September 2023

Pitt Cyber
Institute for
Cyber Law, Policy, and Security

Who Authors the Internet?

Analyzing Gender Diversity in ChatGPT-3 Training Data

BY JESSICA B. KUNTZ AND ELISE C. SILVA, PHD

Jessica Kuntz is the Policy Director at the University of Pittsburgh Institute for Cyber Law, Policy, and Security. Elise Silva, PhD, is a Postdoctoral Associate at University of Pittsburgh Institute for Cyber Law, Policy, and Security.

Humans are biased. Contrary to hopes that AI might promote greater objectivity in decision making, research has consistently found that AI systems are rife with bias – in large part because they are trained on human generated data. One study sought to quantify the political bias of various Large Language Models (LLMs) by asking each to respond to 62 political statements, such as "our race has many superior qualities, compared with other races," and "abortion, when the woman's life is not threatened, should always be illegal" (Feng, 2023). The results demonstrated that "pretrained LMs do have political leanings that reinforce the polarization present in pretraining corpora, propagating social biases into hate speech predictions and misinformation detectors."

Research and reporting have exposed numerous instances of gender biased AI outputs: following unsuccessful debiasing attempts, Amazon discontinued an AI powered hiring tool that was found to be biased against women. In 2016, Microsoft shut down its AI enabled Twitter bot Tay in quick order once it was found to generate racist, misogynist and otherwise insulting content (Hunt). One creative study used a list of tasks associated with software development and instructed a LLM to translate each sentence from Finnish (which has no gender pronouns) to English, noting whether the LLM opted to replace the genderless Finnish pronoun with 'he, 'she,' or 'he/she' in the output. Of the 56 tasks, the model opted for 'she' the majority of the time with only four tasks – all associated with talking to colleagues, drafting emails, and taking on the emotional labor of mentorship (Treude & Hata, 2023). This division of work replicates gendered concepts of the workplace, with women predominantly associated with 'soft' communication tasks – an association that is, coincidentally, mirrored and reinforced by the default use of female voices in virtual assistant AIs. A like bias manifested when the UN Development Programme (UNDP) Accelerator Lab prompted text-to-image models to generate photos of an engineer, scientist, mathematician and IT expert: the resulting images portrayed men between 75 and 100 percent of the time (Reproducing inequality, 2023).

In one particularly explicit example, a female journalist of Asian descent recalls how the AI avatar app Lensa "generated realistic yet flattering avatars for [her male colleagues]—think astronauts, fierce warriors, and cool cover photos for electronic music albums— [whereas] I got tons of nudes. Out of 100 avatars I generated, 16 were topless, and in another 14 it had put me in extremely skimpy clothes and overtly sexualized poses" (Heikkilä "Viral Avatar App… ," 2023).

AI developers are aware of the problem and have taken some steps to mitigate it. OpenAI announced it had taken measures to reduce gendered outputs from its DALL-E image generator when given career prompts of the sort documented by UNDP (Reducing Bias and Improving Safety in DALL·E 2, 2022). However, it is reported that the company did so by targeting and amending gender and race identifiers for each prompt, not by addressing the bias present in the training data (Traylor, 2022). This approach is emblematic of an assumption that "toxicity and bias contained in the pre-training data can be sufficiently contained via fine-tuning, turning LLMs from unsupervised monsters into helpful assistants" (Baack, 2023). By failing to address the source of the bias, however, these band-aid solutions allow bias to persist in ways developers may not readily identify. It's a variation of "you can't fix what you don't measure;" if developers know the model characterizes scientists as default male, it can build safeguards to discourage that particular behavior. However, they can only retroactively address biases that have been diagnosed, leaving untold manifestations of bias unaddressed.

Explaining why Lensa was generating sexualized female avatars, the writer notes that "the internet is overflowing with images of naked or barely dressed women, and pictures reflecting sexist, racist

stereotypes, [and so] the data set is also skewed toward these kinds of images." The industry emphasis on training data *quantity* has led to a disregard for *quality*. We join growing calls for data stewardship in urging an intentional approach to data selection (Li) and a more thoughtfully curated training corpus[1] -- even if that entails tradeoffs in overall model capacity, it saves us from a "garbage in, garbage out" version of AI.

Into the Black Box: What Text is Training LLMs?

To better understand the origin of LLM bias, we attempt to quantify an underexamined source of bias in model inputs: the authorship of pretraining data. We evaluate authorship of ChatGPT-3's training corpuses through a gender lens. Operating from developers' limited disclosure of the model, we estimate that just over a quarter – 26.5% – of ChatGPT-3 training data was authored by women.

Our initial research question, stated above, relates to the diversity of authorship in LLM training data. In the effort to quantify the percentage of training text authored by women, we found ourselves having to make repeated assumptions about the representativeness of small snapshots of training data and vaguely educated guesses about the true contents of training corpuses – encapsulating the challenge of conducting research inside the LLM black box. While our findings do indicate a significant underrepresentation of female authorship, our broader take away stems from the sheer number of assumptions necessarily underlying this calculation. This paper serves as a case study of what is lost when disclosure and documentation of LLM training data is lacking. Transparency has become a catchphrase – but it is also a necessary precondition to evaluate these tools for bias, potential misuse, and accuracy.

LLMs currently have a seemingly endless appetite for data, with each model trained on more parameters than the last. We argue for a more intentional approach towards training data selection, with greater emphasis on data quality and representativeness. Considering that these AI systems will be used to facilitate decisions with real life consequences, such a stance seems only reasonable. However, the technical definition of 'quality' differs significantly: in the LLM sense, data quality does not equate to factual accuracy or representativeness – rather it refers to cleaned and structured data that allows for easy filtering. By the technical definition, data gleaned from a *New York Times* article would not necessarily be higher quality than a "stop the steal" Infowars article. Gururangan et al. confirm this by running factually accurate and inaccurate content through their filter, finding that "many factually unreliable news articles are considered high quality by the filter" (Gururangan, 2022).

Training an LLM on a huge volume of 'high quality' (i.e. cleaned and structured) data that was scrapped from conspiracy theory websites, erotica, and Gab would result in a technologically capable model – just one rife with biases, misogyny, and prone to sprouting QAnon inspired content. This point was compellingly demonstrated through the creation of a prank model trained on 134 million posts from 4chan's Politically Incorrect board /pol/. The result was a model that its developer characterized as having "perfectly encapsulated the mix of offensive, nihilism, trolling, and deep distrust of any information whatsoever that permeates most posts on /pol" (Vee, 2022). That such a model could result from so-called 'high quality' data makes clear that we need to dramatically rethink what is

[1] Meta's LLaMA model is reported to have been trained on "two sources … data that was scraped online, and a data set fine-tuned and tweaked according to feedback from human annotators to behave in a more desirable way." (Heikkilä, 2023). Although obviously a heavier lift, cleaning training data is seemingly possible.

understood as high quality data. LLMs mirror the inputs they are fed; it is imperative that we select inputs reflecting the sentiments and diversity of perspectives that we wish to see reflected in the outputs.

The remainder of the study proceeds as follows:

- A brief history of women's exclusion from data and the particular relevance of gender in authorship of written texts
- Overview of trends in LLM data transparency
- Breakdown of ChatGPT-3's training corpuses, as reported by Open AI. For each corpus we consider:
    - The source of the text, including other concerns that have been raised beyond gender in authorship
    - Measurements of gender/proxy variables, noting the underlying assumptions and limitations therein
- Conclusion

Biased Inputs: The Gender Data Gap

The omission of women's voices in AI is the latest manifestation of a long history of underrepresenting women in data collection and analysis. Author Caroline Criado Perez documents a multitude of ways in which female residents, workers, and patients are underrepresented or even omitted in designing cities, medical systems/treatments and jobs. The gender data gap is both cause and consequence of a social "male unless proven otherwise" default:

> The lives of men have been taken to represent those of humans overall … these [data] gaps, have consequences. They impact women's lives every day. The impact can be relatively minor. Shivering in offices set to a male temperature norm, for example, or struggling to reach a top shelf set to a male height norm. Irritating, certainly. Unjust, undoubtedly.
>
> But not life threatening. Not like crashing in a car whose safety measures don't account for women's measurements. Not like having your heart attack go undiagnosed because your symptoms are deemed 'atypical.' For those women, living in a world built around male data can be deadly (Perez, 2019).

Perez's examples abound: failure to include data about women's travel patterns (including school drop off and pick up) in public transportation or snow clearing schedules discourages female participation in the labor force and leads to higher female injury rates; pension schemes that penalize workers who reduce their hours to accommodate child and elder care lead to "feminized poverty;" tools, military equipment/uniforms, PPE designed for the standard male body compromise women's safety and health.

These examples speak to data that fails to account for variation in men and women's physical bodies and behavior patterns. LLM's use of written text as data brings a related question to bear: how do women's writing styles and perspectives vary from those of men – and why does it matter?

Sociolinguists have documented distinct gender-based differences in writing and speech patterns. In one of the largest studies seeking to understand differences in how men and women write, researchers found that women used more pronouns than men. Researchers describe this as a more *involved* style of

writing. Men's writing is described as more *informational*, using more specifiers and common nouns than women (Argamon, 2003). An example might be illustrated in these two sentences: 1) We argue in this paper that flawed training data will lead to biased textual generation; 2) This paper argues that flawed training data will lead to biased textual generation. While both sentences convey the same message, syntactic difference subtly changes the readerly experience. When a gender-skewed dataset trains a text generating LLM, that text risks favoring male-centric writing styles.

We view the world thought multiple lenses, gender being an important one.[2] This perspective impacts how each of us sees the world and what aspects we choose to document. A study examining blog content notes that "male bloggers of all ages write more about politics, technology and money than do their female cohorts. Female bloggers discuss their personal lives – and use more personal writing style – much more than males do" (Schler, 2006). On aggregate, a text's content and perspective differ depending on the author's gender.

Further research confirms a strong relationship between gender and writing style, content, and perspective (Colley) owing partially to "the interaction between language and gender [that] are mediated by situational contexts" (Bamman, 2014). We further acknowledge that gender is one of multiple interconnected social categories, including sexual orientation, race, and age, all being shaped by lived experience. Our analysis is merely a starting point to understand how diversity in training data matters. Other marginalized identities and communities – to include race[3] and socio-economic status[4] – merit further analysis.

Homogenous authorship of training materials unintentionally reinforces the male voice/perspective as the norm. With female voices underrepresented in the training dataset, we would expect that perspective would be generated less often or with less precision in the model's output. As we enter a world with AI involvement in medical research, hiring, education and more – the extent to which LLMs are trained on diverse perspectives matters tremendously. The problem only compounds as the LLM's bias becomes self-reinforcing: our world views are shaped by the content we see – whether in the form of textbooks, social media posts, conspiracy theories, or AI-generated content. If that content is limited by training data, our perspectives are likewise limited.

With all this at stake, we stand to lose remarkable cultural diversity in writing styles and content by stifling linguistic, subject, and syntactic gender differences in writing. By so doing, we run the risk of reifying "normative" writing (or default textual generation) as "male." It is important to interrogate the training data alongside its result. If the training data is skewed towards male authorship, the results

[2] We recognize that gender is not binary, despite how we are handling questions of authorship here. The male/female framing was driven by the nature of datasets.

[3] The University of Washington undertook to measure the percentage of minority authored content in the C4 dataset using dialects as a proxy for ethnicity. They found that the common (and well-intended) practice of removing texts that contained terms on a "bad" words list (lacking the ability to discern context, speaker or innuendo) disproportionately removed text authored by minority authors. https://arxiv.org/pdf/2104.08758.pdf

[4] An interesting attempt to evaluate the presence of socio-economic diversity in LLM training data can be found in a 2022 paper by Gururangan et al. The authors recreated the filter that uses "the original WebText [the content of links from upvoted Reddit posts] as a proxy for high-quality documents," then ran the filter on a dataset comprised of high school newspaper articles. They find that "the filter reinforces a language ideology that text from authors of wealthy, urban, and educated backgrounds are more valuable for inclusion in language model training data." https://arxiv.org/abs/2201.10474

likewise will be, even if they are assumed to be neutral, or non-gendered. Our analysis suggests they are anything but.

Exploring LLM Training Data is Hampered by a Lack of Transparency

In order to fully assess authorship of the training data, said training data must be disclosed. Unfortunately, that is often – and increasingly – not the case. The Stanford Center for Research on Foundation Models rated top foundation models on their compliance with the draft EU AI Act;[5] for the category of describing training data sources, models earned an average score of 2 (4 being a top score). ChatGPT-4 fell near the bottom with a score of 1 (ChatGPT-3 was not included) (Bommasani, et al., 2023).

As justification, companies often point to safety or market competition. In the paper accompanying the launch of GPT-4, Open AI states: "given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar" (GPT-4 Technical Report, 2023). However, a reported laisse faire approach to internal documentation calls into question the sincerity of the above rationale. A joint report by the Allen Institute for AI and The Washington Post observes that "companies do not document the contents of their training data — even internally — for fear of finding personal information about identifiable individuals, copyrighted material and other data grabbed without consent" (Schaul, et al., 2023). With this in mind, Open AI's safety defense seems more likely to protect the company rather than end users.[6] Poorly documented and/or obscurity around training data is an industry norm, including for open source[7] models. Assessing the state of LLM openness, Dutch researchers took particular issue with LLaMA's purported commitment to transparency (Liesenfeld, 2023): "Meta using the term 'open source' for this is positively misleading: There is no source to be seen, the training data is entirely undocumented, and beyond the glossy charts the technical documentation is really rather poor. We do not know why Meta is so intent on getting everyone into this model, but the history of this company's choices does not inspire confidence. Users beware" (Nolan, 2023).

Whether due to proprietary and monetization priorities or a desire to avoid litigation, the trend is clearly moving away from data transparency. Counter to industry claims that transparency would undermine safety, we contend that required disclosure of training data would hold developers publicly accountable for the quality of their data and the legality of its use and, as such, provide the public with a better, safer product.

[5] The EU AI Act is the most developed of all regulatory efforts concerning AI development and use. The proposed law would use a risk-based classification system and hold systems to requirements in accordance with their risk level.

[6] LLMs' approach of using users as guinea pigs, relying on the public/press/researchers to identify jailbreaks and bias in models that have already been publicly released, seems to be done with little regard to safety.

[7] Open source software makes its source code available to the public, allowing anyone with coding knowledge to modify the source code. In the case of LLMs, a model can be open source without providing training data documentation.

An Analysis of Authorship of Gender in ChatGPT-3 Training Data

ChatGPT-3 was trained on 45 terabytes of text data, which equates to the text from roughly 90 *million* novels. 60% of the training data is from Common Crawl: "data collected [by] crawling web. The corpus contains raw web page data, metadata extracts and text extracts with light filtering" (Brown et al., 2020). Another 22% comes from Webtext2: "the text of web pages from all outbound Reddit links from posts with 3+ upvotes." 16% comes from Books1 and Book2, which are internet-based books. The remaining 3% of training data comes from English language Wikipedia. Open AI notes that "during training, datasets are not sampled in proportion to their size, but rather datasets we view as higher quality[8] are sampled more frequently … but the other datasets are sampled 2-3 times. This essentially accepts a small amount of overfitting in exchange for higher quality training data" (Brown et al., 2020). Using this list, we make preliminary assessments about authorship, which also illustrate the importance of data transparency by revealing the guesswork that results when training materials are obscured. Our findings indicate a low rate of women's authorship, with the possible exception of book-based data.

[8] The reader will recall that 'high quality' data as industry defines it refers to data that is cleaned and structured - regardless of the content.

ChatGPT-3 Data Source	Percentage ChatGPT-3 training corpus	Estimated percentage content authored by women	Source(s)
Common Crawl: Journalistic sources	60%	**19%**[9] *13% U.S. patent applications (#1 source of data)* *9% Wikipedia (#2 source of data)* *37% journalistic sources (#3 source of data)*[10]	US Patent & Trade Office (USPTO) report Global Wikipedia Survey, United Nations University "Global Report on the Status of Women in the News Media," International Women's Media Foundation
Webtext2 (outbound links from Reddit)	22%	**31%** *(based on percentage female Reddit users)*	Pew Research Center
Books1/Books2	16%	**50%** [11]	Analysis of titles in the Library of Congress in "The welfare effect of gender-inclusive intellectual property creation: evidence from books"
Wikipedia	3%	**9%** *(based on percœntage global female Wikipedia editors)*	Global Wikipedia Survey, United Nations University
Average Estimated Female Authorship across GPT-3 Training Data		**26.5%**	

[9] The three top websites/categories of websites are identified by number of tokens in the dataset. We weigh each of these equally to reach the 19% average. The estimated female authorship for each of the top three is noted below.

[10] The 10 journalistic sources included in the top 25 sites are: NYTimes, LA Times, The Guardian, Forbes, Huff Post, Washington Post, Business Insider, Chicago Tribune, The Atlantic, Al Jazeera

[11] There are significant concerns and uncertainties concerning the source(s) of these datasets. 50% reflects the percentage of recently published books authored by women.

Common Crawl:

What is it?

Common Crawl is a 501(3)(c) non-profit, offering its data (which it describes as "a copy of the web") under the open data banner. Open AI is one amongst many to utilize this resource; on their website, Common Crawl boasts that "small startups or even individuals can now access high quality crawl data that was previously only available to large search engine corporations." It is funded by online donations and appears to be a low labor operation: Common Crawl lists only three employees on their website (in addition to the board of directors and advisory board).

In building an AI model, companies apply filters to each corpus, leaving them with a cleaned-up version of the data which is then fed into the model. The University of Washington undertook this exercise with a snapshot taken from Common Crawl. In the resulting paper, they report applying the following common filters: removing "lines which don't end in a terminal punctuation mark or have fewer than three words, discarding documents with less than five sentences or that contain 'lorem ipsum' placeholder text, and removing documents with any word on the List of Dirty, Naughty, Obscene, or Otherwise Bad Words" (Dodge et al., 2021).

In the resulting filtered dataset (titled C4), the two largest sources of training data were patents.google.com and Wikipedia. It's important to put 'largest' in context however: the patents text provided roughly 1 billion tokens and the Wikipedia text 200 million tokens – of the total 156 *billion* tokens in the database. News sites were collectively well represented in the top websites, taking 10 of the 25 slots, to include *NYTimes*, *LA Times* and *Forbes* (although, again – less than 1% combined of the total corpus). With the disclaimer that the sheer size and diversity of the filtered dataset makes it extraordinarily difficult to generalize about the corpus as a whole, we consider those three top categories – patent applications, Wikipedia, and published news – from the perspective of authorship.

Corpus Authorship: Patents, Wikipedia and Published News

Patents: Of the patent applications from patents.google.com, the top contributors were the U.S. (comprising 78% of patent applicants), EU Patent Office (5%) and Japan (5%).

- U.S.: Looking at USPTO data from 1990-2019, 13% of U.S. patent owners were women (Setty, 2022). It is possible that a patent attorney would draft the patent application, as opposed to the patent owner him/herself. Noting that "the patent bar requires a hard science background, such as a degree in engineering, chemistry, physics, or biology," USPTO found in 2020 that 22% of USPTO registered attorneys/agents were women (Spector & Brand, 2020).
- Europe: In 2019, the percentage of female patent owners as measured by the European Patent Office reached 13.2% (European Patent Office, 2022). A study by the European Parliament found that the percentage of female lawyers had risen from 35% in 2004 to 43% in 2015; they did not break down the numbers by legal specialty.
- Japan: The data from Japan is even worse. Using data from 1998-2017, only 6% of Japanese patents were held by women (Stylianou & Guibourg, 2019). As of 2012, 13% of Japanese patent lawyers were women (Benrishi, 2013).

Regardless of country of origin, patent ownership and patent law are male dominated. We use the U.S. numbers to estimate female authorship of patent applications, noting that that is likely overly generous

(only Russia, France, Taiwan and China exceed the U.S. on female patent ownership. France is rolled into the EU data; the other three comprise three percent or less of the applications in patents.google.com.

Wikipedia: Wikipedia is deeply gender biased in terms of authorship – globally, a mere 9% of editors are women (Torres, 2016). Given that Wikipedia is represented separately in ChatGPT-3 training data, duplicative content may have been filtered out – unlike a book, however, Wikipedia is a living document such that the Common Crawl corpus may well retain achieved Wikipedia content.

Published news: A 2010 study by the International Women's Media Foundation (IWMF) surveyed 522 media companies across 59 countries and found that 36% of reporters were women (Byerly, 2011). Within the U.S., the data is conflicting: using data from 2012-2016, the Worlds of Journalism study found that 27% of U.S. reporters are female (Vos & Craft, 2016). In 2023, however, Pew found that 46% of reporting journalists are women (the survey also documents significant variance between beats – women comprising 15% of sports journalists, but 64% of health journalists) (Pew Research Center, 2023). While acknowledging the variance, we have opted to use the IWMF number: it has the benefit of capturing non-U.S. reporters while averaging the two metrics for the U.S. market.

Summary/Assumptions: The fact that the top 25 contributing websites comprise roughly one percent of C4 content makes us hesitant to draw broader conclusions. However, if we assume the top three sources are broadly representative of the dataset,[12] we estimate that 19% of Common Crawl text (13% patent owners, 9% Wikipedia editors, 37% published news – all weighted equally) is authored by women.

The necessary guess work and assumptions involved in generating this number drive home a point made elsewhere: transparency on training data is essential if we are to understand the biases of LLMs. There are legitimate queries to be raised about training data, gender of authorship being one of them. If, however, information of the data's composition is not available, those questions remain unanswerable – and resulting harms are obscured.

Webtext2:

What is it?

The 22% of training data from Webtext2 is described as the links from upvoted Reddit posts – as Open AI reasons, scrapping links from upvoted posts served as "a heuristic indicator for whether other users found the link interesting, educational, or just funny" (Radford, 2019). Although ChatGPT-3 does not profess to utilize Reddit's actual message board content in training its model, Reddit's recent efforts to monetize this scraping suggest that other models do (Isaac, 2023). Reddit's value to generative AI models, CEO Steve Huffman stated, is "authentic conversation … there's a lot of stuff on the site that

[12] We recognize that there are several underlying assumptions here, made out of practicality in the face of data limitations. As we have little to no insight into the remaining ~99% of data comprising Common Crawl, we opt to treat the dataset we do have as representative. Some textual content from the internet could not easily be attributed to single author – company websites, for example. Academic articles would also be difficult to categorize, given that many list multiple authors. Blog based content, to the extent it is present, would tend towards female authorship. While one could substitute alternative assumptions here, we believe it is reasonable – in the face of data limitations – to treat the top 25 sources as broadly representative.

you'd only ever say in therapy, or A.A., or never at all." Put this way, we might well be asking whether Reddit users would consent to their data being used to train AI.

Corpus Authorship:

Reddit is known to be a male dominated space: in 2016, Pew found that 31% of Reddit users are women (Barthel, 2016). Knowing this, the Webtext2 data would be disproportionately informed by content Reddit's male users considered worth sharing. Although male users can and do share female-authored content, the content captured in this training dataset underrepresents female interests and perspectives by default of the gender breakdown of Reddit users.

Books1/Books2

What is it?

Of the data used to train ChatGPT-3, the Books1 and 2 corpuses remain the least understood (and have opened the company to legal challenges). Open AI makes no mention of the composition/source of these datasets. In attempts to recreate Books1, researchers posit that is the now defunct BooksCorpus, a dataset consisting of roughly 11,000 free books scraped from Smashwords.com, a site that describes itself as "the world's largest distributor of indie [or self-published] ebooks" (Bandy, 2021). Noting that self-published books do not go through the standard editing process, public libraries have raised concerns that they are a gateway for the uncontrolled dissemination of conspiracy theories and other low-quality content (Woodcock, 2022). Even before digging into the authorship, the content sets off red flags.

A 2021 paper on "documentation debt" in BookCorpus training data finds that "no documentation exists about the dataset's motivation, composition, collection process" (Bandy & Vincent, 2021). The paper goes on to note that BookCorpus violated copyright law,[13] contains significant duplicate text and skews in genre representation. The paper also raises concerns that the online self-publishing industry favors erotica and fanfiction writing, which tends to reinforce gender roles through sexualized content. Although it may not be necessary to train AI on the great American novel, the data feeding BookCorpus are far from representative of broader literary works.

The source of Books2 remains undisclosed; some suspect it comes from "shadow libraries" hosting pirated content, including copyrighted books, textbooks and academic papers (Ulea, 2023). One of several recent lawsuits alleges that "ChatGPT generates summaries of Plaintiffs' copyrighted works—something only possible if ChatGPT was trained on Plaintiffs' copyrighted works" (Poritz, 2023). Hopefully ongoing legal challenges will lead to further disclosure of the source data for these corpuses, although suggestions that developers have, in some cases, opted not to document training data raises the question of whether willful ignorance is a sufficient defense.

Corpus Authorship:

The publishing world has recently reached gender parity, as compared to 1960 when 18% of published books were authored by women (Waldfogel, 2023). The self-publishing industry actually has an

[13] Other LLMs, including LLaMA, utilize the Project Gutenberg dataset, which is comprised of published books now in the public domain. The contents, by default, tend to be older books and from an era dominated by male authors.

overrepresentation of women, although that may be diluted by the finding that BooksCorpus skews to a limited number of prolific authors,[14] raising broader questions about its representativeness. In the absence of other relevant details about the composition of Books1 and 2 datasets, such as publishing dates, we estimate that women contributed 50% to these corpuses. If striving for a corpus characterized by equitable gender authorship, recently published books fit the bill (albeit, at the expense of intellectual property law).

Wikipedia:

What is it?

Wikipedia is an open access encyclopedia and one of the most visited websites both domestically and internationally. As part of the open movement, articles are contributed by volunteer editors. As previously discussed, text drawn from Wikipedia also populates the Common Crawl corpus. Duplicate text is removed during the filtering process; however, as Wikipedia content is constantly evolving, we presume that the actual content of this corpus varies from the Wikipedia content that feeds into Common Crawl.

Wikipedia "has been criticized for exhibiting systemic bias, particularly gender bias against women and geographical bias against the Global South" (Wikipedia, 2023). Like the other training data used to train ChatGPT3 Wikipedia's content skews heavily male. What is of particular note, however, is Wikipedia's proactive efforts to diversify authorship.

Corpus Authorship:

Available data concerning Wikipedia authorship is dated: as of 2011, only 9% of global Wikipedia editors were women, 15% in the US (Torres, 2016). Consequently, we would expect the platform's content to reflect a male bias. However, Wikipedia offers a model for addressing (albeit slowly) representation and inclusion of women in content creation. Grassroots editing collectives like Women in Red are seeking to redress the "content gender gap," driving towards a resource that is "more representative of human knowledge" (WikiEdu, 2020). In eight years, the group has increased the percentage of Wikipedia biographies about women from 15% to 20% (Wikipedia: WikiProject Women in Red, 2023). A like-minded group collaborated with historians and archivists at the National Archives and Records Administration to rewrite the Wikipedia article about the 19th Amendment with a more balanced (and less male centric) narrative. Wikipedia has launched educational programs, supported grassroots training, and invested in outreach programs that encourage diverse editors to join the community (WikiEdu, 2020). With this concerted effort, representation is slowly growing.

Ultimately, Open AI and others LLM developers could learn from Wikipedia's example. Rather than accepting homogenous data as inevitable, they could invest in representative data creation, selection, and usage. In other words, we need not (and should not) settle for biased data.

[14] "Among free books in Smashwords21, the top 10% of authors by word count were responsible for 59% of all words in the dataset" (Bandy, 2021)

Conclusion:

In this paper we have argued the following:

1) The authorship of LLM training data, looking specifically at ChatGPT-3, is not sufficiently diverse to be representative of women authored text.
2) The obscurity of LLM training data makes it impossible to precisely quantify the percentage of texts authored by women. This, in and of itself, is illustrative of the shortcomings of LLM training documentation and transparency.
3) Our estimates suggest women authored texts are underrepresented to the point of impacting the quality of LLM outputs. This is supported by numerous documented examples of LLM's perpetuating gender stereotypes and otherwise generating harmful, sexist content (gender biased hiring algorithms, sexualized avatars, text to picture career prompts, etc.)
4) Using non-representative training data for AI models risks entrenching, even obscuring, enduring cultural biases.

We conclude by urging developers – and policymakers – to make representation a central consideration in the curation of training corpuses. Representative training data would extend beyond gender: for example, others have raised issue that the dominance of English language sources overrepresents the perspectives of English speakers (Vashee, 2023). And of course, relying on web-based content completely omits the perspective of those in developing countries impacted by the digital divide. These biases are just some of the very real limits to generative AI systems as they are currently trained and deployed. As Gururangan et al. assert, "laissez-faire data collection (i.e., filtering large web data sources) leads to data homogeneity."

We fully expect that some readers will take issue with how we collected, analyzed, and presented our estimate that 26.5% of ChatGPT-3 training data was authored by women. While we acknowledge the limitations of our methods, they only reinforce the most salient parts of our argument. The lack of transparency surrounding training data makes analyses like ours exceedingly difficult, thus barring researchers, policy makers, and members of the public from understanding fully how these models work. Transparency is widely accepted as a necessary condition to hold governments accountable; similar thinking informs required disclosure related to salaries, campaign funding and real estate sales. As we are approaching an era where political ads, marketing copy, clickbait articles, and even legislation all contain AI generated content, transparency regarding how these models are trained will help us analyze, in an informed way, their public impact. It will allow us to mitigate harm and accentuate the very real benefits such systems can offer society. The impact of any new technology depends on how we use it. The fundamental question is: will we use AI to automate the world as it is, or to move us closer to the world as we wish it were?

We thank the following reviewers for their thoughtful feedback and recommendations: Jennifer Keating, Associate Professor of English, University of Pittsburgh; Yu-Ru Lin, Associate Professor, University of Pittsburgh School of Computing and Information and Research Director, University of Pittsburgh Institute for Cyber Law, Policy, and Security; Beth Schwanke, Executive Director, University of Pittsburgh Institute for Cyber Law, Policy, and Security; and Annette Vee, Associate Professor of English, University of Pittsburgh.

References

Argamon, S., Koppel, M., Shimoni, A. (2003). Gender, genre, and writing style in formal written texts. *Text - Interdisciplinary Journal for the Study of Discourse*, *23*(3). https://doi.org/10.1515/text.2003.014

Baack, S. (2023, August 2). *The human decisions that shape generative AI*. Mozilla Foundation. https://foundation.mozilla.org/en/blog/the-human-decisions-that-shape-generative-ai-who-is-accountable-for-what/

Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, *18*(2), 135–160. https://doi.org/10.1111/josl.12080

Bandy, J. (2022, January 6). Dirty secrets of BookCorpus, a key dataset in machine learning. *Medium*. https://towardsdatascience.com/dirty-secrets-of-bookcorpus-a-key-dataset-in-machine-learning-6ee2927e8650

Bandy, J. (2021, May 11). Addressing "documentation debt" in machine learning research: A retrospective datasheet for BookCorpus. arXiv.org. https://arxiv.org/abs/2105.05241

Barthel, M. Stocking, G. Holcomb, J. & Mitchell, A. (2020, August 27). *Reddit news users more likely to be male, young and digital in their news preferences.* Pew Research Center's Journalism Project. https://www.pewresearch.org/journalism/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/

Benrishi Increase in the Number of Japanese Patent Attorneys (2013). http://www.benrishi.com/en/patentattorney/pat_introduction1.html#:~:text=Out%20of%20the%20said%20total,933%20(12.1%25)%20for%20female

Bommasani, R. Klyman, K. Zhang, D. Liang, P. (2023, June 15). *Do foundation model providers comply with the draft EU AI Act?* Stanford Center for Research on Foundation Models. https://crfm.stanford.edu/2023/06/15/eu-ai-act.html

Brown, T. B. (2020, May 28). *Language models are few-shot learners*. arXiv.org. https://arxiv.org/abs/2005.14165

Byerly, Carolyn M. (2011). *Global report on the status of women in the news media*. International Women's Media Foundation. https://www.iwmf.org/wp-content/uploads/2018/06/IWMF-Global-Report.pdf

Colley, A., & Todd, Z. (2002). Gender-Linked differences in the style and content of E-Mails to friends. *Journal of Language and Social Psychology*, *21*(4), 380–392. https://doi.org/10.1177/026192702237955

Dodge, J. Sap, M. Marasović, A. Agnew, W. Ilharco, G. Groeneveld, D. Mitchell, M. & Gardner, M. (2021). Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/2021.emnlp-main.98

European Parliament. (2017, August). *Mapping the representation of women and men in legal professions across the EU*. Policy Department for Citizen's Rights and Constitutional Affairs. https://www.europarl.europa.eu/RegData/etudes/STUD/2017/596804/IPOL_STU(2017)596804_EN.pdf

European Patent Office. (2022). *EPO - Fewer than 1 in 7 inventors in Europe are women*. https://www.epo.org/news-events/news/2022/20221108.html

Feast, J. (2020, October 8). 4 Ways to address gender bias in AI. *Harvard Business Review.* https://hbr.org/2019/11/4-ways-to-address-gender-bias-in-ai

Feng, S., Park, C. Y., Liu, Y., & Tsvetkov, Y. (2023). From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. *https://arxiv.org/abs/2305.08283*. https://doi.org/10.18653/v1/2023.acl-long.656

GPT-4 Technical Report (2023, March 27). OpenAI. https://cdn.openai.com/papers/gpt-4.pdf

Gururangan, S. (2022, January 25). *Whose language counts as high quality? Measuring Language Ideologies in Text Data Selection*. arXiv.org. https://arxiv.org/abs/2201.10474

Halpern, S. (2023, March 28). What we still don't know about how A.I. is trained. *The New Yorker.* https://www.newyorker.com/news/daily-comment/what-we-still-dont-know-about-how-ai-is-trained

Heikkilä, M. (2022, December 14). The viral AI avatar app Lensa undressed me—without my consent. *MIT Technology Review*. https://www.technologyreview.com/2022/12/12/1064751/the-viral-ai-avatar-app-lensa-undressed-me-without-my-consent/

Heikkilä, M. (2023, April 20). OpenAI's hunger for data is coming back to bite it. *MIT Technology Review.* https://www.technologyreview.com/2023/04/19/1071789/openais-hunger-for-data-is-coming-back-to-bite-it/

Heikkilä, M. (2023, July 20). Meta's latest AI model is free for all. MIT Technology Review. https://www.technologyreview.com/2023/07/18/1076479/metas-latest-ai-model-is-free-for-all/

Hunt, E. (2019, September 9). Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter. *The Guardian*. https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter

Isaac, M. (2023, April 18). Reddit wants to get paid for helping to teach big A.I. systems. *The New York Times*. https://www.nytimes.com/2023/04/18/technology/reddit-ai-openai-google.html

Li, H. (2023). Data scraping makes AI systems possible, but at whose expense? *Tech Policy Press*. https://techpolicy.press/data-scraping-makes-ai-systems-possible-but-at-whose-expense/

Liesenfeld, A., Lopez, A., & Dingemanse, M. (2023). Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators. *CUI 2023: Proceedings of*

the 5th International Conference on Conversational User Interfaces. https://doi.org/10.1145/3571884.3604316

Mathewson, J. (2020, October 19). *10 years of helping close Wikipedia's gender gap*. Wiki Education. https://wikiedu.org/blog/2020/10/14/10-years-of-helping-close-wikipedias-gender-gap/

Nolan, M. (2023, July 28). Llama and ChatGPT are not Open-Source. *IEEE Spectrum*. https://spectrum.ieee.org/open-source-llm-not-open

Perez, C. C. (2019). *Invisible Women: Data bias in a world designed for men.* https://www.amazon.com/Invisible-Women-Data-World-Designed/dp/1419729071

Pew Research Center. (2023, April 4). *US journalists' beats vary by gender, employment status, race and ethnicity*. https://www.pewresearch.org/short-reads/2023/04/04/us-journalists-beats-vary-widely-by-gender-and-other-factors/

Poritz, I. (2023, June 29). OpenAI legal troubles mount with suit over AI training on novels. *Bloomberg Law*. https://news.bloomberglaw.com/ip-law/openai-facing-another-copyright-suit-over-ai%20training-on-novels

Porter, J. (2023, April 12). China wants homegrown AI to reflect the core values of socialism. *The Verge.* https://www.theverge.com/2023/4/12/23680027/china-generative-ai-regulations-promote-socialism-chatgpt-alibaba-baidu

Radford, A. (2019). Language models are unsupervised multitask Learners. *Open AI.* https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

Reducing bias and improving safety in DALL·E 2. (2022, July 18). https://openai.com/blog/reducing-bias-and-improving-safety-in-dall-e-2

Reproducing inequality: How AI image generators show biases against women in STEM | United Nations Development Programme. (2023, April 3). UNDP. https://www.undp.org/serbia/blog/reproducing-inequality-how-ai-image-generators-show-biases-against-women-stem

Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. W. (2005). Effects of age and gender on blogging. *National Conference on Artificial Intelligence*, 199–205. http://languagelog.ldc.upenn.edu/myl/ldc/schler_springsymp06.pdf

Setty, R. (2022, October 19). Women represent 13% of US patent owners after 30 years of growth. *Bloomberg Law.* https://news.bloomberglaw.com/ip-law/women-represent-13-of-us-patent-owners-after-30-years-of-growth

Spector, Elaine and Brand, Latia. (2020, September 16). A data analysis of diversity in the patent practice by technology background and region. *Landslide Magazine*, *13*(1). https://www.uspto.gov/sites/default/files/documents/Landslide_Diversity_Article_September.pdf

Stylianou, B. C. G. a. N. (2019, October 1). Why are so few women inventors named on patents? *BBC News.* https://www.bbc.com/news/technology-49843990

Tiku, K. S. S. Y. C. N. (2023b, April 19). See the websites that make AI bots like ChatGPT sound so smart. *Washington Post*. https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/?itid=lk_inline_manual_25

Torres, N. (2016, June 2). *Why do so few women edit Wikipedia?* Harvard Business Review. https://hbr.org/2016/06/why-do-so-few-women-edit-wikipedia

Traylor, Jake. (2022, July 27) AI coming to life? Google engineer claims chatbot is sentient. *NBC News.* https://www.nbcnews.com/tech/tech-news/no-quick-fix-openais-dalle-2-illustrated-challenges-bias-ai-rcna39918

Treude, C. (2023, March 17). *She elicits requirements and he tests: Software engineering Gender bias in large language models*. arXiv.org. https://arxiv.org/abs/2303.10131

Ulea, A. (2023, July 10). US comedian Sarah Silverman joins authors in suing Meta and OpenAI over copyright infringement. *Euronews*. https://www.euronews.com/culture/2023/07/10/us-comedian-sarah-silverman-joins-authors-in-suing-meta-and-openai-over-copyright-infringe

Vashee, K. (2023). Making Generative AI effectively multilingual at scale. *ModernMT Blog*. https://blog.modernmt.com/making-generative-ai-multilingual-at-scale/

Vee, A. (2022). Automated Trolling: The Case of GPT-4Chan When Artificial Intelligence is as Easy as Writing. *Interfaces Essays and Reviews in Computing and Culture*, vol 3. https://cse.umn.edu/cbi/interfaces#edgelordy

Vincent, J. (2022, June 8). YouTuber trains AI bot on 4chan's pile o' bile with entirely predictable results. *The Verge*. https://www.theverge.com/2022/6/8/23159465/youtuber-ai-bot-pol-gpt-4chan-yannic-kilcher-ethics

Vos, T., and Craft, S. (2016, July 31). Journalists in the United States. *Worlds of Journalism Study*. https://epub.ub.uni-muenchen.de/34878/1/Country_report_US.pdf

Waldfogel, J. (2023, February). *The Welfare Effect of Gender-Inclusive Intellectual Property Creation: Evidence from Books*. National Bureay of Economic Research. https://doi.org/10.3386/w30987

Webz. (2023, March 22). Large language models: What your data must include. *Webz.io.* https://webz.io/blog/machine-learning/large-language-models-what-your-data-must-include/

Wikipedia. (2023, July 29). *Wikipedia*. https://en.wikipedia.org/wiki/Wikipedia

Wikipedia. (2023, July 29) *Wikipedia:WikiProject Women in Red*. Wikimedia Foundation. https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Women_in_Red

Woodcock, C. (2022, April 20). Ebook Services Are Bringing Unhinged Conspiracy Books into Public Libraries. *Vice*. https://www.vice.com/en/article/93b7je/ebook-services-are-bringing-unhinged-conspiracy-books-into-public-libraries

UNIVERSITY OF PITTSBURGH

INSTITUTE FOR CYBER LAW, POLICY, AND SECURITY