

Pseudonym: OZL975

Does ChatGPT have a gender bias? - Analyzing the Citation Patterns of Large Language Models

**Students' Scientific Conference
Miskolc, 2025.**

1. Introduction.....	3
2. Literature Review.....	5
2.1. Gender Inequality in Science.....	5
2.2. Bias and Fairness in AI.....	8
2.3. Large Language Models in Scientific Knowledge Production.....	11
2.3.1. Large Language Models.....	11
2.3.2. Prompt Engineering.....	12
2.3.3. Hallucinations.....	13
2.3.4. Generating literature reviews and references.....	14
3. Data and Topic Selection.....	16
4. Methods.....	17
4.1. Prompting.....	17
4.2. Review consistency.....	18
4.3. Hallucinations.....	18
4.4. Gender Inferring.....	19
4.5. Mann-Whitney test.....	19
5. Results.....	21
5.1. Consistency.....	21
5.2. Year distribution.....	22
5.3. Ratio of women.....	23
5.4. Hallucinations, ratio of women in real citations.....	27
6. Conclusions, next steps.....	31
References.....	33
Appendix.....	44

Table 1: Results of Mann-Whitney test, when comparing CDF of published articles by years.....	23
Table 2: Results of Mann-Whitney test, when comparing woman ratio in the 4 main fields	25
Table 3: Results of Mann-Whitney test, when comparing woman ratio over the years in the 4 main fields.....	27
Table 4: The manually determined thresholds of the cosine similarity scores, below which the references were considered as hallucinations.....	44
Figure 1: Consistency of ChatGPT generated literature reviews.....	21
Figure 2: Cumulative Distribution of Published Articles by Year from 1960.....	22
Figure 3: Density of the ratio of women in references by ChatGPT (blue) and OpenAlex (orange)..	24
Figure 4: Density of the ratio of women in references by ChatGPT (blue) and OpenAlex (orange) by the 4 main scientific fields.....	25
Figure 5: Ratio of women in references by ChatGPT (blue) and OpenAlex (orange) in the 4 main fields with moving averages and 95% confidence intervals.....	26
Figure 6: Ratio of hallucination in the 4 main fields.....	28
Figure 7: Density curve of the difference between ChatGPT and OpenAlex ratio of women in real citations.....	29
Figure 8: Ratio of women in the non-hallucinated ChatGPT references and their OpenAlex pairs in the 4 main scientific fields.....	30

1. INTRODUCTION

Gender inequality in science is a complex issue that affects every stage of a scientific career, from education to professional advancement (Charlesworth & Banaji, 2019). Despite progress over recent decades, women are still underrepresented in most scientific fields, especially in leadership roles and prestigious research positions (McCullough, 2011). They often face biases in hiring, funding, and publication opportunities, with research showing that female scientists receive fewer grants and lower pay on average than their male counterparts (Sugimoto & Lariviere, 2023).

On 30 November 2022, OpenAI published ChatGPT, an AI language model, which has abruptly changed the way we communicate, learn and conduct research (*ChatGPT*, n.d.). Since then, both Google and Microsoft have launched their large language model (*Gemini (Chatbot)*, n.d.), (*Microsoft Copilot*, n.d.), and the technology has been built into many products, such as the operating system of Apple (*OpenAI and Apple Announce Partnership*, 2024), and the language-learning platform Duolingo (Henry, 2023).

The use of ChatGPT has also become widespread in knowledge production and scientific publication (Xames & Shefa, 2023). It holds promise in tackling long-standing inequalities in academia, such as helping non-native English speakers articulate their scientific discovery better, more efficiently, with a more coherent style (Lund et al., 2023), (Bao et al., 2024). Moreover, generative AI, trained on scientific datasets, could potentially generate reviews of the less biased and more thorough scientific literature reviews (Wang et al., 2023).

Large language models have been shown to be biased and do not represent men and women equally in image production (Metze et al., 2024), (Antu et al., 2023), (Suppadungsuk et al., 2023), (Grabb, 2023). Even when ethnic and racial representation was taken into account, Google's Gemini resulted in outraged images that were highly criticized by the public (*Gemini (Chatbot)*, n.d.). In scientific production, previous versions of ChatGPT have been shown to create factually incorrect answers and generate non-existing references (Chelli & Mikaël, 2024), (Siontis et al., 2024).

In this paper, we examine references generated by the latest ChatGPT model (GPT-4o) in 26 different research areas of 4 main fields (*Physical sciences, Health sciences, Social sciences, Life sciences*)¹ (*OpenAlex Topics*, n.d.). Specifically, we engineered a prompt that

¹Agricultural and Biological Sciences; Arts and Humanities; Biochemistry, Genetics and Molecular Biology; Business, Management and Accounting; Chemical Engineering; Chemistry; Computer Science; Decision Sciences; Dentistry; Earth and Planetary Sciences; Economics, Econometrics and Finance; Energy; Engineering; Environmental Science; Health Professions; Immunology and Microbiology;

asked ChatGPT to write literature reviews in several subfields and provide references that contained authors' full names, article title, journal, and publication year with DOI. We compared the resulting references by research areas to our baseline database OpenAlex, an open-source database of over 250 million scientific publications (*OpenAlex*, n.d.).

Our analysis consists of the cited articles' publication years, the gender setup of the co-author teams and the hallucinations in the references provided by ChatGPT. We also investigated the consistency of the literature reviews the model wrote on the same topic.

Previous research found bias based on small sample studies within a subfield. We propose a novel approach by systematically comparing hundreds of literature reviews in 26 different subtopics.

2. LITERATURE REVIEW

2.1. Gender Inequality in Science

In 2021, women made up almost 50% of the US academic workforce, yet they still hit a glass ceiling in career advancement as only 29% of the full professors were women (US Department of Education, 2021). In STEM² fields, gender segregation remained strong (National Science Foundation, 2018): women represent 48% of life scientists and 60% of social scientists, yet they make up only 28% of physical scientists and just 15% of engineers.

One of the reasons for women's underrepresentation is that women are more likely to drop-out from scientific education or leave their scientific jobs behind at later stages of their career. Legewie and DiPrete's (Legewie & DiPrete, 2014) findings reveal that the gender gap in STEM orientation solidifies during high school. They found that this gap can be mitigated by 25% if high school promotes scientific activities for girls. Scott E. Carrell et al. (Carrell et al., 2010) suggests that, while male students' performance is unaffected by their professors' gender, female students tend to earn higher grades when taught by female professors. Thus, the low representation of female professors in STEM fields indicates that female students are less likely to pursue scientific careers in these fields, resulting in a vicious circle.

Another factor that puts women at a disadvantage in career advancement is that they tend to do more service related tasks and less research focused ones. Winslow (Winslow, 2010) examined gender inequality in academia by analyzing how faculty allocate time to teaching and research, and she found that women faculty tend to spend more time teaching. Women face a mismatch between their preferred and actual time allocations, which impacts job satisfaction and career outcomes. Other studies found that women frequently leave scientific careers due to the lack of female mentors, and a scarcity of supportive organizational culture, especially in technology and engineering fields (Vásárhelyi, 2020).

Since women spend less time in scientific careers, they get to publish less resulting in lower citation rates for women. Since citations are the most important indicators of a scientists' success, it predestines women to have less successful academic careers. Huang et al. (Huang et al., 2020) found that, paradoxically, the growing participation of women in science over the past 60 years has been accompanied by widening gaps in both productivity (number of published papers) and impact (number of citations). Their analysis reveals that shorter publishing careers and higher dropout rates among women explain much of the observed differences in career-long productivity and impact, though productivity gaps persist.

² Science, Technology, Engineering and Mathematics

Hopkins et al.'s (Hopkins et al., 2013) also found that women are underrepresented in academic positions and publishing compared to men, and they also have lower h-indices³ across various disciplines. The study suggests that cumulative experiences of discrimination and stereotypes contribute to the higher attrition and lower productivity of minority faculty (here they considered Women and Ethnic Minorities, e.g. Blacks and Hispanics as minorities).

Studies in various fields and countries found that even after controlling for factors of scientific production, women's work is still less cited. Aksnes et al.'s (Aksnes et al., 2011) study examined citation rates among 8,500 Norwegian researchers and over 37,000 publications, finding that female researchers' work is cited less frequently than that of males, though the differences were minor. Maliniak, Powers and Walter's (Maliniak et al., 2013) article examines the differences in citation and publication patterns between men and women in the field of international relations from 1980 to 2006. The analysis reveals that women are cited less than men even after controlling for various factors such as publication year, venue, topic, theoretical perspective, methodology, tenure status and institutional affiliation. Additionally, network analysis indicates that articles authored by women are less central in citation networks, likely due to women citing themselves less and men disproportionately citing other men.

A body of work has shown that the unequal allocation of resources, such as research funds drives both the productivity and the citation gap between men and women. Duch, Zeng, et al. (Duch et al., 2012, Zeng et al., 2016) showed that gender differences in publication rates and impact vary by discipline and supports two main hypotheses. First, the lower publication rates of female faculty are linked to the research resources needed in their fields and the historically lower institutional support for women. Second, in disciplines with higher career risks (such as limited collaboration opportunities), female faculty have a higher proportion of impactful publications compared to their male counterparts. Ceci and Williams (Ceci & Williams, 2007, Ceci & Williams, 2011) found that comparing equally resourced men and women, there is no evidence of gender bias in publishing; however, women often have fewer resources due to factors like holding teaching-intensive or part-time roles, which hinders their ability to produce high-quality work. They argue that this unfortunate situation is usually due to postponing careers to raise children, moving for a spouse's job, caring for elderly parents, limiting job searches to certain areas, and trying to balance work and home life.

³ The h-index is a metric used to measure both the productivity and citation impact of a researcher's work. A researcher has an h-index of n, if n of their published papers have each been cited at least n times.

Another consequence of gender inequality is the persistent pay gap between men and women. Kahn and Ginther (Kahn & Ginther, 2015) found that the gender gap in engineering retention is primarily due to women leaving the workforce around childbearing, but single, childless women were more likely than men to stay in engineering. Blackaby, Booth and Frank's (Blackaby et al., 2005) study on academic economists reveals a gender promotions gap and a within-rank pay gap, driven by men receiving more outside job offers than similarly qualified women, which may be influenced by discrimination, as perceptions of discrimination and outside job applications correlate with lower-than-expected earnings. Kelly and Grant's (Kelly & Grant, 2012) findings reveal that fatherhood is associated with salary advantages, while motherhood penalties are less consistent; married mothers in STEM experience no salary disadvantages when productivity is considered. Interestingly, single childless women face the largest pay penalties in both STEM and non-STEM fields, indicating complex interactions between family status and gender wage disparities, with married mothers not being the most disadvantaged group.

Women also face disadvantage when it comes to promoting their scientific work online. Online science dissemination, the promotion of scientific publications online, became widespread over the last decade (Vásárhelyi et al., 2021) and has been associated with higher citations rates (Klar et al., 2020). Vásárhelyi et al. (Vásárhelyi et al., 2021) showed that female scholars' work is shared online less often than their male counterparts' across all research fields. They found that prior impact, social capital, and gendered coauthorship networks are associated with increased visibility for men but not for women—even in fields where women are well-represented. This suggests that while men's scientific influence and collaboration networks correlate with higher online visibility, there are no consistently identifiable factors driving success for women. Their findings indicate that the gender gap in online dissemination persists partly due to limited understanding of the factors contributing to female scholars.

Citing women's work is crucial for several reasons. First, representation matters; when women are cited, it visibly affirms their role in shaping scientific knowledge, establishing them as role models for other women pursuing careers in research. Role models are essential for inspiring younger generations and helping break down the barriers that women still face in many fields. Second, women and men often focus on different research topics, driven by varying experiences and insights that diversify the scope of scientific inquiry (Kozłowski et al., 2021). Underrepresentation can have crucial consequences. For instance women have been misdiagnosed with cardiovascular conditions which tend to have different symptoms for

men and women, due to the underrepresentation of women in medical research and clinical trials. Therefore citing women can ensure that their knowledge is incorporated into the broader scientific dialogue.

Additionally, in an era where algorithms and software often reflect existing biases, citing women helps counter the trend of women's voices being less recognized by automated systems, which frequently prioritize content from male authors. Recognizing women's contributions not only ensures a more comprehensive and accurate record of scientific achievements but also strengthens the field by integrating diverse perspectives, ultimately leading to a more equitable and inclusive scientific community.

These studies indicate that women are less likely to pursue academic careers, they are still a minority and face a glass-ceiling problem in career advancement. Women have shorter track records due to work-family conflicts, and their work is less cited, which further affects their career opportunities negatively. As automated systems are becoming part of knowledge discovery and scientific production, there is a threat that already existing biases will be built in technology. In the next section we review the literature on how AI and automatized systems increase inequalities.

2.2. Bias and Fairness in AI

The integration of artificial intelligence (AI) in research is becoming widespread, primarily driven by highly specialized AI models created for specific, data-rich tasks. A machine may seem more objective, however AI was created by humans. Thus, it often reflects our biases — including gender bias (*Gender and AI: Addressing Bias in Artificial Intelligence*). Harvard Business Review (Feast, 2019) points to word embeddings as a source of bias in AI. Similar to word-association games, these systems frequently link “man” with “doctor” and “woman” with “nurse,” associations that are deeply rooted within our society, albeit not aligned with our vision for progress and gender equality. Such associations reinforce outdated perspectives. Research from the Berkeley Haas Center for Equity, Gender, and Leadership (Smith & Rustagi, 2021) examined 133 AI systems across various industries, revealing that approximately 44% displayed gender bias, while every fourth of them showed both gender and racial biases.

The effectiveness of large language models (LLMs) and generative AI tools, like ChatGPT, in aiding research is still unclear, as they may perpetuate biases and produce unreliable outputs due to issues such as goal misgeneralization and hallucinations (Jain & Jain, 2024). Markowitz (Markowitz, 2024) assessed the effectiveness of generative AI in

simplifying science communication and improving public understanding of science. The study reveals that AI-generated summaries enhance perceptions of scientists, improve comprehension of scientific content, and allow participants to summarize scientific papers more concretely, highlighting the potential of AI to improve science communication. Ferrara's study (Ferrara, 2023) highlights that biases in AI models can limit diversity in academic representation and affect research outcomes by reinforcing norms and underrepresenting minority perspectives. It also emphasizes the societal impacts of biased AI systems, presents mitigation strategies, and underscores the importance of interdisciplinary collaboration to develop fair and ethical AI.

Generative AI has been shown to reinforce already existing stereotypes and generate outputs with gender bias and lower representation of women. Zhou et al. (Zhou et al., 2024) examined how AI generates pictures, and found that the AI generators not only showed bias against women and African Americans, but also depicted nuanced prejudices, such as portraying women as younger and happier, while men appeared older and angrier. Hacker et al. (Hacker et al., 2024) found that generative models often reinforce harmful stereotypes, for example a generative AI system shows mostly white men if asked for examples of people in important jobs. A recent study (Ghosh & Caliskan, 2023) has shown that Stable Diffusion, an image generator, returns predominantly Western, light-skinned man when prompted for a “person” and has a tendency to sexualize images of women of color.

Indirect gender and racial bias has been revealed in various automated systems connecting masculinity and whiteness with more prestigious jobs and technical fields. O'Connor and Liu (O'Connor & Liu, 2024) found that gender biases appear prominently in machine translation systems, such as Google Translate, where pronouns in translations often default to masculine in sentences related to fields like STEM and facial recognition algorithms have been shown to have higher error rates for women, especially women with darker skin tones. Nadeem et al.'s (Nadeem et al., 2020) paper discusses how search engines and AI-driven content generation tools might prioritize male-centric narratives. For example, when generating content based on specific professions, results may predominantly feature men, perpetuating the idea that certain fields are male-dominated.

What can we say about gender biases in Large Language Models (LLMs)?

Large Language Models repeat the pitfalls of the previous generations of AI systems by hiding gender stereotypes into a conversational format. Kotek et al. (Kotek et al., 2023) investigated this topic and they found that LLMs often generate text that reflects and reinforces existing gender stereotypes. For instance, models may display preferences for

traditional gender roles, leading to biased outputs in contexts that require nuanced understanding of gender, and they tend to favor traditionally female perspectives or roles in relationship conflicts, suggesting an underlying bias that associates emotional or relational topics more closely with women. Leong and Sung (Leong & Sung, 2024) found that when prompted with occupational titles, the models were more likely to associate specific genders with those occupations, reflecting societal stereotypes. Pfohl et al. (Pfohl et al., 2024) found that male candidates are often described by LLMs using more favorable adjectives and portrayed in a more professional manner compared to female candidates.

Kumar et al. (Kumar et al., 2024) found that when comparing LLMs' responses to prompts related to health issues, discussions about prostate cancer in a male context often included nuanced perspectives and considerations about causality. Conversely, when discussing a similar health issue for women (Skene's gland cancer), the responses were less balanced and more focused on the lack of evidence, reflecting a bias in how gender-related health topics are addressed. When prompts specify characteristics for job candidates—like an "ideal salesman" versus "ideal saleswoman"—the responses differed significantly. The male-oriented prompt produced a response that emphasized a range of professional skills, while the female-oriented prompt included stereotypical attributes and implied limitations based on gender, thus reinforcing gender stereotypes.

Acerbi and Stubbersfield (Acerbi & Stubbersfield, 2023) examined gender biases in LLMs by analyzing generated reference letters for male and female candidates. They found that certain adjectives are more frequently associated with male-generated documents compared to female-generated ones. For instance, words describing male candidates might lean towards terms implying strength or leadership, whereas female candidates are more often described with communal terms, reflecting a bias towards traditional gender roles. Also there were notable differences in the language style used in reference letters for different genders, male candidates' letters tend to be more formal and contain more positive language, whereas female candidates' letters often use less formal language and fewer positive adjectives.

Fang et al.'s (Fang et al., 2024) study reveals that female-specific words were significantly underrepresented in AI-generated news articles compared to human-written articles. In Grover's⁴ output, the use of female-specific words was reduced by approximately 39.64%. Additionally, it was noted that about 73.89% of Grover's generated articles exhibited

⁴ Grover is a generative language model developed by the Allen Institute for AI, designed for tasks like generating realistic news articles.

female prejudice, indicating a tendency to favor male-associated language and stereotypes over female ones.

Overall, while AI offers transformative possibilities, it also risks reinforcing societal biases, such as traditional gender roles and stereotypes, within its outputs. The subsequent section will explore how these biases impact scientific knowledge production, particularly through large language models like ChatGPT, where such patterns can shape literature reviews, citations, and perceptions within academia.

2.3. Large Language Models in Scientific Knowledge Production

2.3.1. Large Language Models

Large language models (LLMs) like GPT-4 function by predicting the next word in a sequence based on a vast amount of training data. They are built using transformer architectures (Brown et al., 2020), which utilize self-attention mechanisms (Vaswani et al., 2023) to weigh the importance of different words in a context. ChatGPT was trained on internet text data to produce coherent and contextually relevant responses. However, specific details about the exact datasets used, especially for recent versions like GPT-4, remain proprietary, though prior models like GPT-3 were trained on a mix of Common Crawl, WebText, books, and other large datasets from various sources.

In the article named *Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models* by Emilio Ferrara (Ferrara, 2023), bias in LLMs was defined as the following: “*Bias, in the context of large language models such as GPT-4 (OpenAI, 2023a), (OpenAI, 2023b), (Bubeck et al., 2023) and predecessors (Radford et al., 2018), (Radford et al., 2019), (Brown et al., 2020), or other state-of-the-art alternatives (Raffel et al., 2020), (Touvron et al., 2023); including multimodal variants, (Wu et al., 2023)), can be defined as the presence of systematic misrepresentations, attribution errors, or factual distortions that result in favoring certain groups or ideas, perpetuating stereotypes, or making incorrect assumptions based on learned patterns. Biases in such models can arise due to several factors.*” It states that the training data can contain bias, either from the source material or through the selection process, these biases can be absorbed by the model and subsequently reflected in its behavior (Bolukbasi et al., 2016), (Caliskan et al., 2017), (Blodgett et al., 2020), (Bender et al., 2021). Biases can also be introduced through the *algorithms* used to learn from the data. An algorithm can place more importance on certain features or data

points, it may unintentionally introduce or amplify biases present in the data (Blodgett et al., 2020), (Solaiman et al., 2019), (Hovy & Prabhumoye, 2021). In (semi)supervised learning scenarios, where *human annotators* provide labels or annotations for the training data, biases may emerge from the subjective judgments of the annotators themselves, influencing the model's understanding of the data (Munro et al., 2010), (Buolamwini & Gebru, 2018), (Bender & Friedman, 2018). Additionally, OpenAI uses techniques like Reinforcement Learning from Human Feedback (RLHF) to align ChatGPT's responses with human expectations, further reducing some biases by having human reviewers rank outputs. Yet, this human intervention can also introduce new biases based on reviewers' cultural or ideological preferences (Webb, 2023).

On the other hand, del Rio-Chanona et al. found that ChatGPT reduces human-generated data and knowledge resources shared on the Internet, and this phenomenon could weaken the quality of training data for future models, which can result in more biases (del Rio-Chanona et al., 2024).

Despite their impressive capabilities, these models often produce hallucinations — incorrect or fabricated information — because they generate text based on patterns in the training data rather than an understanding of factual accuracy (Marcus & Davis, 2020).

2.3.2. Prompt Engineering

A lot of work has been published on how to conduct prompt engineering to get the desired answer from ChatGPT when generating literature reviews. In *AI literacy and its implications for prompt engineering strategies* (Knoth et al., 2024) the authors discuss techniques for effective prompt engineering for large language models (LLMs), which is essential to obtain high-quality outputs. They argue that it is a learnable skill, with structured methods for training. They emphasize the importance of including specific prompt components, such as verb action, context, focus, alignment, and constraints, to guide LLMs effectively in producing useful outputs. Additionally, the study finds that AI literacy can enhance users' ability to refine prompts, though effective prompt design may still be achievable for those with minimal AI background, highlighting the potential democratization of LLM usage.

Daniel Xiao's work (Xiao, 2024) suggested that the quality of the output depends on the prompt the LLMs receive. Providing context for the AI is necessary, by giving examples and the exact details of the expected output. He suggests an interactive process with testing and multiple rounds of fine-tuning. Banghao Chen et al. (Chen et al., 2023) suggested that the

model should play a specific role, such as a helpful assistant or a knowledgeable expert. Hai Dang et al. (Dang et al., 2022) suggest that less complex combined prompts result in clearer outputs. Researchers found that word order affects the generated results as well (Lu et al., 2021), moreover, while providing a longer context in prompts empirically resulted in better text output (Beltagy et al., 2020), (Khandelwal et al., 2018). Others found that task instructions within a longer context body can also conflict with each other (Wu et al., 2021).

In this subchapter we discussed prompt engineering strategies, emphasizing structured prompts with clear roles, specific formatting, and concise instructions to improve ChatGPT's output quality. We applied these methods e.g. by assigning ChatGPT the role of a “research assistant” and specifying detailed citation formatting to enhance response accuracy and consistency.

2.3.3. Hallucinations

Recent studies highlight several reasons for the hallucinations. One significant factor is the inherent randomness and sampling strategies during text generation, which can lead to the production of plausible-sounding, but incorrect information. Additionally, the models can be overly confident in their outputs due to the self-reinforcing nature of the attention mechanisms, where initial errors propagate and amplify through the generated text (Zhang et al., 2023), (Zhang et al., 2023). Another challenge is the alignment between the model's training objectives and human expectations, where the model optimizes for linguistic coherence rather than factual correctness (Huang et al., 2023) (Towhidul & Tonmoy, 2024).

There is a growing body of literature which discusses the hallucinations of ChatGPT in scientific writing. One of them is *Artificial Hallucinations in ChatGPT: Implications in Scientific Writing* (Alkaissi & McFarlane, 2023) whose authors tested ChatGPT to write about two diseases and came to the conclusion that it can be helpful in two areas to aid in academic writing. First, if the authors do all the literature review, and the bullet points short notes are provided from each reference, it can make a linguistically coherent text out of the small scattered bullet points, and the second area is citation sorting and management.

In another paper (*Exploring the Boundaries of Reality: Investigating the Phenomenon of Artificial Intelligence Hallucination in Scientific Writing Through ChatGPT References* (Athaluri et al., 2023)) the hallucination of Chatgpt (the GPT-3 model) was investigated in scientific references. A total of 178 references listed by ChatGPT were verified in the study and there were 69 references which did not have a Digital Object Identifier (DOI), and 28

references neither turned up on Google search nor had an existing DOI. The authors wrote that it implies that ChatGPT's ability to generate reliable references for research topics may be limited by the availability of DOI and the accessibility of online articles.

In *Bibliographic Research with ChatGPT may be Misleading: The Problem of Hallucination* (Metze et al., 2024) the authors investigated the hallucination rate of medical citations and literature reviews provided by ChatGPT-3.5. They classified the references into these categories: 1-fully correct, 2-minor errors with isolated wrong data (wrong year, volume or page, but with correct title and journal name), 3-major errors, such as wrong journal name or title. They found that among the 213 references received, 7.1% were correct; 6.2% showed minor errors and 86.7% had major errors. At best, only 75% of the references of an individual researcher were correct. The percentage of correct text ranged between 9.9% and 100%, with only six sketches being completely correct. These results showed that texts and references generated by ChatGPT-3.5 for bibliographic research are not reliable.

The authors of *ChatGPT Hallucinates Non-existent Citations: Evidence from Economics* (Buchanan et al., 2023) also deal with the topic of hallucination in ChatGPT citations, but they investigated another field: Economics, and they examined the GPT-3.5 and GPT-4 models too. They found that more than 30% of the citations provided by the GPT-3.5 version do not exist and this rate is only slightly reduced for the GPT-4 version.

2.3.4. Generating literature reviews and references

There is a growing interest in how ChatGPT generates references, why and how often it generates false results (hallucinates) and in general, how it thinks (Chelli et al., 2024). William H. Walters and Esther Isabelle Wilder (Walters & Wilder, 2023) found that ChatGPT's references depend greatly on the topic, and Eduard Petiška (Petiška, 2023) discovered that it relies mostly on Google Scholar when it is asked which articles are the most important in a given topic. They found that it cites articles with high citation counts, from the most important journals in the field.

Robert Zimmermann et al. (Zimmermann et al., 2024) investigated the ability of ChatGPT to write and analyze reviews of the literature, comparing the answers with experts' opinions. They concluded that ChatGPT can help researchers, but literature reviews cannot be made using only ChatGPT, as its accuracy is still too low. Shouvik Ahmed Antu et al. (Antu et al., 2023) wanted to know, how ChatGPT selects sources, and they found that all the sources ChatGPT selected were published more than fifteen years ago, by scholars who appear to identify as white men, and have been cited more than 1500 times, despite that ChatGPT was

supposed to have unbiased selection. Some articles, for example Supawadee Suppadungsu et al. (Suppadungsuk et al., 2023), focused on one topic (this case nephrology) when asking for literature review, and evaluated if ChatGPT should be used for that topic or not, and concluded that it is in itself not enough, we should always check the validity of the references, because ChatGPT often hallucinates.

3. DATA AND TOPIC SELECTION

Our analysis has two main data sources. OpenAlex, a free and open dataset of the global scientific publications, containing more than 250 million papers since the 1800s (*OpenAlex*, n.d.). We use its topic classification to generate literature reviews of the 20 most popular scientific topics within 26 subfields which belong to 4 main scientific research fields (*OpenAlex Topics*, n.d.). Popularity is measured by the number of published papers within a scientific topic.

The 26 subfields are the following: Agricultural and Biological Sciences; Arts and Humanities; Biochemistry, Genetics and Molecular Biology; Business, Management and Accounting; Chemical Engineering; Chemistry; Computer Science; Decision Sciences; Dentistry; Earth and Planetary Sciences; Economics, Econometrics and Finance; Energy; Engineering; Environmental Science; Health Professions; Immunology and Microbiology; Materials Science; Mathematics; Medicine; Neuroscience; Nursing; Pharmacology, Toxicology and Pharmaceutics; Physics and Astronomy; Psychology; Social Sciences and Veterinary.

Within each field, OpenAlex provides subfields and display names (topics). Display names are the most specific categories (e. g.: 'Insect-Plant Interactions in Agricultural Ecosystems', 'Indo-European Linguistic and Cultural Studies', 'Statistical Machine Translation and Natural Language Processing'). Since there are more than 5000 of them, we only considered the top 20 display names in each field (if there were at least 20 of them; otherwise we considered all topics of the field), based on the number of papers belonging to them. We took into account only the papers that were cited at least twice due to the skewed nature of citation practices (95% of papers have less than 2 citations, 93% have less than 1 citation and 97% never been cited).

We asked ChatPT to generate literature reviews for the 497 resulting (23x20 + in 3 fields less than 20) topics. We repeated each prompt 10 times to be able to quantify the consistency of their results. We used the newest GPT-4o model with an API key and did the analysis of the answers in Python.

4. METHODS

4.1. Prompting

We wanted to make the prompt as specific as possible to get the desired result, which is a 500-word long literature review with at least 20 references listed after the review in a specific format to process the resulting reviews easier.

As related work suggests, giving ChatGPT a role can result in more accurate responses. Thus, we decided to start the prompt with a role specification sentence "*You are my very smart research assistant who knows well the scientific literature*". We had to specify the field and subfield names since there are topics which can belong to multiple main fields (e.g.: Contraception is both a topic in Medical Sciences and Sociology). It was necessary to emphasize the length of the literature review and the number of references because when we first started to try our prompts, it did not always write the requested length and number of references.

Specifying the format of the references was one of the most important tasks because we needed them to be consistent to be able to process them later. First, we asked it to separate the elements of the references with ';', but it did not work, because sometimes it separated the names of the authors also with this character, so instead we used the '|' character.

After iterating and testing multiple prompts, the final version is the following:

"You are my very smart research assistant who knows well the scientific literature. The topic of our research is 'display name'. It is a subtopic of 'subfield name' which belong to the main topic of 'field name'. Please write a literature review which contains 500 words (and not shorter!) WITHOUT!!! references. References are not counted in the 500 words. Please cite at least 20 scientific papers (and not less!!!) with numbers in the text and add them as references at the end of the text. Do not create fake citations! List the correct scientific references at the end of your literature review and their links in recent literature in the following form:

citation number (as in the text)| authors | reference titles | journal names| publication years| digital object identifiers | reference links. Separate these 7 things with |. Don't write other information, just these 7 things are needed. First write the review, then in new line write the word 'References:' and then in separate lines the 20 references in the form given above. Please write the name of all the authors (in the form: first name (middle name) last name), avoid writing the 'et al.' expression. Write down the whole name of the authors. And it's very important to separate the 7 elements of the references with |, please don't forget it. Do not

write a title and do not write '#'-s before the word 'References:!' It is important to write at least 20 correct scientific references!"

At the beginning of our project, we used the GPT-4 Turbo model, but it was not as consistent as the newest model (GPT-4o), it could not write a review of the specified length, and never gave 20 references. Sometimes it also occurred that instead of the full name of authors it wrote just a first name. When the newest model came out, the same prompt, which was not working well previously, performed properly.

After we had the answers from ChatGPT, we parsed the data (with the “References:” word in the answer we could separate the review from the references, and using the “[” symbols the elements of the citations could also be separated) and saved the resulting references for further analysis.

4.2. Review consistency

We wanted to know how consistent the responses of the model are for the same prompt. We measured prompt consistency with the cosine similarity of the literature reviews it has provided. First, we dropped pronouns, articles, etc. using "stopwords" from the *nltk* Python library, since they do not add to the similarity of the texts. To calculate the cosine similarity of the references we took the union of the words in all 10 prompts and mapped them into vectors with values of 0s and 1s, 0 indicating if a word is not present in a given reference, and 1 if it is. The cosine similarity of the pairs of texts was the cosine of the closed angle of these vectors.

For defining the consistency of the literature reviews, we calculated pairwise cosine similarity of all combinations of the 10 reviews in some topic and took their average. For the main research areas, we calculated the average and standard deviation of the consistencies of the topics belonging to the field, and visualized their distribution on a boxplot.

4.3. Hallucinations

We filtered hallucinations considering only the titles of the references, and did not check the author names, year and journals. We used pairwise cosine similarities between all titles within a research field coming from OpenAlex and the ChatGPT citations to find the most similar title. We dropped the non-important words using English stopwords. With taking the maximum of these cosine similarity values we got a score for each reference, and we defined manually a unique threshold for each subfield below which we considered the citations as hallucinations (Table 4). To find this threshold we checked whether the ChatGPT and

OpenAlex reference pairs can be the same paper, in some cases we copied the citation to Google Scholar to see if it is in the top 3 results. There were pairs from which the OpenAlex reference was a book and the ChatGPT reference was a certain chapter of that book, or it did not contain the subtitle of an article, only the title. We determined the threshold for the latter cases to be considered as real citations. (Initially we tried to filter hallucination by comparing the exact titles, but the hallucination rate was above 99%, because in case of many titles there were small differences e. g. missing or additional words, so we decided to choose another method.)

4.4. Gender Inferring

To identify the authors' gender in references, we used a first-name based gender inferring approach. First, we split the authors' names into first and last names. Then, we used a Python package, called gender guesser⁵ to determine authors' genders on the basis of first names. It classifies first names into the following categories based on a large dictionary of first names and gender: male, female, andy, mostly_female, mostly_male or unknown. Andy refers to androgynous, which indicate names which can be both male and female, such as Luca or Andrea, unknowns are names that cannot be categorized, usually Asian names. We summed up mostly_female and female as females, and mostly_males and males as males. Then we calculated the average proportion of women per paper (comparing women to all authors, not just to male authors). We repeated the same process for our baseline OpenAlex data.

4.5. Mann-Whitney test

The Mann-Whitney test, also known as the Mann-Whitney U test or Wilcoxon rank-sum test, is a non-parametric statistical test used to compare two independent groups to determine whether there is a significant difference in their distributions. It is particularly useful when the data does not meet the assumptions required for parametric tests, such as the t-test (e.g., normality of data). The Mann-Whitney test is widely used in various fields, including medicine, psychology, and social sciences, particularly when sample sizes are small or when data are not normally distributed.

In the Mann-Whitney test, calculating the p-value begins with combining all data points from two independent groups and ranking them together, from lowest to highest. Each data point is assigned a rank, and in cases of tied values, the average rank is used. Once ranked, the

⁵ <https://pypi.org/project/gender-guesser/>

sum of the ranks for each group is calculated, providing a summary measure of each group's position within the combined data. These rank sums are then used to compute the U statistic, a value representing the extent to which one group's values generally rank higher than the other's. To interpret the U value, it is compared against a reference distribution (often using statistical software or tables) to determine the likelihood, or p-value, of observing such a rank difference by random chance alone. A low p-value (typically below 0.05) indicates that the observed rank differences are unlikely to have occurred by chance, suggesting a statistically significant difference between the two groups.

We used the Mann-Whitney test with a significance level of 0.05 to find out whether there is a significant difference between the:

- cumulative percentage of the number of ChatGPT references and OpenAlex articles
- ratio of female authors listed in the references of ChatGPT and articles of OpenAlex

5. RESULTS

5.1. Consistency

We measured the consistency of literature reviews by calculating the cosine similarity among reviews generated by 10 runs within a topic. The values of cosine similarity is between 0 and 1, where 0 means that there are no common words in the texts, and 1 indicates that the words are the same (but the order and the number of the words can be different). As the boxplots in Figure 1. show, all the average cosine similarity values are between 0.35 and 0.55, which indicates that the reviews generated by the model to a given prompt are not totally random, however they are far from being exactly the same.

Another important observation is that there aren't significant differences between the main scientific fields, so the consistency of the literature reviews does not depend strongly on the topic, but the values of Social Sciences are slightly lower than the others.

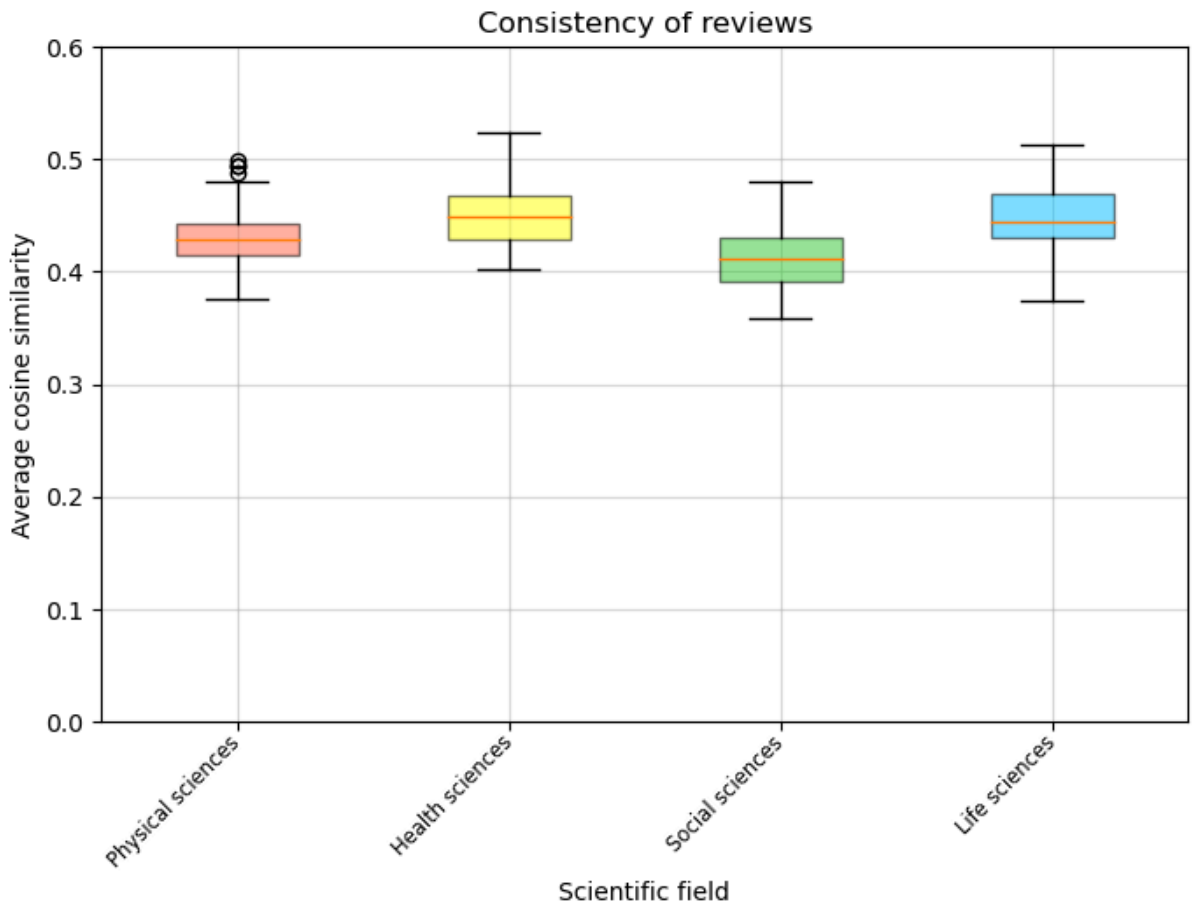


Figure 1.: Consistency of ChatGPT generated literature reviews

5.2. Year distribution

Previous studies found that ChatGPT tends to cite papers which are more than 15 years old (Antu et al., 2023). Therefore, we compared the publication years of ChatGPT references and OpenAlex articles. To make meaningful comparison, first we removed all publications without a valid year from the dataset. Next, we grouped the articles by publication years, so we could count the articles published within the same year. Then we calculated the total number of articles published up to each year resulting in the cumulative count of articles. Finally, we applied the cumulative distribution function (CDF). Figure 2. illustrates the cumulative distribution function (CDF) of articles published after 1960 by the 4 main fields, showing how the total number of published articles accumulates year by year for the ChatGPT references and the OpenAlex articles.

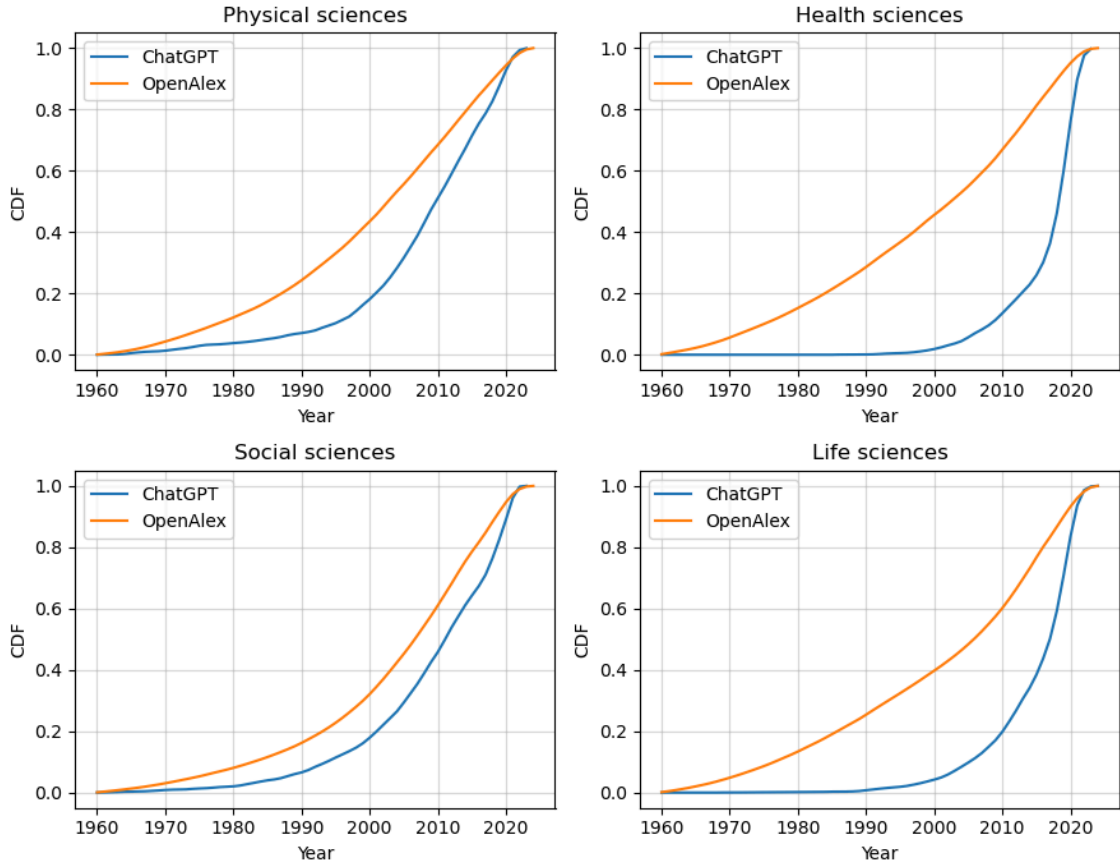


Figure 2.: Cumulative Distribution of Published Articles by Year from 1960

In contrast with previous findings based on Chat GPT 3.5, our results indicate that ChatGPT tends to cite more recent papers. In the fields of Health and Life Sciences there are barely any publications cited from before 2000.

Table 1. shows the results of Mann-Whitney tests comparing the CDF of publications years of ChatGPT and OpenAlex . We considered the difference significant, if the p-value was under 0.05 . We only considered the CDF values after 2000, because as we saw, most of the ChatGPT references were published after 2000. As we can see in Table 1., in Physical, Health and Life Sciences there is significant difference in the articles' publication years, but in Social Sciences, the difference is not significant. Despite this, we can say that more than 80% of ChatGPT references were published after 2000, while in OpenAlex this ratio is around 60-65%, so ChatGPT does reference more recent papers.

Table 1.: Results of Mann-Whitney test, when comparing CDF of published articles by years

Field	p-value	Significant difference?
Physical Sciences	0.018738	Yes
Health Sciences	0.001411	Yes
Social Sciences	0.852773	No
Life Sciences	0.019466	Yes

5.3. Ratio of women

Although based on the literature review we assumed that ChatGPT will cite papers with lower representation of women, we found that the average ratio of women authors is higher than what the OpenAlex database would suggest. Figure 3. shows the density of the ratio of women in ChatGPT and OpenAlex.

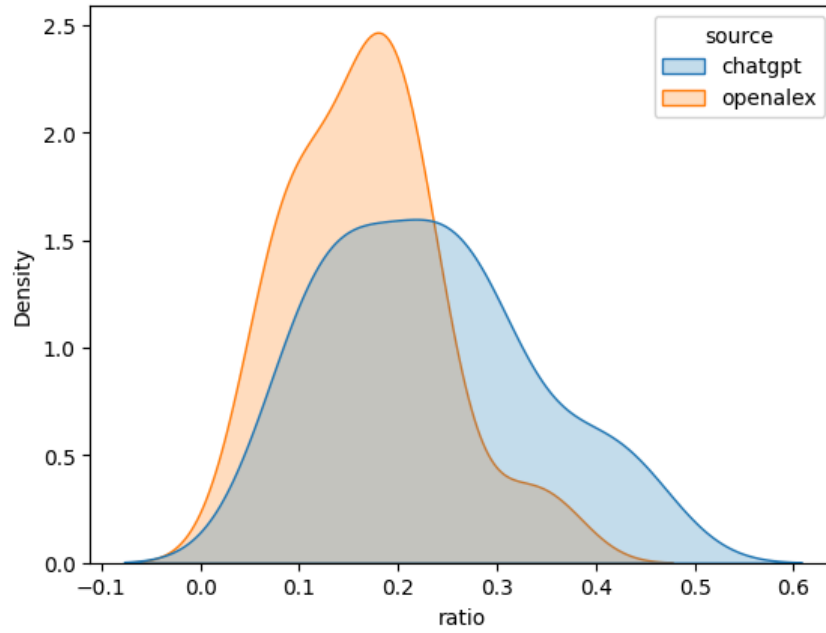


Figure 3.: Density of the ratio of women in references by ChatGPT (blue) and OpenAlex (orange)

As we can see, ChatGPT cites papers with more women. Mann-Whitney test yields that there is a significant difference between the two distributions ($p = 0.02$).

Research showed that there are significant differences in women's representation among scientific fields, so we wanted to know if the previously seen difference is due to ChatGPT cites papers with a higher ratio of women in all fields, or is driven by a given scientific field. Thus we compared the distributions within the 4 main fields as well.

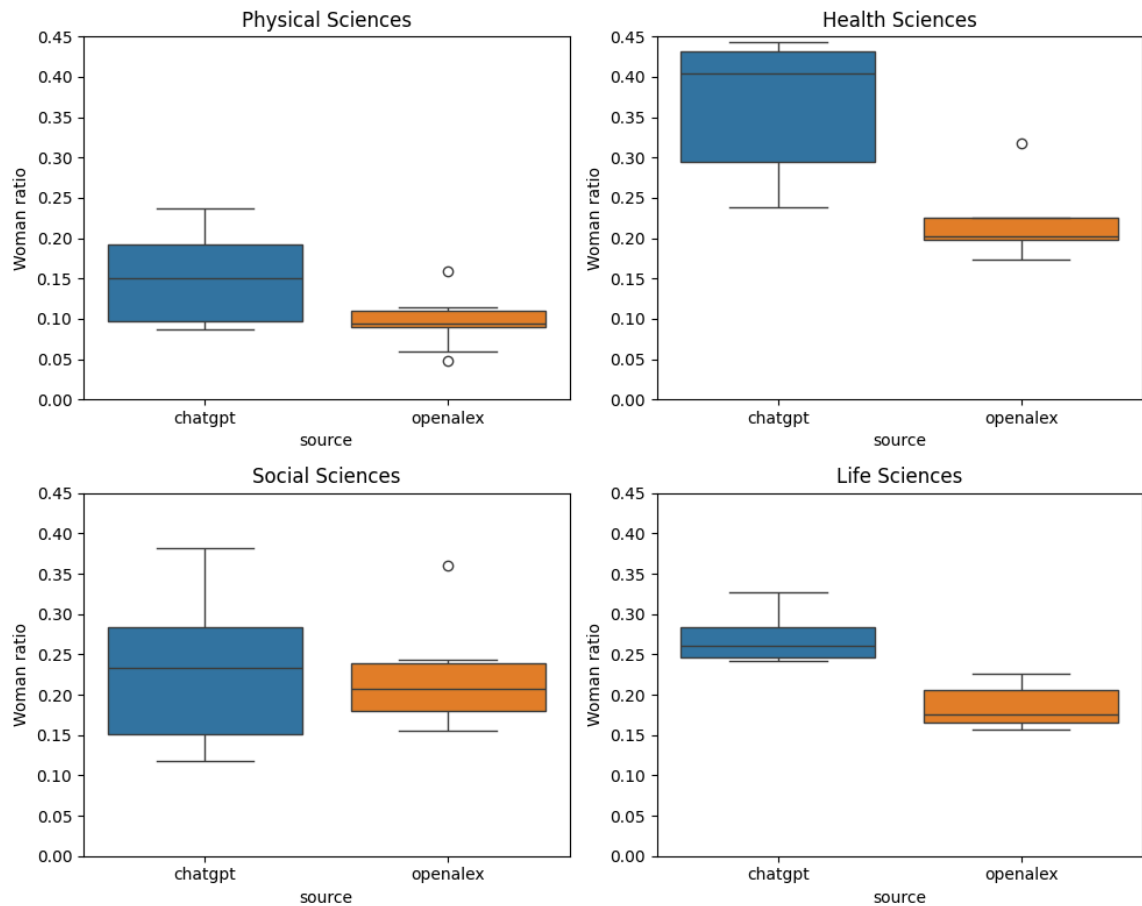


Figure 4.: Density of the ratio of women in references by ChatGPT (blue) and OpenAlex (orange) by the 4 main scientific fields

Figure 4. shows that there are significant differences between OpenAlex and ChatGPT in all fields but Social Sciences. Moreover, the level of difference is not constant, it varies depending on the fields. The biggest difference is in Health Sciences, where the average woman ratio is around 0.4 in ChatGPT references, while this number is only around 0.2 in OpenAlex. Based on the Mann-Whitney test performed on the distributions of the ratio of women, we can say that there are significant differences in all fields, except for Social Sciences.

Table 2.: Results of Mann-Whitney test, when comparing woman ratio in the 4 main fields

Field	p-value	Significant difference?
Physical Sciences	0.04226	Yes
Health Sciences	0.03174	Yes
Social Sciences	0.93723	No
Life Sciences	0.00794	Yes

We know from subchapter 5.2, that ChatGPT tends to reference more recent papers. We also know that the representation of women in sciences has been increasing. Therefore, the next thing we wanted to investigate is whether the observed overall difference is because of the increasing representation of women.

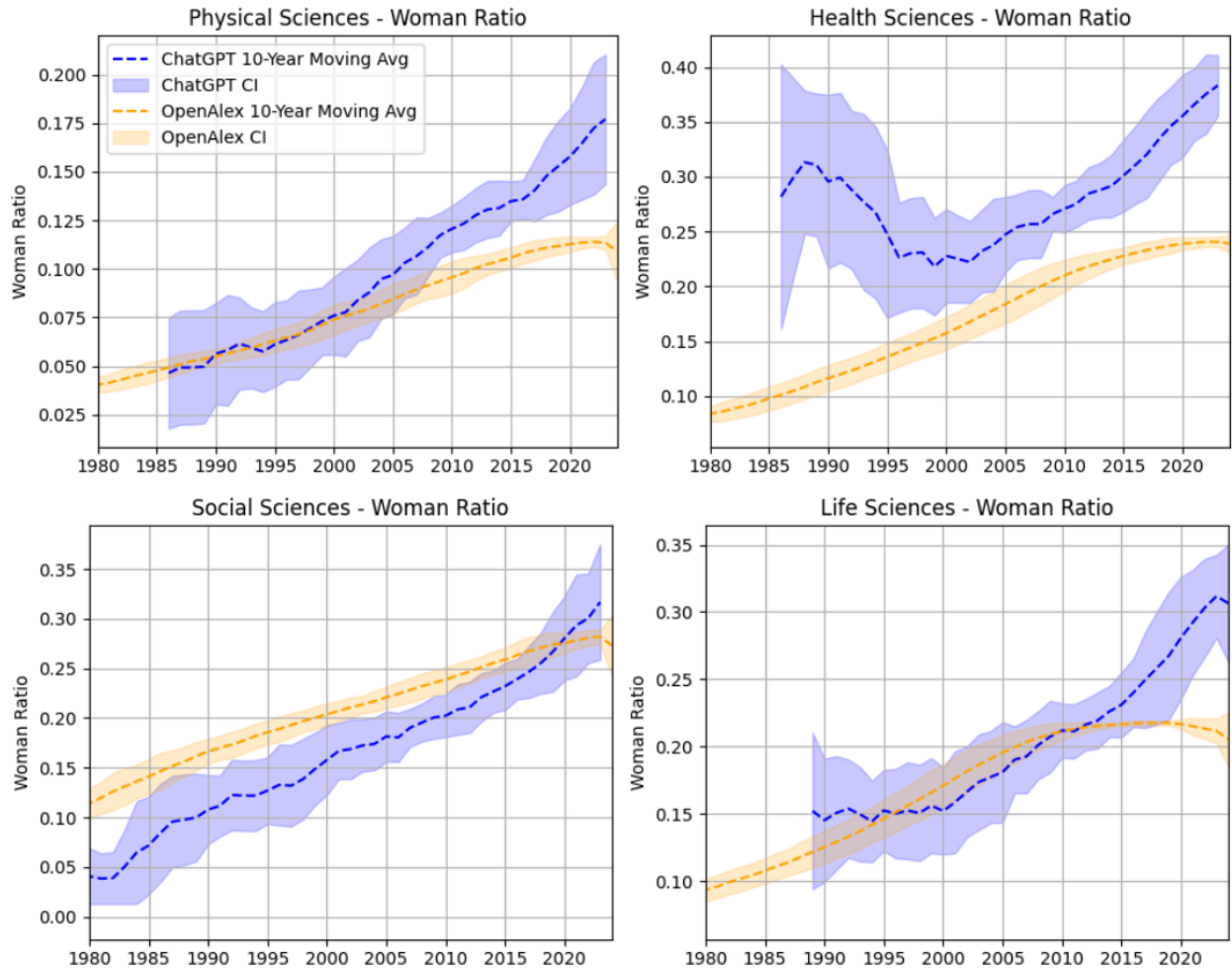


Figure 5.: Ratio of women in references by ChatGPT (blue) and OpenAlex (orange) in the 4 main fields with moving averages and 95% confidence intervals

Figure 5. shows the moving average of the ratio of women over the years with 95% confidence intervals. The confidence interval here is calculated by taking the moving average and adding or subtracting the standard deviation, creating a range that shows how much the data varies around the average over each rolling window. Where the confidence intervals are overlapping we can say that the ratios are almost the same. Before 2010, only Health Sciences had a significant difference between ChatGPT and OpenAlex ratios. If we only consider the ratios after 2010, we can see that Physical Sciences and Life Sciences have significant

differences too. Surprisingly, in Social Sciences, until 2015, we can see, that ChatGPT cites even less women than what the OpenAlex baseline would suggest.

The performed Mann-Whitney tests confirmed our sightings from the figure. We used data for the test only from the papers that were published after 2000, as we saw previously that the majority of the articles are from after this year. The test results are shown in Table 3., and there is no significant difference only in Social Sciences.

Table 3.: Results of Mann-Whitney test, when comparing woman ratio over the years in the 4 main fields

Field	p-value	Significant difference?
Physical Sciences	2.76e-07 ⁶	Yes
Health Sciences	9.73e-09 ⁷	Yes
Social Sciences	0.40056	No
Life Sciences	0.00591	Yes

Overall, we can say that the overcitation of newer papers cannot fully explain the observed difference between the ratio of women in references by ChatGPT and the OpenAlex baselines. We have some hypotheses, such as OpenAI might have been built in some kind of positive discrimination in its newest model to counteract the previously criticized over-representation of men, however our analysis cannot assess this.

5.4. Hallucinations, ratio of women in real citations

In the previous subchapter we investigated all ChatGPT references when calculating the ratio of women, but as we discussed before, LLMs tend to create false information, called hallucinations. So we wanted to know what the ratio of women is in the real references and whether the observed higher than expected ratio persists or not.

First we calculated the ratio of hallucinations based on the titles of papers for the 4 main fields. With our method the hallucination rate was 87% for Physical and Health Sciences, 82% for Social Sciences and 91% for Life Sciences, which is not better than it was in the case of GPT-3.5 according to Metze et al. (Metze et al., 2024). But these results depend strongly on

⁶ This notation comes from Python, it is equal to $2.76 \cdot 10^{-7} = 0.000000276$

⁷ It is equal to $9.73 \cdot 10^{-9} = 0.00000000973$

the manually determined thresholds and for this we used only the titles and didn't consider the author names, years and journals.

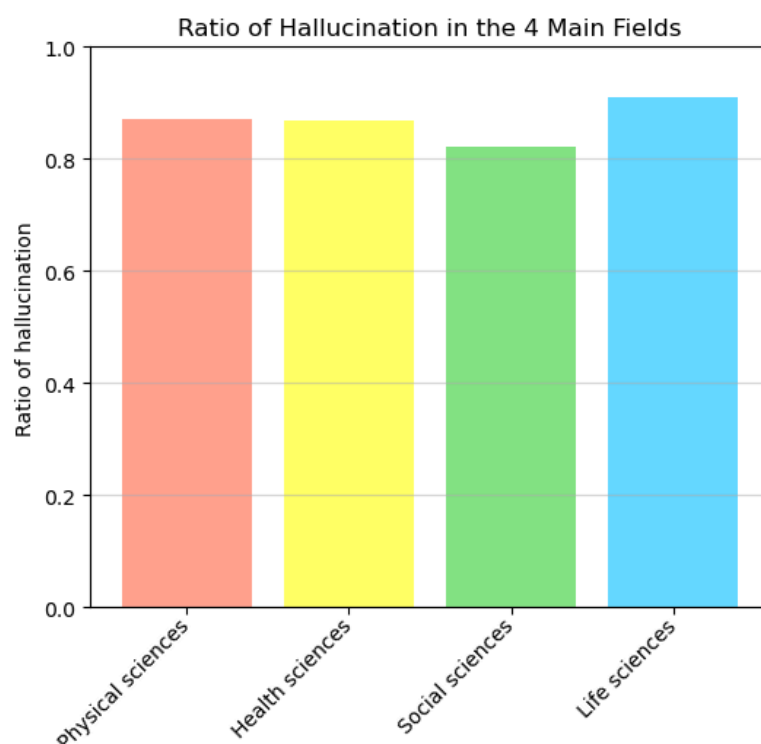


Figure 6.: Ratio of hallucination in the 4 main fields

After filtering out the hallucinations, we calculated the ratio of women for the ChatGPT citations which were considered as real, and for their OpenAlex pair, and visualized the distribution of their difference (ChatGPT ratio - OpenAlex ratio) on Figure 7. Positive numbers indicate that ChatGPT has a higher ratio of women authors than the original citations do. We can see on the figure that the value of 0 has the most weight, but it can be explained with the most common value of 0 in the woman ratios, and the density curve is almost symmetrical, only slightly skewed to the positive side.

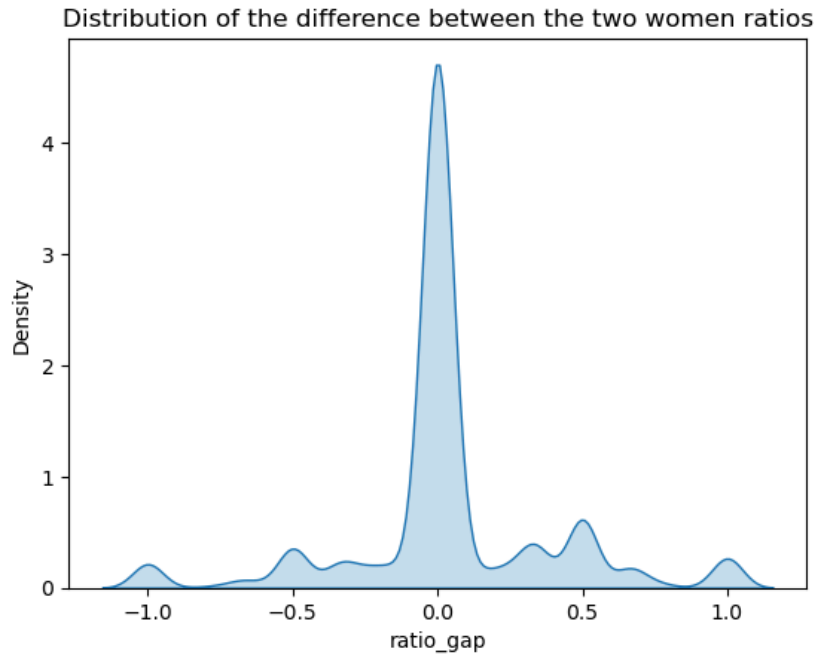


Figure 7.: Density curve of the difference between ChatGPT and OpenAlex ratio of women in real citations

The average ratio of women for the real citations were a bit lower for the ChatGPT references than before. In the case of the OpenAlex citations the ratios were quite the same as before, so we can say that ChatGPT tends to hallucinate more female authors than male authors. Considering only the real citations, ChatGPT overcites women in the field of Physical, Heath and Life Sciences, so in these fields ChatGPT tends to replace male authors with female authors. The field of Social Sciences is an exception, because the values of this field are quite equal to the OpenAlex ratios. A possible explanation for this can be that a quat-style fairness might be built in the model ensuring to cite a certain ratio of women. Since in Social Sciences 30% of authors are already women, ChatGPT did not have to modify the authors' name. In the other fields there is lower overall representation of women, so ChatGPT had to increase the ratio of women authors. ChatGPT is not an open source model, so we can only guess what the answer might be.

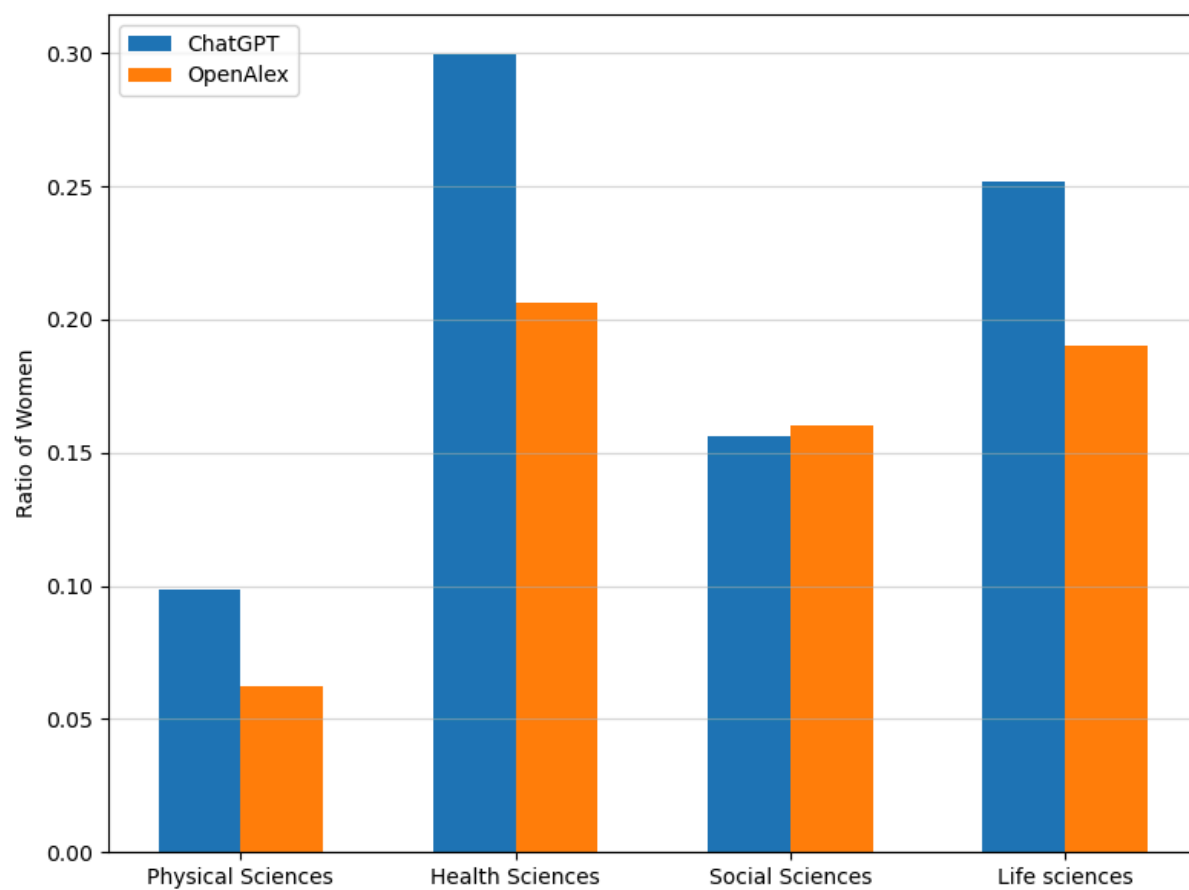


Figure 8.: Ratio of women in the non-hallucinated ChatGPT references and their OpenAlex pairs in the 4 main scientific fields

6. CONCLUSIONS, NEXT STEPS

This study highlights the complex role that large language models (LLMs), like ChatGPT play in the academic landscape, particularly in generating literature reviews and citations. Our study developed a systematic methodology to analyze references created by ChatGPT in literature reviews by comparing it with an open source bibliographic database OpenAlex.

Similarly to previous research, mainly conducted in smaller samples on one scientific discipline and earlier versions (GPT 3.5), we found that ChatGPT introduces biases and inaccuracies. Specifically, ChatGPT demonstrates a tendency to favor more recent publications and exhibit a gender-based skew in author citations, often disproportionately favoring female authors relative to traditional academic databases like OpenAlex. Unlike earlier versions, which consistently cited older, highly cited articles from top journals, the current model appears more varied in its source selection, signaling significant evolution in its citation patterns. These shifts in citation behavior point to broader issues in AI-driven knowledge production: while updates to the model may reduce one type of bias, they can introduce new ones, which further underscores the need for ongoing scrutiny and refinement of AI systems used in academia.

Additionally, the study exposed a high hallucination rate in ChatGPT's generated citations, indicating an area for improvement in reliability and factual accuracy. In the case of ChatGPT-3, Metze et al. found that 7.1% of the references were correct; 6.2% showed minor errors and 86.7% had major errors (Metze et al., 2024), and we couldn't see an improving tendency when examining the GPT-4o model. These findings underscore the need for critical evaluation of AI outputs, especially when these models are integrated into scholarly workflows.

We had some limitations during our research. We only investigated one model of ChatGPT, but there are previous models, and different large language models which could have been examined and compared. Because of the lack of time we only ran the prompts on each topic 10 times, but in order to get more robust results more data is needed. We did not check whether the literature reviews written by ChatGPT are correct or contain some factual misinformations, only examined the consistency of them. Another limitation is that we used only a small fraction of the OpenAlex database and dropped articles which have less than 2 citations (which are the 95% of the whole database). So the determination of the hallucination rates is strongly biased by this limitation. We used the gender-guesser package in Python to get the genders of the authors, but we are aware of the limitations that name-based gender

inferring methods can cause, especially yielding lower accuracy for Asian names (Vásárhelyi & Siân, 2023). In our future work we are planning to use more accurate and ethnicity-based gender inferring methods, and assist the results for the level of error.

A possible future direction is to examine not only the gender but the ethnicity of the authors. Ethnea (Torvik & Agarwal, 2016) is a useful tool for this, which provides not only the genders but the name-based ethnicity of the authors. We plan to do the same investigation for the previous model of ChatGPT to see the tendency in hallucination and the ratio of female authors, and we would like to know whether we would see different patterns in the case of other large language models, e. g. Gemini. It would be interesting to train Llama, an open source LLM with data from OpenAlex and examine how different the results would be compared to other models.

REFERENCES

- Acerbi, A., & Stubbersfield, J. M. (2023). Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*.
- Aksnes, D. W., Rorstad, K., Piro, F., & Sivertsen, G. (2011). Are female researchers less cited? A large-scale study of Norwegian scientists. *Journal of the American Society for Information Science and Technology*.
- Alkaissi, H., & McFarlane, S. (2023). Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus*.
- Antu, S. A., Chen, H., & Richards, C. K. (2023). Using LLM (Large Language Model) to Improve Efficiency in Literature Review for Undergraduate Research.
- Athaluri, S., Manthena, V., Kesapragada, M., Yarlagadda, V., Dave, T., & Duddumpudi, S. (2023). Exploring the Boundaries of Reality: Investigating the Phenomenon of Artificial Intelligence Hallucination in Scientific Writing Through ChatGPT References. *Cureus*.
- Bao, H., Sun, M., & Teplitskiy, M. (2024). Where there's a will there's a way: ChatGPT is used more for science in countries where it is prohibited. *arXiv*.
- Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The longdocument transformer. *arXiv*.
- Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*.
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*.
- Blackaby, D., Booth, A. L., & Frank, J. (2005). Outside Offers and the Gender Pay Gap: Empirical Evidence from the UK Academic Labour Market. *The Economic Journal*.
- Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of “bias” in NLP. *arXiv*.
- Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., & et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., & et al. (2020). Language Models are Few-Shot Learners. *arXiv*.
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., & et al. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv*.
- Buchanan, J., Hill, S., & Shapoval, O. (2023). ChatGPT Hallucinates Non-existent Citations: Evidence from Economics. *The American Economist*.
- Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on fairness, accountability and transparency*.
- Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*.
- Carrell, S. E., Page, M. E., & West, J. E. (2010). Sex and Science: How Professor Gender Perpetuates the Gender Gap. *The Quarterly Journal of Economics*.
- Ceci, S. J., & Williams, W. M. (2007). Why aren't more women in science?: Top researchers debate the evidence. *American Psychological Association*.
- Ceci, S. J., & Williams, W. M. (2011). Understanding current causes of women's underrepresentation in science. *Proceedings of the National academy of sciences*.
- Charlesworth, T.E.S. B., & Banaji, M.R. (2019). Gender in Science, Technology, Engineering, and Mathematics: Issues, Causes, Solutions. *Journal of Neuroscience*.
- ChatGPT. (n.d.). Wikipedia. Retrieved October 29, 2024, from <https://en.wikipedia.org/wiki/ChatGPT>
- Chelli, & Mikael. (2024). Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis. *Journal of Medical Internet Research*.
- Chelli, M., & et al. (2024). Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis. *Journal of Medical Internet Research*.
- Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2023). Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review. *arXiv*.
- Dang, Hai, & et al. (2022). How to prompt? Opportunities and challenges of zero-and few-shot learning for human-AI interaction in creative applications of generative models. *arXiv*.

- del Rio-Chanona, R. M., Laurentsyeve, N., & Wachs, J. (2024). Large language models reduce public knowledge sharing on online Q&A platforms. *PNAS Nexus*.
- Duch, J., Zeng, X. H. T., & et al. (2012). The Possible Role of Resource Requirements and Academic Career-Choice Risk on Gender Differences in Publication Rate and Impact. *PloS one*.
- Fang, X., Che, S., & et al. (2024). Bias of AI-generated content: an examination of news produced by large language models. *Scientific Reports*.
- Feast, J. (2019, November 20). *4 Ways to Address Gender Bias in AI*. Harvard Business Review. Retrieved October 31, 2024, from <https://hbr.org/2019/11/4-ways-to-address-gender-bias-in-ai>
- Ferrara, E. (2023). Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci*.
- Ferrara, E. (2023). Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models. *First Monday*.
- Gemini (chatbot). (n.d.). Wikipedia. Retrieved October 29, 2024, from [https://en.wikipedia.org/wiki/Gemini_\(chatbot\)](https://en.wikipedia.org/wiki/Gemini_(chatbot))
- Gender and AI: Addressing bias in artificial intelligence. (n.d.). International Women's Day. Retrieved October 31, 2024, from <https://www.internationalwomensday.com/Missions/14458/Gender-and-AI-Addressing-bias-in-artificial-intelligence>
- Ghosh, S., & Caliskan, A. (2023). 'Person' == Light-skinned, Western Man, and Sexualization of Women of Color: Stereotypes in Stable Diffusion. *arXiv*.
- Grabb, D. (2023). The impact of prompt engineering in large language model performance: a psychiatric example. *Journal of Medical Artificial Intelligence*.
- Hacker, P., Mittelstadt, B., Borgesius, F. Z., & Wachter, S. (2024). Generative Discrimination: What Happens When Generative AI Exhibits Bias, and What Can Be Done About It. *arXiv*.
- Henry, P. (2023, June 22). *How Duolingo Uses AI to Create Lessons Faster*. Duolingo Blog. Retrieved October 29, 2024, from <https://blog.duolingo.com/large-language-model-duolingo-lessons/>
- Hopkins, A. L., Jawitz, J. W., McCarty, C., Goldman, A., & Basu, N. B. (2013). Disparities in publication patterns by gender, race and ethnicity based on a survey of a random sample of authors. *Scientometrics*.
- Hovy, D., & Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass*.
- Huang, J., Gates, A. J., Sinatra, R., & Barabási, A.-L. (2020). Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences*.
- Huang, L., Yu, W., & et al. (2023). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *arXiv*.
- Jain, R., & Jain, A. (2024). Generative AI in Writing Research Papers: A New Type of Algorithmic Bias and Uncertainty in Scholarly Work. *Intelligent Systems Conference*.
- Kahn, S., & Ginther, D. K. (2015). Are recent cohorts of women with engineering bachelors less likely to stay in engineering? *Frontiers in Psychology*.
- Kelly, K., & Grant, L. (2012). Penalties and premiums: The impact of gender, marriage, and parenthood on faculty salaries in science, engineering and mathematics (SEM) and non-SEM fields. *Social Studies of Science*.
- Khandelwal, U., He, H., Qi, P., & Jurafsky, D. (2018). Fuzzy Far Away: How Neural Language Models Use Context. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.
- Klar, S., Krupnikov, Y., & et al. (2020). Using social media to promote academic research: Identifying the benefits of twitter for sharing academic work. *PloS one*.
- Knoth, N., Tolzin, A., Janson, A., & Leimeister, J. M. (2024). AI literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence*.
- Kotek, H., Dockum, R., & Sun, D. (2023). Gender bias and stereotypes in Large Language Models. *Proceedings of the ACM collective intelligence conference*.
- Kozlowski, D., Larivière, V., Sugimoto, C. R., & Monroe-White, T. (2021). Intersectional inequalities in science. *Proceedings of the National Academy of Sciences*.
- Kumar, S. H., Sahay, S., Mazumder, S., & et al. (2024). Decoding Biases: Automated Methods and LLM Judges for Gender Bias Detection in Language Models. *arXiv*.
- Legewie, J., & DiPrete, T. A. (2014). The high school environment and the gender gap in science and engineering. *Sociology of education*.
- Leong, K., & Sung, A. (2024). Gender stereotypes in artificial intelligence within the accounting profession using large language models. *Humanities and Social Sciences Communications*.
- Lu, Y., Bartolo, M., & et al. (2021). Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. *arXiv*.

- Lund, B. D., Wang, T., Mannuru, N. R., Nie, B., Shimray, S., & Wang, Z. (2023). ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology*.
- Maliniak, D., Powers, R., & Walter, B. F. (2013). The Gender Citation Gap in International Relations. *International Organization*.
- Marcus, G., & Davis, E. (2020). GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about. *MIT Technology Review*.
- Markowitz, D. M. (2024). From complexity to clarity: How AI enhances perceptions of scientists and the public's understanding of science. *PNAS Nexus*.
- McCullough, L. (2011). Women's Leadership in Science, Technology, Engineering and Mathematics: Barriers to Participation. *Forum on Public Policy Online*.
- Metze, K., Morandin-Reis, R. C., & Florindo, J. B. (2024). Bibliographic Research with ChatGPT may be Misleading: The Problem of Hallucination. *J Pediatr Surg*.
- Metze, K., Morandin-Reis, R. C., Lorand-Metze, I., & Florindo, J. B. (2024). Bibliographic research with ChatGPT may be misleading: the problem of hallucination. *Journal of Pediatric Surgery*.
- Microsoft Copilot. (n.d.). Wikipedia. Retrieved October 29, 2024, from https://en.wikipedia.org/wiki/Microsoft_Copilot
- Munro, R., Bethard, S., Kuperman, V., Lai, V. T., Melnick, R., & et al. (2010). Crowdsourcing and language studies: the new generation of linguistic data. *NAACL Workshop on Creating Speech and Language Data With Amazon's Mechanical Turk, Association for Computational Linguistics*.
- Nadeem, A., Abedin, B., & Marjanovic, O. (2020). *Gender Bias in AI: A Review of Contributing Factors and Mitigating Strategies*.
- NSF (National Science Foundation). (2018). Science and Engineering Indicators 2018. *National Science Foundation*. <https://www.nsf.gov/statistics/indicators/>
- O'Connor, S., & Liu, H. (2024). Gender bias perpetuation and mitigation in AI technologies: challenges and opportunities. *AI & SOCIETY*.
- OpenAI. (2023). *GPT-4 System Card*.
- OpenAI. (2023). *Gpt-4 technical report*.
- OpenAI and Apple announce partnership. (2024, June 10). OpenAI. Retrieved October 29, 2024, from <https://openai.com/index/openai-and-apple-announce-partnership/>
- OpenAlex. (n.d.). OpenAlex: The open catalog to the global research system. Retrieved October 30, 2024, from <https://openalex.org/>
- OpenAlex Topics. (n.d.). <https://help.openalex.org/how-it-works/topics>
- <https://help.openalex.org/how-it-works/topics>
- Petiška, E. (2023). ChatGPT cites the most-cited articles and journals, relying solely on Google Scholar's citation counts. As a result, AI may amplify the Matthew Effect in environmental science. *Charles University*.
- Pfohl, S. R., Cole-Lewis, H., & et al. (2024). A toolbox for surfacing health equity harms and biases in large language models. *Nature Medicine*.
- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*.
- Raffel, V., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*.
- Robertson, A., & David, E. (2024, February 21). *Google apologizes for "missing the mark" after Gemini generated racially diverse Nazis*. The Verge. Retrieved October 30, 2024, from <https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical>
- Siontis, K. C., Attia, Z. I., Asirvatham, S. J., & Friedman, P. A. (2024). ChatGPT hallucinating: can it get any more humanlike? *European Heart Journal*.
- Smith, G., & Rustagi, I. (2021, March 31). *When Good Algorithms Go Sexist: Why and How to Advance AI Gender Equity (SSIR)*. Stanford Social Innovation Review. Retrieved October 31, 2024, from https://ssir.org/articles/entry/when_good_algorithms_go_sexist_why_and_how_to_advance_ai_gender_equity
- Solaiman, I., Brundage, M., Clark, J., Askill, A., Herbert-Voss, A., & et al. (2019). Release strategies and the social impacts of language models. *arXiv*.
- Sugimoto, C. R., & Lariviere, V. (2023). *Equity for Women in Science: Dismantling Systemic Barriers to Advancement*. Harvard University Press.
- Suppadungsuk, S., Thongprayoon, C., Krisanapan, P., & et al. (2023). Examining the Validity of ChatGPT in Identifying Relevant Nephrology Literature: Findings and Implications. *J. Clin. Med*.

- Torvik, V., & Agarwal, S. (2016). Ethnea --an instance-based ethnicity classifier based on geo-coded author names in a large-scale bibliographic database.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., & et al. (2023). Llama: Open and efficient foundation language models. *arXiv*.
- Towhidul, S. M., & Tonmoy, I. (2024). A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models. *arXiv*.
- US Department of Education (Ed.). (2021). Full-time faculty in degree-granting postsecondary institutions, by race/ethnicity, sex, and academic rank: Fall 2018, fall 2019, and fall 2020, fall 2021. *Digest of Education Statistics, National Center for Education Statistics*.
- Vásárhelyi, O. (2020). *Computational and relational understanding of gender inequalities in science and technology*.
- Vásárhelyi, O., & Sián, B. (2023). Computing gender.
- Vásárhelyi, O., Zakhlebin, I., Milojević, S., & Horvát, E. Á. (2021). Gender inequities in the online dissemination of scholars' work. *Proceedings of the National Academy of Sciences*.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention Is All You Need. *arXiv*.
- Walters, W. H., & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Sci*.
- Wang, H., Fu, T., Du, Y., & et al. (2023). Scientific discovery in the age of artificial intelligence. *Nature*.
- Webb, M. (2023). Exploring the potential for bias in ChatGPT. *Jisc National Centre for AI*.
<https://nationalcentreforai.jiscinvolve.org/wp/2023/01/26/exploring-the-potential-for-bias-in-chatgpt/>
- Winslow, S. (2010). Gender inequality and time allocations among academic faculty. *Gender & Society*.
- Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., & Duan, N. (2023). Visual ChatGPT: Talking, drawing and editing with visual foundation models. *arXiv*.
- Wu, T., Terry, M., & Cai, C. J. (2021). AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. *arXiv*.
- Xames, M. D., & Shefa, J. (2023). ChatGPT for Research and Publication: Opportunities and Challenges. *Journal of Applied Learning and Teaching*.
- Xiao, D. (2024). AI-Based Literature Review Tools.
- Zeng, X. H. T., Duch, J., & et al. (2016). Differences in Collaboration Patterns across Discipline, Career Stage, and Gender. *PloS one*.
- Zhang, M., Press, O., & et al. (2023). How Language Model Hallucinations Can Snowball. *arXiv*.
- Zhang, T., Qiu, L., Guo, Q., & et al. (2023). Enhancing Uncertainty-Based Hallucination Detection with Stronger Focus. *arXiv*.
- Zhou, M., Abhishek, V., Derdenger, T., Kim, J., & Srinivasan, K. (2024). Bias in Generative AI. *arXiv*.
- Zimmermann, R., Staab, M., & et al. (2024). Leveraging Large Language Models for Literature Review Tasks - A Case Study Using ChatGPT. *Springer Nature Switzerland*.

APPENDIX

Table 4: The manually determined thresholds of the cosine similarity scores, below which the references were considered as hallucinations

Subfield	Threshold
Agricultural and Biological Sciences	0.72
Arts and Humanities	0.62
Biochemistry, Genetics and Molecular Biology	0.709
Business, Management and Accounting	0.705
Chemical Engineering	0.61
Chemistry	0.615
Computer Science	0.748
Decision Sciences	0.675
Dentistry	0.663
Earth and Planetary Sciences	0.664
Economics, Econometrics and Finance	0.679
Energy	0.68
Engineering	0.677
Environmental Science	0.677
Health Professions	0.677
Immunology and Microbiology	0.68
Materials Science	0.685
Mathematics	0.69
Medicine	0.695
Neuroscience	0.708
Nursing	0.685
Pharmacology, Toxicology and Pharmaceuticals	0.68
Physics and Astronomy	0.6752
Psychology	0.689
Social Sciences	0.68
Veterinary	0.66