de/aureka-blog/2024/12/26/warum-gpt-fuer-kultureinrichtungen-im-jahr-2025-wichtig-ist (accessed on 29 June 2024).

Macfarlane, B., Zhang, J., & Pun, A. (2014). Academic integrity: A review of the literature. *Studies in Higher Education*, *39*(2), 339–358. [CrossRef]

Markov, T., Zhang, C., Agarwal, S., Eloundou, T., Lee, T., Adler, S., Jiang, A., & Weng, L. 2023 August 22. *New and improved content moderation tooling. [via Wayback Machine]*. Available online: https://web.archive.org/web/20230130233845mp_/https://openai.com/blog/new-and-improved-content-moderation-tooling/ (accessed on 28 June 2024).

Martin, L., Whitehouse, N., Yiu, S., Catterson, L., & Perera, R. (2024). Better call GPT, comparing large language models against lawyers. *arXiv*, arXiv:2401.16212.

McCabe, D. L., & Pavela, G. (1997). Ten principles of academic integrity for faculty. *The Journal of College and University Law*, *24*, 117–118.

McGee, R. W. (2023). *Is chat gpt biased against conservatives? an empirical study (February 15)*. Available online: https://ssrn.com/abstract=4359405 (accessed on 5 February 2025). [CrossRef]

MedCalc Software. (2018). *MEDCALC. Comparison of proportions calculator version 22.032*. MedCalc Software. Available online: https://www.medcalc.org/calc/comparison_of_proportions.php (accessed on 5 February 2025).

Merritt, E. 2023 January 25. *Chatting about museums with ChatGPT*. American Alliance of Museums. Available online: https://www.aam-us.org/2023/01/25/chatting-about-museums-with-chatgpt (accessed on 5 February 2025).

Metz, C. (2025, January 27). What is DeepSeek? And how is it upending A.I.? *The New York Times*. Available online: https://www.nytimes.com/2025/01/27/technology/what-is-deepseek-china-ai.html (accessed on 5 February 2025).

Millidge, B. (2023, July 23). LLMs confabulate not hallucinate. *Beren's Blog*. Available online: https://www.beren.io/2023-03-19-LLMs-confabulate-not-hallucinate (accessed on 5 February 2025).

Morocco-Clarke, A., Sodangi, F. A., & Momodu, F. (2024). The implications and effects of ChatGPT on academic scholarship and authorship: A death knell for original academic publications? *Information & Communications Technology Law*, *33*(1), 21–41.

Motoki, F., Pinho Neto, V., & Rodrigues, V. (2023). *More human than human: Measuring chatgpt political bias*. Available online: https://ssrn.com/abstract=4372349 (accessed on 5 February 2025).

Nicholson, J. M., Uppala, A., Sieber, M., Grabitz, P., Mordaunt, M., & Rife, S. C. (2021). Measuring the quality of scientific references in Wikipedia: An analysis of more than 115M citations to over 800 000 scientific articles. *The FEBS Journal*, *288*(14), 4242–4248. [CrossRef]

Onishi, T., Wang, H., Bansal, M., Gimpel, K., & McAllester, D. (2016). Who did what: A large-scale person-centered cloze dataset. *arXiv*, arXiv:1608.05457.

OpenAI. (2025). *Models*. Available online: https://platform.openai.com/docs/models (accessed on 4 February 2025).

Pascoe, B. (2014). *Dark emu black seeds: Agriculture or accident?* Magabala Books.

Qi, X., Zhu, Z., & Wu, B. (2023). The promise and peril of ChatGPT in geriatric nursing education: What we know and do not know. *Aging and Health Research*, *3*(2), 100136. [CrossRef]

Rao, A. S., Pang, M., Kim, J., Kamineni, M., Lie, W., Prasad, A. K., Landman, A., Dryer, K., & Succi, M. D. (2023). Assessing the utility of ChatGPT throughout the entire clinical workflow. *medRxiv, 2023*. [CrossRef]

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, *3*, 121–154. [CrossRef]

Rozado, D. (2023). The political biases of chatgpt. *Social Sciences*, *12*(3), 148. [CrossRef]

Rutinowski, J., Franke, S., Endendyk, J., Dormuth, I., & Pauly, M. (2023). The self-perception and political Biases of ChatGPT. *arXiv*, arXiv:2304.07333. [CrossRef]

Sarraju, A., Bruemmer, D., Van Iterson, E., Cho, L., Rodriguez, F., & Laffin, L. (2023). Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based Artificial Intelligence model. *Jama*, *329*(10), 842–844. [CrossRef]

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017, May 22–26). *Membership inference attacks against machine learning models*. 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA.

Sng, G. G. R., Tung, J. Y. M., Lim, D. Y. Z., & Bee, Y. M. (2023). Potential and pitfalls of ChatGPT and natural-language artificial intelligence models for diabetes education. *Diabetes Care*, *46*(5), e103–e105. [CrossRef]

Spennemann, D. H. R. (2023a). ChatGPT and the generation of digitally born "knowledge": How does a generative AI language model interpret cultural heritage values? *Knowledge*, *3*(3), 480–512. [CrossRef]

Spennemann, D. H. R. (2023b). Children of AI: A protocol for managing the born-digital ephemera spawned by Generative AI Language Models. *Publications*, *11*, 45. [CrossRef]

Spennemann, D. H. R. (2023c). Exhibiting the Heritage of Covid-19—A Conversation with ChatGPT. *Heritage*, *6*(8), 5732–5749. [CrossRef]

Spennemann, D. H. R. (2023d). Exploring ethical boundaries: Can ChatGPT be prompted to give advice on how to cheat in university assignments? *Preprint*, 1–14. [CrossRef]

Spennemann, D. H. R. (2023e). What has ChatGPT read? References and referencing of archaeological literature by a generative artificial intelligence application. *arXiv*, arXiv:2308.03301. [CrossRef]

Spennemann, D. H. R. (2023f). Will the age of generative Artificial Intelligence become an age of public ignorance? *Preprint*, 1–12. [CrossRef]

Spennemann, D. H. R. (2024). Will artificial intelligence affect how cultural heritage will be managed in the future? Conversations with four genAi models. *Heritage*, *7*(3), 1453–1471. [CrossRef]

Spennemann, D. H. R., Biles, J., Brown, L., Ireland, M. F., Longmore, L., Singh, C. J., Wallis, A., & Ward, C. (2024). ChatGPT giving advice on how to cheat in university assignments: How workable are its suggestions? *Interactive Technology and Smart Education*, *21*(4), 690–707. [CrossRef]

Surameery, N. M. S., & Shakor, M. Y. (2023). Use chat gpt to solve programming bugs. *International Journal of Information Technology & Computer Engineering (IJITC)*, *3*(01), 17–22.

Tirumala, K., Markosyan, A., Zettlemoyer, L., & Aghajanyan, A. (2022). Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, *35*, 38274–38290.

Trichopoulos, G., Konstantakis, M., Alexandridis, G., & Caridakis, G. (2023a). Large language models as recommendation systems in museums. *Electronics*, *12*, 3829. [CrossRef]

Trichopoulos, G., Konstantakis, M., Caridakis, G., Katifori, A., & Koukouli, M. (2023b). Crafting a museum guide using GPT4. *Bid Data and Cogntiive Computing*, *7*(3), 148. [CrossRef]

Wen, J., & Wang, W. (2023). The future of ChatGPT in academic research and publishing: A commentary for clinical and translational medicine. *Clinical and Translational Medicine*, *13*(3), e1207. [CrossRef]

Wylie, A. (2002). *Thinking from Things: Essays in the philosophy of archaeology*. University of California Press.