Gender Representation Across Online Retail Products

Dana Pessach
Amazon.com
Tel-Aviv, Israel
danapess@amazon.com

Barbara Poblete
Department of Computer Science, University of Chile
Santiago, Chile
Amazon.com
Seattle, WA, USA
barbara@poblete.cl

ABSTRACT

We present a broad characterization of gender representation in a large heterogeneous sample of retail products. In particular, we study online product textual information, such as titles and descriptions. Our goal is to understand from a semantic perspective, differences and similarities in how girls (women) and boys (men) are represented. We perform a comparative analysis of the language used in gendered products (i.e., products that mention exclusively either of these two genders), and additionally compare it to products that are explicitly gender neutral or inclusive. We found that the adjectives, skills, occupations, and values described in gendered products tended to reinforce stereotypes. Some of these stereotypes are aligned with historical findings from research on traditional off-line retail stores, and others are new owing to the up-to-date product dataset our research is based on. By leveraging additional existing resources we were able to gain insight into how certain product descriptions reflect stereotypes that are related to soft-skills and hierarchical occupational information. Conversely, we found that a large segment of products present explicitly as gender neutral or inclusive. We explore whether the language used by gender-inclusive products can be useful to improve stereotypes reflected in gendered product text. Specifically, we study its effect in word embedding fairness through debiasing techniques.

CCS CONCEPTS

• **Information systems** → **Electronic commerce**; **Information retrieval**.

ACM Reference Format:
Dana Pessach and Barbara Poblete. 2024. Gender Representation Across Online Retail Products. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 3–6, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3630106.3658947

1 INTRODUCTION

Increasingly, people want their e-commerce purchases to be socially responsible. This effort does not only apply to customers, but also to manufacturers. For example, it is more and more common to find products that identify as "climate pledge friendly" [3] or as sustainable options. Similarly, we can see other emergent societal drives for responsible purchases. One such drive is parents and individuals seeking products that present an equitable gender representation or that are gender neutral [67]. This trend has been met by important commercial brands that have, for instance, vowed to remove gender stereotypes from their products [70], and large retailers moving away from gender-based product classifications [36]. Even state legislators have pushed to require stores to have non-gendered toy sections [1, 12].

The preference towards more equitable and neutral gender representation has emerged over time as stereotypes are becoming challenged worldwide. This is supported by decades of research that evidenced significant under-representation of women in the media [25], as well as disparities in gendered toys in stores [70, 72]. For example, finding from these studies have shown that toys traditionally targeted towards girls had higher tendency of reinforcing stereotypes related to physical appearance and domestic skills, while toys for boys were found more belligerent or more STEM oriented.

Gender differences, however, are not limited to toys and have been well-documented across different areas. In particular, systems that make use of large language models (e.g., BERT [44], GPT-2/3/4 [4, 60]) have shown to encode gender bias and as well as other social biases [2, 16, 68, 79]. To illustrate how large language models can reflect gender stereotypes in children's toys, Fig. 1 shows responses provided by ChatGPT [2] to the questions *"Which toys are relevant for developing skills for little girls/boys?"*[1] The generated answer displays clear differences based on gender, suggesting items such as *dolls*, *kitchen sets*, and *arts & crafts* for girls, and *cars*, *action figures*, and *sports equipment* for boys.

Nowadays, the vast offer of toys and other retail products promoted and sold online constitute an unprecedented opportunity to better understand how gender is currently represented in retail. This insight is particularly useful and necessary, as information retrieval and recommendation systems are increasingly making use of large language models, including domain-specific applications, such as e-commerce [10, 11, 32].

In this work, we present a characterization of how gender is represented across a large heterogeneous sample of online products. Overall, we analyze textual data such as product title and description, from roughly 8, 9 million products available on the Web. Our approach is not focused on specific products, but rather on types of products and their product categories. We center our research mostly on the language and semantics used in the text

[1] Results captured on 2022-12-15.

FAccT '24, June 3–6, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658947

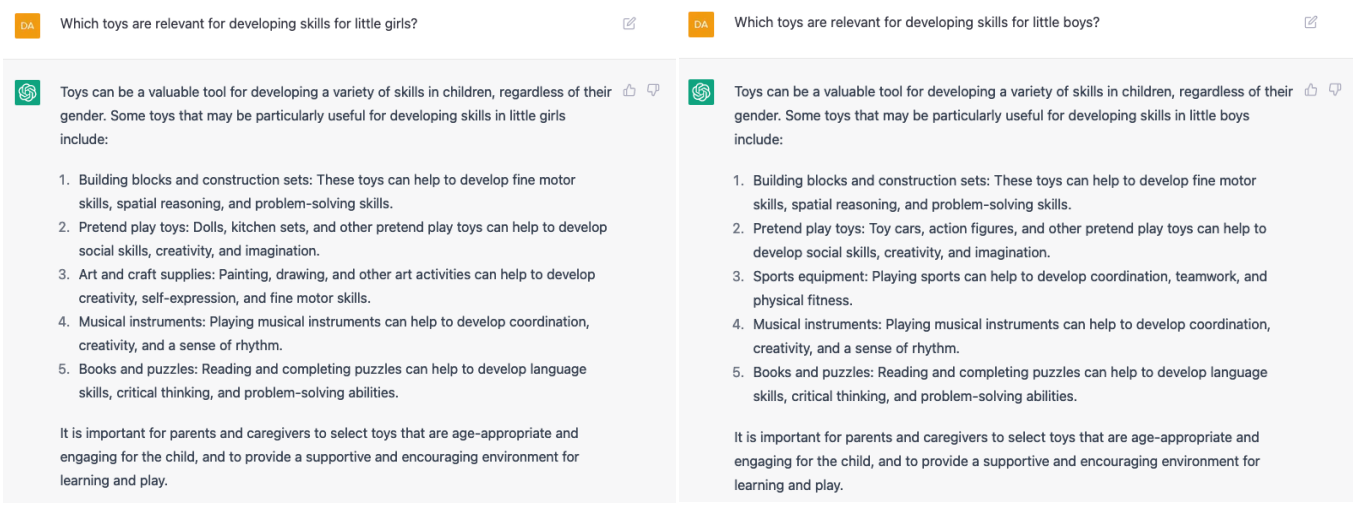


Figure 1: ChatGPT responses to which toys are relevant for developing skills for little girls (left), and for little boys (right).

that describes products claiming to be specific to a particular gender. Similarly to prior work, we study children's toys, however, we further expand that scope to include adult's products as well. In addition, we explore potential biases that could emerge in language models that were to be derived exclusively or partially from such data. Furthermore, we investigate how gender neutral products could be leveraged to mitigate such effects using existing debiasing techniques.

Our findings show that in the sample of children's products that were, so called, "gendered" (i.e., products that mentioned exclusively one gender, or conversely, that explicitly use certain gender-inclusive terms), 45% were targeted to girls, 19% to boys, and 36% were gender inclusive. In adult's products this distribution corresponded to 51% products for women, 31% for men, and 28% were gender inclusive. It is important to note that our dataset gender neutral products are underrepresented, since we mostly focused on studying gender when it is explicitly mentioned in text (more on this in Section 6). Additionally, we observed that for children's products most gender differences were concentrated in certain product categories, such as *video games* and *sports* favoring more boys, and *beauty* and *jewelry* favoring more girls. Moreover, toys such as *toy cars* and *toy guns*, alluded more often to boys, while *dolls & accessories* and *arts & craft kits* more to girls. On the other hand, gender-inclusive products were more frequent among *puzzles* and *educational toys*. Among other things, we also looked into soft skills that were mentioned in product text, observing that products for girls more often mentioned concepts such as "creativity", "confidence", "responsibility", "paying attention", "compassion", "communication", "planning", "social skills", "listening", and "empathy". For boys more common concepts were "flexibility", "creative thinking", "interacting", "critical thinking", "teamwork", and "problem solving".

As part of our analysis, we studied bias that emerges when gendered product data is used to train word embeddings. We use this as a means to gain insight into how gender stereotypes may transfer to any other type of language model when focusing on the product domain. We show how gender-inclusive data can effectively be used to mitigate bias, without apparently distorting utility.

Overall our study provides an unprecedented quantitative view of the language used to describe retail products, such as children's toys and other traditionally gendered retail items. Our work contributes by broadly expanding the scope of prior work and providing new insight into gender representation. This insight can help, for example, inform design choices when creating *natural language processing* (NLP) tools for various domains, specially e-commerce.

2 BACKGROUND AND RELATED WORK

We discuss briefly some relevant advances in the areas that are most related to our current study. This overview is provided as context for our current work, but is by no means complete due to space restrictions and the extensive amount of qualitative social science's research on children's toys.

2.1 Gender Representation

Women have been traditionally underrepresented in the media, specifically in television and in movies [15, 72] Similarly, decades of qualitative research on toys for girls and toys for boys show important gender-based differences. As early as 1975 Rheingold and Cook [64] observed that boys had more spatial-temporal toys, including STEM toys, sports equipment and military toys. Girls on the other hand, had dolls and doll accessories, including other domestic items, such as stoves and dishes. Other, more recent works, have come across very similar findings without significant changes [17]. Nevertheless, studies involving parents have shown that these increasingly desire gender neutral toys for their children [45, 48]. Gender-stereotypical references in products have been shown to have effects on children's beliefs, as well as adults' gender-based expectations [27, 33, 34, 47].

In relation to toys promoted and/or sold online, Auster & Mansbach [7] studied in 2012 gender marketing on the Disney store website for approximately 600 toys. This study revealed that toys

were targeted as follows - *"Bold colored toys, predominantly red, black, brown, or gray toys, and those that were action figures, building toys, weapons, or small vehicles typified toys for boys. Pastel colored toys, predominantly pink or purple toys, and those that were dolls, beauty, cosmetics, jewelry, or domestic-oriented typified toys for girls".* In addition, Azmi et al. [8] studied 87 social media advertisement posted during 2019 on Mattel's[2] official Facebook page, yielded similar findings as well as more marketing targeted to girls. Raj and Ekstrand [62] looked at e-commerce search results and query suggestions for toys. They reported gender differences, similar to those discussed in prior work, depending on the gender used in the search query.

Our work differs from prior research, first in the scale of our study, which covers 8.9 million products, allowing us to carry out a quantitative analysis. Secondly, we study a broad range of products that not only includes toys, but other retail products targeted towards men and women. Also, by studying textual product data, we investigate an important source for potential biases in e-commerce systems that rely on language models.

2.2 Bias and Fairness in NLP and LLMs

Natural Language Processing (NLP) and Large Language Models (LLMs) have recently experienced great advancements, driven by technological development and the availability of large corpora on the Web. An increasing number of language models pretrained on extremely large datasets, are being made widely accessible to the public and to developers through open-source libraries [38]. Along with this, there is also a rise in concern regarding how such models can be prone to encode more significant biases than those based on carefully and neutrally curated data [16, 68, 75, 79].

Multiple studies evidence gender bias in natural language corpora. For instance, Wikipedia, which is a popular textual data source, suffers from important under-representation of women and other populations [71]. This disparity has also been observed in other diverse textual data sources, such as for online news articles [25], the film industry [41, 72], high-school textbooks [6], computer science education materials [54], and court decisions [59]. This has a massive impact on any language model trained from this kind of data, and their downstream applications. Examples of biased results from derived applications, have been found in machine translation [23, 31, 73], coreference resolution [65, 79, 81], sentence encoding [51], semantic role labeling [80], and recommender systems [26, 69].

One prominent line of work in the field of fairness in NLP is that of fairness in word embeddings [18, 20, 46, 51, 58]. These models are designed such that the vectors will indicate something about the meanings and relationships between words (i.e., words with similar meanings have vectors that are close in the vector space). Pretrained word embeddings, trained on vast amounts of data, are broadly used in many NLP applications, such as in search engines, machine translation, resume filtering, job recommendation systems, online reviews, and more. Previous studies have shown that, similarly to large language models, there are inherent social biases and gender stereotypes in pretrained word embedding models [18–20, 51, 58, 82].

[2]Mattel is an American toy manufacturing company, mostly known for making the "Barbie" doll.

Several recent studies have developed methods for mitigation of unfairness in NLP tasks. Techniques for improving fairness can be broadly divided into three categories: *pre-process*, *in-process* and *post-process* [58]. Pre-process techniques involve changing the training data before feeding it into the machine learning algorithm so that a subsequent model will be more fair [21, 22, 30, 49, 50, 78]. For example, [19, 81] propose pre-process mechanisms for reducing bias by perturbing or removing documents that are used for training and are traced as the origin for the word embedding bias.

In-process techniques involve modifying machine learning algorithms to account for fairness during training time [13, 14, 43, 74, 76, 77]. Zhao et. al [82] suggest an in-process mechanism, referred to as *Gender-Neutral Global Vectors (GN-GloVe)*, to reduce bias.

Post-process techniques perform post-processing of the output scores of the model to make decisions more fair [24, 28, 39, 55]. Bolukbasi et al. [18] suggest a post-process mechanism for removing gender bias, referred to as *hard-debiasing*. Their method first identifies a *gender dimension*, which is determined by a set of words that indicate gender definitions (e.g., "he"/"she"). Second, it inherently defines *neutral words* (such as occupations) and then zeroing the projection of all of these neutral words with respect to the gender direction (so that the bias of neutral words is now zero by definition) by re-embedding the word w:

$$\vec{w} := (\vec{w} - \vec{w_B})/\|\vec{w} - \vec{w_B}\| \quad (1)$$

where $\vec{w}$ is the embedding of the selected word and $\vec{w_B}$ is the projection of w with respect to the gender direction.

One concern regarding this kind of approach is that bias may not effectively removed but rather just "hidden". In particular, Gonen an Goldeberg [35] found that for debiasing methods such as hard-debiasing, gender stereotypical words might still be easy to cluster together. Moreover, they showed that implicit gender could be recovered by a downstream classifier just using the hard-debiased vectors without any additional information.

Our work is centered in the e-commerce domain, we focus on identifying and understanding bias of word embeddings trained using textual data from online products targeted to a specific gender. We also evaluate the debiasing effect of leveraging data from products that identify explicitly as gender-neutral, using pre-processing and post-processing techniques.

3 DATA DESCRIPTION

For this research, we combined different data sources, including *skills and occupations resources*, as well as *online product data*. We detail each next:

Skills and Occupations resources. We used two resources for the analysis of skills and occupations. The first is ESCO[3] a European taxonomy that classifies skills, competences and occupations. More specifically, we utilized the ESCO occupation data[4], and we used the ESCO API[5] to retrieve the skills related to each occupation. ESCO was originally designed for job market analysis [5, 29, 57] and maps different occupations, assigning skills to each occupation. We used ESCO to identify professions and skills that are mentioned in

[3]https://esco.ec.europa.eu/en/about-esco/what-esco
[4]V1.0.8 https://esco.ec.europa.eu/en/use-esco/download
[5]https://esco.ec.europa.eu/en/use-esco/use-esco-services-api

relation to gender in product text. The purpose of this was to study these associations and their differences based on gender. Secondly, and in addition to ESCO, we used a lexicon provided by Fareri et al.[29], that contains a list of soft skills. We used this lexicon to identify soft skills mentioned in text related to children's products.

Online product data. In order to study the language used in online products we created a text dataset of product titles and descriptions. Additionally, for each product we also included product category, product type, and colors if these were mentioned. The products included in our dataset are only, so called, *"gendered"* products found on the website Amazon.com.[6] Other products (i.e., non-gendered products) were excluded from this particular study. In particular, for the purpose of our current analysis we define gendered products to be those that explicitly mentions male or female genders in their text, or that, alternatively, explicitly mention certain gender-inclusive terms, such as unisex terms or both male and female genders at the same time. In particular, our dataset contains products whose title matches one of the following 6 categories: 1) **boys:** i.e., products mentioning exclusively the terms "boy" and "boys", 2) **girls:**, i.e., products mentioning exclusively "girl(s)", 3) **children's gender inclusive:** i.e., mentioning "boys" & "girls", and/or gender neutral terms like "children", "kids", and "toddlers", 4) **men:** i.e., products mentioning "men", "men's" and "man", 5) **women:** i.e., products mentioning "women('s)", and 6) **adult's gender inclusive:**, including both "men" & "women", and/or gender neutral terms such as "unisex" (in non-children's products).

Overall, our data contains roughly 8.9 million *gendered products* with text in English obtained in June, 2022. In this sample, 48% corresponded to products targeted to females, 25% to males, and 27% to gender-inclusive. Currently, the scope of our work is that of characterizing differences between male and female genders. This, as well as our focused sampling on gendered products, results in gender-neutral and non-gendered products–which are the vast majority of products–being underrepresented in our data. We discuss these and other limitations more in detail in Section 6.

In addition, although our dataset is not public as it was derived from the internal Amazon catalog data, it was sampled representatively to contain products available online publicly, and that have reviews. We expect that similar studies to ours can be conducted using publicly existing datasets of online retail products, such as that of provided by McAuley et al. [52].[7]

4 EXPLORATORY ANALYSIS

In this section we present an exploratory analysis that characterizes the products contained in our gendered dataset described in Section 3. This analysis studied differences in product groups according to the genders that are mentioned in product titles and descriptions. Specifically, we look at gender differences in the distribution of products, as well as in relation to skills and occupations. We note that due to the constraints that we placed on our dataset, this analysis only reflects characteristics of gendered products. This is further discussed in Section 6.

[6] https://www.amazon.com/
[7] https://snap.stanford.edu/data/amazon/productGraph/

4.1 Distribution of gendered products

First, we study how gender mentions are distributed in different product categories. For children's products, overall, we found that for gendered products 45% were targeted to girls, 19% to boys, and 36% were gender inclusive. Figure 2 shows this distribution in children's product categories. We observe that an important portion of products are gender inclusive (i.e., label "both" in the figure). However, some product categories had considerably more exclusive mentions to boys (i.e., label "boys") than to girls (i.e., label "girls"), these were *video games* and *sports*. For girls, these categories were in *beauty* and *jewelry*. In the case of gender inclusive products, these were found in higher proportion in *tools* and *musical instruments*. In addition, the most frequent colors in boy's products were *black*, *"shark"*, *"camo gray"*, and *navy*, whereas in girl's products these were *pink*, *rose*, *red*, *gold*, *"mermaid"*, *"floral"* and *"rainbow"*.

Figure 3 shows a similar breakdown that only takes into consideration children's *toys*. We observe more disproportionate exclusive mentions to boys in *toy vehicles (cars)* and *toy guns*. In contrast, for girls these products are *dolls & clothing* and *arts & craft kits*. The most relative mentions for gender inclusive, on the other hand, were found in *swings*, *board games* and *puzzles*.

In addition, we identified the most distinctive terms for each gender, using class based TF-IDF analysis [37, 42, 66]. This yielded the following terms characterized more exclusively girls: *dress, princess, unicorn, pink, doll, headband, mermaid, accessories, little, rose, fancy, purse, clothes, purple, skirt, bag, hair, rainbow, jewelry*. The following terms characterized more exclusively boys: *car, dinosaur, truck, remote-control, blue, construction, vehicle, monster, building, superhero, speed, video game*. For gender inclusive, the most characteristic terms were: *fidget toy, educational toy, stress-relief, outdoor, animal, wooden, water, fun, sensory, learning*.

We also studied the distribution of gendered products for adults, shown in Figure 4. Here, men are more represented in *video games*, *(digital) music/videos*, *apps*, *automotive (cars)* and *watches*. Women, on the other hand, in *beauty*, *jewellery*, *electronics*, *baby*, *cameras*, *furniture*, *home & kitchen*, and *luggage (bags)*.

4.2 Skills and Occupational Gender Differences

We characterize gender differences according to professional groups and types of skills. We leveraged ESCO, described in Section 3, to extract hierarchical occupational information (i.e., occupations and occupation groups) from gendered products.

For each occupation in ESCO we computed the percentage of products that mentioned that occupation from each of our product classifications for adults: *women*, *men* and *gender inclusive*. Table 1 shows the results of this process for occupations that were found to have gender bias in this analysis. We observe that occupations with stereotypical association to women were mentioned more in products associated to this gender. These professions included: *nurse, teacher, fashion designer, artist, and cleaner*. Similarly, stereotypical male professions, such as *barber, hunter, pilot, and carpenter* were also mentioned more in products that identified as being for men. In addition, some occupations that were equally directed to both genders were, for example, *scientist* and *barista*.

For example, when considering only products that mentioned the occupation "teacher", 11.96% of products were targeted to women,

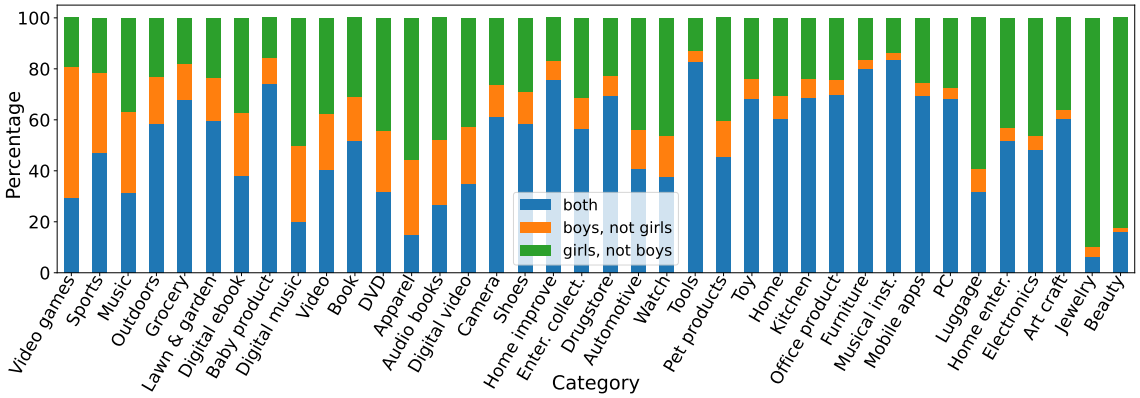


Figure 2: Distribution of gendered products in different product categories for children. Overall, girls were much more mentioned than boys, however this varied depending on the product category. For example, for *video games* and *sports*, boys had the highest representation in relation to girls.

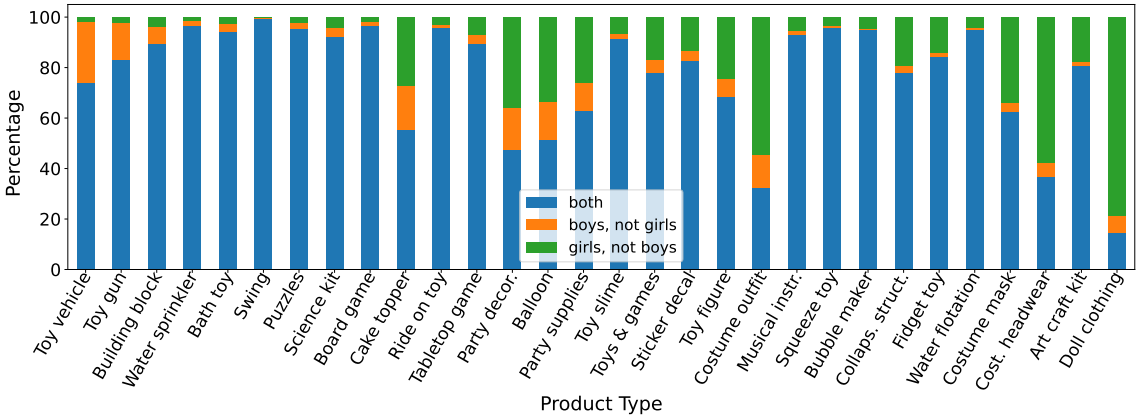


Figure 3: Distribution of gendered products within children's toys. Most exclusive mentions to boys are concentrated in *toy vehicles* and *toy guns*. For girls, these are *dolls & clothing* and *arts & craft kits*.

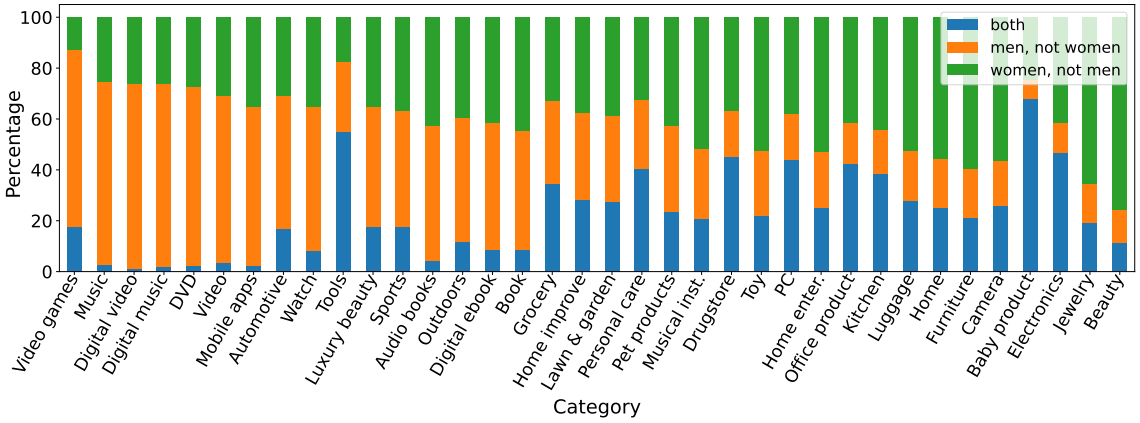


Figure 4: The distribution of gendered products for adults.

Table 1: Occupation differences based on mentions in gendered products.

Occupation	Occupation Group	% both	% men	% women	% diff	Ratio	Bias
real estate agent	Business and administration associate professionals	0.15%	0.01%	0.11%	0.10%	1141.79%	Female
accountant	Business and administration professionals	0.10%	0.02%	0.24%	0.22%	1046.96%	Female
cleaner	Cleaners and helpers	5.62%	0.95%	1.58%	0.64%	167.06%	Female
silversmith	Handicraft and printing workers	0.01%	0.11%	0.54%	0.44%	515.30%	Female
dental hygienist	Health associate professionals	0.04%	0.01%	0.09%	0.08%	1314.39%	Female
nurse	Health professionals	20.44%	0.89%	13.54%	12.65%	1518.83%	Female
doctor	Health professionals	4.63%	3.50%	6.14%	2.64%	175.42%	Female
veterinarian	Health professionals	0.13%	0.04%	0.22%	0.18%	601.07%	Female
midwife	Health professionals	0.01%	0.00%	0.07%	0.06%	2071.15%	Female
writer	Legal, social and cultural professionals	0.86%	0.37%	2.17%	1.80%	587.49%	Female
artist	Legal, social and cultural professionals	1.61%	1.91%	2.96%	1.04%	154.66%	Female
social worker	Legal, social and cultural professionals	0.15%	0.02%	0.20%	0.18%	1194.90%	Female
beautician	Personal service workers	0.04%	0.00%	0.08%	0.08%	2389.79%	Female
esthetician	Personal service workers	0.01%	0.00%	0.07%	0.06%	2071.15%	Female
fashion designer	Science and engineering professionals	0.65%	0.16%	1.45%	1.29%	899.02%	Female
teacher	Teaching professionals	6.09%	2.91%	11.96%	9.05%	410.72%	Female
plumber	Building and related trades workers, excluding electricians	0.06%	0.15%	0.01%	-0.14%	6.93%	Male
woodworker	Building and related trades workers, excluding electricians	0.16%	0.20%	0.01%	-0.19%	2.66%	Male
carpenter	Building and related trades workers, excluding electricians	0.32%	2.54%	0.08%	-2.46%	3.30%	Male
electrician	Electrical and electronic trades workers	0.32%	0.73%	0.06%	-0.68%	7.54%	Male
butcher	Food processing, wood working, garment and other craft and related trades workers	0.17%	0.26%	0.04%	-0.22%	16.93%	Male
podiatrist	Health professionals	0.25%	0.30%	0.02%	-0.28%	7.88%	Male
hunter	Market-oriented skilled forestry, fishery and hunting workers	1.35%	7.82%	2.12%	-5.71%	27.03%	Male
blacksmith	Metal, machinery and related trades workers	0.14%	0.14%	0.01%	-0.13%	5.69%	Male
welder	Metal, machinery and related trades workers	0.28%	0.43%	0.04%	-0.39%	8.51%	Male
rigger	Metal, machinery and related trades workers	0.31%	1.03%	0.01%	-1.03%	0.51%	Male
mechanic	Metal, machinery and related trades workers	1.42%	2.99%	0.80%	-2.19%	26.64%	Male
sergeant	Non-commissioned armed forces officers	0.06%	0.42%	0.06%	-0.36%	13.69%	Male
bartender	Personal service workers	0.47%	0.35%	0.04%	-0.31%	11.27%	Male
barber	Personal service workers	4.75%	8.36%	0.71%	-7.66%	8.43%	Male
forester	Production and specialised services managers	0.04%	0.11%	0.01%	-0.10%	9.37%	Male
firefighter	Protective services workers	0.98%	1.61%	0.45%	-1.16%	27.80%	Male
handyman	Refuse workers and other elementary workers	1.60%	1.52%	0.07%	-1.44%	4.84%	Male
astronaut	Science and engineering associate professionals	0.59%	1.37%	0.85%	-0.52%	62.00%	Male
pilot	Science and engineering associate professionals	1.55%	5.11%	0.91%	-4.20%	17.89%	Male
engineer	Science and engineering professionals	0.33%	1.05%	0.17%	-0.88%	15.73%	Male

2.98% to men, and 6.09% were gender inclusive. However, in a closer look, the occupational analysis surfaced additional gender differences within certain product categories. For instance, within products related to teachers, which were more targeted to women, we found the exception in products such as *sports photos* that were targeted to male teachers. Another example was in *tools*, which mentioned more often men than women, mostly in occupations such as *handyman, carpenter, mechanic, electrician*, etc.; nevertheless, the exception to this was occupations such as *nurse*, for which tools were more targeted to women.

We also characterized genders based on skill groups associated to the occupations mentioned in products. For this, we used the ESCO API to identify all the skills that were related to each occupation found in products. From these skills we only kept those that were deemed as essential to the occupation according to the taxonomy. Then, skills were grouped based on the hierarchical skill groups in the taxonomy. We show the results of this process in Table 2, which indicates the percentage of products for each skill group based on gender. Some of the skill groups that were more present in occupations found in women's products were *"teaching and training"*, as well as *"providing health care or medical treatments," "advising and consulting," "supervising people"* and *"demonstrate consideration"*, among others. For men, these groups included *"protecting and enforcing," "handling animals," "monitoring, inspecting and testing," "using precision instrumentation and equipment"* and *"engineering and engineering trades"*, among others.

4.3 Skill Gender Differences in Children's Products

We investigate soft skill differences in gendered products for children. To achieve this, we used the soft skills lexicon presented by Fareri et al. [29] and SpaCy's [40] PhraseMatcher functionality, which allowed us to identify soft skills mentioned in products according to their gender classification.

According to this process, the leading soft skills identified in girl's toys were: *"creativity," "confidence," "responsibility," "pay attention," "compassion," "communicate," "planning," "social skills," "listening"*, and *"empathy."* For boys these were: *"flexibility," "creative thinking," "interact," "critical thinking," "teamwork"*, and *"problem solving."* In addition, leading soft skills found for gender inclusive products were: *"autonomy," "curiosity, "brainstorming," "asking questions," "strategic thinking," "self confidence," "learning"*, and *"using imagination."*

5 GENDER DIFFERENCES IN WORD EMBEDDINGS

The exploratory analysis presented in Section 4, indicates that there are significant differences in gender representation found in online products. Some of these stereotypes are aligned with historical associations of gendered children's toys in traditional retail stores, and others are newly surfaced by our study.

Given these findings, a way to further investigate these gender differences is to look into how these could potentially be transferred to language models that were to be derived from this kind of data.

Table 2: Skill groups according to occupations mentioned in gendered products, based ESCO taxonomy.

Skill Group	% Both	% Men	% Women	Diff	Ratio	Bias
teaching and training	6.01%	1.82%	6.83%	5.00%	374.65%	Female
providing health care or medical treatments	7.28%	1.57%	6.54%	4.97%	416.01%	Female
advising and consulting	5.69%	2.12%	5.11%	2.99%	241.13%	Female
supervising people	2.75%	1.46%	4.11%	2.65%	281.93%	Female
demonstrate consideration	2.82%	0.55%	2.82%	2.27%	512.42%	Female
liaising and networking	4.32%	2.33%	4.37%	2.04%	187.38%	Female
education	0.93%	0.65%	2.04%	1.39%	313.30%	Female
organising, planning and scheduling work and activities	2.04%	0.74%	2.07%	1.34%	281.84%	Female
leading and motivating	1.43%	0.65%	1.90%	1.25%	291.51%	Female
developing objectives and strategies	2.70%	1.19%	2.30%	1.11%	193.03%	Female
analysing and evaluating information and data	2.70%	1.25%	2.33%	1.08%	186.03%	Female
accessing and analysing digital data	1.56%	0.16%	1.18%	1.02%	745.51%	Female
working with others	2.72%	1.66%	2.51%	0.85%	151.49%	Female
solving problems	1.68%	0.53%	1.27%	0.74%	239.59%	Female
creating artistic, visual or instructive materials	1.21%	1.60%	2.28%	0.68%	142.49%	Female
counselling	0.81%	0.08%	0.64%	0.56%	848.46%	Female
managing information	0.79%	0.07%	0.61%	0.54%	879.16%	Female
welfare	0.78%	0.05%	0.58%	0.53%	1079.63%	Female
making decisions	0.79%	0.10%	0.60%	0.50%	604.19%	Female
adapt to change	0.80%	0.10%	0.59%	0.49%	572.41%	Female
mathematics and statistics	0.01%	0.02%	0.03%	0.01%	126.97%	
demonstrate willingness to learn	0.01%	0.01%	0.01%	0.00%	139.31%	
using digital tools to control machinery	0.02%	0.06%	0.06%	-0.00%	97.22%	
management skills	0.03%	0.04%	0.04%	-0.00%	92.60%	
tending plants and crops	0.00%	0.02%	0.00%	-0.02%	4.58%	Men
installing interior or exterior infrastructure	0.02%	0.06%	0.00%	-0.06%	4.98%	Men
deal with uncertainty	0.02%	0.10%	0.00%	-0.09%	2.72%	Men
driving vehicles	0.08%	0.28%	0.05%	-0.23%	17.89%	Men
installing, maintaining and repairing electrical, electronic and precision equipment	0.11%	0.27%	0.03%	-0.24%	11.35%	Men
positioning materials, tools or equipment	0.09%	0.40%	0.06%	-0.34%	14.82%	Men
operating mobile plant	0.04%	0.36%	0.01%	-0.35%	3.81%	Men
working with machinery and specialised equipment	0.23%	0.49%	0.11%	-0.38%	21.57%	Men
operating aircraft	0.12%	0.54%	0.08%	-0.46%	14.37%	Men
architecture and construction	0.30%	0.57%	0.09%	-0.48%	16.16%	Men
forestry	0.10%	0.82%	0.18%	-0.64%	21.43%	Men
agriculture	0.08%	0.93%	0.05%	-0.88%	5.57%	Men
installing, maintaining and repairing mechanical equipment	0.20%	1.04%	0.15%	-0.89%	14.57%	Men
building and repairing structures	0.28%	1.10%	0.14%	-0.96%	12.80%	Men
law	0.63%	2.12%	0.97%	-1.15%	45.85%	Men
biological and related sciences	0.20%	1.38%	0.18%	-1.20%	13.33%	Men
environment	0.20%	1.64%	0.35%	-1.29%	21.47%	Men
transport services	0.34%	1.59%	0.22%	-1.37%	13.89%	Men
personal services	0.90%	1.73%	0.31%	-1.41%	18.21%	Men
engineering and engineering trades	0.70%	1.82%	0.36%	-1.46%	20.02%	Men
using precision instrumentation and equipment	0.64%	2.04%	0.51%	-1.53%	25.20%	Men
monitoring, inspecting and testing	1.60%	3.89%	1.49%	-2.40%	38.22%	Men
handling animals	0.62%	4.24%	0.73%	-3.51%	17.23%	Men
protecting and enforcing	13.36%	19.87%	13.65%	-6.22%	68.71%	Men

As a proof-of-concept of that idea, in this section we study word embeddings trained on different portions of our gendered dataset. This approach has three advantages, the first, that it allows us to gain insight into biased terms by exploring them in the embedding space. The second, that it can be used as a proxy to understand biases that could potentially transfer into other kinds of language models (e.g., generative language models, among others) and their downstream applications. The third advantage, is that we can explore the use of gender inclusive curated data as means to debias, using existing word embedding debiasing techniques.

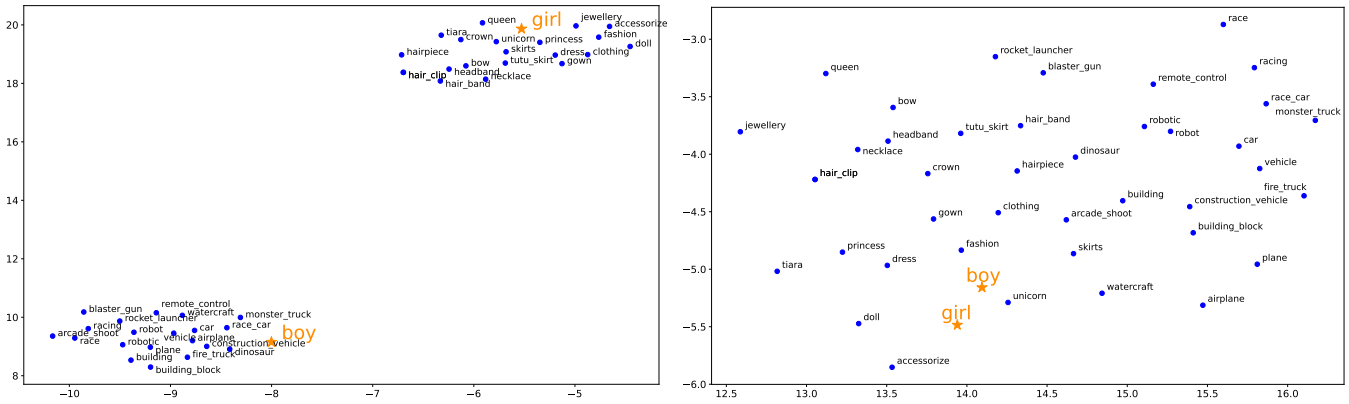
Since our approach is centered on gendered products, resulting embeddings represent products that explicitly mention gender or gender inclusive terms as defined in our dataset in Section 3. Therefore, our embeddings show a filtered view of differences between genders only in gendered products. We discuss this more in detail in Section 6.

5.1 Debiasing using gender neutral data

In this section, we evaluate the effect of using gender neutral product data as a means for pre-process debiasing. We additionally combine this with a hard-debiasing technique to measure further bias improvements.

In particular, we compare two different versions of product word embeddings. The first embedding, which will constitute the baseline for gendered products, consists of training an end-to-end embedding model using the *complete* gendered dataset for children's toys. Specifically for this model we use products in the classifications *"girls"*, *"boys"* and *"children's gender inclusive"*, defined in Section 3. Secondly, we train a so called *neutral* embedding, trained only on the gender inclusive portion of the same data.

In particular, we trained embeddings using the Gensim [63] implementation of Word2Vec [56], with 30 epochs, learning rate of 0.03 and vector sizes of 300. The pre-processing steps included,



(a) Embedding from all children's toys in our dataset

(b) Embedding from neutral children's toys in our dataset.

Figure 5: Product embeddings from children's toys data.

concatenation of each product's text to include title, description and additional textual details. Stop words and non-alphabetic characters were removed, and sentences were lemmatized using SpaCy. We then used the Gensim Phrases package to detect multi-word expressions (e.g., "social skill"). Note that there were no duplicates in the collection, since these were removed when creating the initial dataset.

In Figure 5a we present the visual result, using UMAP [53], of the embedding based on the complete gendered data. This visualization shows a comparison between the terms "boy" and "girl" in relation to their nearest products. In this figure, we can observe that there is a clear separation between products according to gender. Here, we can see that the term "girl" is very close to products with terms such as "queen," "jewellery," "unicorn," "princess" and "fashion", among others. The term "boy", on the other hand is closer to "vehicle," "monster truck," "race car," "construction," and "dinosaur".

Figure 5b, on the other hand, shows the terms for the embedding that was trained on the neutral data. In this case, we can observe that product terms do not show a clear gender based separation as before, and are not as clustered as in Figure 5a. More in detail, when measuring the differences between the similarities of skills and adjective keywords for "boys" and "girls," this was reduced to close to 0. For example, for "STEM toys"[8] the absolute difference between the similarities to "girl" and to "boy" was reduced from 0.169 to 0.039. For "Science toys", the difference also reduced, from 0.118 to 0.018.

To evaluate more comprehensively the fairness of both embeddings we use WEAT [20]. WEAT is a measurement of fairness of word embeddings, which works by measuring the relationship between two sets of *target* words (e.g., female words like "she," "her," "woman," "girl" vs. male words like "he," "him," "man," "boy") and two sets of *attribute* words (including terms related to skills, adjectives, products, occupations, etc.). The closer the measure is to 0, the less biased the model is. We use the WEAT implementation provided by the WEFE package [9].

[8]STEM: Science Technology Engineering and Mathematics.

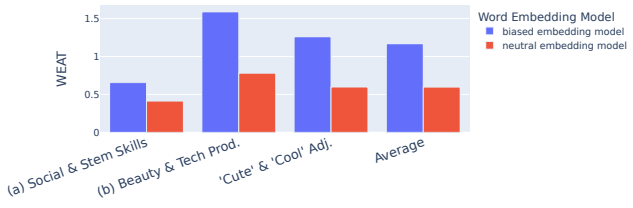


Figure 6: WEAT comparison for children's toys.

The evaluation of WEAT showed that, as expected, the neutral embedding was more fair than the first model trained on the entire gendered dataset. In particular, the range of differences between similarities of vocabulary words to "girl" and "boy" was reduced from (-0.227,0.334) in the gendered embedding to (-0.150,0.123) in the neutral embedding. Moreover, in the neutral model the similarity between the words "girl" and "boy" was increased from 0.795 to 0.877 (by 110%). Overall, the average WEAT measure was reduced from 1.17 to 0.60 (by ≈ 50%), shown in Figure 6. The chart compares the WEAT bias measure between the two models–with and without neutral pre-processing. The queries we used to assess the measure were: (a) social & STEM skills: using female terms (e.g, "girl," "she," "her") and male terms (e.g, "boy," "he," "his") as *target* keywords with respect to social skills terms (e.g., "creativity," "social skill") and STEM skills (e.g., "stem," "engineering") as *attributes*; (b) beauty & technical products: female terms and male terms with respect to beauty terms (e.g, "fashion," "jewellery") and technical terms (e.g., "robotic," "building block"); (c) 'cute' & 'cool' adjectives - female terms and male terms with respect to 'cute' terms (e.g, "cute," "lovely," "adorable") and 'cool' terms (e.g., "cool," "interactive" and "motorized").

Based on these results, the neutral embedding is able to achieve a more unbiased perception of the product data. Suggesting that there is a well established segment of gender inclusive products that can be leveraged for training balanced models on real concurrent non-manipulated data.

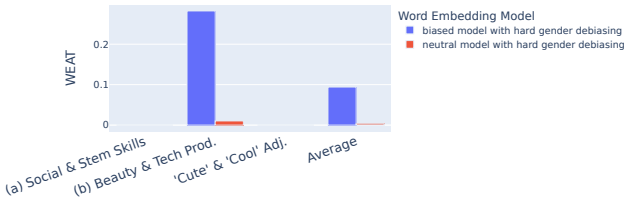


Figure 7: WEAT comparison with post-process hard debias.

Combining pre-processing and post-processing. We also combine the proposed pre-process technique with the post-process *hard-debiasing* [18] mitigation technique.Figure 7 shows the results. By applying hard-debiasing on the neutral embedding we are able to obtain more fair results (i.e., WEAT is lower), compared to applying each of the methods on their own. Note that although the combined technique achieves better fairness, the utility of adding hard-debiasing may be lower since this method distorts the resulting embedding. In this regard, using only the pre-process method has the advantage of being trained on a dataset of existing products which should not distort the utility as much.

6 DISCUSSION & LIMITATIONS

In this work we present a preliminary characterization of gender differences in e-commerce products. This analysis is useful to better understand gender representation in retail, as well as to inform design choices in e-commerce systems that rely on product data.

Due to the vast volume of data available as part of online products and given our goal of studying gender representation, we made the choice to focus our research on products explicitly to male and female genders. In addition to those products, we included products that were explicitly directed to both genders equally, or that intentionally used gender inclusive terms. We understand that the way in which our dataset was sampled has several consequences on our study. In particular, the portrayal of gender is limited to the gender binary notion that includes only male and female, and excludes existing additional gender classifications, such as nonbinary and transgender, just to name a few. In this regard, we view our work as part of initial studies towards understanding gender representation in e-commerce, which should further be extended in the future.

For simplicity, our dataset was restricted to products containing keywords that matched the terms described in Section 3, and to simple variations of those terms. This could be further expanded to include gendered products that can be identified using keywords not included currently in our filter, such as for example for "women" terms like "feminine," "lady," "girly," etc., which can also be found in product names. The selection of terms used, although not comprehensive, can be seen as a way to sample representatively the majority of products targeted to male and female genders.

Given that our focus was placed on gender characterization, our analysis selectively focuses on gendered items. Our coverage of gender inclusive products was limited only to some gender neutral terms for adults and children (i.e., "unisex", "children('s)", "kid(s/'s)", "toddler(s/'s)" and products that mentioned both genders at the same time). Gender inclusive products in our dataset were selected to contrast how language differed when products explicitly targeted both genders or declared being unisex. In this regard, many of products not considered in our study, which do not mention any gender at all, could potentially also be considered as gender inclusive. When considering all of online products, the products that are targeted to male and female genders is very small. Hence, our dataset very likely significantly undersamples gender-neutral products in relation to their actual prevalence in the complete universe of online products. However, the number of gender inclusive products in our data is significant and even larger than the number of male targeted products that we were able to find.

We also studied bias using word embeddings, presented in Section 5. We compared an embedding derived from children's products from our dataset, which included gendered *and* gender inclusive products, to another embedding based only on the gender inclusive portion of the data. In this particular analysis, we found the "neutral" embedding to have sufficient coverage of gendered products, allowing us to compare both embeddings. This suggests that gender inclusive data can be used as a resource to train (gender) unbiased embeddings. However, it should be noted that when considering the complete universe of online products, neutral data will probably not be able to provide sufficient coverage of all products. Despite this, we believe that gender inclusive products can be used as an alternative to improve highly biased portions of the data for training purposes. This may be a better alternative in some cases to data augmentation [81]–which refers to duplicating the data, replacing for example, "girl" with "boy" and vice versa–without the introducing artificial data that does not resemble reality (similar to what happened with Gemini image generation [61]).

7 CONCLUSIONS

In this paper, we presented a broad characterization of gender representation in a large set of gendered online products. We performed a comparative analysis of the language used in products related to genders, as well as the language related to products that are gender inclusive. We found that skills and occupations represented in gendered products tend to reinforce classical stereotypes.

We further observe that skills that are nurtured in toys that are marketed to a specific gender, are also manifested in gendered products for adults, showing early-age stereotypes reinforced in adulthood. We also found an important segment of products that are gender neutral or inclusive, which shows a segment of products using diverse and inclusive language.

Understanding stereotypical differences in data, such as those based on gender, is relevant when designing NLP tools and systems that interact with this information. In this work, we explored an approach of utilizing gender neutral products for debiasing word embeddings. More specifically, we showed how pre-processing the underlying data to include more neutral representations can improve the fairness of the resulting embeddings. In addition, we suggested a combined approach of pre-processing and post-processing debiasing method for fair embeddings, which improves embedding fairness even further. This work is only a first step since it is important to explore these techniques for other types of language models.

REFERENCES

[1] 2021. Assembly Bill No. 1084 Gender neutral retail departments. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202120220AB1084 (accessed: 2023-01-02).

[2] 2023. OpenAI Guides - Embeddings - Limitations & Risks. https://platform.openai.com/docs/guides/embeddings/limitations-risks (accessed: 2023-03-28).

[3] 2024. The Climat Pledge. https://www.theclimatepledge.com/us/en/Signatories (accessed: 2024-04-16).

[4] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[5] David Alfonso-Hermelo, Philippe Langlais, and Ludovic Bourg. 2019. Automatically learning a human-resource ontology from professional social-network data. In *Canadian Conference on Artificial Intelligence*. Springer, 132–145.

[6] Mohadeseh Amini and Parviz Birjandi. 2012. Gender Bias in the Iranian High School EFL Textbooks. *English Language Teaching* 5, 2 (2012), 134–147.

[7] Carol J Auster and Claire S Mansbach. 2012. The gender marketing of toys: An analysis of color and type of toy on the Disney store website. *Sex roles* 67 (2012), 375–388.

[8] Nor Jijidiana Azmi, Isyaku Hassan, Radzuwan Ab Rashid, Zulkarnian Ahmad, Nor Azira Aziz, and Qaribu Yahaya Nasidi. 2021. Gender stereotype in toy advertisements on social networking sites. *Online Journal of Communication and Media Technologies* (2021).

[9] Pablo Badilla, Felipe Bravo-Marquez, and Jorge Pérez. 2020. WEFE: The Word Embeddings Fairness Evaluation Framework. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*. International Joint Conferences on Artificial Intelligence Organization, 430–436. https://doi.org/10.24963/ijcai.2020/60

[10] Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. Tallrec: An effective and efficient tuning framework to align large language model with recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 1007–1014.

[11] Keqin Bao, Jizhi Zhang, Yang Zhang, Wang Wenjie, Fuli Feng, and Xiangnan He. 2023. Large language models for recommendation: Progresses and future directions. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*. 306–309.

[12] Adam Beam. 2021. California law requires gender-neutral area in some stores. https://apnews.com/article/business-gavin-newsom-california-state-legislature-legislature-6ee331cbf5eb7a22c046f5ed528b42f9 (accessed: 2023-01-02).

[13] Yahav Bechavod and Katrina Ligett. 2017. Learning fair classifiers: A regularization-inspired approach. *arXiv preprint arXiv:1707.00044* (2017), 1733–1782.

[14] Yahav Bechavod and Katrina Ligett. 2017. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044* (2017).

[15] Alison Bechdel. 1986. *The essential dykes to watch out for*. Houghton Mifflin Harcourt.

[16] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 610–623.

[17] Judith E Owen Blakemore and Renee E Centers. 2005. Characteristics of boys' and girls' toys. *Sex roles* 53 (2005), 619–633.

[18] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*. 4349–4357.

[19] Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2019. Understanding the Origins of Bias in Word Embeddings. In *International Conference on Machine Learning*. 803–811.

[20] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[21] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*. 3992–4001.

[22] L Elisa Celis and Vijay Keswani. 2019. Improved Adversarial Learning for Fair Classification. *arXiv preprint arXiv:1901.10443* (2019).

[23] Prafulla Kumar Choubey, Anna Currey, Prashant Mathur, and Georgiana Dinu. 2021. GFST: Gender-Filtered Self-Training for More Accurate Gender in Translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 1640–1654.

[24] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 797–806.

[25] Jamell Dacon and Haochen Liu. 2021. Does gender matter in the news? detecting and examining gender bias in news articles. In *Companion Proceedings of the Web Conference 2021*. 385–392.

[26] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated experiments on ad privacy settings. *Proceedings on privacy enhancing technologies* 2015 (2015), 92–112.

[27] Lisa M Dinella and Erica S Weisgram. 2018. Gender-typing of children's toys: Causes, consequences, and correlates. *Sex Roles* 79 (2018), 253–259.

[28] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. 2018. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency*. 119–133.

[29] Silvia Fareri, Nicola Melluso, Filippo Chiarello, and Gualtiero Fantoni. 2021. SkillNER: Mining and mapping soft skills from any text. *Expert Systems with Applications* 184 (2021), 115544.

[30] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 259–268.

[31] Joel Escudé Font and Marta R Costa-Jussa. 2019. Equalizing gender biases in neural machine translation with word embeddings techniques. *arXiv preprint arXiv:1901.03116* (2019).

[32] Luke Friedman, Sameer Ahuja, David Allen, Terry Tan, Hakim Sidahmed, Changbo Long, Jun Xie, Gabriel Schubiner, Ajay Patel, Harsh Lara, et al. 2023. Leveraging large language models in conversational recommender systems. *arXiv preprint arXiv:2305.07961* (2023).

[33] Megan Fulcher and Emily F Coyle. 2018. Working at play: Gender-typed play and children's visions of future work and family roles. (2018).

[34] Megan Fulcher and Amy Roberson Hayes. 2018. Building a pink dinosaur: The effects of gendered construction toys on girls' and boys' play. *Sex Roles* 79 (2018), 273–284.

[35] Hila Gonen and Yoav Goldberg. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 609–614.

[36] Emanuella Grinberg. 2015. Target to move away from gender-based signs. https://edition.cnn.com/2015/08/08/living/gender-based-signs-target-feat (accessed: 2024-04-16).

[37] Maarten Grootendorst. 2020. c-TF-IDF. https://github.com/MaartenGr/cTFIDF. (accessed: 2022-05-16).

[38] Satish Chandra Gupta. 2023. ChatGPT Alternatives That Deserve Your Attention. https://www.ml4devs.com/newsletter/021-chatgpt-google-bard-lamda-meta-llama/ (accessed: 2023-03-28).

[39] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.

[40] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. (2017).

[41] Lesley Istead, Andreea Pocol, and Sherman Siu. 2022. Evaluating Gender Bias in Film Dialogue. In *Natural Language Processing and Information Systems: 27th International Conference on Applications of Natural Language to Information Systems, NLDB 2022, Valencia, Spain, June 15–17, 2022, Proceedings*. Springer, 403–410.

[42] Thorsten Joachims. 1996. *A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization*. Technical Report. Carnegie-mellon univ pittsburgh pa dept of computer science.

[43] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 35–50.

[44] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*. 4171–4186.

[45] Marlene Kollmayer, Marie-Therese Schultes, Barbara Schober, Tanja Hodosi, and Christiane Spiel. 2018. Parents' judgments about the desirability of toys for their children: Associations with gender role attitudes, gender-typing of toys, and demographics. *Sex roles* 79 (2018), 329–341.

[46] Siyu Liao, Rongting Zhang, Barbara Poblete, and Vanessa Murdock. 2023. Bias Invariant Approaches for Improving Word Embedding Fairness. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (, Birmingham, United Kingdom,) *(CIKM '23)*. Association for Computing Machinery, New York, NY, USA, 1400–1410. https://doi.org/10.1145/3583780.3614792

[47] Lynn S Liben, Kingsley M Schroeder, Giulia A Borriello, and Erica S Weisgram. 2018. Cognitive consequences of gendered toy play. (2018).

[48] Kornelia Lipowska and Ariadna Beata Łada-Maśko. 2021. When parents go shopping: Perspectives on gender-typed toys among polish mothers and fathers from big cities. *Children* 8, 9 (2021), 744.

[49] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. 2016. The variational fair autoencoder. *International Conference on Learning Representations (ICLR)* (2016).

[50] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309* (2018).

[51] Chandler May, Alex Wang, Shikha Bordia, Samuel Bowman, and Rachel Rudinger. 2019. On Measuring Social Biases in Sentence Encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 622–628.

[52] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-Based Recommendations on Styles and Substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Santiago, Chile) *(SIGIR '15)*. Association for Computing Machinery, New York, NY, USA, 43–52. https://doi.org/10.1145/2766462.2767755

[53] L. McInnes, J. Healy, and J. Melville. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints* (Feb. 2018). arXiv:1802.03426 [stat.ML]

[54] Paola Medel and Vahab Pournaghshband. 2017. Eliminating gender bias in computer science education materials. In *Proceedings of the 2017 ACM SIGCSE technical symposium on computer science education*. 411–416.

[55] Aditya Krishna Menon and Robert C Williamson. 2018. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*. 107–118.

[56] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2* (Lake Tahoe, Nevada) *(NIPS'13)*. Curran Associates Inc., Red Hook, NY, USA, 3111–3119.

[57] Peter Mirski, Reinhard Bernsteiner, and Dania Radi. 2017. Analytics in human resource management the OpenSKIMR approach. *Procedia computer science* 122 (2017), 727–734.

[58] Dana Pessach and Erez Shmueli. 2022. A Review on Fairness in Machine Learning. *Comput. Surveys* 55, 3, Article 51 (feb 2022), 44 pages. https://doi.org/10.1145/3494672

[59] Alexandra Guedes Pinto, Henrique Lopes Cardoso, Isabel Margarida Duarte, Catarina Vaz Warrot, and Rui Sousa-Silva. 2020. Biased language detection in court decisions. In *International Conference on Intelligent Data Engineering and Automated Learning*. Springer, 402–410.

[60] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. (2019).

[61] Prabhakar Raghavan. 2024. Gemini image generation got it wrong. We'll do better. https://blog.google/products/gemini/gemini-image-generation-issue/ (accessed: 2024-04-16).

[62] Amifa Raj and Michael D Ekstrand. 2022. Fire Dragon and Unicorn Princess; Gender Stereotypes and Children's Products in Search Engine Responses. *arXiv preprint arXiv:2206.13747* (2022).

[63] Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic* 3, 2 (2011).

[64] Harriet L Rheingold and Kaye V Cook. 1975. The contents of boys' and girls' rooms as an index of parents' behavior. *Child development* (1975), 459–463.

[65] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender Bias in Coreference Resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. 8–14.

[66] Claude Sammut and Geoffrey I. Webb (Eds.). 2010. *TF–IDF*. Springer US, Boston, MA, 986–987. https://doi.org/10.1007/978-0-387-30164-8_832

[67] Lisa Selin Davis. 2021. Why gender-neutral holiday presents matter for your children. https://www.cnn.com/2021/12/20/health/gender-neutral-presents-wellness/ (accessed: 2024-04-16).

[68] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326* (2019).

[69] Tom Simonite. 2015. Probing the Dark Side of Google's Ad-Targeting System. https://www.technologyreview.com/s/539021/probing-the-dark-side-of-googles-ad-targeting-system/ (accessed: 2022-07-21).

[70] Rachel Treisman. 2021. Lego says it will work to rid its toys of harmful gender bias. https://www.npr.org/2021/10/12/1045244110/lego-toys-survey-gender-bias-stereotypes (accessed: 2023-01-02).

[71] Francesca Tripodi. 2021. Ms. Categorized: Gender, notability, and inequality on Wikipedia. *New Media & Society* (2021), 14614448211023772.

[72] Danielle Turchiano. 2019. Geena Davis Talks 'This Changes Everything' Doc and 'Conscious Gender Bias' in Behind-the-Scenes Hiring. https://variety.com/2019/film/features/geena-davis-this-changes-everything-documentary-interview-1203286574/ (accessed: 2023-03-28).

[73] Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2019. Getting gender right in neural machine translation. *arXiv preprint arXiv:1909.05088* (2019).

[74] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. 2017. Learning Non-Discriminatory Predictors. In *Conference on Learning Theory*. 1920–1953.

[75] Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdulnour, et al. 2024. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health* 6, 1 (2024), e12–e22.

[76] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1171–1180.

[77] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *Artificial Intelligence and Statistics*. 962–970.

[78] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International Conference on Machine Learning*. 325–333.

[79] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender Bias in Contextualized Word Embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 629–634.

[80] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

[81] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. 15–20.

[82] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning Gender-Neutral Word Embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 4847–4853.